Classification of vocal fold vibration as regular or irregular in normal, voiced speech

by

Kushan Krishna Surana

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2006

Author . . . .
Department of Electrical Engineering and Computer Science
February 3, 2006

Certified by . . . .
Janet Slifka
Research Scientist
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Classification of vocal fold vibration as regular or irregular in normal, voiced speech

by

Kushan Krishna Surana

Submitted to the Department of Electrical Engineering and Computer Science
on February 3, 2006, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

Irregular phonation serves an important communicative function in human speech and occurs allophonically in American English. This thesis uses cues from both the temporal and frequency domains — such as fundamental frequency, normalized RMS amplitude, smoothed-energy-difference amplitude (a measure of abruptness in energy variations) and shift-difference amplitude (a measures of periodicity) — to classify segments of regular and irregular phonation in normal, continuous speech.

Support Vector Machines (SVMs) are used to classify the tokens as examples of either regular or irregular phonation. The tokens are extracted from the TIMIT database, and are extracted from 151 different speakers. Both genders are well represented, and the tokens occur in various contexts within the utterance. The train-set uses 114 different speakers, while the test-set uses another 37 speakers. A total of 292 of 320 irregular tokens (recognition rate of 91.25% with a false alarm rate of 4.98%), and 4105 of 4320 regular tokens (recognition rate of 95.02% with a false alarm rate of 8.75%) are correctly identified. The high recognition rates are an indicator that the set of acoustic cues are robust in accurately identifying a token as regular or irregular, even in cases where one or two acoustic cues show unexpected values.

Also, analysis of irregular tokens in the training set (1331 irregular tokens) shows that 78% occur at word boundaries and 5% occur at syllable boundaries. Of the irregular tokens at syllable boundaries, 72% are either at the junction of a compound-word (e.g "outcast") or at the junction of a base word and a suffix. Of the irregular tokens which do not occur at word or syllable boundaries, 70% occur adjacent to voiceless consonants mostly in utterance-final location. These observations support irregular phonation as a cue for syntactic boundaries in connected speech, and combined with the robust classification results to separate regular phonation from irregular phonation, could be used to improve speech recognition and lexical access models.

Thesis Supervisor: Janet Slifka
Title: Research Scientist

Acknowledgments

This thesis was supported by NIH # DC02978.

I would like to thank my advisor, Dr. Janet Slifka for her guidance and support. I've had the pleasure of working with her as a UROP and a graduate student, and have found it to be an invaluable experience.

I would also like to thank the Speech Group, specifically Professor Kenneth Stevens, Joseph Perkell, Stefanie Shattuck-Hufnagel, Helen Hanson, Sartrajit Ghosh, Seth Hall, Arlene Wint, Lan Chen, Yoko Saikachi, Tomas Bohm and Steven Lulich for their support. The Speech Group is truly a fantastic place to work and I will always cherish the time I spent there.

I've been fortunate to have had contact with some amazing people at MIT. There are far too many mention, but I would like to express my thanks to Professors George Verghese and Ronald Parker.

Finally, I would like to thank my family. To my parents for their love, and my brothers Kanishka and Kunal for their unconditional support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Currently, robust systems to classify and detect irregular phonation do not exist. This thesis aims to address this issue and builds upon existing studies of irregular phonation. In this chapter, irregular phonation is defined and described in terms of its segmental and acoustic correlates in the speech waveform. Some common types of irregular phonation are also described and the specific aim of this thesis is detailed.

1.1 Irregular phonation

The source-filter model of speech production, as set up by Fant (1960), proposes that human speech is a consequence of the generation of one or more sources of sound and the filtering of these sounds by the vocal tract. One type of sound source results from the vibration of the vocal folds and is the result of a delicate balance of the subglottal air pressure that drives the folds apart, and the muscular, elastic and Bernoulli forces that bring them together. Sounds produced in this manner are generally referred to as voiced sounds.

Normal, voiced speech, or regular phonation, is characterized by the quasi-regular vibration of the vocal folds. Although the vocal folds will oscillate quasi-regularly in general when the variables transglottal pressure, vocal fold tension, and vocal fold adduction — among others — are in particular ranges, irregularities in the vocal fold vibrations are observed for certain combinations of the values of these variables.

These irregularities in vocal fold vibration lead to the observation of irregularities in the speech waveform, and are more pronounced than the small-scale cycle-to-cycle variations observed in quasi-periodic, normal, voiced speech.

The small-scale variations during normal, voiced speech mentioned above have been enumerated and defined by Titze (1995, p. 338-340):

- **jitter**: "a short-term (cycle-to-cycle) variation in the fundamental frequency of a signal."
- **shimmer**: "a short-term (cycle-to-cycle) variation in the amplitude of a signal."
- **perturbation**: "a disturbance, or small change, in a cyclic variable (period, amplitude, open quotient, etc.) that is constant in regular periodic oscillation."
- **tremor**: "a 1-15 Hz modulation of a cyclic parameter (*e.g.*, amplitude or fundamental frequency), either of a neurologic origin or an interaction between neurological and biomechanical properties of the vocal folds."

Papers dealing with the subject of voice quality and phonation often use the terms *"modal"* and *"periodic"* interchangeably with regular phonation. Similarly, *"nonmodal"* and *"aperiodic"* are often used to denote irregular phonation. This thesis avoids the use of these terms as they are not synonymnous with regular or irregular phonation. For example, nonmodal phonation includes irregular, aperiodic phonation such as vocal fry as well as regular, periodic phonation such as breathy voice. Regions in the speech waveform with very low frequency, periodic glottal pulses are also not typical of the quasi-periodic pulses in the phonation for a given speaker at a given time with the auditory impression of a "...rapid series of taps, rather like the sound of a stick being run along a railing."(Catford, 1977, p.98). These regions are classified as irregular in this thesis, in spite of being periodic.

Based on an initial survey of the literature and the specific aims of the system, a specific definition for irregular phonation was formulated to contrast it with regular phonation and its small-scale variations:

> **"A region of speech is an example of irregular phonation if the speech waveform displays either an unusual difference in time or amplitude over adjacent pitch periods that exceeds the small-scale jitter and shimmer differences or an unusually wide-spacing of the glottal pulses compared to their spacing in the local environment, indicating an anomaly from the usual, quasi-periodic behavior of the vocal folds."**

Irregular phonation occurs in a number of contexts in American English, ranging from a single glottal closure accompanying a consonantal segment to a change in voice source characteristics over a region encompassing several segments or even syllables. Irregular phonation also commonly occurs allophonically in certain contexts. For example, in American English, vowel initial words may be produced irregularly at onset (e.g "elephant") (Dilley & Shattuck-Hufnagel, 1995); in syllable-final environments, voiceless stop consonants, particularly /t/, may be realized as a glottal stop (e.g. in "hat rack") (Pierrehumbert, 1995); and allophonic irregular phonation may often be associated with vowels adjacent to a glottal stop, with languages differing in the duration of this allophonic irregularity (Blankenship, 2002).

The study of irregular phonation is also relevant for languages other than American English. Gordon and Ladefoged (2001) completed a survey which shows how languages use irregular phonation contrastively to distinguish among word forms. Hausa and certain other Chadic languages use irregular phonation contrastively for stops. Some other Northwest American Indian languages, e.g., Kwakwála, Montana Salish, Hupa, and Kashaya Pom, contrast irregular and regular voicing among their sonorants. Laver (1980) and others have also suggested that certain languages use irregular phonation to signal a speaker's turn. For example, irregular phonation may mark the end of a turn in London Jamaican (Local, Wells & Sebba, 1985).

In acoustic terms, irregular phonation is generally associated with irregularly spaced pitch periods and is often accompanied by other characteristics, such as full damping, low F0, breathiness or low amplitude (Ladefoged, 1971; Fischer-Jorgenson, 1989; Klatt & Klatt, 1990; Pierrehumbert & Talkin, 1992). These characteristics are believed to contribute to the perceptual impression of a glottal gesture or disturbance in the regular voice quality (Rozyspal & Millar, 1979; Hillenbrand & Houde, 1996; Pierrehumbert & Frisch, 1997).

Various theories and studies have tried to explain the physiological basis for irregular phonation. One theory suggests that from the perspective of vocal fold dynamics, regular and irregular phonation may be distinguished based on the entrainment or lack of entrainment of natural vibratory modes of the vocal folds, called eigenmodes (Berry, 2001). Slifka (2000) conducted a study which suggests that as the glottal configuration moves from one setting to another, it could move through regions of instability leading to irregular phonation. Hanson, Stevens, Kuo, Chen & Slifka (2001) have tried to explain the physiological variations during irregular phonation by exploring how the glottal waveform and vocal tract transfer function are affected by the various patterns of complete/incomplete/nonsimultaneous closing of the vocal folds during phonation. These studies contrast the incomplete closing of the vocal folds in irregular phonation to regular phonation which has been defined as phonation in which full contact occurs between the vocal folds during the closed phase of a phonatory cycle (Titze, 1995).

1.2 Types of irregular phonation

The articulatory mechanism may affect the kinds of irregular vocal fold vibrations produced. Over the years, researchers have classified irregular phonation into sub-groups based on a combination of physiological, perceptual and acoustic characteristics. Various terms have been used interchangeably to describe these sub-groups with papers dedicated to establishing a taxonomy of irregular phonation (Gerrat & Kreiman, 2001). This section describes some of of these terms.

- **Creaky phonation** : "...typically associated with vocal folds that are tightly adducted but open enough along a portion of their lengths to allow for voicing" (Gordon & Ladefoged, 2001, p.386). This is often accompanied by irregularly spaced pitch periods and decreased acoustic intensity relative to regular phonation (Gordon & Ladefoged, 2001, p.387). An example of creaky phonation is shown in Figure 1-1 (a).

- **Vocal fry** : It is usually defined as a train of discrete, laryngeal excitations of extremely low frequency, with almost complete damping of the vocal tract between excitations (Hollien, Moore, Wendahl & Michel, 1966) giving the auditory impression of a "...rapid series of taps, rather like the sound of a stick being run along a railing."(Catford, 1977, p.98). One of vocal fry's distinct characteristics is that the vocal folds tend to vibrate so slowly that individual vibrations can be perceived (Colton & Casper, 1996). Vocal fry is characterized by a very short open period and a very long period where the vocal folds are completely adducted (Blomgren, Chen, Ng & Gilbert, 1998, p.2650). Zemlin (1988, p.166) reported that examination of vocal fry with high speed photography revealed that "...the folds are approximated tightly, but at the same time they appear flaccid along their free borders, and subglottal air simply bubbles up between them at about the junction of the anterior two-thirds of the glottis". An example of vocal fry is shown in Figure 1-1 (b).

- **Glottalization** : Titze (1995, p.338) has defined it as "...transient sounds resulting from the relatively forceful adduction or abduction of the vocal folds [with the perceptual impression of] a voice that contains frequent transition sounds (clicks)." Huber (1992) defines this term as an initial vibratory cycle clearly demarcated from the rest of the periodic glottal vibrations, which is in contrast to the more common reference of glottalization occurring at other locations in the speech signal, including in phrase final position. An example of glottalization is shown in Figure 1-1 (c).

- **Diplophonia** : "...simultaneous production by the voice of two separate tones" (Ward, Sanders, Goldman & Moore, 1969, p.771). Titze (1995, p.337) restricts the two tones to be dependent, the frequency of one tone an octave lower than the other, but this study assumes no such rational dependence. An example of diplophonia is shown in Figure 1-1 (d).

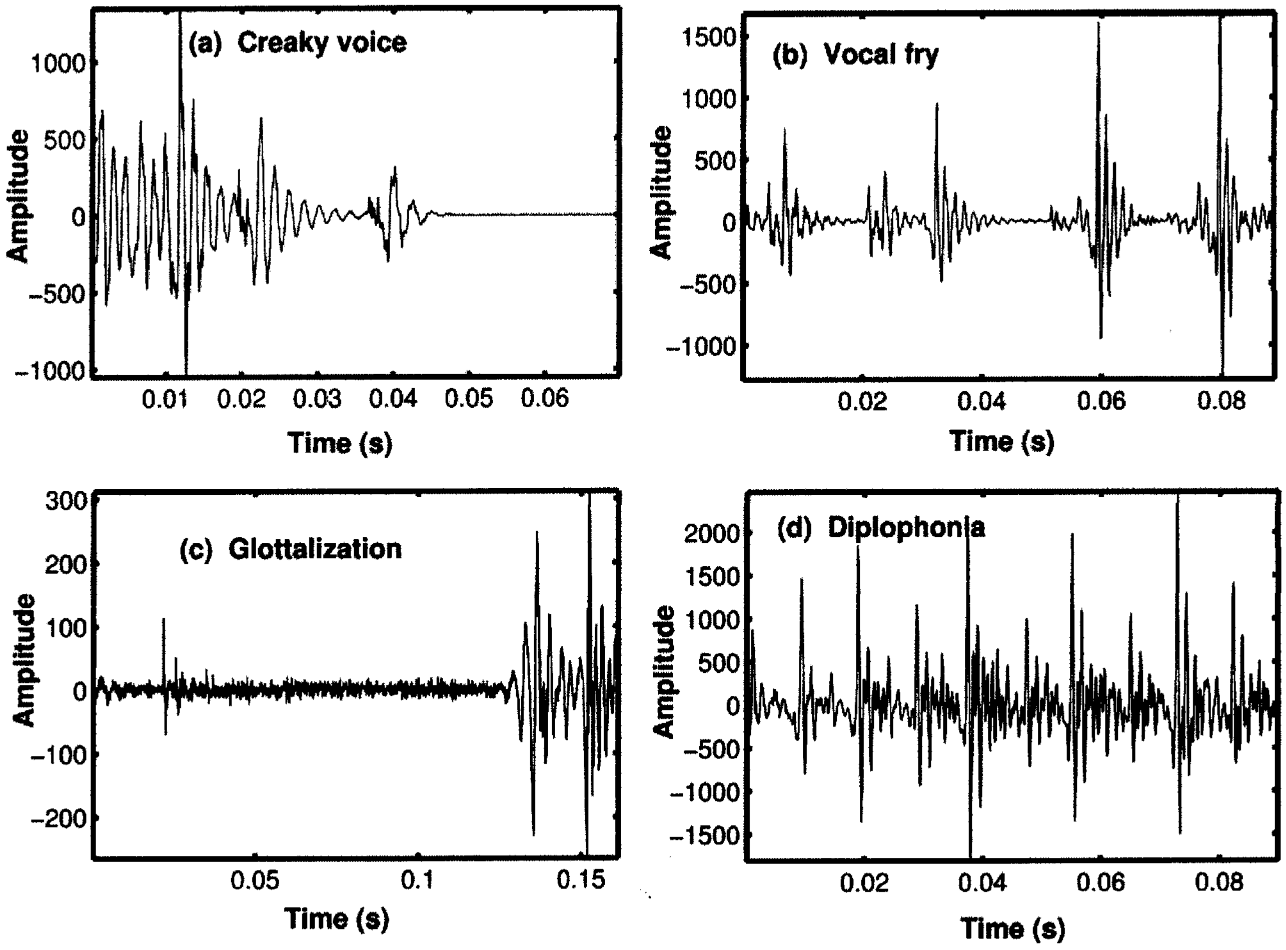

Figure 1-1: Some different types of irregular phonation: (a) Creaky voice (b) Vocal fry (c) Glottalization (d) Diplophonia. (Source of waveforms: TIMIT, 1990)

The examples above offer a glimpse into the range of variations in irregular phonation in normal speech. Some of the definitions offer concrete physiological characteristics associated with a particular type of irregular phonation, but a lot more remains to be understood regarding the physiological mechanism of irregular phonation production. Detailed models of vocal fold functions such as those developed by Titze & Talkin (1979) and studies done by Hanson *et. al.* (2001) and Slifka (2005) may help enhance our understanding about irregular phonation.

1.3 Specific Aim

This thesis attempts to use signal-processing techniques, in either the temporal or frequency domain, to analyze the systematic variations in examples of irregular vocal fold vibration that distinguish them from examples of regular vocal fold vibration. Acoustically, this translates to proposing a set of acoustic cues capable of distinguishing between regions of periodic, glottal pulses and (1) regions of aperiodic pulses, (2) single aperiodic pulses or, (3) regions of atypically large spacing between adjacent glottal pulses (as compared to the glottal pulse spacing in the local environment).

Another aim of this thesis is to study the context for occurrences of irregular phonation. In other words, given an occurrence of irregular phonation, is there a specific context where it is more likely to occur than others? The following contexts are observed: (1) utterance-boundaries, (2) word-boundaries, (3) syllable-boundaries, (4) voiceless-stops /p/, /t/ and /k/ and (5) vowel-medial locations.

Chapter 2

Motivation

Irregular phonation in the form of regions of creakiness, period doubling, irregular pitch periods and amplitude modulation can occur in the speech of normal as well as pathological talkers (Docherty, 2001, p.364). This chapter details the benefits of an accurate classification system to distinguish between regular and irregular phonation.

2.1 Lexical Access From Features (LAFF) Project

The work done in this thesis falls under the purview of the Lexical Access From Features (LAFF) Project (Stevens, 2002), which proposes a model where words are represented in the mental lexicon as a sequence of segments, each of which is described by a set of binary distinctive features. Although the results of this work are widely applicable, this section of the thesis focuses exclusively on the role this thesis plays within the LAFF Project. In order to adequately describe this role, a brief overview of the project is required.

2.1.1 Theory

The LAFF Project considers words to be represented as sequences of segments (also referred to as phonemes), each of which can be defined by a set of binary distinctive features (Jakobson, Fant, and Halle, 1952). These features specify the phonemic contrasts that are used in a particular language so that a change in one feature leads to a different word. The project proposes the existence of a universal set of features, with every language defined by a unique subset drawn from these features. The end goal of the project is to decompose any utterance into a series of feature bundles, assign a probabilistic estimate to the features within a segment and arrive at a hypothesis for the underlying word sequence. In order to achieve this goal, the initial focus is on building a system which can correctly identify all the distinctive features for American English.

As a first step towards the goal of arriving at a feature set, the LAFF Project aims to identify "landmarks" for the acoustic waveform. Landmarks are regions in the acoustic waveform which either show a peak in low-frequency amplitude, a low-frequency minima or acoustic discontinuities. These landmarks are detected based on amplitude changes in various energy bands (Stevens, 2002; Lin 1995; Slifka, Stevens, Manuel, Shattuck-Hufnagel, 2004).

The type of landmark region provides evidence for a broad class of distinctive features called "articulator-free" features. These features refer to the general characteristic of the constrictions within the vocal tract and the acoustic consequences of these constrictions. There is another class of features called "articulator-bound" features which are derived from the acoustic cues sampled near the landmark region. The articulator-bound features provide information about the action of the particular articulator used in producing the phoneme. Table 2.1 (Slifka *et. al*, 2004) shows a list of distinctive features for American English grouped by articulator-free and articulator-bound classes.

Each phoneme is characterized by a unique combination of these articulator-free and articulator-bound features. The feature set is arranged in a heirarchical structure, which implies that the entire feature set does not need to be specified since some features can be inferred from others. Table 2.2 (Stevens, 2002) shows the lexical representations of the words "debate", "wagon" and "help" to illustrate this point.

Each distinctive feature is considered to have a defining articulatory action and a correspoding acoustic correlate. An example is the feature [back] for vowels. For

Table 2.1: List of distinctive features for American English grouped by articulator-free and articulator-bound classes (Slifka *et. al*, 2004)

Articulator-free	**Articulator-bound features**		
features	Vowel and glide	Consonant	
Vowel	High	Lips	Lateral
Consonant	Low	Tongue blade	Rhotic
Continuant	Back	Tongue body	Nasal
Sonorant	Adv. Tongue root	Round	Stiff vocal folds
Strident	Spread glottis	Anterior	

Table 2.2: List of distinctive features for the words "debate", "wagon" and "help" (Stevens, 2002)

	debate						**wagon**						**help**			
	d	ə	b	e	t		w	a	g	ə	n		h	e	l	p
vowel		+		+				+		+				+		
glide							+						+			
consonant	+		+		+				+		+				+	+
stressed		-		+				+		-				+		
reducible		+		-				-		+				-		
continuant	-		-		-				-		-				-	-
sonorant	-		-		-				-		+				+	-
strident																
lips			+													+
tongue blade	+				+						+				+	
tongue body									+							
round			-				+									-
anterior	+				+						+				+	
lateral															+	
high		+		-			+	-	+	-				-		
low		-		-			-	+	-	-				-		
back		-		-			+	-	-	+				-		
adv. tongue root				+			+	-						-		
spread glottis													+			
nasal											+					
stiff vocal folds	-		-		+				-							+

[+back] vowels, the tongue body is displaced back to form a narrowing in the pharyngeal or posterior oral cavity. The acoustic consequence is a second-formant frequency that is low and close to the first-formant frequency. Vowels classified as [-back], on the other hand, are produced with the tongue body forward and a high second-formant frequency (Stevens, 2002).

It is clear that based on the model proposed by the LAFF Project, landmark detection is inarguably the first and most important step in finding the underlying word sequence of an utterance.

2.1.2 Relevance of irregular phonation

In addition to landmarks, there are other regions in the utterance which might show characteristics similar to those observed for landmarks (i.e. peaks, valleys or discontinuities in certain frequency ranges). Some of these regions are classified as "acoustic events". Irregular phonation is one example of such an acoustic event. The presence of irregular phonation can hence result in incorrect landmark indentification. Due to the frequent occurrence of irregular phonation in normal speech, detecting this particular acoustic event and distinguishing it from landmarks is especially important. To cite one example where irregular phonation is responsible for incorrect landmark identification, regions of irregular phonation are often classified as vowel landmarks by the landmark identifier. If the regions of irregular phonation are correctly identified, the misclassification of landmarks could be greatly reduced, which would result in more accurate articulator-free feature identification and also lay a more robust framework for articular-bound feature identification.

