Automatic Voice Disorder Recognition using Acoustic Amplitude Modulation Features

by

Nicolas Malyska

B.S., Electrical Engineering
B.S., Computer Engineering
University of Florida, 2000

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2004

Signature of Author..
Department of Electrical Engineering and Computer Science
February 2, 2004

Certified by ....
Thomas F. Quatieri
Senior Member of Technical Staff; MIT Lincoln Laboratory
Faculty of MIT SHBT Program
Thesis Supervisor

Accepted by..
Prof. A. C. Smith
Chair, Department Committee on Graduate Students
Department of Electrical Engineering and Computer Science

This work was sponsored by the Air Force Research Laboratory, under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Automatic Voice Disorder Recognition using Acoustic Amplitude Modulation Features

by

Nicolas Malyska

B.S., Electrical Engineering
B.S., Computer Engineering
University of Florida, 2000

Submitted to the Department of Electrical Engineering and Computer Science
on February 2, 2004 in partial fulfillment of the
Requirements for the Degree of Master of Science in
Electrical Engineering

ABSTRACT

An automatic dysphonia recognition system is designed that exploits amplitude modulations (AM) in voice using biologically-inspired models. This system recognizes general dysphonia and four subclasses: hyperfunction, A-P squeezing, paralysis, and vocal fold lesions. The models developed represent processing in the auditory system at the level of the cochlea, auditory nerve, and inferior colliculus. Recognition experiments using dysphonic sentence data obtained from the Kay Elemetrics Disordered Voice Database suggest that our system provides complementary information to state-of-the-art mel-cepstral features.

A model for analyzing AM in dysphonic speech is also developed from a traditional communications engineering perspective. Through a case study of seven disordered voices, we show that different AM patterns occur in different frequency bands. This perspective challenges current dysphonia analysis methods that analyze AM in the time-domain signal.

Thesis Supervisor: Thomas F. Quatieri
Title: Senior Member of Technical Staff; MIT Lincoln Laboratory
Faculty of MIT SHBT Program

Acknowledgements

First, I would like to thank Tom Quatieri, my thesis advisor, for sharing his time and energy with me for the last year and half. I appreciate that Tom encourages new ideas and paradigms and creates a research environment where we can be creative without getting lost. Tom, I look forward to working with you in the coming months and years—thanks.

I would like to thank Cliff Weinstein and Marc Zissman for allowing me to work with such an inspiring group of teammates at Lincoln Laboratory. I'm excited about continuing into the future as a part of Group 62.

Finally, I would of course like to thank my family and friends for all of their support and encouragement. Thanks for your insistence that I do what it takes to get what I want out of life. Thanks to my dad, for his insightful questions and my mom for always being there to talk. To my brother William, an enlisted Marine: thanks for helping me to keep things in perspective and please come home safely. To Becky: thanks for your encouragement, patience, and caring through the last few months.

Contents

Chapter 1

Introduction

The ability to recognize characteristic voice qualities is an intriguing human trait. With this ability, we can obtain information such as a speaker's identity, state of health, and degree of fatigue by listening to only several seconds of speech. The acoustic properties that convey these often subtle elements are only starting to be understood. This project is motivated by the desire to develop features which capture these *rich* voice qualities.

In comparison to the many voice factors that human listeners exploit, the elements used by automatic recognition systems in speech technology are surprisingly limited. It is the hypothesis of this thesis that current automatic recognition systems—in speech, speaker, and language recognition for example—are not designed to take advantage of the rich properties of the human voice. One area of automatic recognition that has only recently begun to emerge is automatic speech disorder, or *dysphonia*, recognition. A dysphonia is a disorder of the speech production mechanisms in the larynx with perceptual, acoustic, and physical correlates. Examples of these disorders include excessive tension of the laryngeal muscles and the presence of abnormal masses of tissue on the vocal folds.

For the investigation of voice quality features, using dysphonic voices evaluated in an automatic voice disorder recognition system is a natural choice. These voices, which are often distinctly rough, hoarse, or breathy, provide acoustic evidence that a person is not well. The problem of dysphonia recognition is particularly interesting because, unlike the speaker recognition problem, it is largely dependent on speaker differences in the glottal source, rather than differences in the vocal tract resonances. Dysphonic speech also may represent the *extremes* of acoustic phenomena occurring in normal voices such as the irregular nature of glottalization [38]. At least on a first pass, the baseline for our experiments, mel-cepstral features, are usually not thought to well represent source characteristics [34]. Therefore, the disordered voice recognition task is a challenge to the ability of our models to represent characteristics of the glottal source.

We hypothesize that dysphonic voices will provide examples of acoustic phenomena more subtly present in normal voices. By improving features to represent these nuances, especially periodic fluctuations in the amplitude envelope, or *amplitude modulations*, we hope to improve the overall representation of both normal and pathological voices in automatic recognition systems. Teager and Titze highlight the importance of understanding the speech signal as a glottal source carrier modulated by physiological inputs such as muscle movements, vortices of air, and the motion of laryngeal tissues. Titze, for example, states that at present "we don't know how to measure or classify the multiplicity of perturbations and modulations that are observed simultaneously." [46, 47] This thesis explores models that accurately capture a subset of such modulation phenomena.

1.1 Thesis Outline

1.1.1 Chapter 2

This work begins by rigorously defining the term dysphonia—its perceptual and acoustic nature and also its origins in the larynx. We address two main topics (1) the aural perception of a voice disorder and (2) changes in the larynx that create voice disorders. The question of acoustics and perceptual cues are discussed in terms of a review of the literature, building a preliminary taxonomy of voice qualities that are analyzed acoustically for amplitude modulations in chapter 3. Likewise, we organize the physical properties of voice disorders into a hierarchy that describes how the physiology of these patients differs from normal subjects and from one another.

Also in this chapter, voice quality and the physiology of dysphonia are connected. We discuss evidence that certain combinations of acoustic properties correlate with certain physical voice disorders. A review of ideas and methods in clinical practice and engineering design from the literature provides strong motivation for automatic dysphonia recognition. The chapter ends with a discussion of one commercially available dataset, the Kay database [1], for studying dysphonia. We describe how this speech corpus is organized along with several problems that are inherent to its structure.

1.1.2 Chapter 3

As the second chapter defines dysphonia, chapter 3 defines amplitude modulation. The organization is a progression from traditional models of AM synthesis and analysis in communications engineering, to a theoretical relation of these concepts with speech signals, and finally to an analysis of AM in speech. Throughout this chapter, we derive and reinforce the concepts that (1) a bandpass analysis of speech signals is equivalent to demodulation, (2) a band-dependent analysis of speech can yield different patterns of fluctuations at different frequencies, (3) analysis filter bandwidth is critical to the analysis process, and (4) amplitude modulation in speech depends on both frequency and amplitude relationships of spectral components. These points are highlighted by presenting the most extreme cases from an acoustic survey of over 350 dysphonic voices.

1.1.3 Chapter 4

With the concepts of dysphonia and amplitude modulation defined, chapter 4 introduces three biologically-inspired models for the extraction of AM patterns. This chapter addresses the motivation, design, and implementation of each of these approaches, showing that the different models have complementary properties. We compare and contrast the responses of the models using a combination of synthetic signals and real speech.

1.1.4 Chapter 5

Chapter 5 discusses the motivation, design, implementation, and testing of a Gaussian-mixture-model-based dysphonia recognition system. A system capable of recognizing five voice problems—general dysphonia, hyperfunction, anterior-posterior squeezing, and several types of vocal folds lesions—is introduced. Extending the work of chapter 4, we show the process by which features are extracted from the auditory models, how classification is performed with a Gaussian mixture model, and how recognition results are fused. The final portion of the chapter shows the results of recognition experiments with the disordered speech database presented in chapter 2.

1.1.5 Chapter 6

This chapter draws conclusions about AM in speech and our dysphonia recognition system based on these principles. In particular, we discuss key findings of the research with a particular focus to how they may be improved in the future.

1.1.6 Appendix A

The appendix describes a dysphonic speech-synthesis experiment with the Klatt formant synthesizer [23] motivated by a study in the literature [2]. Sustained vowels for three of the patients analyzed in this thesis, DAS10, KAH02, and JAB10 are synthesized and the resulting signals are compared to the original utterances. We discuss differences observed between normal and dysphonic voices that make synthesis challenging.

1.2 Thesis Contributions

There are a number of areas in which this work is different from previous approaches. First, an explicit amplitude modulation detection technique is applied to automatic dysphonia recognition. This research also refines some of the biologically-inspired methods on which it is based [7, 36], aiming to create improved features for AM extraction. While implied by previous research [14, 46, 47], this thesis introduces the first specific use of frequency-band-dependent AM in dysphonic speech. We do not focus on modulations visible in the time-waveform; rather, through bandpass analysis, we view modulations in each frequency band separately.

The dysphonia recognition experiments of this thesis are also different in that we present results for a large range of voice disorders: normal/pathological cases, paralysis, hyperfunction, A-P squeezing, and several types of growths on the vocal folds. The notable exception is work performed on paralysis by [10]. In our work we also present, and attempt to work around, several limitations of the Kay database that we have not found previously discussed in the literature. Since we also rigorously define how our experiments are constructed, a possible spin-off of this work may be a more standard approach selecting data is when testing dysphonia recognition.

Finally, with regard to recognition, our results suggest that our biologically-inspired features provide speech characteristics complementary to standard mel-cepstrum. A comparative analysis is presented of the relative performance of the features derived from variations of the auditory models and also illustrates advantages gained by optimally fusing recognition results from different extracted features.

Chapter 2

Dysphonia

A major part of this work addresses the automatic detection of voice disorders, or *dysphonia* in speech. A dysphonia is a disorder of the mechanisms of phonation in the larynx caused by elements of diseased tissue, misuse and/or abuse of the speech production mechanisms, and psychological and/or stress factors [30]. Most patients have not only one, but a combination of problems making up to a dysphonia, some of which involve a "vicious cycle" [49] of further complications in the larynx. Voice disorders manifest as a combination of acoustical voice qualities in tandem with physical changes in the larynx.

In this chapter, we first introduce motivation for general acoustic and perceptual voice categories derived from the literature. Then we define different physical voice disorders that affect the larynx including aspects of how they arise, the effect they have on speech production, and how they relate to one another. We provide evidence for a correlated relationship between these two aspects of dysphonia. We also provide background information on the design and contents of the database used in dysphonia recognition experiments.

2.1 Acoustic and Perceptual Characteristics of Voice Quality

A preliminary taxonomy for voice quality derived from the literature is described below with reference to Figure 2-1. Each section describes a particular term, its synonyms, where it has been studied, and what it means in terms of acoustic and perceptual characteristics. It is important to realize that this taxonomy is only a basic effort to bring together many different terms and fit them together; a well agreed upon taxonomy for the speech community is still very much the subject of ongoing research [14]. There are surely further distinctions beyond the branches of the tree explored (for example encompassing breathy phonation). Also the taxonomy categories should not be taken as too rigorous; many combinations of the categories, for example a mix of modal and breathy speech or the coexistence of breathy and glottalized voice qualities, are possible.

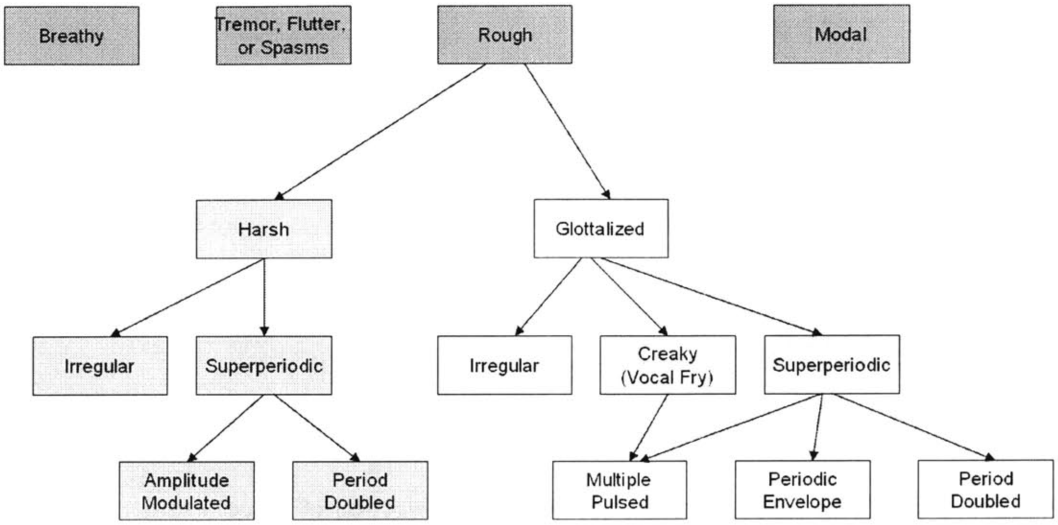

Figure 2-1. A preliminary taxonomy for voice quality. Note that, although this tree is derived from reports in the literature, a well agreed upon categorization system is still an open problem in the field.

2.1.1 Modal Voice

Modal speech is loosely defined as the "usual or baseline kind of phonation" [14]. It is speech at a normal fundamental frequency for the sex of the speaker, regular in both amplitude and pitch period, and without the nonmodal characteristics defined in the rest of the taxonomy [24]. Normal voices can be described as a mix of varying degrees of modal and nonmodal qualities [27]. Acoustically, "the volume-velocity waveform or area function produced by a modal model is zero during the closed phase, and its first derivative has a discontinuity at the moment that closure occurs"[16]. Also, in the ideal model, "the spectrum falls off at 6 dB/octave" [16].

2.1.2 Rough Voice

Roughness is the percept, or mental state generated by a physical observation, thought to be associated with an "uneven, bumpy quality" [47]. As depicted in Figure 2-1, roughness may be broken down into two perceptual subcategories, harsh and glottalized voice, primarily dependent on the fundamental frequency [27]. As will be shown, acoustically roughness can have many component aspects. Rough speech may contain aspects of all its subcategories, thus making some utterances difficult to describe. It is important to realize that some aspects of roughness may seem "irregular," but we will show that many such signals do, in fact, have a periodic variation.

2.1.3 Breathy Voice

The percept of breathiness has to do with the "sound of breathing (expiration in particular) during phonation" [47]. Acoustically, breathiness manifests as a decreased harmonic energy-to-noise ratio and an increase in the energy of the first harmonic [24] and usually contains a significant voiced component. The physiology behind such acoustics are relatively well-studied and have to do with the tension of the glottis as well as the area which remains open during the glottal cycle.

2.1.4 Voice with Tremor, Flutter, and Spasms

Acoustically, tremor, flutter, and spasms are low frequency variations in the amplitude of speech on the order of 4-13 Hz [30]. As will be seen in the next section, these sounds are highly linked to several neurological speech disorders that cause periodic and aperiodic variations in the amplitude of the speech waveform. This category is different from the amplitude modulation subcategory of glottalization and harsh voice in that it occurs at much lower frequencies. As we will see in Chapter 3, it is not uncommon to have variation in the fundamental frequency accompany this activity.

2.1.5 Glottalized Voice

Perceptually, glottalized or laryngealized voice includes "salient auditory impressions of glottal gesture, roughness, or creakiness" [38]. Acoustically, its main distinction is that it has a dramatically lower frequency (averaging about 50 Hz) than both modal voicing and harsh voice [5]. Glottalized speech can be broken down into three (primarily acoustic) components—period-to-period irregularity, creak, and superperiodic phonation. Glottalized voices have been shown to consist of varying combinations of these three elements [14, 38].

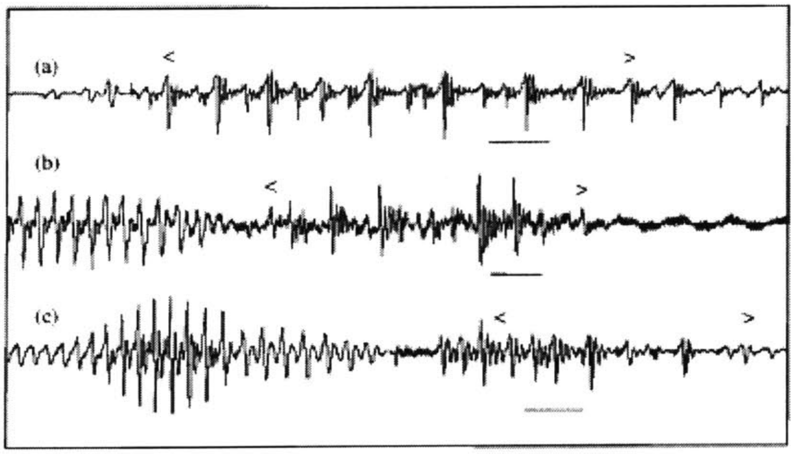

Figure 2-2. Examples of the aperiodic category of glottalization from Redi and Shattuck-Hufnagel [38]. Activity is indicated by the region within the angle bracket; the bar indicates 10 ms.

In the literature, for example, Redi and Shattuck-Hufnagel divide glottalization into four categories: (1) *aperiodicity* or "irregularity in duration of glottal pulses from period to period," (2) *creak* or "prolonged low fundamental frequency accompanied by almost total damping", (3) *diplophonia* or "regular alteration in shape, duration, or amplitude of glottal periods", and (4) *glottal squeak* or "a sudden shift to relatively high sustained *f0*, which [is] usually very low amplitude" [38]. Figures 2-2, 2-4, and 2-6 show characteristic examples of the first three types; each exhibits distinct amplitude fluctuations with time. Note from the taxonomy figure that we include similar subdivisions, using slightly modified names which are defined shortly. For a complete review of glottalization and different perceptual, acoustical, and physiological correlates as well as other possible categorization schemes, see [15].

2.1.6 Harsh Voice

Harshness is defined by Laver as an unpleasant, rasping quality of speech [27]. Acoustically speaking, it is loosely characterized as having glottal pulses which decay much more quickly than those for modal voices [24]. Harshness is primarily different that glottalization in that it has a significantly higher (close to modal) fundamental frequency [27]. Although subcategories of harshness have not been specifically studied, the taxonomy of rough voice quality with near modal fundamental frequency has been described as including both irregular and superperiodic components [14]. We have thus chosen this term to refer to any roughness at or above the fundamental frequency, but there is no solid motivation for doing so.

2.1.7 Irregular Voice

Acoustically, signals that are irregular vary in amplitude and/or period from one pitch period to the next [38]. This term differs from "superperiodic" in that there is no time pattern to the varying components. Terms for "irregular" in the literature include "aperiodic" [38], "rough" [11], and "creaky" [38], the last two terms having additional meanings in our taxonomy. For harsh voice, irregularity has been shown by [14] as indicated in Figure 2-3.

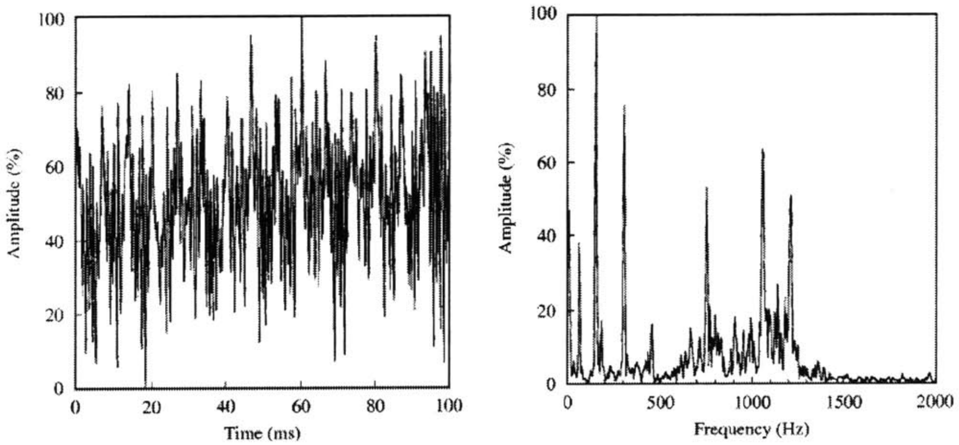

Figure 2-3. Figure from [14] showing an example of aperiodic behavior in a normal voice. Observe that there does not exist a clear fundamental and there seems to be significant "interharmonic noise" [14] in the spectrum.

2.1.8 Vocal Fry or Creak

Vocal fry is the element of sharp, short glottal pulses in voicing that creates the perception of low frequency individual pulses in speech [27], [5]. Some authors refer to the entire category of glottalization as vocal fry, but it has been deemed more descriptive to include separate aperiodic and superperiodic components . Creak may also be combined with superperiodicity to create a multiple pulse behavior whereby there are several pulses in quick succession followed by extended inactivity [14].

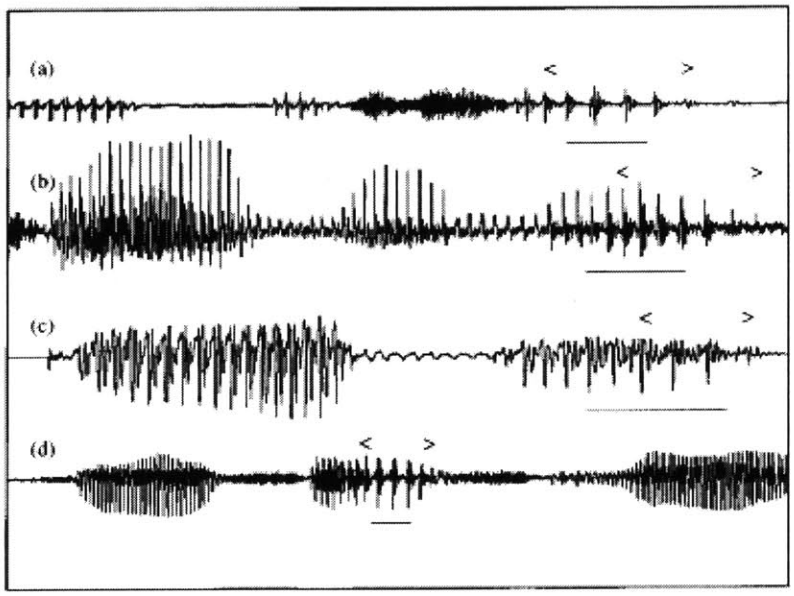

Figure 2-4. Examples of the creak category of glottalization from Redi and Shattuck-Hufnagel [38]. Activity is indicated by the region within the angle bracket; the bar indicates 10 ms.

2.1.9 Superperiodic Voice

Superperiodic speech is defined as having a "repeating pattern that extends over more than one apparent glottal cycle" [14]. This category is analogous to the diplophonic component of glottalization described in [38] and shown in Figure 2-5. [14] shows an example of further decomposition of superperiodic speech into periodic envelope and period doubling types. Perceptually, these qualities are "described as 'rough' or 'bitonal'". Other subcategories of superperiodicity include period tripling and quadrupling as well as the multiple pulsing often found in vocal fry [14]. As this category corresponds to clear amplitude variation with time, we will investigate it further in the chapter 3.

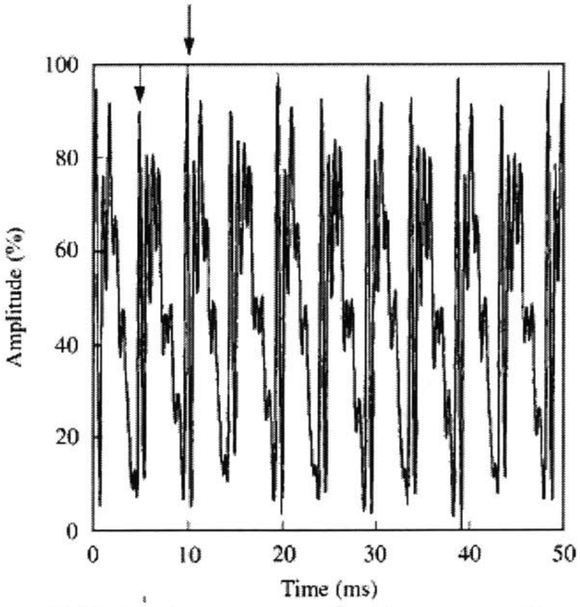

Figure 2-5. Example from [14] showing an example of superperiodic period-doubling behavior in a normal voice.

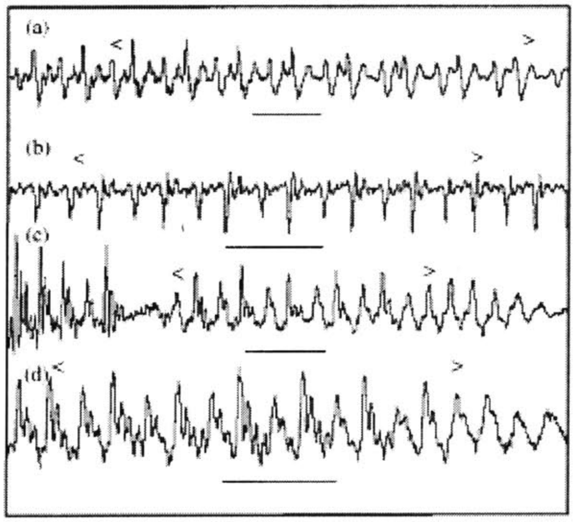

Figure 2-6. Examples of the diplophonia category of glottalization from Redi and Shattuck-Hufnagel [38]. Activity is indicated by the region within the angle bracket; the bar indicates 10 ms.

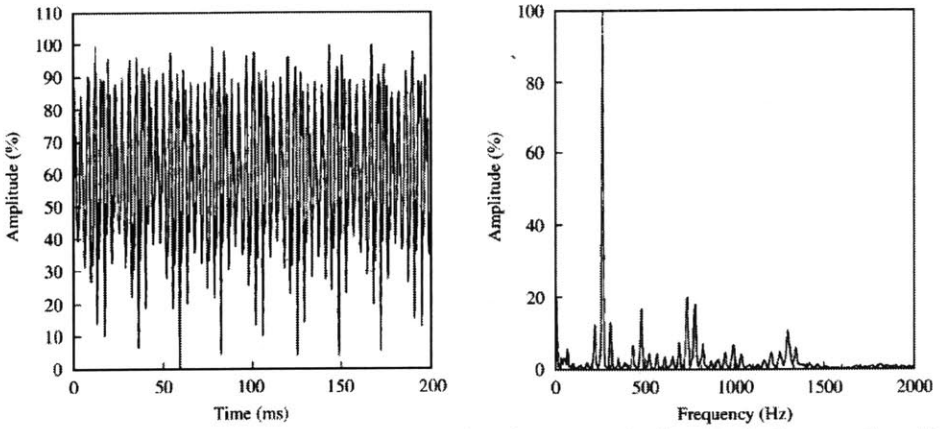

Figure 2-7. Figure from [14] showing an example of superperiodic "periodic envelope" behavior in a normal voice. The right panel shows harmonics of the fundamental frequency along with subharmonics spaced by 44 Hz defining the frequency of the envelope.

Garratt and Kreiman mention that physical evidence exists for a periodic envelope category through the action when "the two vocal folds vibrate at different (but close) frequencies" [14] shown in Figure 2-7. There have also been attempts to model this from a multiple vibrational mode point of view due to nonlinearities in the glottal source mechanism. In a similar fashion, Berry [3] proposes such a model of vocal fold vibration, multiple modes, supported by actual glottal vibration experiments depicted in Figure 2-8. His work supports there being several characteristic frequencies of the vocal folds, which in a normal voice, quickly move towards to one frequency, the fundamental. In nonmodal phonation, however, including superperiodic vibration, he is able to show that the so-called *eigenmodes* of vibration converge to different frequencies. In period doubling and glottalization, these stable states are multiples of a common frequency, whereas for irregular vibration, the modes are not harmonically related [3]. In other words, Berry is suggesting that the notion of a single fundamental frequency, belonging to a single source is a fallacy and that one can explain more complicated voice types using interacting vibrations.

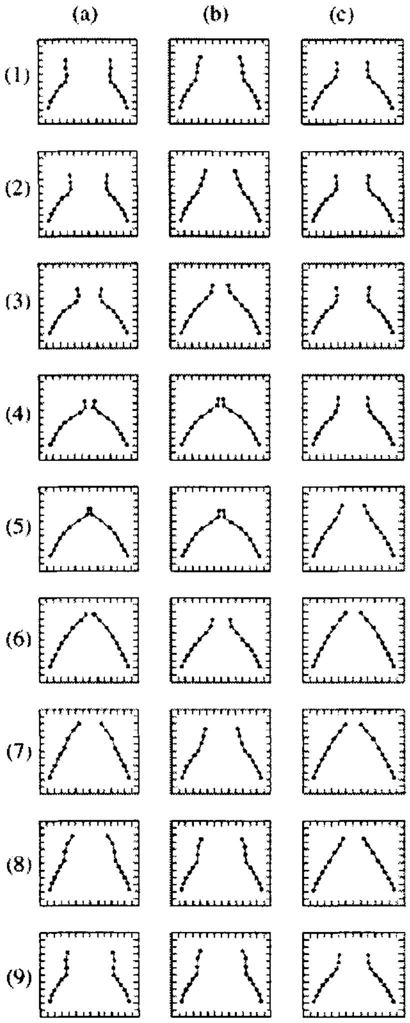

Figure 2-8. Progression of three simultaneous modes (a), (b), and (c) of vocal fold vibration observed in a human larynx removed from a body of work by Berry [3]. Each cross-section is taken in the coronal plane of the layryx with the dots separated by approximately 1 mm. Berry hypothesizes that eigenmodes as in this example combine to produce both modal and nonmodal voice properties.

2.2 Physiological Manifestation of Voice Disorders

Having introduced how voice qualities, many of which are present with dysphonia, sound and how they are characterized acoustically, this section presents physical aspects of dysphonia. The following describe several major classes of voice disorders, their putative causes, and interrelationships. Our discussion will follow a the schematic of interrelationships between clinical terms shown in Figure 2-9. As with the previous section, this picture is not meant to be rigorous, but it allows visualization of some of the complex interrelationships we will address.

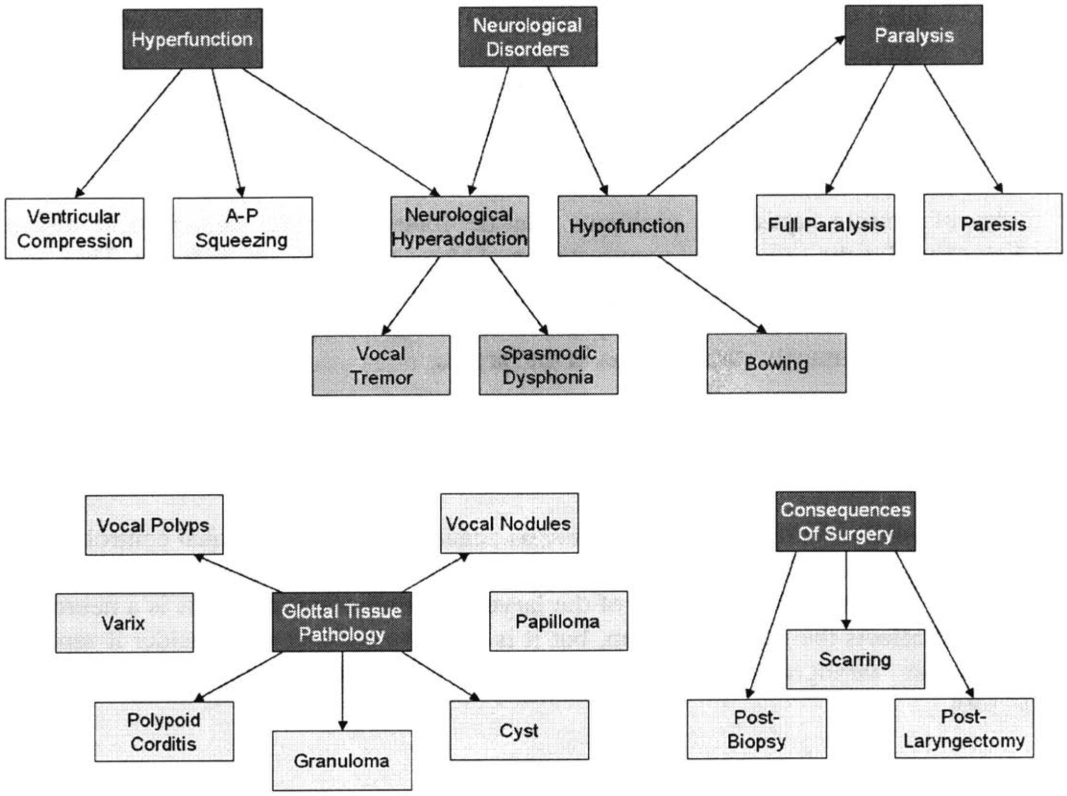

Figure 2-9. Organizational chart to guide discussion of the physical aspects of dysphonia. The arrow indicates that the node is well-described as a subclass of its parent.

2.2.1 Hyperfunction

This class of disorder is caused when excessive tension is present in the muscles of speech production, causing primarily the vocal folds to be too forcefully adducted [30]. Clinicians can observe certain specifics about hyperfunction during their stroboscopic investigation: "muscle tension dysphonia can be suspected when excessive glottic and supraglottic contraction can be identified during phonation by video laryngoscopy" [49]. Clinicians may note the general location of the tension—in *A-P squeezing*, for example, the tension is in a front-back direction, whereas for ventricular compression, the false vocal folds are squeezed together during phonation.

Hyperfunction is not usually considered to be the result of a conscious effort on the part of the speaker. That is, muscle-tension dysphonia is considered a *psychogenic* voice disorder; it comes about as the result of disordered emotions such as stress [30]. As stated by Wilson *et al.*: "when long-standing and severe, muscle tension dysphonia can lead to vocal cord structural lesions including vocal cord polyploid change (Reinke's edema), vocal cord polyps, vocal cord ('singer's') nodules, arytenoid ulcers and granuloma." [49]. Even before the development of such lesions, this class of disorders may negatively impact a patient by causing irritation of the larynx, which tends to promote even more tension to produce a normal voice.

One might also link abuse and misuse of the voice such as strained vocal style and screaming into this category. Such voice states differ, though, in that they are under the conscious control of the speaker. It is likely, however, that after the damage caused by initial laryngeal trauma, a subject will begin to exhibit some hyperfunction in order to compensate for the slight difficulty with phonation. For example, after cheering loudly at a football game, a fan may have slight laryngeal trauma. Instead of resting his voice, he continues to speak and must strain his voice in order to be heard. Such strain, in turn, yields even more trauma. It should be noted that patients with hyperfunction tend to respond well to voice therapy.

Other forms of muscle squeezing problems can be found with the neurological disorders discussed shortly. But the causes and manifestation of these disorders are different. They do not tend to respond in the same way to voice therapy as do stress and misuse caused by hyperfunction. They also tend to be more prone to cause spasms, whereas these hyperfunction disorders seem to cause mostly static tension in the larynx.

2.2.2 Paralysis

Paralysis and its relative, partial paralysis or *paresis* , causes a loss of positional control of one or both vocal folds [49]. Paralysis is actually best described as a subclass of *hypoadduction*, the abnormal lack of tension in the structures of the larynx. By definition, paralysis is a neurological disorder since it affects the nervous system, but it is so prevalent that we consider it separately. Minifie notes that "laryngeal nerve paralysis is one of the most common neurological voice disorders" [30].

2.2.3 Glottal Tissue Pathology

These problems manifest as masses, irritation, and/or swelling in the region on and surrounding the vocal folds. Masses, such as nodules, polyps, cysts, granulomas, or tumors, are thought to cause difficulties with phonation by interfering with the proper motion of the vocal cords, often preventing complete closure and also causing irregular vibration. Irritation and swelling, such as that caused by stomach acid reflux disorders, alcohol use, or smoking may also interfere with the proper operation of the vocal fold tissues and may even lead to the creation of more advanced lesions [16, 30, 49]. Misuse/abuse habits such as screaming, cheering, talking loudly, coughing, using a pressed voice are also a problem: "vocal nodules, vocal polyps, and contact ulcers are examples of voice disorders associated with vocal misuse/abuse and vocal hyperfunction" [30].

There exist far too many types of tissue pathology to offer a complete list. However, some of the most prevalent are:

(1) *Vocal nodules*. Nodules are like calluses on the vocal folds, often occurring bilaterally. They are due to trauma due to straining and misusing the voice. It is not unusual for nodules to be found in the same patient as hyperfunction, as one tends to lead to the other [49].

(2) *Vocal fold polyps*. Polyps are the "focal structural manifestation of chronic vocal cord irritation" [49]. As fluid filled lesions, they appear like blisters on the cords and impede the proper vibration of the vocal folds. A polyp may fill with blood and become hemorrhagic.

(3) *Polyploid coditis (Reinke's edema)*. This disorder leads to severe hoarseness with the vocal folds developing swelling and redness; the cords look "floppy and bag-like" [49].

(4) *Cyst*. A cyst appears as a round mass on the vocal fold and arises from "an obstructed mucous gland or as a result of a congenital lesion" [49].

(5) *Varix.* A varix is an enlarged vein on the vocal fold which can impede proper vibration of the folds.

(6) *Papilloma.* This disease is viral in origin, forming "a benign, tumor-like condition that may recur throughout childhood" [30].

(7) *Granuloma.* Granulomas "represent an exuberant focal granular response to laryngeal truama, such as intubation trauma or chronic microtrauma during phonation in patients with certain vocal styles" [49]. A granular response is defined as one resulting in the build-up of tissue.

2.2.4 Neurological Disorders

These problems are caused by disease at different levels of the nervous system. As a general class of disease, neurological disorders often affect many systems in the body but can specifically "affect the larynx and respiratory system and be reflected in a disordered voice." [30] Neurological disorders can occur at many levels of the speech chain—motor neuron, muscle, and in the brainstem and brain. The causes of the problems can include brain trauma, stroke, or a wide variety of chronic illnesses. As it has a tendency to affect normal movement of the articulators, a neurological disorder is also often an "entire motor speech disorder" [30], or *dysarthria.* In general, the effects of neurological disorders can be subdivided into abnormal excessive activity, or *hyperadduction*, and abnormal lack of activity, or *hypofunction.*

Paralysis is probably the most common type of hypofunction, although others, such as lack of tone causing bowing of the vocal fold, are also prevalent. Disorders that cause hypofunction often yield a breathier voice quality as attaining full closure of the glottis is difficult. "Hyperadduction occurs in many neurological disorders of phonation" [30]; chorea, tics, and dystonia are common examples of the manifestation of hyperadduction. These disorders are involuntary. Spasmodic dysphonia is a characteristic example of a dystonia causing hyperadduction and appears as a "choked, strain-strangled phonation and, in some cases, vocal tremor" [30]. Another affliction, Parkinson's disease—causes *tremor*, or low frequency amplitude variations in the voice, as a symptom. "Imprecise articulation and disordered rate are observed in patients with Parkinson's disease" [30]. Tremor is in range of 4-8 Hz. Higher rates (10-13 Hz) exist for some other neurological disorders and are called flutter. [30]

2.2.6 Consequences of Surgery

Surgery involving the mechanisms of voice production may also change a patient's voice dramatically. With cancer, for example, whole sections of the larynx must be biopsied or removed entirely. This can understandably affect voice quality positively or negatively. One complication of surgery for our purposes is that it can repair voice quality while a condition remains. A good example of this is with paralysis on one side, which can be aided by moving the paralyzed vocal fold inward surgically. Thus the patient continues to be diagnosed with paralysis, but with a much improved voice quality [49].

2.3 The Connection between Dysphonia and Voice Quality

Having presented basic overviews of what the rich qualities of speech sound like as well as the types of physical changes that make up voice disorders, the question arises of how to connect the two areas. A major portion of this thesis centers around the assumption that the two are correlated.

2.3.1 Clinical Usage

Our first indication that speech perception—and consequently acoustics—is useful for predicting dysphonia is that clinicians commonly use the sound of a voice to determine the pathology. In a clinical environment, patient examinations encompass many different aspects including a medical history, video stroboscopic investigation, subjective evaluation of voice quality, and objective acoustic measurements [49]. Often, a change in voice quality is the first indication of a problem, bringing a patient to the clinic for the first time. Perceptual evaluation is so commonly used that most hospitals use forms including scales to describe a patient's voice. For example, the popular GRBAS voice rating scale allows a clinician to listen to a voice and rate it using five perceptual rating scales—general evaluation, roughness, breathiness, asthenicity, and strain [9]. Although there is some debate as to the actual efficacy of scales like this, the fact remains that listening to voice is a large part of the clinical process.

