Acoustic and Linguistic Interdependencies Of Irregular Phonation

by

Kimberly F. Dietz
B.S., Electrical Engineering
Massachusetts Institute of Technology, 2009

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

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

MAY 2010

Signature of Author.....
Kimberly F. Dietz
Department of Electrical Engineering and Computer Science
May 21, 2010

Certified by.....
Thomas F. Quatieri
Senior Member of Technical Staff; MIT Lincoln Laboratory
VI-A Company Supervisor

Certified by.....
Stefanie Shattuck-Hufnagel
Principal Research Assistant; MIT Research Laboratory of Electronics
M.I.T. Thesis Supervisor

Accepted by.....
Dr. Christopher J. Terman
Senior Lecturer, MIT Department of Electrical Engineering and Computer Science
Chairman, Department of Committee on Graduate Theses

This work was sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002. The opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Acoustic and Linguistic Interdependencies Of Irregular Phonation

by

Kimberly F. Dietz
B.S., Electrical Engineering
Massachusetts Institute of Technology, 2009

Submitted to the Department of Electrical Engineering and Computer Science
On May 21, 2010
In partial fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Irregular phonation is a commonly occurring but only partially understood phenomenon of human speech production. We know properties of irregular phonation can be clues to a speaker's dialect and even identity. We also have evidence that irregular phonation is used as a signal of linguistic and acoustic intent. Nonetheless, there remain fundamental questions about the nature of irregular phonation and the interdependencies of irregular phonation with acoustic and linguistic speech characteristics, as well as the implications of this relationship for speech processing applications.

In this thesis, we hypothesize that irregular phonation occurs naturally in situations with large amounts of change in pitch or power. We therefore focus on investigating parameters such as pitch variance and power variance as well as other measurable properties involving speech dynamics.

In this work, we have investigated the frequency and structure of irregular phonation, the acoustic characteristics of the TIMIT Acoustic-Phonetic Speech Corpus, and relationships between these two groups. We show that characteristics of irregular phonation are positively correlated with several of our potential predictors including pitch and power variance. Finally, we demonstrate that these correlations lead to a model with the potential to predict the occurrence and properties of irregular phonation.

VI-A Company Thesis Supervisor: Thomas F. Quatieri

Title: Senior Member of Technical Staff; MIT Lincoln Laboratory

M.I.T. Thesis Supervisor: Stefanie Shattuck-Hufnagel

Title: Principal Research Assistant; M.I.T Research Laboratory of Electronics

Acknowledgements

I am deeply grateful to my advisors, to the Human Language Technology Group at the MIT Lincoln Laboratory, and to my friends and family. This thesis would not have been possible without their support and encouragement.

Above all, I am indebted to my official and unofficial thesis supervisors. I could not have asked for a more passionate or supportive group to work with. Tom Quatieri, Stefanie Shattuck-Hufnagel, and Nick Malyska have consistently impressed me with their dedication, insight, and knowledge; meeting with them always led to new ideas and inspirations. I would also like to thank Bob Dunn for sharing his insights about pitch-modification with us.

It has been my pleasure to be a part of the community at the MIT Lincoln Laboratory and in the Human Language Technology Group. From seminars and presentations on Laboratory research to salsa lessons and lunchtime concerts, I have enjoyed the many opportunities to explore. Within my office, I would like to thank full- and part-time labmates Tian Wang, Nancy Chen, Zahi Karam, Tom Baran, Dan Rudoy, and Daryush Mehta for an always interesting environment. Daryush, having just completed his PhD thesis, was several times the reassuring voice of experience. I would also like to thank Doug Jones, Joel Acevedo-Aviles, Linda Kukolich, Yi-Hsin Lin, and Amy Englehart for the great lunchtime conversations and camaraderie.

To my Undergraduate Advisor who continued to offer his support even in my M.Eng. year, George Verghese …

To Anne Hunter, Niki Huhn, and Brandon Moore, who provided a sounding board when I was stressed …

To Jo-Ann Graziano, for last minute editing ...

To my family – Mom, Steph, Abby, and Matt – who have always supported me …

To my Dad ...

To Kevin, who keeps me focused …

Thank you! I would not have made it without you.

And finally, I have thanked him already, but one credit is not enough. Thank you wholeheartedly to Tom Quatieri, for his incredible dedication to his students and the encouragement he has given me throughout this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When people speak, their vocal folds vibrate, sending pulses of air through the vocal tract and shaping them into the sounds we recognize. The vocal folds generally vibrate at a regular rate, generating something we hear and call pitch (or measure and call average fundamental frequency). However, the vocal folds can, and often do, vibrate irregularly. This may mean an irregular spacing of pulses, irregular pulse heights, or vibrating at a frequency so low that individual pulses are distinguishable to the human ear. When the vocal folds vibrate irregularly in any of these ways, the result is termed *irregular speech*. 'Irregular' here does not refer to the spacing between the pulses or their heights, but to the fact that the pattern produced is substantially different from the one produced under most circumstances.

1.1 Problem Statement and Motivation

The ability to predict (1) *the occurrence of irregular phonation* and (2) *its properties* would lead to important advances in the analysis, modification, and synthesis of speech. Current methods of *analyzing* speech have difficulty when they encounter sections of speech that are irregularly phonated; they are unable to cope using standard models. Without knowing the characteristics of irregular phonation, realistic *modification* (in terms of pitch, speed or identity) of speech is hindered. *Synthesizing* realistic speech also depends on an inclusion of this phenomenon – without including episodes of irregular phonation, synthetic speech feels unnatural.

1.2 Hypothesis

We posit that the relationship to the variation of pitch and power is due largely to the physical constraints on the speech apparatus. Slifka has found that 'the physical reality of managing vocal fold vibration' leads to an association between irregular phonation and silence within speech [1-2]. In this thesis, we hypothesize that these same 'physical realit[ies]' are responsible for the association between irregular phonation and variation in pitch and power. When pitch and power vary, rapid reconfiguration of parameters which affect its production is necessary, including the spacing and tension of the vocal folds, the level of contraction of the surrounding muscles, and the rotation and translation of the arytenoid cartilages. During each set of adjustments, the probability of producing irregular phonation is increased. This results in a stochastic increase in episodes of irregular phonation.

The relationship to dialect and gender (in at least English) can be attributed to the allophonic nature of words pronounced with and without irregular phonation. Because of this, a manner of speaking which includes more or less irregular phonation can be adopted by a group. In the case of gender, physical differences in average length and mass of the vocal folds can play a role in the extent and nature of irregular phonation as well.

Linguistic dependencies of irregular phonation have also been found previously. We propose that something similar to Slifka's "preferred location" for irregular phonation near silence can aid us in understanding the connection between linguistic uses of irregular phonation and measurable properties of speech such as pitch and power variance. Linguistic use of irregular phonation as a signal has two sources: irregular phonation is 'available' to hold extra information (since words with and without irregular phonation are allophonic in English) and irregular phonation has tacitly understood correlates in terms of variance as well as in terms of silence.

1.3 Methodology

We address the challenge of irregular phonation description and classification on a large scale, using a database of 530 speakers, 151 of whom are fully phonetically pre-labeled with information about irregular phonation. We draw on pre-existing tools for analysis. These include a sinewave-based pitch and voicing extraction tool [3], an implementation of the Maximum Entropy Deconvolution (MED) for pulse location [4], and a fusion method which combines several pulse location methods for increased accuracy. We use these tools to gain information about the frequency and nature of irregular phonation and its co-incidence with other defining characteristics of speech: measurable acoustic properties, timing and location, and speaker demographics. Overall we note a relationship between irregular phonation, the variation of pitch and power, dialect, and gender.

1.4 Summary of Contributions

This thesis provides results in two distinct areas: description and prediction. We begin by *describing* the TIMIT Acoustic-Phonetic Continuous Speech Corpus in terms of average pitch, pitch variance, and occurrence of irregular phonation. We form a hypothesis that places of pitch and power variation form a "preferred location" for irregular phonation, akin to Slifka's descriptions of regions of silence as "preferred locations for irregular phonation. In the realm of *prediction*, we show that short- and long-term acoustic properties of speech are predictive of irregular phonation, as is demographic information. We report correlations for individual relationship as well as a model for prediction of the occurrence of irregular phonation at the sentence level, based on acoustic and demographic properties.

1.5 Thesis Outline

This thesis is organized as follows. In Chapter 2 we provide background on the description, and detection of irregular phonation. In Chapter 3 we discuss our choice of corpus and characterize its pitch and the frequency of occurrence of irregular phonation within it. In Chapter 4 we present work related to the frequency of occurrence of irregular phonation. In Chapter 5 we discuss the small-scale properties of irregular phonation. In Chapter 6 we use relationships described in

previous chapters to build a model which predicts the occurrence of irregular phonation and its average inter-pulse spacing. We discuss the implications of relative coefficient magnitudes and measures of significance in this model. Finally, in Chapter 7 we conclude and offer suggestions for future work.

Chapter 2

Background: Speech Production and Irregular Phonation

In Chapter 2, we provide a brief overview of the complexity of the speech production system and the source-filter system often used to describe it more simply. We then describe and define irregular phonation. We discuss known correlates to irregular phonation: acoustic, demographic, and linguistic. These findings give us insight into whether, at its root, its production is physically or linguistically motivated. We discuss the notion of a 'preferred location' in linguistics and put forth the notion that pitch and power variance, in addition to silence, may be such a location.

2.1 Speech Production

2.1.1 True Speech Production

Speech production is a complex process, involving physical systems located from the midsection upwards. Moving from the bottom up, the system can be divided into subglottal, laryngeal, and farther forward subsystems.

The subglottal subsystem, including the abdominal muscles, intercostals, diaphragm, lungs, branching airways, and trachea provide a variable supply of pressurized air to the glottis. The larynx, further subdivided into the vocal folds, supporting structures for the vocal folds, extrinsic laryngeal muscles, and laryngeal structures above the vocal folds, produces vibration in response to the air flowing through from below. The properties of the vocal folds (such as tension, position, and mass per unit length) can effectively be modified by adjustments of the surrounding muscles and cartilages. The vocal tract above the larynx is characterized by the pharynx, soft palate, nasal cavity, oral cavity, tongue, mandible, lips, the mechanical properties of the wall of the vocal tract, and the overall length and volume of the vocal tract [5].