The identification of irregular phonation could also help in the detection of utterance and word boundaries and lay the groundwork for estimating the prosodic structure of an utterance (see Chapter 8 for further discussion) — both of which are relevant to the LAFF Project.

2.2 Applicability beyond LAFF Project

Huber (1992, p.503) has conducted experiments to show that human listeners use irregularity in speech signals for segmentation purposes. These results are collaborated by Blomgren, Chen, Ng & Gilbert (1998) who also observed that listeners were consistently able to perceive glottal fry. Huber's research (1992), with that of Kreiman (1982), show that irregular phonation is an important demarcation cue in connected speech, used to support the segmentation of the continuous speech utterance into relevant information units in American English. Their research suggests that a better understanding of irregular phonation is essential to develop accurate and robust automatic speech recognition systems and human-like speech synthesis systems.

Irregular vocal phenomenon is also used to convey linguistic and nonlinguistic information. Gordon & Ladefoged (2001, p.383) note that a difference in phonation type might indicate a contrast between otherwise identical lexical items and boundaries of prosodic constituents in many languages. Their statement is substantiated by research done by Dilley, Shattuck-Hufnagel & Ostendorf (1996), Pierrehumbert & Talkin (1992) and Pierrehumbert (1995) who state that irregular phonation could be exploited as a cue for recognizing prosodic patterns. This could improve automatic detection of prosodic markers, both for corpus transcription and for speech understanding applications (Dilley, Shattuck-Hufnagel & Ostendorf, 1996, p.438).

The detection of irregular phonation is also of interest for pathological speech. There are numerous medical conditions that affect voice quality. Many such conditions have their origins in the vocal system and the tools available for the detection of these speech pathologies are invasive or require expert analysis. Hence, a reliable, accurate and non-invasive automatic system for recognizing and monitoring speech abnormalities is one of the necessary tools in pathological speech assessment (Dibazar, Narayanan & Berger, 2002).

Since irregular phonation often interrrupts the periodicity of the speech segment, a key understanding of it will also aid in developing better F0 estimation algorithms. If an algorithm were to be developed to correctly identify regions of irregularity, incorrect F0 estimates for those regions could be avoided.

The relatively frequent occurence of irregular phonation in normal speech across languages, combined with its usefulness in terms of the acoustic cues it provides, makes its comprehensive study essential towards establishing a complete model of speech production and in developing robust algorithms for pitch detection, speech-synthesis and automatic speech recognition.

Chapter 3

Prior Work

There exists a wide range in the rate of occurrence of irregular phonation across individual speakers (Huber, 1988; Dilley *et. al.*, 1996; Dilley & Shattuck-Hufnagel, 1995). Irregular phonation also occurs more often at certain locations in an utterance over others. For example, Redi & Shattuck-Hufnagel (2001) found a higher rate of irregular phonation on words at the ends of utterances than on words at the ends of utterance-medial intonational phrases. In spite of the speaker-to-speaker and the context-to-context variations of irregular phonation, an ideal classification system trained to distinguish between regular and irregular phonation should be speaker-independent and context-independent.

There have not been many automatic classification systems proposed to classify regular and irregular phonation. To the author's knowledge, only Ishi (2004) and Kiessling, Kompe, Niemann, Nöth & Batliner (1995) have addressed this topic explicitly.

3.1 Kiessling, Kompe, Niemann, Nöth & Batliner, 1995

Kiessling *et. al*, 1995 proposes a recognition scheme for classifying frames of irregular phonation (referred to as "laryngealization" in the paper) from regular phonation using two approaches; the first based in the frequency domain, and the second in the temporal domain. The database used in the study contained 1329 sentences from 4 speakers (3 female, 1 male) for a total of 30 minutes of speech. Frames of irregularity were labeled by two trained phoneticians resulting in 1191 frames.

The first approach in this study used cues from the spectrum, the cepstrally smoothed spectrum and the cepstrum of the speech waveform to disinguish between regular and irregular phonation. These cues were extracted based on the observation that the spectra and cepstra of irregular phonation differ from regular phonation (for example, a lack of a regular harmonic structure was observed in the cepstrally smoothed spectrum of irregular segments as compared to regular segments). Based on these differences, the following five cues were proposed:

- the sum of the vertical distances of neighboring extrema in the cepstrally smoothed spectrum below 1700 Hz.
- the average vertical distance of neighboring extrema in the cepstrally smoothed spectrum below 1700 Hz.
- the location of the absolute maximum in the cepstrum.
- the height of the absolute maximum in the cepstrum.
- the quotient of the largest and the second largest maximum in the cepstrum of the center-clipped signal (to eliminate the influence of the vocal tract).

These five cues were combined with normal mel-cepstral coefficients to train a phone component recognition system. The system was originally set to distinguish between 40 different phones using 11 mel-cepstral coefficients per frame and a Gaussian classifier, automatically clustering into 5 clusters per phone and a full covariance matrix. For all the phones which had more than 100 frames labeled as irregular in the database, a new additional phone label was introduced increasing the number of phones from 40 to 51. The 40 regular phones were mapped into one class and the remaining 11 irregular phones into another. The first portion of the experiment was speaker-dependent for multiple speakers and yielded a recognition rate of 80% with a false alarm rate of 8% for irregular phonation. The second portion of the experiment used three speakers for training and one for testing to obtain a recognition rate of 67% with a false alarm rate of 7% for irregular phonation.

The second approach in this study used time domain cues. The approach proposed an inverse filtering technique using artificial neural networks. The output of the neural network was classified into three classes: unvoiced, regular voiced and irregular voiced. The sample values of the neural-network filter output were used as input for another artificial neural network trained to discriminate between the three classes. This approach resulted in a 65% recognition rate with a false alarm rate of 12% for irregular phonation. The paper does not mention if these results are speaker-dependent or speaker-independent.

3.2 Ishi, 2004

Ishi, 2004 attempts to classify irregular phonation, referred to as "creaky voice", from regular and aspirated segments of speech using the ratio of the first two peaks of the autocorrelation function of the glottal excitation waveform as a primary cue.

In the study, the speech signal was first high-pass filtered at 60 Hz in order to prevent the waveform from gradually rising or falling. A 1^{st}-order LPC-analysis was applied to the speech waveform. The estimated coefficient is referred to as the adaptive pre-emphasis coefficient (APE) in the study. The speech signal was then pre-emphasized using the APE, and subsequently 18^{th}-order LPC-analysis was applied on the pre-emphasized signal. The obtained LPC coefficients were used for inverse filtering of the high-pass filtered speech signal. The residual signal was treated as the glottal excitation waveform.

The glottal excitation waveform was low-pass filtered at 2 kHz, before estimating the autocorrelation function (ACF) to make ACF peak detection easier. The window-size for the ACF was chosen in two steps. First, the ACF was estimated in an 80 ms window. The time lag of the maximum peak was extracted and multiplied by four to be used as the new window size, restricting the window size to lie between 16 ms and 80 ms. The obtained ACF function was normalized using the following expression,

$$NAC(L) = \left(\frac{N}{N-L} \times \frac{R_{xx}(L)}{R_{xx}}\right)$$

where N is the number of samples in the frame window, L is the number of samples of the autocorrelation lag and R$_{xx}$ is the autocorrelation function.

The study proposes a clear periodicity in the ACF for regular phonation with the NAC peaks close to 1, and no small peaks between time lag 0 and the first big peak, due to the regular structure of the glottal excitation waveform. For creaky voice, the study notes either the observation of one or more smaller peaks between time lag 0 and the first big peak due to the difference in amplitude over successive glottal pulses in the glottal excitation waveform while for vocal fry, the study notes the presence of a big peak with a narrow width due to the impulse-like shape of the glottal excitation waveform. Based on these visual observations from the NAC of the glottal excitation waveform, the first two peaks from time lag 0 in the NAC, called P1 and P2, are used to characterize different phonation types. A threshold of 0.2 was used to detect peaks in the NAC.

The following cues are proposed based on these two peaks P1 and P2:

Peak magnitude (NAC) value ratio $NACR = 1000 \times \frac{NAC(P2)}{NAC(P1)}$

Peak position (time lag) ratio $TLR = 2000 \times \frac{TL(P2)}{TL(P1)}$

Peak width ratio $WR = 1000 \times \frac{W(P2)}{W(P1)}$

Maximum peak magnitude $NAC_{max} = NAC(P_{max})$

Maximum peak position $TL_{max} = TL(P_{max})$

Maximum peak width $W_{max} = W(P_{max})$

Table 3.1 and 3.2 show the expected values of these cues for regular and irregular phonation.

Table 3.1: Expected values of the cues for regular and double periodic irregular phonation (Ishi, 2004)

	NACR	**TLR**	**WR**	**$\mathbf{NAC_{max}}$**
(Single) Periodicity regular	$\cong 1000$	$\cong 1000$	$\cong 1000$	$\cong 1000$
(Double) Periodicity irregular	> 1000	$\neq 1000$	< 1000	< 1000

Table 3.2: Expected values of the cues for low fundamental frequency irregular phonation (Ishi, 2004)

	TL_{max}	W_{max}
Low frequency irregular phonation	Big	Small

The study uses a dataset containing 404 phrase-final syllables segmented from natural spontaneous speech of a single female adult speaker. Each syllable of the dataset was labeled as either Creaky(C), Modal(M) or Aspirated(A) by looking at the waveform and hearing the segments, leading to a dataset of 5619 frames.

A preliminary evaluation, using a decision tree for each of the three categories, resulted in 91.5% of correct identification of the frames in all the categories. Specifically for the creaky category, the deletion error was 13.7% while the substitution error was 7.9% .

3.3 Comments

Although both studies show some promise in the classification of regular and irregular phonation, a few limitations in the studies must be pointed out. Both the studies used a limited number of speakers — Kiessling's study used four speakers while Ishi's study used only one female speaker. Since irregular phonation is expected to show a high degree of inter-speaker variability, the limited number of speakers is of concern. In addition, Kiessling *et. al.*'s results are speaker-dependent since the same speakers are used for training and testing the system, while Ishi's study is both speaker-dependent and context-dependent because only a female speaker is used to gather data and the regions of irregular phonation occur solely in phrase-final position. As stated at the beginning of this chapter, a robust classification scheme should make the classification of regular and irregular phonation speaker-independent and context-independent.

Essentially, both studies have provided preliminary evidence that differences exist between regular and irregular phonation. Kiessling *et. al.* (1995) perform their analysis in the frequency domain while Ishi's study (2004) is in the temporal domain. The differences between regular and irregular phonation will be further explored in this thesis in the hope of building a more general classification scheme for distingushing regular phonation from irregular phonation — one that is both speaker-independent and context-independent.

Chapter 4

Speech corpora

4.1 Choice of Database

Speech materials used in this study come from the TIMIT corpus (1990), a phonetically-labeled database of isolated utterances, recorded with a 16 kHz sampling rate. The database includes time-aligned orthographic, word, and phone transcriptions. The database consists of a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States (TIMIT, 1990). The speech material is subdivided into portions of training and testing, making the choice of training and testing data self-evident. In this study, only a subset of the database is used — those utterances produced by speakers from the dialect regions "Northern" (drl) and "New England" (dr2).

The TIMIT database is well-known within the speech community and one of its uses is to provide speech-data for the development of automatic speech recognition systems. An important reason for choosing the TIMIT database is the large amount of data it provides for multiple speakers from different regions. This is especially important for irregular phonation, where inter-speaker variation is common. In addition, both males and females are well-represented in the database. Also, the database consists of read, continuous speech which is more consistent with speech that we encounter everyday without extraneous supra-segmental effects one might find in another database — for example, the BUFM database (Ostendorf *et.al.*, 1995). Hence, an algorithm trained and tested on this data-set has wider applicability for improving existing speech-recognition, speech-parsing and speech-synthesis systems. Finally, TIMIT is a well-know corpus which has been used extensively for speech research allowing easier reproduction and corroboration of the results obtained in this thesis.

Once the TIMIT database was selected based on the reasons mentioned above, the two dialect regions chosen were scanned for regions of regular and irregular phonation.

Vowels generally show a quasi-regular structure in normal, voiced speech. Tokens of regular phonation were specifically extracted from stressed vowels in the database, since they are generally characterized by long duration and are less susceptible to co-articulation. The symbols for the vowels used as regions of regular phonation are enumerated along with an example word in Table 4.1.

The TIMIT database contains the phone label, 'q' or glottal stop, which is used to label an allophone of /t/ or to mark an initial vowel or vowel-vowel boundary. The criteria for applying this label 'q' is not tied to the acoustic realization, as is the case in this study, and is not used to label all possible cases of irregular phonation. For these reasons, the irregular tokens were hand-labeled. The labeling was conducted by analyzing the waveform in both the temporal and frequency domains and by hearing the speech-waveform repeatedly when required. As stated in Chapter 1, regions within the speech-waveform are labeled as irregular under the following conditions:

- **if adjacent glottal pulses show unusual irregularities in time or amplitude**

- **if the spacing between adjacent glottal pulses is unusually large, compared to the spacing of the glottal pulses in the immediate local environment.**

4.2 Database characteristics

Figure 4-1 shows the distribution of regular and irregular tokens based on their duration using a boxplot (box-and-whiskers plot). The boxplot is a useful way of plotting

Table 4.1: List of vowels used to denote regions of regular phonation along with an example word

\iy\	beet
\ey\	bait
\ae\	bat
\aa\	bott
\aw\	bout
\ay\	bite
\ao\	bought
\oy\	boy
\ow\	boat
\uw\	boot
\ux\	toot
\er\	bird
\axr\	butter

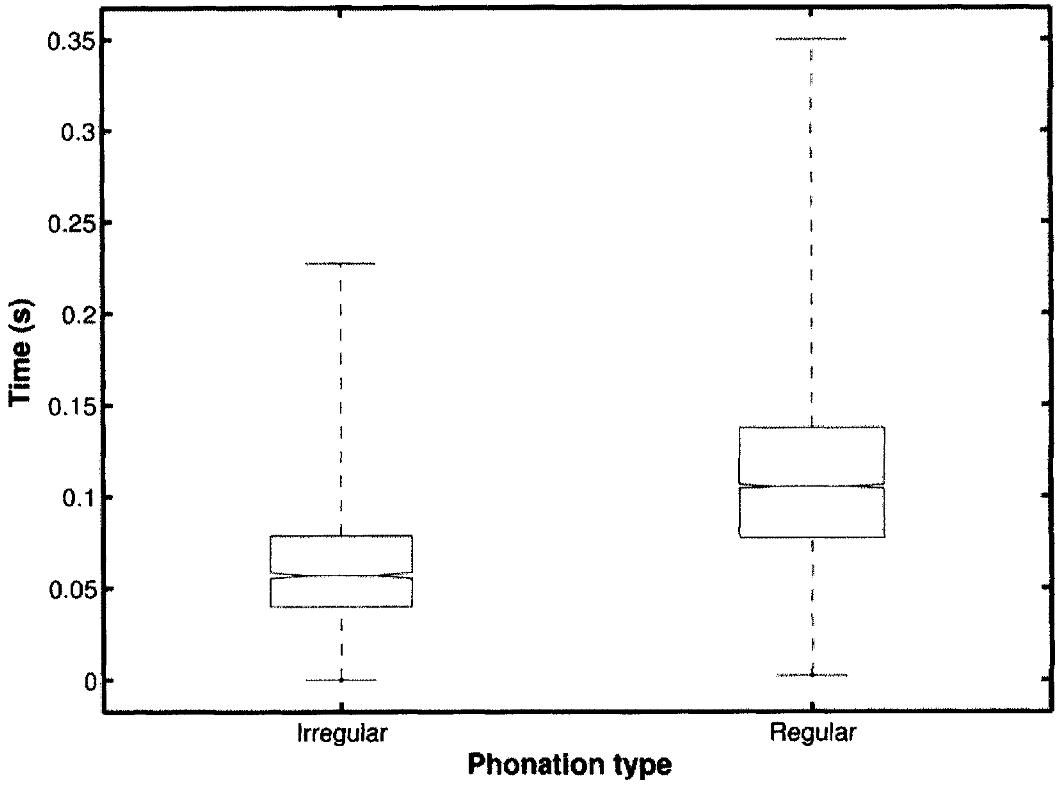

Figure 4-1: Duration of regular and irregular tokens

Table 4.2: Number of regular and irregular tokens based on duration of tokens

Duration of tokens (s)	No restriction	.030 s	.040 s	.050 s	.060 s
Number of regular, male tokens	5554	5458	5345	5154	4890
Number of regular, female tokens	2642	2597	2562	2491	2378
Total number of regular tokens	**8196**	**8055**	**7907**	**7645**	**7268**
Number of irrregular, male tokens	794	735	607	473	339
Number of irrregular, female tokens	609	544	444	363	291
Total number of irregular tokens	**1403**	**1279**	**1051**	**836**	**630**

the five quantiles of the data. The ends of the whiskers show the position of the minimum and maximum of the data whereas the edges and line in the center of the box show the upper and lower quartiles and the median. The whiskers show the behavior of the extreme outliers. Table 4.2 shows the number of tokens for regular and irregular phonation, broken down by gender, based on the duration of the tokens.

Chapter 5

Method

This thesis uses a knowledge-based approach, rather than solely a data-driven one, to develop a set of acoustic cues that can separate regular phonation from irregular phonation. Different methods were explored to compute and normalize these cues. The separation of these cues was subsequently tested using various statistical classifiers in smaller pilot studies, and a process of iteration resulted in the final choice of four acoustic cues which can distinguish between regular and irregular phonation. This chapter describes these acoustic cues.

5.1 Cue selection

Fundamental frequency, normalized root mean square amplitude, smoothed-energy-difference amplitude and shift-difference amplitude are the four cues chosen to distinguish regular phonation from irregular phonation. Their method of computation and a detailed overview on the rationale behind choosing these cues will be presented in the following section. These cues are chosen based on the observation that irregular phonation is accompanied by clear peculiarities in the signal in the form of either a lack of periodicity, strong variations of the amplitude, very long pitch periods or special forms of the damped wave which are not observed in regular phonation.

Table 5.1: Expected behavior of an ideal F0 estimator to distinguish between regular and irregular phonation.

	F0 output
Irregular (abnormal spacing)	$<$ 72 Hz (Blomgren *et al.*, 1998)
Irregular (lack of structure)	0 Hz (i.e. No fundamental frequency estimate)
Regular	86 Hz - 170 Hz (males) (Blomgren *et al.*, 1998) 175 Hz - 266 Hz (females) (Blomgren *et al.*, 1998)

5.1.1 Fundamental Frequency (F0)

This thesis essentially aims to detect two broad categories of irregular phonation — the first type shows distinct irregularities in time or amplitude and is characterized by a lack of structure in the waveform while the other type has abnormal spacing between adjacent glottal pulses relative to the glottal pulse spacing in the local environment. Both these descriptions differ from the quasi-periodic structure of the waveform for regular phonation. This distinction suggests that fundamental frequency could be a valuable cue in separating regular phonation from irregular phonation. Table 5.1 lists the ideal behavior for a F0 estimator to classify regular phonation from irregular phonation showing the expected F0 ranges for the two types of irregular phonation as well as gender-based, expected F0 ranges for regular phonation.

The absence of a robust F0 estimator which applies to both regular and irregular phonation is a roadblock in using this cue. Most estimators are specifically designed to compute F0 estimates for examples of regular phonation. This thesis uses an F0 estimator, based on the filtered-error-signal-autocorrelation sequence to minimize formant interaction, which can provide a reasonable level of separation in the F0 estimates for both regular and irregular phonation (see Table 5.2). A detailed overview of this method is available in Markel & Gray (1976), but the algorithm is briefly outlined here.

In order to compute the F0 estimate, the speech segment is first low-pass filtered using a 12^{th}-order Chebyshev filter with the stop-band ripple 30 dB down and the stopband edge frequency at 1000 Hz. The segment is then pre-emphasized using a 500 Hz single-pole high-pass filter which boosts amplitudes at higher frequencies. This step counteracts the net decrease in amplitude of -6 dB/octave at higher frequencies (resulting from the sum of a -12 dB/octave decrease in amplitude from the voicing source and a +6 db/octave rise due to the radiation characteristics) during speech production. After processing the resulting segment through a Hamming window of equal length, the autocorrelation sequence for the segment is found. The Levinson-Durbin recursion algorithm is used to find a set of coefficients that model the vocal tract as an all-pole filter using what is commonly referred to as the "autocorrelation method". The coefficients from the Levinson-Durbin algorithm model the vocal tract as a transfer function in the form,

$$H(z) = \left(\frac{1}{1 - \sum_{n=1}^{12} a_n \times z^{-n}} \right)$$

where a represents the coefficients from the Levinson-Durbin algorithm.

The original segment is filtered using the coefficients from the Levinson-Durbin algorithm to yield the error signal, which is an indicator of the glottal activity at the source. The autocorrelation sequence for the error signal forms the basis for the F0 computation. The autocorrelation sequence is first normalized by the peak amplitude at zero lag. Subsequently, all peaks greater than 0.46 over a range from 2.5 ms to half the length of the autocorrelation sequence are selected. The choice of 0.46 as a threshold value is not mentioned in the Markel & Gray algorithm and was selected based on analysis documented in Table 5.2. Specifically, choosing 0.46 as a threshold value results in reasonable F0 estimates for a majority of the regular and irregular tokens.

The choice of a particular peak's index provides an estimate for the fundamental period of the segment. The F0 estimate is calculated by taking the inverse of the fundamental period. The steps involved in choosing the correct peak index have been itemized below:

- If no peaks > 0.46, then the F0 estimate is 0.
- If only one peak is > 0.46, then the associated index is estimated as the fundamental period.

Table 5.2: Number of F0 estimates below 72 Hz for regular and irregular tokens using different threshold values for the peak-detector in the F0 estimator. Ideally, a majority of the irregular tokens, but very few regular tokens, should have F0 values less than 72 Hz.