2.3.2 Existence of Objective Tools Mapping Acoustics to Dysphonia

In addition to the link between perceived voice quality and voice disorders, there is a movement to integrate *objective acoustic measures* into clinical practice. These methods analyze speech automatically and provide the clinician with a numerical value related to the severity of the voice. Although there are many variations available in the literature, a large number of clinically used objective acoustic measurements for dysphonic voices fall loosely into one of two different categories—perturbation measures and glottal noise measures. Perturbation measures such as shimmer and jitter measure the irregularity of the speech signal from one pitch period to the next, usually in the time domain. In contrast, glottal noise measures, such as harmonic-to-noise ratio (HNR), attempt to separate out the speech signal into two components, a harmonic component and a noise component. These latter class of techniques has been attempted in both the spectral and time domains [19, 29]. Three common measures are further described below:

(1) *Jitter.* Jitter is defined as a "short term (cycle-to-cycle) perturbation in the fundamental frequency of the voice" [47]. Several automatic and hand-assisted techniques exist to measure jitter in sustained phonation [20, 29]. These methods are currently used in both clinical and speech research settings and are of comparable reliability to perceptual methods for normal speakers but less so for pathological speakers [4, 37]. Jitter is usually thought to correspond well with the percept of roughness, but has also been linked with breathy and hoarse phonation types [11].

(2) *Shimmer.* Shimmer is cycle-to-cycle change in the amplitude of successive voice periods [47]. As with jitter, both automatic and hand-assisted techniques have been created to measure shimmer in sustained phonation [21, 29].

(3) *Harmonic-to-Noise Ratio.* This measure represents the amount of energy coming from periodic parts of the speech signal compared with the energy of aperiodicity, as coming from aspiration for example. Several specific algorithms exist but all are based on this idea [19, 29]. Although harmonic-to-noise ratio is reviewed more in terms of its

importance for the percept of breathiness [18, 24], it may also be related to roughness [11].

It is also unclear exactly how useful existing automatic techniques are for classifying specific types of dysphonia. Work has been done to produce techniques including the hoarseness diagram described in [12] and the Dysphonia Severity Index [50] which are capable of better combining the three automatic methods described above. These approaches seem to yield numbers that can differentiate *physical* larynx properties, but the work to date is limited to only certain voice problems such as paralysis and post-surgical recovery. Another limitation as discussed in [32] is that most of the previous work with these objective perturbation features has been done with sustained vowels rather than continuous running speech. Although voiced portions of speech can be extracted prior to processing [12], most perturbation measures are still relatively sensitive, relying on accurate estimates of the fundamental frequency. Thus, there are issues with measuring modulation patterns for even moderately irregular signals, such as those seen with many dysphonic cases.

Overall, it appears probable that there is a connection between the acoustics of the human voice and the condition of the larynx. In later chapters we will work to build features that focus on time variation of amplitude and attempt to use them to predict physical diagnoses.

2.4 The Kay Disordered Voice Database

The clinical voice database being studied in this thesis, called the Kay Disordered Voice Database [1] includes a wide variety of clinical diagnoses and physician comments to accompany each of the speech samples. This database was generated from recordings taken during clinical visits by on the order of 600 patients to the Massachusetts Eye and Ear Infirmary. Both continuously sustained 1 second vowel utterances of /a/ as well as continuous speech in the form of 12-second recordings of the *rainbow passage* are included. The rainbow passage is a diagnostic sentence which begins "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch...."

The Kay database includes a set of data including patient age, sex, smoking habits, visit date, as well as a group of diagnoses and notes taken by a clinician reviewing the case. Because of the way in which they were created—through physician review of files—the notations range from extremely general to very specific. It is often the case that one note will be a synonym of the others. For example, *hyperfunction*, the abnormal overuse of the speech production muscles might be noted for the same patient as *A-P squeezing*, which is a specific type of hyperfunction involving compression in the forward-backward direction.

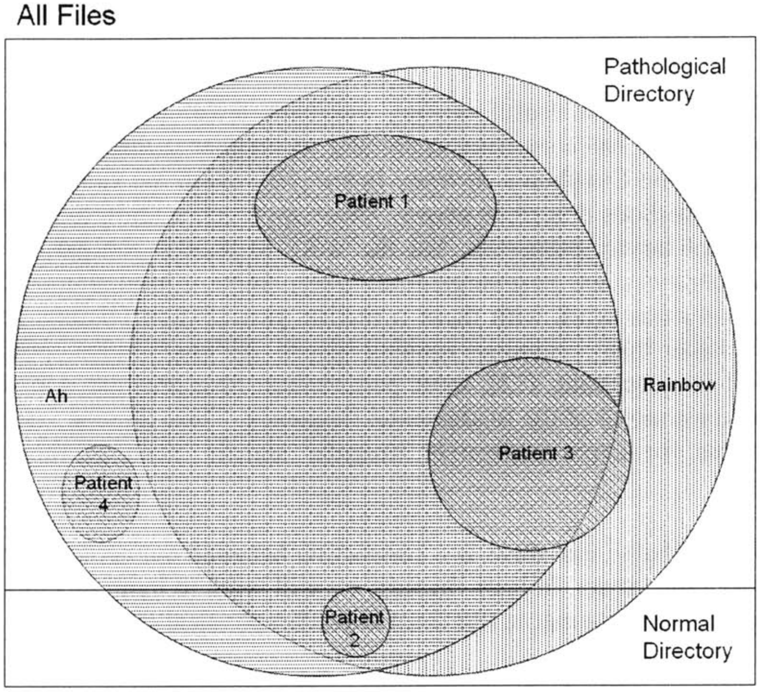

Figure 2-10. Diagram of patients in the Kay database. Patients can have multiple files, each from a different visit date. As depicted, the patient may be normal or pathological and also may have Rainbow passage and/or vowel recordings for each file.

How meaningful and specific a certain voice classification also varies. At one end of the range, there are observable physical problems which either exist or do not exist—for example, total paralysis or surgically removed anatomical structures. There are also disorders which can be determined with high confidence through physical examination but can differ in degree of pathology. For example, vocal nodules, similar to calluses on the vocal folds, can be diagnosed after years or just as they are forming. Unfortunately, the Kay database being used does not have a good system for indicating severity for most conditions, including all cyst patients into the same category and all partial paralysis, or *paresis*, cases into the same group. At the other end of the range are behavioral conditions that may exist to some degree even in normal patients.

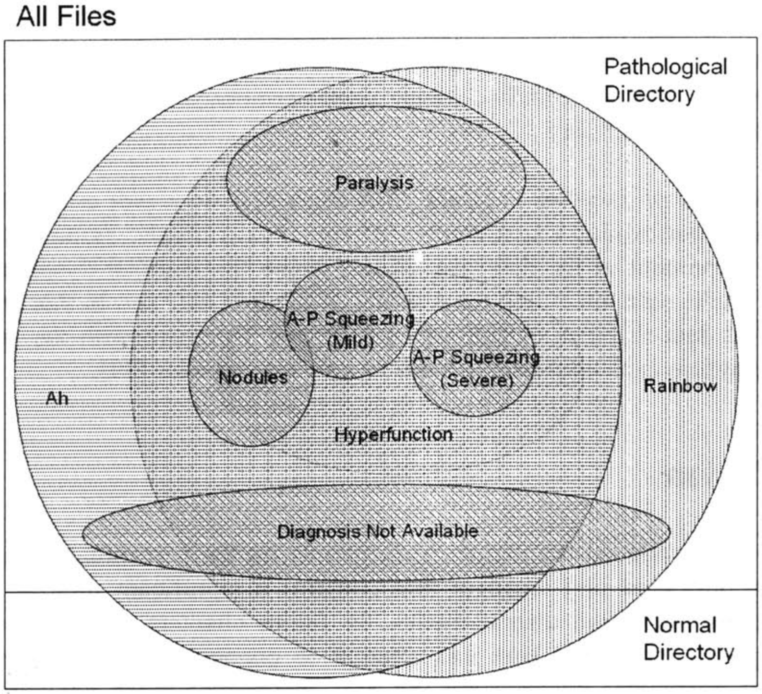

Figure 2-11. Diagram of showing the typical organization of diagnoses found in the Kay database. Each file can have multiple diagnoses, even to the extent that having one diagnosis implies another one.

Figures 2-10 and 2-11 detail the organization of the database and show the nature of patients and diagnoses within this structure. First, a single patient can have multiple files and each may or may not contain utterances from the rainbow passage and the vowel /a/. The two main *pathological* and *normal* directories do not have overlapping patients. As can be seen, diagnoses can overlap extensively, even entirely. Some patients do not have a diagnosis or accompanying information at all and are shown in Figure 2-11 as "Diagnosis Not Available."

For this thesis, we cannot change the methodology used to record the clinical diagnoses and the descriptions in the database. We can, however, take into account the generality and meaning of different disorders when choosing classes for automatic recognition. First, we can keep in mind those general categories used to define voice disorders—hyperfunction, paralysis, tissue pathology, neurological disorders and, surgical results. We can also look to previous recognition research in order to determine what sets of data are most appropriate to study. On the whole, there are a number of characteristics of the Kay database that could affect our study:

(1) *Different recording site for normals.* Normals are recorded at least partly at different sites than the pathological voices. This introduces the possibility of slight differences in the recording environment and microphones used. Unfortunately, we do not implement a direct control for this possibility, however channel normalization techniques are used.

(2) *Not enough cases.* Another consideration with the database that there are not enough of them—between 700 and 800 total—recorded to make strong statistical conclusions in dysphonia recognition. The problem is worse for certain voice classes; there are only 53 normal speakers, for example.

(3) *Reliability of physician diagnosis.* It is not known how reliable physician diagnoses and comments are for each specific diagnosis category. As discussed above, certain diagnoses also have an inherent gradient which the database indicates poorly, leading to speculation that accurately noting them is difficult. Other diagnoses, such as paralysis, seem easier to diagnose in a formal way.

(4) *Large overlap of dysphonia classes.* This characteristic is part of the nature of dysphonia and makes performing experiments difficult. For example, assume that one dysphonia affects 90 percent of the database. Then assume that another dysphonia also affects 90 percent of the data, including the residual of the first data set. In this example, there is no set of test data that lies outside of both groups. Compare with the speaker recognition problem where each utterance belongs to only one speaker [39].

2.5 Conclusion

In this chapter we introduced the nature of dysphonia on the perceptual, acoustic, and physical levels. We also presented one database that contains a significant number of test cases and reviewed its organization and characteristics. In future chapters, we refer extensively to the aspects of dysphonia presented here and attempt to connect them to other ideas. Chapter 3 analyzes several dysphonic voices from the Kay disordered voice database using amplitude modulation theory. We continue this analysis with biologically-motivated designs in chapter 4 and, finally, in chapter 5, we experiment with automatic machine recognition of dysphonia.

Chapter 3

Amplitude Modulation

This chapter serves to define the term *amplitude modulation*, or *AM*, as it relates to dysphonia. First we motivate a definition for amplitude modulation in speech from classical sinusoidal AM definitions presented in communications engineering. Specifically, we describe signal synthesis in AM communications systems as well as the corresponding demodulation process. We then present a model for how speech differs from ideal sinusoidal AM and how this complicates the demodulation problem.

In this chapter, we also analyze amplitude modulations in dysphonic speech, showing evidence for their existence and basic nature. First, our study shows that both the frequency spacing and relative amplitudes of sinusoidal spectral components influence amplitude modulations. We find that the frequency of envelope variation often varies with the region of the spectrum being analyzed. We additionally present evidence for two other phenomena—amplitude modulated noise sources and the transduction of frequency modulation to amplitude modulation. This exploration indicates that dysphonic signals often contain a wealth of amplitude modulations. In later chapters, we develop models that better characterize the frequency-band-dependent envelope variations in dysphonia.

3.1 Amplitude Modulation of Sinusoids

Amplitude modulations have been studied from a communications viewpoint as a means to efficiently send signals from one location to another through transmission lines and radio waves. As depicted in 3-1, these systems allow multiple bandlimited signals to be displaced in frequency so that they can share the same transmission channel. This is a form of *frequency division multiplexing*, the simultaneous transmission of multiple signals by assigning each a frequency band. The limited commodity in a communications system is often the total bandwidth, the full spectrum allocated to a specific application. When using frequency-division multiplexing, an AM system allows multiple sources—radio stations for instance—to be broadcast from a source to a receiver at the same time

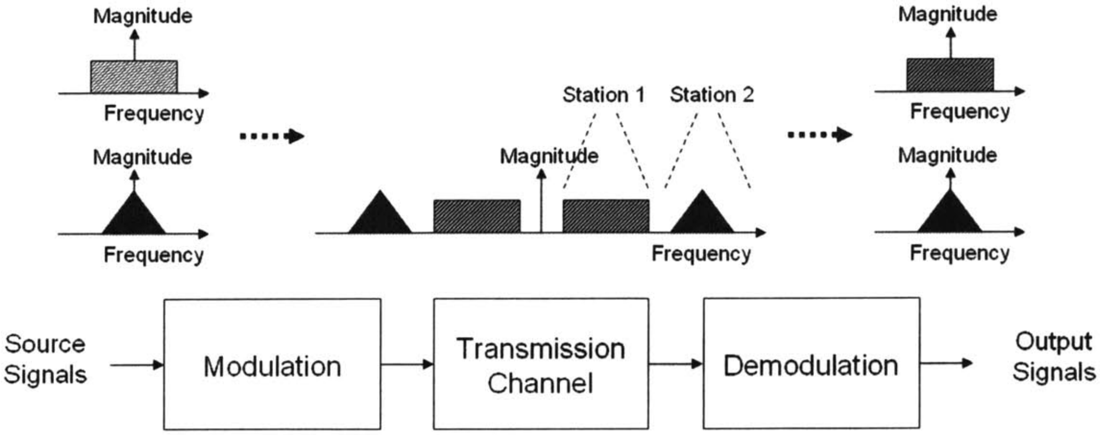

Figure 3-1. Schematic drawing of an AM transmission/receiver system. The idea behind such a system is to allow multiple transmissions to share the same transmission medium, for example radio waves or a wire.

3.1.1 Synthesis Model for AM Sinusoids

In general, the idea behind synthesizing amplitude modulations in a communications system is to take a series of *bandlimited* signals and transpose their spectra to higher frequencies. This process amounts to modulating the amplitude envelope of a sinusoidal carrier by the source signal [6].

In mathematical form, amplitude modulation of a single bandlimited signal is then defined as [6]:

$$g(t) = A_c[1 + m(t)]$$

$$s(t) = A_c[1 + m(t)]\cos(\omega_c t)$$

where $g(t)$ is the envelope multiplied by the cosine carrier with radian frequency w_c to create the modulated signal $s(t)$. The bandlimited source signal, $m(t)$, is used to create the envelope by adding it to 1 and scaling the sum by a constant A_c. Here $m(t)$ is defined to be between -1 and 1 such that $g(t)$ always remains positive. Also, ω_c is assumed to be no less than twice the highest frequency component of the original signal. The process of synthesizing AM is depicted in Figure 3-2, where a 30 Hz signal is used to modulate a 500 Hz carrier. The envelope of the output fluctuates, or *beats*, with time at 30 Hz.

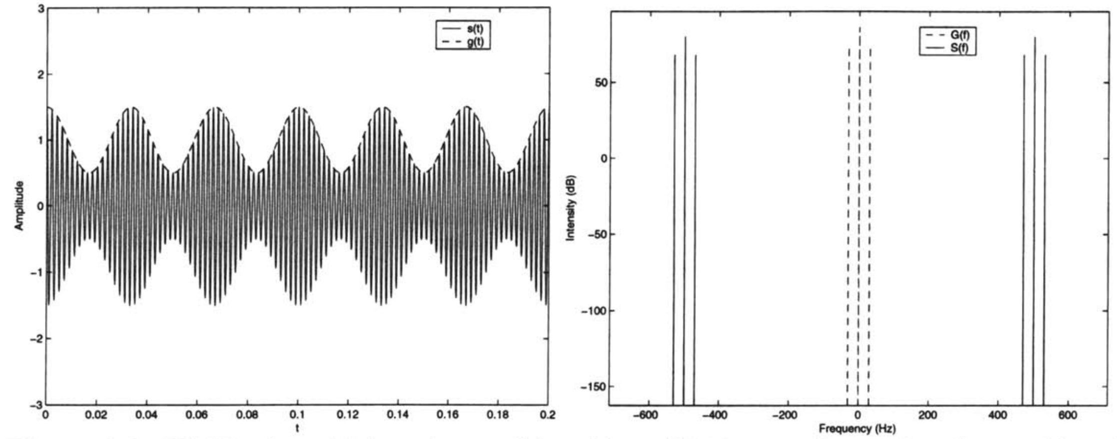

Figure 3-2. 500 Hz sinusoidal carrier modulated by a 30 Hz waveform *g*(*t*). The resulting time waveform *s*(*t*) beats at 30 Hz, the frequency imposed by the envelope. In the frequency domain, the original spectrum is simply shifted by the carrier frequency.

As can be seen in the figure, this interpretation of amplitude modulation is as a multiplicative envelope on a sinusoidal carrier. Another interpretation of the same waveform is as three line components in the frequency domain—one due to the carrier and the others due to the *sidebands*. This results from the equivalency of AM to convolution in the frequency domain. In a typical AM system, many modulated signals are summed together to allow them to transmit them on the same channel without frequency overlapping. Figure 3-3 shows both the previous 500-Hz carrier modulated by 30 Hz summed with a 700-Hz carrier modulated by a 50-Hz modulator.

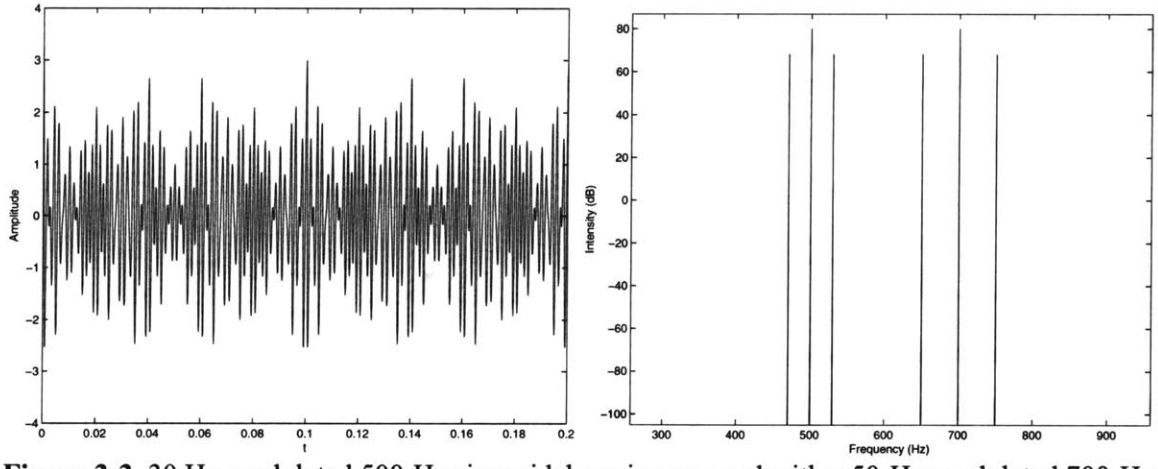

Figure 3-3. 30 Hz-modulated 500-Hz sinusoidal carrier summed with a 50-Hz-modulated 700-Hz sinusoidal carrier. Although the time envelopes are obfuscated in the time domain, the original spectra remain separated in the frequency domain.

3.1.2 Analysis Model for AM Sinusoids

A signal can be amplitude modulated then by multiplying it by a sinusoid at a carrier frequency. In order to reconstruct the original carrier signal, one needs to bring the center of the signal spectrum back to zero frequency, also known as *demodulating to baseband*. In communications systems containing many signals broadcast simultaneously, each modulating a different carrier, the strategy is to (1) isolate all frequencies affected by a certain carrier using one or more bandpass and lowpass filters and (2) remove the effect of the carrier using one of several methods of *detection* [6]. This general approach, depicted in Figures 3-4 and 3-5, assumes that there is *no overlap* between the spectra of neighboring channels before demodulation.

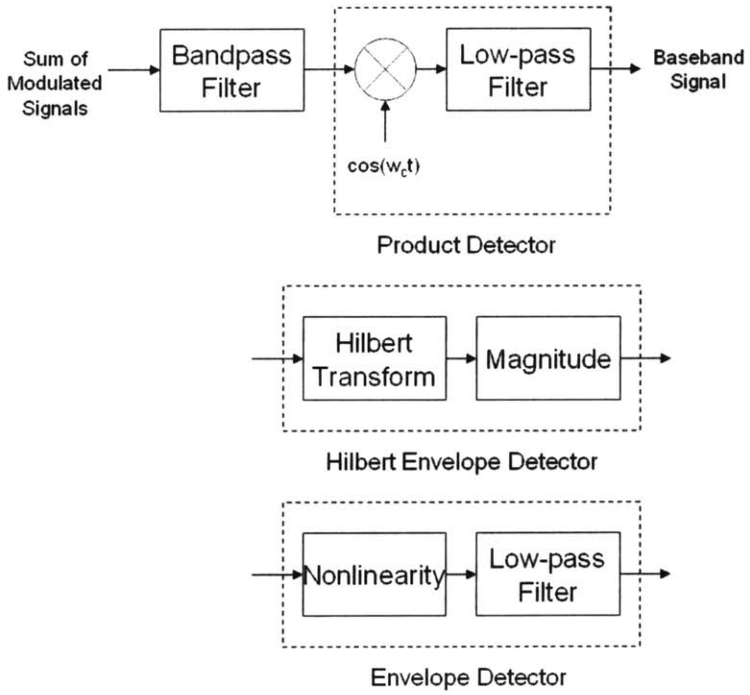

Figure 3-4. Diagram of demodulation of a sum of modulated sinusoids into a baseband signal using a combination of bandpass filter and one of several detector types.

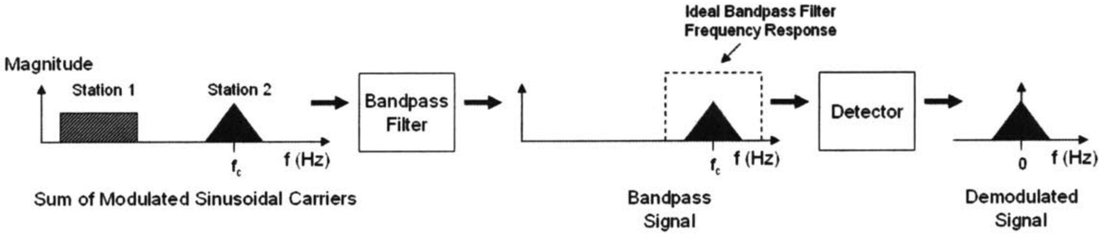

Figure 3-5. Schematic drawing showing the demodulation of two modulated sinusoidal carriers.

As shown in Figure 3-4, there are many types of detection including the product detector, the Hilbert envelope, and even simple rectification and low-pass filtering. The goal of an ideal detector is to perfectly recover *g*(*t*) from the bandpass signal *s*(*t*) [6]. Certain detectors require finding the correct carrier frequency or knowing it *a priori*—this is called coherent detection. With the use of a product detector in an AM radio, for example, part of this process of *tuning* often relies on a human to choose the proper carrier frequency using a frequency dial. Envelope detectors, on the other hand, employ noncoherent detection, whereby the carrier frequency does not need to be known in order to recover the original signal.

Thus, the process of demodulation in general is one of (1) separating by bandpass filtering an interesting portion of the signal and (2) bringing the signal to baseband using a detection scheme. This series of steps should be familiar from many well-studied speech analysis methods, including the spectrogram. A perspective of the short-time Fourier transform as a series of demodulators is shown in Figure 3-6. This representation is governed by[31]:

$$X[n,\lambda) = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m}$$

where $0 \leq \lambda < 2\pi$ denotes the analysis frequency in radians per second and $w[m]$ is the windowing function.

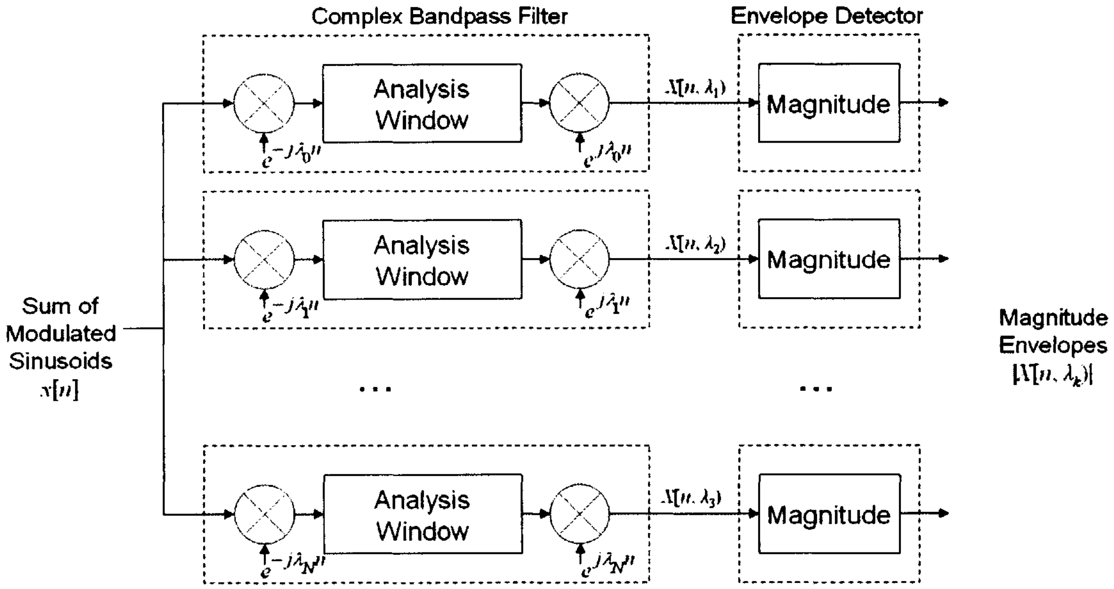

Figure 3-6. View of the short-time Fourier transform as a series of demodulators. Adapted from [31].

A problem with using such general spectral analysis techniques for the purpose of demodulation is that there is no guarantee that the filters will pass all the sidebands of a particular station. Thus we run into a well-known fundamental tradeoff—long analysis windows improve capturing the sidebands of a signal while short windows allow interference from the sidebands of neighboring channels. Figure 3-7 compares the spectrograms using three different window sizes to demodulate two amplitude-modulated sinusoids. With the longest envelope, the signal is not demodulated as the sidebands are each interpreted as carriers. With the middle envelope, we get demodulation at 30 Hz and 50 Hz as desired. The envelope of the outputs using the shortest window case contains amplitude variations that beat at both the modulation frequencies as well as at the difference frequency between sideband and carrier components of two different stations. This observation is connected to the well known rule-of-thumb that there is beating in the envelope of a signal at *difference frequencies*. That is, with overlapping modulation channels, two different types of interactions can produce magnitude variations on the output of an analysis channel. One interaction is due to amplitude modulation of a single sinusoidal carrier as expected from the theory. The other type is due to components of two different stations beating together in a phenomenon we will call *artifact AM*. From a synthesis point of view, artifact AM is defined as AM not resulting from the application of envelope $g(t)$ to a sinusoidal carrier; instead it results from *summing* two amplitude-modulated carriers. From an analysis point of view, the difference between these two cases is often difficult or impossible to determine—both will be denoted as

"AM" in the analysis performed in this thesis, regardless of their origin. It is important to realize that time variation of the resulting envelope occurs for both cases.

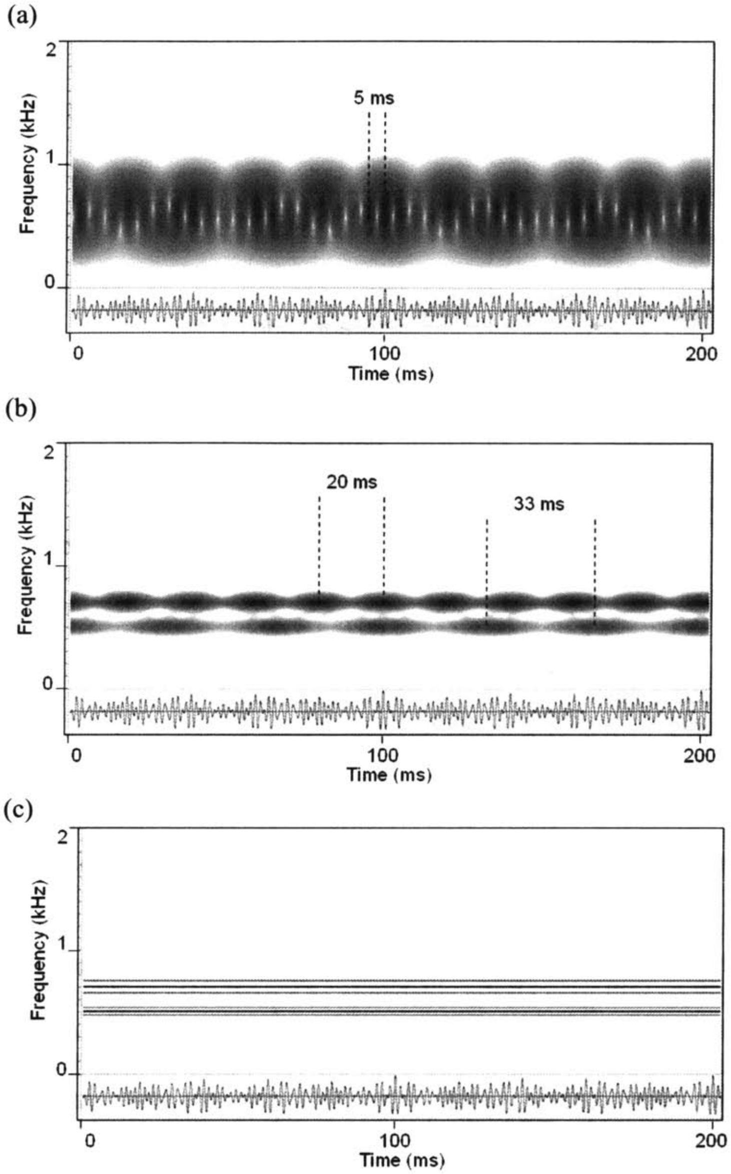

Figure 3-7. Illustration of the effect of changing the size of the Fourier analysis window on demodulation performance. Notice that time-variation in the output envelopes can occur at frequencies other than the two modulation frequencies. From top to bottom, the Hamming window lengths are (a) 5 ms, (b) 16.6 ms, and (c) 200 ms, respectively.

3.2 Amplitude Modulations in Voice

Throughout the discussion of voice thus far, we have been hesitant to use the term *amplitude modulation* to describe the time-variation observed in real speech signals. The purpose of this section and the next is to connect the theory of sinusoidal AM presented above to the properties of real voice. In the previous chapter, we described a significant body of work on voice quality, often showing evidence of time-varying amplitude patterns. Further, we have shown evidence from the literature that these acoustic properties influence what a human perceives—that is, that time variations of amplitude are important to whether we perceive a particular speaker as, for example, breathy, rough, or hoarse. Examples were presented that indicate that observed properties are in fact connected to actual physical changes of the larynx accompanying voice disorders.

Based on the synthesis techniques described in the first section, one could hypothesize a model of the human voice as created by a series of summed amplitude-modulated sinusoidal sources with non-overlapping bandwidths. At a basic level, such a model must be able to generate speech, as we know from the mathematics of Fourier synthesis that *any* real signal can be created as the sum of an infinite number of constant sinusoidal components. There is also an attraction to a model that views sources and their amplitude modulations as separate actions, much in the way a source and a filter are viewed as separate blocks in traditional speech modeling [22].

The human voice, however, is almost surely not well modeled in this way. First, the physical sources of sound produced by the glottis and by turbulent airflow in the vocal tract are for the most part not well described as single sinusoids. For example, it is typical to represent the glottal excitation as a harmonic sequence of sinusoids, and aspiration and frication sources as wideband noise sources. Also the carrier sources may be varied with respect to frequency as well as time. Although we did not present the theory, such a frequency modulated, or FM, signal is not beyond the capabilities of many communications systems, but adds complications that are beyond the scope of this thesis. Suffice to say that the Hilbert envelope detector in previously described models allows AM and FM to be separately demodulated [34, 35].

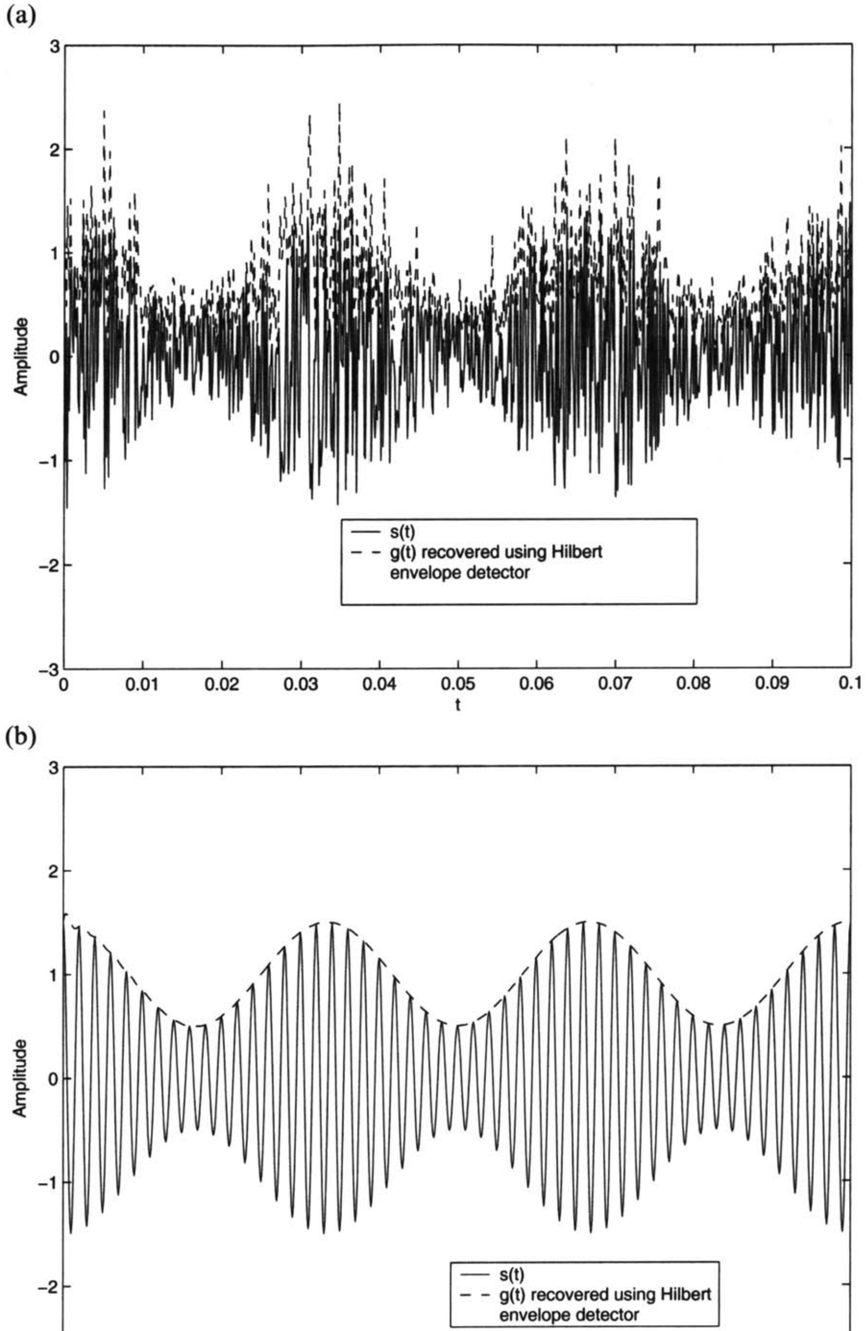

Figure 3-8. Irregular fluctuations in the magnitude of the Hilbert envelope of uniformly distributed white noise modulated at 30 Hz (a) compared with the envelope for a 30-Hz modulated sinusoid (b). Notice that a noise carrier has a noisy envelope due to beating between its own components.

In addition to the problem of complex carrier spectra, there is also almost surely an overlap between the carriers and sidebands of physiologically plausible speech sources. For instance, in the production of a voiced fricative like /v/, both the noise source and the voicing source have spectral components across a large range of frequencies and overlap significantly. As shown in the AM theory discussion, such phenomena create envelopes that vary at the difference frequencies of the interacting components. When sources are non-sinusoidal, there may also be modulation between components *within* the carrier [7]. As depicted in Figure 3-8 for a 30-Hz modulated uniformly-distributed white noise source, these interactions appear as corruption of the envelope.

A final difference between our ideal AM sine-wave model and real speech is that most speech models represent the vocal tract as an additional filterbank. The poles of this set of filters, known as the formant frequencies, determine the overall shape of the spectrum created by sounds produced at the glottis. Additionally, sources arising at other points in the vocal tract undergo other filtering determined by their location in the vocal tract [43]. The time-varying nature of this filterbank combined with the movement of the carrier frequencies, allow for the possibility of amplitude modulations generated by changes in the relation of spectral components with the filters. As we will see in the next section, acoustics consistent with this phenomena exist in dysphonia.

Many factors, then, including the unknown number and spectra of the original sources, interactions between the carriers and sidebands belonging to each channel, and interactions with the vocal tract filters make the general demodulation problem for speech a difficult and currently unsolved problem. Thus, the work presented later in section 3.3 does not seek to present the demodulation of speech into amplitude-modulated physiologically-plausible sources. Rather, it proposes to connect the acoustics of voice quality presented in the previous chapter with AM patterns seen using bandpass analysis. This investigation differs from previous work in the literature in that it acknowledges that *different amplitude modulation behavior can occur in different regions of the spectrum*. That is, we argue that spectrally separate modulation patterns may tend to become hidden when viewing a complete time-domain waveform. Figure 3-9 shows an example of this phenomenon using a sum of five AM sinusoids—patterns which are obscured in the time domain appear in different frequency bands in the spectral domain. By performing a demodulation step—the short-time Fourier transform—first, we are able to analyze AM behavior on a band-by-band basis.

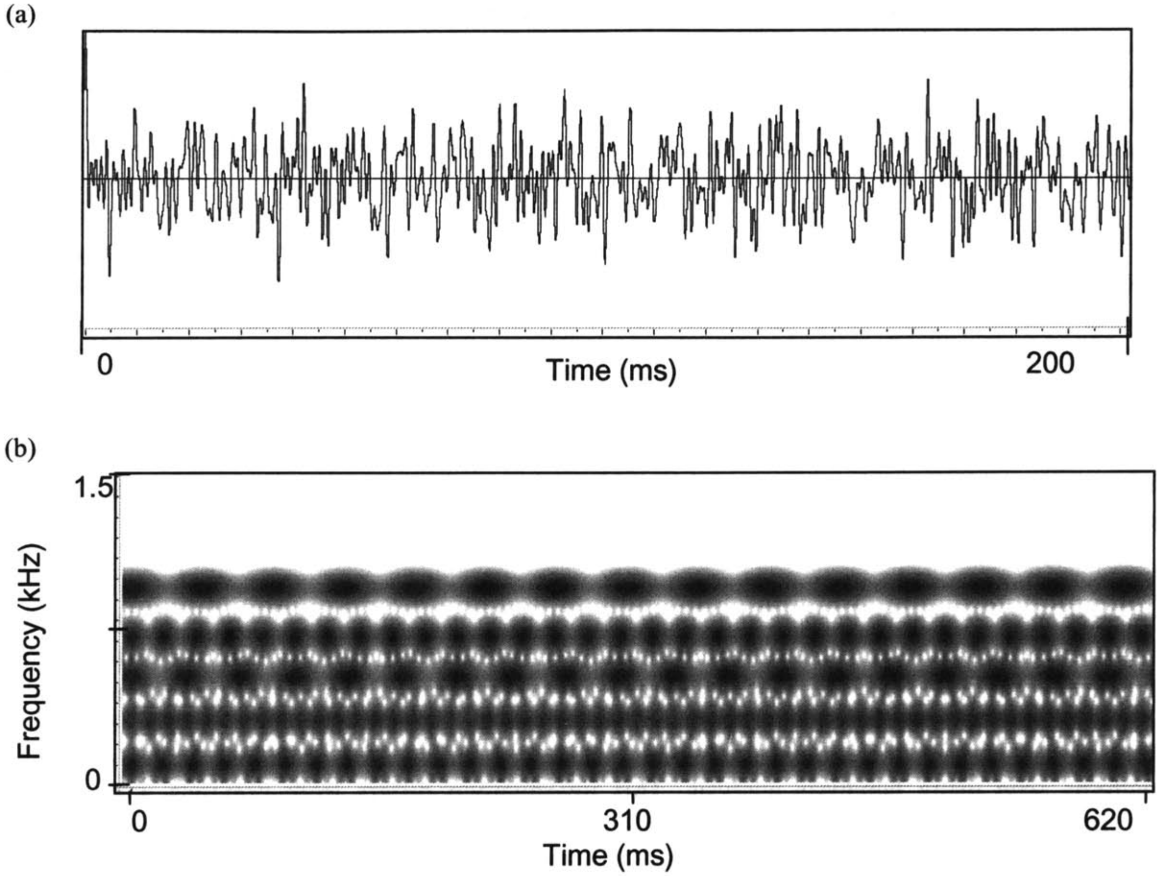

Figure 3-9. Sum of five amplitude-modulated sinusoids showing aperiodic variations in the time-domain (a) but strong periodic patterns in the spectral domain envelopes (b) using a 14.3-ms Hamming window.