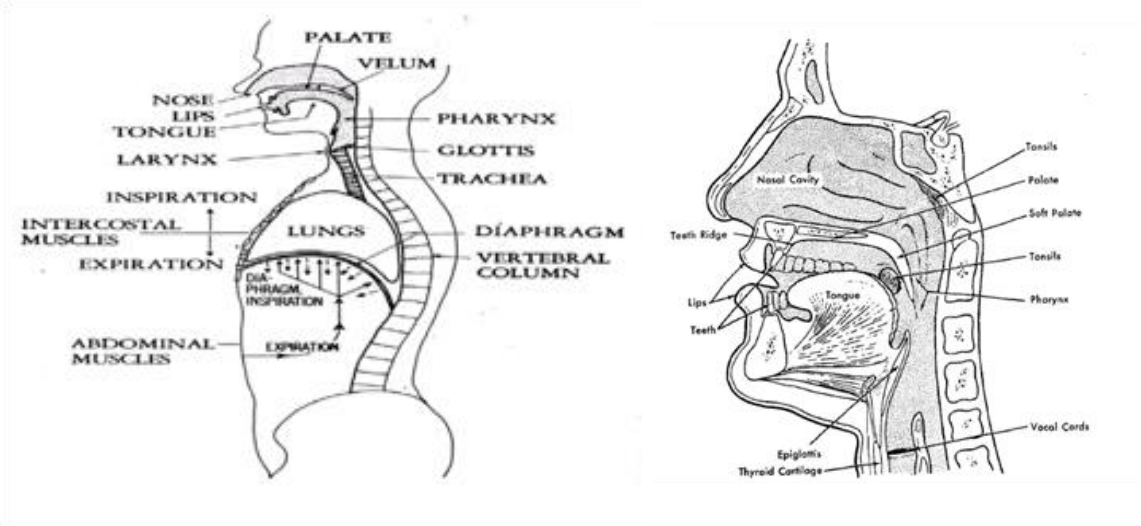

Figure 1. Composite diagram of speech system components. Speech system schematic by Stevens (Fig. 1.2) [5]. Cross-sectional view of the vocal tract by Coker, Denes, and Pinson (Fig.4.7) [6].

2.1.2 Modeled Speech Production

Characterizing all of the properties of the speech production system and emulating them would be a difficult task. The interactions are many and nonlinear in real life. For practical purposes, we lump many components together and use a linear source-filter model to understand speech. We essentially divide this complex system into two key components and focus on their interaction.

Gunnar Fant developed the canonical source-filter model of speech production. [7]. His model states that speech is made up of two key components. The first is the generation of sound sources. Air is pushed from the lungs to the vocal folds, which vibrate and form it into pulses – the *source*. The second component of speech generation is shaping by the vocal tract. The sound source is encounters the pharynx, tongue, teeth, lips, and nasal passages. The specific path available to it depends both on both permanent and temporary physical parameters (such as the size of their vocal tract and the instantaneous position of the tongue). Its movement through the vocal tract shapes the sound source into its final form. The *system* formed by the physical space the source moves through modifies the source. We describe both system and source mathematically, and are therefore able to describe source-system interactions with precision. A diagram representing key components of this model appears in **Figure 2**.

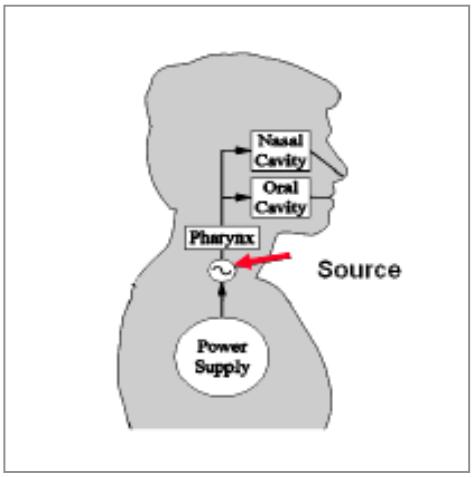

Figure 2. Key components of the source-filter model of speech production (adapted from Fig. 1-1, Malyska [8]).

Specifically, the source-filter model describes the signal as the sum of two pieces, a voicing source (air driven by the vocal folds) and a noise source (air which moves through the vocal tract without causing the vocal folds to vibrate). The system is modeled as a time-varying filter with terms modeling vocal tract resonances and the effect of lip radiation (as described by acoustic phonetic theory). A diagram of the source-filter model is shown in Figure 3. This model is useful as a conceptual framework for the understanding of speech production. It also leads to effective mathematical manipulation of the system. One advantage of this model is that it partitions our task. Rather than constantly accounting for dozens of factors, we model only two components.

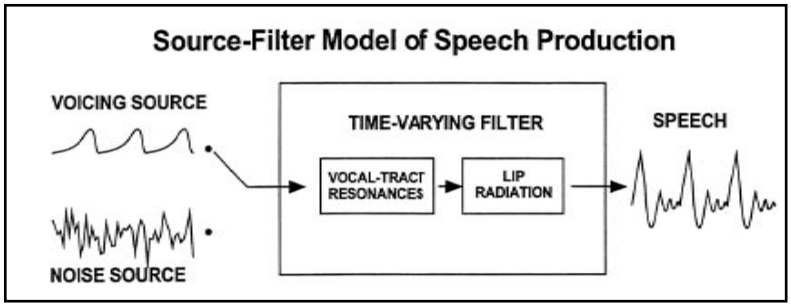

Figure 3. The source-filter model of speech generation [9].

2.2 Descriptions of Irregular Phonation – Physical, Perceptual, and Acoustical

2.2.1 Physical Descriptions

In reality, fine-level details of speech are not well-modeled by the source-filter system. The behavior of the speech production system depends on the interplay between many factors, including the amount of tension in the vocal folds, their position (pressed tightly together, held far apart, or somewhere in between), and transglottal pressure. Different configurations of the

system can lead not only to *modal* voiced speech and *unvoiced* speech (the "regular" modes), but also to irregularly voiced speech or speech which otherwise does not sound "regular". It is the speech we hear when the vocal folds are vibrating with a quasi-regular pitch period.

Irregular phonation is caused by the interplay of various physiological factors. Important factors include the amount of tension in the vocal folds, their position (pressed tightly together, held far apart, or somewhere in between), and transglottal pressure. Janet Slifka has given evidence that these factors tend to naturally align in favor of irregular phonation at word and syllable boundaries in spoken speech. She identifies two methods of producing irregular phonation, whether intentionally or incidentally: a speaker can either tightly adduct their vocal folds or abduct them. As Slifka uses our definition of irregular phonation, her findings are directly applicable to our research [1].

The fundamental frequency of irregular phonation has been described and measured in previous papers. Hollien and Wendahl [10] describe a phenomenon very similar to our 'irregular phonation' which they term *vocal fry*. They say that "[v]ocal fry (1) is a normal mode of laryngeal production; (2) it consists of a register of very low fundamental frequencies, and (3) it consists of a train of relatively discrete laryngeal pulses with nearly complete damping between successive glottal excitations." Hollien and Michel [11] further study whether or not vocal fry constitutes a phonational register, defined as "a series or range of consecutive (vocal) fundamental frequencies of similar quality; in addition, there should be little or no overlap in fundamental frequency between adjacent registers." They use a corpus which includes 12 males and 11 females. Their relevant findings are summarized in **Table 1**. Their "group range" indicates smallest and largest quantities measured across the group; their "mean range" averages the smallest and largest quantities of each speaker. We see that males and females have an extremely similar fundamental frequency range while producing irregular phonation, despite the fact that they have much more divergent ranges during modal phonation.

	Males	**Females**
Group Range of Irregular Region	7 – 78 Hz	2 – 78 Hz
Mean Range of Irregular Region	24 – 52 Hz	18 – 46 Hz
Group Range of Modal Region	71 – 561 Hz	122 – 798 Hz
Mean Range of Modal Region	94 – 287 Hz	144 – 538 Hz

Table 1. Result of a study on fundamental frequency of distinct registers.

2.2.2 Perceptual Descriptions

A litany of buzzwords have sprung up around irregular phonation and related phenomena, many of them perceptual. While these are useful for building intuition about the phenomenon, the pure number of words used can be confusing. Episodes of irregular phonation are often described as sounding 'rough' [12], 'creaky' [8], or similar to a "… rapid series of taps, like a stick being run

along a railing" [13]. Perhaps the most telling of these terms is "creaky." Irregular phonation, at a basic perceptual level, can be thought of as a brief, creaky sound that happens in speech.

Basic elements of perceptual voice analysis used in the setting of voice-disorder clinics include roughness, breathiness, hoarseness, and strain. These perceptual labels are used to aid physicians in the work of voice study and therapy [14-15]. To understand why a voice might sound rough, breathy, hoarse, or strained, we need to look further into how speech is produced. A perception of roughness translates acoustically to irregular pitch periods, and therefore irregular phonation. Breathiness constitutes an extra amount of air moving through the vocal tract without vibration. Breathiness alone does not make speech irregular by our definition. Hoarseness is a combination of roughness and breathiness, and is therefore irregular phonation. Finally, a perception of strain occurs when the vocal folds are pressed tightly together.

2.2.3 Acoustical

There is some tolerance in the perceptual definitions of "regular" speech, thus it is sometimes termed "quasi-regular". Titze categorized acceptable variation during regular speech as follows [16]:

- 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"

Irregular voiced speech occurs when the vocal folds are vibrating, but with more extreme irregularity than described as 'acceptable variation' above. Unvoiced speech occurs without vibration of the vocal folds, as in pronunciation of many consonants and during whispering.

An example of irregular phonation at the end of an utterance is shown in **Figure 4**. The irregular phonation was labeled (by Surana [17]) as beginning at 3.011 seconds and lasting until 3.072 seconds. We see from the figure "quasi-regular" spacing for the first 10 pulses, followed by irregular phonation for the remainder of the utterance.