Threshold	**No. of regular tokens < 72 Hz (out of 8055 tokens)**	**No. of irregular tokens < 72 Hz (out of 1279 tokens)**
0.40	584	810
0.41	652	840
0.42	736	873
0.43	838	893
0.44	925	914
0.45	1020	936
0.46	**1157**	**951**
0.47	1264	977
0.48	1394	1005
0.49	1514	1022
0.50	1668	1047

- If more than one peak is > 0.46, then a test is conducted to determine if all the peak indices are proportional to each other within a threshold of 0.02. If the peaks indices do meet this criteria, then the second peak index is estimated as the fundamental period. The first peak is ignored since its choice leads to halving of the actual F0 value.

- if all the above-mentioned criteria fail, the maximum peak above the threshold value is selected and its index determined as the fundamental period.

Figure 5-1 illustrates the F0 computation on a regular and an irregular token respectively.

5.1.2 Normalized Root Mean Square Amplitude

Most of the examples of irregular phonation encountered during labeling match descriptions of vocal fry. The number of glottal pulses per unit time for vocal fry is less than the number for regular phonation due to the abnormal spacing of glottal pulses. Other types of irregular phonation show a similar behavior where irregulari-

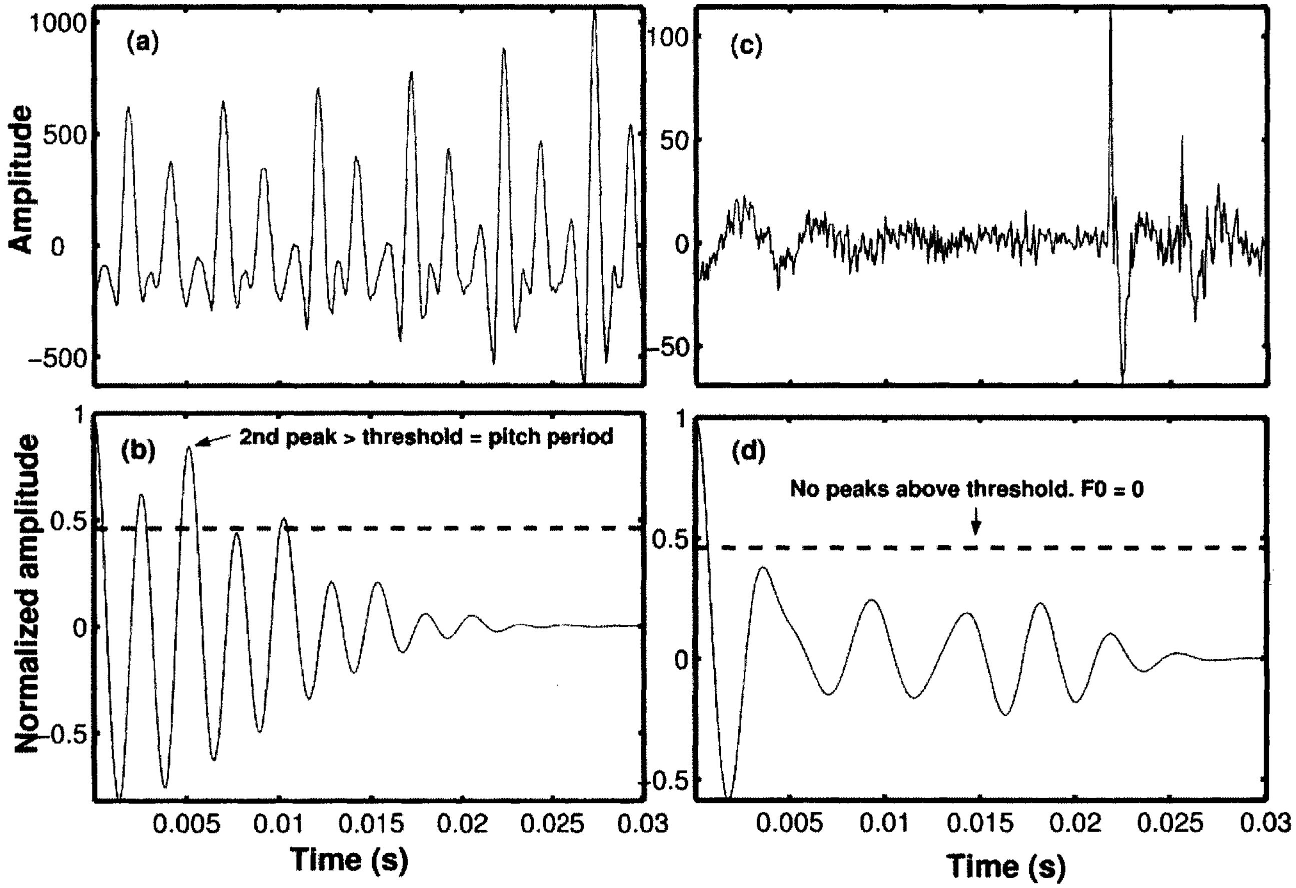

Figure 5-1: **(a)** Example of a regular token. **(b)** The autocorrelation function for (a). **(c)** Example of an irregular token. **(d)** The autocorrelation function for (c). The horizontal line indicates the threshold value of 0.46 used in the F0 computation. **(b)** has multiple peaks greater than 0.46; the fundamental period is correctly chosen by the second peak greater than 0.46. In contrast, **(d)** has no peaks greater than 0.46 and the F0 estimate equals the default value of 0.

ties in the spacing between glottal pulses lead to a lower number of glottal pulses per unit time compared to regular phonation. This observation suggests that the average signal amplitude estimated over a fixed time window should be greater for a regular segment than for an irregular segment. Figure 5-2 illustrates this hypothesis using an example of a regular and an irregular token.

Root mean square (RMS) amplitude is a common tool used in signal processing to estimate the average amplitude of a signal. The result for the RMS amplitude of the token is normalized by the RMS amplitude of the entire speech signal from which the regular or irregular token is extracted to account for inter-speaker variation in signal amplitude. The assumption using this method of normalization is that the speaker uses the same "speaking level" over the course of the utterance. The mathematical formulation to compute this cue is,

$$A_{RMS} = \frac{(\frac{1}{L}\sum_{n=1}^{L-1} s[n]^2)^{0.5}}{(\frac{1}{N}\sum_{n=1}^{N-1} S[n]^2)^{0.5}}$$

where s[n] is the regular or irregular token, S[n] is the entire speech signal or utterance in the case of the TIMIT database, N is the length of the entire speech signal in samples and L is 30 ms of the regular or irregular token in samples.

5.1.3 Smoothed-energy-difference amplitude

Most examples of irregular phonation in the data-set either match descriptions of vocal fry with widely spaced glottal pulses or show abruptness in the time-domain waveform. This abruptness can be manifested in the form of an "impulse-like" triangular pulse, a sudden change in amplitude of a glottal pulse or the appearance of an additional glottal pulse within the normal glottal cycle. It is hypothesized that all these behaviors should be characterized by a rapid transition of energy within the irregular segment. Regular phonation, on the other hand, will not generally show such rapid variations in energy. In order to test this hypothesis quantitatively, the smoothed-energy-difference amplitude cue was formulated.

First, the 512-point Fast Fourier transform (FFT) for the token is computed. A

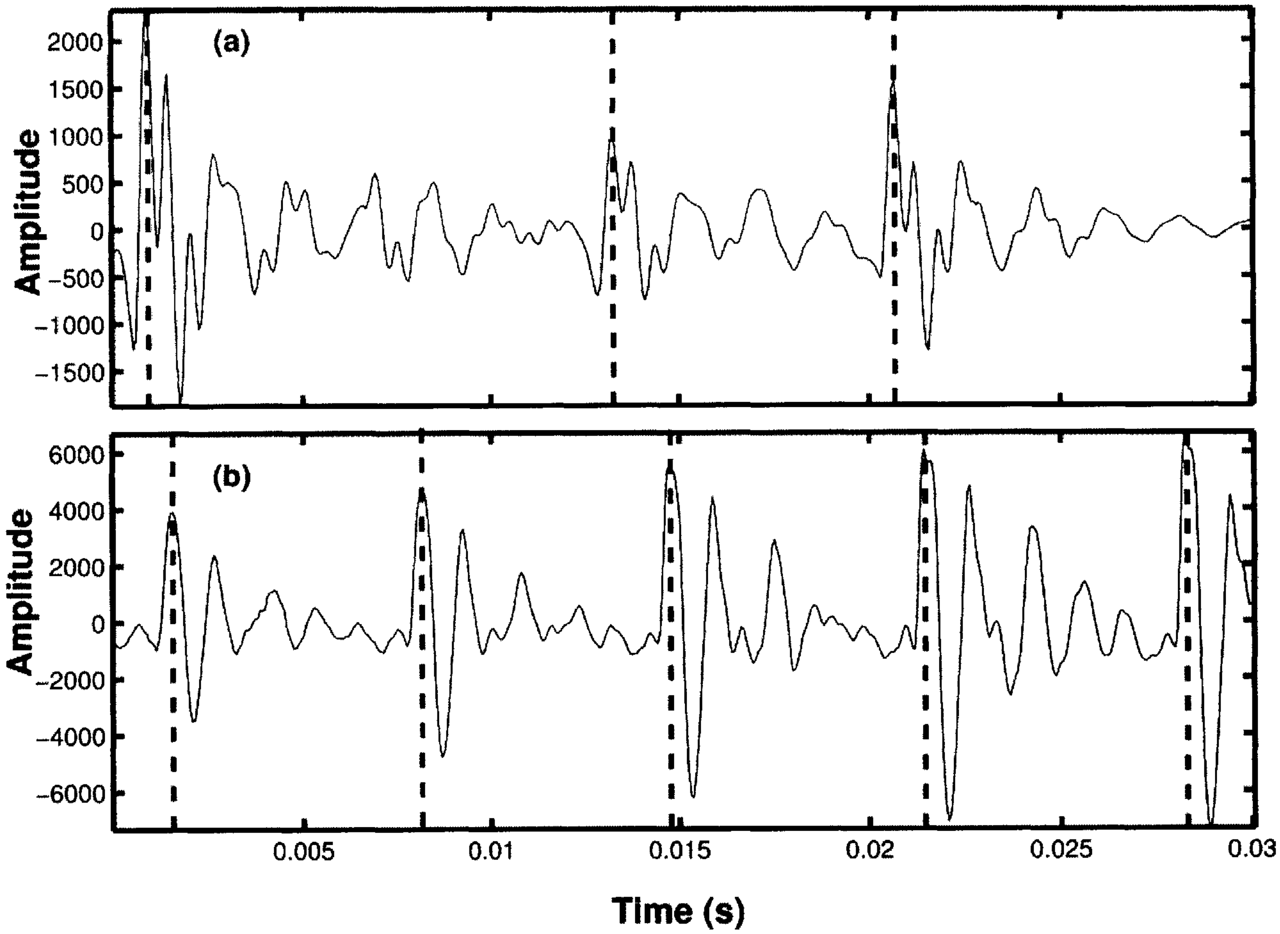

Figure 5-2: (a) Example of an irregular token. (b) Example of a regular token. Both are taken from the same speaker and are of the same duration to avoid inter-speaker variablity in signal amplitude. The dashed vertical lines indicate the glottal pulses in the token. (b) has five glottal pulses, compared to three for (a) and hence a higher average signal amplitude.

Hamming window size of 16 ms was chosen to window the token while calculating the FFT in the form,

$$X[k] = \sum_{n=1}^{512} x[n]w[n]e^{-jw_0 n}$$

where x[n] is the input segment, w[n] is the 16 ms Hamming window and X[k] is the FFT of x[n].

Since the FFT is symmetric in frequency, only the first 256 points of the FFT are analyzed. The window is shifted by 1 ms and the FFT is calculated recursively across all the segments in the token. These values are squared eventually giving a matrix of size $(256 \times nFrames)$, where $nFrames$ is the number of frames in the token where the FFT has been computed. This matrix contains the energy information for the token across all frequencies. The energy in each frame is averaged between 300 Hz - 1500 Hz and the $10 * log_{10}$ of the values are used to compute a matrix of size $(1 \times nFrames)$ giving one averaged energy value per frame for the token.

The choice of lower frequencies is valid since most of the energy in vowels, which are used as examples of regular phonation, is concentrated in this range of frequencies with the first formant rarely dipping below 300 Hz. The upper limit of 1500 Hz was chosen since it resulted in the best separation in the smoothed-energy-difference values for regular and irregular tokens as documented in Table 5.3.

The energy values found previously were averaged in time using different smoothing window sizes — the initial choices being 10 ms and 16 ms respectively. The choice of (10 ms, 16 ms) was based on the rationale that both these window sizes would include at least one glottal pulse in the time-domain waveform for regular phonation resulting in a small difference between the two smoothed-energy waveforms. However, in many cases of irregular phonation with widely-spaced glottal pulses, a 10 ms window size would not encompass one glottal pulse resulting in a larger difference in energy between the pair of smoothed-energy waveforms.

The difference between the smoothed averaged energy values using the two window sizes, called the smoothed-energy-difference waveform, helps separate abrupt variations in energy from smoothly-varying variations in energy. Since the energy in

Table 5.3: Number of smoothed-energy-difference estimates below 2 for regular and irregular tokens using different higher frequency bands. Ideally, a majority of the regular tokens, but very few irregular tokens, should have smoothed-energy-difference values less than 2.

Upper frequency (Hz)	**No. of regular tokens < 2 (out of 8055 tokens)**	**No. of irregular tokens < 2 (out of 1279 tokens)**
900	5881	147
950	5887	150
1000	5886	149
1050	5889	149
1100	5890	149
1150	5895	147
1200	5905	148
1250	5911	147
1300	5910	145
1350	5913	145
1400	5914	144
1450	5916	144
1500	**5921**	**143**

regular phonation is smoothly varying, few peaks are expected in its smoothed-energy-difference waveform. On the other hand, the smoothed-energy-difference waveform should show a more jagged structure for irregular phonation.

Inadvertent peaks might be produced at the beginning and end of the smoothed-energy-difference waveform due to filtering artifacts when averaging by different window lengths. In order to avoid these erroneous peaks, $max(smoothing\_window\_size)/2 + 1$ samples from the beginning and end of the waveform are excluded from analysis. Figures 5-3 and 5-4 show typical smoothed-energy-difference waveforms for regular and irregular phonation respectively. Taking advantage of the presence of jagged peaks in the smoothed-energy-difference waveform for irregular tokens and their absence in regular tokens, the smoothed-energy-difference cue is the largest peak in the smoothed-energy-difference waveform.

The lower window size was decreased to 8 ms and 6 ms respectively keeping the upper smoothing window size fixed at 16 ms, as shown in Table 5.4, to see if this change resulted in a greater separation between regular and irregular tokens. A trade-

Table 5.4: Number of smoothed-energy-difference estimates below 2 for regular and irregular tokens using different lower smoothing window sizes kepping the upper smoothing window size at 16 ms. Ideally, a majority of the regular tokens, but very few irregular tokens, should have smoothed-energy-difference values less than 2.

Lower smoothing window size (ms)	**No. of regular tokens < 2 (out of 8055 tokens)**	**No. of irregular tokens < 2 (out of 1279 tokens)**
10	7293	389
8	6619	222
6	**5921**	**143**

off is observed with a decrease in the smoothing window size not only reducing the number of irregular tokens, but also the number of regular tokens, with small-valued peaks. Ideally, a majority of the regular tokens but very few irregular tokens, should have small smoothed-energy-difference values. In order to weigh this cue towards the correct identification of irregular tokens — since it is hypothesized that the F0 and the normalized RMS cues should be robust in identifying regular phonation — a choice was made to accept this trade-off and the lower window size is chosen to be 6 ms.

5.1.4 Shift-difference amplitude

The method to compute this cue, referred to as shift-difference amplitude in this thesis, is largely based on work done by Kochanski, Grabe, Coleman & Rosner (2005) with minor modifications. It is a measure of aperiodicity and Kochanski *et. al.* (2005) used it to detect prominence in speech. In the context of this thesis, shift-difference amplitude is used to distinguish regular and irregular phonation since irregular phonation can often be characterized by a lack of periodicity in contrast to regular phonation.

The computation is based on the difference between adjacent segments of the time-domain waveform. The method assigns values close to 0 to regions of perfect periodicity and values in the vicinity of 1 for aperiodic segments.

To compute this cue, the audio signal has low frequency noise and DC offsets

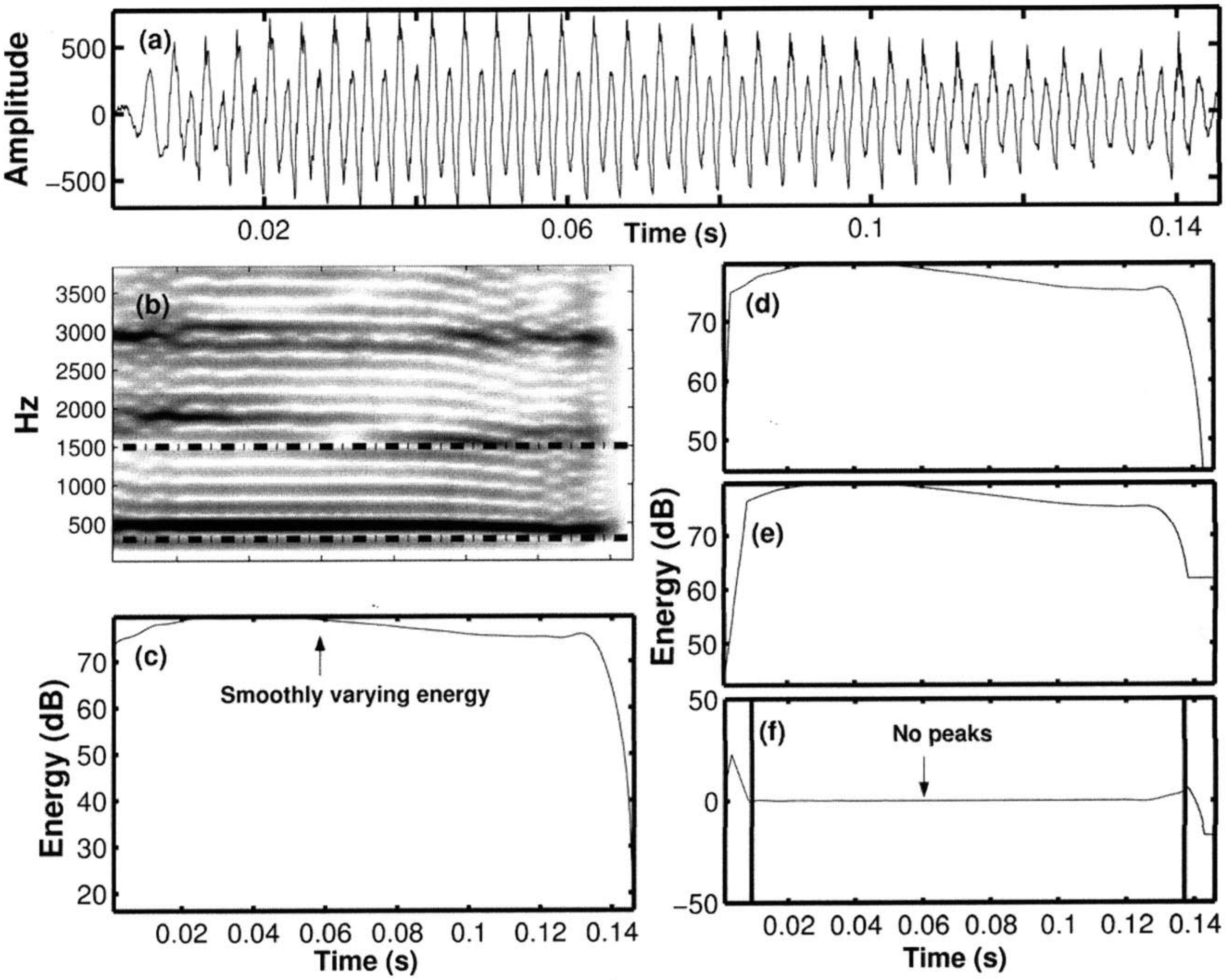

Figure 5-3: **(a)** A typical regular token. **(b)** The spectrogram for the token showing the energy in the signal at lower frequencies. The horizontal dashed lines show the limits of 300 Hz and 1500 Hz over which the energy is averaged per frame. **(c)** The averaged energy waveform. **(d)** The averaged energy waveform after being smoothed in time using a window size of 6 ms. **(e)** The averaged energy waveform after being smoothed in time using a window size of 16 ms. **(f)** The difference between the two smoothed waveforms. The regions to the left and right of the vertical lines are left out from analysis due to artificial peaks created during the averaging process.