3.3 Case Study of AM in Dysphonia

In developing an AM interpretation of the human voice, it is important to study the acoustic signatures of several dysphonias and, through spectral analysis, tie them into AM theory. In the following chapters, the goal will be to develop models that better extract the frequencies of voice AM and apply them to automatic recognition systems. As mentioned, the task of connecting the physical operation of the larynx to acoustic characteristics is largely an unsolved problem. Thus, we only speculate about the physical sources causing the AM we observe, but when possible, we will highlight hypotheses that exist in the speech community to explain a phenomenon.

In this section, seven different examples of the sustained vowel /a/ are investigated. This particular set of voices is the result of a survey by the author of over 350 sustained vowel utterances from the previously discussed Kay Disordered Voice Database [1]. The examples are selected primarily (1) because the patterns of amplitude fluctuation are extreme compared to other dysphonic utterances both acoustically and using spectrographic analysis and (2) because they demonstrate distinct types of time-varying phenomena as discussed in the previous chapter. Since there are many types of dysphonia, the chosen examples are meant to be used as acoustic evidence, not as a general survey of all possible dysphonia.

All spectrograms and spectra for this portion of the study were created using the WaveSurfer software package. Except where mentioned, only the first 5 kHz of the frequencies are shown in the each graph and a standard high-pass, *pre-emphasis*, filter is used. Many variations of analysis window length are investigated in order to best find evidence of amplitude modulations in the speech. Thus, some spectrograms have high time resolution and others high frequency resolution, depending on what is needed to highlight the features being discussed. The same variation is true for the grey-scale map chosen—some figures saturate to show only the varying amplitude regions and others show more clearly the entire spectrum. As discussed in the theory section, selection of bandwidth is a critical step in the demodulation process, with different patterns apparent when using different filters. In the next chapter, we introduce a biologically-inspired choice of analysis filter bandwidths and spacing that mimic aspects of the human auditory process.

3.3.1 AM Interpretation of Superperiodic Structure

As described in chapter 2, many forms of dysphonic voice qualities exhibit time-varying patterns in their envelopes, forming the *superperiodic* category. Recall that superperiodic speech can involve categories such as diplophonia, with a repeating pattern of glottal pulse height every two cycles, and periodic envelope, where a single underlying sinusoid appears to have an amplitude envelope applied to it.

Figure 3-10 gives an example from speaker JMJ04 showing several different patterns of superperiodic behavior. In the left box, there is evidence for a strong pulse every two cycles, yielding a repeating pattern of about 100 Hz. As shown in Figure 3-11, this pattern agrees with interaction between sub-harmonic components which are multiples of about 105 Hz. Here, the fundamental which can be seen strongly in the spectrum is at around 215 Hz. This observation is consistent with a view of the modal glottal source—a series of harmonically-related harmonic components—each modulated by an envelope with frequency 105 Hz. The 215-Hz harmonic sequence can be considered the carrier, with the amplitude envelope defined by the 105-Hz modulator.

The right box of the spectrogram indicates a change in behavior about halfway through the signal, with a strong pulse occurring every three glottal cycles. This behavior corresponds to a repeating pattern at about 70 Hz as is evident in the spectrum of Figure 3-12. It is interesting to note that the modal fundamental frequency component, here 215 Hz, remains strong, while the subharmonics are relatively weaker. As the subharmonics grow stronger, so does the difference between the magnitude of the largest and smallest glottal peaks in the time domain.

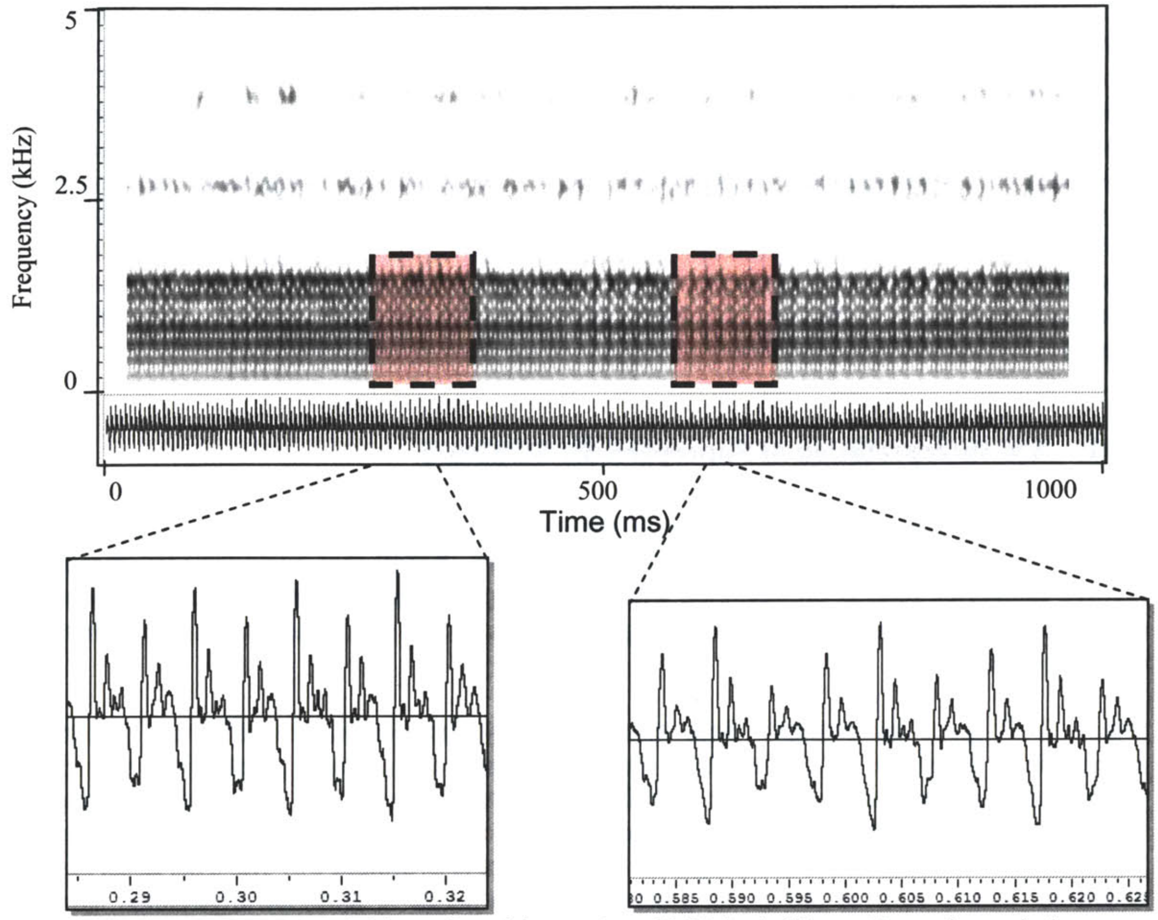

Figure 3-10. Spectrogram of /a/ produced by patient JMJ04. A 10-ms Hamming window was used in generating this figure. The period of the variation in the left box is every two glottal pulses—on the order of 100 Hz— and that in the right box is every three glottal pulses—about 67 Hz.

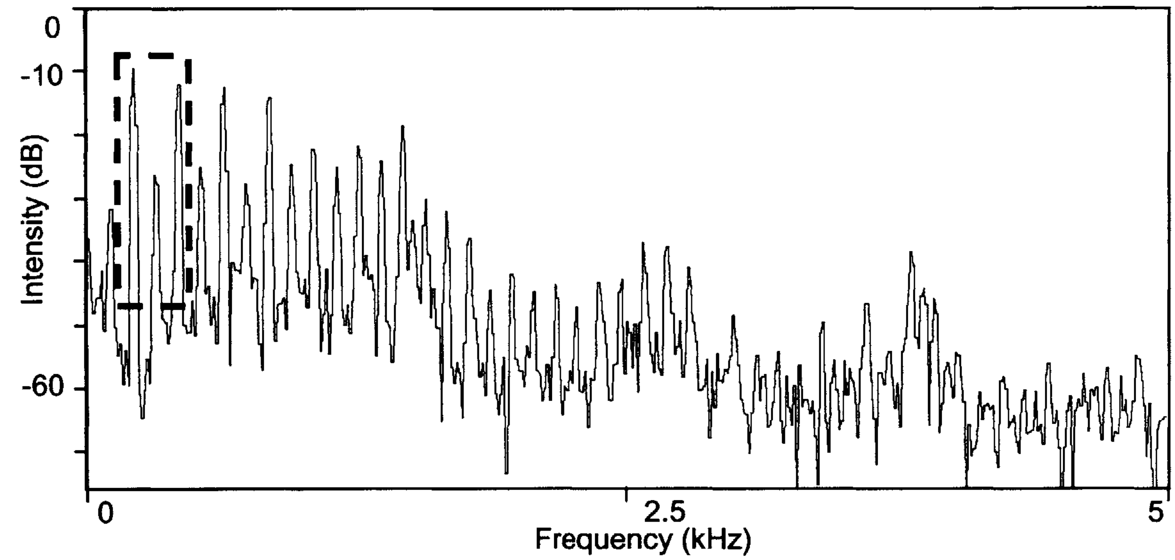

Figure 3-11. Spectrum of /a/ at 300 ms produced by patient JMJ04. An 82-ms Hamming window was used in generating this figure. The box on the left highlights three components of alternating height, measured at about 215, 314, and 418 Hz respectively. Observe that these are all multiples of the subharmonic at 105 Hz.

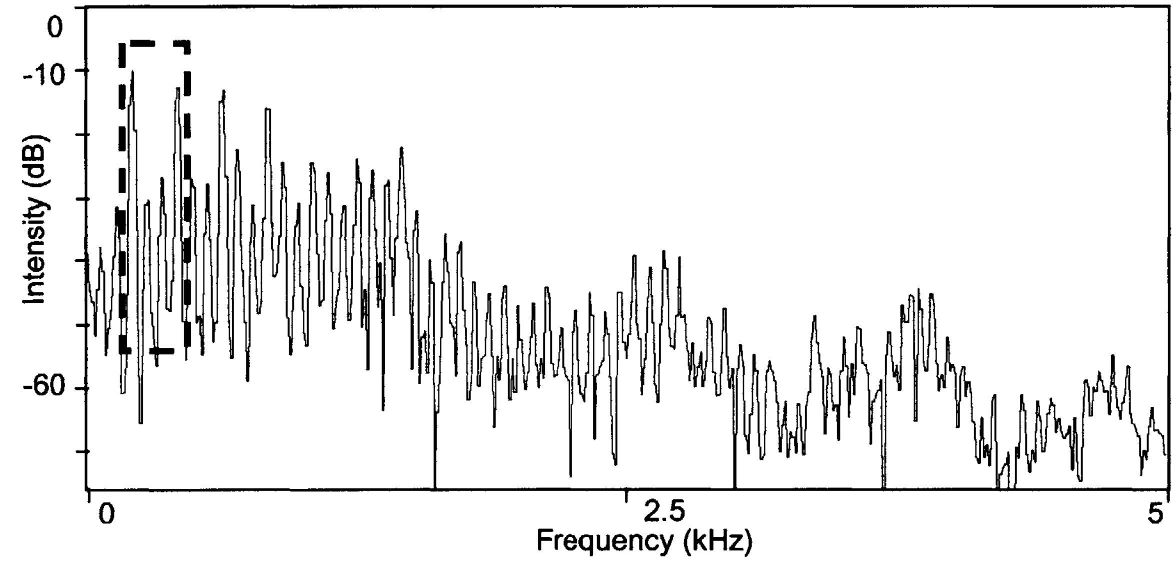

Figure 3-12. Spectrum of /a/at 600 ms produced by patient JMJ04. An 82-ms Hamming window was used in generating this figure. The box on the left highlights four components—one large, two small, and one large—measured at about 211, 278, 356, and 418 Hz respectively. Observe that these are all multiples of the subharmonic at 70 Hz.

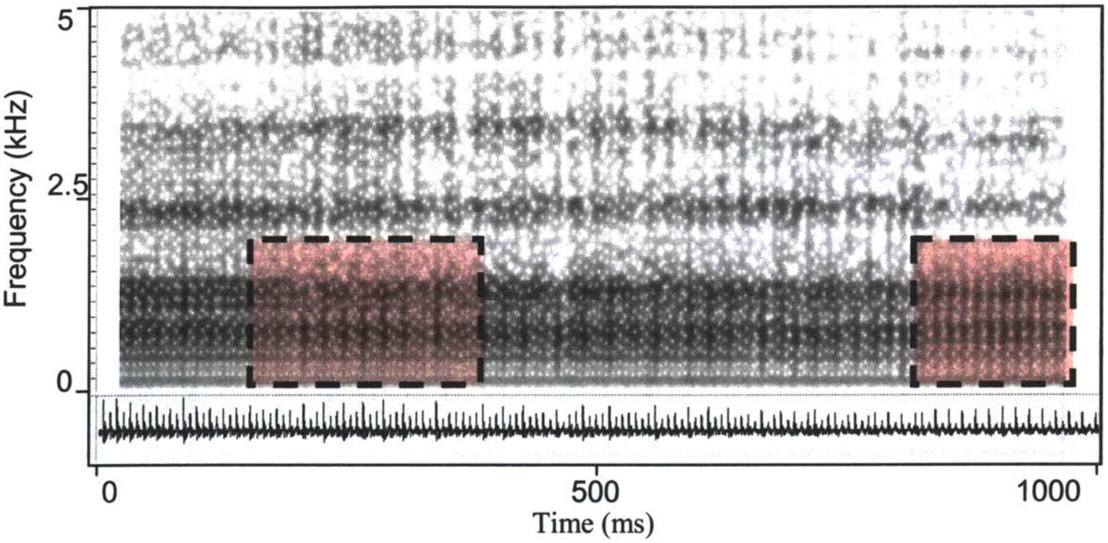

Figure 3-13. Spectrogram of /a/ produced by patient DAS10. A 14.2-ms Hamming window was used in generating this figure. The period of the variation in the left box is every three glottal pulses—on the order of 50 Hz— and that in the right box is every two glottal pulses—about 77 Hz. In the center of the utterance there is an irregular pattern of superperiodic behavior.

These observations for patient JMJ04 agree with those for patient DAS10 in Figure 3-13. Several differences, including a lower fundamental of about 150 Hz, a region of irregularly spaced pitch periods in the middle, and more quickly damped pulses are apparent. The far right pitch-doubled area, in fact, is almost completely transformed into a waveform of half the fundamental frequency. Overall, the waveform in this region is close to the description of creak, with highly damped pulses at half the fundamental, shown in Figure 3-14. Thus, DAS10's vowel production is a good example of a combination of at least two subcategories of glottalization—creakiness and superperiodicity.

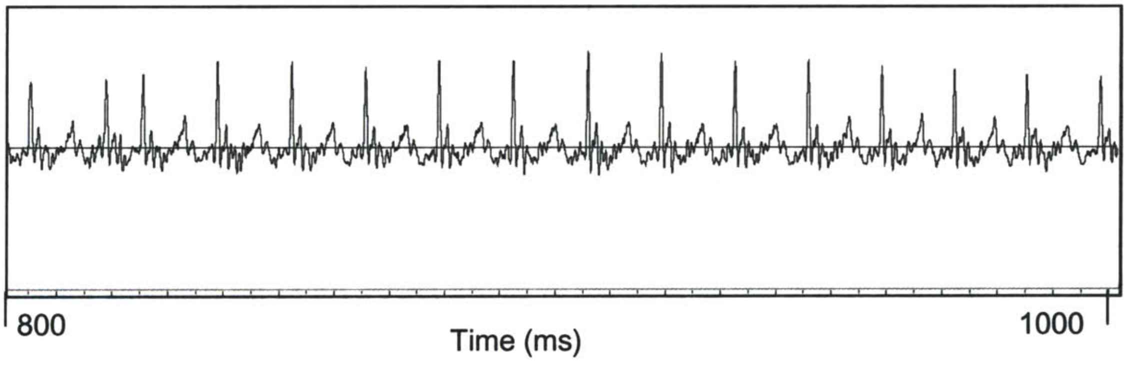

Figure 3-14. Waveform of /a/ produced by patient DAS10 between 800 and 1000 ms. Note the similarity of this region to creak.

Some voices, such as for patient JXS01, tend to be glottalized throughout their voicing. According to the database, this 70-year-old male smoker suffers mainly from unilateral left paralysis as well as various forms of hyperfunction—the latter presumably to compensate for the inability to use one of the cords. An example of creak from this utterance, highlighted in the right box of Figure 3-16 and the waveform in Figure 3-15, shows strong, quickly decaying pulses. It is interesting to note that the same time-varying pattern is seen up to about 4 kHz, due to the high-bandwidth impulsive nature of the glottal pulses in creak.

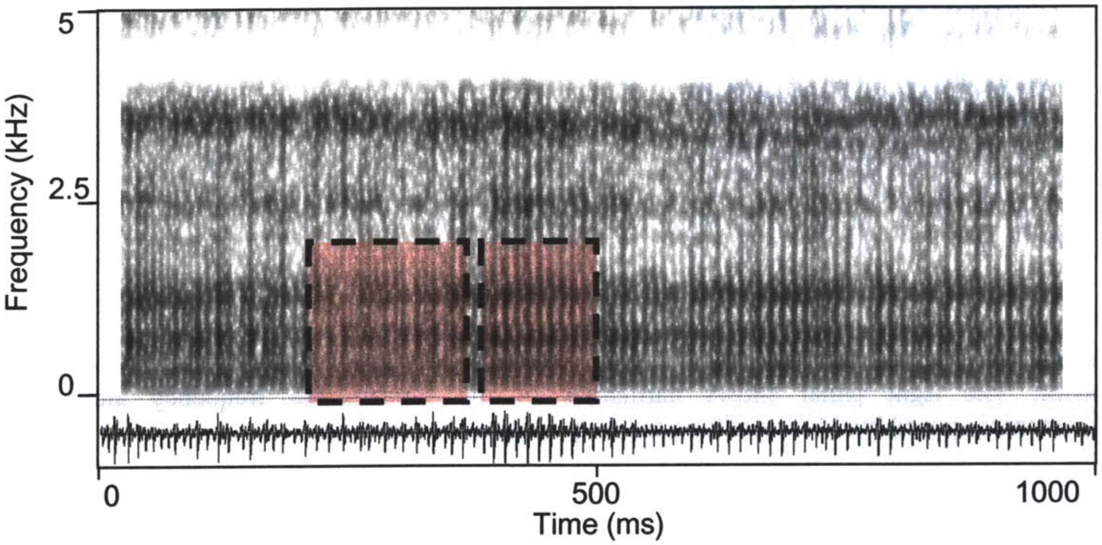

Figure 3-15. Spectrogram of /a/ produced by patient JXS01. A 10-ms Hamming window was used in generating this figure. The box on the right indicates a region of creak, with a relatively low 91-Hz pitch, while the left box shows irregular glottalization.

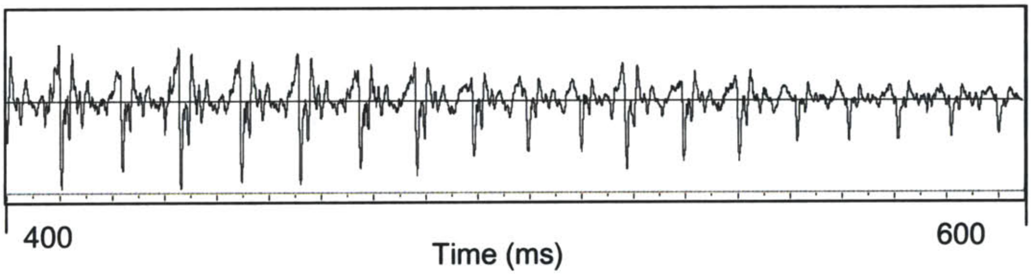

Figure 3-16. Waveform of /a/ produced by patient JXS01 between 400 and 600 ms. Observe the regularly spaced low-frequency pattern of highly damped glottal pulses at about 90 Hz. There is some irregularity here in the height of the pulses.

Another category of superperiodic behavior is that which appears to be a single fundamental modulated by a lower frequency envelope. Gerratt and Kreiman discuss this type of signal in their taxonomy of acoustical voice properties as "a relatively high-frequency wave… modulated by a much lower frequency envelope." [14] Figure 3-17 is repeated from chapter 2 to show what is mean by this description. In the right panel, we see a spectrum with many harmonically related subharmonics which interact as reflected in the time-variations at the period of the envelope. In this example, the apparent fundamental frequency of the waveform underneath the envelope envelope, is reflected in the spectral domain as a prominent line component.

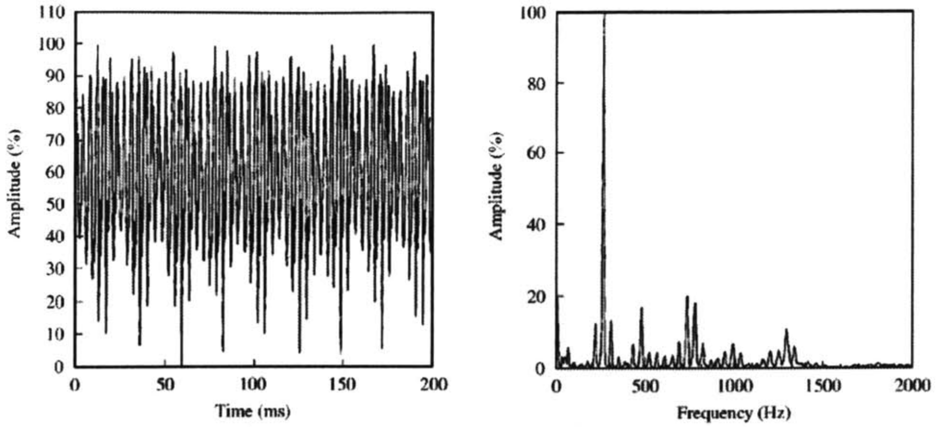

Figure 3-17. Figure from [14] showing an example of superperiodic "periodic envelope" behavior. The right panel shows harmonics of the fundamental frequency along with subharmonics spaced by 44 Hz defining the frequency of the envelope.

Patient HXL28 provides another example of this "periodic envelope" type of periodicity as demonstrated in the waveform of Figure 3-19. Notice the slowly-varying envelope modulating a waveform with a fundamental frequency of 328 Hz. This fundamental, when combined with the envelope, results in a harsh complex voice quality descriptively described as a "chainsaw voice" after the sound of a power saw as it cuts through a tree. In the spectrogram of this voice shown in Figure 3-18, we see two distinct frequencies at which the envelope varies with time—one around 30 Hz, and the other around 60 Hz. These observations match the spectrum analysis shown in Figure 3-20, where there are difference frequencies agreeing with both these values. One suitable cause of the 328 Hz pitch, a low frequency pole amplifying the 328 Hz component, is also highlighted in the spectrum. Low-frequency poles are used in dysphonic speech synthesis approach by [2] discussed in appendix A. Note that this spectral peak is sharper than F1, which is in the neighborhood of 1000 Hz and isolates the 328 Hz component and its two sidebands.

One question that becomes apparent is how the periodic envelope voice type is fundamentally different than the period doubled or tripled examples previously given. Both cases have envelopes governed by the difference frequencies of the subharmonics, with the fundamental frequency governed by a large spectral component. One hypothesis is that the difference lies in the pole around 300 Hz, with the amplitude modulation being inherent to the glottal source when period doubling and tripling are exhibited and a property the vocal tract filter when the periodic envelope voice type occurs .

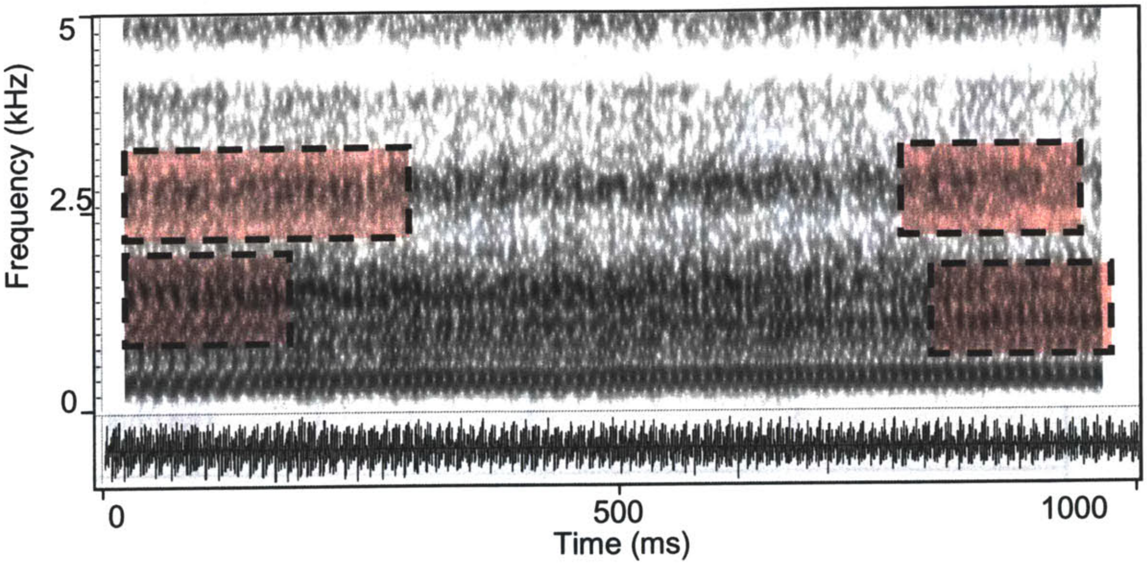

Figure 3-18. Spectrogram of /a/ produced by patient HXL28. A 10-ms Hamming window was used in generating this figure. The top left boxes indicate modulations around both 30 Hz and 60 Hz, while the bottom left box indicates variation at about 64 Hz. The boxes on the right indicate similar frequency-dependent activity. Also note the strong activity around 300 Hz indicating a pole in the speech spectrum.

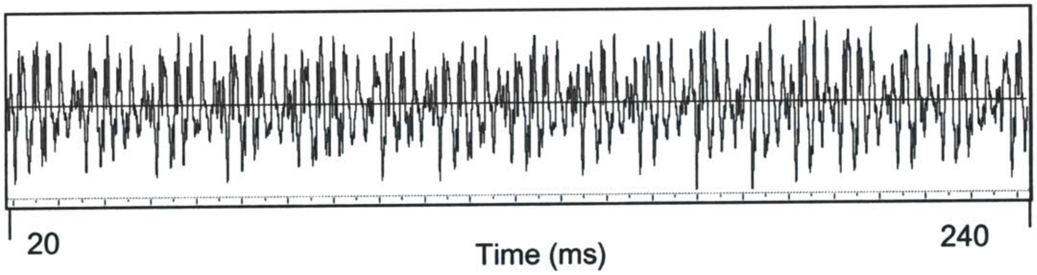

Figure 3-19. Waveform of /a/ produced by patient HXL28 between 0 and 240 ms. The signal exhibits a 328-Hz fundamental modulated by a 64-Hz envelope.

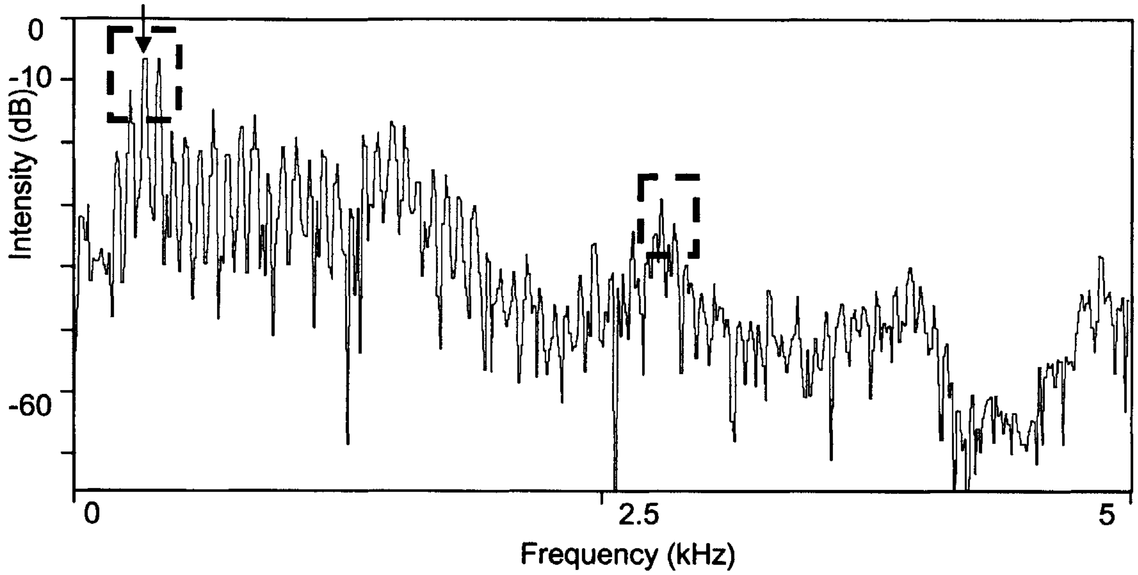

Figure 3-20. Spectrum of /a/ at 0 ms produced by patient HXL28. An 82-ms Hamming window was used in generating this figure. The box on the left highlights a region suggesting a low-frequency pole in the spectrum which amplifies the 328-Hz fundamental, indicated with the arrow. The box on the right indicates a region with three beating components, one pair separated by 64 Hz, the other by about 32 Hz.

3.3.2 AM Interpretation of Aperiodic Signals

As previously defined, aperiodic or irregular signals, involve glottal pulses with little or no apparent pattern in the length of the periods between them or in their heights. In this section, we investigate three kinds of aperiodic signals—irregular glottalization, aperiodic interactions of harmonics, and amplitude modulated noise. Note that the last category is not explicitly described in the voice quality literature. It probably best fits hoarseness, which has been defined as the combination of breathiness and roughness [11].

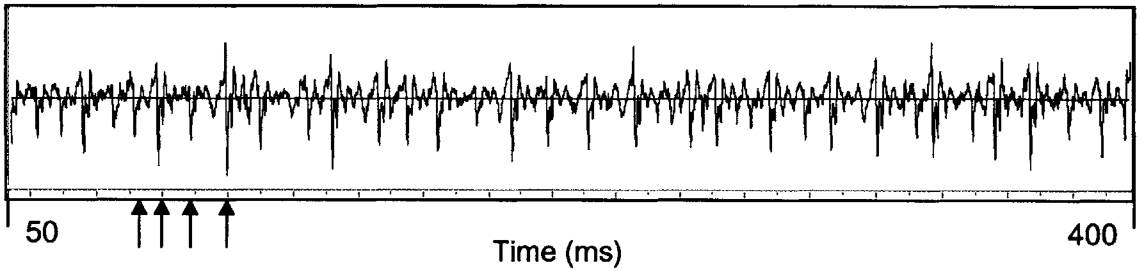

Figure 3-21. Waveform of /a/ produced by patient JXS01 between 50 and 400 ms. The pitch periods are distinct as in creak but irregularly spaced and with varying amplitudes. Arrows indicate several such pitch periods.

Figure 3-21 shows a waveform from the previously-introduced patient JXS01 showing irregular glottalization. Note that variations in both the height and spacing of the pulses does not indicate a periodic signal. Another form of aperiodic phonation is shown in Figure 3-22, a waveform produced by patient KAHO2, a 73-year-old nonsmoker. The perceptual nature of this utterance is described as a "chainsaw voice" in the same way as HXL28. Notice, however, that the time waveform does not contain a similar periodically-modulated high-frequency sinusoid, although there is indication of periodic patterns of high energy. When the spectrogram in Figure 3-23 is taken, a more complicated pattern emerges. Around the first formant, we see repeating pulses between about 50 and 60 Hz. Then, at the third formant, another pattern emerges, consisting of periodicity in the envelope at half that frequency. Finally, we see a third distinct pattern of approximately 100-Hz amplitude variation centered at about 200 Hz.

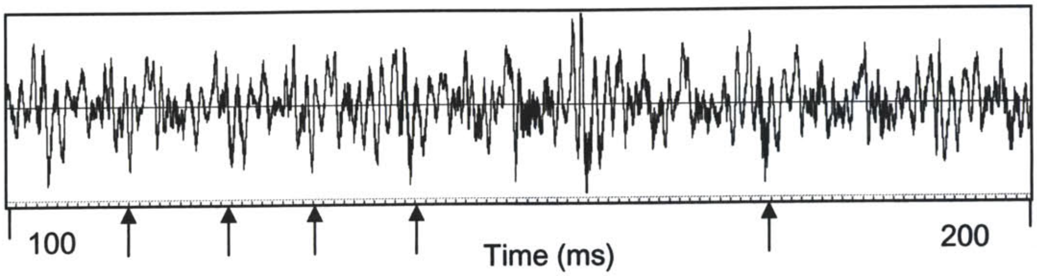

Figure 3-22. Waveform of /a/ produced by patient KAH02 between 100 and 200 ms. Unlike in creaky voice, it is difficult to pick out individual glottal pulses here. Large pulses every 5-to-10 ms suggest a complex underlying pattern; several of these are indicated with arrows.

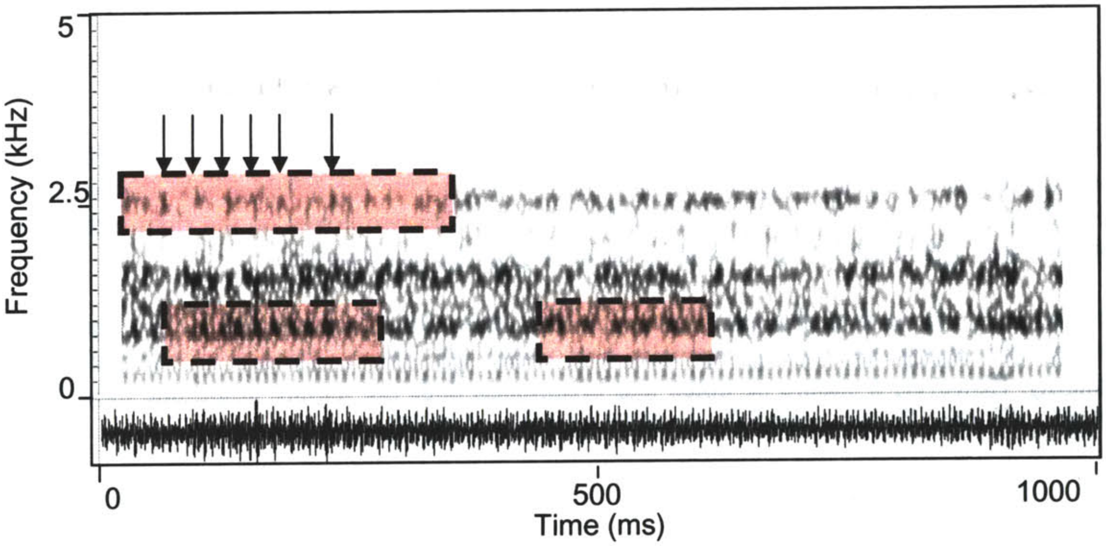

Figure 3-23. Spectrogram of /a/ produced by patient KAH02. A 10-ms Hamming window was used in generating this figure. The top left box and accompanying arrows indicate regular patterns of variation which are different from the activity at the first formant frequency in the bottom left box. In contrast to the 30-to-40-Hz pattern around F3, there is a 50-to-60-Hz pattern at F1. The right box highlights amplitude modulations at about 100 Hz occurring near 200 Hz. Different bands show different modulations.

The spectrum taken around 100 ms, shown if Figure 3-24, can be used to explain much of the amplitude modulation behavior seen in the spectrogram. First, there is a difference frequency between harmonics of about 29 Hz yielding a 30-Hz difference frequency in the region of F3 and a 60-Hz difference around F1. The spectrum also includes a set of prominent line components at 179, 290, and 460 Hz. The three components are not multiples of one another, but rather are all multiples of about 29 Hz. Other components at 29-Hz multiples have less magnitude than at these three frequencies.

As is demonstrated in Appendix A through a dysphonic speech synthesis experiment, one hypothesis is that the amplified components at 179, 290, and 460 Hz are not harmonics, but instead the fundamental frequencies of individual modes of glottal fold vibration as modeled by Berry [3]. Regardless of the origin of the phenomenon, it is hidden in the time waveform and requires further analysis.

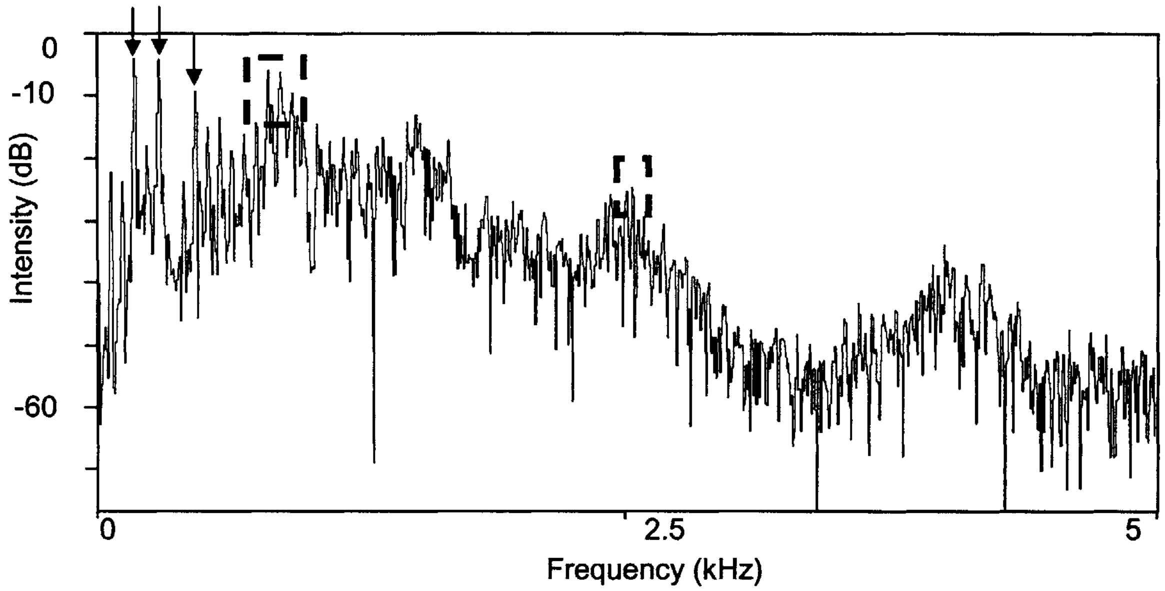

Figure 3-24. Spectrum of /a/ at 100 ms produced by patient KAH02. A 164-ms Hamming window was used in generating this figure. The arrows on the far left highlight the prominent sinusoids at 179, 290, and 460 Hz discussed in the text. The two boxes on the right indicate the harmonic spacings related at 60 Hz and 30 Hz as seen in the spectrogram.

The last form of aperiodicity investigated is that of amplitude modulated noise. Unlike the previous two categories, this group falls best under the breathy header in our preliminary taxonomy. However, it should be noted that breathiness does not require amplitude modulation; elements of it have been successfully synthesized without it [24].

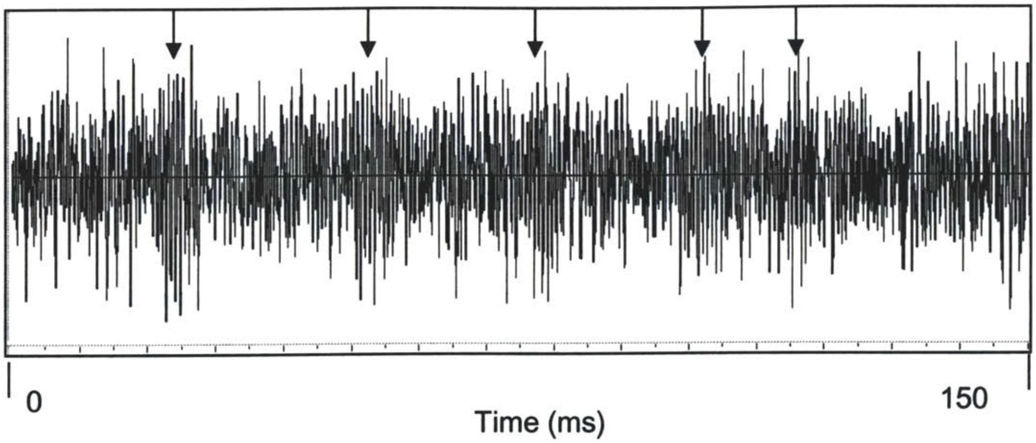

Figure 3-25. Waveform of /a/ produced by patient DXC22 between 0 and 150 ms. Notice that there is no obvious pitch period here, and the waveform appears as random fluctuations. There are four areas of activity indicated by arrows that constitute evidence for AM noise as discussed in the text.