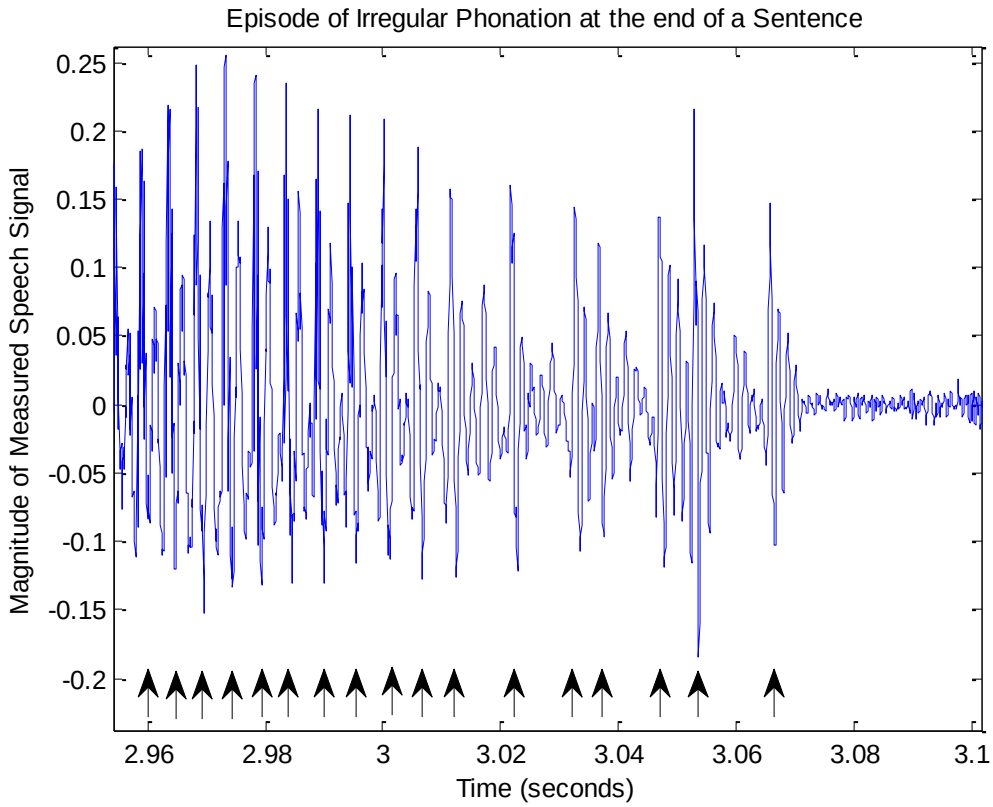

Figure 4. Example of irregular phonation. Taken from the end of sentence SA1 by speaker fjsp0 in TIMIT. Arrows indicate approximate locations of glottal pulses. Note that, between the large pulses of the region of irregular phonation, there are what appear to be smaller pulses. These may or may not register as pulses through the use of various automatic methods of pulse detection. Even if they register as pulses, this sample will be classed as irregular due to the uneven pulse amplitude.

2.2.4 For this Thesis

For the purposes of this thesis, we deal with *irregular phonation* as described by Surana in his thesis [17]. Irregular phonation includes phonation with a noticeably different quality than the speaker normally uses. This includes not only irregularly voiced phonation, but also phonation with an unusually low fundamental frequency, such that it sounds "like a series of taps. [13]"

The definition presented in Surana's 2006 Master's thesis and which we also adopt for use in this thesis is:

> "*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.*"

2.3 Known Correlated Properties of Speech

2.3.1 Speaker, Dialect, and Gender Dependence

Speaker Dependence of Irregular Phonation

The amount of irregular phonation detected in an individual has been shown to be highly variable. Some speakers speak "with almost continuous creak," [18], some have been unable to produce it at all [11], and most fall along a spectrum between these extremes.

Furthermore, acquaintances of subjects have been shown to perceive these differences in production of irregular phonation and to use them in speaker identification tasks. In one study, researchers recorded speech samples from subjects. They then manipulated the short segments of the recorded utterances, creating and removing episodes of irregular phonation. Acquaintances of the subjects listened to the original and altered versions, attempting to determine which sounds most like a sample of the subject's voice. These acquaintances consistently chose the original sentence [19]. In another series of experiments, timing and amplitude features derived from glottal events were shown to carry speaker-dependent information which was successfully used to improve the accuracy of results obtained with a leading speaker identification algorithms [20].

Dialect Dependence of Irregular Phonation

Some dialects of similar languages have in the past been shown to be distinguishable in part on the basis of how "creaky" they are. This is, for instance, the case in British English [21]. Due to the fact that irregularly phonated words are allomorphic in most languages, irregular phonation is 'freer' to vary across populations than it otherwise would be.

Gender Dependence of Irregular Phonation

Gender has often been posited as an indicator of the propensity of a speaker for irregular phonation. In general, it has been assumed that males would be more prone to irregular phonation [21], due to the fact that their regular phonation is characterized by a lower fundamental frequency than that of females. However, in actuality which gender produces more irregular phonation in studies is definition dependent. According to one study which defines four voice types – 'creaky voice,' 'creak,' 'glottalization,' and 'diplophonia,' where we would distinguish only 'irregular phonation,' there are differences in phonation patterns based upon gender. Males produce more 'creaky voice' than females, whereas females produce more 'creak' than males. Levels of 'glottalization' and 'diplophonia' are similar across genders [22]. While we are not concerned here with the precise definitions used in this work to characterize irregular phonation, we note that the nature of irregular phonation has been shown to differ across groups of subjects (e.g. subjects grouped by gender).

2.3.2 Linguistic Dependence – Structure and Intent

The appearance of irregular phonation has been found to be related to the linguistic structure of speech, in the form of grammatically significant locations such as pitch accents and intonational phrase boundaries, pragmatic structure, and regions of silence. While physical constraints may partially explain these correlations, linguistic intent is also at play in at least some instances.

Dilley, et. al. [23] and Pierrehumbert [24] have found that, "glottalization [roughly, irregular phonation] of word-initial vowels is influenced by full vs. intermediate intonational phrase boundaries and pitch accent on the target syllable or word." Intonational phrase boundaries are perceptually defined boundaries between segments of speech with a single pitch and rhythm contour. This research demonstrates that the speaker's choice of intonational phrase boundaries significantly affects whether or not irregular phonation occurs during word-initial vowels. Specifically, they found that at the beginning of a full intonational phrase irregular phonation occurs on both full and reduced vowels, while on an intermediate intonational phrase irregular phonation occurs on only full vowels. This research suggests that irregular phonation appears differentially in grammatically significant locations.

A 2004 study by Grivičić and Nilep [25] similarly suggests that irregular phonation is a cue to pragmatic understanding of dialog. This study examines the relationship between using the word 'yeah,' irregular phonation, and speakers' 'turns' in a conversation. The study found that people who said 'yeah' with regular phonation continued to speak afterward 60 percent of the time. However, people who said 'yeah' with irregular phonation only continued to speak afterward 20 percent of the time. This implies that speakers who want to give someone else a turn can proactively initiate a change in their speech quality. The fact that the other speakers began to speak after hearing their partners produce these episodes of irregular phonation implies that they subconsciously understood the signal.

Slifka [1] has noted that irregular phonation in speech can be both prompted by physical realities and used as a linguistic cue. Furthermore, she has shown links between the two. Specifically, she has posited that irregular phonation has a preferred function for cueing silence, a role related to the cueing of boundaries and hence a linguistic cue. She has shown that this cue is not randomly chosen but rather a natural outcome of the positioning of unplanned irregular phonation. According to her work, as silence is approached or passed, the configuration of physiological factors around the glottis is such that the unplanned occurrence of irregular phonation becomes more likely. This, coupled with the fact that speakers are able to artificially increase the probability of the production of irregular phonation by tightening (or loosening) the vocal folds, leads to the use of irregular phonation as a signal of silence in many languages.

We believe that the fact that irregular phonation may be used to cue not only silence, but also other naturally co-occurring circumstances. Specifically, we believe that Slifka's theory may apply to variance in pitch and power as well as silence.

2.4 Summary

The speech system is an extraordinarily complex system, much more so than the models with which we usually describe it. Irregular phonation in particular is not well-suited to description and manipulation under current methods. We described irregular phonation both as "a series of taps" and as variations from the normal speech pattern above an acceptable "small-scale" level. We discussed the way in which this physical phenomenon has been adopted for linguistic use and assert that pitch and power variance are potentially "preferred location[s]" for irregular phonation.

Chapter 3

The Speech Corpus: Choice and Characterization

The choice of a speech corpus is an important one which shapes the scope and nature of the work. Here, we discuss our choice of the TIMIT Acoustic-Phonetic Continuous Speech Corpus and characterize the speech found in the database in terms of pitch and frequency of irregular phonation.

3.1 Choice of Corpus - TIMIT

In choosing our corpus, we considered the amount of recorded material, the number of unique speakers, whether the speech was spontaneous or read, and what labeling had been performed (prosodic, IP (irregular phonation), and phonetic labels were desired). Our process is shown in Table 2 to aid others in choosing between these corpora in the future. The corpora are listed in ascending size. With ascending size tends to come descending labeling, due to the amount of time required to label speech. We chose TIMIT because of its substantial size advantage. Unfortunately, it is not labeled prosodically, which limits us to acoustic and phonetic deductions about irregular phonation.

TIMIT was developed as a joint effort between the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). The work was sponsored by the Defense Advanced Research Projects Agency – Information Science and Technology Office (DARPA-ISTO), in order to provide a large corpus of speech for use in developing acoustic-phonetic knowledge and developing and evaluating automatic speech recognition systems [26]. The database consists of 530 speakers who speak 10 sentences each. The sentences are chosen such that two are common to all speakers in the database, 450 are spoken by 7 speakers in the database, and 1890 appear are spoken only once. The two commonly spoken sentences are designed as dialect *shibboleths*, meant to reveal differences between the speakers based on their geographic location. The 450 multi-speaker sentences are phonetically-compact, designed to 'provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest.' The single-appearance sentences are selected from previous text corpuses and meant to diversify the content of the sentences.

Corpus	Size (Time)	Num. Speakers	Spontaneous?	Prosodic Labels?	IP Labels?	Phonetic Labels?
Maptask	12 minutes	4	Yes	Yes*	Yes*	Yes*
BU FM	6 hours	6	Read	Yes (?)	No	Yes (?)
TIMIT	50+ hours	630	Read in lab	No	Yes**	Yes
Switchboard, Callhome	Huge	Lots	Yes	No (very few)	No	Yes (2 hours)

Table 2. Comparison of corpora. A single asterisk indicates that the data needs a small amount of hand-correction to be usable. A pair of asterisks means that labeled data does not need corrections, but not all data is labeled. In the TIMIT database, 151 of the 530 speakers are fully labeled for irregular phonation.