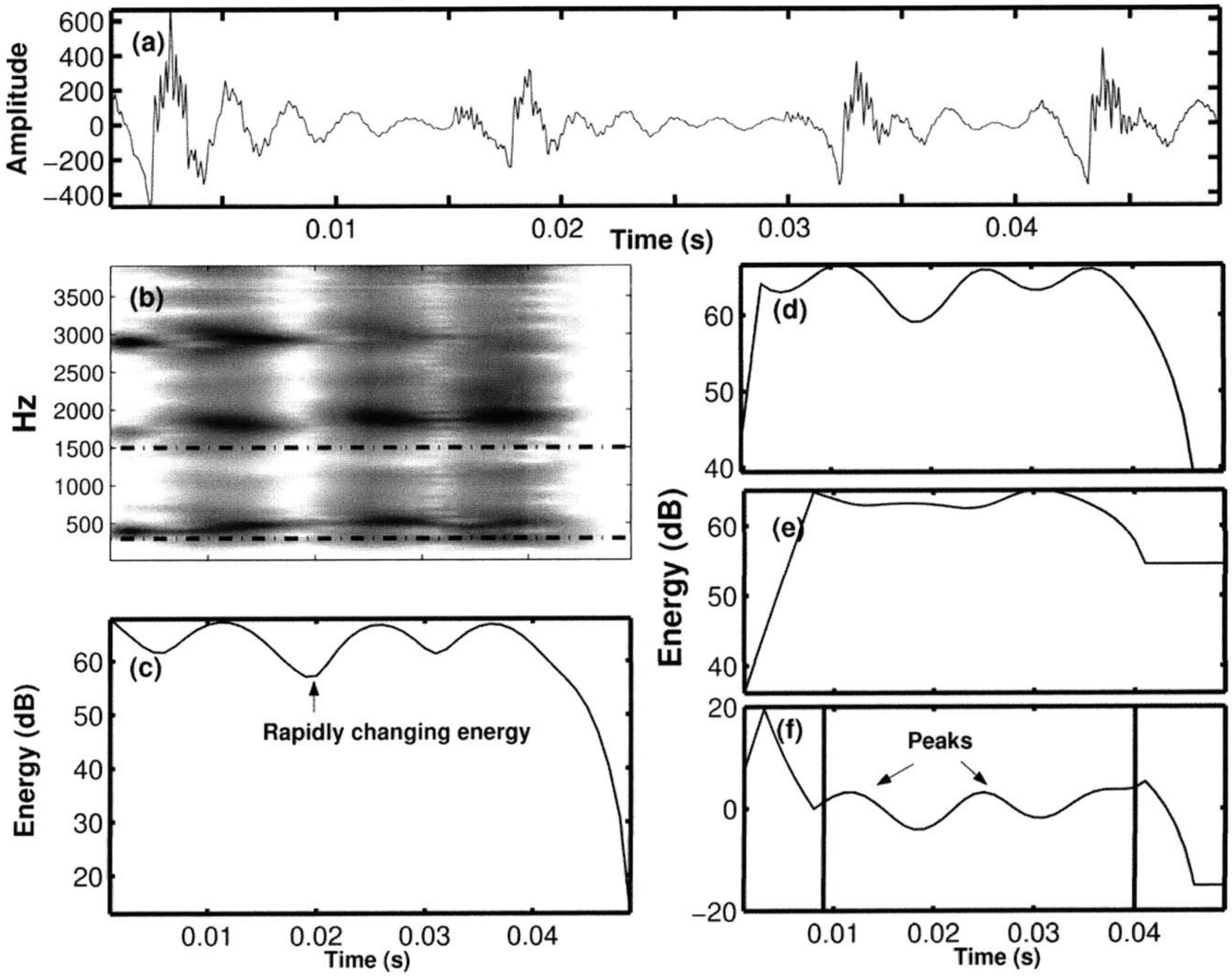

Figure 5-4: **(a)**A typical irregular token. See figure 5-3 for an explanation on how to interpret panes **(b)**, **(c)**, **(d)**, **(e)** and **(f)**.

removed with a 50 Hz 4^{th}-order time-symmetric Butterworth high-pass filter and is then passed through a 500 Hz single-pole high pass filter for pre-emphasis. The aperiodicity measure is calculated by taking 10 ms of the middle section of the token, windowing it by a Gaussian with 20 ms standard deviation, and comparing it to other sections shifted by 2 ms to 10 ms in increments of the sampling rate. If the segment is periodic, then one of the shifted windows will match the original window resulting in a minimum difference. The value of this cue is proportional to the value of the difference between the shifted windows leading to the term "shift-difference" method.

For each possible shift, between 2 ms and 10 ms to the left and right,

$$d_\delta[n] = (s[n+\delta/2] - s[n-\delta/2])^2$$

is computed where s[n] is the middle section of the filtered segment at time n. The middle section is multiplied to itself to give $P[n] = s[n]^2$. This value is a measure of the energy in the original filtered segment. Both $d_\delta[n]$ and P[n] are convolved with 20 ms standard deviation Guassians to yield $\tilde{d}[n]$ and $\tilde{P}[n]$.

$$\hat{d}[n] = min_\delta\{\tilde{d}_\delta[n]\}$$

is the minimum difference over all the shifts δ. In order to normalize the output, the shift-difference amplitude cue is

$$(\hat{d}[n]/\tilde{P}[n])^{0.5}$$

.

The steps above have been originally outlined by Kochanski *et al.* (2005). The only change made in this implementation is in the extent of the shifts which are between 2 ms and 10 ms, instead of between 2 ms and 20 ms. This change is due to the classification of abnormally wide-spaced glottal pulses as irregular in this thesis, in spite of being periodic. If shifts as high as 20 ms were to be allowed, the shift-difference amplitude for these specific instances of irregularity would result in estimates more

Table 5.5: Expected behavior of the cues for regular and irregular segments.

	F0	**Normalized RMS**	**Smoothed-energy-diff.**	**Shift-diff.**
Regular	Higher	Higher	Lower	Lower
Irregular	Lower	Lower	Higher	Higher

consistent for regular tokens.

Table 5.5 is a summary of the expected contrast in the values of these cues for regular and irregular tokens.

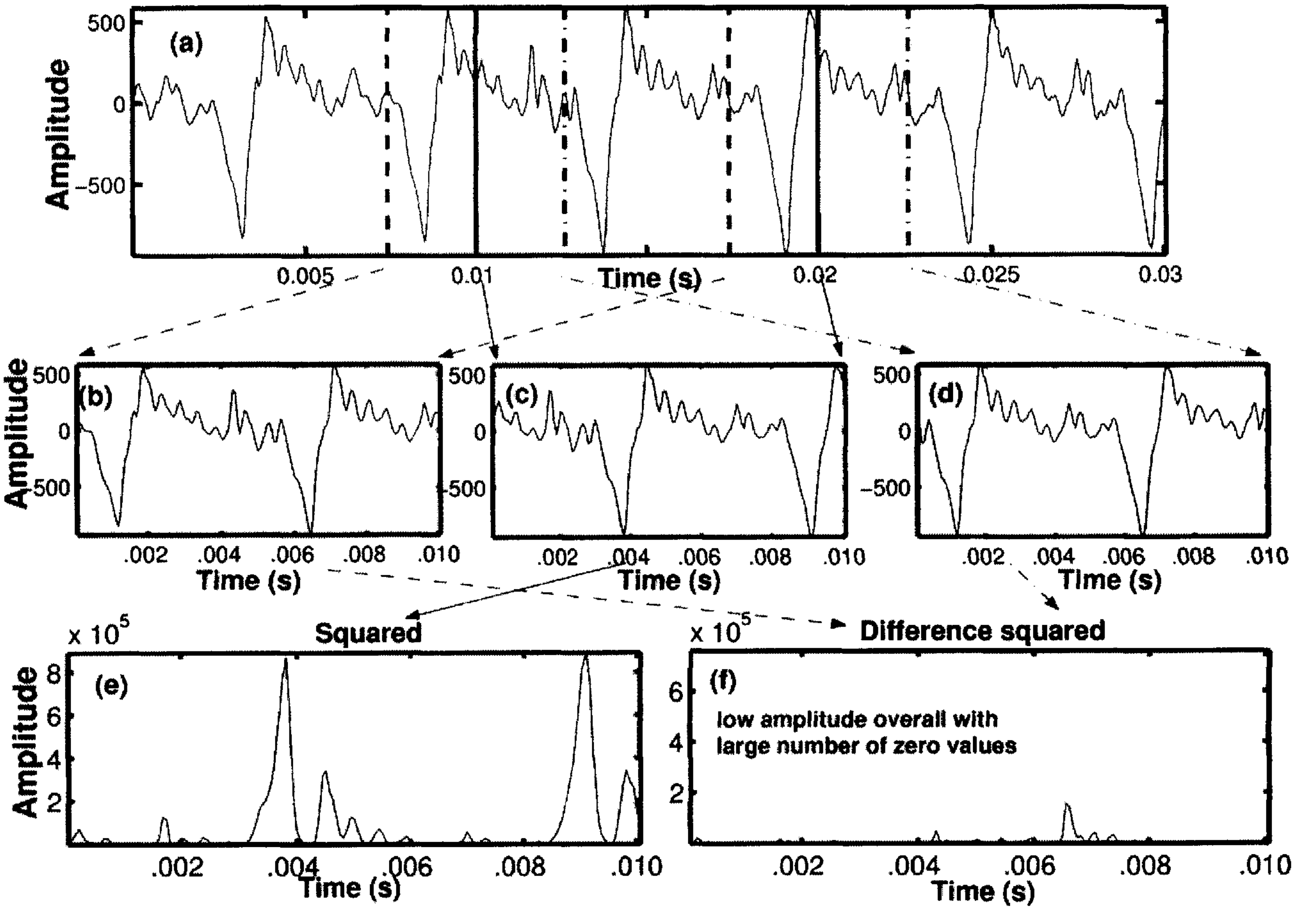

Figure 5-5: (a) A typical regular token after high-pass filtering and pre-emphasis. The solid line shows the middle segment of the token, while the dashed line marks the shifted segments after shifting the window by 2.6 ms in either direction. (b) The left-shifted segment. (c) The middle segment. (d) The right-shifted segment. (e) The squared middle segment. (f) The squared difference between (b) and (d). The output shift-difference magnitude is the $\sqrt{(f)/(e)}$ which for this particular time-shift is 0.09, consistent with expectations for regular tokens.

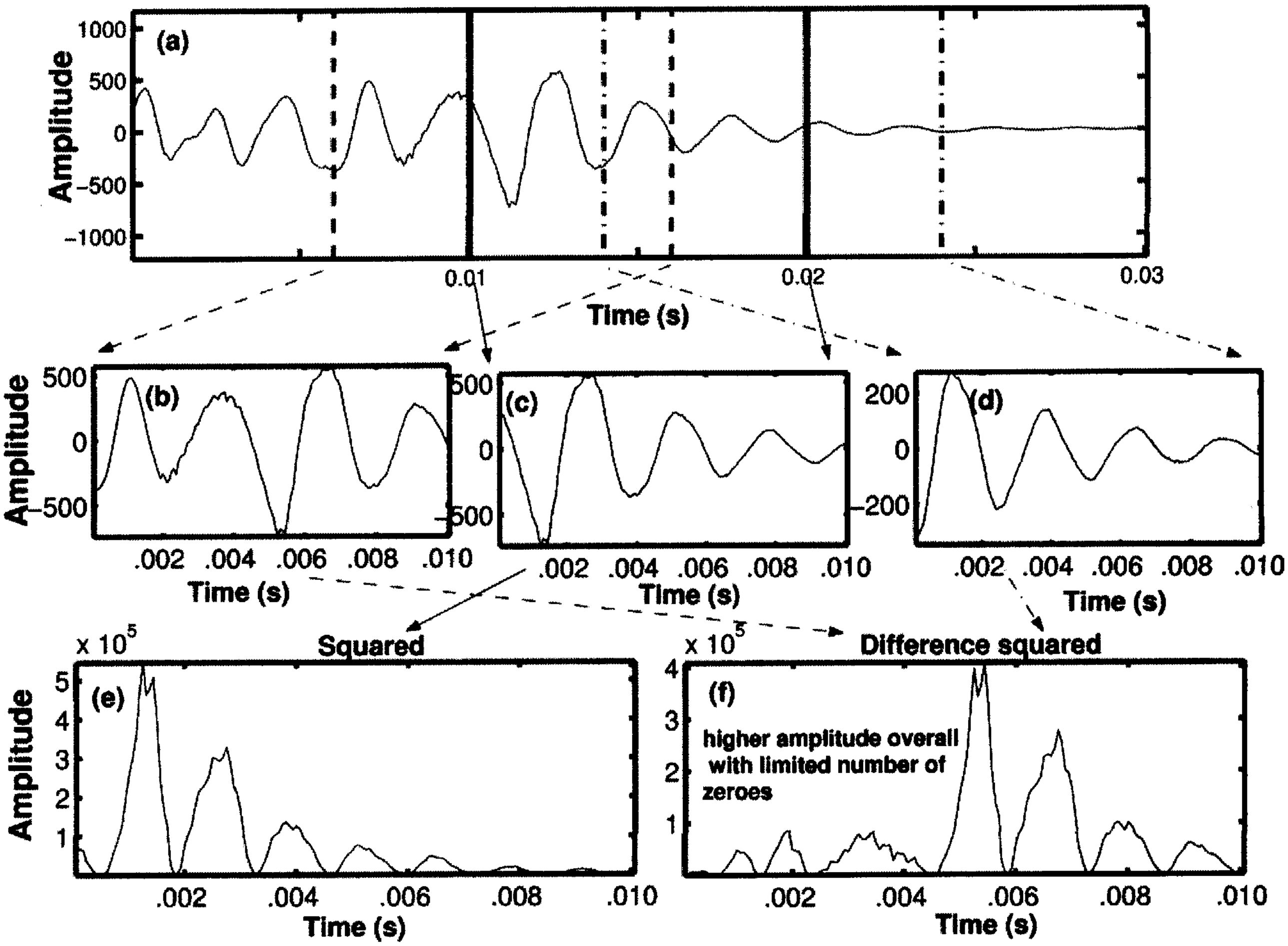

Figure 5-6: **(a)** A typical irregular token after high-pass filtering and pre-emphasis. The solid line shows the middle segment of the token, while the dashed line marks the shifted segments after shifting the window by 4 ms in either direction. **(b)** The left-shifted segment. **(c)** The middle segment. **(d)** The right-shifted segment. **(e)** The squared middle segment. **(f)** The squared difference between (b) and (d). The output shift-difference magnitude is the $\sqrt{(f)/(e)}$ which for this particular time-shift is > 1, consistent with expectations for irregular tokens.

Chapter 6

Analysis

This chapter analyzes the distribution patterns of fundamental frequency, normalized root-mean-square amplitude, shift-difference amplitude and smoothed-energy-difference amplitude for all the regular and irregular tokens from the dataset. In addition to displaying the distribution patterns, a two-sample t-test with the null hypothesis that the means of the acoustic cues are equal for regular and irrregular phonation is used to test the significance of their separation. Finally, a failure analysis is conducted to understand the behavior of the tokens where the acoustic cues fail to separate regular and irregular phonation.

6.1 Overview

This thesis proposes a token-based recognition scheme, in favor of a frame-based recognition scheme, to classify regular phonation from irregular phonation.

All the labeled tokens, for both regular and irregular phonation, were decomposed into smaller units of 30 ms to compute fundamental frequency, normalized RMS amplitude and shift-difference amplitude. In the case of fundamental frequency and shift-difference amplitude, the minimum of these values was selected as the output cue value for the token. However, in the case of normalized RMS amplitude, the mean of the values was taken as the output cue value for the token.

The decomposition of the tokens into 30 ms segments was necessary for the above-

Table 6.1: Two-sample t-test on the four acoustic cues with the null hypothesis that the means are equal (df 9332).

	t-statistic	**p-value**
F0	32.16	$<< 0.001$
Normalized RMS	36.59	$<< 0.001$
Smoothed-energy-difference	-61.92	$<< 0.001$
Shift-difference	-74.99	$<< 0.001$

mentioned cues to avoid misleading cue values. The reasons are cue-specific and are expanded below:

- **Fundamental frequency (F0) & Shift-difference amplitude** - the decomposition prevented the normal variations in fundamental frequency and periodicity from biasing the fundamental frequency and shift-difference amplitude estimates.

- **Normalized RMS amplitude** - the decomposition prevented unequal token lengths from affecting the amplitude.

The smoothed-energy-difference amplitude was not explicitly calculated over smaller 30 ms segments because its manner of computation intrinsically decomposes the token into smaller segments.

6.2 Distribution pattern

Figure 6-1 shows the distribution of the four acoustic cues for regular and irregular tokens. The significance of the separation between regular and irregular phonation for each acoustic cue was quantitatively tested using a two-sample t-test, the results of which are shown in Table 6.1.

The t-statistics for all four acoustic values are large in magnitude with small p-values (df 9332) indicating that a significant separation exists between the means of the acoustic cues for regular and irregular phonation.

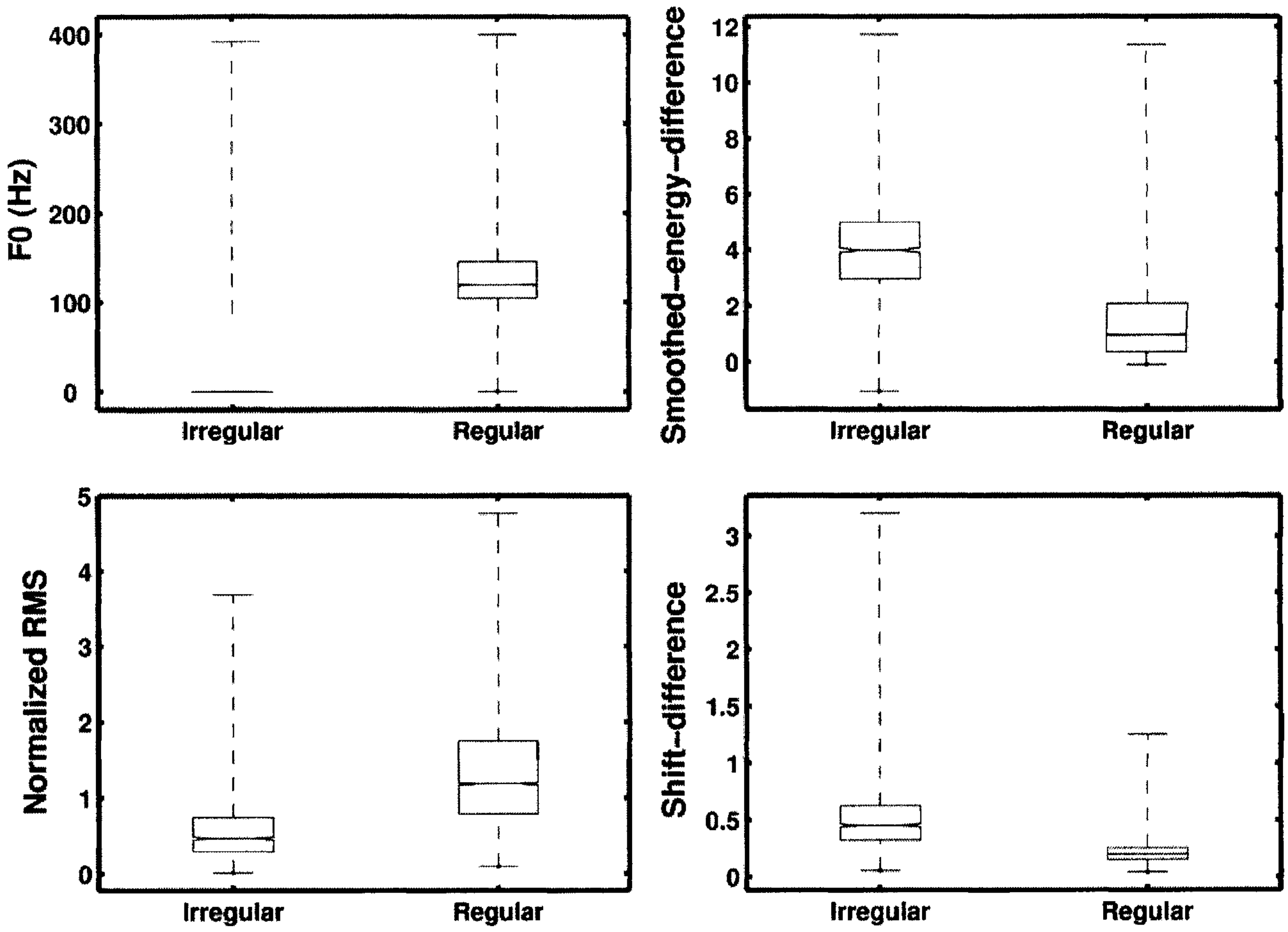

Figure 6-1: Distribution of the acoustic cues for regular and irregular tokens.

6.3 Failure analysis for each cue

According to Figure 6-1, the values of each acoustic cue for irregular phonation are sometimes more consistent with those for regular phonation and vice-versa. A detailed analysis of the tokens from which these acoustic cue values are computed is conducted in this section for each cue separately. However, only two representative examples of regular and irregular phonation respectively are discussed and presented in this section for each cue.

6.3.1 Fundamental Frequency (F0)

Irregular tokens

The F0 estimates for irregular phonation are expected to be either very low or 0 (indicating that the signal has no periodic structure). However, as shown in Figure 6-1, some F0 estimates for irregular phonation are as high as 400 Hz. The tokens associated with these unexpectedly high F0 values were analyzed, showing that most of the irregular tokens which result in high F0 values match descriptions of vocal fry with widely spaced glottal pulses as shown in Figure 6-2.

This observation suggests that the F0 algorithm only distinguishes between regular phonation and examples of irregular phonation showing a lack of structure in the waveform. These types of irregular phonation are characterized by a lack of periodicity and will often have no peaks greater than the threshold resulting in a F0 estimate with value 0.

The algorithm fails to correctly estimate the fundamental frequency for tokens that match the description of vocal fry with a F0 $<$ 72 Hz (Blomgren *et. al.*, 1998). The reason for this failure is that the F0 computation chooses either the first peak (when only one peak is $>$ *threshold*) or the second peak (when multiple peaks are $>$ *threshold* and proportionally aligned) in the autocorrelation function to estimate the fundamental period. The choice of either of these peaks is based on the expected F0 ranges in regular phonation. It is not valid as a fundamental frequency estimate for widely spaced glottal pulses which should have a fundamental period much larger than indicated by the first or second peaks.