DXC22 is a 48-year-old male post-biopsy and irradiation case with swelling of the vocal folds and hyperfunction. Perceptually, the voice is similar to white noise with a subtle roughness. As can be seen in Figure 3-25 the time waveform contain periodic bursts of energy but has no obvious pitch period. As indicated by the arrows, there are four regions of increased energy, which we will call amplitude-modulated noise. This view is furthered by the spectrogram in Figure 3-26, which contains several regions of time-varying patterns. Overall, more research is required in this area, but the results thus far support the use of improved models to extract frequency-dependent patterns "hidden" in the time waveforms.

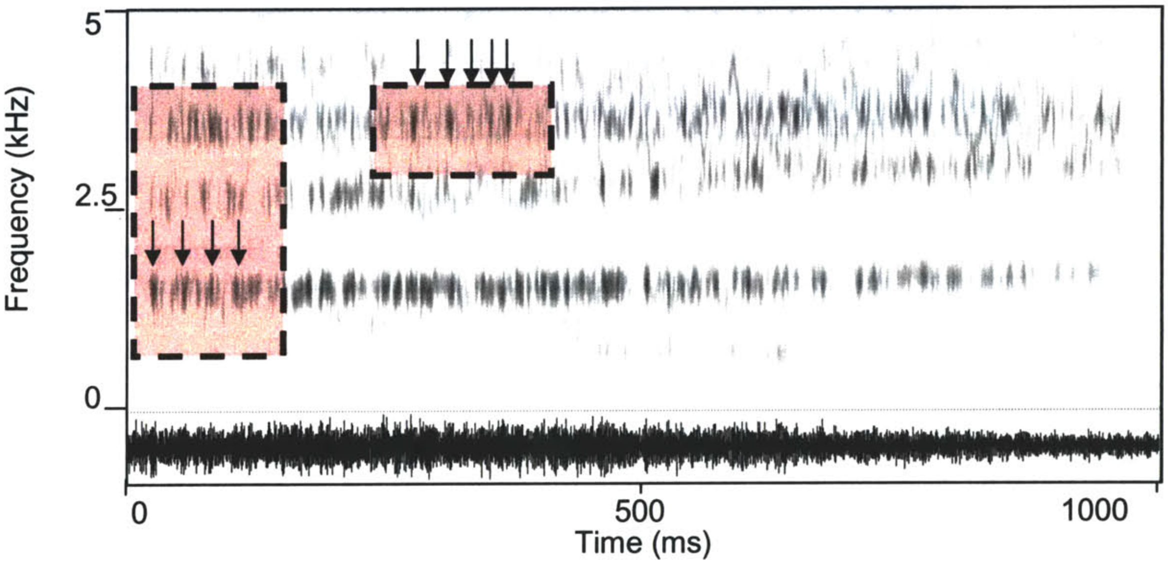

Figure 3-26. Spectrogram of /a/ produced by patient DXC22. A 5-ms Hamming window was used in generating this figure. The two boxes and accompanying arrows indicate regions with repeating patterns. At signal onset, the repetition at F1 is about 40 Hz whereas, in the middle of the utterance, the variation at F2 is between about 50 and 60 Hz.

3.3.3 Transduction of FM to AM

The last group of amplitude modulations studied involve tremor and flutter, both variations in the amplitude envelope of a signal below 15 Hz. As discussed in chapter 2, these problems are most often related to neurological speech disorders, where nervous control of the articulators is compromised.

The generation of AM from FM is mentioned in section 3.2 as a possible consequence of a system with time-varying source frequency and/or a time-varying formants. Figure 3-27, recorded by patient JAB08, a 69-year-old post-biopsy case with scarring and a vocal tremor illustrates this phenomenon. As indicated in the spectrogram, the amplitude envelope of this patient increases whenever there is a dip in the fundamental. Figure 3-28 compares the spectrum at 313 and 424 ms in the utterance. When the 5th harmonic is closer to the first formant, the amplitude is increased; likewise, it decreases when it is further away. We propose this case as an example of the transduction of FM to AM similar to that discussed in [35].

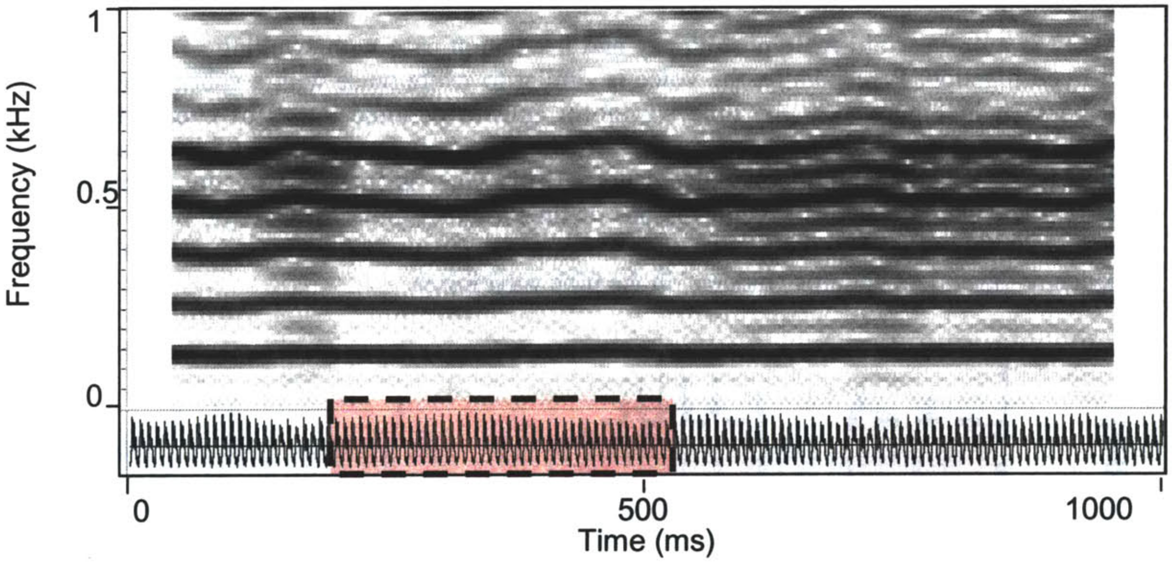

Figure 3-27. Spectrogram of /a/ produced by patient JAB08. A 66.6-ms Hamming window was used in generating this figure. The box highlights a portion the time-varying envelope of the waveform that changes with variations of the fundamental frequency.

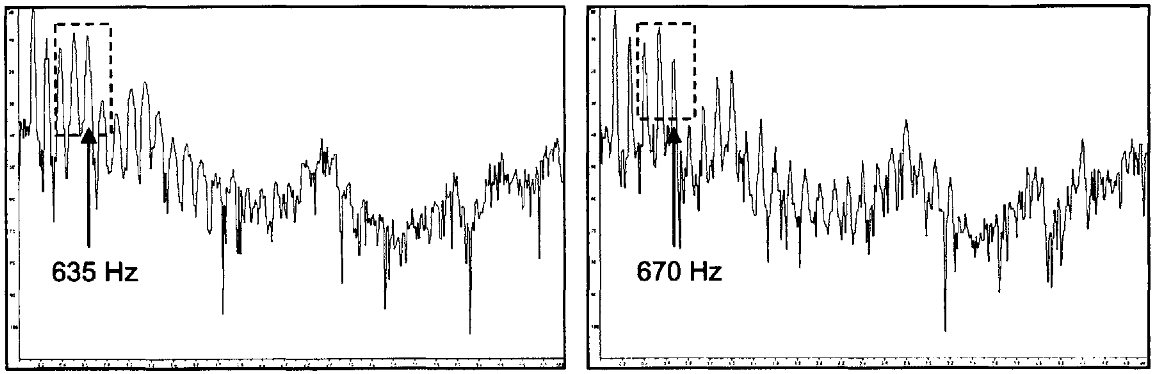

Figure 3-28. Spectra of /a/ at 313 and 424 ms generated by patient JAB08. An 82-ms Hamming window was used in generating these figures. The box highlights the region of spectral components that change as the pitch frequency varies. Note that the 5th harmonic is larger in the first panel and becomes smaller in the second as it increases about 35 Hz in frequency.

3.4 Conclusions

In this chapter, we defined amplitude modulations both in a communications engineering sense and with relation to real speech. We propose that AM in real speech is different from the traditional definitions in that it has (1) non-sinusoidal sources as carriers, (2) overlapping sidebands each of unknown bandwidth, and (3) post-filtering of the summed AM sources by the vocal tract. These issues combine to make demodulation of the speech sources a difficult and currently unsolved problem.

With the AM theory developed, we analyzed the speech of several different dysphonic speakers in order to observe how interactions of components in the frequency domain can produce a wide range of voice qualities. Four broad and overlapping classes that we observed were (1) interacting harmonic components, (2) irregular pitch pulses, (3) amplitude-modulated broadband noise, and (4) the transduction of FM to AM.

The first class was characterized by line components in the frequency domain. AM theory dictates that modulations occur at difference frequencies between individual frequencies; this was shown in our real speech examples. An additional observation is that the relative magnitudes of different harmonic components are important to the patterns seen in the spectrogram and in the time domain. JMJ04, HXL28, and KAH02, for example, all had regions exhibiting subharmonics which were multiples of about 30 Hz but with very different relative component strengths. In JMJ04, in the time region characterized by period-tripling, the spectral line components alternate in a pattern of two small 60-Hz components followed by one large component. In HXL28, we see evidence of 30-and 60-Hz subharmonics but with the multiples of 60 Hz around 300 Hz amplified. Finally, in KAH02, we see prominent low-frequency spectral line components mainly at 179, 290, and 460 Hz, all approximately multiples of about 29 Hz. The physiological causes of such phenomena are unknown.

This chapter also presented examples of irregular glottalization, amplitude-modulated noise, and FM-to-AM transduction. Spectral correlates for the last class were found but the acoustic theory to explain these types of modulations needs to be developed. We do not yet have an AM model that explains the time-fluctuations in demodulated broadband noise or frequency-modulated sources.

In this chapter, we demonstrated a connection between the bandpass filters used for demodulation and the resulting time-varying patterns in the envelope. Specifically, we showed that there is a tradeoff between narrow and wide filters. Narrow filters restrict the overlap of two stations but can filter out sidebands while wide filters restrict the loss of sideband components but can pass overlapping stations. In this chapter we used the short-time Fourier transform, a constant-bandwidth analysis method, for demodulation. In chapter 4, we present three models with biologically-motivated analysis-filter bandwidths. We then show evidence that each of these models present complementary information about AM.

Chapter 4

Biologically-Inspired Models to Represent Amplitude Modulations

As discussed in the previous chapters, we propose that amplitude modulations are an important property to extract from speech for dysphonia recognition, as well as other applications such as speaker recognition. The purpose of this chapter is to describe the three classes of human auditory system models—*mel-filtering*, *adaptive nonlinearity*, and *modulation filtering*—used in this thesis and how amplitude modulations are represented by each of them. In chapter 3, analysis filter bandwidth was found to be important in extracting amplitude modulation patterns from synthetic waveforms and speech. There are many possible analysis filter configurations, but we propose that a human-inspired filterbank is a natural choice. Humans seem to be adept at determining voice characteristics such as those that identify a certain speaker [41] or indicate perceptual voice quality such as roughness [25], suggesting that mimicking the human analysis filtering process is a logical approach.

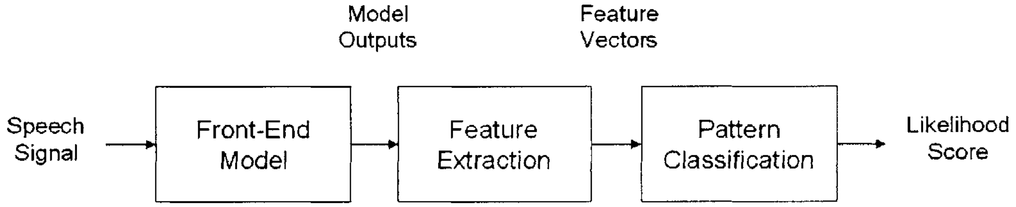

Figure 4-1. Overview of a dysphonia recognition system. A speech signal enters and is processed by an auditory front-end model to produce a set of output signals. From these outputs, features are extracted and sent to an automatic classifier which determines how likely it is that an utterance indicates a certain voice disorder.

Each auditory model in this chapter acts as *front-end model*, the first of three stages that make up our dysphonia recognition systems, as depicted in Figure 4-1. The goal of the front-end model is to convert an acoustic waveform into a new representation better suited to capture application-dependent properties. In speaker recognition applications, for example, this stage is used to obtain features that best characterize a certain talker's voice. In dysphonia recognition, features should capture acoustic properties that make a certain voice disorder stand out from other voice disorders as well as from normal speech.

The first model studied, mel-filtering, is a component of a technique called the *mel-cepstrum* used to build the baseline system in speaker recognition [34] which has also performed effectively in dysphonia recognition tasks [10]. Two previously existing biologically-inspired designs [36] are also introduced in this chapter along with new improvements to allow them to better capture amplitude modulations. The intent of these models is to represent the response of the auditory system to amplitude modulations on two different levels: (1) the cochlea and auditory nerve and (2) the inferior colliculus or *ICC*. We will study how modulations are represented by the various techniques for both synthetic stimuli and real voices and argue that the two levels provide somewhat complementary information about AM.

4.1 Mel-Filtering

For a number of applications, including speaker recognition [34, 40] and dysphonia recognition [10], mel-filtering models have performed very competitively in comparison to other methods. In this sense, they are referred to in this thesis as *state-of-the-art*. There do not yet exist features which consistently outperform mel-filtering models in these two tasks. Many elements of this feature set come from the work of Davis and Mermelstein, who applied the model to the problem of automatic speech recognition [8]. Some elements used in this technique, including the logarithmically increasing spacing and bandwidths of the analysis filters with frequency are biologically inspired.

Generally, the effect of mel-filtering can be thought of as a smoothing over a high-resolution DFT. Equivalently, this can be pictured as the obscuring of pitch and other source characteristics. Quatieri [34] notes that the outputs of mel-filtering "are likely to contain some source information, e.g., the spectral tilt of the STFT magnitude influenced by the glottal flow derivative" but that their primary content is information "from the vocal tract system function." The class of speech characteristics, including amplitude modulations, however, is not known rigorously in the representation by mel-filtering.

4.1.1 Model

As shown in Figure 4-2, the action of the front-end mel-filtering model is to take the magnitude spectrum of the DFT, multiply it in the frequency domain using triangular weighting functions, and compute the filter output energy [34]. The weighting function applied to the analysis filters can have several different variations. Davis and Mermelstein [8] use a method with triangles spaced from 100 Hz up to 1000 Hz and then logarithmically spaced with a ratio of 1.1487 between filters with center frequencies between 1000 Hz and 4000 Hz as shown in Figure 4-3.

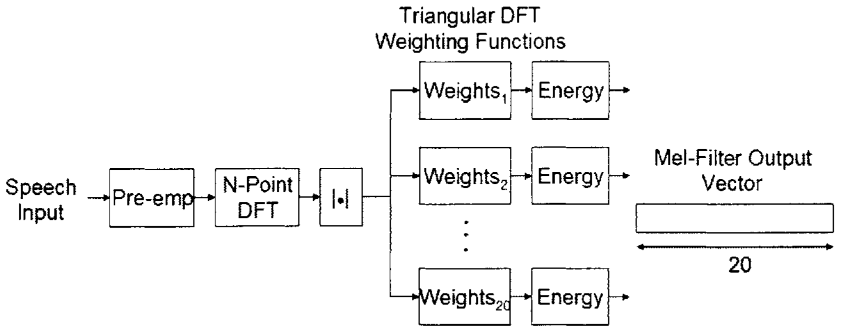

Figure 4-2. Schematic view of the mel-filtering model. The DFT weighting function has spacing and bandwidths based on the mel-scale as described in the text and accompanying figures.

Quatieri [34] uses a similar method where the center frequencies of the triangles have a modified mel-scaling which yields linear filter spacing up to 1000 Hz and logarithmic spacing above 1000 kHz. Although Quatieri's formulation is quite general, the shape of the bands is suggested to be triangular with widths equivalent to human-derived critical bands (as in Figure 4-3) after the work by Davis and Mermelstein described above [8]. Quatieri's description is different than these authors in that (1) the triangular filters are normalized for equal energy, or summed squared magnitude, (2) the weighting functions are triangular in the spectral magnitude domain, not in the spectral energy domain, and (3) the spacing and bandwidths of the triangular functions above 1000 Hz are slightly different, with a ratio of 1.1 between neighboring center frequencies, as shown in Figure 4-4.

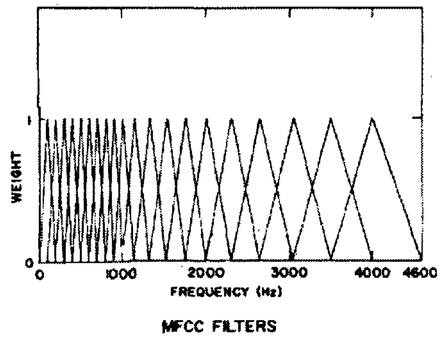

Figure 4-3. A plot of the mel-frequency filters from [8].

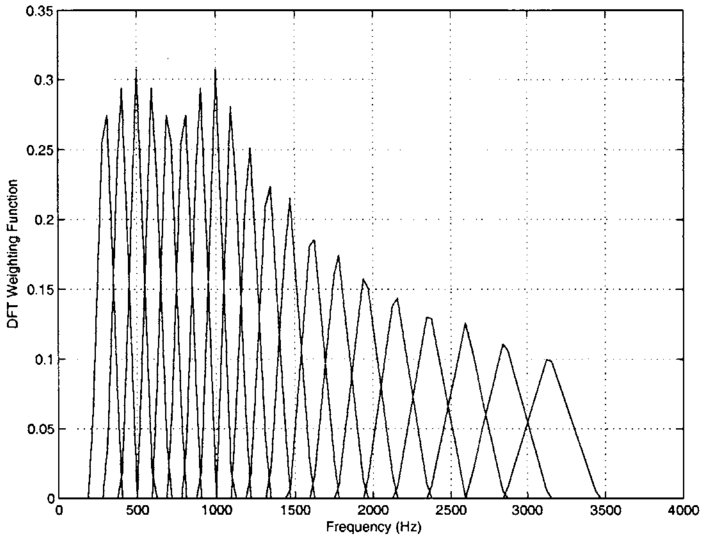

Figure 4-4. Triangular weighting functions used to weight the DFT magnitude spectrum. This particular set of functions is for the standard 256-point DFT implementation of the technique as implemented in this thesis.

In this thesis, we use Quatieri's formulation as a baseline except that the triangular weighting functions are normalized to equal summed *magnitude*, not equal summed energy. Preliminary speaker recognition experiments have shown comparable results using all three of these filter variations. We also include a pre-emphasis filter stage with impulse response and frequency response shown in Figure 4-5. The entire procedure is outlined below:

(1) The input waveform is windowed using 20-ms triangular functions, each overlapping the previous one by 10 ms.
(2) The DFT (implemented as an FFT) is taken to transform the windowed signal into the spectral domain.
(3) The pre-emphasis filtering function is applied to the DFT, acting to highpass-filter the signal.
(4) The DFT magnitude bins are weighted according to the 20 triangular weighting functions and the weighted sums are calculated.
(5) Each of the outputs is squared and the logarithm is taken.

(a)

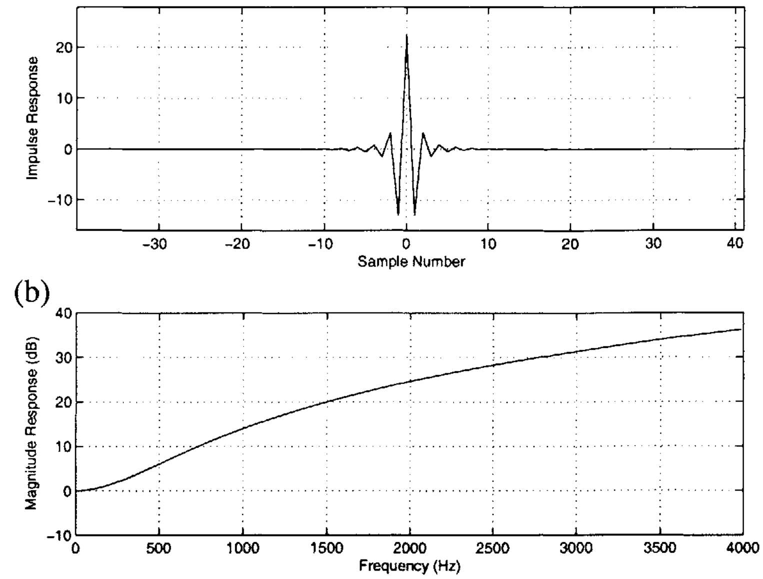

Figure 4-5. Impulse response (a) and frequency response (b) of the pre-emphasis filter used in the mel-filtering computation.

4.1.3 Representations of Amplitude Modulations

As discussed in previous chapters, amplitude modulation occurs in both normal and dysphonic speech and is hypothesized to be important to the voice qualities that one perceives. For this reason, it is useful to analyze both real and synthetic signals using each of our feature extraction methods. In this chapter, we revisit several of the speech examples seen in previous chapters using the new approaches. We also show the response of the models to amplitude modulated tones and noise.

The output of the mel-filtering model for two different synthetic amplitude modulated signals are shown in Figures 4-6 and 4-7. In order to eliminate the transient response to the stimulus onset, only the middle 1 second of a 3 second presentation of the signals are displayed. In addition, the outputs of the model are sampled every 5 ms. Figure 4-6 shows the response to a single 1000-Hz sinusoid modulated by a 30-Hz sinusoid as described by:

$$y[n] = \cos(2\pi 1000n / f_s)(1 + 0.5\cos(2\pi 30n / f_s)$$

Figure 4-7 shows the response to a white noise source modulated by a 27-Hz sinusoid as described by:

$$y[n] = q[n](1 + 0.5\cos(2\pi 27n / f_s)$$

where q[n] is a uniformly distributed white noise sequence.

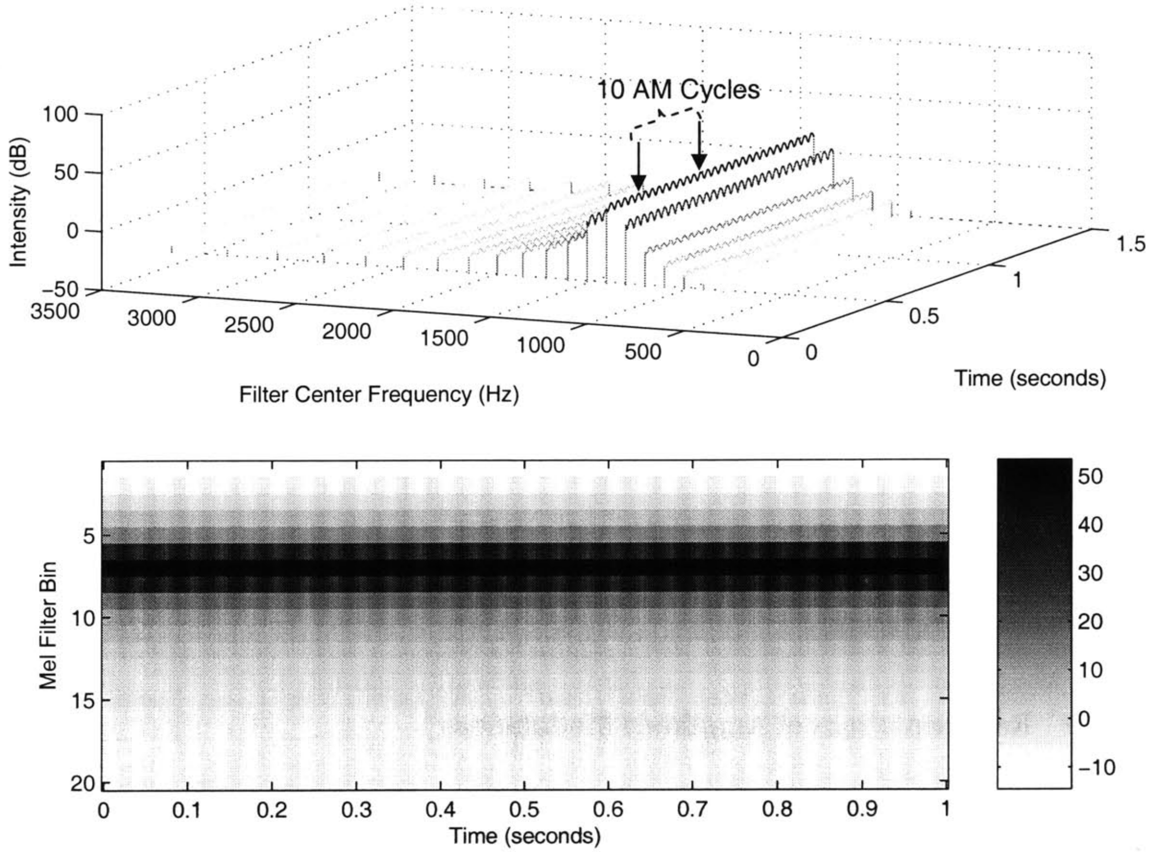

Figure 4-6. Response of the mel-filters to a 1000-Hz sinusoidal carrier amplitude modulated by a 30-Hz sinusoid. The perspective in the top panel is skewed such that the peak is actually directly above the 1000-Hz line. The arrows in the top panel indicate 10 cycles of the time-varying amplitude envelope.

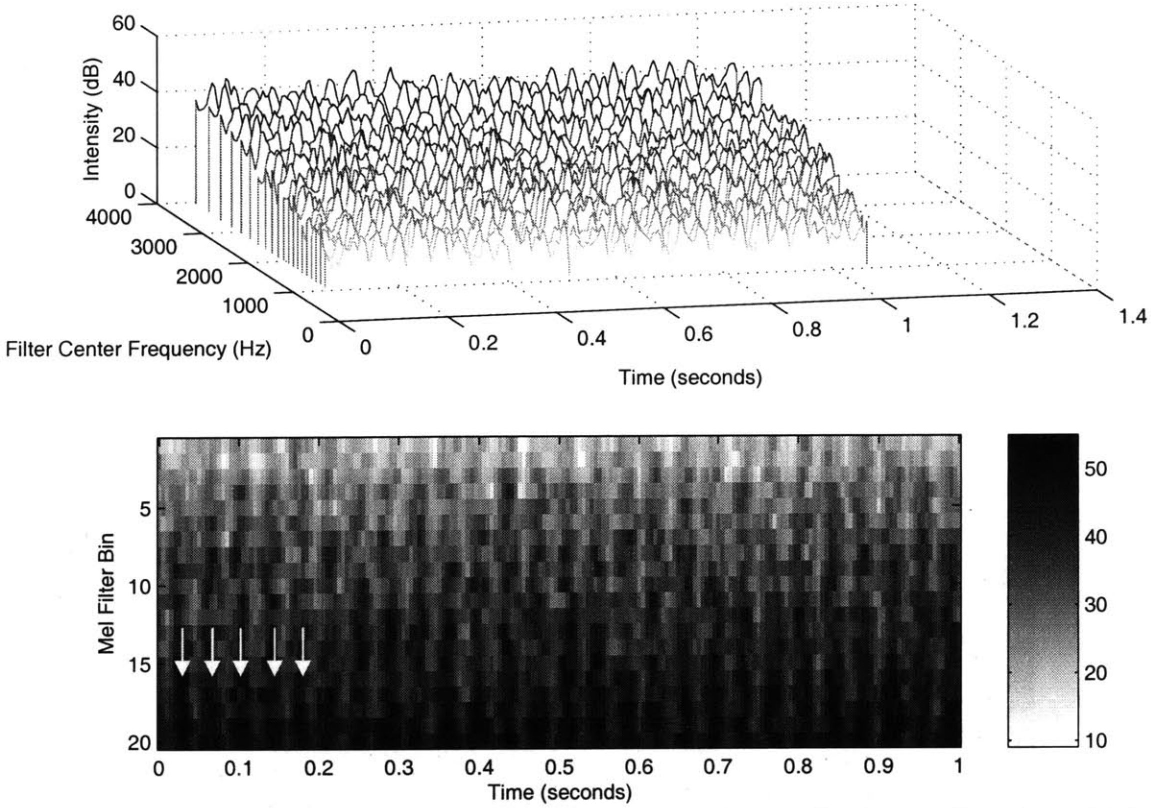

Figure 4-7. Response of the mel-filters to a white-noise carrier amplitude-modulated by a 27-Hz sinusoid. Arrows indicate peaks of the modulation envelope—the envelope is blurred at low mel-frequencies.

In Figure 4-6, the sinusoid causes the peak response to occur at 1000 Hz. The envelope of this signal can be seen to fluctuate according to the 30 Hz modulating waveform, as would be expected. However, amount of modulation is not *explicitly* captured by the mel-filtering model. That is, the fact that 30-Hz modulation is used must be inferred from the *time variation* on one of the channels. In Figure 4-7, the spectrum has more noise due to the broadband source, but we can still observe a weak periodic variation due to the 27-Hz modulator. Here we also can observe the effect of the pre-emphasis filter, which boosts the high frequencies.

Figure 4-8 shows the response of the DPK model to the middle 800 ms of the sustained vowel utterance from DAS10, introduced fully in chapter 3. We can see some dynamic activity, but the details are blurred. In particular, we know from previous analysis that we expect a series of strong glottal pulses in the representation. Although there is some pulse-like activity near 100 ms, the activity is less clear than in the original DFT analysis.

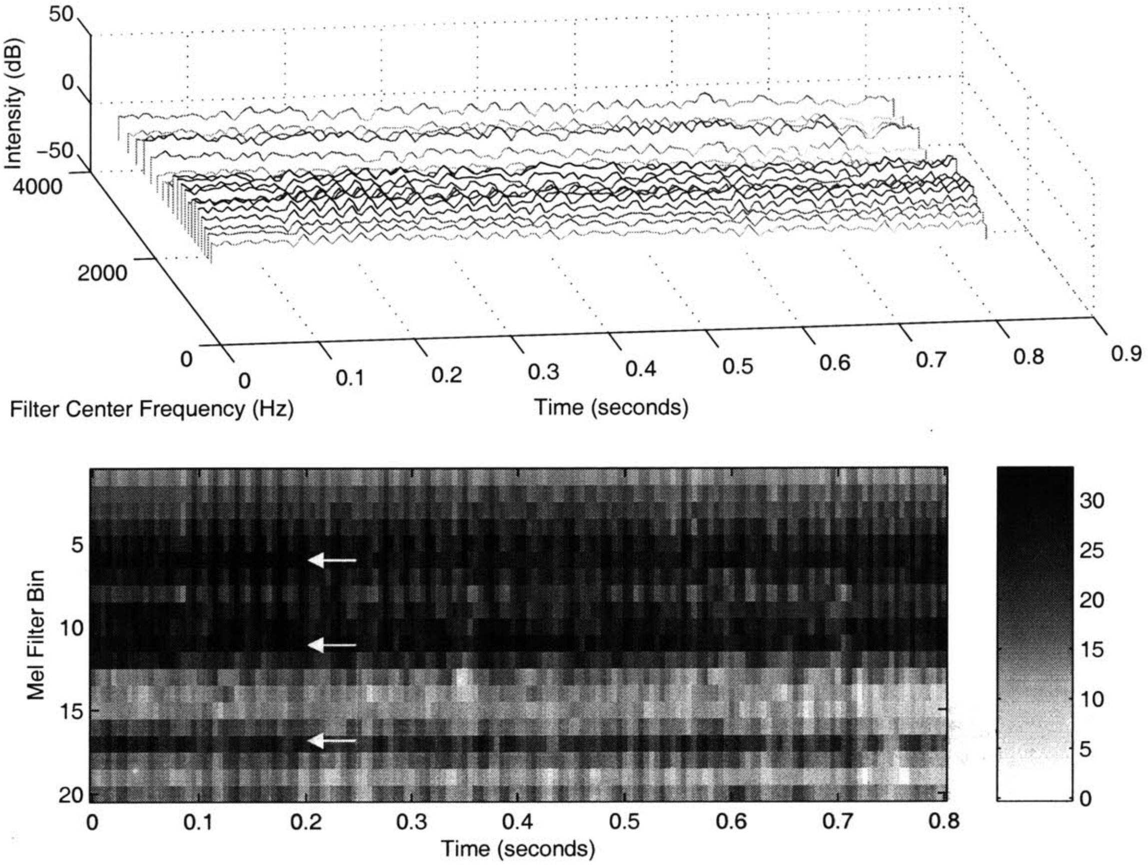

Figure 4-8. Response of the mel-filters to the voice signal DAS10. The formant frequencies are apparent in this representation as indicated by the arrows. An example of modulation can be seen left of the arrows in a region corresponding to period-tripling.

4.2 DPK Nonlinear Adaptation Model

As part of this thesis, a published model of the auditory periphery [7, 45]—referred to as the *DPK* model after the authors Dau, Puschel, and Kohlrausch—is used to represent the response of the human auditory system at the level of the cochlea and auditory nerve. In addition to a mel-scaled auditory filterbank, a nonlinear adaptation stage, enhancing the response to transient acoustic activity, is introduced. This model has been successfully used in previous research to improve both automatic speech recognition [44, 45] and speaker recognition [36] systems, adding a performance gain over mel-filtering models.

4.2.1 Model

The cochlea, which is the organ in humans responsible for the transduction sound waves into neural impulses, can be roughly described as a group of overlapping bandpass filters. In contrast to the linear filtering performed for the mel-filters, it is known that the response properties of the outputs of the cochlea are nonlinear. An interesting question, then, is whether the inclusion of such nonlinear elements in the feature extraction stage of an automatic recognition system can improve the recognition accuracy. It is known that the acoustic components of speech contain distinctive dynamics—rapid bursts of energy for stop consonants, for example, as well as abrupt changes in energy at the beginning and ends of words. In chapters 2 and 3, we saw that disphonic voices can also have rapid transients. Creak and irregular glottalization are two forms of dysphonic speech exhibiting sharp glottal pulses.

Adaptation is defined as the changing response over time to an input stimulus. In a typical model of adaptation in the auditory system such as [28], the response to a transient signal is initially large, eventually decaying to a lower, nearly constant, response. The DPK adaptive model studied in this thesis has two major properties: (1) a near-linear response to very quickly varying parts of signals and (2) a compressive transformation of slowly-changing components. As will be discussed, these characteristics are accomplished using five stages of serially arranged feedback loops, each with a different time constant depicted in Figure 4-11. Time constants are derived in the original paper [45] using a curve fit: "by combining several linear [on a dB scale] decays with different slopes, we get a piecewise approximation to the well known forward-masking curve." In this way, the model is phenomenological, fitting data to human perceptual studies rather than modeling actual neural responses as is done in [28, 42].

A primary biological motivation for the compressed response to static stimuli is the nonlinear nature of the cochlea. Three prominent factors are (1) the nonlinear amplification of acoustic stimuli by the cochlea, (2) the nonlinear opening of transduction channels in the stereocilia, and (3) the saturating response of the transduction to neural spike rate function for a single auditory nerve fiber. Details of these systems are well-studied, but are not modeled explicitly in the DPK model. Rather the authors of the DPK model state that "the output of five consecutive adaptation loops is then [the 32nd root of the input], which approximates the logarithm of the input signal." [45]. Logarithmic compression is commonly used to model the static nonlinearity in basic models of the auditory periphery; it is also present to some extent in the mel-filtering model.

In the human, the adaptive portion of the nonlinearity is thought to occur at the synapse of the inner hair cell with the auditory nerve, primarily through complicated dynamics of the depletion of available neurotransmitter [28]. Adaptation is not yet completely understood, but most of the literature seems to point to only two time constants on a short (less than 1 second) time scale. These are typically referred to as the "rapid" and "short-term" adaptation effects, with the first having a time constant around 1-10 ms and the second on the order of 60 ms [48]. The DPK model's use of five time constants reflects the decision to have a stationary signal be transformed logarithmically.

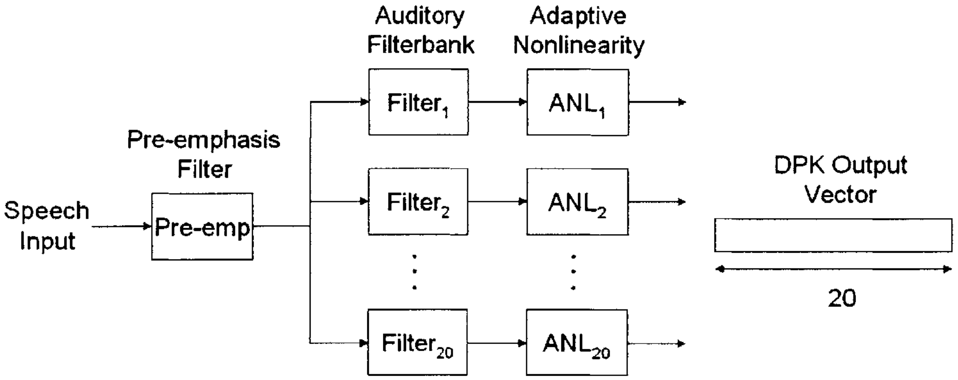

Figure 4-9. Schematic view of the DPK model. The pre-emphasis filter is the same as described for mel-filtering. The auditory filterbank has spacing and bandwidths based on the mel-scale as

described in the text and accompanying figures. The nature of the adaptive nonlinearity is to enhance transient activity and suppress the response to static portions of the stimulus.

The organization of the DPK model, as depicted schematically in Figure 4-9, is in part similar to that of the mel-filtering model. As in mel-filtering, the first portion of this system consists of pre-emphasis followed by an auditory-like filterbank. In this case, however, the filterbank is implemented through convolution instead of weighting of the magnitude spectrum. The filter responses are calculated using an overlap-add paradigm as described in [31]. The incoming signal is first split into frames using 20-ms triangular windows, each overlapping the previous by 50 percent and then the 512-point DFT is taken. Pre-emphasis, as shown in Figure 4-5 is applied to the signal in the frequency domain. Then the DFT is multiplied by the 512-point DFT of each filter, and the IDFT is taken of the result to yield a 512-point time-domain sequence. These time-domain blocks are summed together at each sample to reconstruct the filtered signal.

Twenty symmetric Gaussian filters make up the analysis filterbank shown in Figure 4-10. Each of these filters has an impulse response of 32.5 ms. Between 300 Hz and 1000 Hz, the filter center frequencies are linearly spaced, whereas they are logarithmically-spaced above this point. This mel-spacing scheme is equivalent to that used for the triangular weighting functions implemented in the mel-filtering model. Each of the filters is normalized to equal energy.

The nonlinear adaptation portion of the design consists of five loops in series as shown in Figure 4-11. For quickly-varying stimuli, the response to a change in the output is linear; for more slowly varying input, the output approaches the 32nd root. Thus, a nonlinear compression is introduced for the static regions of the input signal. In the diagram, the response of each of the five loops can be seen to be controlled by a free parameter, β, which determines the time constant of the response decay. In the thesis, β_1 – β_5 are equal to 0.90, 0.99, 0.997, 0.9985, and 0.9991 respectively.

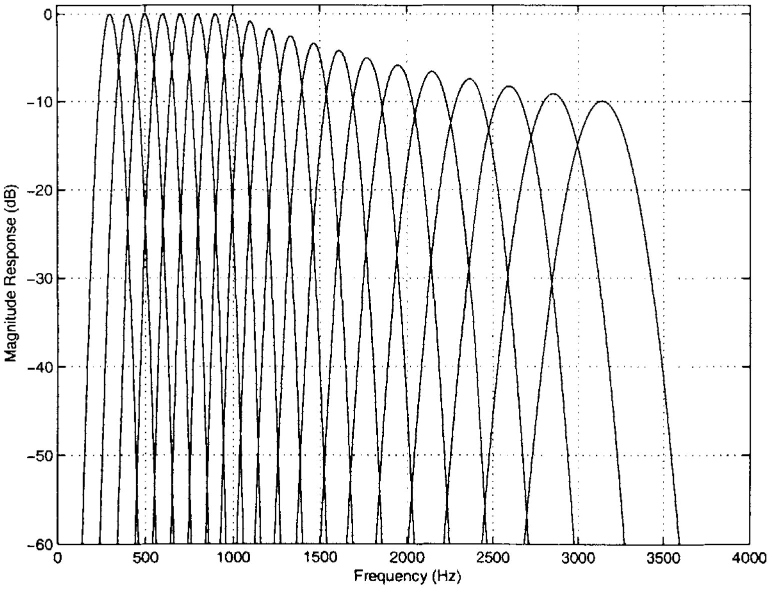

Figure 4-10. Frequency response of the Gaussian filters used to mimic the human cochlear response in the DPK model.