A deciding factor in our use of TIMIT was that much of it has been labeled for irregular phonation. Further, the labeling was performed by Surana according to the definition of irregular phonation that we use here [17]. Therefore, one source of error (inconsistent definitions of IP) is removed from contention. Surana's labeling was automatically performed by a method developed by Surana, then reviewed and hand-corrected to ascertain the efficacy of the method and to ensure accurate results. The TIMIT database is divided into dialect regions of the United States. Two of these dialect regions, 'New England' and 'Northern' were fully labeled by Surana and are used throughout the thesis. They are also referred to in this thesis as dialect region 1 and dialect region 2, respectively. The remaining regions, 'North Midland,' 'South Midland,' 'Southern,' 'New York City,' 'Western,' and 'Army Brat' are used for large-scale database characterizations but not analyses of irregular phonation.

3.2 Characterization of Corpus – Pitch, Variance, Irregular Phonation

Before seeking to predict regular or irregular phonation, we seek a basic understanding of the underlying distribution. What is the average pitch of a male or female speaker? Are male and female pitch roughly normally distributed? If so, with what variance? What percentage of sentences in the general population include irregular phonation?

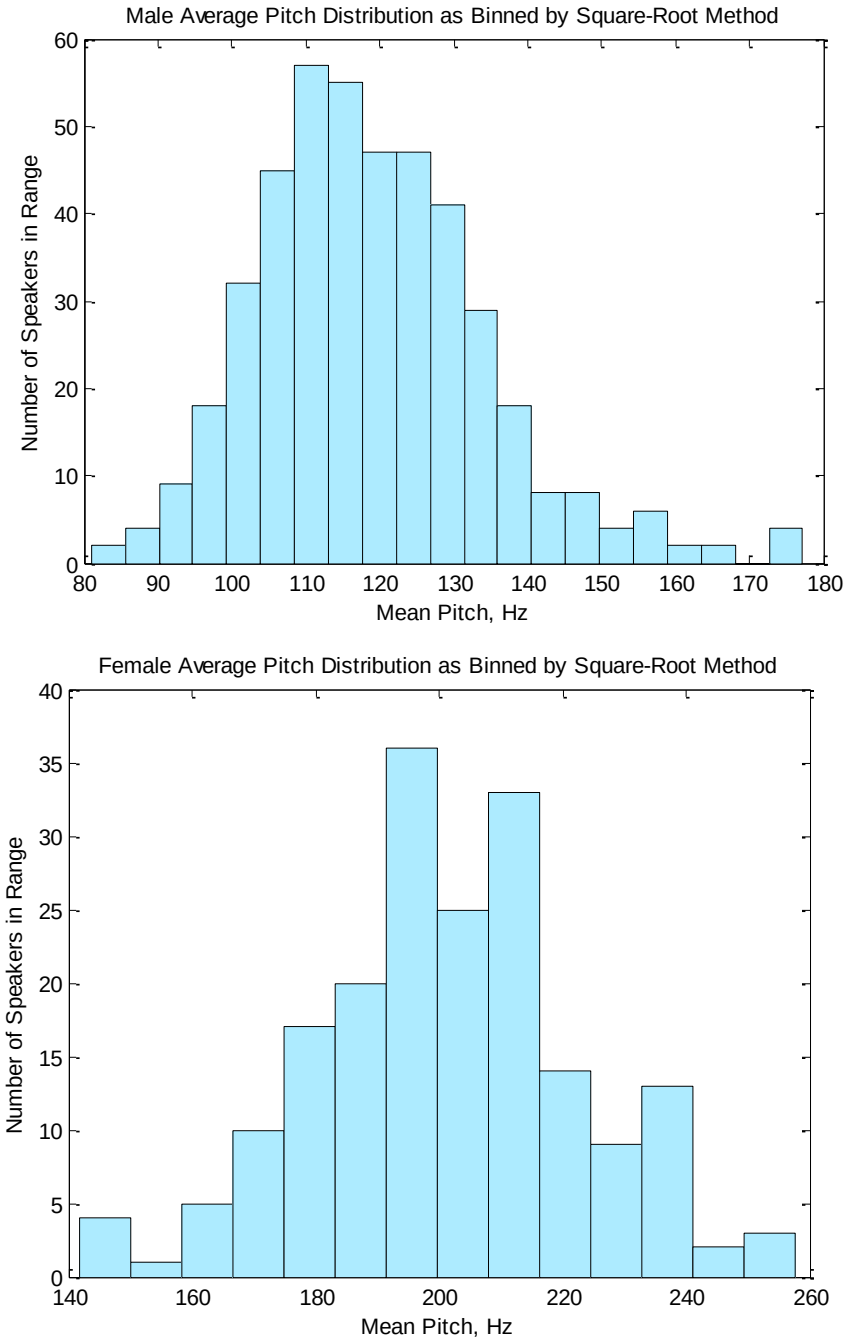

Figure 5. Distribution of average pitch for males (upper) and females (lower) in TIMIT. For each of the 438 male speakers and 192 female speakers in the database, 10 sentences were available. We used our pitch and voicing extraction tool to calculate an estimate of the pitch and whether the speech was voiced for each sample (16k sampling rate). We use a voiced cutoff of 0.75 (out of a maximum score of 1) to determine whether a frame is voiced. We take the mean of the pitch measurements for all voiced points during the 10 sentences. This leads to one average pitch data point per speaker. In these histograms, binning – for which there is no agreed-upon optimal algorithm – is done according to the square-root method, whereby the number of bins in the histogram is equal to the square root of the number of samples.

3.2.1 Distribution of Pitch

It has been widely noted in the literature [20] that the mean pitch of speech is related to the voice quality – a term describing perceptual characteristics such as 'rough' or 'creaky' – of the speech. Therefore, our investigation of pitch characteristics has a twofold benefit, both describing regular speech and providing a clue to the occurrence of irregular speech.

We use an in-house sinewave-based pitch and voicing extraction tool [3] to determine these parameters. In this way, we have calculated an average pitch for each of the 530 speakers in the corpus. **Figure 5** displays histograms of the results for males and females.

We fit normal distributions to the data, and the mean and variance calculated are shown in **Table 3**. Standard deviation is also determined from variance and shown in this table. Standard deviation lends itself to more intuition, as its units of Hz are easier to grasp and manipulate internally. The fit of the data is shown graphically in **Figure 6** and appears to be well approximated by the normal distribution.

	Males	**Females**
Mean (Hz)	119.1	201.4
Variance (Hz^2)	242.0	469.4
Standard Deviation (Hz)	15.6	21.7

Table 3. Average pitch distribution in TIMIT.

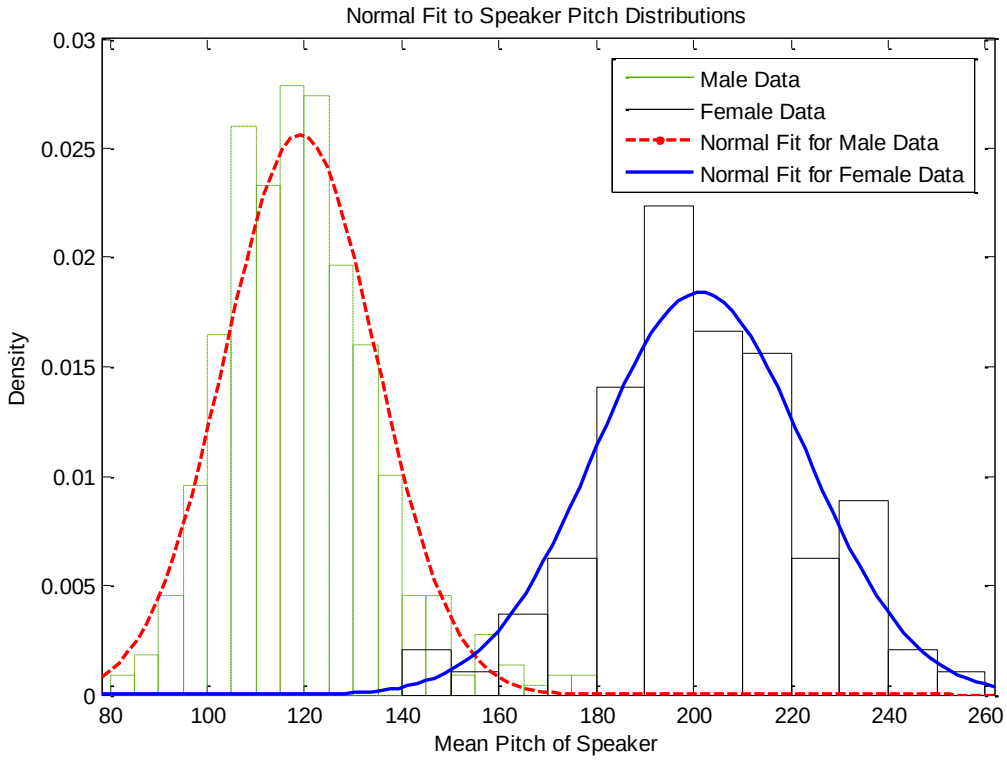

Figure 6. Normal distribution fitted to male and female mean pitch data.

3.2.2 Distribution of Variance

Next, we seek to characterize the variation in pitch found in our database. Irregular phonation has been found to have a 'preferred role' as an indicator of silence, due to 'the physical reality of managing vocal fold vibration. [1]' We believe that similar physical realities lead logically to the association of highly variable pitch with irregular phonation. The more quickly adjustments are made to the vocal folds and their surrounding supports and muscles, the greater the chance that the vocal folds will be, however briefly, in a position which makes it difficult to sustain regular phonation. Therefore, we quantify the variation in pitch found in our corpus.

Variance, as a unit, can be difficult to grasp. To give a sense of the range of variance in our data, we present two example sentences in **Figure 7**. Both are versions of the sentence with the text "She had your dark suit in greasy wash water all year." Each is produced by a different female speaker from the New England dialect region. The sentence which has a smoother pitch contour is measured to have a variance of only 515 Hz^2, whereas the distribution which shows more pitch movement is measured to have a variance of 1776 Hz^2. The dark points in the plot indicate data which was used in this calculation. The light points in the plot indicate the pitch measurements which were not used due to the fact that the 'voicing measure' – a predictor of whether a pitch can reasonably be calculated for a given sample – was too low (we used a cutoff of 0.75). The voicing measure prevents us from using most of the unreasonably high or low pitch points in our calculations. Nonetheless, some of these points are included. Luckily, their impact is low – for the majority of the TIMIT sentences, there are 30 to 60 *thousand* samples taken. A few samples which are off will not have a substantial impact on our calculations.