It is hypothesized that the misleadingly high F0 estimates for vocal fry will be offset by other cues such as normalized RMS amplitude and shift-difference amplitude to distinguish it from regular phonation. Since there are a fewer number of glottal pulses per unit time in vocal fry, its normalized RMS amplitude will be lower than examples of regular phonation. The shift-difference amplitude is expected to be large, since the glottal pulses are spaced far apart, in contrast to regular phonation. For the two examples in Figure 6-2, the F0 estimates are **(356 Hz, 340 Hz)** respectively. However, these misleadingly high F0 estimates are offset by the normalized RMS amplitude values, which are **(0.18, 0.44)** respectively, and the shift-difference-amplitude values, which are **(0.83, 0.52)** respectively. The values for these cues follow the trend enumerated in Table 5.5. Specifically, the normalized RMS and shift-difference amplitude values for these irregular examples are 1 stdv. below (for normalized RMS amplitude) and 4 stdv. above (for shift-difference amplitude) their respective means for regular tokens.

Regular tokens

In addition to unexpected F0 values for irregular tokens, regular tokens sometimes show low fundamental frequency values outside of expected ranges. On further analysis, it was found that a vast majority of these F0 values are equal to zero, which is the default F0 estimate when the algorithm fails to find a fundamental period.

This failure is correlated to the choice of the threshold value in the F0 computation as shown in Table 5.2. Increasing the threshold value increases the number of irregular tokens with no fundamental frequency estimate, but also does the same for regular tokens. Figure 6-3 shows two example of regular phonation with F0 estimates equal to 0 due to the high threshold value. As in the case of the inappropriate F0 estimates for irregular phonation, it is expected that the remaining cues will prove sufficient in distinguishing these particular regular tokens from irregular tokens. For the two examples shown in Figure 6-3, the normalized RMS amplitude, shift-difference amplitude and smoothed-energy-difference amplitude are **(0.66, 0.30, 0.63)** and **(2.84,**

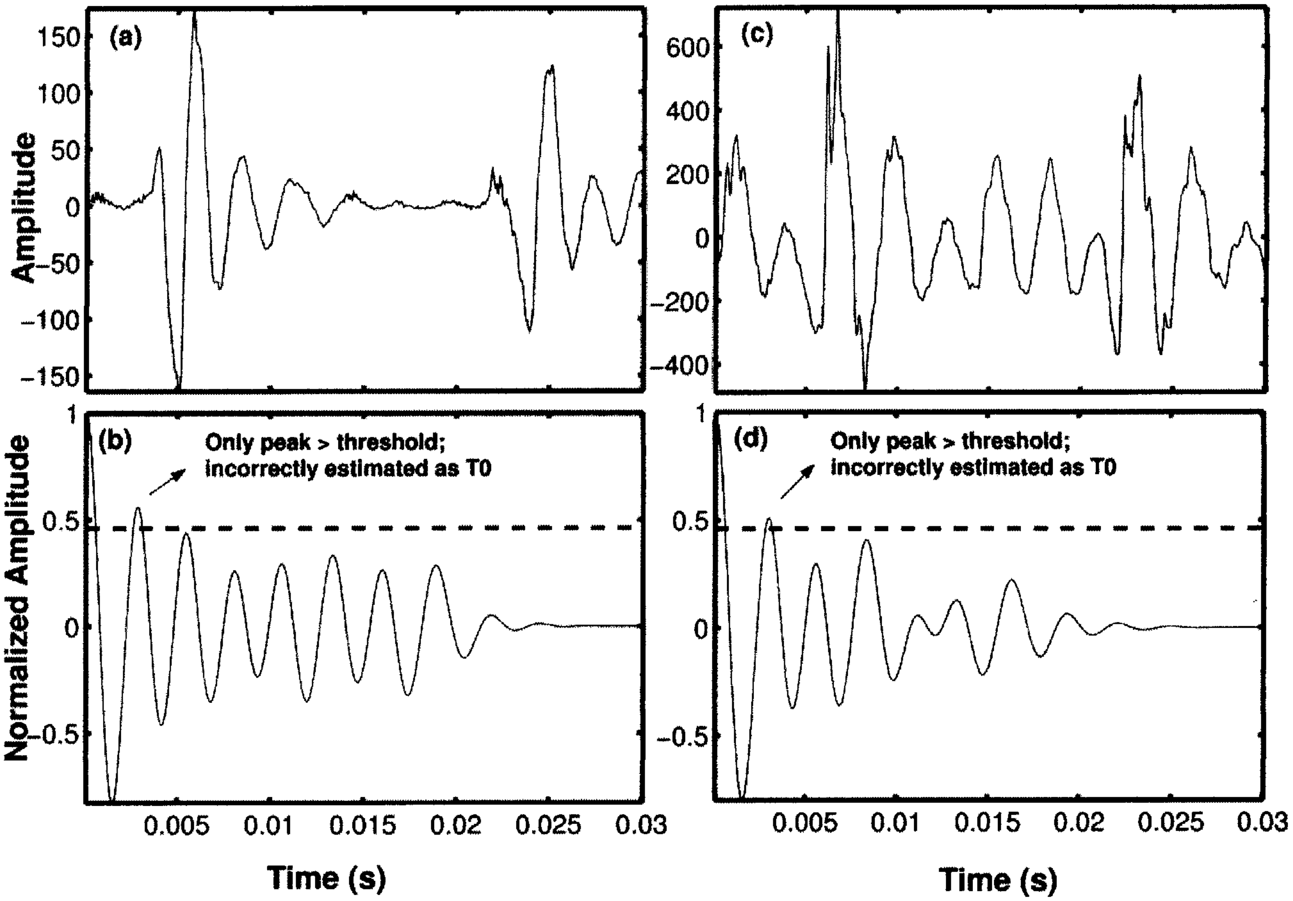

Figure 6-2: **(a)** Example of vocal fry. **(b)** The autocorrelation function for (a). **(c)** Another example of vocal fry. **(d)** The autocorrelation function for (c). The horizontal line indicates the threshold value of 0.46 used in the F0 computation. In both **(b)** and **(d)**, the fundamental period is inappropriately chosen by the only peak greater than 0.46.

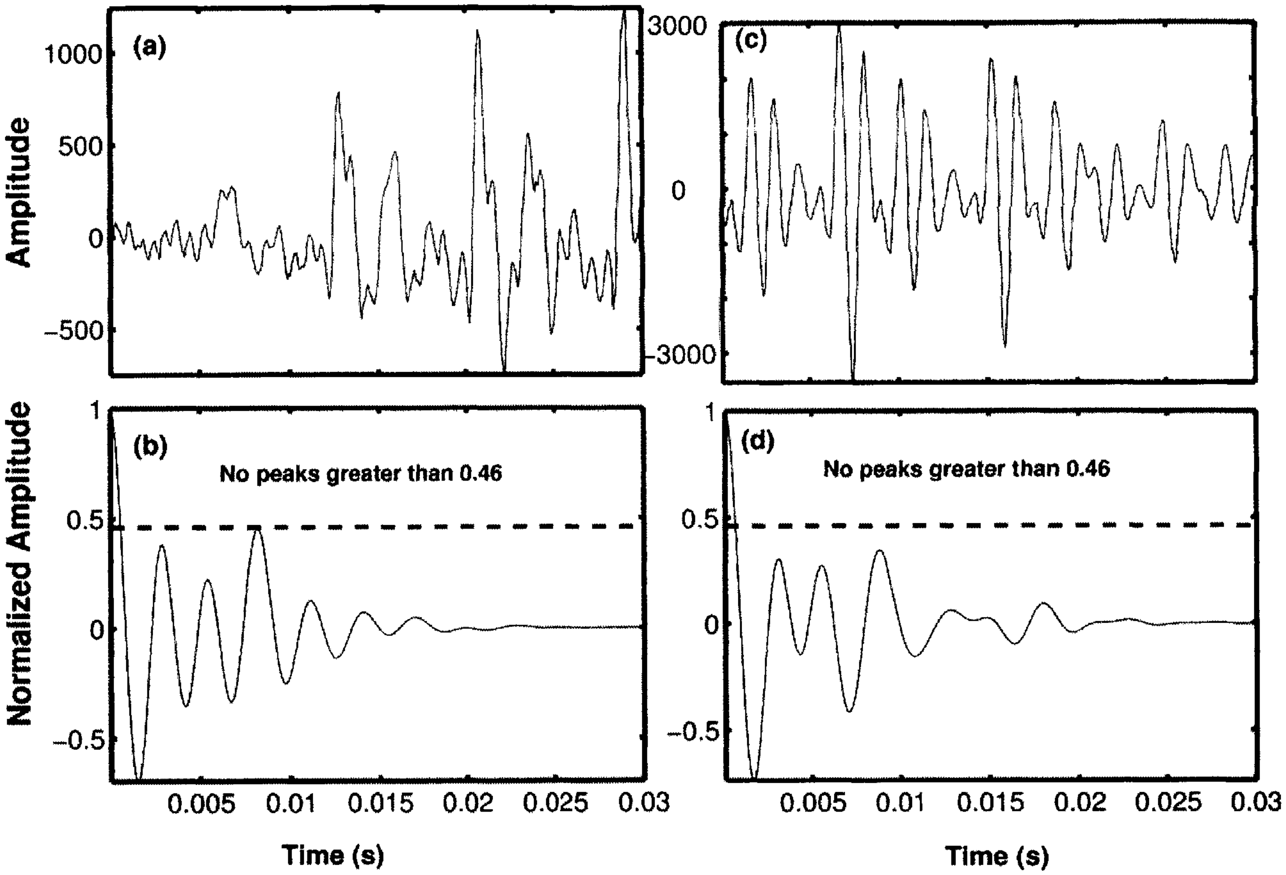

Figure 6-3: **(a)** Example of a regular token. **(b)** The autocorrelation function for (a). **(c)** Another example of a regular token. **(d)** The autocorrelation function for (c). The horizontal line indicates the threshold value of 0.46 used in the F0 computation. In both **(b)** and **(d)**, none of the peaks are greater than 0.46 leading to misleading F0 estimates of 0.

0.22, 2.29) respectively. For the first example, the normalized RMS cue (with value 0.66) is lower than expected and is almost equal to the mean of the normalized RMS value for irregular tokens. However, the shift-difference amplitude and the smoothed-energy-difference amplitude for this example are approximately 1 stdv. below and 2 stdv. below their mean values for irregular tokens respectively. For the second example, the normalized RMS amplitude is 4 stdv. above, the shift-difference amplitude is 2 stdv. below, and the smoothed-energy-diffference amplitude is 1 stdv. below their respective means for irregular tokens.

Additional Comments

While analyzing the fundamental frequency ranges for regular and irregular phonation, it was observed that vocal fry for female speakers was sometimes characterized by a fundamental frequency as high as 100 Hz. Some of these instances were individually analyzed and their pitch periods manually corroborated given the tendency of the F0 algorithm to fail when dealing with instances of vocal fry.

While these F0 values were in contrast to the F0 values of the female speaker during regular phonation and resulted in a perceptual impression of irregularity, the result was surprising given the ranges of vocal fry obtained by earlier studies (McGlone, 1967; McGlone & Shipp, 1971; Blomgren *et. al.*, 1998). Blomgren *et. al* (1998) reported a small difference between the F0 values for vocal fry between males (range of 24 Hz - 77 Hz) and females (range of 24 Hz - 72 Hz). Figure 6-4 shows that females can produce vocal fry with F0 values higher than the range proposed in Blomgren *et. al.* (1998) and more consistent with F0 values for regular, male phonation. While an in-depth analysis of the fundamental frequency ranges for vocal fry is beyond the scope of this thesis, these examples of vocal fry showing higher F0 values for female speakers should be further studied.

6.3.2 Normalized RMS amplitude

Irregular tokens

An analysis of irregular tokens with an inappropriately high normalized RMS amplitude shows that most of them are characterized by a high first formant amplitude as shown in Figure 6-5. The F0, shift-difference amplitude and smoothed-energy-difference amplitude for (a) and (b) in Figure 6-5 are **(0, 0.60, 2.86)** and **(0, 0.30, 2.83)** respectively. These cue values offset the misleading normalized RMS amplitudes **(3.69, 3.23)** for the two irregular examples. For the first example, F0 is 2 stdv. below, shift-difference amplitude is 4 stdv. above and smooothed-energy-difference amplitude is 1 stdv. above their respective means for regular tokens. The separation is similar for all the cues in the second example except for the shift-difference

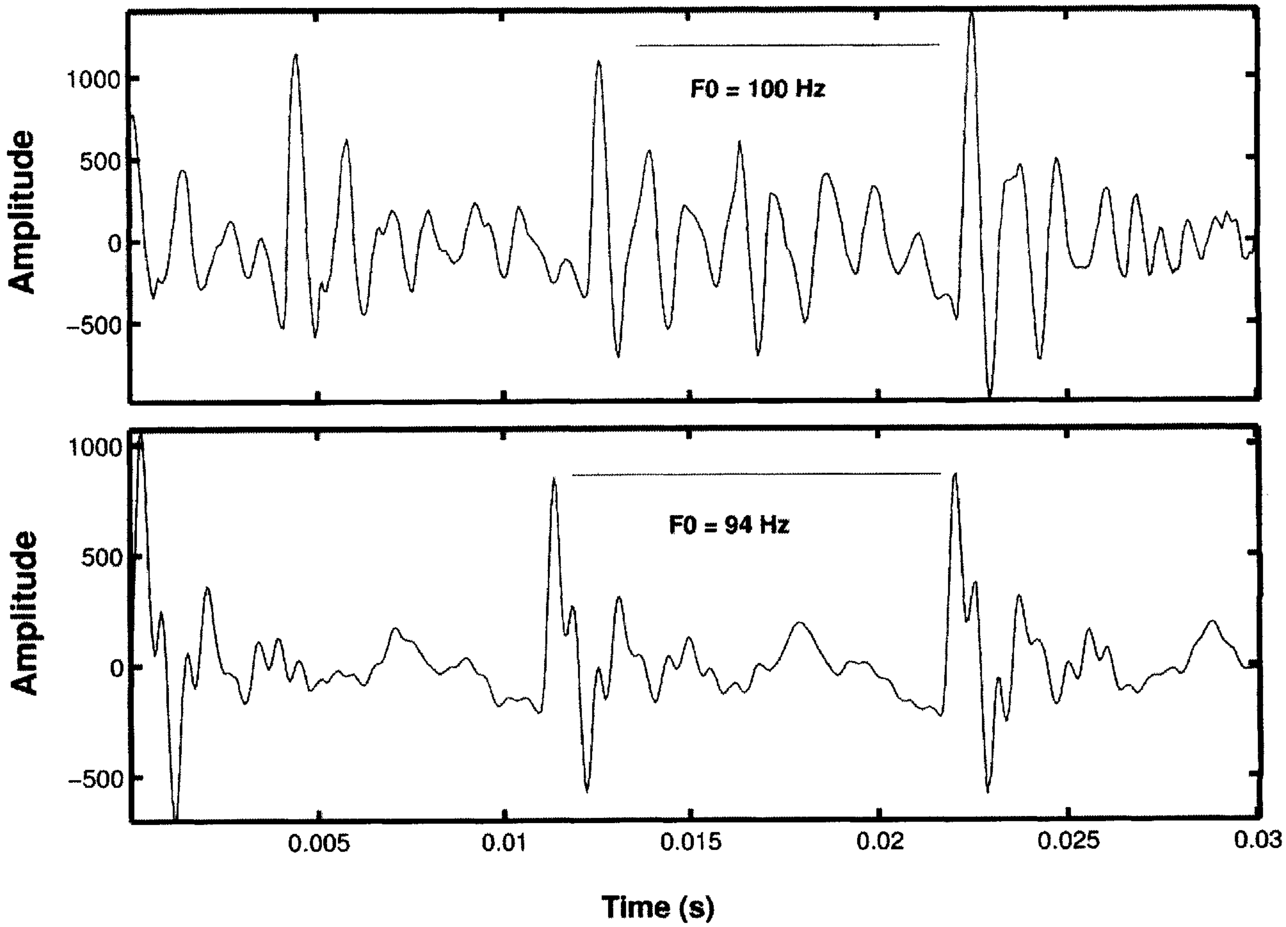

Figure 6-4: Examples of vocal fry with F0 estimates as high as 100 Hz for two different female speakers. (Source of waveform: TIMIT,1990)

amplitude which is 1 stdv. above its mean for regular tokens.

Regular tokens

Some examples of regular tokens have normalized RMS values that are also in lower than expected ranges. Figure 6-5 illustrates two such examples. Most of these regular tokens are characterized by a low amplitude not consistent with the "speaking-level" of the entire speech-waveform. Hence, normalizing the RMS amplitude of the token by the RMS amplitude of the signal does not produce the desired result of increasing the normalized RMS amplitude in these specific cases. In other words, the normalization method accounts for inter-speaker variation, not intra-speaker variation, in amplitude of the time-domain waveform.

The two examples shown in (c) and (d) of Figure 6-5 have F0, shift-difference amplitude and smoothed-energy-difference values of **(96, 0.25, 2.04)** and **(182, 0.37, 1.97)**, offsetting the inappropriately low normalized RMS estimates of **(0.27, 0.55)** respectively. For the first regular example, F0 is 0.5 stdv. above, shift-difference amplitude is 1 stdv. below and smooothed-energy-difference amplitude is 1 stdv. below their respective means for irregular tokens. The separation is similar for the second example, except for F0 which is 1 stdv. above its mean value for irregular tokens.

6.3.3 Smoothed-energy-difference amplitude

Irregular tokens

Inappropriately low smoothed-energy-difference amplitude values are found for some irregular tokens. The unifying characteristic of these examples is that they have only one or two glottal pulses in the token and either match descriptions of vocal fry or glottalization. Figure 6-6 shows an example of vocal fry with only two glottal pulses. The limited number of glottal pulses results in fewer transitions of energy within the speech waveform as seen by the lone undulation in the energy waveform in (c) of Figure 6-6. This behavior is not characteristic of the jagged structure expected in the

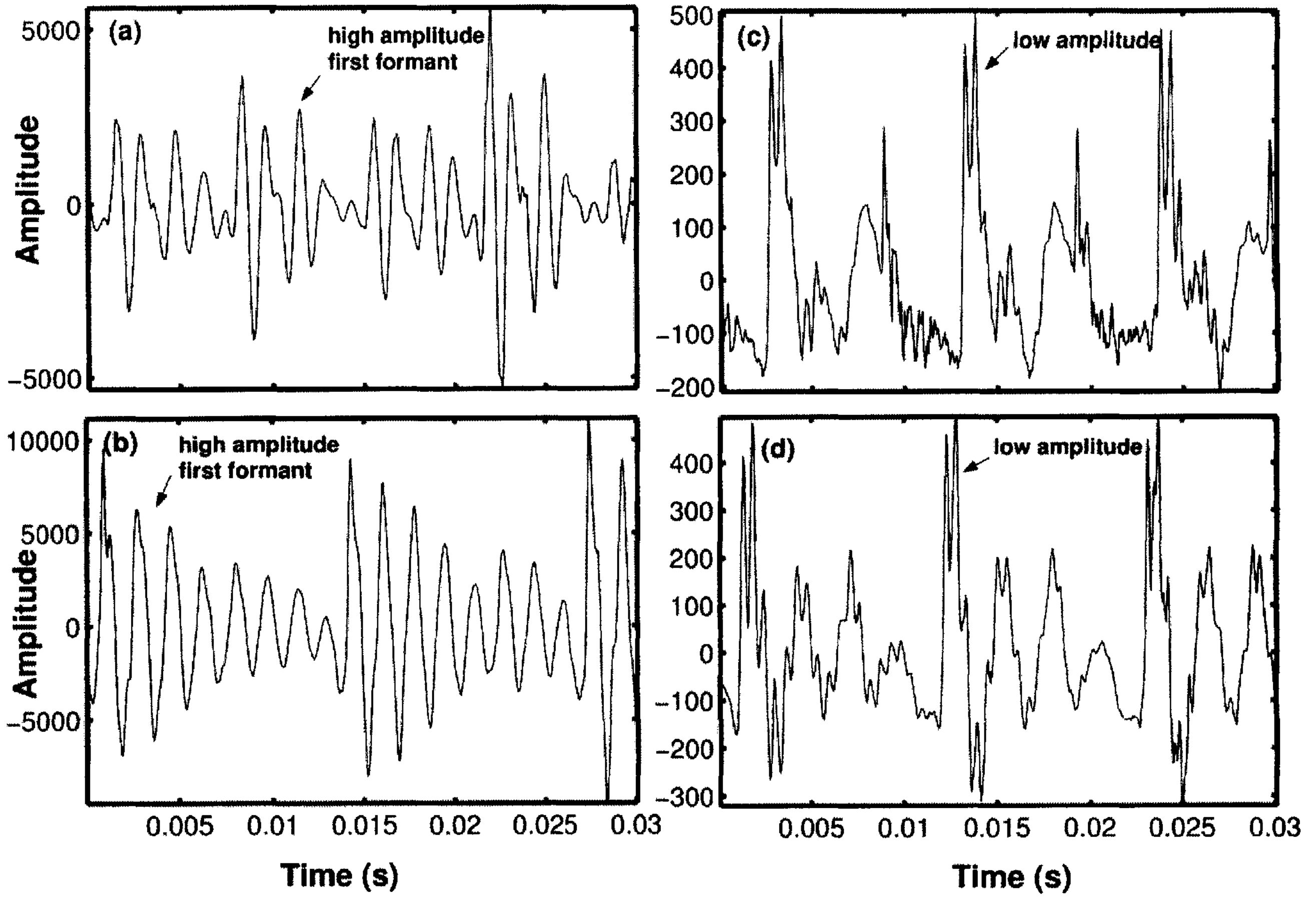

Figure 6-5: Illustration of misleading normalized RMS estimates for regular and irregular phonation. (a) and (b) show examples of irregular tokens with inappropriately high normalized RMS values. (c) and (d) show examples of regular tokens with inappropriately low normalized RMS values.