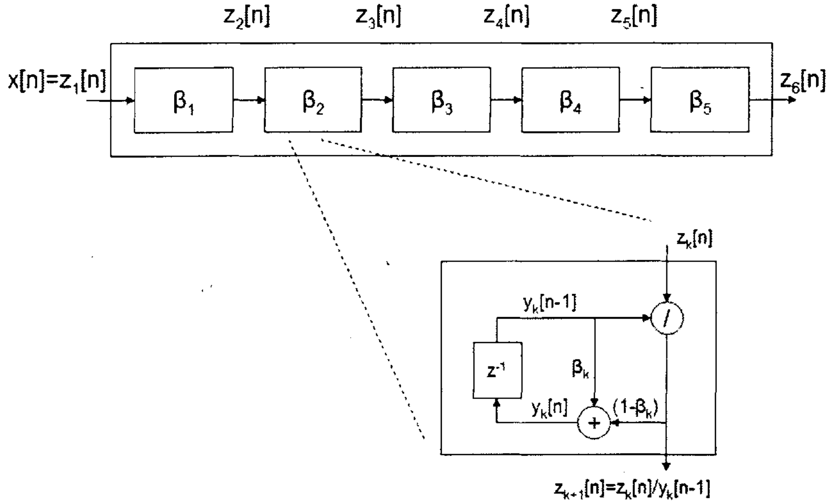

Figure 4-11. Architecture of the five-stage adaptive nonlinearity used in the DPK model. Each β denotes a constant which controls the response characteristic of a given stage. Each loop is a first-order recursive low-pass filter. A larger β tends to yield a slower adaptation. The steady-state response of each stage is the square root of the input, whereas the change in the response to a quickly-changing component is the current input divided by the square root of the previous input.

4.2.2 Representations of Amplitude Modulations

Figures 4-12 and 4-13 depict the output of the DPK model to the same two amplitude-modulated stimuli discussed for the mel-filtering examples. Despite the similarities between the mel-filtering and DPK models, the responses look remarkably different. The first observation in Figure 4-12 is that the high-frequency output channels have a distinct response, even though the stimulus energy does not reach that far in frequency. The response energy at all frequencies is much larger relative to that at the stimulus frequency than for the mel-filtering model. One might describe the response as noisy, with small changes in signal resulting in large transient spikes. An attempt is not made to explain how this behavior arises, but it is likely linked to sensitivity of the nonlinearities when the amount of incoming signal is very low. The modulated noise in Figure 4-13 is an example of how it is difficult to discern the envelope imposed on the noise carrier from the model outputs. Modulations are observed at the 27-Hz rate, but also many sharp and seemingly irregular spikes are seen. These observations suggest that the DPK method may produce arrant spikes in response to a low-amplitude and/or noise-source inputs.

Another observation is that the shape of the response to the 30-Hz modulator appears sharper and more defined than for mel-filtering in Figure 4-12. Variation at this rate is visible near 1000 Hz, especially in the two-dimensional rendering. The peaks and troughs of the modulating waveform are easy to see, whereas they appear more subtle in the mel-filtering plot. The generation of spikes at center frequencies away from the carrier is probably an artifact of the system and was not observed with natural stimuli. Near the 4000 Hz center frequency, for example, we see spikes occurring at a rate of about 10 Hz.

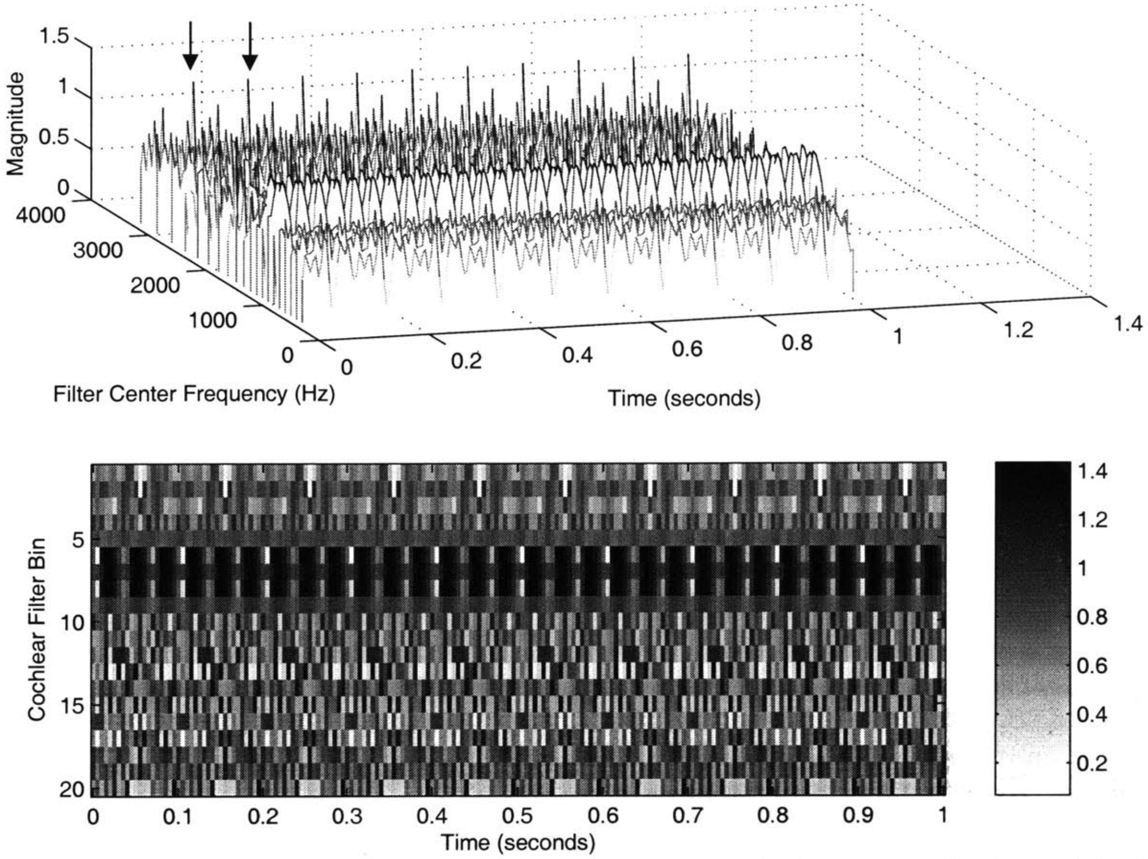

Figure 4-12. Response of the DPK model to a 1000-Hz sinusoidal carrier amplitude-modulated by a 30-Hz sinusoid. The arrows highlight the high-frequency pulses of unknown cause noted in the text. Fluctuations near the 1000-Hz carrier are enhanced compared with the outputs of mel-filtering.

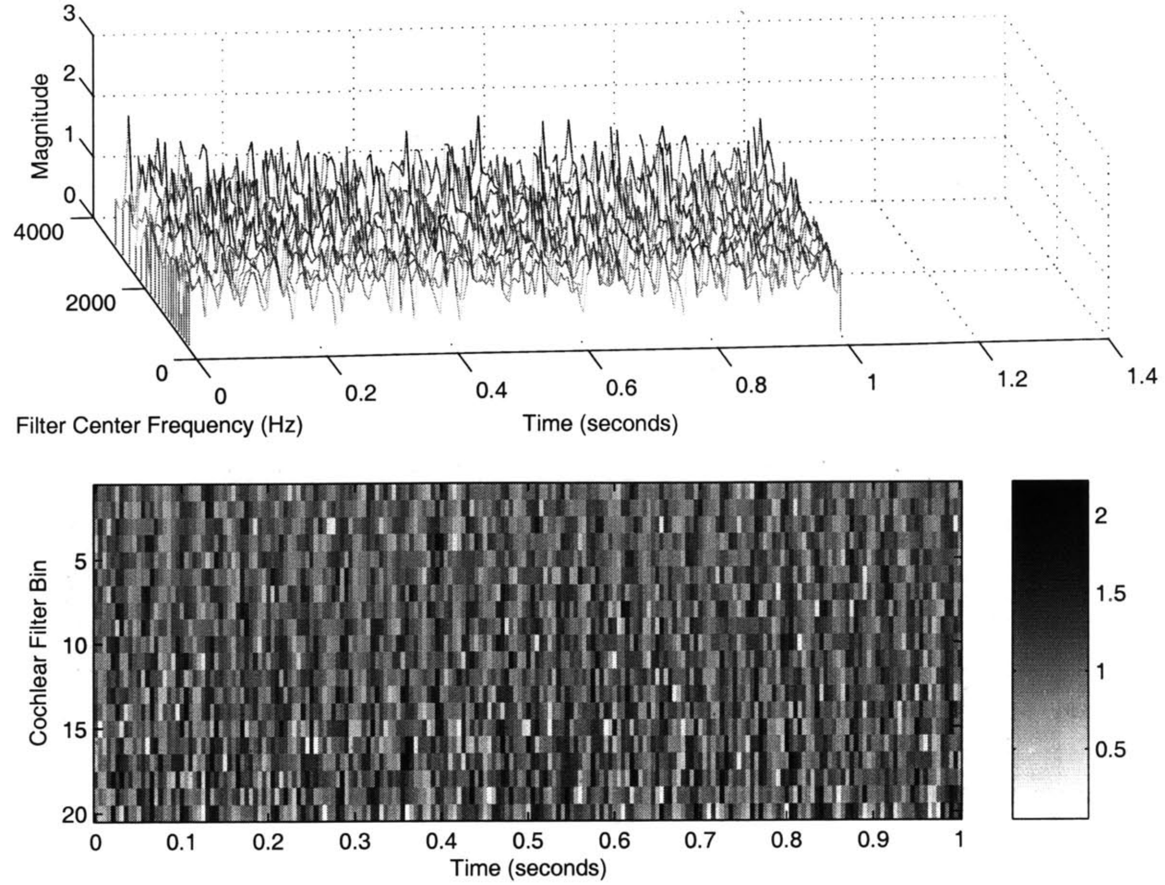

Figure 4-13. Response of the DPK model to a white-noise carrier amplitude-modulated by a 27-Hz sinusoid. There is a lack of a clear periodically-fluctuating envelope due to the sensitivity of the DPK model to the noise carrier signal.

Figure 4-14 shows the response of the DPK model to the middle 800 ms of the utterance by patient DAS10, described in chapter 3. Recall that this particular voice has a creaky quality, with large pulses occurring at various intervals. In the figure, we see that these pulses, especially at 350 ms for example, yield appropriately sharp responses in the DPK model output.

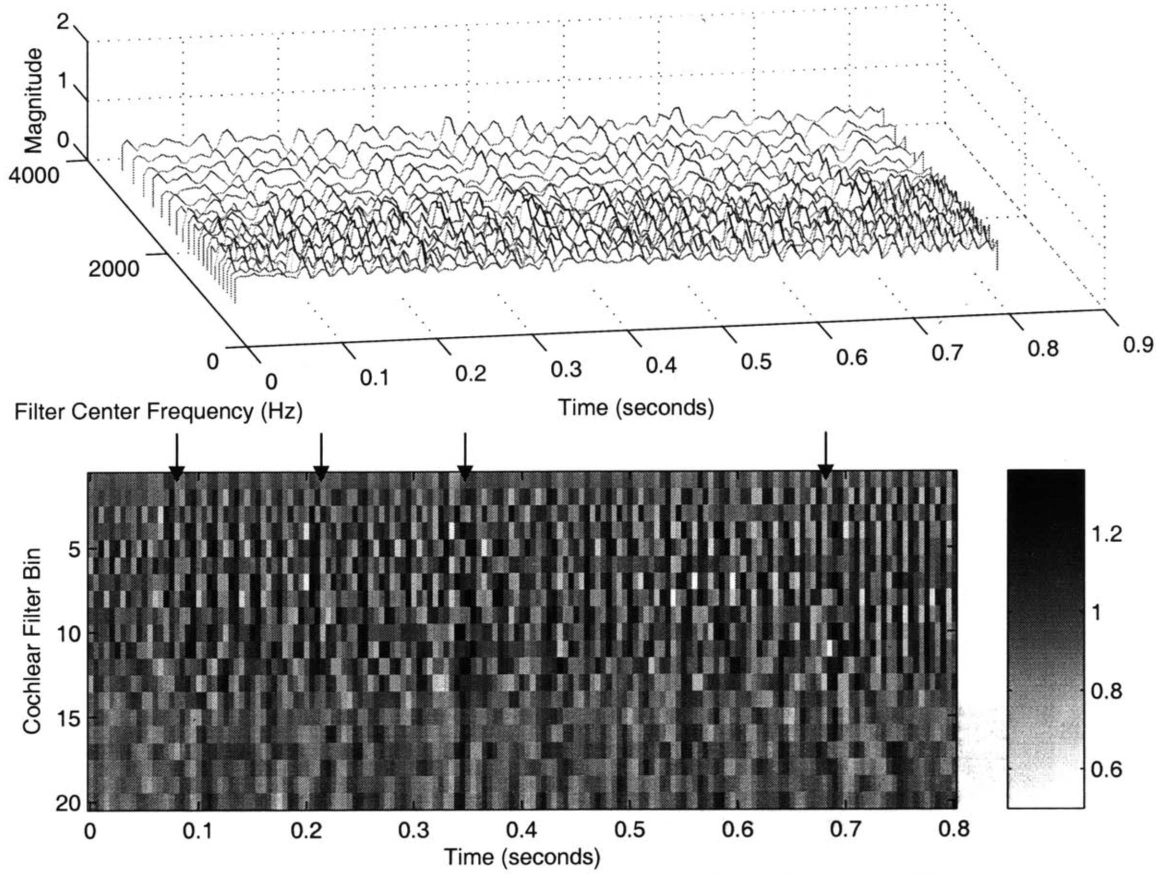

Figure 4-14. Response of the DPK features to the sustained vowel in file DAS10. Large pulses in the stimulus—plotted in earlier analyses—give sharp responses in the output. The arrows highlight several points where the DPK model accentuates large transients.

4.3 ICC Modulation Filterbank Model

The biological motivation for human sensitivity to the amplitude modulation of auditory stimuli comes from work with a structure in the auditory midbrain called the *inferior colliculus*, or *ICC*. This structure is thought to contain regions that are sensitive to amplitude modulations from about 10-to-1000-Hz as detailed by [26]. This work was exploited in the literature by the same authors of the DPK model to arrive at a model containing modulation-sensitive channels [7].

The purpose of the biologically-motivated ICC technique is to explicitly represent the amplitude modulation content of a signal by having modulations appear as excitation in a specific output band. This process adds a second analysis stage to the general demodulation process discussed in chapter 3. Using the same bandpass analysis filter structure as the DPK model, the model extracts a series of time-varying envelopes which are then frequency analyzed by a bank of *modulation filters*.

4.3.1 Model

The ICC model represents modulations at different points across the frequency spectrum. As depicted in Figure 4-15, the first stage of the system uses the DPK model described in section 4.2, including pre-emphasis, the same group of twenty overlapping bandpass filters, and identical compressive nonlinearity stages. Each of the twenty DPK model outputs is then processed through yet another filterbank, referred to as the modulation filterbank. This set of thirteen biologically-motivated modulation filters, detailed in Figures 4-16, 4-17, and 4-18, extracts modulations within the range specified by the passband. This method separates out each carrier signal using the auditory filterbank and then determines the amplitude modulations applied to this carrier by using the modulation filterbank [36].

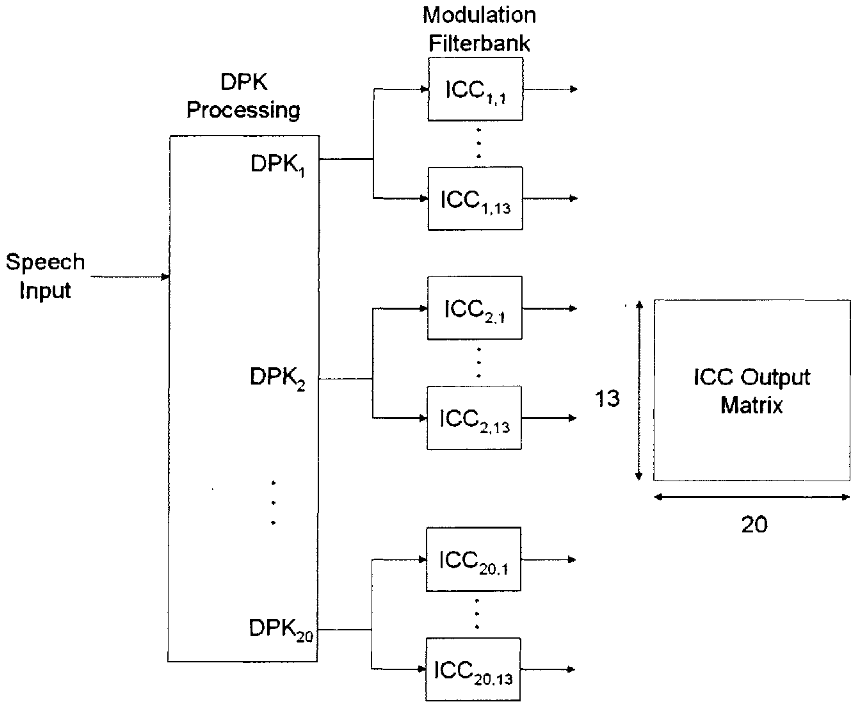

Figure 4-15. Schematic drawing of the baseline ICC model. The DPK block indicates a one-input, twenty-output system equivalent to that described in section 2.2, containing pre-emphasis filtering, auditory filtering, and adaptive nonlinearity blocks. The modulation filter subscripts indicate the DPK output number followed by the modulation channel number.

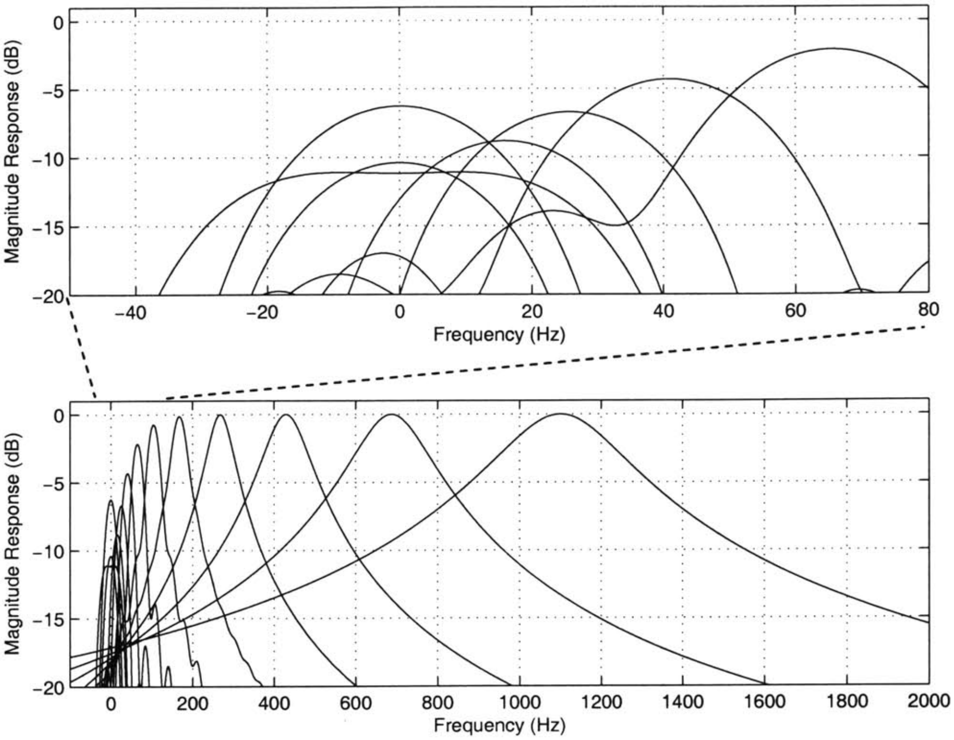

Figure 4-16. Frequency response of the fast modulation filterbank used to process each of the outputs of the auditory front-end. Only the three lowest frequency filters have real impulse responses, while the others are imaginary and have a one-sided frequency response. The figure shows that the improved speed of this method discussed in the text is traded off with distorted low-frequency filters.

Three different versions of the modulation filters are used in the experiments in this thesis. They are denoted the *fast*, *accurate*, and *narrow-spacing* versions. As with the DPK implementation, all convolutions are performed using a block convolution overlap-add method with 512-point and 2048-point DFTs as appropriate. For the *fast* filters, designed after previous work [36], thirteen modulation filters are used, having center frequencies of 0, 5, 10, 16, 25.6, 40.96, 65.536, 104.85, 167.8, 268.4, 429.5, 687.1, and 1099.5 Hz. Each of these filters has an impulse response that is cut off at 32.5 ms for efficiency. As detailed in the top panel of Figure 4-16, the response of the first three filters are symmetric, whereas the remaining filters are complex and thus yield the analytic signal. The speed increase from using shorter impulse responses comes at a price—the low frequency modulation filters are blurred, yielding a low resolution.

The second version of the modulation filters, the *accurate* version uses 225 ms of each filter impulse-response instead of 32.5 ms. This change requires a longer FFT to be used and makes the implementation slower by about a factor of four. As depicted in Figure 4-17, this understandably causes the filters to be narrower in the frequency domain and to exhibit less side-band activity. The third, *narrow-spacing* filters shown in Figure 4-18, are similar to the accurate filters, but are active across a much smaller frequency range and have correspondingly narrower bandwidths. In order to prevent a strong response at DC and at frequencies out of the range of many noted dysphonia amplitude modulations, the thirteen center frequencies are placed at frequencies 12, 14.4, 17.28, 20.7, 24.9, 29.9, 35.8, 43, 51.6, 61.9, 74.3, 89.2, and 107 Hz. The highest modulation channel is at nearly one-tenth of the highest frequency in the fast and accurate filters. Neither the accurate or narrow-spacing filters use the nonlinear stage of the DPK model. The auditory filter outputs are used directly in order to preserve the spectral characteristics of each band for the modulation filterbank analysis; the choice is a tradeoff between reduction of spectral spread and enhancement of transient activity.

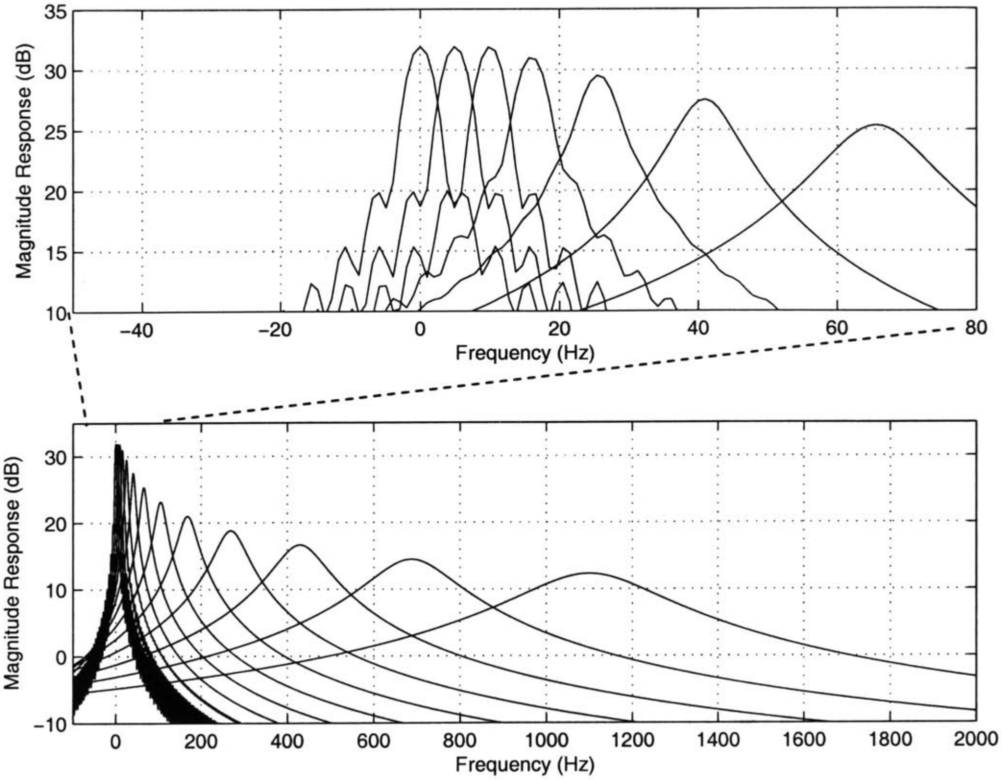

Figure 4-17. Frequency response of the accurate modulation filter-bank used to process each of the outputs of the auditory front-end. This version uses 225-ms impulse response length as compared to the 32.5 ms used for the fast ICC model. As shown, only the lowest frequency filter has a real impulse response; the others are imaginary and have only a one-sided frequency response. This figure shows that the reduced speed of this method is traded off with more accurate low-frequency filters.

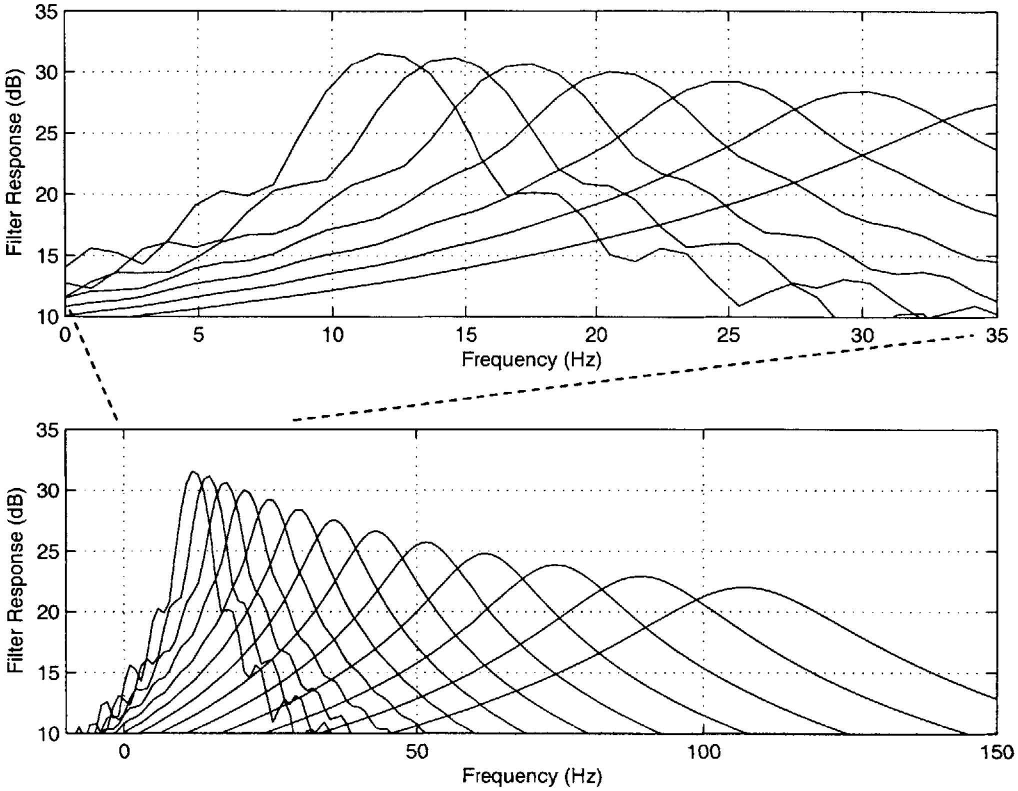

Figure 4-18. Frequency response of the narrow-spacing modulation filter-bank used to process each of the outputs of the auditory front-end. This version uses 225-ms impulse response length as well as significantly narrower spacing when compared to the fast ICC. All filters have imaginary impulse responses and are thus one-sided in the frequency domain. Compared with the accurate filters, the center frequencies as well as the bandwidths are reduced and the filters are more-closely spaced.

4.3.2 Representations of Amplitude Modulations

Figures 4-19 and 4-20 depict the response of the narrow-bandwidth ICC system to the same modulated sinusoidal and noise carriers used for the mel-filtering and DPK studies. Again, only the middle one second of the three second stimuli is shown in order to focus the discussion on the response to amplitude modulations and not onsets and offsets of the signals. The responses to both of the above synthetic stimuli show peaks around the 6th modulation filter bin having a center frequency at 29.9 Hz. This corresponds to the modulation frequencies in the stimuli. From the 2D plots, especially for the noise carrier, one can also observe time variations in the trajectory at the modulation frequency which is somewhat unexpected. We speculate that this phenomenon is due to the interference of the additive DC term with the modulating sinusoid during the analysis stage—*artifact AM* as discussed in the chapter 3.

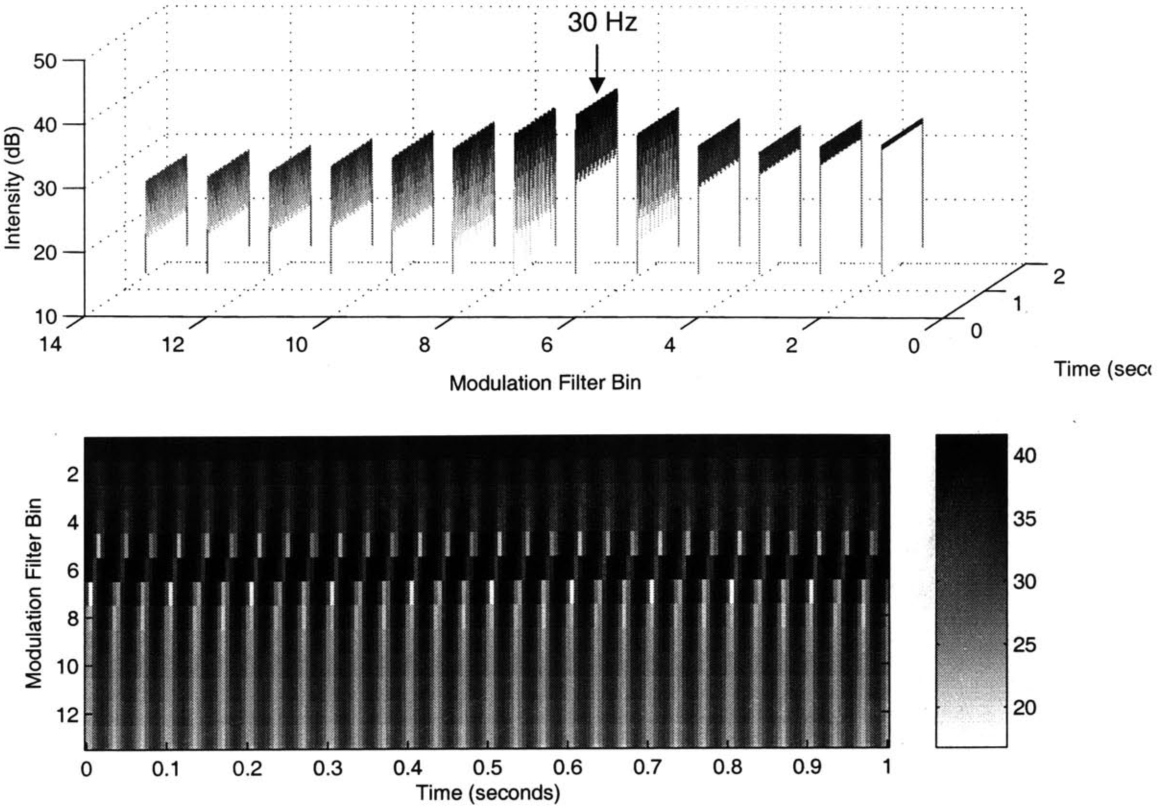

Figure 4-19. Response of the narrow-spacing ICC model to a 1000-Hz sinusoidal carrier amplitude modulated by a 30-Hz sinusoid. The modulation-filter center frequencies are at 12, 14.4, 17.28, 20.7, 24.9, 29.9, 35.8, 43, 51.6, 61.9, 74.3, 89.2, and 107 Hz. The peak in the filter energy occurs near 30 Hz.

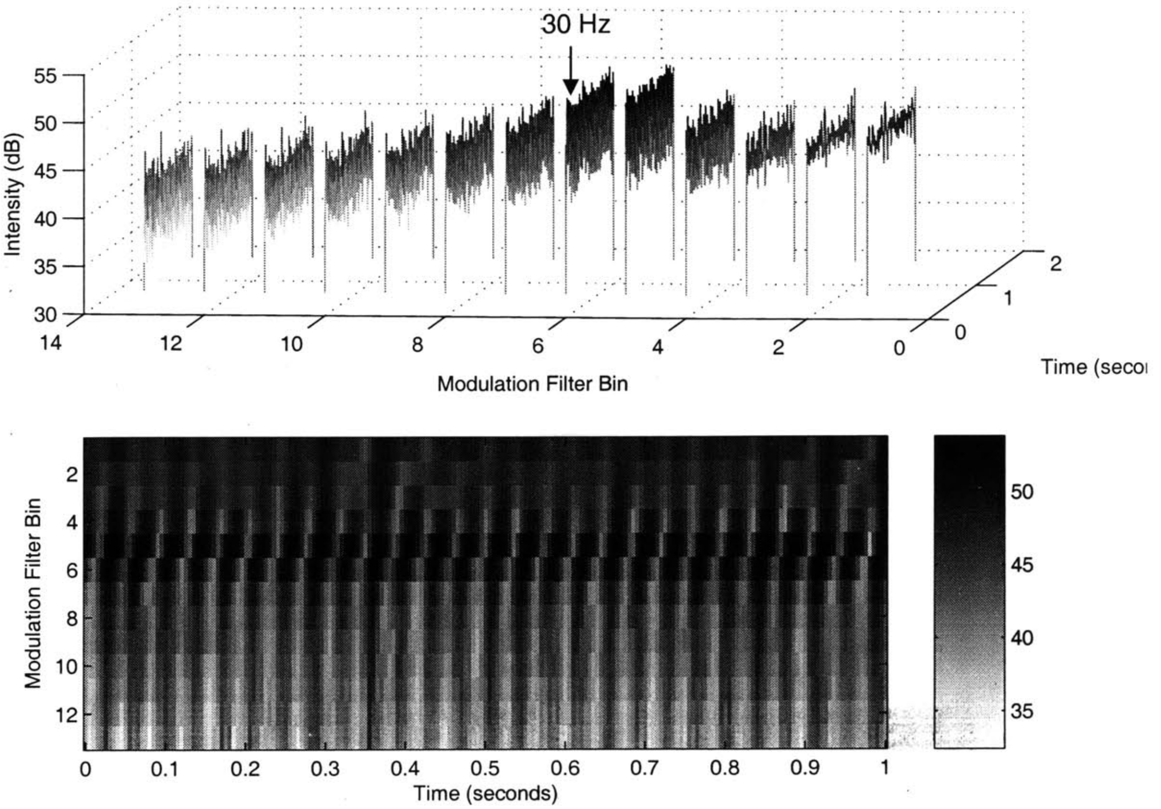

Figure 4-20. Response of the narrow-spacing ICC model to a white noise carrier amplitude modulated by a 27-Hz sinusoid. The modulation-filter center frequencies are at 12, 14.4, 17.28, 20.7, 24.9, 29.9, 35.8, 43, 51.6, 61.9, 74.3, 89.2, and 107 Hz. The peak in the filter energy occurs near 30 Hz.

Figures 4-21, 4-22, and 4-23 compare narrow-spacing ICC model representations of the middle 800 ms of three dysphonic voices. These three voices—JAB08, DAS10, and KAH02—were analyzed in chapter 3 and are synthesized in appendix A. The prior analysis revealed amplitude modulations on different scales for each utterance. With JAB08, there was strong AM below 10 Hz as well as some period-doubling at the end of the utterance. Figure 4-21 shows time-varying activity, with strong modulation near 12 Hz throughout the utterance. There is also evidence of 30-Hz modulation at 300 ms in the figure and energy near 60 Hz at both 0 and 500 ms.

Recall that DAS10 contains glottalization-like responses of several varieties, changing throughout the stimulus. In Figure 4-22, the modulation activity moves from 70 Hz at the signal onset to a strong region near 50 Hz, then to a region of complicated AM over many frequencies, and finally returns to a steady modulation around 70 Hz. Activity at the lower frequencies is complicated, with the trend being a general decrease with time. It is also interesting to note that the large irregular glottalization pulses in the middle of the stimulus seem to cause correspondingly large response pulses throughout most of the modulation frequencies—see the response at 270 ms and 370 ms.

Figure 4-23 shows the response to KAH02, that was an example of a dysphonic voice with complex interacting harmonics. The ICC response shows modulations near 50 Hz and 30 Hz that agree with the difference frequencies between line components as seen in the previous analysis. Overall, the ICC model reveals details about how the signal changes with time and that are not apparent in the mel-filter representation or the DPK model outputs.

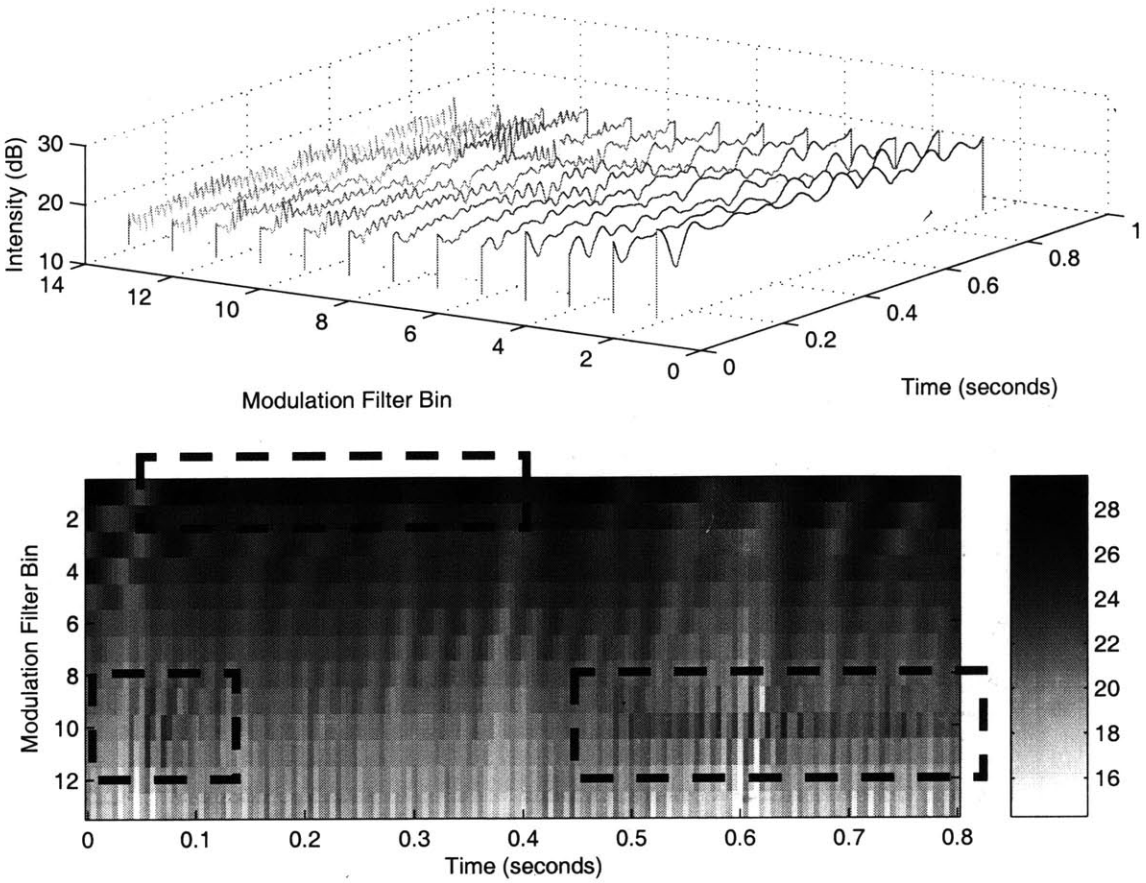

Figure 4-21. Response of the narrow-spacing ICC model to a sustained vowel produced by patient JAB08. The modulation-filter center frequencies are at 12, 14.4, 17.28, 20.7, 24.9, 29.9, 35.8, 43, 51.6, 61.9, 74.3, 89.2, and 107 Hz. The dashed boxes highlight regions of high modulation activity in the image.