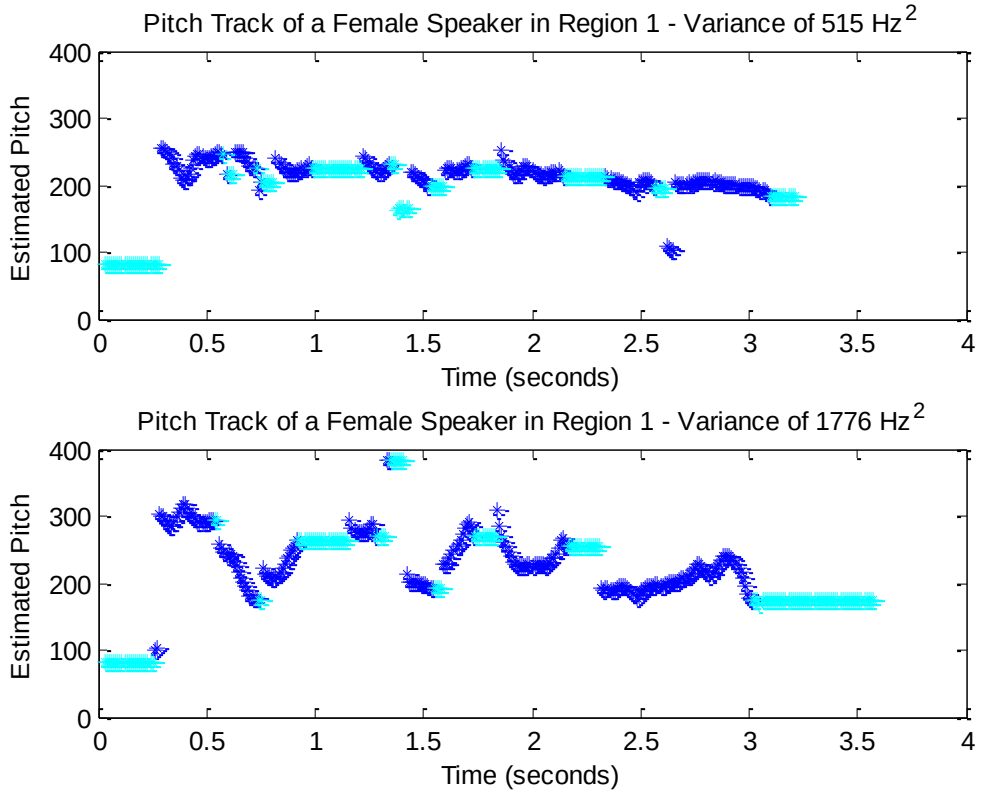

Figure 7. Variance in Pitch Over a Single Sentence - Comparison Between Two Speakers. (The second speaker is the same speaker used to generate the example or irregular phonation shown in **Figure 4**).

In Figure 8 we see the distribution of variance measured in the TIMIT population. We fit normal distributions to these male and female variance distributions. We report our results in **Table 4** and display them graphically in **Figure 9**. The normal distribution fits less well in this case.

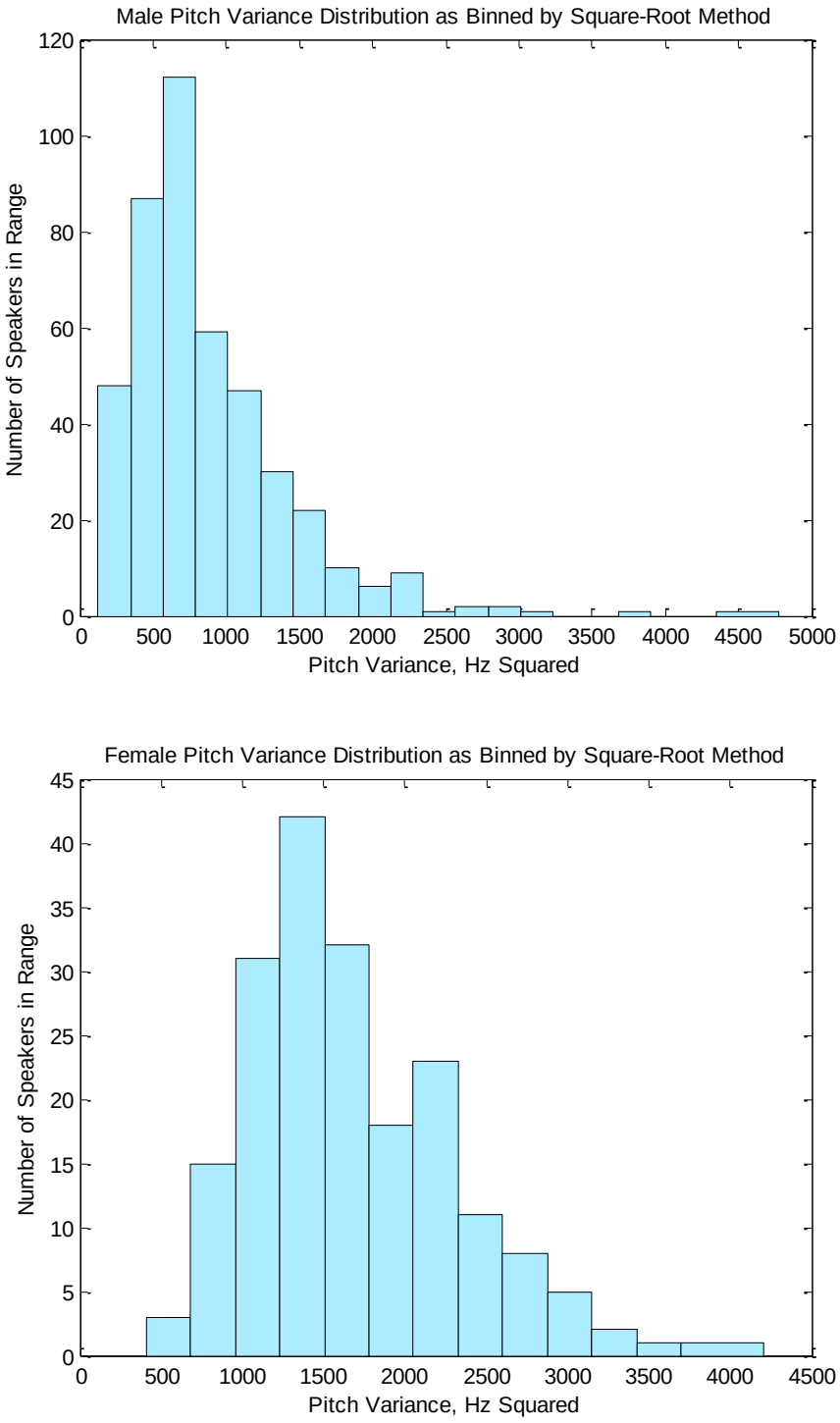

Figure 8. Distribution of variance for males (upper) and females (lower) in TIMIT. For each of the 438 male speakers and 192 female speakers in the database, estimates of the pitch during only voiced regions of speech are obtained as described in Figure 5. For each speaker, we take the variance of the pitch measurements for all samples determined to be voiced in the 10 sentences. This leads to one pitch variance data point per speaker. In this histogram, binning – for which there is no agreed-upon optimal algorithm – is done according to the square-root method, whereby the number of bins in the histogram is equal to the square root of the number of samples.

	Males	Females
Mean (Hz)	119.1	201.4
Variance (Hz^2)	242.0	469.4
Standard Deviation (Hz)	15.6	21.7

Table 4. Pitch variance distribution in TIMIT.

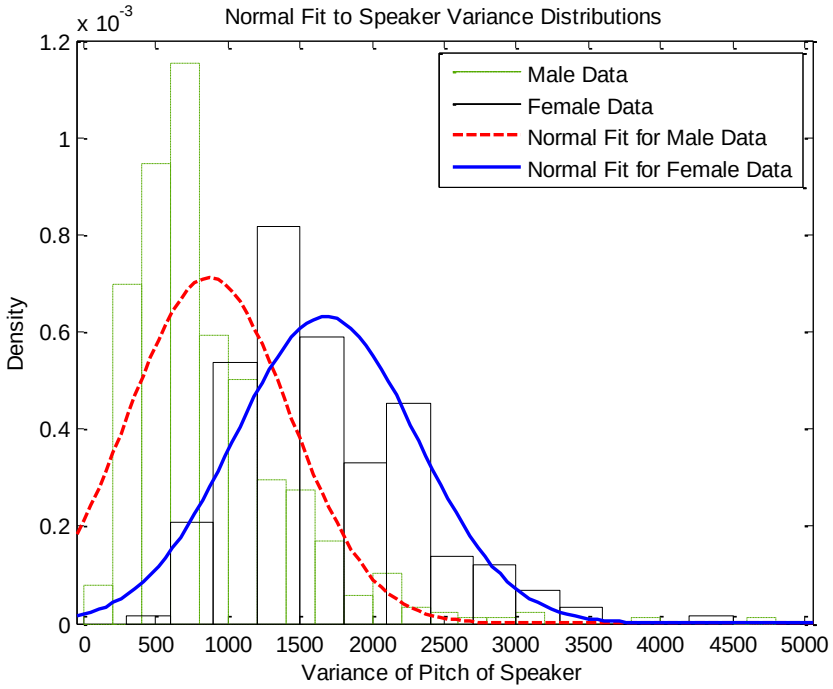

Figure 9. Normal distribution fitted to male and female variance data.

3.2.3 Overall Frequency of Irregular Phonation

Statistics have been reported on the frequency of irregular phonation in various populations. Generally these statistics do not cover a large group of individuals. Therefore, it is of interest to characterize our database in terms of overall propensity to irregular phonation.

Based on the phonetic labeling of the TIMIT database and the extension of labeling of irregular phonation in regions 1 and 2 by Surana, [17] we were able to calculate the amount of time spent on production of irregular phonation, as a percentage of total time spent speaking. Results for regions 1 and 2 of the TIMIT database are shown in **Figure 15**. As seen there, females were observed to spend more time on production of irregular phonation than were males.