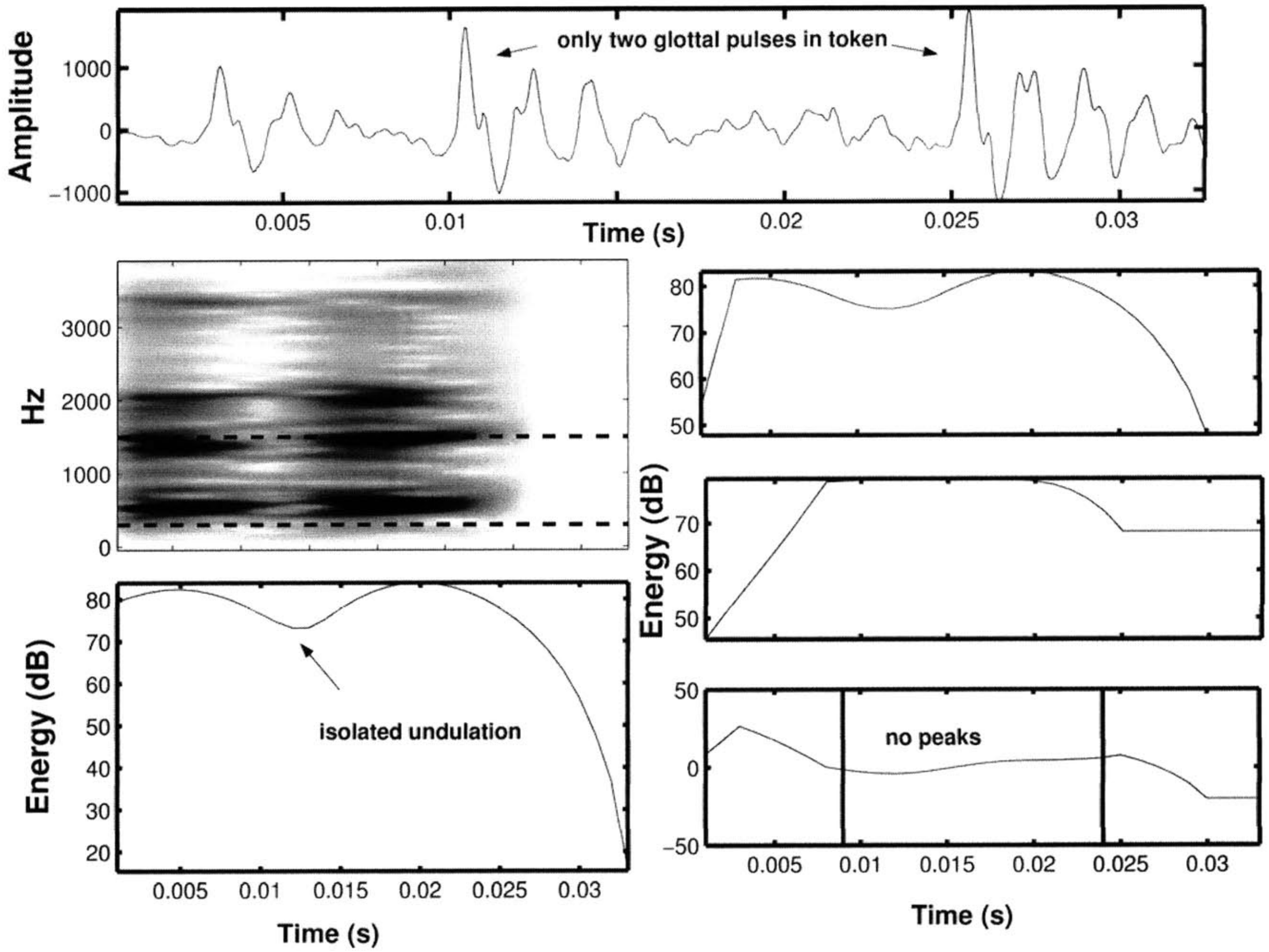

Figure 6-6: Illustration of misleading smoothed-energy-difference estimates for irregular phonation. **(a)** Example of vocal fry with only two glottal pulses. See Figure 5-3 for an explanation on how to interpret the remaining panes.

energy waveform for irregular tokens. Hence, smoothing the energy-waveform using different window sizes does not result in a peak value.

The misleading cue values are expected to be offset by the F0, normalized RMS and shift-difference amplitudes. The example in Figure 6-6 has a F0 value of **0**, a normalized RMS amplitude of **1.53** and a shift-difference amplitude of **0.71**. Although the normalized RMS amplitude for this cue is inappropriately high and is greater than the mean of the normalized RMS cue for regular tokens, the F0 is 2 stdv. below and the shift-difference amplitude is 5 stdv. above their respective means for regular tokens.

Regular tokens

Among the 2134 instances of regular tokens with inappropriately high smoothed-energy-difference values (> 2), 1930 instances were generated by male speakers. This suggests that a strong correlation exists between speaker gender and inappropriate smoothed-energy-difference estimates for regular phonation. Another common characteristic of these examples is a high second formant frequency as seen in Figure 6-7.

These characteristics suggest that the combination of the wider spacing of the glottal pulses for male speakers compared to female speakers and the low amplitude because of high F2 leads to high smoothed-energy-difference estimates for regular tokens. Specifically, the lower window size of 6 ms will encompass a glottal pulse for a female speaker, and hence the difference in the averaged energy for smoothing-window sizes of 6 ms and 16 ms will be small. For male speakers, it is unlikely that the lower window size of 6 ms will encompass a glottal pulse, which combined with the low amplitude because of the high second formant frequency, results in a rapid transition of energy for different smoothing window sizes and a high smoothed-energy-difference output.

The example in Figure 6-7 has F0, normalized RMS and shift-difference amplitudes of **(110, 1.29, 0.31)** respectively. The F0 is 0.5 stdv. above, the normalized RMS is 1 stdv. above and the shift-difference amplitude is 1 stdv. below their respective means for irregular tokens. These values should offset the misleadingly high smooothed-energy-difference estimate of **2.06** in this example.

6.3.4 Shift-difference amplitude

Irregular tokens

Irregularities in the form of isolated glottal pulses are the main reason for inappropriately low values of shift-difference amplitude for irregular tokens as shown in Figure 6-8. If the isolated pulse occurs in the middle of the token, the difference between the shifted segments is negligible resulting in an unexpectedly low shift-difference esti-

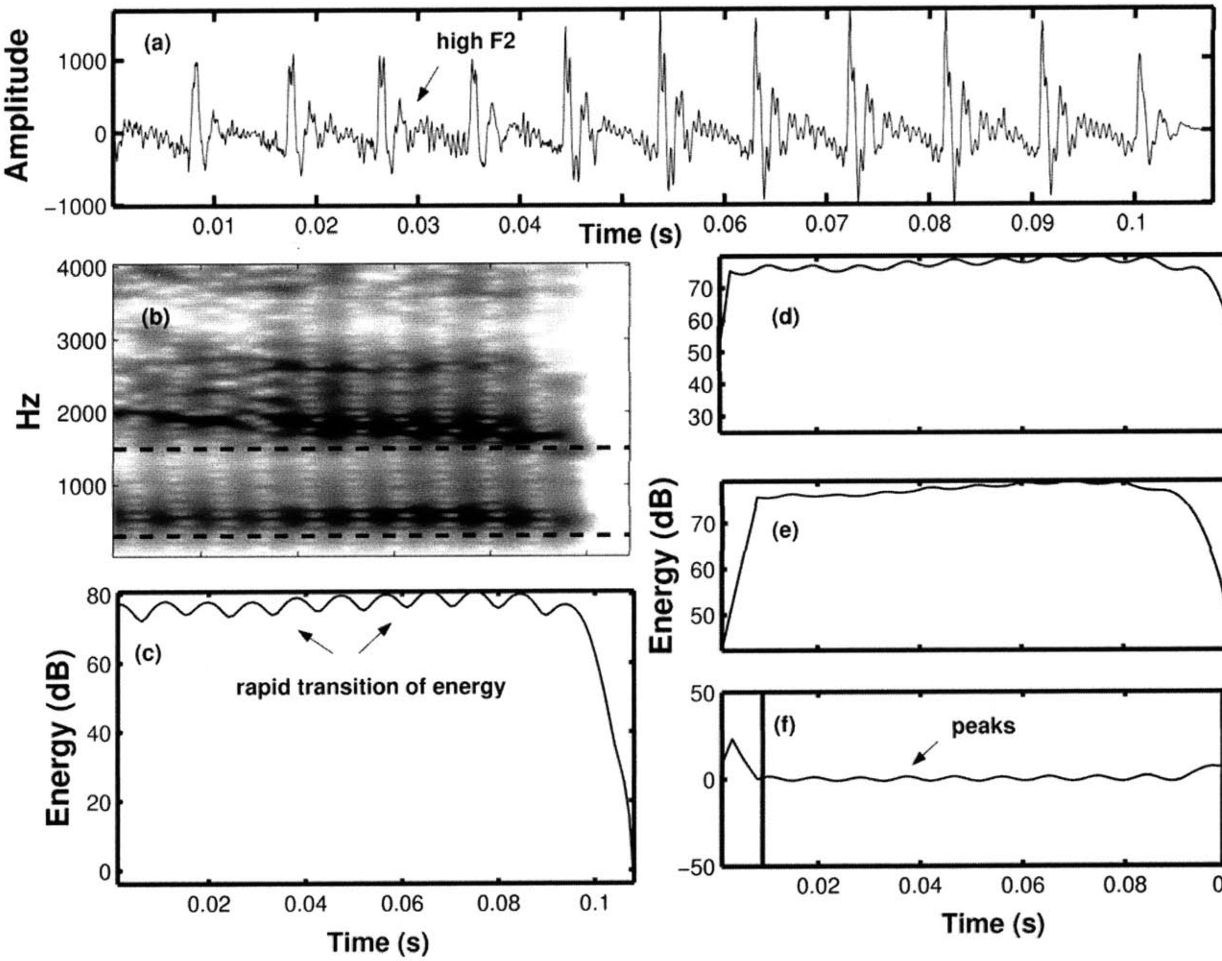

Figure 6-7: Illustration of misleading smoothed-energy-difference estimate for regular phonation. **(a)** A regular token. See Figure 5-3 for an explanation on how to interpret the remaining panes.

mate for the irregular token. Again, the remaining three cues are expected to classify the token as irregular. For the example in Figure 6-8, the F0, normalized RMS amplitude and smoothed-energy-difference amplitude are **(0, 0.3289 and 6.4789)** respectively. The F0 is 2 stdv. below, the normalized RMS amplitude is 1.5 stdv. below and the smoothed-energy-difference amplitude is 4 stdv. above their respective means for regular tokens.

Regular tokens

Some regular tokens show inappropriately high shift-difference amplitudes. Out of these tokens, a few are characterized by decaying amplitude (perhaps they occur in utterance-final position) as in (a) of Figure 6-9. The difference between adjacent segments within the same token is therefore larger than expected compared to other regular tokens. Other cases are borderline regular as shown in (b) of Figure 6-9 and do not show a completely regular structure in the time-domain waveform. The F0, normalized RMS amplitude and smoothed-energy-difference amplitude for the two cases are **(115, 0.84, 2.29)** and **(119, 1.04, 0)** respectively which should offset the inappropriately high shift-difference values of **(0.54, 0.57)** respectively. For the first example, the F0 and normalized RMS amplitude are 0.5 stdv. above their respective means for irregular tokens, while the smoothed-energy-difference amplitude is 1 stdv. below its mean for irregular tokens. The second example shows a similar separation from the irregular tokens except for the shift-difference amplitude cue which is 2 stdv. below its mean for irregular tokens.

6.3.5 Summary

Table 6.2 presents a summary of the most common causes of failure for each cue for both regular and irregular tokens. It must be emphasized that for tokens where one cue fails to generate the expected value, the other cues are generally expected to provide information to adequately classify the token as regular or irregular as illustrated in the previous section.

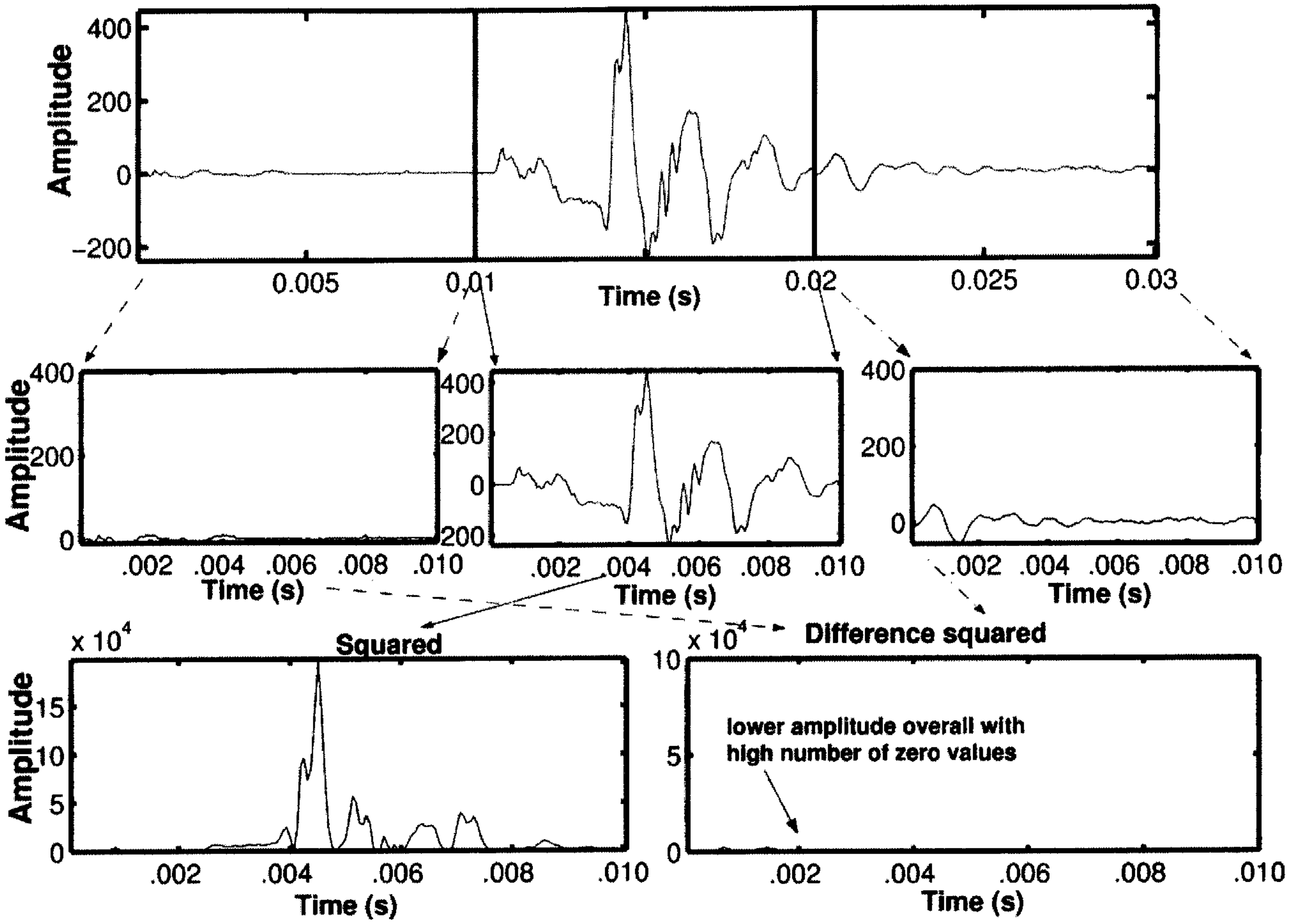

Figure 6-8: Illustration of misleading shift-difference estimates for irregular phonation. **(a)** A typical irregular token after high-pass filtering and pre-emphasis. The solid line shows the middle segment of the token, while the dashed line marks the shifted segments after shifting the window by 10 ms in either direction. **(b)** The left-shifted segment. **(c)** The middle segment. **(d)** The right-shifted segment. **(e)** The squared middle segment. **(f)** The squared difference between (b) and (d). The output shift-difference magnitude is the $\sqrt{(f)/(e)}$ which for this particular time-shift is inappropriately low at 0.05.

Table 6.2: Common causes of failure for each cue for regular and irregular tokens.

	Cause of failure	
	(Regular tokens)	**(Irregular tokens)**
F0	Low threshold value	Widely spaced glottal pulses
Normalized RMS	Low amplitude	High F1 amplitude
Smoothed-energy-diff	High F2	Only one or two glottal pulses in token
Shift-diff	Borderline regular Decaying amplitude	Isolated glottal pulses

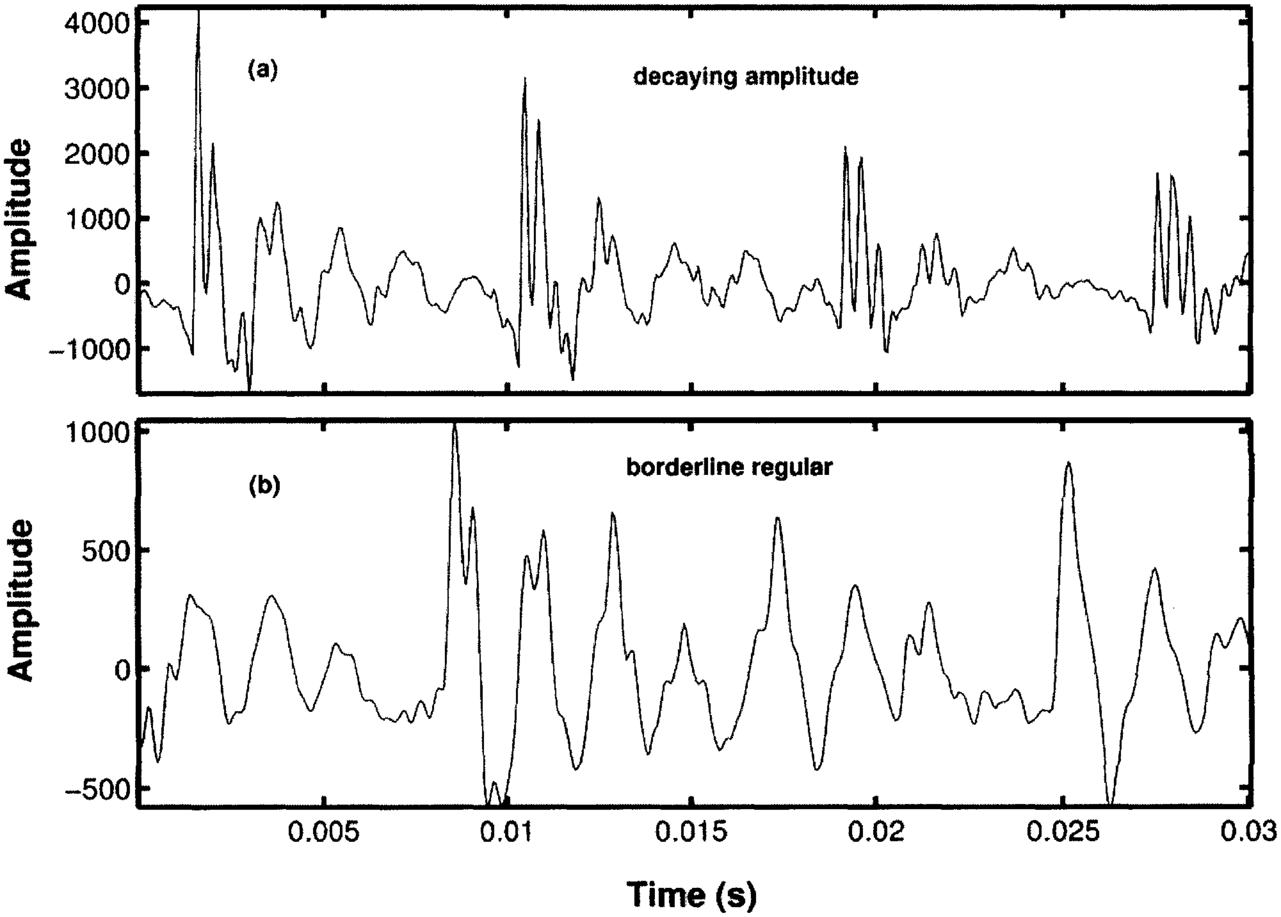

Figure 6-9: Two examples of regular tokens with misleading shift-differencc estimates. (a) A regular token with decaying amplitude. (b) A borderline-regular token.

6.4 Failure analysis for all cues

There are a few cases where all four acoustic cues return unexpected values for both regular and irregular tokens.

Most of the irregular tokens that show such behavior are either examples of vocal fry produced by females with F0 values higher than the expected vocal fry range (as outlined by Blomgren *et. al.*,, 1998) or examples of diplophonia with glottal pulses alternating in amplitude as shown in Figure 6-10.

The incorrect estimates due to examples of vocal fry with high F0 values could be eliminated by making the classification system gender-dependent. Essentially, these tokens are irregular relative to other regular tokens produced by females since the spacing between the glottal pulses is wider relative to the regular glottal pulse spacing for females. However, this wider spacing of glottal pulses for females is consistent with the regular spacing of the glottal pulses for males as seen in (a) of Figure 6-10. The values of the acoustic cues for these particular examples of vocal fry for females are therefore consistent with the acoustic cue values of regular tokens for males.

The examples of diplophonia with misleading acoustic values show a very regular structure except for the alternating glottal pulse amplitudes as seen in (b) of Figure 6-10. It can be surmised that additional cues are needed to distinguish these particular types of irregular phonation from regular phonation.

Regular tokens with unexpected acoustic cue values are mostly examples which should not have been classified as regular in the first place as seen in (c) and (d) of 6-10. Since the regular tokens have not been individually hand-labeled and are extracted from stressed vowels, these occasional mislabels are expected and the resulting incorrect classification is acceptable.