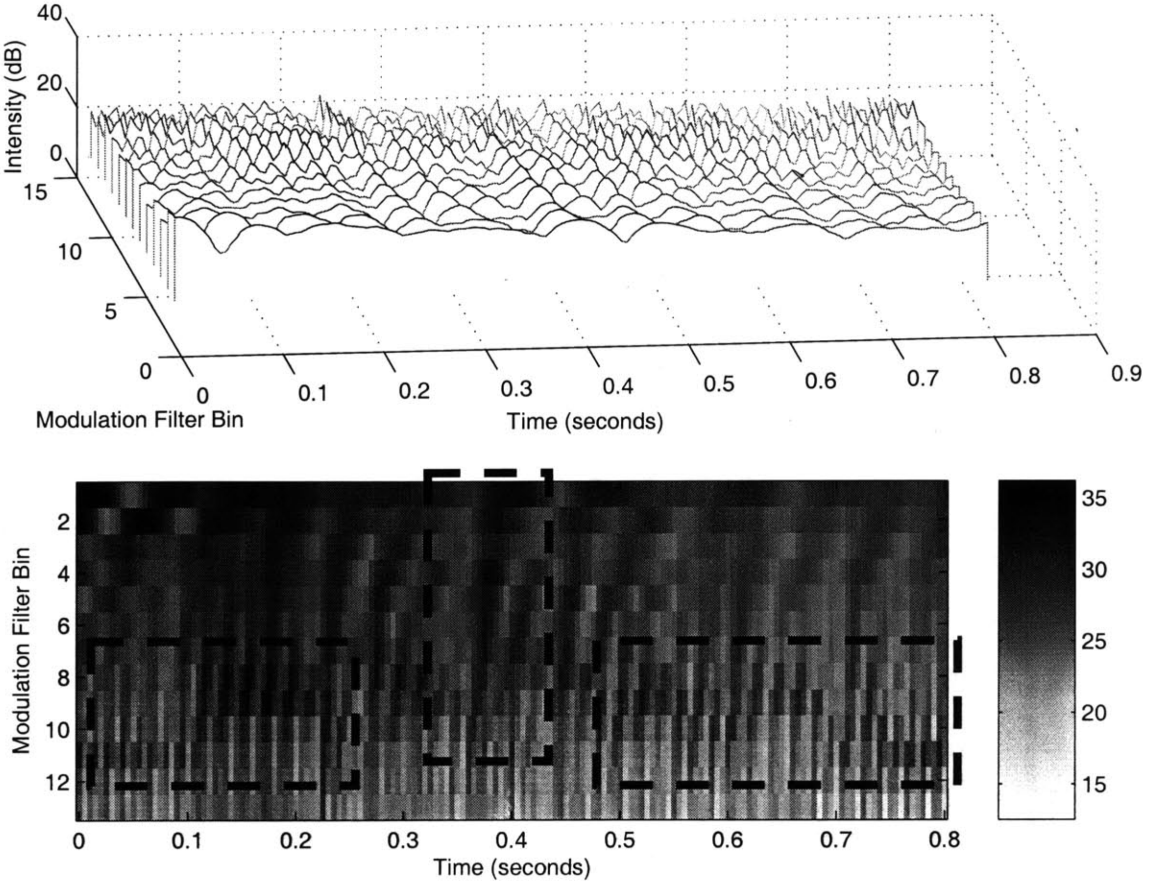

Figure 4-22. Response of the narrow-spacing ICC model to a sustained vowel produced by patient DAS10. The modulation-filter center frequencies are at 12, 14.4, 17.28, 20.7, 24.9, 29.9, 35.8, 43, 51.6, 61.9, 74.3, 89.2, and 107 Hz. The dashed boxes highlight several regions of high modulation. The modulation begins near 75 Hz, becomes lower and then increases again near the signal end.

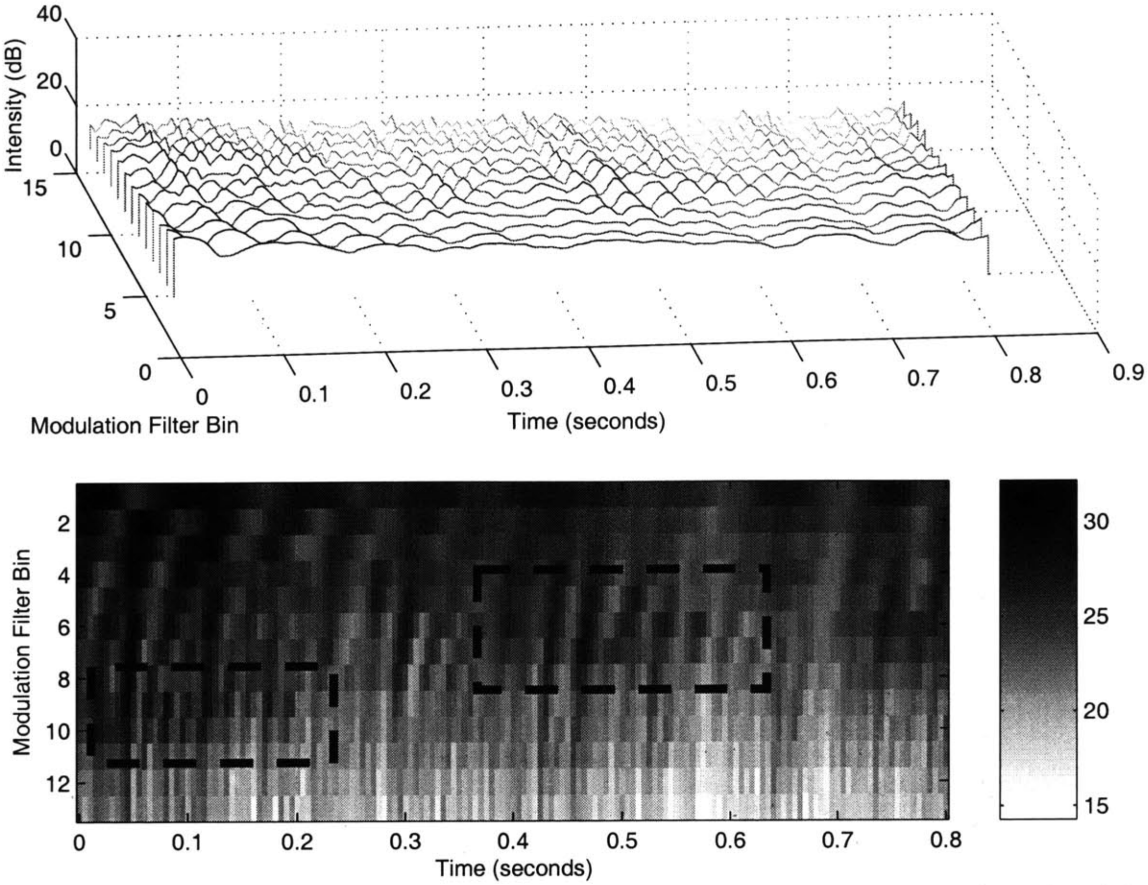

Figure 4-23. Response of the narrow-spacing ICC model to a sustained vowel produced by patient KAH02. The modulation-filter center frequencies are at 12, 14.4, 17.28, 20.7, 24.9, 29.9, 35.8, 43, 51.6, 61.9, 74.3, 89.2, and 107 Hz. The dashed boxes highlight active regions near the 50-Hz and 30-Hz modulation frequencies.

4.4 Conclusions

This chapter has addressed three different classes of biological-inspired front-end models for the analysis of amplitude modulations in speech. The first, mel-filtering, is used with the standard mel-cepstral feature extraction technique. The second, the DPK nonlinearity model, captures quickly-changing envelopes while suppressing those that remain static. Finally, the ICC modulation-filtering model is designed to capture time variations of envelopes extracted using an auditory-motivated analysis filterbank.

Each model has specific traits that make it unique and useful. The mel-filtering model, which is based on short-time Fourier transform analysis has the advantage of speed and represents spectral components such as formant frequencies. Its downside is that it is based on uniform bandwidth filters and blurs time information at high frequencies compared to models based on variable bandwidth filters [34]. In contrast, the DPK model uses a logarithmic bandwidth filterbank in combination with an adaptive nonlinearity stage to enhance fast AM in a signal. Compared with mel-filtering, the DPK approach does not blur quickly varying signals but is sensitive in frequency regions of low energy and/or noise. The DPK representation gives transient activity, for example with creak, but blurs out spectral formants. The ICC model adds an additional analysis step to the DPK method's logarithmic filterbank, allowing individual modulation components to be separated, automating what was observed in the spectrograms of chapter 3. The ICC representation has the advantage of characterizing amplitude modulation information but, at the same time, removes an explicit representation of formants and other spectral characteristics.

In the next chapter, we incorporate our three models into an automatic dysphonia recognition system. The system is designed to use the AM extraction methods presented in this chapter to capture the acoustic fluctuations in disordered speech. In order to exploit the complementary advantages of each technique, we also introduce a technique with which to optimally combine the results obtained with the different models.

Chapter 5

Automatic Dysphonia Recognition

A dysphonia recognition system consists of front-end modeling, feature extraction, and pattern classification stages as depicted in Figure 5-1. This architecture is also common to other recognition systems such as speech, speaker, and language recognition systems [39]. In our work, the front-end is one of the auditory models discussed in previous chapters. The purpose of this stage is to transform a speech signal into a lower-dimensional representation that provides speech attributes such as formant structure and amplitude modulations. From this front-end, the model outputs are sent to a feature extraction stage which samples the signals and performs post-processing tasks such as channel normalization. Pattern classification is the final stage, where a stream of feature vectors is analyzed in order to compute the likelihood that the original speech is from a patient with a certain voice disorder.

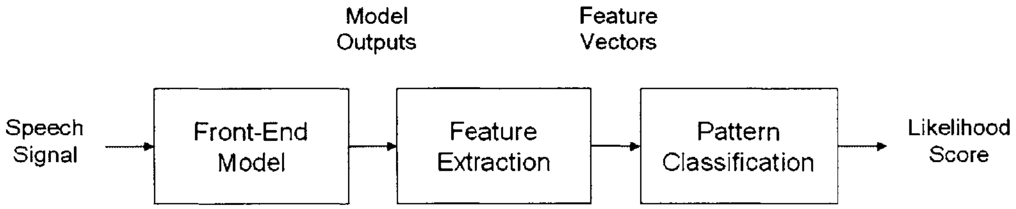

Figure 5-1. Overview of a dysphonia recognition system. A speech signal enters and is processed by an auditory model to produce a series of output signals. From these outputs, features are extracted and sent to an automatic classifier which estimates how likely it is that an utterance indicates a certain voice disorder.

The purpose of this chapter is to discuss the application of the auditory models described in chapter 4 to the task of automatic dysphonia recognition. Specifics on the setup of each experiment are given as well as a full description of the techniques to combine results obtained using different features extraction methods. A comparative analysis is presented of the relative performance of the features derived from variations of the auditory models and also illustrates advantages gained by optimally fusing recognition results from different extracted features.

5.1 Feature Extraction

Feature extraction is the process of sampling the outputs of a front-end model such as mel-filtering, nonlinear adaptation (DPK), or inferior colliculus modulation filtering (ICC) and using various post-processing schemes to enrich the data. This section describes the feature extraction stages for each of the three main models of this thesis. Details of feature extraction are varied within experiments and may differ slightly from the descriptions given below. The intent is to provide the most commonly used feature extraction methodology for each model.

5.1.1 Mel-Cepstral Feature Extraction

In chapter 4, we described the front-end filtering model used in the mel-cepstrum as a series of triangular filters. In addition to this stage, the standard mel-cepstrum involves computing a quantity called the inverse discrete cosine transform (IDCT) of the log energies of the model outputs [34].

In Davis and Mermelstein's [8] description, *mel-frequency cepstral coefficients* are calculated as:

$$MFCC_i = \sum_{k=1}^{20} \log\{E_{mel}(i,k)\}\cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{20}\right] \qquad i=1,\ 2,\ \cdots,\ M$$

where i is the index of the coefficient, M denotes the number of coefficients being taken—Davis and Mermelstein use both 6 and 10—and $E_{mel}(i,k)$ is the energy of the k^{th} mel-filter output. In order to calculate the output energy of a given auditory filter, k, the squared discrete Fourier transform (DFT) magnitude of the input is weighted using one of the 20 triangular functions shown in chapter 4 and the sum is taken across the DFT bins.

Quatieri [34] has a similar formulation to that used in this chapter's experiments given as:

$$C_{mel}[n,m] = \frac{1}{R}\sum_{k=0}^{R-1} \log\{E_{mel}(n,k)\}\cos(\frac{2\pi}{R}km)$$

As above, $E_{mel}(n,k)$ denotes the magnitude squared of the k^{th} output of the mel-filter bank at sample n; m indicates the cepstral coefficient number and R equals the total number of coefficients. Quatieri's description of feature extraction is different from Davis and Mermelstein's in that the number of filters and cepstral coefficients are equal.

Feature extraction for the mel-cepstral features is performed by first uniformly sampling the mel-cepstral output every 10 ms, equivalent to every 80 samples at a sampling rate of 8000 Hz. The zeroth bin of each mel-cepstral feature vector, corresponding to a multiplicative scale factor of the speech signal, is then removed. For the dysphonia recognition experiments described in following chapters, this output is then processed using silence removal, channel normalization, and feature differencing stages. The first step selects which output frames contain speech and which contain silence so that the final feature sets reflect only speech regions.

Two different types of channel normalization were performed. The first type, *cepstral mean subtraction*, or *CMS*, calculates and subtracts the mean value of each feature across all the voiced frames [34]. In the cepstral domain, such a subtraction corresponds to removing any constant filtering of the signal, which is presumed to arise from recording-dependent properties of the signal path. For example, such processing is intended to remove the relative filtering effects of different telephone microphones and transmission lines. *RASTA*, the second type of channel normalization, filters the trajectory of each feature with time, emphasizing those fluctuations at rates between about 0.26 Hz and 12.8 Hz [17]. Such modulations are important to the linguistic information contained in the signal while other fluctuations are viewed as noise. For the mel-cepstrum, such processing has been shown to improve speech recognition in the presence of a convolutional channel.

The last stage of the feature extraction system, referred to as the delta-feature calculation, finds the difference between the values of neighboring frames. The goal is to capture signal changes with time. In the implementation used in this thesis, frames 20 ms forward and 20 ms backward are taken into account when calculating these delta features through a form of interpolation. The result of these three processing stages are 19 mel-cepstral features and 19 delta-cepstral features for a total of 38 features per output vector.

5.1.2 Dau, Püschel, and Kohlrausch (DPK) Feature Extraction

In the DPK feature extraction method, features are uniformly extracted from the output of the DPK model every 5 ms, corresponding to every 40 points at a sampling rate of 8000 Hz. As described for the mel-cepstra, the IDCT is calculated so as to mimic the latter stages of mel-cepstral processing. At this point, the zeroth bin of each vector is discarded and speech detection, channel normalization—including RASTA and CMS—and delta-features are used. The resulting 19 DPK and 19 delta features yield a total of 38 features as output by the feature extraction stage.

5.1.3 Inferior Colliculus (ICC) Feature Extraction

Extracting a relatively short vector from the ICC model is not simple. In particular, as shown in Figure X, the technique yields a very large (20 by 13 = 260 element) matrix for each frame. Such an output is inefficient and in order to interact best with automatic systems, must be minimized. For this thesis, two different techniques are used to reduce the model output, based on methods introduced in [36]. The first approach is to sum across all of the auditory channels yielding 13 outputs, one for each modulation channel. One interpretation of such a value is as a "common modulation" rate across auditory channels. This technique is designated *ICC summation* and is described by the following formula where $Y_s[m,n]$ is the sum for the m^{th} modulation filter channel and the n^{th} sample and $X_l[m,n]$ is the magnitude of the signal at this point :

$$Y_s[m,n] = \sum_{l=1}^{R} X_l[m,n]$$

The second approach is to calculate the centroid of the energy across modulation channels, yielding one output for each of the twenth auditory filter bands. This output can be interpreted as an indicator of the "information rate" in the signal at a given time, with more active modulation bands contributing more to the centroid value. This technique is designated *ICC centroid* and is described by the following formula where $Y_c[m,n]$ is the centroid for the m^{th} auditory filter channel and the n^{th} sample, W_l is a weight proportional to the center frequency of the l^{th}

modulation filter, and $X_l[m,n]$ is the magnitude of the output of the l^{th} modulation channel of the m^{th} auditory filter:

$$Y_c[m,n] = \frac{\sum_{l=1}^{R} W_l X_l^2[m,n]}{\sum_{l=1}^{R} X_l^2[m,n]}$$

As with the DPK extraction technique, a feature vector is obtained from all versions of the ICC model output every 5 ms. The outputs of the different ICC summation and ICC centroid models, however, are often extracted differently. For the ICC summation models, the cepstrum is taken and the first bin of the feature vector discarded to take advantage of some of the presumed benefits of the cepstrum observed for mel-cepstral and DPK features. Speech detection, both forms of channel normalization, and delta features are used for these features as well. In contrast, because the output of the ICC centroid model is not directly the output of a filtering process, the cepstrum is not taken and RASTA is not implemented. CMS is applied, however, under the hypothesis that unchanging elements of the signal do not contain useful information about the speech signal. In total, then, the ICC summation method yields 12 modulation features and 12 delta features for a total of 24 outputs and the ICC centroid method yields 20 features plus 20 delta features for a total of 40 features per frame.

5.2 Pattern Recognition

The purpose of the classification stage in a recognition system is to label an utterance as coming from one or more of several groups. Figure 5-2 shows an overview of this process for dysphonia recognition, with a single utterance being assigned a percentage chance of containing each of several different voice disorders. As discussed in chapter 2, an example voice is classified as being part of each category. This is different from other recognition tasks where possible classes for an utterance are mutually exclusive.

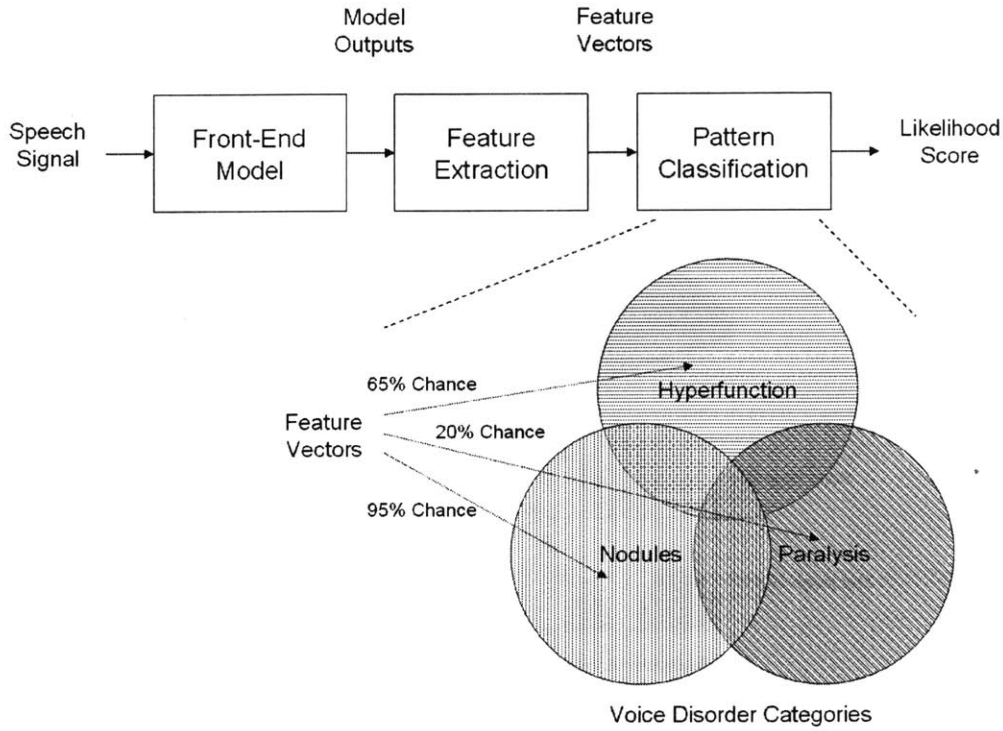

Figure 5-2. Overview of recognition system for dysphonia. The feature vectors from an utterance are obtained and the chance that they contain each of a set of three voice disorders is calculated. As discussed in chapter 2, voice disorder types are not mutually exclusive and a patient may suffer from them all.

5.2.1 Gaussian Mixture Model (GMM) Classifier

In this thesis, we use a pattern recognition scheme based on a Gaussian mixture model or *GMM*, of the probability density function (pdf) of feature vectors. The recognition system is a modified version of the existing Lincoln Laboratory GMM-based speaker recognition system [40] which is widely used for speaker and language recognition tasks. In order to describe the basic theory of a GMM-based recognizer, let us devise a mock recognition experiment depicted in Figure 5-3. Here, we show a hypothetical machine that measures the weight of a person and attempts to determine whether or not the object on the machine is an adult human,18-years-old or above, using this information. In this example, our feature set will simply be one weight measurement.

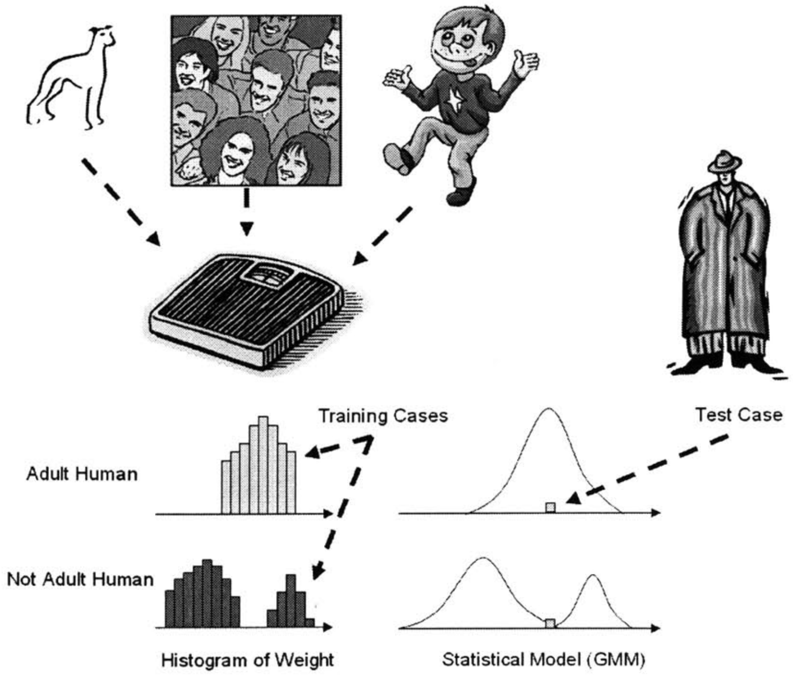

Figure 5-3. Hypothetical example demonstrating the action of a GMM-based pattern recognition system. The machine in the figure uses a single measurement of the weight of each subject and determines whether it is an adult human.

On the top left, we see the histogram of the measured weights of a large set of adults with Gaussian distributions fit to this data. This model is called the *target* model. In the bottom left, there is the histogram and GMM of measured weights from a large set of non-adult humans including children, animals, and other *background* cases. Using these models of targets and background cases, the goal is to correctly classify a *test* case, shown with relation to the models. In order to obtain a value related to the likelihood that the test case is a human adult, we first calculate a quantity called the log-likelihood-ratio value, computed as [34]. From the ratio of probabilities

$$\frac{P(\text{weight from adult human})}{P(\text{weight not from adult human})} = \frac{P(\lambda_C \mid X)}{P(\lambda_{\bar{C}} \mid X)} = \frac{p(X \mid \lambda_C)P(\lambda_C)/P(X)}{p(X \mid \lambda_{\bar{C}})P(\lambda_{\bar{C}})/P(X)}$$

we form the log-likelihood ratio

$$\Lambda(X) = \log[p(X \mid \lambda_C)] - \log[p(X \mid \lambda_{\bar{C}})]$$

where λ_C denotes the PDF of weights for adult humans and X is the weight of the test case. In order to make a decision as to whether a weight indicates an adult human or not, we must set a value, c, called the *criteria* or *threshold*, which the log-likelihood-ratio is compared against using the following rule:

$$\begin{array}{ll} \Lambda(X) \geq c, & \text{accept as adult human} \\ \Lambda(X) < c, & \text{reject} \end{array}$$

If a large number of test cases are used, we can obtain a *miss probability*—when an adult human is mistakenly classified as the other group—and a *false alarm probability*—when a non-adult-human subject is mistakenly classified as adult human. In order to characterize such an experiment, we use a plot called a detection error tradeoff or *DET* curve. This plot, which graphs miss probability versus false alarm probability for each possible value of *c*, allows us to characterize how well the system is able to recognize the adult human class. We can also use a value called the equal error rate, or *EER*, to represent the system—this is defined as the total percent correct at the value of *c* when miss and false alarm probabilities are closest together. We will use all three of these tools extensively in this thesis.

5.2.2 Extension of the GMM to Dysphonia Recognition

The above example illustrates in part our approach to dysphonia recognition. First, for each voice disorder, the voice database is split into two groups—those files from patients with a certain disorder and those files not from patients with the disorder. Detecting a voice differs from the adult human detector example, however, in that we are concerned with detecting a wide range of voice disorders. In this way, we design a system as a series of binary decision units as shown in Figure 5-4. The relations used for the classifier units are similar to above [34]:

$$\frac{P(\text{weight from adult human})}{P(\text{weight not from adult human})}=\frac{P(\lambda_C \mid X)}{P(\lambda_{\bar{C}} \mid X)}=\frac{p(X \mid \lambda_C)P(\lambda_C)/P(X)}{p(X \mid \lambda_{\bar{C}})P(\lambda_{\bar{C}})/P(X)}$$

with the log-likelihood ratio calculated as

$$\Lambda(X)=\log[p(X \mid \lambda_C)]-\log[p(X \mid \lambda_{\bar{C}})]$$

and the detection decision made as

$$\begin{array}{ll}\Lambda(X)\geq c, & \text{accept as voice disorder n}\\ \Lambda(X)<c, & \text{reject}\end{array}$$

The large vector $X=\{x_0,x_1,\ldots,x_{M-1}\}$ is the stream of multidimensional features generated by the test utterance. It is important to realize that there are several other approaches to constructing a recognition system than the one presented such as those in [40]. Our choice of binary decisions is due to factors discussed in chapter 2 including the small number of available utterances in the Kay database and the overlapping nature of dysphonia classes for one file.

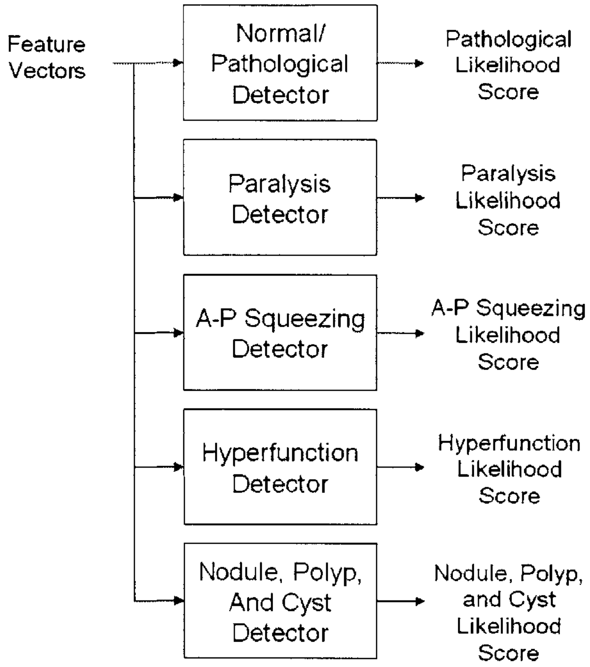
Feature Vectors
Normal/ Pathological Detector
Pathological Likelihood Score
Paralysis Detector
Paralysis Likelihood Score
A-P Squeezing Detector
A-P Squeezing Likelihood Score
Hyperfunction Detector
Hyperfunction Likelihood Score
Nodule, Polyp, And Cyst Detector
Nodule, Polyp, and Cyst Likelihood Score

Figure 5-4. System of binary voice disorder detection units used to produce the likelihood of a certain diagnosis. Details on the selection of each data set are provided in section 5.2.3.

5.2.3 Training and Testing Procedure

In order to build target and background models for the Kay database [1], list files grouping the speakers into categories for training and testing were generated using a custom Matlab program. The program first builds a system of records from the spreadsheet diagnosis file provided with the dysphonic speech files. This database includes a separate record containing all files for a specific patient number. In many cases, no patient number is provided for a file, and so these files were each put into their own record. Each file includes the patient sex, age, visit date, clinician diagnoses, smoking history, native language, and origin. For some files, little or none of this information is provided. Also noted is whether the utterance contained in the file is of normal or pathological speech, and whether it is a sustained vowel or rainbow passage utterance. Recall the more complete description of the database given in chapter 2.

The software developed creates collections of files based on search criteria, and these collections could be combined using AND, OR, XOR, and NOT set operators in order to arrive at collections of files with very particular properties. The ability to split a collection of files into a user-specified number of groups was also included.

Five categories of dysphonia were chosen for recognition experiments. The first two, normal/pathological and anterior-posterior squeezing had been tested previously in the literature with a GMM system [10] and were used as baseline experiments. The other three categories, *hyperfunction*, *paralysis*, and *nodules, polyps, and cysts* were chosen as representative of major categories revealed within the voice disorder discussion of chapter 2. All audio was digitally converted from a proprietary PCM format to 8 kHz PCM wav files. For the experiments described below, the "rainbow passage" sentence files are used, each less than 12 seconds in length. The process used to build each of the training and test sets was:

(1) Only the Rainbow passage sentences are used in our experiments, but the database was still restricted in this step to those files having *both* rainbow passage and sustained vowel recordings. The purpose of this step was to allow the possibility of future comparisons between recognition performance of sustained vowels and continuous speech.

(2) All the files containing the desired set of diagnoses, listed in Table 5.1, were extracted as the *target* files. Any pathological files also containing the note "normal" were discarded.

(3) All files in the database without a patient number and diagnosis information were removed. About a third of the database fit this category.

(4) The *background set* consisted of any files not found in step 2 or eliminated in step 3. This group included the 53 utterances recorded from normal patients.

(5) 80% of each of the target and background sets were used for training, 20% were used for testing. There was no overlap between patients or files in the training and testing sets or between the target and background classes.

Table 5.1. The disordered voice database used contains many diagnoses. The phrases in this table are synonyms for the categories used in our dysphonia recognition system. In order to build lists of files for each of the five classes listed on the left, we searched the database for the terms listed on the right. For example, in order to build the paralysis dataset, we used files given the clinical diagnoses "paralysis" or "paresis."

Class	**Diagnoses Corresponding to this Class**		
Pathological	Pathological		
A-P Squeezing	"A-P compression"	"AP compression"	"A-P squeezing"
Hyperfunction	"hyperfunction		
Paralysis	"paralysis"	"paresis"	
Nodules, Polyps	"vocal nodules"	"vocal fold polyp"	"cyst"

5.2 Fusion Techniques

A combination of Perl and Matlab tools were built to accomplish an optimum linear fusion of scores resulting from different auditory models. As depicted in Figure 5-5, such a technique amounts to arriving at the flat surface (for example, a line or a plane) in the N-dimensional space of scores which best discriminates scores based on a lowest EER criteria.

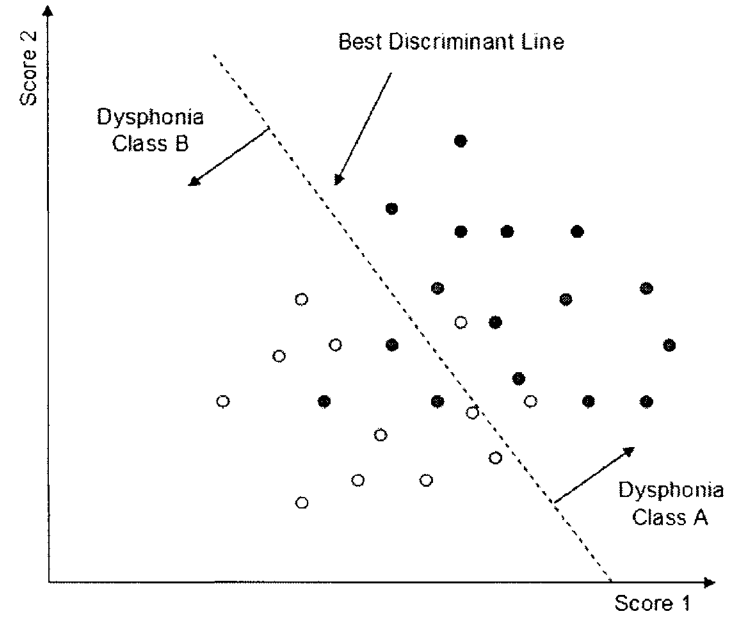

Figure 5-5. Schematic depicting the task of finding optimal score weights as the problem of finding the angle of the line which best discriminates two classes of data. We considered the best surface to be the one yielding the lowest EER.

An exhaustive search in a multi-dimensional plane is used to arrive at the lowest possible EER defined for this section as the average of percent miss and percent false alarm when the percent miss and percent false alarm are closest. It is realized that there are possibly existing linear regression techniques to make this process more efficient, but these were not used in this thesis. The algorithm involves calculating the linear combination of the score on each dimension of the line (more generally, a multi-dimensional plane) for which the lowest average EER is obtained. This set of weights is incremented by steps of constant degrees (in polar coordinates) and the minimum value of the resulting average EER versus degrees function is found. To compute the best EER at each angle, the threshold of the linear sum is calculated using a function-minimizing technique built into the Matlab software. As we are looking for an optimal combination of weights, there are no separate training and testing phases to this process. This optimal combination shows the limit at which such a system can perform.

The steps of this procedure for the two-class problem are as follows:

(1) Iterate slope of discriminant line using 4000 uniform steps in angle between 0 and 2π. The angle was chosen to search the space of weights in a way that best characterized the rotation of the decision surface.

(2) Calculate EER by sweeping the threshold. The threshold value for which the % error for the two classes was closest was found. The percent-error rates for each of the classes at this point were then averaged to yield a value, denoted as the *average EER.*

The relation of the two weights to the iterated angle, θ, is illustrated in Figure 5-6 with $w_1 = \cos(\theta)$ and $w_2 = \sin(\theta)$. This diagram also shows the method to construct the descriminant line $w_1x + w_2y = c$, where c is the value of the decision criteria, x is the value of the first score and y is the value of the second. In our dysphonia recognition problem, if $w_1x + w_2y \geq c$, we classify the test point as the target, and if $w_1x + w_2y < c$, we classify the test point as a background case. In other words, an (x, y) score point is classified as a target only if it is in the region

$$y > -\frac{w_1}{w_2}x + \frac{c}{w_2}$$

This principle can be generalized to higher dimensions. For example, three scores can be optimally combined by finding the decision surface $w_1x + w_2y + w_3y = c$.

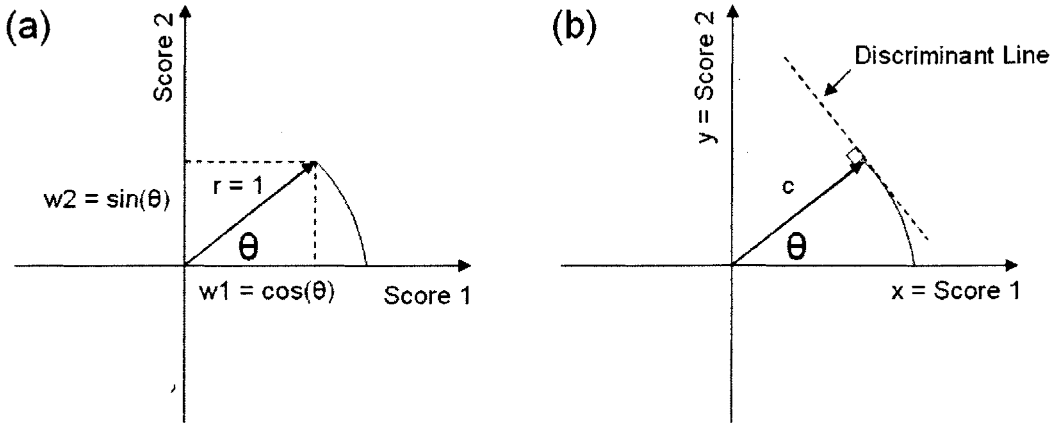

Figure 5-6. Geometry used to determine the best discriminant line for the two-class recognition problem. (a) Diagram of how $w1$ and $w2$, the linear weights for score 1 and score 2 respectively, were obtained as θ was swept through 4000 increments from 0 to 2π. (b) Every possible discriminant line can be parameterized by the angle θ and c, the tangential distance from the origin. The equal error rate for each θ tested was calculated by finding the c value at which the difference was smallest between the percent of Class A labeled as Class B and the percent of Class B labeled as Class A.

As an example, Figure 5-7 shows a polar plot of average EER versus angle for a fusion experiment involving mel-cepstrum and fast ICC centroid features (given later in Figure 5-). The slope of the discriminant surface, denoted by the dotted line, is actually *perpendicular* to the angle (of the linear regression) plotted. As can be seen in the figure, a slightly lower EER results for a linear regression between the extremes of only mel-cepstral (0 degrees in the figure) and only fast ICC centroid (90 degrees). The angle of the regression line for the best average EER is close to 48 degrees, corresponding to weights of 0.4685 and 0.5315.

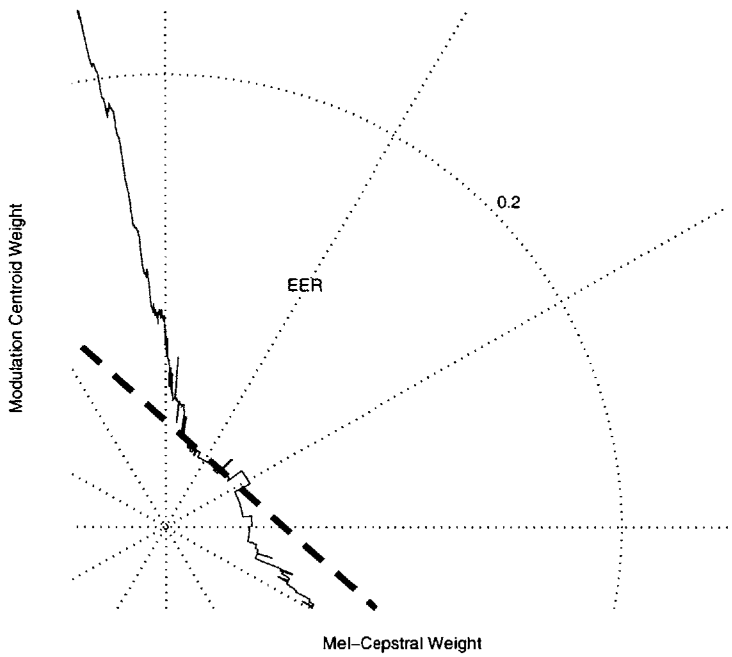

Figure 5-7. Effects of varying feature weight on average EER for a simple linear fusion. Quadrant 1 of a polar plot is shown. Each point on the plot can be described using (r, θ), where r is the EER at the point and θ is the angle perpendicular to the discriminant line producing that EER. Combinations of weights resulting in points closer to the origin of the plot are said to give higher performance. The approximate angle of the optimum decision surface is shown with the dotted line.

5.4 Dysphonia Recognition Experiment Results

We performed recognition experiments using a 64-mixture GMM[1] system for each of five dysphonia categories: nomal/pathological, A-P squeezing, hyperfunction, paralysis, and vocal-fold lesions. With the testing and training sets described in section 5.3, we ran five different tests, each using a different set of files for training and testing. This technique, called a *jackknife* [32], allows us to use every voice as both a test and training case.

The results of the dysphonia recognition experiments are plotted in the set of DET curves on the following six pages. The top panel on each pages shows performance without score fusion and the bottom panel shows results using score fusion with the mel-cepstrum. Four feature sets—mel-cepstrum, DPK nonlinearity, fast ICC centroid, and fast ICC summation—are tested for each of the five systems. Additionally, accurate and narrow-spacing ICC summation features are presented for the normal-pathological case[2].

Figures 5-8 through 5-11 present results for the normal/pathological test. Without fusion, the features do not appear to have a performance advantage over the mel-cepstrum. The fast ICC centroid, accurate ICC summation, and narrow-spacing ICC summation approach do, however, improve performance when fused with mel-cepstrum. This is evidence that the fast ICC centroid,

[1] A preliminary experiment (not shown) was run to choose the best-performing number of mixtures.
[2] Due to time restrictions, these features have not been run on the other dysphonias.

the accurate ICC summation, and the narrow-spacing ICC summation provide complementary information to the mel-cepstrum alone.