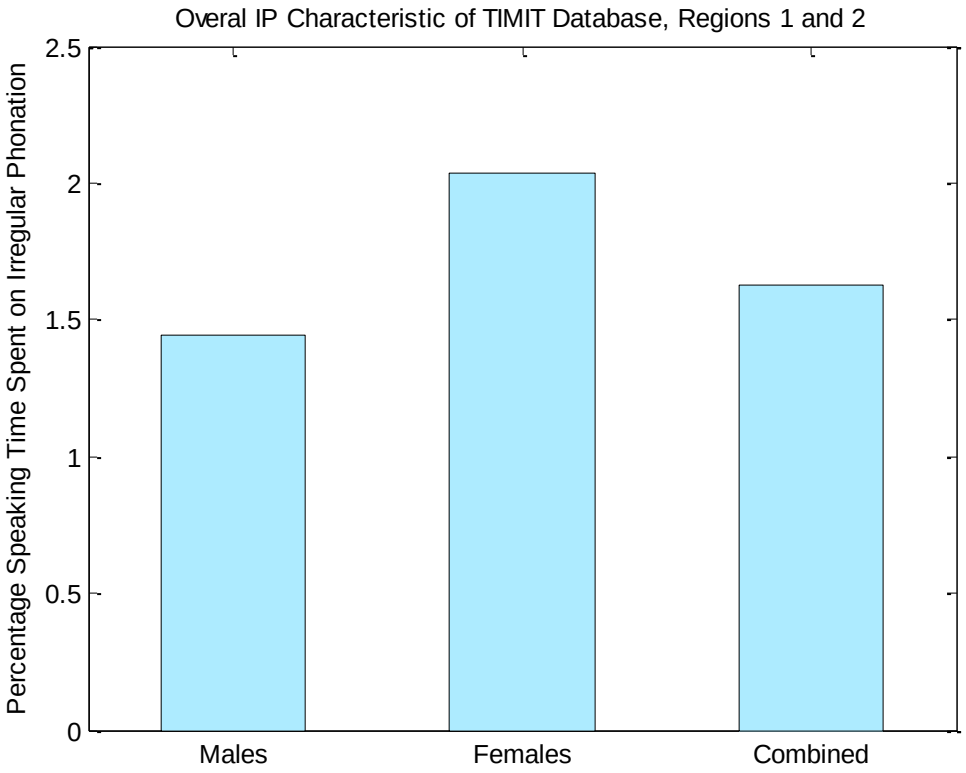

Figure 10. Percentage time spent on irregular phonation. This calculation uses dialect regions 1 and 2 of the TIMIT database, due to the fact that they are fully labeled for occurrences of irregular phonation.

3.3 Summary

We chose to use the TIMIT Acoustic-Phonetic Continuous Speech Corpus in this work. We did so to maximize amount of data we had to work with while still maintaining the advantage of labeling. We found that the mean pitch for males and for females was roughly normally distributed, while the pitch variance for the genders was not. We found that irregular phonation occurred more frequently in female subjects' speech than in that of males.

Chapter 4

The Frequency of Irregular Phonation: Dependence on Prosody

In Chapter 4 we probe the correlation between irregular phonation and readily measured aspects of prosody. We know that physical acts and circumstances (such as pulling the vocal folds apart for lax irregular phonation or pushing them together for *tense* irregular phonation) underlie the phenomenon. The physical requirements for irregular phonation may be met more frequently as people speak with different pitch and power patterns. Absolute pitch and power may be important because speaking in certain ranges of either may cause tensioning and placement of the vocal folds to require more or less precision. Variance in pitch and power is likely important because it requires adjustments. Whenever the configuration in the larynx is changing, its many components must simultaneously adjust, and on sum may pass through a state in which the vocal folds are not able to vibrate regularly. With these possibilities in mind, we investigate the relationship between irregular phonation and prosody. We begin by examining the effect of pitch (both mean and variance). We then conduct parallel experiments with power. Finally, we present a case study of irregular phonation which occurs at the end of an utterance (known as final creak.)

4.1 Average Pitch and Pitch Variance

It has been widely noted in the literature that the mean pitch of speech is related to the voice quality of the speech [20]. We probe this relationship with studies of how mean pitch and variance of pitch are related to the amount of speaker time spent on irregular phonation.

For each subject in district regions 1 and 2 of TIMIT, we calculated their average pitch during voiced segments (determined using a cutoff value of 0.75, as described in Chapter 2.) Using the same pitch points, we also calculated the variance of pitch among these speakers. We used all 10 sentences for each subject to determine their mean pitch and pitch variance, giving us approximately 30 seconds worth of data per subject. Using the prosodic labeling of irregular phonation provided by Surana, we computed the amount of time spent on irregular phonation and the total amount of time spent speaking for each speaker. We then compare these to average pitch and pitch variance in turn. When we make these comparisons, we specify their strength and validity using the Pearson product-moment correlation coefficient (the r-value) and the probability of obtaining results at least as extreme, assuming that correlation does not exist (the p-value). Our results are shown below.

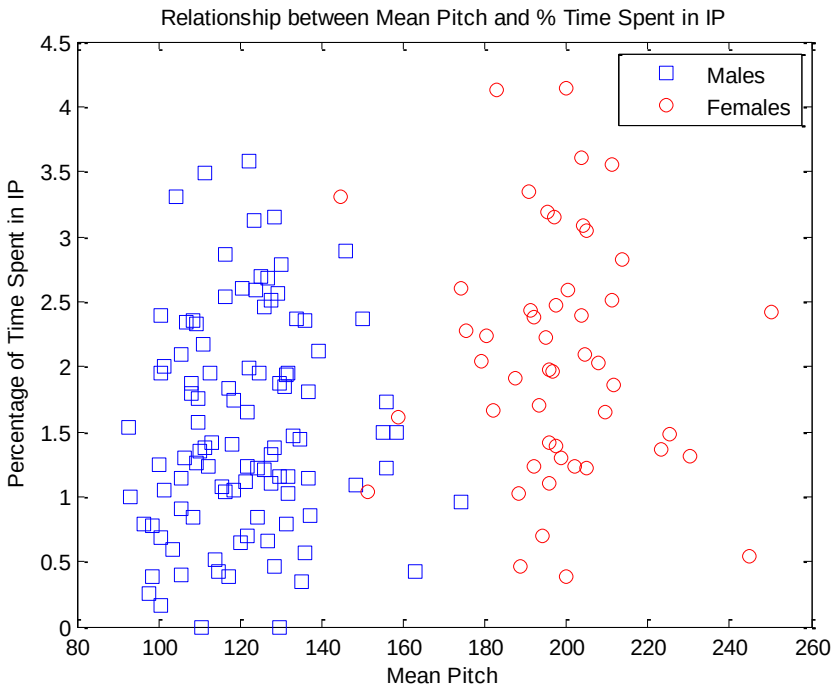

Figure 11. Mean pitch and percentage time spent on irregular phonation. r=0.2598, p=0.0013.

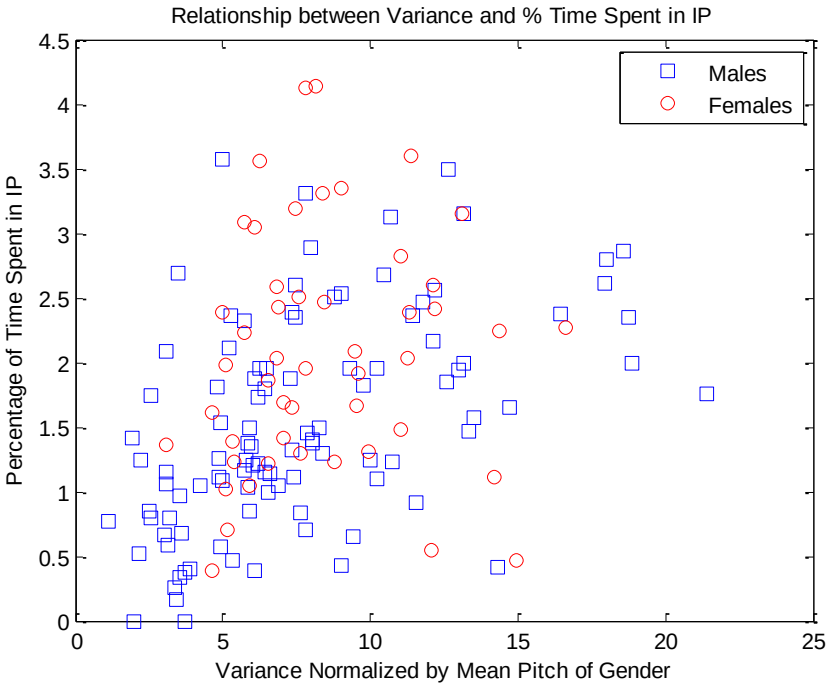

Figure 12. Pitch variance and percentage time spent on irregular phonation. r=0.3757 (p not calculated). The variance measure for each speaker is divided by the mean pitch of the speaker's gender (119.1 Hz for males and 201.4 Hz for females) in order to make the values easier to compare across gender.

Figure 11 shows the average pitch of a speaker and the percentage of their speaking time they spend on irregular phonation. There are two main groupings in this graph, corresponding to male and female speakers. The correlation between mean pitch and percentage of time spent on irregular phonation is much stronger for the group as a whole than the equivalent for either gender group. The individual gender results have low r-values and the associated p-values are not low enough to make them statistically significant at the 95% significance level. The results for mean pitch versus amount of time spent on irregular phonation as a whole, on the other hand, boast a correlation coefficient of 0.2598. They are significant at the 95% level and beyond, with a p value of 0.0013. This seems to indicate that while mean pitch can be used to predict irregular phonation, it is more a proxy for gender than anything else.

The results of our comparison of variance with irregular phonation, displayed in **Figure 12**, show a strong relationship between the two. This is especially true for the population as a whole and for the combined population. The overall population has an r value of 0.3757 and is statistically significant at the 95% level. The male population has an r value of 0.4885 at the 95% significance level. The female population has an r value of .0805 but is not statistically significant.

4.2 Average Power and Power Variance

We next perform parallel experiment relating to power. We expect to find that the mean and variance of the amount of power measured in a speaker will be positively correlated with the percentage of time spent on irregular phonation. We expect the variance of power to be more strongly correlated than its mean, for two reasons. The first reason is related to our overall thesis: as mentioned previously, we speculate that a possible factor in generating irregular phonation is that the speaker moves to or through a configuration which naturally causes irregular phonation to have a greater likelihood of appearing. We expect this effect to be more substantial than the effect of the differential difficulty of controlling the vocal folds at different mean levels of pitch and power. The second reason has to do with experimental setup. We note that factors such as the distance to the microphone of each speaker were not rigidly controlled for in the recording of the TIMIT database. This means that the amount of power recorded cannot be directly measured and used as such. To control for such factors, we normalized each speaker's file by the maximum strength of the original signal recorded.

Our results are shown below. The male correlation measures were significant, but the female and combined measures were not. The mean power for males and the percentage of time the same males spent on irregular phonation were found to have a Pearson correlation coefficient of -0.1653, which was significant at the 90% level. The power variance for males and the percentage time the same males spent on irregular phonation were found to have a Pearson correlation coefficient of -0.2023, which was found to be significant at the 95% level.