The existence of a category between regular and irregular phonation must also be acknowledged. A few tokens in the data set fell under this category and should not be classified as either regular or irregular. These tokens show a waveform structure which is neither periodic nor aperiodic. Future work on this topic could include a third category of tokens which are neither clearly regular nor clearly irregular, based

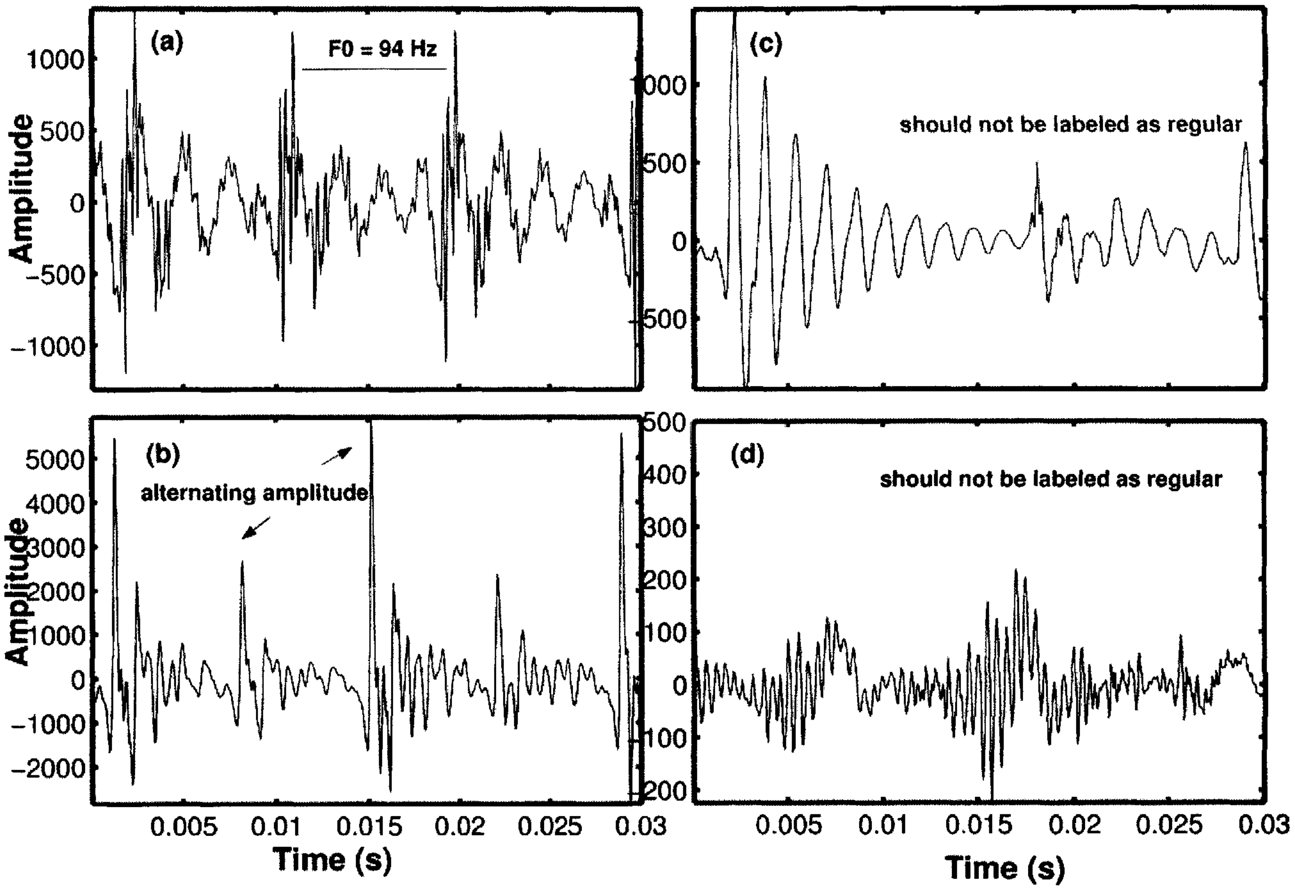

Figure 6-10: Examples of regular and irregular tokens with misleading estimates for all four acoustic cues. (a) and (b) show common irregular phonation examples, while (c) and (d) show common regular phonation examples, which lead to failure. (a) An example of vocal fry produced by a female with a relatively high F0 value compared to the expected vocal fry range. (b) An example of diplophonia with glottal pulses alternating in amplitude. (c) & (d) Examples of tokens which should not be labeled as regular.

on the definitions in Chapter 1.

Chapter 7

Classification

This chapter gives a brief background on Support Vector Machines (SVMs) and outlines the results obtained in the classification of regular and irregular phonation using SVMs.

7.1 Support Vector Machines

7.1.1 Theory

SVMs are learning machines for pattern classification and regression tasks based on statistical learning theory (Vapnik, 1995). Given a set of training vectors $\{\mathbf{x}_i\}_{i=1}^{l}$, and the corresponding class labels $\{y_i\}_{i=1}^{l}$ such that

$$y_i \in \{-1, +1\} \textbf{ and } \mathbf{x}_i \in \Re^n$$

SVMs select a set of support vectors $\{\mathbf{x}_i^{SV}\}_{i=1}^{SV}$ that is a subset of the training set $\{\mathbf{x}_i\}_{i=1}^{l}$ and find an optimal decision function

$$f(\mathbf{x}) = sign(\sum_{i=1}^{N_{SV}} y_i \alpha_i K(\mathbf{x}_i^{SV}, \mathbf{x}) - b)$$

where K is an em a priori chosen kernel function. The weights α_i, the set of support vectors $\{\mathbf{x}_i^{SV}\}_{i=1}^{N_{SV}}$ and the bias term b are found from the training data using quadratic optimization methods. A Gaussian kernel is used in this study to classify regular and irregular phonation. For the gaussian kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma|\mathbf{x}_i. - \mathbf{x}|^2)$$

The experiment in this study was carried out using the OSU SVMs Toolbox (http://www.ece.osu.edu/˜maj/osu_svm/).

7.1.2 RBF (Gaussian) kernel

The Gaussian kernel requires two parameters: C and γ. It is not known beforehand which C and γ will be the best for a particular problem. In order to find the best (C,γ) so that the classifier can accurately predict the unknown testing data, 3-fold cross-validation was used (Hsu, Chang, Lin).

The first 960 tokens of both regular and irregular phonation were used as the training set during cross-validation. In 3-fold cross-validation, the training set is divided into 3 subsets of equal size. Sequentially, one subset is tested using the classifier trained on the remaining 2 subsets. Thus, each instance of the whole training set is predicted once. The cross-validation accuracy is the percentage of data which are correctly classified.

Using a "grid-search" on C and γ, various pairs of values were tried. The (C,γ) which resulted in the best classification rates equal to (256, .0312).

7.1.3 Results

The training and test data need to be scaled to prevent cues in higher numeric ranges from dominating the smaller numerical ranges which could lead to numerical difficulties. In order to prevent this problem, the distribution for each cue in the training set is converted to zero mean, unit variance for both regular and irregular tokens. Since the test-data also needs to be normalized appropriately, the distribution of each of the cues in the test-set is normalized using the mean and variance of the associated cue in the training set for both regular and irregular tokens.

Figure 7-1 shows the Receiver Operating Characteristic (ROC) curves for the correct classification of irregular tokens based on the SVM outputs. An ROC curve is a graphical representation of the trade off between the false negative and false positive rates for every posssible threshold. The diagnostic test is successful when the ROC curve climbs rapidly towards the upper left hand corner of the graph. This means that (1 - the false negative rate) is high and the false positive rate is low. The test is unsuccessul when the ROC curve follows a diagonal path from the lower left hand corner to the upper right hand corner. This behavior means that every improvement in the false positive rate is matched by a corresponding decline in the false negative rate.

The rise of the ROC curve to the upper left hand corner can be quantified by measuring the area under the curve — the larger the area, the better the diagnostic test. If the area is 1.0, the test is ideal, because it achieves both 100% sensitivity (synonymnous with the true positive rate) and 100% specificity (synonymnous with the true negative rate). If the area is 0.5, then the test has effectively 50% sensitivity and 50% specificity. This is a test that is no better than flipping a coin. In practice, a diagnostic test has an area somewhere between these two extremes. The quality of the test is judged by the proximity of the area under the curve to 1.

For each ROC curve in Figure 7-1, only 959 irregular tokens are used for training since the number of irregular tokens is limited in the data-set. However, the number of regular tokens used for training is increased from 959 → 1500 → 2500 → 3500. Figure 7-1 shows an improvement in the classification scheme as the training size of regular tokens is increased, but this improvement decreases after 2500 samples.

The same test-set, consisting of 4320 regular tokens and 320 irregular tokens, is used in the SVM as the training size of regular tokens is increased. The unequal size of the test-set should not affect the performance of the SVM and is merely an artifact of having an unequal number of regular and irregular tokens.

The area under the ROC curves is close to 1 using 2500 regular tokens for training as shown in table 7.1 showing that the SVMs can classify regular and irregular phonation well. Using a threshold of 0 in this case, a recognition rate of 91.25% is obtained

Table 7.1: Area under ROC curve using 959 irregular tokens and different number of regular tokens for training the SVM. The same test-set, consisting of 4320 regular tokens and 320 irregular tokens, is used in all the cases.

No. of regular tokens	**Area under ROC**
959	0.74
1500	0.87
2500	**0.93**
3500	0.90

for irregular phonation with a false alarm rate of 4.98% while regular phonation is classified with a recognition rate of 95.02% and a false alarm rate of 8.75%. The threshold value can be adjusted based on the requirements of the system to increase the recognition rate for a particular class, but this would also result in an increase in the false error rates for that class.

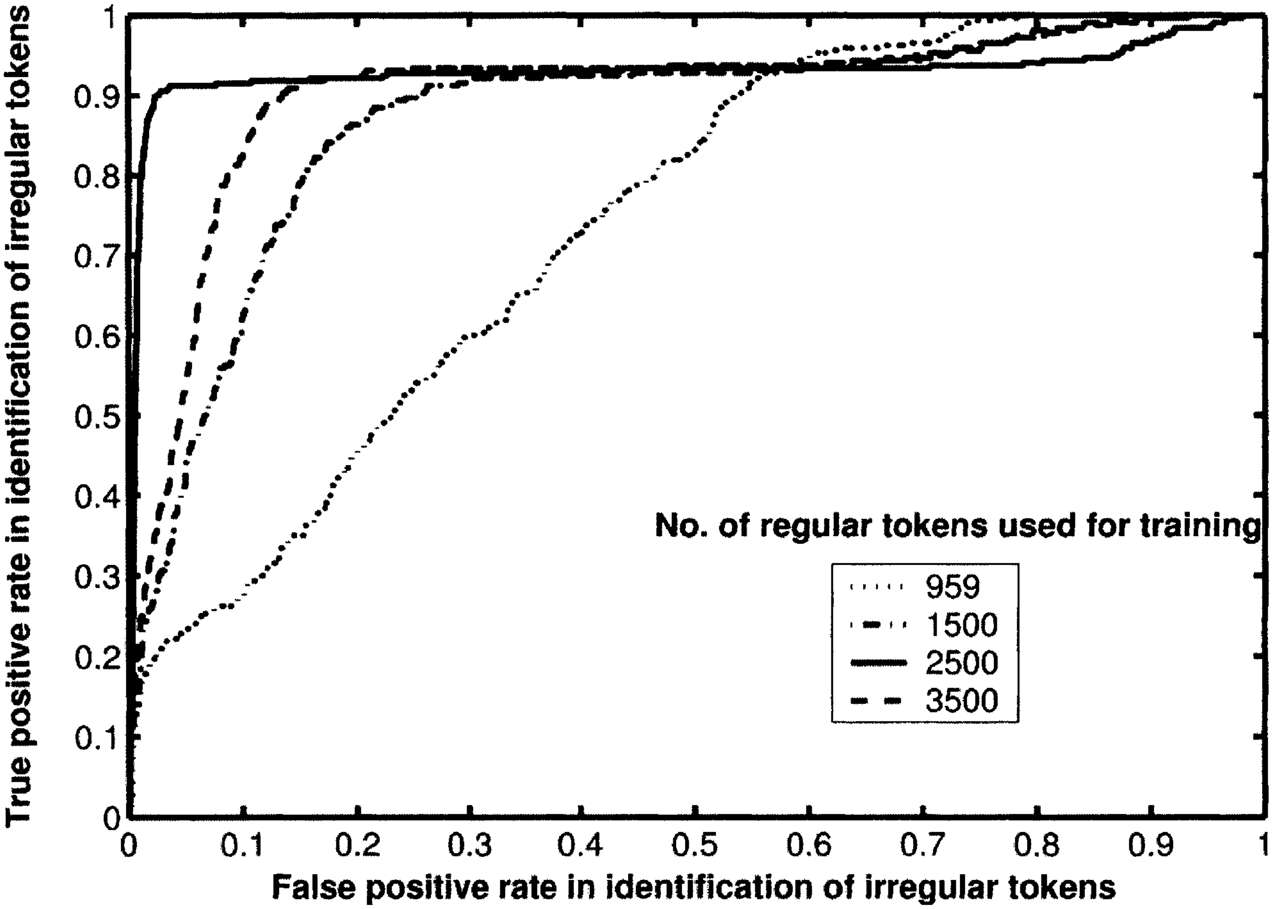

Figure 7-1: ROC curves for the classification of irregular tokens

Chapter 8

Irregular phonation as a segmentation cue

The large scale labeling associated with the thesis offered an opportunity to study aspects of the communicative function of irregular phonation in a speaker-independent setting. Specifically, it enabled a study on the efficacy of irregular phonation as a reliable cue towards the segmentation of continuous speech in American English.

8.1 Introduction

A large body of research exists regarding the range of acoustic cues used to mark boundaries in the speech stream. These cues serve a segmentation purpose for various types of units — including syllables, words, phrases, utterances and dialogues. In American English, these cues include the aspiration of voiceless stop consonants in syllable-initial position (for stressed syllables), segmental lengthening prior to a major prosodic boundary such as the utterance, and signal amplitude changes in the vicinity of a silent pause as the speaker suspends the sound source. In particular, prior work has focused on specifying the factors which determine the likelihood that a word boundary will be marked with irregular phonation. In general, these factors may arise from a segmental context and/or a prosodic environment. For example, irregular phonation tends to occur at word boundaries between vowels (Umeda, 1978), and at syllable final /t/ and sometimes /p/ (Pierrehumbert, 1994). The occurrence of irregular phonation at word-initial vowels and its relationship with the prosodic structure of the utterance has also been explored (Dilley & Shattuck-Hufnagel, 1995; Pierrehumbert, 1995). Their studies show that irregular phonation at word-initial vowels occurs more often at the beginnings of intonational phrases, and to a greater degree if the word is pitch-accented.

As stated, these studies focus on determining the factors that influence the likelihood that a word boundary will be marked with irregular phonation. In this chapter, a related question is addressed with a slightly different focus — given the presence of irregular phonation, what is the likelihood of a word boundary at that location? Similarly, if irregular phonation does not occur at a word boundary, in what context does it occur? The results directly support the use of automatically detected regions of irregular phonation in spoken language systems. First, these irregular regions can help determine the probability of a word-boundary location. Also, with limited additional context, the probability estimate for a word-boundary can be strengthened. Specifically, two cases are examined in more detail — voiceless stop consonants and vowel-vowel junctions.

The ends of utterances (and phrases) have also been observed to be marked with irregular phonation (Lehiste, 1979; Kreiman, 1982). Given the structure of the TIMIT database as isolated utterances, utterance-initial and utterance-final irregular phonation as well as syntactic level phrase-initial and phrase-final irregular phonation are examined.

8.2 Data set

Only the training data set was used for this portion of the study resulting in 1331 hand-labeled irregular tokens from 114 speakers. The word transcription of the TIMIT database was used to determine word and utterance boundaries. Regions of irregular phonation were classified in relation to the syntactic boundaries of syllable, word, phrase and utterance. Phone-related instances for /p/, /t/, /k/ and

Table 8.1: Syntactic boundary labels for irregular token occurrence.

Word level	Phrasal level	Stops	Other
Word-final	Utt-final	p	Vowel-vowel
Word-initial	Utt-initial	t	
Syll-final	Phrase-final	k	
Syll-initial	Phrase-initial		
	Last phonation in utt.		
	First phonation in utt.		

vowel-vowel sequences were classified using the TIMIT phonetic transcription. A summary of the classification categories is given in Table 8.1 for four categories — word level, phrasal level, voiceless stop consonants and vowel-vowel boundaries. Within the word-level and phrasal-level categories, the labeling is mutually exclusive. For example, a word-initial occurrence of irregular phonation is not counted as syllable-initial. Similarly, an utterance-initial occurence of irregular phonation is not marked as phrase-initial.

8.3 Results

Figure 8-1 shows the percentage, as well as the absolute number, of irregular tokens that occur at word and syllable boundaries. Out of all the irregular tokens, 78% occur at word boundaries — 45% at word-final locations and 33% at word-initial locations. An additional 5% of the irregular tokens occur at syllable boundaries. These tokens were re-analyzed, leading to two main observations. First, of the 69 irregular tokens occurring at syllable boundaries, 50 occurred (72%) either at the junction of a compound word (e.g. "outcast") or at the junction of a base word and a suffix. For example, irregular phonation was noted at the end of 'equip' in 'equipment'. Secondly, 52 of the 69 irregular tokens (75%) at syllable boundaries conincided with a voiceless stop location (either /p/, /t/ or /k/).

Figure 8-2 shows the percentage, as well as the absolute number, of irregular tokens at phrasal boundaries. Combined, 48% of the irregular tokens occur at phrasal

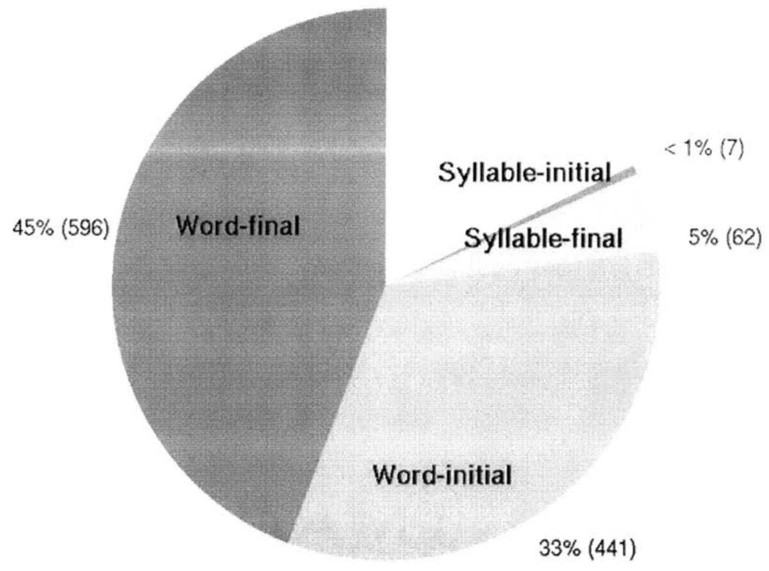

Figure 8-1: Breakdown of irregular phonation at word and syllable boundaries. The absolute number is shown next to the percentage within brackets. (Based on 1331 tokens)

boundaries — 27% at utterance boundaries while another 21% at syntactic phrase boundaries. In Figure 8-2, the irregular tokens occurring at the last instance of phonation within the utterance are combined with the utterance-final tokens. Similarly, the irregular tokens at the first place of phonation within the utterance are combined with the utterance-initial tokens.

Figure 8-3 shows the percentage, as well as the absolute number, of the irregular tokens that occur at voiceless stop /p/, /t/ or /k/ and at vowel-vowel junctions. A total of 24% of the irregular tokens occur at voiceless stop consonants and 10% occur at vowel-vowel junctions.

A further study of the irregular tokens at voiceless stop locations and vowel-vowel junctions was conducted in relation to word-boundaries (Figure 8-4). All the irregular tokens at vowel-vowel junctions occur at word-boundaries, i.e. either in word-initial or word-final position. For the irregular tokens at voiceless stops, 268 of the 326 occur at word-final position while another 44 occur at syllable-final position. All 44 of the syllable-final irregular tokens for voiceless stops occur either at the junction of a compound word or at the junction of a base word and a suffix.