Figures 5-12 and 5-13 show the A-P squeezing test. We see much reduced performance for all four features. In Figure 5-13, the mel-cepstrum fused with the fast ICC centroid model seems to add complementary information. The EER of this fusion is 38.72 percent compared with 42.68 percent for mel-cepstrum. Hyperfunction recognition results are shown in Figures 5-14 and 5-15. The systems perform on the order of mel-cepstrum: the best EER for this recognition is 38.40 percent.

Figures 5-14 and 5-15 show the paralysis recognition experiments. Near the EER, the fast ICC centroid shows increased performance relative to mel-cepstrum before fusion. With fusion, the curves suggest that the fast ICC centroid and summation features add complementary informaion to mel-ceptrum. The fast ICC centroid curve is consistently to the left of the mel-cepstral curve. Vocal fold lesion detection shown in Figures 5-16 and 5-17 is another case where the fast ICC summation curve is to the left of the mel-cepstral curve. In this case, however, the fast ICC summation curve is also the furthest to the right in the individual experiments. The EER score for the fast ICC is 36.86 percent compared with 41.86 percent for the mel-cepstrum alone.

The evidence presented needs to be verified on a larger data set as the 95-percent-confidence error bars for the above measurements are large: between about 6 and 12 percentage points wide for the normal/pathological case. This issue arises because the number of target and background cases is limited. The normal/pathological class has 397 target and 53 nontarget files, A-P squeezing has 164 target and 266 nontarget files, hyperfunction has 263 target and 161 nontarget files, paralysis has 81 target and 365 nontarget files, and vocal fold lesions has 43 target and 407 nontarget files.

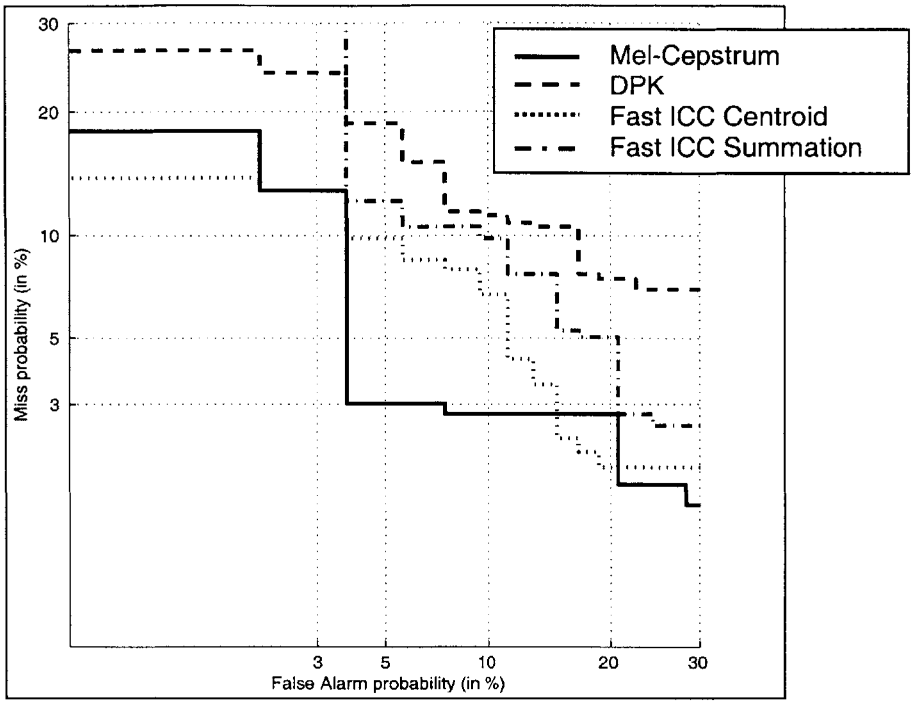

Figure 5-8. DET curves using mel-cepstrum, DPK, and fast ICC features. The mel-cepstral features perform with an EER of 3.77% while DPK, fast ICC centroid, and fast ICC summation yield EERs of 9.43%, 8.06%, and 9.82% respectively. The task is normal/pathological classification.

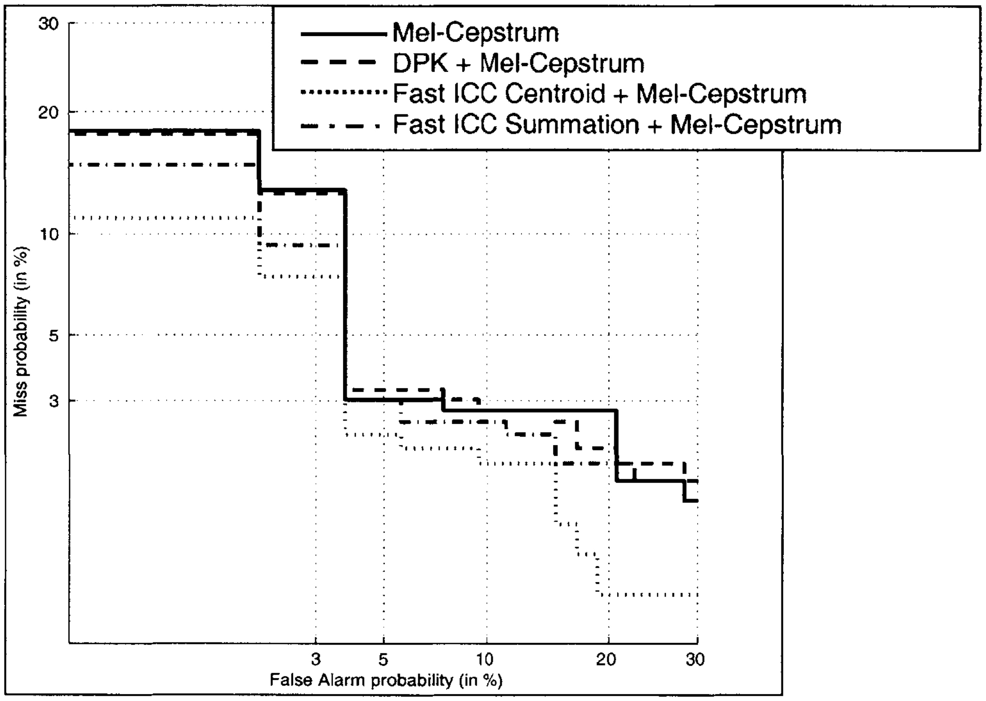

Figure 5-9. DET curves using DPK, fast ICC centroid, and fast ICC summation features fused with mel-cepstrum. The mel-cepstral model has a 3.77% EER. The fusions between mel-cepstrum and DPK, fast ICC centroid, and fast ICC summation have EERs of 3.77%, 3.77%, and 3.77% respectively. The task is normal/pathological classification.

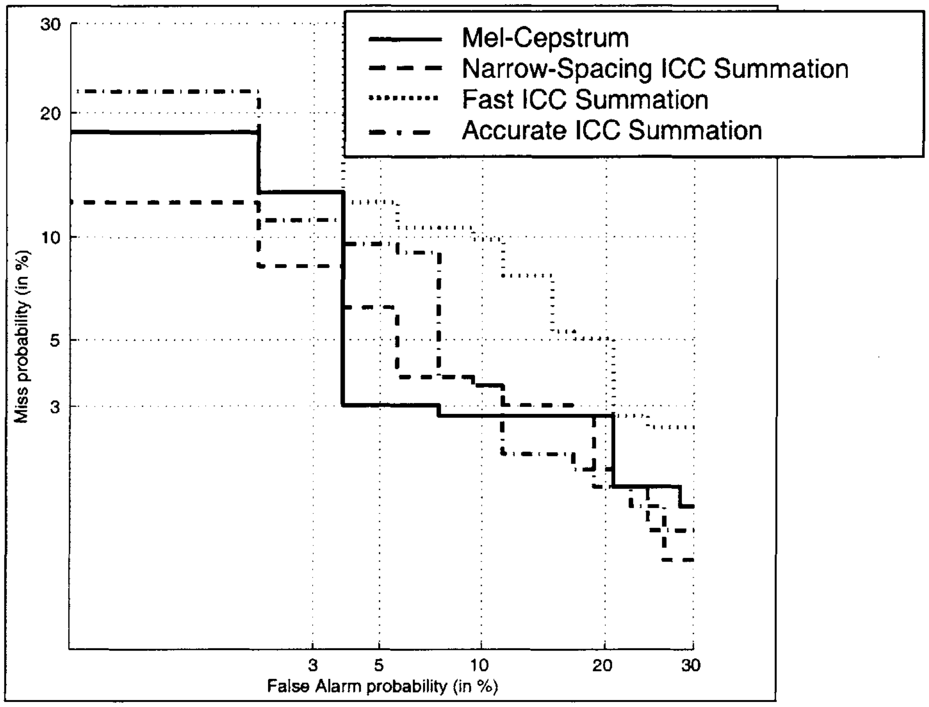

Figure 5-10. DET curves using the fast, accurate, and narrow-spacing models of the ICC. The mel-cepstral model has a 3.77% EER and the narrow-spacing ICC, fast ICC, and accurate ICC summations result in EERs of 5.66%, 8.06%, and 7.55% respectively. The task is normal/pathological classification.

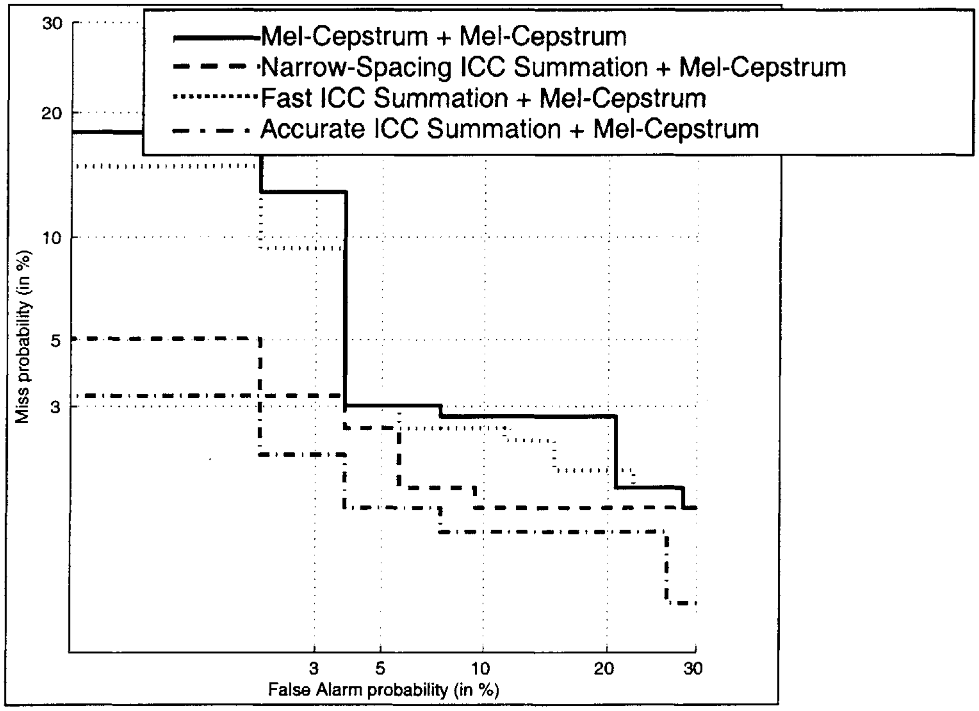

Figure 5-11. DET curves using narrow-spacing ICC, fast ICC, and accurate ICC summation features fused with mel-cepstrum. The mel-cepstral model has a 3.77% EER. The fusions between mel-cepstrum and narrow-spacing ICC, fast ICC, and accurate ICC summation have EERs of 2.02%, 3.77%, and 3.27% respectively. The task is normal/pathological classification.

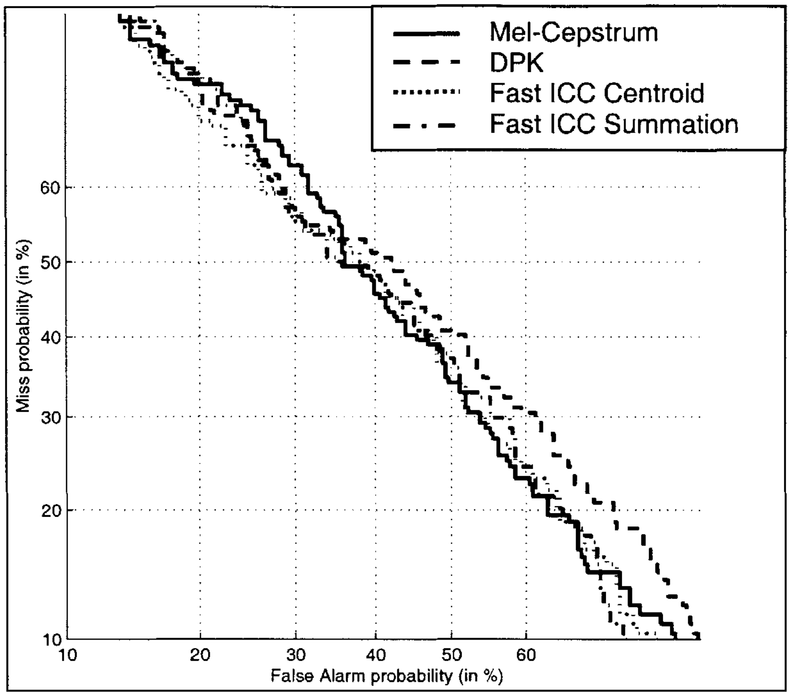

Figure 5-12. DET curves comparing DPK, and fast ICC features. The mel-cepstral features perform with an EER of 42.68% while DPK, fast ICC centroid, and fast ICC summation yield EERs of 45.73%, 43.61%, and 44.51% respectively. The task is A-P squeezing classification.

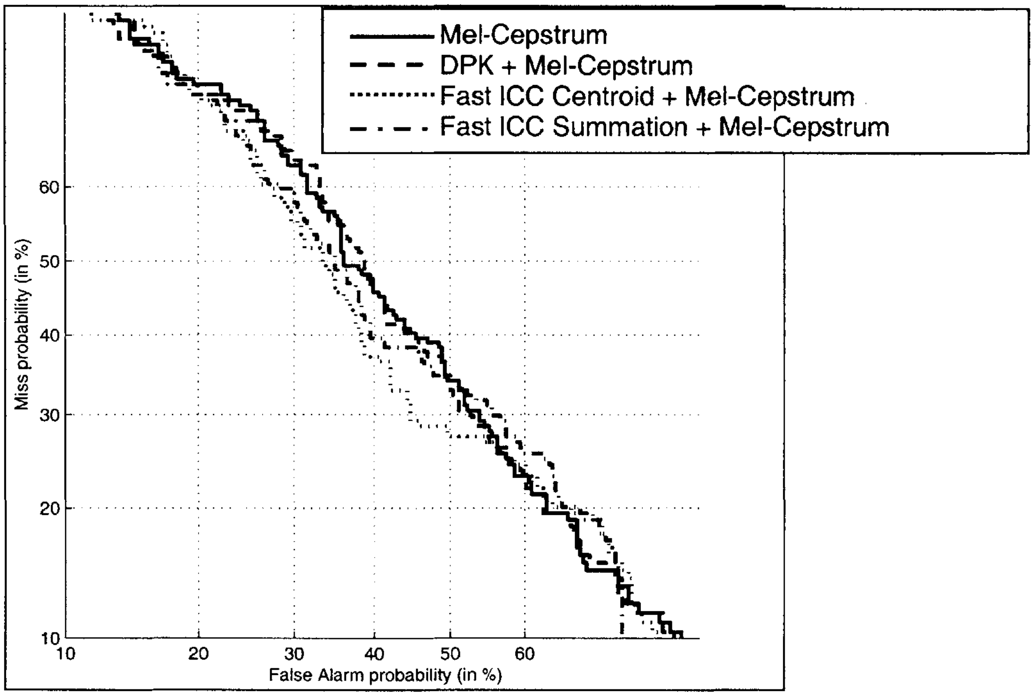

Figure 5-13. DET curves using DPK, fast ICC centroid, and fast ICC summation features fused with mel-cepstrum. The mel-cepstral model has a 42.68% EER. The fusions between mel-cepstrum and DPK, fast ICC centroid, and fast ICC summation have EERs of 41.46%, 38.72%, and 39.63% respectively. The task is A-P squeezing classification.

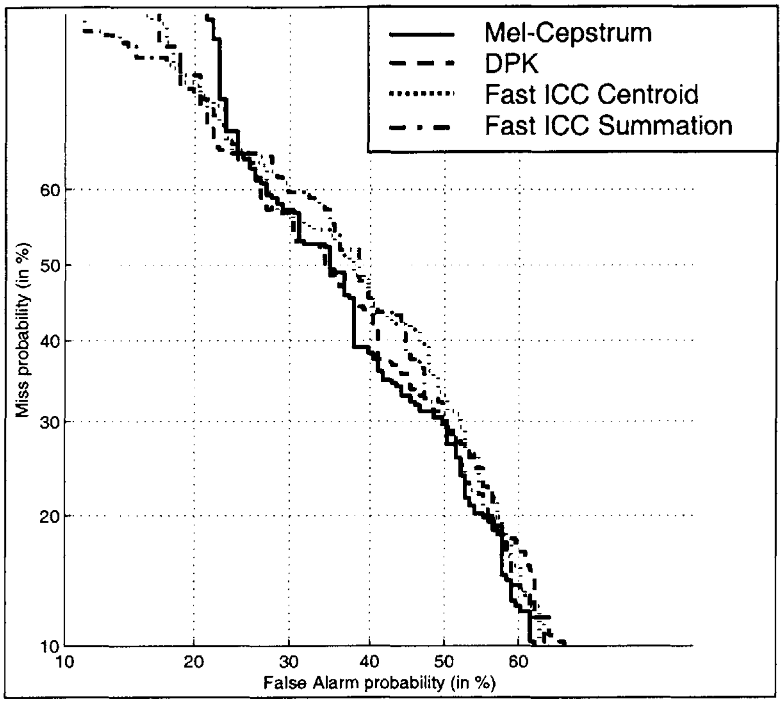

Figure 5-14. Comparison of mel-cepstrum, DPK, and fast ICC features. The mel-cepstral features perform with an EER of 39.16% while DPK, fast ICC centroid, and fast ICC summation yield EERs of 40.99%, 42.86%, and 43.35% respectively. The task is hyperfunction classification.

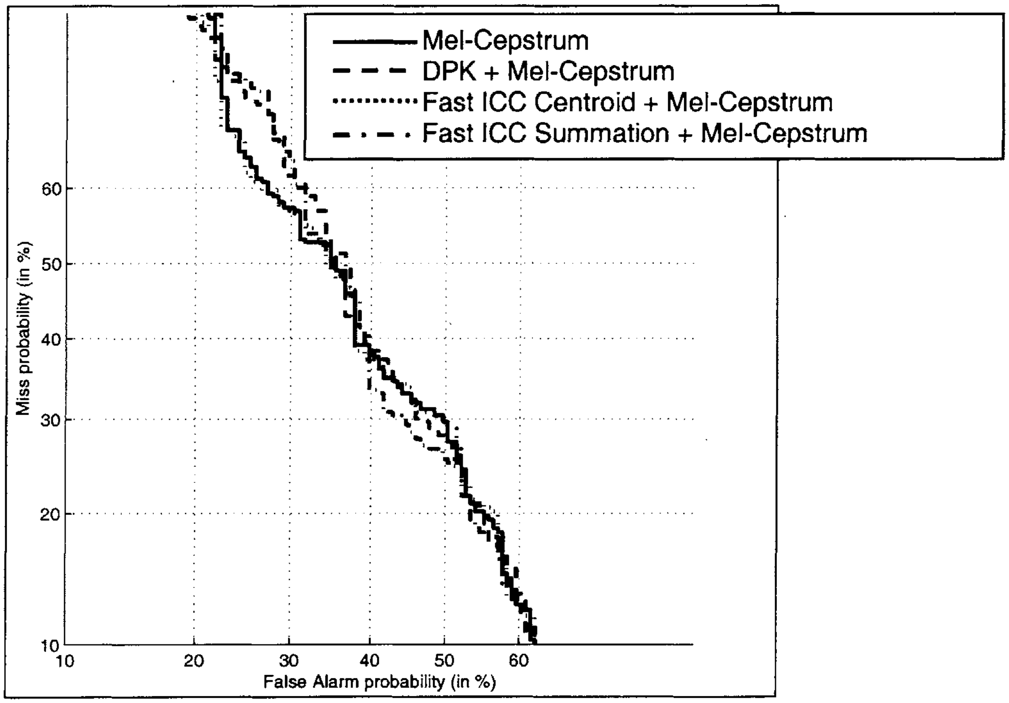

Figure 5-15. DET curves using DPK, fast ICC centroid, and fast ICC summation features fused with mel-cepstrum. The mel-cepstral model has a 39.16% EER. The fusions between mel-cepstrum and DPK, fast ICC centroid, and fast ICC summation have EERs of 39.54%, 38.40%, and 39.75% respectively. The task is hyperfunction classification.

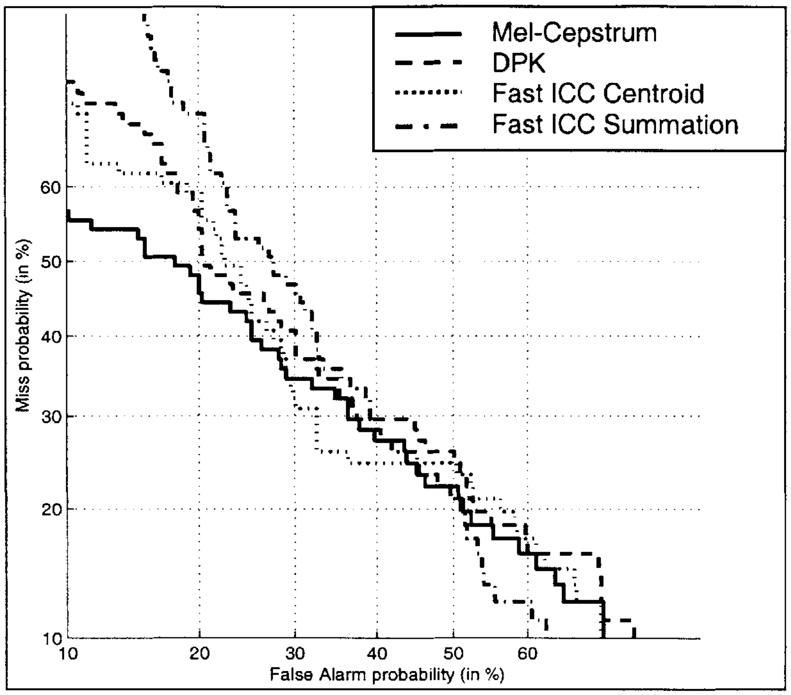

Figure 5-16. Comparison of mel-cepstrum, DPK, and fast ICC features. The mel-cepstral features perform with an EER of 33.33% while DPK, fast ICC centroid, and fast ICC summation yield EERs of 34.57%, 42.86%, and 43.35% respectively. The task is paralysis classification.

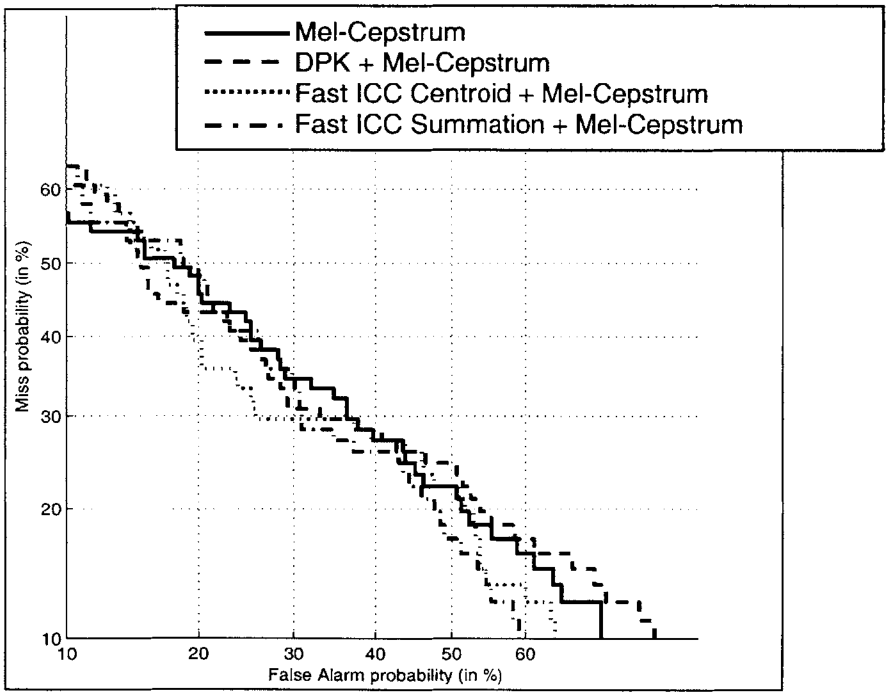

Figure 5-17. DET curves using DPK, fast ICC centroid, and fast ICC summation features fused with mel-cepstrum. The mel-cepstral model has a 33.33% EER. The fusions between mel-cepstrum and DPK, fast ICC centroid, and fast ICC summation have EERs of 30.86%, 29.63%, and 30.81% respectively. The task is paralysis classification.

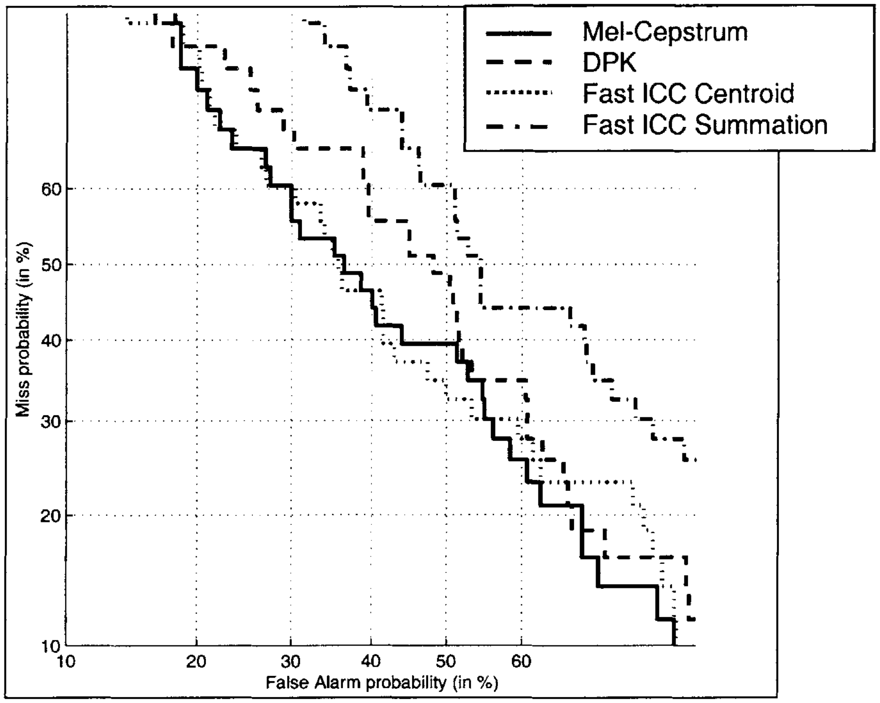

Figure 5-18. Comparison of mel-cepstrum, DPK, and fast ICC features. The mel-cepstral features perform with an EER of 41.86% while DPK, fast ICC centroid, and fast ICC summation yield EERs of 48.84%, 41.52%, and 52.83% respectively. The task is polyp, nodule, and cyst classification.

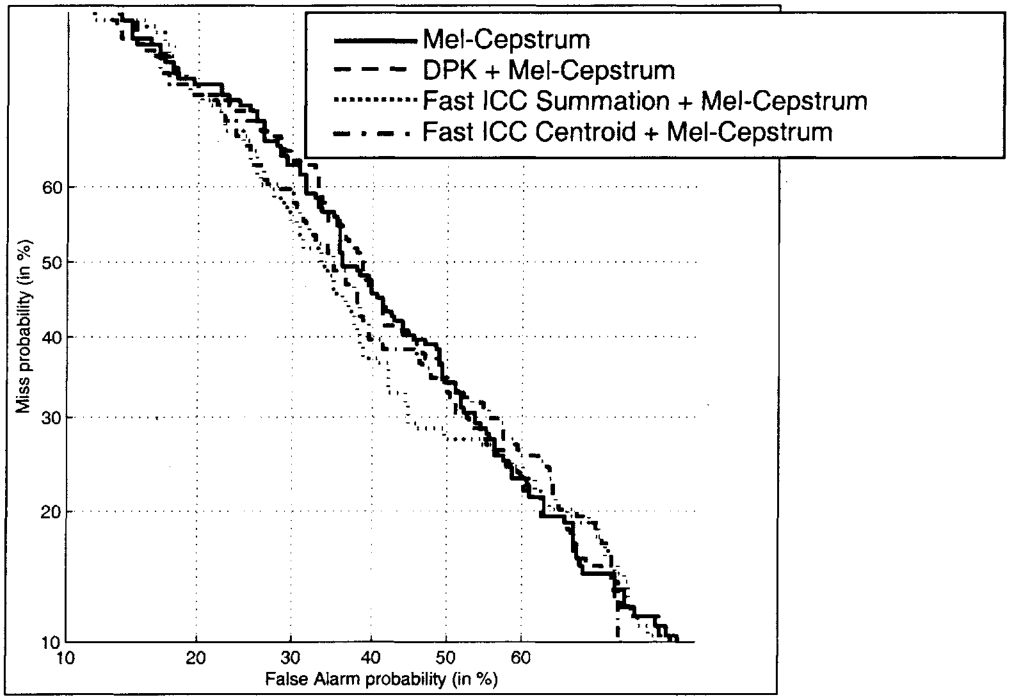

Figure 5-19. DET curves using DPK, fast ICC centroid, and fast ICC summation features fused with mel-cepstrum. The mel-cepstral model has a 41.86% EER. The fusions between mel-cepstrum and DPK, fast ICC centroid, and fast ICC summation have EERs of 39.56%, 39.53%, and 36.86% respectively. The task is polyp, nodule, and cyst classification.

5.5 Conclusions

In this chapter, we integrated the biologically-inspired features of chapter 4 into an automatic dysphonia recognition system. Our system recognizes five disorders—general pathology. hyperfunction, A-P squeezing, paralysis, and vocal fold lesions. The experimental results suggest that biologically-inspired methods provide complementary information to one another as proposed in chapter 4.

Dysphonia recognition using the Kay database is reported in the literature by several groups. Dibazar and Narayanan [10] describe a normal/pathological classification system consisting of a GMM classifier and using mel-cepstral features on both the rainbow passage and sustained vowels. They report a 2.54 percent EER for this experiment, which is close to our best fused score of 2.02 percent with the narrow-spacing ICC summation features. With sustained vowels, Dibazar and Narayanan's best EER is 1.70 percent. These authors also perform A-P squeezing recognition on 163 A-P squeezing subjects, dividing this group into four different severity levels—52 minor, 57 mild, 39 moderate, and 19 severe. In this experiment, they test sustained vowel utterances only, achieving 73 percent classification [10].

Objective clinical speech methods, such as shimmer and jitter discussed in chapter 2, can be used as a model for feature extraction. Godino-Llorente *et al.* use the Kay Multi-Dimensional Voice Program to extract perturbation features including jitter, shimmer, and harmonic-to-noise ratio. Using these features in a sustained-vowel dysphonia normal/pathological recognition experiment, they obtain 5.13 percent EER in their highest-performing system. Parsa and Jamieson devise feature extraction methods capable of extracting jitter, shimmer, and harmonic-to-noise ratio in both sustained vowels and the rainbow passage. Their highest-performing features, based on the harmonic-to-noise ratio, yield classification rates of 96.5 percent for sustained vowels and 95.6 percent for the Rainbow passage [32]. Our results suggest improved performance with relation to these studies. Our best fused EER for the normal-pathological case was 2.02 percent using the narrow-spacing ICC summation model.

Chapter 6

Conclusions and Future Work

6.1 Summary

This thesis has motivated and demonstrated the use of biologically-inspired models to capture amplitude modulations in dysphonic speech. In chapter 2, we began by discussing dysphonia, first presenting a taxonomy for voice quality derived from the literature in terms of acoustic and perceptual traits and later providing a connection between these characteristics and physiological disorders. From this work, we found reports in the literature of time-varying acoustic amplitude patterns in voice. We also highlighted complications of the Kay Disordered Voice Database [1] used in this thesis as the dysphonia recognition corpus.

In chapter 3, we derived an interpretation of time-varying acoustic patterns in voice using AM theory from communications engineering. The model we developed views bandpass analysis, such as used by the short-time Fourier transform, as a form of signal demodulation. Through a case study of seven utterances taken from the Kay database, we found that the interaction of spectral components can create different patterns in different frequency bands. We also observed cases where both the spacing between components and their relative frequencies influence fluctuations seen in the time domain.

Chapter 4 described the motivation, design, and implementation of biologically-inspired models for the demodulation and enhancement of AM in speech. We presented three models—mel-filtering, Dau, Puschel, and Kohlrausch (DPK) nonlinear adaptation, and inferior colliculus (ICC) modulation filtering—previously used in the fields of speech, language, and speaker recognition. We supported the assertion that these models extract complementary voice properties by studying their responses to synthesized AM sinusoids and noise as well to dysphonic speech. The mel-filtering model was shown to represent spectral formant information but blurred transient activity. In contrast, the DPK model traded formant information for the property of enhancing quickly-varying components of the voice. The ICC model explicitly represented amplitude modulations in the signals through a second bandpass analysis stage.

Chapter 5 linked the auditory models in chapter 4 with feature extraction and pattern classification systems to build an automatic dysphonia recognition system. Using the dysphonia database discussed introduced in chapter 2, we created a system to recognize five different speech diagnoses—general pathology, hyperfunction, anterior-posterior (A-P) squeezing, paralysis, and a class of vocal fold lesions. Through the results of our experiments, we saw evidence for improved performance over mel-cepstra alone when fusing systems based upon the three biologically-inspired models. The results support our hypothesis that the auditory-system models extract complementary acoustic information about voice quality.

6.2 Contributions of this Thesis

This thesis advances the connection of amplitude-modulation theory to the analysis of voice quality. Specifically, we describe bandpass analysis systems from the perspective of amplitude demodulation and relate this to speech. Speech has differences from AM signals in traditional communications engineering including (1) potentially overlapping source bandwidths, (2) non-sinusoidal carrier sources and (3) an unknown numbers or source bandwidths. In our work we find that different AM patterns occur in different frequency bands. This perspective challenges current dysphonia analysis methods that analyze AM in the time-domain signal. A frequency-dependent analysis differs, for example, from the recent classification schemes based on fluctuations in time waveforms as performed by Titze [47] and by Garratt and Krieman [14]. Although previous authors have addressed the existence of amplitude modulation patterns in speech and mentioned spectral correlates, there has been found no previous attempt to classify speech based on different time-varying envelopes in different frequency bands.

This thesis has contributed to methods used for automatic dysphonia recognition by providing new AM-sensitive features. It also discusses the methods used to perform recognition experiments on multiple dysphonia types, and implements the automatic recognition of five different classes of dysphonia. This has not been reported previously in the speech literature. Additionally, we show that features created using biologically-inspired AM-sensitive models may provide complementary information to state-of-the-art mel-cepstral features. This strengthens the case for their use, initially presented for speaker recognition in [36].

6.3 Future Work

Chapter 2 presented an analysis model for AM in speech, but a corresponding AM synthesis model for speech has yet to be developed. In future research, we plan to extend current models of the dysphonic voice [3, 16] to understand the modulations they produce. By developing the capabilities to analyze and synthesize speech *concurrently*, we propose to iteratively improve our methods. That is, both the analysis and synthesis models will begin with simple cases; as synthesis abilities improve, we propose to develop more complicated analysis methods to understand them. Appendix A contains a first step toward this goal, adapting the well-understood Klatt synthesizer model to three case-studies of dysphonic voice.

The recognition experiments in chapter 5 remain inconclusive until we can obtain a database with a larger number of speakers. The Kay database has a superset called VoiceBase that contains near 10,000 patients that we are attempting to obtain and use. Additionally, a study of dysphonia recognition performance by clinicians would be useful to better understand limits of machine performance. A similar study, performed for speaker recognition could guide this research[41].

Appendix A

Synthesis of Dysphonic Voices Using the Klatt Synthesizer

A.1 Introduction

In the clinical evaluation of the voice, the condition of the larynx is commonly studied using visual examination with optical tools [49], perceptual rating scales [9], and objective acoustic measures [29, 33, 50]. Partially due to the invasiveness of the first and problems with the reliability and utility of the last two of these methods [37], there continues to be a search for new techniques by which to determine the state of the voice production mechanisms. Recent efforts have pointed to analysis-by-synthesis of dysphonia, the acoustical manifestation of an underlying voice disorder, as being a possible improvement upon previous acoustical evaluation mechanisms. In this scenario, a clinician attempts to match the quality of a synthesized utterance to a patient's voice using a limited set of parameters [2, 13]. The hope is that the values of these parameters will provide a more robust representation of the underlying quality of the voice than existing objective and subjective measures.

Aside from the applications in the clinical domain, a better understanding of how to synthesize voice disorders effectively may contribute to knowledge about how current speech synthesis models might be made more realistic. It becomes apparent from the literature that the line between normal and abnormal voices is not a strong one [25]. It can be hypothesized that the characteristic roughness or breathiness affecting a normal speaker's voice is really very similar, perhaps even in its mechanism of production, to that of a dysphonia. Thus, one of the goals of a better understanding of the synthesis of dysphonic voices is to motivate new tools for the analysis of both normal and pathological voices.

The following questions are addressed by the current study:

(1) Can the procedure reported in [2] to synthesize dysphonic tokens of the vowel /a/ be replicated and applied to the new data effectively?
(2) Can analysis of the signal through synthesis help indicate what makes the voices sound abnormal?

A.2 General Methodology

In the present study, three different voices denoted JAB08, KAH02, and DAS10 were investigated. The tokens were all of the sustained vowel /a/ and were extracted from the KAY Disordered Voice Database recorded in a clinical environment at the Massachusetts Eye and Ear Infirmary [1]. The set of voices used were chosen after surveying more than 300 voice samples in the database. JAB08 was selected as a strong example of tremor, a muscle control disorder; KAH02 was picked as a distinctly hoarse voice with many aperiodic components; DAS10 was chosen because it sounded extremely creaky and demonstrated repeating patterns of both large and small glottal pulse. This small sample of the data was not intended to capture the full range of possible dysphonic characteristics, but was hoped to provide some insight into the challenges of synthesizing disordered speech.

Synthesis was performed using the KLSyn formant synthesizer [23] with the KLGlott88 voicing source [22]. All voice samples were lowpass filtered and downsampled to a 10 kHz sampling rate using the software program Wavesurfer and were brought into SpeechStation 2 for analysis. All spectral measurements were performed using the SpeechStation spectrum tool, with a 51.2 ms Hamming window. In order to aid in the measurement of pitch and formant frequencies, SpeechStation's formant and pitch trackers were occasionally activated. Formants were measured by cursor directly off of the spectra, using a 14^{th} order LPC envelope as an aid.

The procedure used as a first pass to synthesize the voices created and analyzed in this study was modeled after [2], in which a series of 24 dysphonic utterances were synthesized and compared against the originals in a perceptual study. In this paper, seven steps were outlined for the synthesis of a disordered voice. Below are these steps, slightly modified:

(1) *Match F0, formant frequencies and bandwidths, and spectral tilt.* The goal of this step was to match the gross spectral characteristics of a speech utterance at a single point in the vowel, neglecting details of the individual amplitudes of the harmonics. Additionally, the spectral tilt, TL, of the utterance was modified here in order to match the overall slope of the spectrum, with larger TL values leading to less high frequency energy.

(2) *Adjust source amplitudes AV, AH, and AF.* The amplitudes of each of the source mechanisms were adjusted to constant values in this step. AV, AH, and AF control the amplitude of voicing, aspiration, and frication sources, respectively.

(3) *Adjust open quotient to match H1 amplitude.* The amplitude of the first harmonic, which tends to reflect the breathiness of a speech signal [24], was adjusted using the parameter OQ. OQ defines the open-quotient of the glottal pulse. This step was a first pass by which to mimic the intensity of H1 in the spectrum of the speech signal and often did not make H1 large enough.