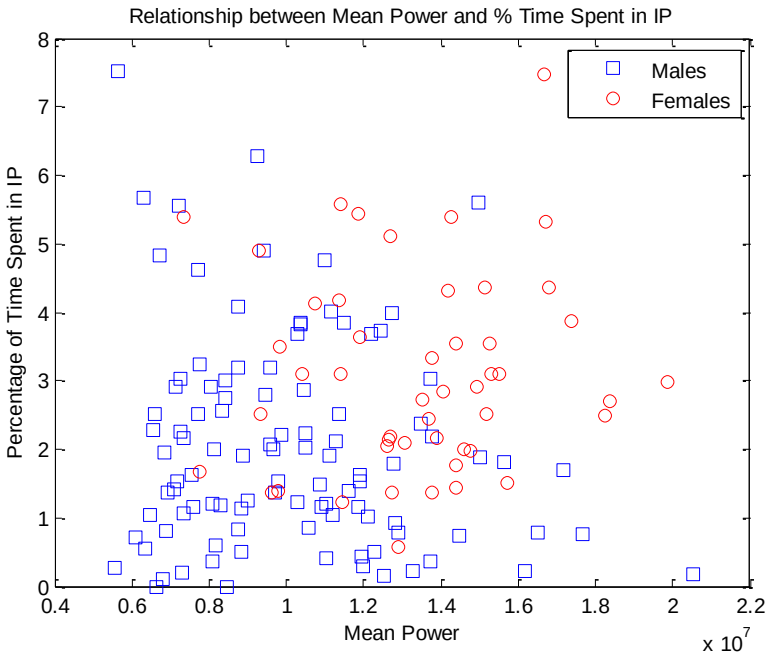

Figure 13. Mean power vs. percentage time spent on irregular phonation.

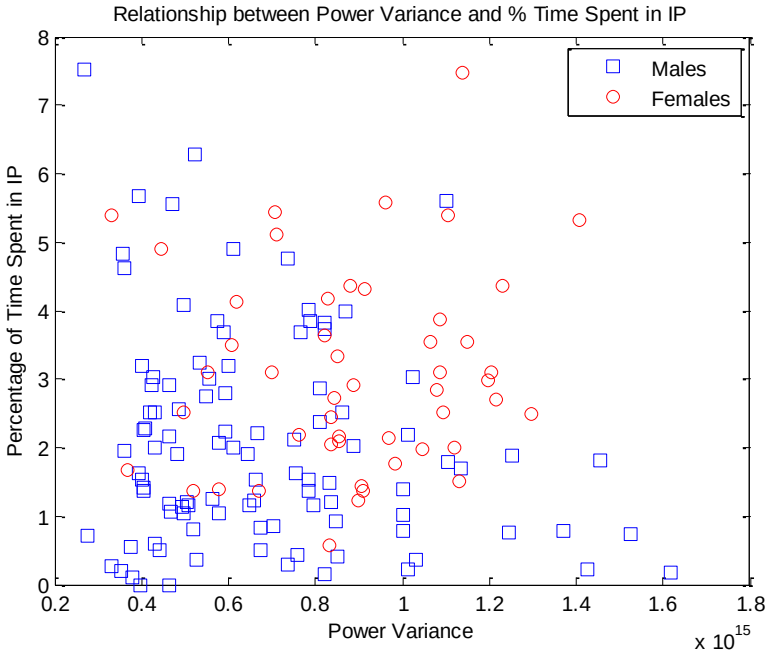

Figure 14. Power variance vs. percentage time spent on irregular phonation.

4.3 Pitch and Power Case Study: Utterance Final Irregular Phonation

We next take a more in-depth look at prosodic effects on irregular phonation localized to a small part of a sentence. We limited this case study to a single location in order to decrease the number of potentially obfuscating variables. We chose to study final creak, a subset of irregular phonation which occurs specifically at the end of an utterance. We took baseline measurements, measures of pitch, power, and duration.

4.3.1 Baseline Measurements

We first took baseline measurements on the New England and Northern dialect regions of the TIMIT database. This grouping includes 102 males and 49 females, each of whom speak ten sentences, for a total of 1510 sentences. Of these, two sentences -- labeled in **Figure 15** as sentences 1 and 2 -- are said in common among the speakers. Sentence 1 reads, "She had your dark suit in greasy wash water all year." Sentence 2 reads, "Don't ask me to carry an oily rag like that." We calculated that final creak occurred in 33% of the entire set of 1510 sentences. Females exhibited final creak more frequently than did males, with a significance level of 95%. See **Figure 15**.

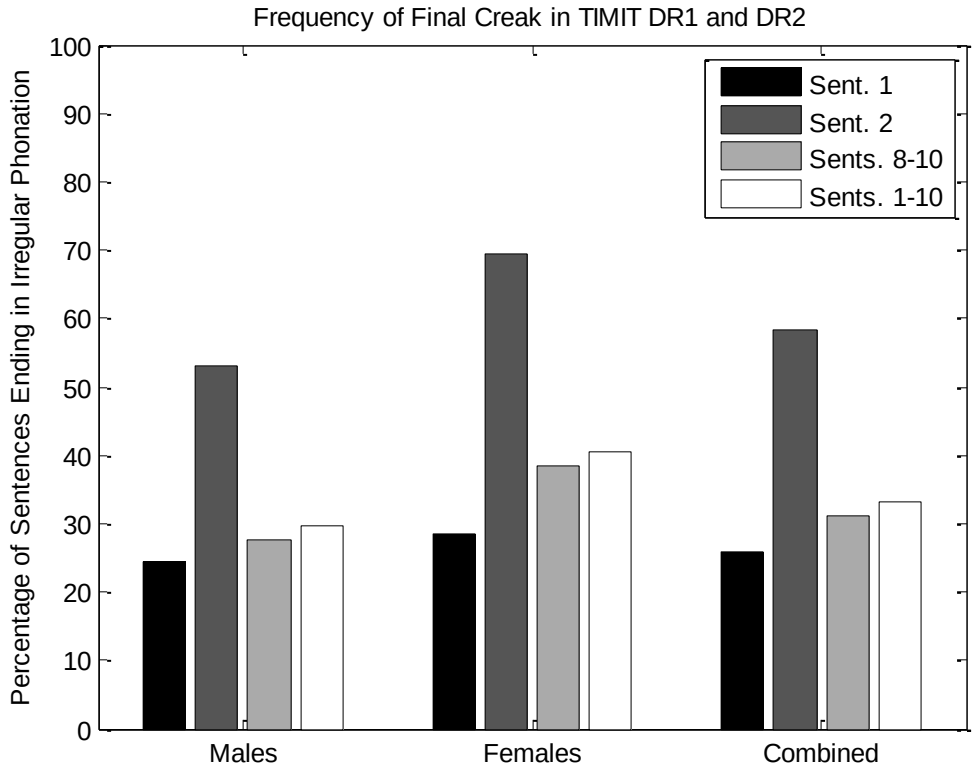

Figure 15. Final creak in the TIMIT population, regions 1 and 2.

We next studied pitch effects on sentence 1. We limited ourselves to a single sentence so that the majority of the pitch trajectories across the set of speakers would follow the same pattern. At the end of the word 'year' in this sentence, most of their pitch slopes move downward approximately linearly. Therefore, instead of measuring pitch variance, we measured the pitch slope. We fit a first-order line to each subject's pitches. After determining the best line in a least squares sense, we examined the distribution of slopes for the group with final creak and for the group without final creak, irregular phonation which specifically occurs at the end of an utterance. We found that the subjects who exhibited final creak tended to have more highly sloped pitch trajectories as they approached the end of the sentence than did those who did not exhibit final creak. **Figure 16** shows a cumulative distribution of the slopes for each group. Roughly 50% of the sentences with final creak, but only 20% of the sentences without final creak, showed a slope of -1 or less.

4.3.2 Pitch Slope

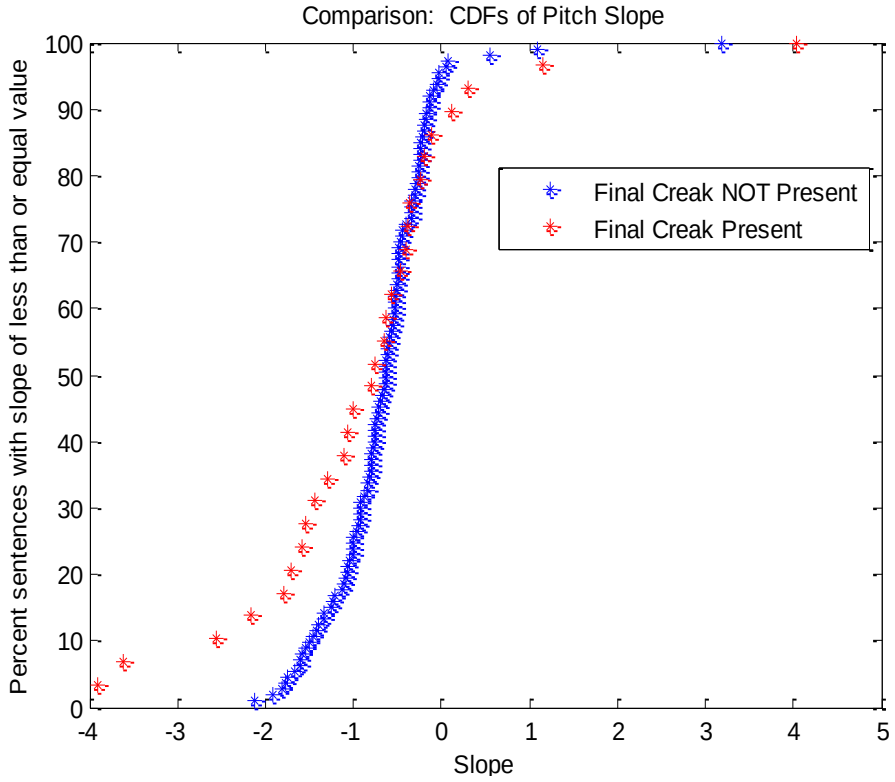

Figure 16. The cumulative distribution function for the approximated slope of the final word of the sentence for cases in which final creak is and is not present. The blue curve indicates the cases in which final creak is not present; the red curve indicates the cases in which final creak is present.