Additional analysis was conducted on cases of irregular phonation which do not

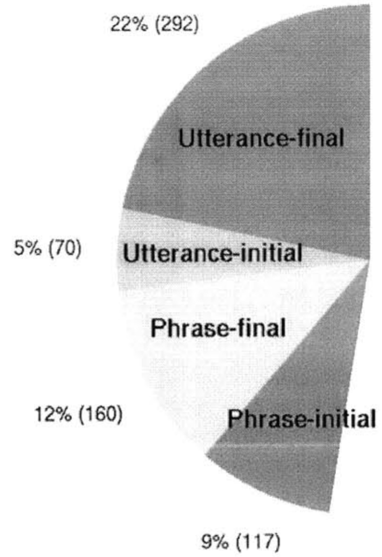

Figure 8-2: Breakdown of irregular phonation at syntactic phrase and utterance boundaries. The absolute number is shown next to the percentage within brackets. (Based on 1331 tokens)

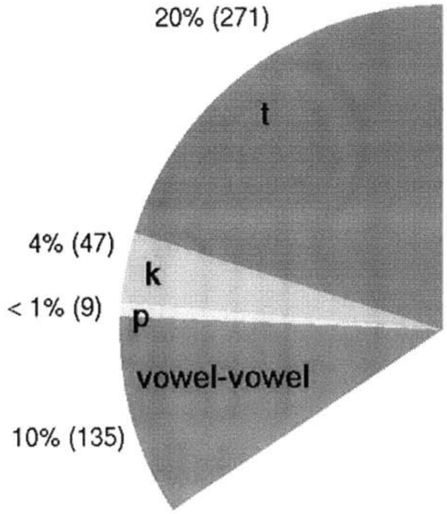

Figure 8-3: Breakdown of irregular phonation at voiceless stops and vowel-vowel boundaries, The absolute number is shown next to the percentage within brackets. (Based on 1331 tokens)

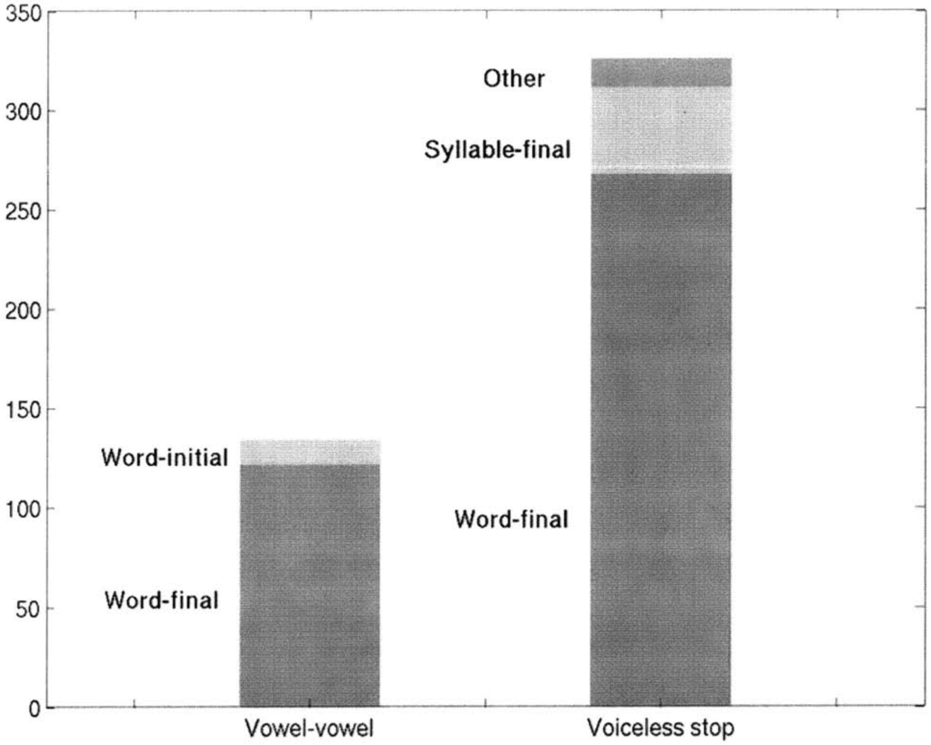

Figure 8-4: Breakdown of irregular phonation at word level boundaries for vowel-vowel junctions and voiceless stops.

Table 8.2: Contexts in which irregular phonation at word-medial position occur in the data set.

Before or after a voiceless consonant
Before or after a voiced consonant
Before or after a sonorant consonant
Function word 'a'
Other

coincide with either a word or syllable boundary in order to determine the context in which the irregular phonation occurs. Table 8.2 lists the five broad contexts in which these irregular tokens occur.

A total of 225 irregular tokens occur at neither word nor syllable boundaries. Figure 8-5 shows their distribution among the five categories listed in Table 8.2. Of the 225 irregular tokens not at word-boundaries, 158 occur adjacent to a voiceless consonant. Analyzing these tokens showed that 130 of these occur either in utterance-final location or before a pause in the utterance. In Figure 8-5, utterance-final voiced consonants (stops & fricatives) are grouped with the voiceless consonants since such realizations are largely devoiced. One such example is shown in Figure 8-6 (a) where the word "subject" occurs in utterance-final location and the irregular token precedes the voiceless stop /k/.

Of the 9 irregular tokens which occur adjacent to a voiced consonant, 6 are at utterance-initial or phrase-initial position. And, of the 31 irregular tokens next to sonorant consonants, 16 occur either at the last word of the utterance or at pre-pausal locations. The 7 irregular tokens at function words encompass the entire word and hence are classified as neither word-initial nor word-final. The remaining 20 irregular tokens classified under the "Other" category include 10 tokens which show irregular phonation in vowel-medial position. This particular behavior is observed across multiple speakers. Figure 8-6 (b) shows one particular example where the irregular token occurs within the vowel /ae/ in the word "packing".

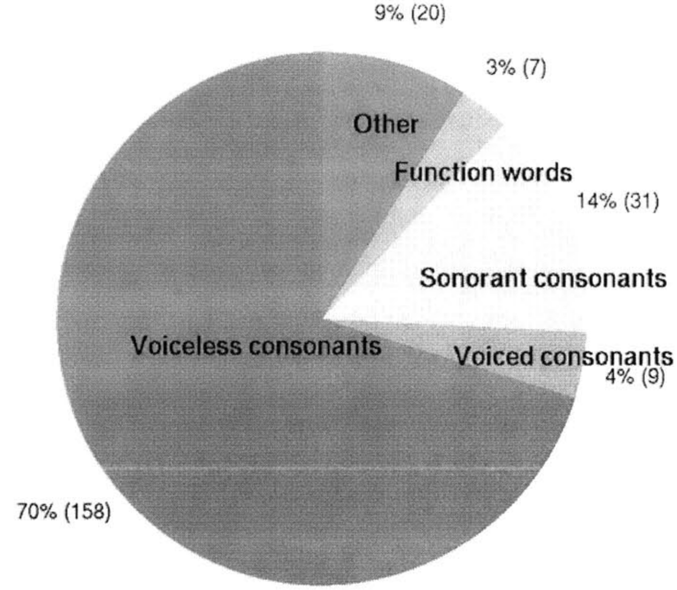

Figure 8-5: Breakdown of irregular phonation which does not occur at word or syllable boundaries.

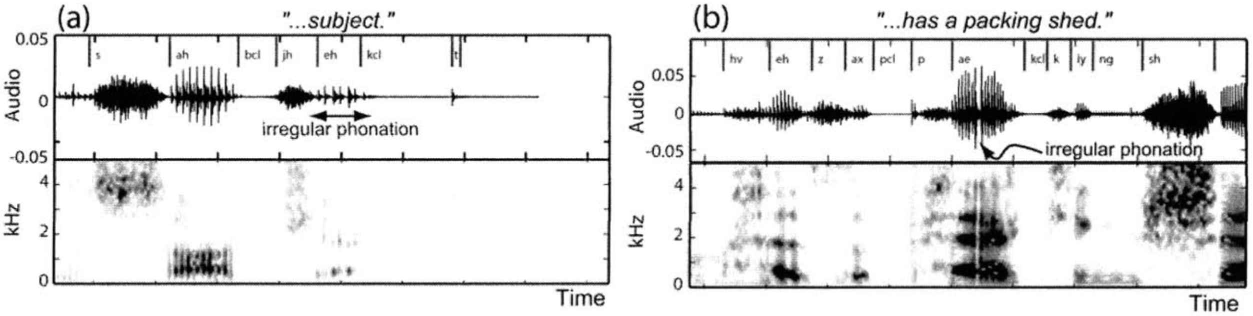

Figure 8-6: Two examples of irregular phonation which do not occur at word boundaries. (a) is an example of an irregular token adjacent to a voiceless consonant in utterance-final location while (b) shows an irregular token in vowel-medial position.

8.4 Discussion

This chapter addresses the question of whether or not all detected instances of irregular phonation in American English are associated with a boundary location. The results are collected in a speaker-independent analysis across 114 different speakers and show that 78% of the irregular tokens occur at a word boundary. Batliner *et al.* (1993) examined instances of irregular phonation for German speech. One-third of the database consisted of real spontaneous utterances gained from human-human clarification dialogues, while the rest consisted of the same utterances read by the same speakers nine months afterwards. From a total of 1191 irregular portions of speech, 58% occurred in word-initial position and 18% occurred at the end of a word. The results of the present study for American English are highly consistent with the results of Batliner *et al.* (1999), and support the conclusion that irregular phonation is a strong acoustic cue for the detection of word boundaries.

These results are applicable to the development of spoken language systems for lexical access or automatic speech recognition. The detection of a subset of the word boundaries in a speech stream (based on robust acoustic cues such as irregular phonation and regions of silence) can provide segmentation of the speech stream into limited regions for proposing a cohort of word candidates. Appropriately limiting the search region prevents the cohort from growing unmanageably large. The results of the present study are in conjunction with an effort towards the development of a system for automatic classification of regions of phonation as either regular or irregular discussed earlier in Chapters 5, 6 and 7.

A secondary question examined the 22% of the tokens of irregular phonation which did not occur at a word boundary and asked if there are still consistent observable trends in relation to other types of boundaries. Of these 22% of the tokens (294 in total), 50 were found to occur at syllable boundaries located at the junction of a compound word or between a base word and a suffix (such as -ment, -ly, or -en). An additional 130 tokens, which do not occur at a syllable boundary, occur in the vicinity of a voiceless (or devoiced) consonant at the end of an utterance, and 6 tokens, occur following a nominally voiced stop consonant at the start of an utterance (/b/, /d/ or /g/).

Recently, the physiological correlates to irregular phonation in utterance-final location, for utterances ending with a vowel, have been quantitatively studied by Slifka (2005). She found that when the end of the utterance coincides with the speaker taking a breath, the conditions associated with the respiratory actions to finish one breath and prepare for the next inhalation tend to give rise to a particular type of irregular phonation - one that is produced with relatively widely abducted vocal folds or produced as the vocal folds are in the process of continuing to abduct. This configuration yields irregular phonation which is highly damped and is in contrast to definitions of glottalization associated with tightly adducted vocal folds. In the present data, 58% of the tokens not occurring at a word or syllable boundary occurred in the vicinity of the end of an utterance. For example, in an utterance ending in the word 'subject,' the utterance ends with a voiceless consonant production, but the last instance of phonation in the utterance is irregular. In such cases, a physiological basis similar to that in Slifka (2005), may create conditions conducive to irregular phonation.

Overall, for the 22% of irregular tokens which do not occur at a word boundary, 63% of them do occur in a boundary-related environment (such as syllable or utterance). These results further support the conclusion that, if in a spoken language system, an instance of irregular phonation is detected, the probability of a speech boundary at that location should be very high. The type of the boundary will depend on additional analysis which might include acoustic cues related to the specific nature of the irregular phonation, other acoustic cues related to the prosodic structure (such as duration and intonation), or the information regarding the segmental context.

Chapter 9

Conclusion

This thesis has presented a knowledge-based classification scheme to distinguish between regular and irregular phonation with accuracy rates greater than 90%. With four cues (fundamental frequency, normalized RMS amplitude, smoothed-energy-difference amplitude and shift-difference amplitude), a clear separation between regular and irregular tokens using Support Vector Machines (SVMs) was found. The results are especially significant since the proposed system uses tokens from multiple speakers — 114 different speakers for training and 37 different speakers for testing. Given the high inter-speaker variation of irregular phonation, the high accuracy rates using multiple speakers proves the robustness of the proposed cues. In addition, both male and female speakers are well represented in the data-set and the regular and irregular tokens used for training and testing occur in various contexts (i.e. utterance-initial, phrase-final, utterance-final etc.). These characteristics of the data-set are in contrast to the characteristics of those used in previous studies which were either speaker-specific, gender-specific and/or context-specific. This thesis has demonstrated that it is possible to make the classification of regular and irregular phonation gender-, speaker- and context-independent with high accuracy rates.

Additionally, this thesis has also conducted an in-depth study on instances of irregular phonation and their place of occurrence. It is observed that irregular phonation occurs often at word-boundaries. Those instances of irregular phonation which do not occur at word-boundaries are usually associated with some other speech-boundary.

Given that regions of phonation can be classified as regular or irregular with a high degree of consistency, these results confirm that regions of irregular phonation can reliably serve as a segmentation cue for speech recognition, speech parsing and speech-synthesis. Future studies also offer the possibility of combining several cues with irregular phonation to build the prosodic structure of an utterance in a spoken language system.

Bibliography

[1] Abramson, A.S. & Tingsabadh, K. "Thai final stops : cross-language perception," *Phonetica*, Vol. 56, 111-122, 2000.

[2] Batliner, A., Burger, S., Johne, B., & Kiessling, A. "MÜSLI: A classification scheme for laryngealizations," In *Proc. ESCA Workshop on prosody*, 176-179, Lund, September, 1993.

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

[4] Blankenship, B. "The timing of nonmodal phonation in vowels," *Journal of Phonetics*, Vol. 30, 163-191, 2002.

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

[6] Catford, J.C., "Phonation types: The Classification of some laryngeal components of speech production," In *Jones, Daniel (dedic)*, 1964.

[7] Catford J.C., *Fundamental Problems in Phonetics.* Bloomington: Indiana University Press: 99, 1977.

[8] Colton, R.H. & Casper, J.K. *Understanding Voice Problems.* Baltimore, MD: Williams & Wilkins, 1996.

[9] Dibazar A.A., Narayanan S. & Berger M., "Feature analysis for automatic detection of pathological speech," In *Proc. Engineering Medicine and Biology Symposium*, vol. 1, 182-183, 2002.

[10] Dilley, L. & Shattuck-Hufnagel, S. "Variability in glottalization of word onset vowels in American English," In *Proceedings of the XIIIth international congress of phonetic sciences*, Stockholm, vol. 4, 586-589, 1995.

[11] Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M. "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, Vol. 24, 423-444, 1996.

[12] Docherty, G. (Editor) *Preface - Journal of Phonetics*, Vol.29, 4, 2001.

[13] Dolanský L., Tjernlund P. "On certain irregularities of Voiced Speech Waveforms," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-16, No.1, March, 1968.

[14] Fant, G.C.M. *Acoustic Theory of Speech Production.* Gravenhange, The Netherlands: Mouton & Co., 1960.

[15] Fischer-Jorgensen, E. "Phonetic analysis of the stod in Standard Danish," *Phonetica*, Vol. 46, 1-59, 1989.

[16] Gerratt R.B. & Kreiman J. "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, Vol. 29, 365-381, 2001.

[17] Gordon, A. & Ladefoged, P. "Phonation types: a cross linguistic overview," *Journal of Phonetics*, Vol. 29, 383-406, 2001.

[18] Hagen, A. "Linguistic functions of glottalization and their language specific use in English and German," *Master's thesis in Computer Science*, Institut für Mathematische Maschinen und Datenverarbeitung, Nürnberg / Massachusetts Institute of Technology, Cambridge.

[19] Hanson, H. M., Stevens, K., Kuo, H.J., Chen, M. & Slifka, J. "Towards models of phonation," *Journal of Phonetics*, Vol. 29, 451-480, 2001.

[20] Henton, C.G. & Bladon, A. "Creak as a sociophonetic marker," In *Language, speech and mind: studies in honor of Victoria A. Fromkin*(L. Hyman & C. Li,editors). London: Routledge, 3-29, 1987.

[21] Hillenbrand, J.M. & Houde, R.A. "The role of f_0 and amplitude in the perception of intervocalic glottal stops," *Journal of Speech and Hearing Research*, Vol. 39, 1182-1190, 1996.

[22] Hollien, H., Moore, P., Wendahl, R.W. & Michel, J. "On the nature of vocal fry," *Journal of Speech and Hearing Research*, Vol.9, 245-247, 1966.

[23] Huber, D. "Aspects of the communicative function of voice in text intonation," Ph.D. dissertation, Chalmers University, Göteborg/Lund, 1988.

[24] Huber, D. "Perception of aperiodic speech signals," *International Conference on Spoken Language Processing*, Vol.1 , 503-505, 1992.

[25] Ishi, T.C. "Analysis of autocorrelation-based parameters for creaky voice detection" In *ISCA Archive.* Presented at Speech Prosody 2004; Nara, Japan; March 23-26, 2004.

[26] Jakobson, R.C., Fant, G.M. & Halle, M. "Preliminaries to speech analysis: The distinctive features and their correlates," Technical Report 13, Acoustics Laboratory, MIT, 1952.

[27] Javkin H., Maddieson I. "An Inverse Filtering Study of Burmese Creaky Voice," *UCLA Working Paper in Phonetics*, Vol. 57, 115-130, 1983.

[28] Kiessling, A., Kompe, R., Niemann, H., Nöth, E. & Batliner, A. "Voice source state as a source of information in speech recognition: Detection of laryngealizations," In *Speech Recognition and Coding — New Advances and Trends"* (Rubio-Ayuso & Lopez-Soler, editors). Springer, 329-332, 1995.

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

[30] Kohler, K.J. "Glottal stops and glottalization in German," *Phonetica*, Vol. 51, 38-51, 1994.

[31] Kohler, K.J. "Investigating unscripted speech: implications for phonetics and phonology," *Phonetica*, Vol. 57, 85-94, 2000.

[32] Kochanski G., Grable E., Coleman, J. & Rosner, B. "Loudness Predicts Prominence; Fundamental Frequency Lends Little," *Journal of the Acoustical Society of America*, Vol. 11(2), 1038-1054, 2005.

[33] Kreiman, J. "Perception of sentence and paragraph boundaries in natural conversation," *Journal of Phonetics*, Vol. 10, 163-175, 1982.

[34] Kreiman, J. & Gerratt, B.R. "Measuring voice quality," In *Handbook of voice quality measurement* (R.Kent & M.J.Ball,editors). San Diego: Singular Publishing Group, 73-102, 1999.

[35] Kreiman, J., Gerratt, B.R. "Sources of listener disagreement in voice quality assessment," *Journal of the Acoustical Society of America*, Vol. 108, 1867-1879, 2000.

[36] Ladefoged, P. *Preliminaries to linguistic phonetics.* Chicago: University of Chicago Press, 1971.

[37] Ladefoged, P. "The linguistic use of different phonation types," In *Vocal fold physiology: Contemporary research and clinical issues* (D. Bless & J. Abbs, editors). San Diego: College Hill Press, 351-360, 1983.

[38] Ladefoged, P. "Discussion of Phonetics: A Note on some terms for Phonation Types," In *Vocal Fold Physiology: voice production* (Osamu Fujimura, editor). New York, NY: Raven Press, 373-375, 1988.

[39] Ladefoged, P. & Maddieson, I. *The Sounds of the World's Languages.* Oxford: Blackwell, 1996.

[40] Local, J.K., Kelly, J. & Wells, W.H.G. "Phonology for conversation: phonetics aspects of turn delimitation in London Jamaican," *Journal of Pragmatics*, Vol. 9, 309-330, 1985.

[41] Laver, R. "Phonatory settings," In *The phonetic description of voice quality.* Cambridge University Press, 1980.

[42] Ostendorf, M., Price, P. & Shattuck-Hufnagel, S. "The Boston University radio news corpus," Boston University ECS Technical Report ECS-95-001, 1995.

[43] Pierrehumbert, J. & Talkin, D. "Lenition of \h\ and glottal stop," In *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (G. Docherty & D.R. Ladd, editors). Cambridge: Cambridge University Press, 90-127, 1992.

[44] Pierrehumbert, J. "Knowledge of variation,", invited talk at CLS 30. In *Papers from the 30th regional meeting of the Chicago Linguistics Society,* Chicago: University of Chicago, 1994.

[45] Pierrehumbert, J. "Prosodic effects on glottal allophones," In *Vocal Fold Physiology: voice quality control* (Osamu Fujimura & Minoru Hirana, editors). San Diego, CA: Singular Publishing Group, 39-60, 1995.

[46] Pierrehumbert, J. & Frisch, S. "Synthesizing allophonic glottalization," In *Progress in speech synthesis* (J.H. van Santen, R.W. Sproat, J. Olive, & J. Hirschberg, editors). New York: Springer, 9-26, 1997.

[47] Redi, L., Hufnagel S.S. "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, Vol. 29, 407-429, 2001.

[48] Rozyspal, A.J. & Millar, B.F. "Perception of jitter and shimmer in synthetic vowels" *Journal of Phonetics*, Vol. 7, 343-355, 1979.

[49] Slifka J. "Respiratory constraints on speech production at prosodic boundaries," Ph.D. dissertation, Massachusetts Institute Of Technology, Cambridge MA, 2000.

[50] Slifka, J. "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of Voice*, Vol. 4, 319, 2005.

[51] Stevens, K. *Acoustic Phonetics.* Cambridge, MA: MIT Press, 1998.

[52] Stevens, K. "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of Acoustical Society of America*, Vol. 111, No. 4, 1872-1891, 2002.

[53] Ma, J.,Zhao, Y. "OSU Support Vector Machines (SVMs) Toolbox" .

[54] "TIMIT Acoustic-Phonetic Continuous Speech Corpus", National Institute of Standards and Technology Speech Disc 1 -1.1, NTIS Order No. PB91 -5050651996, October 1990.

[55] Titze, I.R. & Talkin, D.T. "A theoretical study of the effect of various laryngeal configurations on the acoustics of phonation," *Journal of Acoustical Society of America*, Vol. 66, 60-74. 1979.

[56] Titze, I.R. "Definitions and Nomenclature related to voice quality," In *Vocal Fold Physiology: voice quality control* (Osamu Fujimura & Minoru Hirana, editors). San Diego, CA: Singular Publishing Group, 335-342, 1995.

[57] Umeda, N. "Occurence of glottal stops in fluent speech," *Journal of the Acoustical Society of America*, Vol. 64, No. 1, 88-94, 1978.

[58] Ward, P.H.,Sanders, J.W., Goldman, R. & Moore, G.P. "Diplophonia," *Annals of Otology, Rhinology and Laryngology*, Vol. 78, 771-777, 1969.

[59] Vapnik, V. "The nature of statistical learning theory," Springer Verlag, 1995.

[60] Watkins, J. "Can phonation types be reliably measured from sound spectra? Some data from Wa and Burmese," *SOAS Working Papers in Linguistics and Phonetics*, Vol. 7, 321-339, 1997.

[61] Zemlin, W.R. *Speech and Hearing Science: Anatomy and Physiology.* Eaglewood, NJ: Prentice-Hall, 1988.