(4) *Adjust harmonics below F1 using nasal and tracheal poles and zeros.* As will be shown, many disphonic utterances exhibit complicated structure in the harmonics below F1, which can be important to how they are perceived. One method by which we achieved this was to insert additional poles and zeros into the system to boost or reduce the intensities of specific components. The basic procedure was to place a pole and zero at the same frequency, making the bandwidth of the pole smaller than that of the pole if a spectral prominence was desired or making the bandwidth of the zero smaller if a spectral trough was sought.

(5) *Mimic F0 time variation using F0, FL, and DI.* Time variations of frequency parameters across the utterance were handled in this step to mimic characteristics such as deviations of F0 from the nominal value and period-doubling. F0 is the pitch of the synthesized voicing source, FL or *flutter* imposes periodic fluctuation of F0, and DI is diplophonia which causes every other pitch period to be larger.

(6) *Mimic AV time variation.* Some utterances included variations in the amplitude of the voicing source over time. Depending on the utterance, this step was often performed in conjunction with step 5 to yield acceptable amplitude and frequency characteristics.

(7) *Add additional pole-zero pairs if required.* If pole-zero pairs were still available after step 4, they could be used to mimic higher frequency resonances and zeros, such as those created by subglottal coupling and nasality.

In practice, the above steps tended to be performed as an iterative process, with the first pass being followed by a more detailed adjustment of each of the parameters. The first pass, for example, did not consider the variation of formant frequencies and bandwidths across an utterance with time. In later iterations, however, the utterance was often divided into distinct sections and the details of each division were individually modified in the synthesizer.

A.3 Synthesis of Tremor

The patient JAB08, discussed in this section, is a 69 year old male being examined post-biopsy, with scarring on the vocal folds. The records also indicate that the patient presented with a vocal tremor. Tremor is characterized as a 4-to-8 Hz variation in the amplitude envelope of a speech waveform [30]. As can be seen in the spectrogram and waveform of Figure 1, the utterance demonstrated some variation in amplitude at this frequency as well as some evidence of diplophonia, especially around 150 ms and 700 ms. Overall, the voice is perceptually somewhat breathy, but as can be seen in Figure 1 and Figure 3, there is little noise to be modeled with an aperiodic source.

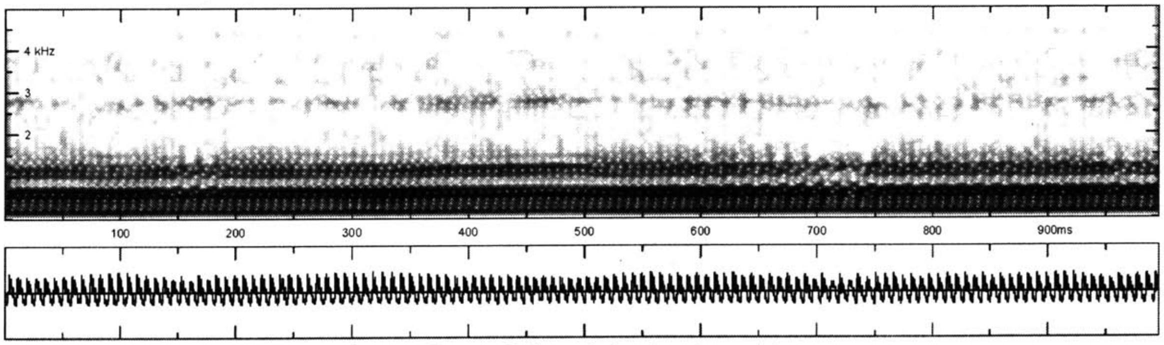

Figure A-1. Spectrogram and waveform of the original voice with tremor.

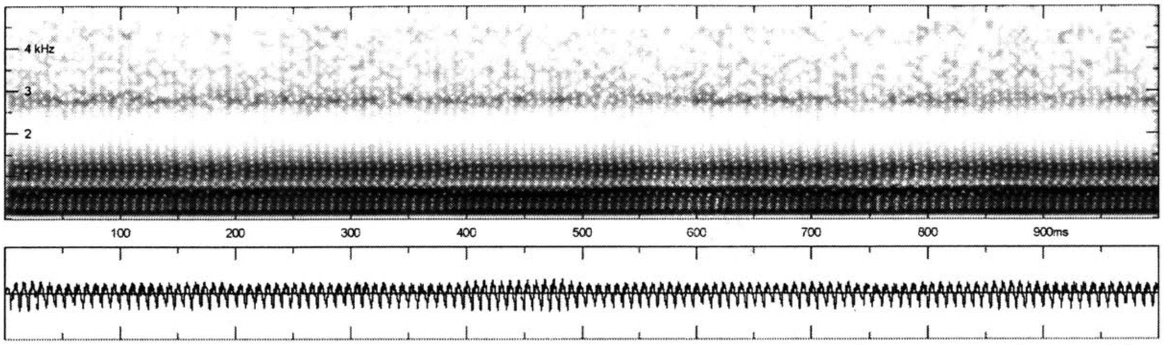

Figure A-2. Spectrogram and waveform of the synthesized voice with tremor. The resulting waveform shows modulation in the amplitude envelope even though the synthesis uses a constant AV value.

A.3.1 Spectral Characteristics

Overall, two issues led to a rather complicated synthesis of the spectrum. First, as can be seen in the left panel of Figure 3, the first harmonic of the speaker had an intensity which was too high to synthesize using only the OQ parameter. If OQ was made too large, it tended to make the upper frequencies of the spectrum unnaturally low. In order to fix this issue, a narrow-bandwidth pole-zero pair was placed at 150 Hz in order to boost the spectrum in the neighborhood of the first harmonic. As can be seen in the right panel of Figure 3, using such a narrow pole, we were able to match the amplitude of the first harmonic quite closely, without distorting the rest of the spectrum. A well-known physiological correlate to a resonance at this frequency is not known, and the difference more likely results from the behavior of the disordered glottis that is not well-modeled by the KLGlott88 source.

The second problem occurred when attempting to synthesize the trough present in the waveform around 800 Hz. Normally, one expects to see tracheal resonances at 550 Hz, 1300 Hz, and/or at 2100 Hz, with the zero slightly above or below these frequencies [22]. In the current example, a corresponding pole to the 800-Hz zero is not apparent, and we synthesized it as a broad pole at 900 Hz due to the hump observed near F2 around this frequency. In retrospect, however, it is more likely that the pole should have been positioned close to 550 Hz. Figure 3 shows a side-by-side comparison of the original spectrum with the final synthesized version.

A.3.2 Time-Varying Characteristics

In order to characterize the time-varying features of the vowel, we first split the utterance into four regions. Diplophonia, controlled by the parameter DI, was then applied to the entire second area, 130 to 205 ms, and as a time-varying contour over the fourth, from 570 to 1000 ms. The resulting effect can be observed in Figure 2, whereby the output in these regions exhibits the characteristic large pulse, small pulse, large pulse pattern of diplophonia

All sections of the utterance exhibited some fluctuations in the fundamental frequency. The amount of flutter, parameter FL, was chosen to capture the variation between 123 and 133 Hz over the entire utterance. Centered around the pitch of 128 Hz, this yielded a flutter value of 65 using the equation

$$f_0 = \left(\frac{3FL}{50}\right)\left(\frac{F0}{100}\right)$$

derived from the flutter equation given by [22]. Surprisingly, this frequency variation seemed to eliminate the need for a separate time-varying manipulation of AV. This example thus appears to be an interesting case of frequency modulation of a harmonic across a formant peak leading to amplitude modulation in the resulting waveform. Although it should be noted that Figure 2 contains the added effects of formant frequency modulation, discussed below, a version of the spectrogram with only flutter managed to cause the same behavior.

The final time-varying change that was made was to F1, which was observed in the original spectrogram to vary periodically between about 530 and 620 Hz. In order to synthesize this property, F1 was set to a value of 546 Hz at the measured F1 minima and 610 Hz at the maxima. As can be observed in Figure 2, the resulting F1 track has 5 minima including both endpoints; one clear example is at 480 ms. It is suspected that the condition of vocal tremor may cause slight opening and closing of the mouth resulting in some low-high variation, which would be expected to result in the observed F1 movement.

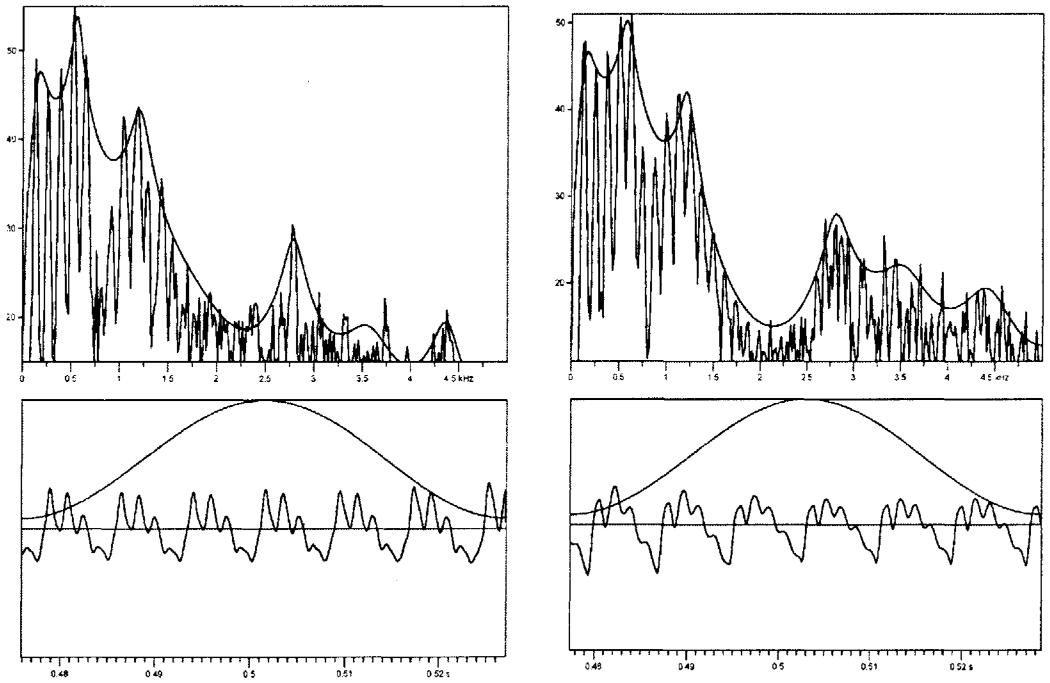

Figure A-3. Spectrum and waveform for original (left) and synthesized (right) tremor speech signals.

A.4 Synthesis of a Creaky Voice

The second voice that was synthesized, DAS10, is produced by a 40 year-old male patient reported to have granulation tissue as well as to exhibit abnormal amounts of tension in the muscles of the larynx. Perceptually the voice was very "creaky," defined as exhibiting an element of sharp, short glottal pulses in the voicing, creating the perception of low-frequency individual pulses [27], [5]. This percept agrees well with acoustic evidence found in the spectrogram shown in Figure 4, which indicates clearly defined glottal pulses. In addition, the utterance contained sections of diplophonic pulses as well as more complicated repeating patterns of activity. In the literature such behavior is called "superperiodic" voicing [14]. Such a voice has obvious similarities to diplophonic glottalization, which exhibits a pattern of one large pulse followed by one small one, seen in sections of normal speech [15, 38].

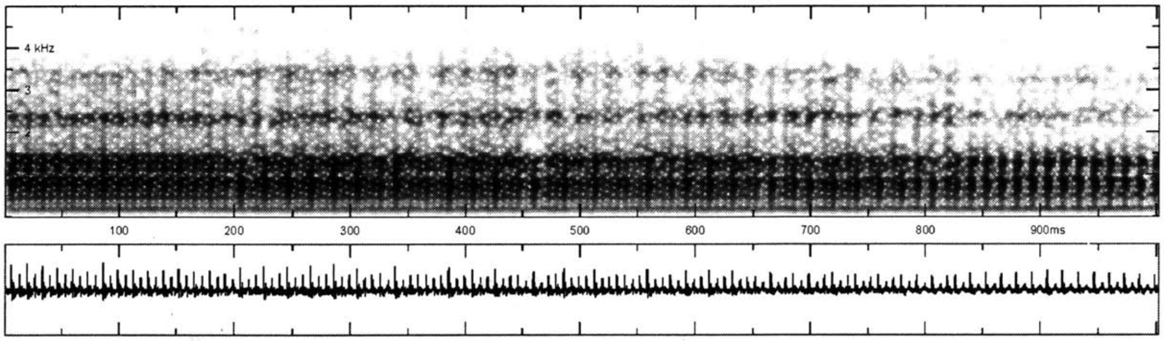

Figure A-4. Original creaky voice spectrogram and waveform.

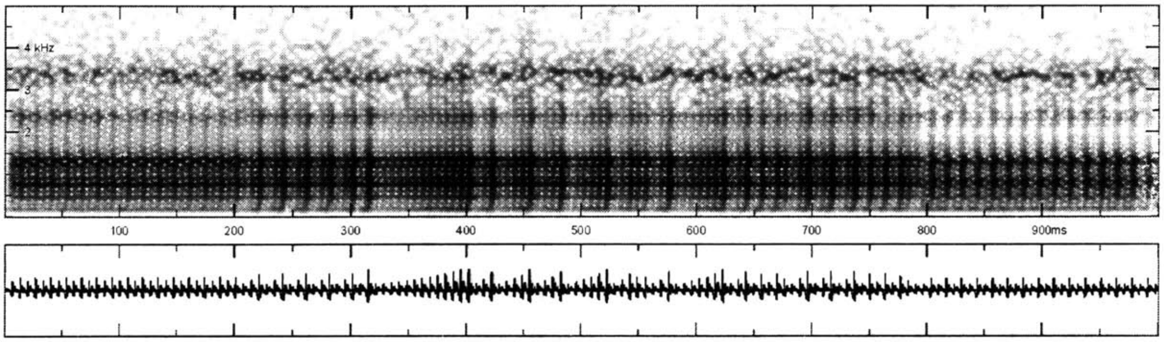

Figure A-5. Spectrogram and waveform of final synthesized creaky voice.

A.4.1 Spectral Characteristics

Achieving spectral characteristics for this utterance was relatively straight-forward. Two additional poles-zero pairs, one operating to increase the response at 450 Hz and the other acting to decrease the response at 2000 Hz were created in order to match the original spectrum. These frequencies are not far from 550 Hz and 2100 Hz, common locations of subglottal resonances [22]. Figure 6 shows the spectra taken early in the utterance, when the zero at 2000 Hz does not appear as active. In contrast, in Figure 7, the zero is seen in both the original and synthesized waveforms. Note that F3 bandwidth was also changed between these two locations.

One issue remains unresolved is making the glottal-source derivative waveform more "peaky" in the time domain. That is, if one compares the waveforms in the two panels of Figure 7, for example, the one on the left has sharper pulses that appear to be of larger amplitude relative to the decaying response. The perceptual influence of this feature is difficult to gauge, but such highly damped responses have been reported in the literature [15]. The exact nature of this damping phenomenon does not seem to be directly related to the steady-state formant bandwidths as none of them seem to be particularly wide.

A.4.2 Time-Varying Characteristics

As can be observed in Figure 4, the original waveform consisted of several distinct sections of time-varying amplitude modulations. Specifically, they were as follows:

(1) *1 to 200 ms*. Moderate diplophonia.

(2) *200 to 310 ms*. This section exhibited a superperiodic pattern of one large glottal pulse followed by three smaller ones each of slightly increased amplitude.

(3) *310 to 840 ms*. A complicated superperiodic amplitude pattern was observed here, consisting of both large and small pulses. One might call the amplitude behavior here irregularity, although the periods seemed to occur at regular intervals without much variation.

(4) *840 to 1000 ms*. Period-doubling, an extreme version of diplophonia was seen in this region, characterized by sharp pulses followed by a much smaller pulse one period later.

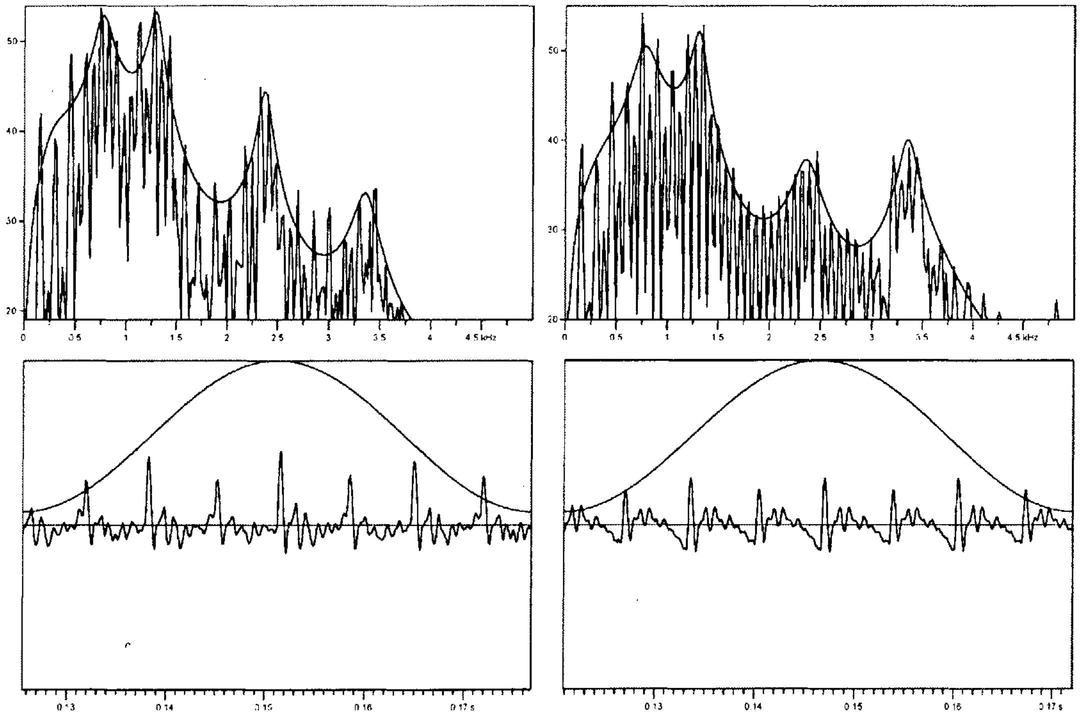

Figure A-6. Comparison of spectra taken around 150 ms for the original (left) and synthesized (right) voices.

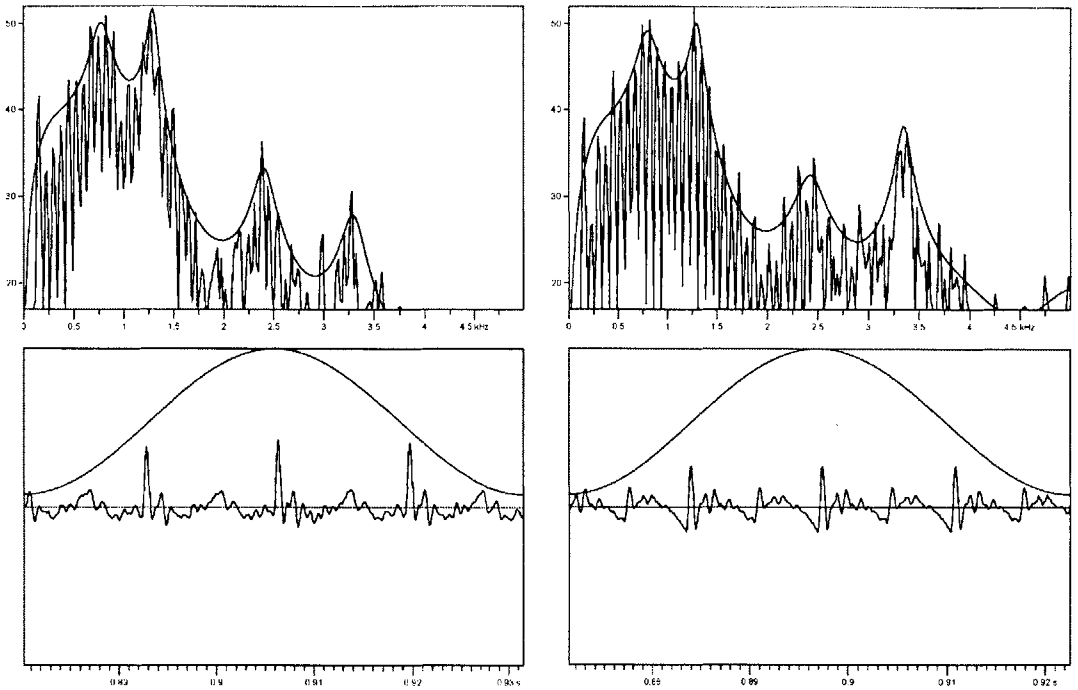

Figure A-7. Comparison of spectra taken around 900 ms for the original (left) and synthesized (right) voices.

Of these regions, the two exhibiting diplophonia were synthesized using the parameter DI as well as slight variations of the other spectral parameters, some of which are described in the spectral characteristics section above. The first step in synthesizing the more complicated regions was to note the times of each pulse in the synthesized waveform. Using the sequence of large and small pulses derived from the original waveform, a detailed path was then created for AV. The peak amplitude was set immediately before each large pulse and set lower immediately after. AV in between each of the large pulses was ramped upward in order to mimic a slight increasing trend in the small pulses between large pulses in the original waveform. As can be observed by comparing Figs. 4 and 5, we were able to mimic some of the characteristics of the original pulse pattern with synthesis.

Perceptually, the rate and spacing of the pulses sounded about correct. Unfortunately, the synthesized waveform which resulted from this procedure yields a very mechanical sounding result and might be described as "less rich" or "less sharp" than the original. A frication noise source was added in the region of F4, which helped to add naturalness to the utterance, but did not improve it greatly. One problem may be that the pulses are too regular in amplitude, whereas the original waveform has random period-to-period amplitude variation, known as shimmer. Another option is to try adding flutter, but in test experiments, this also did not seem to correct the problem.

A.5 Synthesis of a Hoarse Voice

The last voice, KAH02, was recorded from a 73 year-old female but the record does not contain an indication of the diagnosis. Perceptually, the distinguishing characteristic of this voice is its extreme noisiness including several obvious, but difficult to describe, tonal components. During the listening survey of the database, the author noticed that several voices have this same "chainsaw voice" characteristic.

As can be observed in Figure 8, the original signal does not exhibit extensive changes over time. Therefore, we focused on attempting to replicate the complicated spectral characteristics as seen in the average spectrum of Figure 10. In particular (1) the nature of the noise source and (2) the complicated harmonic structure were troubling.

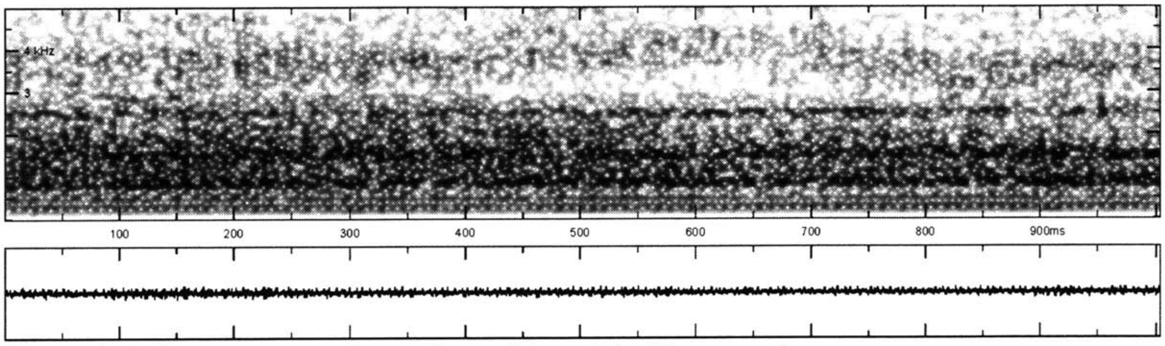

Figure A-8. Spectrogram and waveform of the original hoarse voice.

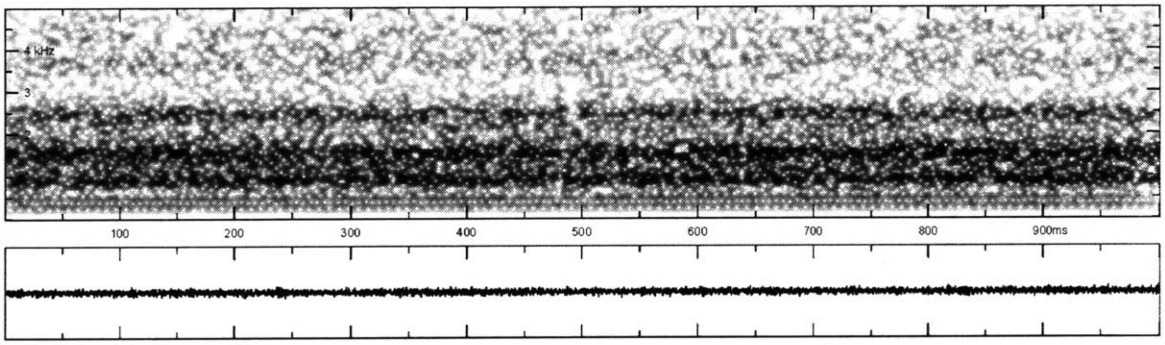

Figure A-9. Spectrogram and waveform of the synthesized hoarse voice.

The first of these issues was a problem mainly because the addition of an aspiration source did not seem to create a realistic match to the original spectrum. After some trial and error, we chose to use the synthesizer's frication noise source processed using the observed formant resonances and bandwidths. The major complication with doing this was that the frication source was only able to be processed by formants above F1. In order to be able to synthesize frication noise through F1, we added a 6th formant and moved the frequencies and bandwidths of each existing formant to the next one up. In this way, F2 became the former F1, F3 became the former F2, and so on. F1 was left as a dummy formant and received no frication source. Once the formant structure was in place, we were able to synthesize the spectrum of the frication by

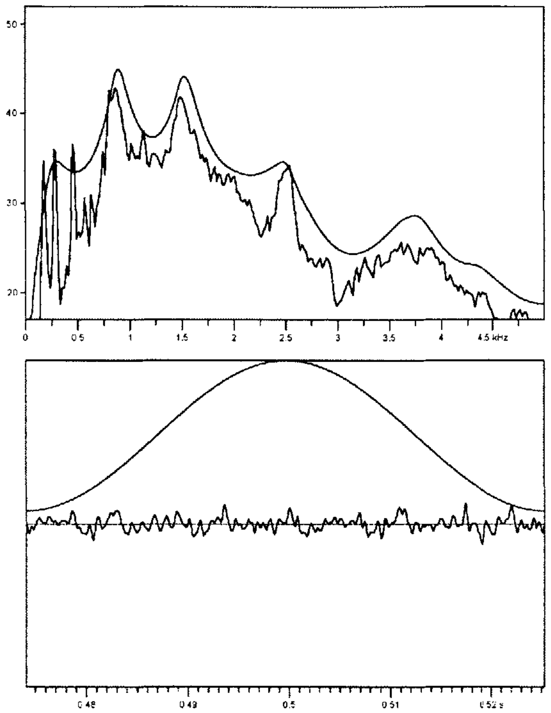

Figure A-10. Average spectrum of the original hoarse voice across the entire file (left) compared with the average spectrum of the synthesized voice (right).

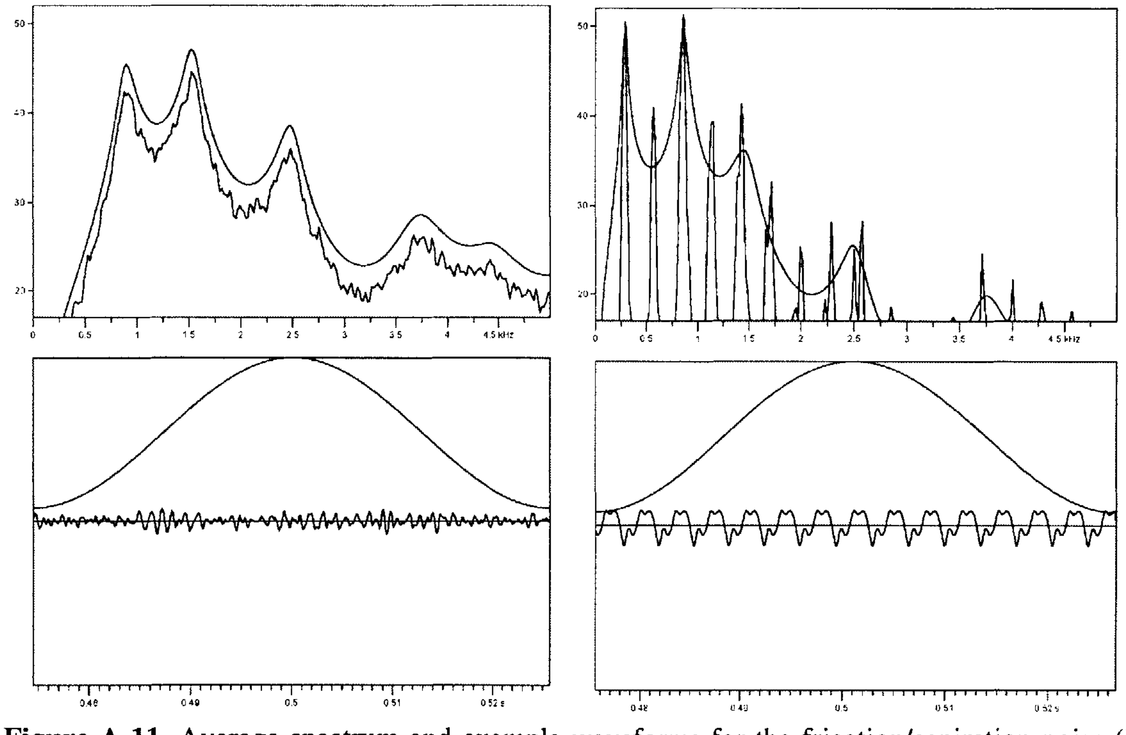

Figure A-11. Average spectrum and example waveforms for the frication/aspiration noise (left) and 280 Hz F0 components of the total waveform (right).

individually manipulating the amplitudes of each formant peak using the frication amplitude parameters. As shown in Figure 11, this technique produced a very realistic match, both visually and acoustically, to the noise portion of the original waveform.

The second problem that was addressed is mimicking the three harmonics present at 168, 280, and 448 Hz, all of which were multiples of 56 Hz. Most of the low-frequency energy is at these frequencies with very little activity in between. The first attempt to model this behavior used a single voicing source along with a series of pole-zero pairs to eliminate harmonics of 56 Hz other than those which were desired. Unfortunately this solution proved too difficult with the tools at hand, and the approach was redesigned. A more successful technique is to separately synthesize each of the voicing components as a separate waveform and then to form a weighted sum of the result with the frication waveform. The spectrum of the 280-Hz component is depicted in Figure 11 and the sum of all components is compared with the original in Figure 10. The weighting added 100% of the frication sound to 20% of the 280-Hz signal, 15% of the 448 Hz signal, and 10% of the 168 Hz signal. This combination was derived through trial and error and resulted in a synthesized voice that was quite realistic.

The question arises as to whether the proposed synthesis method is physiologically plausible. If the underlying fundamental frequency really is around 56 Hz, how does one explain the suppression of most of its harmonics? For example, the amplitudes of components at 112 Hz, 224 Hz, and many other harmonics are extremely difficult to see if it is possible at all. Another hypothesis is that, as synthesized, the glottis of speaker KAH02 has several stable "modes" at which it operates. Each of these modes would act as a somewhat independent glottal source, as is discussed by [3]. In that paper, two modes are used as a possible explanation for diplophonia as synthesized earlier in this paper. Perhaps, with the current speakers, there are somehow three stable harmonically-related modes acting as individual sources. Further investigation of this idea is beyond the scope of the current effort.

A.6 Discussion and Conclusions

As shown in this paper, the dysphonic voices studied have introduced many interesting characteristics. Some of these are summarized below.

(1) Large first harmonic

(2) Complicated harmonic structure

(3) Noise during voicing that cannot be synthesized as aspiration

(4) Superperiodic behavior including diplophonia and other regular patterns of glottal pulses

(5) Time variation of the fundamental frequency

(6) Time variation of formant frequencies

(7) Amplitude modulation created using modulation of the fundamental frequency

Table A.1 summarizes synthesizer parameters used in voice synthesis

As noted in [2], the Klatt synthesizer seemed to fail at modeling "unsteady" voices and some details of the spectra. DAS10, the creaky voice, was the most difficult utterance for us to synthesize realistically. Additionally, we observed the same failure of OQ to properly account for the low frequency characteristics of dysphonic utterances as [2] did. In addition to all of the suggestions for improvements to the synthesis tools that [2] provide, we would add the ability to explicitly specify the repeating pattern of large/small pulses created by the diplophonia parameter.

The time required to create each of the above utterances using existing tools is surprising. A rough estimate is 10-15 hours per vowel, which is in accord with the synthesis in the literature on which we based our work. Therefore, we conclude that, presently, synthesis by clinicians with a high level of detail is probably out of reach. Clinically useful synthesis seem to currently focus on the harmonic-to-noise ratio of the speech signal as demonstrated in [13], arguably a much simpler synthesis task. Therefore, one reason to research improving synthesis of dysphonic voices is to increase the ease with which a practitioner can use the tools. The synthesized utterances resulting from this project, especially JAB08 and KAH02, seem perceptually accurate, and the ability to obtain similar results quickly could prove very useful.

Table A.1. Relevant synthesis parameters used for creating each of the three voices. Note that, as discussed in the text, the hoarse voice shows two different parameter sets—one for the noise component and the other for each of the voiced components.

	Tremor	Hoarse		Creaky
		AF	AV	
F0	128	0	see text	150
AV	60	0	60	see text
OQ	70	50	80	90
TL	22	20	40	5
FL	65	0	0	0
DI	0-10	0	0	0-21
AH	36	30	0	40
AF	0	65	0	55
F1	546-610	280	887	759-782
B1	70	40	140	180
F2	1190	887	1523	1297-1300
B2	200	140	160	130
F3	2783	1523	2479	2375-2390
B3	150	160	200	140-250
F4	3500	2479	3744	3304-3366
B4	500-600	200	350	250
F5	4400	3744	4400	4204-4207
B5	450	350	600	1000
F6	4990	4400	5000	4990
B6	1000	600	600	0
FNP	150	▨	▨	2000
BNP	25	▨	▨	200
FNZ	150	▨	▨	2000
BNZ	100	▨	▨	100
FTP	900	▨	▨	450
BTP	400	▨	▨	100
FTZ	800	▨	▨	450
BTZ	80	▨	▨	200
A2F	▨	55	▨	▨
A3F	▨	55	▨	▨
A4F	▨	45	▨	48
A5F	▨	35	▨	▨
A6F	▨	30	▨	▨
B2F	▨	140	▨	▨
B3F	▨	160	▨	▨
B4F	▨	200	▨	250
B5F	▨	350	▨	▨
B6F	▨	600	▨	▨

Bibliography

[1] "Disordered Voice Database," Version 1.03 ed: Kay Elemetrics Corp., 1994.

[2] P. Bangayan, C. Long, A. A. Alwan, J. Kreiman, and B. R. Gerratt, "Analysis by synthesis of pathological voices using the Klatt synthesizer," *Speech Communication*, vol. 22, pp. 343-368, 1997.

[3] D. A. Berry, "Mechanisms of modal and nonmodal phonation," *Journal of Phonetics*, vol. 29, pp. 431-450, 2001.

[4] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech and Hearing Research*, vol. 39, pp. 126-134, 1996.

[5] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *Journal of the Acoustical Society of America*, vol. 103, pp. 2649-2658, 1998.

[6] L. W. Couch, *Digital and Analog Communication Systems*, 5th ed. Upper Saddle River, NJ: Prentice Hall, 1997.

[7] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation .1. Detection and masking with narrow-band carriers," *Journal of the Acoustical Society of America*, vol. 102, pp. 2892-2905, 1997.

[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, pp. 357-366, 1980.

[9] M. S. DeBodt, F. L. Wuyts, P. H. VandeHeyning, and C. Croux, "Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality," *Journal of Voice*, vol. 11, pp. 74-80, 1997.

[10] A. A. Dibazar and S. Narayanan, "A system for automatic detection of pathological speech," presented at Conference Signals, Systems, and Computers, Asilomar, CA, 2002.

[11] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates of vocal quality," *Journal of Speech and Hearing Research*, vol. 33, pp. 298-306, 1990.

[12] M. Frohlich, D. Michaelis, H. W. Strube, and E. Kruse, "Acoustic voice analysis by means of the hoarseness diagram," *Journal of Speech Language and Hearing Research*, vol. 43, pp. 706-720, 2000.

[13] B. R. Gerratt and J. Kreiman, "Measuring vocal quality with speech synthesis," *Journal of the Acoustical Society of America*, vol. 110, pp. 2560-2566, 2001.

[14] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, pp. 365-381, 2001.

[15] A. Hagen, "The linguistic functions of glottalizations," Erlangen University/MIT Speech Group, 1997.

[16] H. M. Hanson, K. N. Stevens, H. K. J. Kuo, M. Y. Chen, and J. Slifka, "Towards models of phonation," *Journal of Phonetics*, vol. 29, pp. 451-480, 2001.

[17] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578-589, 1994.

[18] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research*, vol. 37, pp. 769-778, 1994.

[19] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech and Hearing Research*, vol. 39, pp. 311-321, 1996.

[20] Y. Horii, "Fundamental-frequency perturbation observed in sustained phonation," *Journal of Speech and Hearing Research*, vol. 22, pp. 5-19, 1979.

[21] Y. Horii, "Vocal shimmer in sustained phonation," *Journal of Speech and Hearing Research*, vol. 23, pp. 202-209, 1980.

[22] D. H. Klatt, "Description of the cascade/parallel formant synthesizer," in *Chapter 3 of a book in preparation.*

[23] D. H. Klatt, "Software for a cascade-parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971-995, 1980.

[24] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820-857, 1990.

[25] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality - Review, tutorial, and a framework for future research," *Journal of Speech and Hearing Research*, vol. 36, pp. 21-40, 1993.

[26] G. Langner and C. E. Schreiner, "Periodicity coding in the inferior colliculus of the cat .1. Neuronal mechanisms," *Journal of Neurophysiology*, vol. 60, pp. 1799-1822, 1988.

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

[28] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America*, vol. 79, pp. 702-711, 1986.

[29] P. Milenkovic, "Least mean-square measures of voice perturbation," *Journal of Speech and Hearing Research*, vol. 30, pp. 529-538, 1987.

[30] F. D. Minifie, "Introduction to Communication Sciences and Disorders." San Diego, CA: Singular Publishing Group, 1994.

[31] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1999.

[32] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech," *Journal of Speech Language and Hearing Research*, vol. 44, pp. 327-339, 2001.

[33] Y. Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *Journal of the Acoustical Society of America*, vol. 105, pp. 2532-2535, 1999.

[34] T. F. Quatieri, *Discrete-Time Speech Signal Processing : Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[35] T. F. Quatieri, T. E. Hanna, and G. C. Oleary, "AM-FM separation using auditory-motivated filters," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 465-480, 1997.

[36] T. F. Quatieri, N. Malyska, and D. Sturim, "Auditory signal processing as a basis for speaker recognition," presented at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain, NY, 2003.

[37] C. R. Rabinov, J. Kreiman, B. R. Gerratt, and S. Bielamowicz, "Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter," *Journal of Speech and Hearing Research*, vol. 38, pp. 26-32, 1995.

[38] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, pp. 407-429, 2001.

[39] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, pp. 173-192, 1995.

[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[41] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, pp. 249-266, 2000.

[42] H. A. Schwid and C. D. Geisler, "Multiple reservoir model of neurotransmitter release by a cochlear inner hair cell," *Journal of the Acoustical Society of America*, vol. 72, pp. 1435-1440, 1982.

[43] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Mass.: MIT Press, 1998.

[44] B. Strobe and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, 1997.

[45] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 106, pp. 2040-2050, 1999.

[46] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modeling*, vol. 55, *NATO Adv. Study Inst. Series D*, H. W.J. and A. Marchal, Eds. Bonas, France: Kluwer Academic Publishers, 1990, pp. 241-262.

[47] I. R. Titze, "Workshop on acoustic voice analysis. Summary statement," National Center for Voice and Speech, Denver, CO 1995.

[48] L. A. Westerman and R. L. Smith, "Rapid and short-term adaptation in auditory-nerve responses," *Hearing Research*, vol. 15, pp. 249-260, 1984.

[49] W. R. Wilson, J. B. Nadol, Jr., and G. W. Randolph, *Clinical Handbook of Ear, Nose, and Throat disorders*. New York, NY: The Parthenon Publishing Group, 2002.

[50] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning, "The Dysphonia Severity Index: An objective measure of vocal quality based on a multiparameter approach," *Journal of Speech Language and Hearing Research*, vol. 43, pp. 796-809, 2000.