We next examine the change in power and how it can be predictive of irregular phonation. Our work focused on one sentence recorded by 151 subjects: 'She had your dark suit in greasy wash water all year'. Approximately 20% of the subjects in our sample exhibited final creak at the end of the word 'year'.

4.3.3 Power Change

We next examined the movement of power for subjects who did and did not exhibit final creak. We expected the power trajectories to be more variable for speakers with final creak. We calculated the amount of power as a function of time for the last word ('year') of each speaker. Then, we normalized the amounts of power (due to the fact that some speakers may have been positioned closer to their microphones) so that each speaker had the same amount of average power. We used the pre-existing phone labels provided with the TIMIT database to time-align the words as pronounced by different speakers. Each speaker began speaking at time 0, finished the vowel sound and moved on the 'r' at time 1. We then average the results of all speakers who did and who did not produce irregular phonation, to form composite results, shown in **Figure 17**. These results support our intuition: The sentences ending in irregular phonation do have a greater 'swing' in power over the last word of the sentence. The 'typical' speaker's power dipped, peaked, and trailed off, regardless of whether they produced irregular phonation. But the typical producer of irregular phonation swings lower initially and peaks higher as they approach the end of the word. This shows that they would have a higher amount of calculated variance. We conclude that variance in power is associated with the production of final creak.

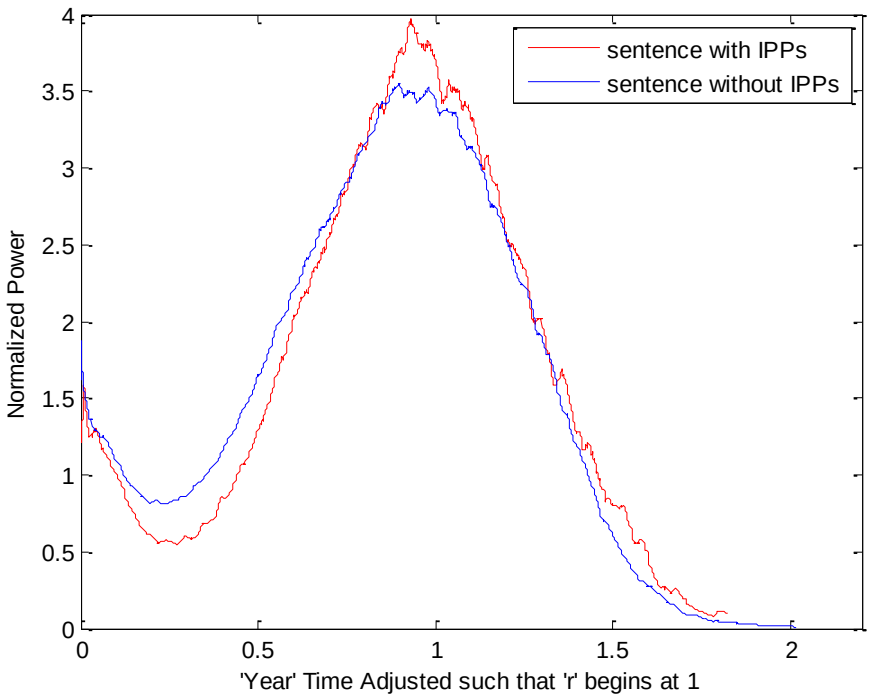

Figure 17. The average power curves are shown for sentences with final creak (red) and without (blue). These power curves span the final word a sentence, which is 'year.' They are aligned in time such the final phone of the word begins at the same point, at adjusted time 1.

4.3.4 Phone Duration

Our last comparison was of phone duration, to see whether speakers who exhibited irregular phonation also increase or decrease the amount of time they spent on any particular portion of the word. The word year can be divided into three parts, roughly speaking: 'y', 'ea', and 'r'. Different speakers realize each part differently. While all speakers were transcribed as producing

the phone '-y' for the 'y' of the word, 'ea' was alternately realized as '-ih', '-iy', or '-ix' and 'r' was alternately pronounced as '-ax', '-ah', '-axr', -'er', '-ar', or '-r' as the phones were transcribed. Additionally, some speakers exhibited irregular phonation during the word and others did not. We calculated the amount of time each speaker spent on the three parts of the word. We plotted composite distribution functions for each piece of the word, showing the distribution of amounts of time spent by subjects who did and did not exhibit final creak. We calculated the amount of time spent on the 'r' of the word in two ways. The first was time spent only on the phone associated with 'r'. The second included time spent on that phone and time spent on irregular phonation at the end of the word.

Figure 18 shows the composite results of the phone duration study. The distribution of time spent on the first and second phones of the word is similar for speakers who did and did not exhibit final creak. We see that if we choose to exclude irregular phonation from the phone associated with 'r' the duration of the final phone tends to be much shorter for sentences with final creak. However, if we include it as an irregular component of the phone, there is little difference in phone length between the two types of utterances.

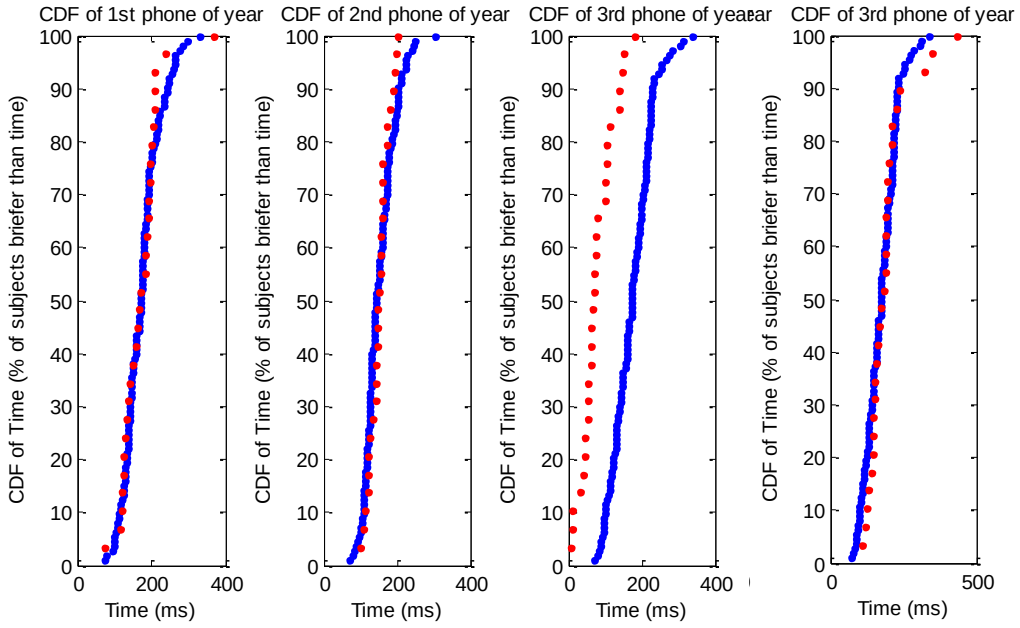

Figure 18. The length of each phone of the word 'year.' The first and second subplot, reading left to right, show the first two phones of the word (roughly, 'y' and 'ea'). The third subplot shows the length of the third phone including regular phonated parts only. The fourth subplot shows the length of the third phone including all phonation, regular or irregular.

4.3.5 Composite

Finally, we present a pair of examples in **Figure 19** and **Figure 20** which demonstrate the three experiments performed above: pitch slope, power trajectory, and phone length. We examine the word 'year' as produced by two speakers, and overlay measures of pitch, power, and phone duration.

Comparing the two, we see that the best fit line for the pitches is more sloped for the sentence with final creak. We also see that there is a more dramatic drop in power at the end of the word in the sentence with final creak. Examining the phone lengths, we see that the result is heavily dependent on where we decide that the phone associated with 'r' ends. If we say that the phone is finished when irregular phonation begins to occur, our measured duration for that phone is much shorter than if we say that it includes the irregular phonation which is occurring at the end of the sentence.

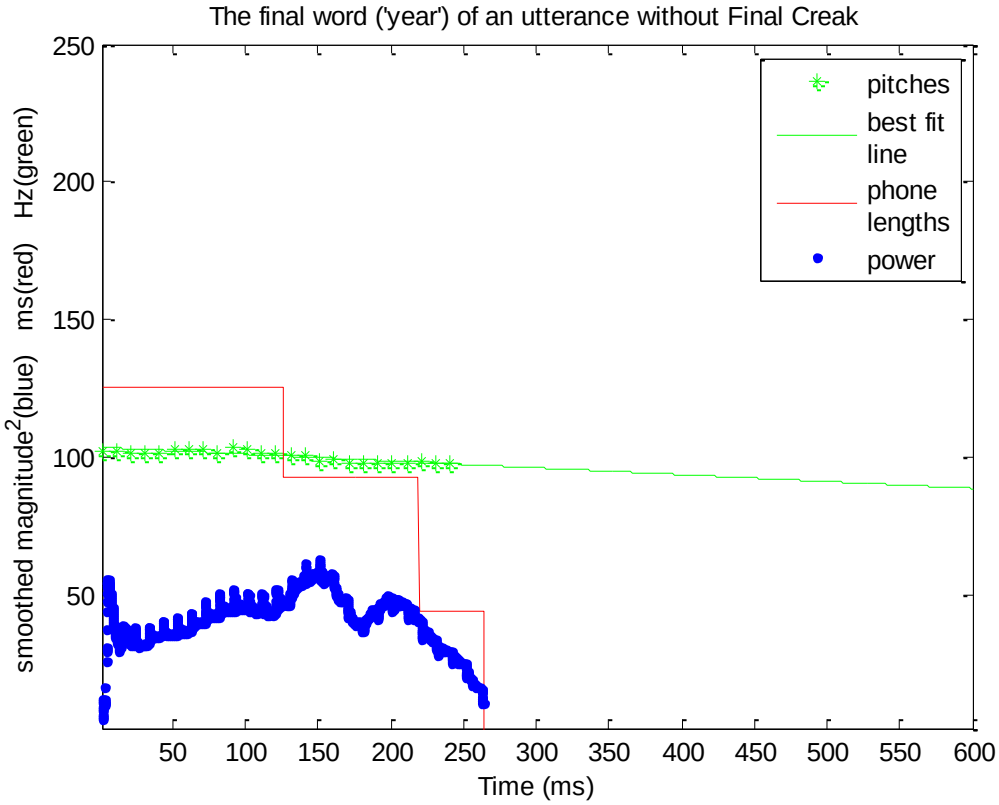

Figure 19. Illustration of pitch, smooth power, and phone duration in the word 'year' for specific TIMIT. case.

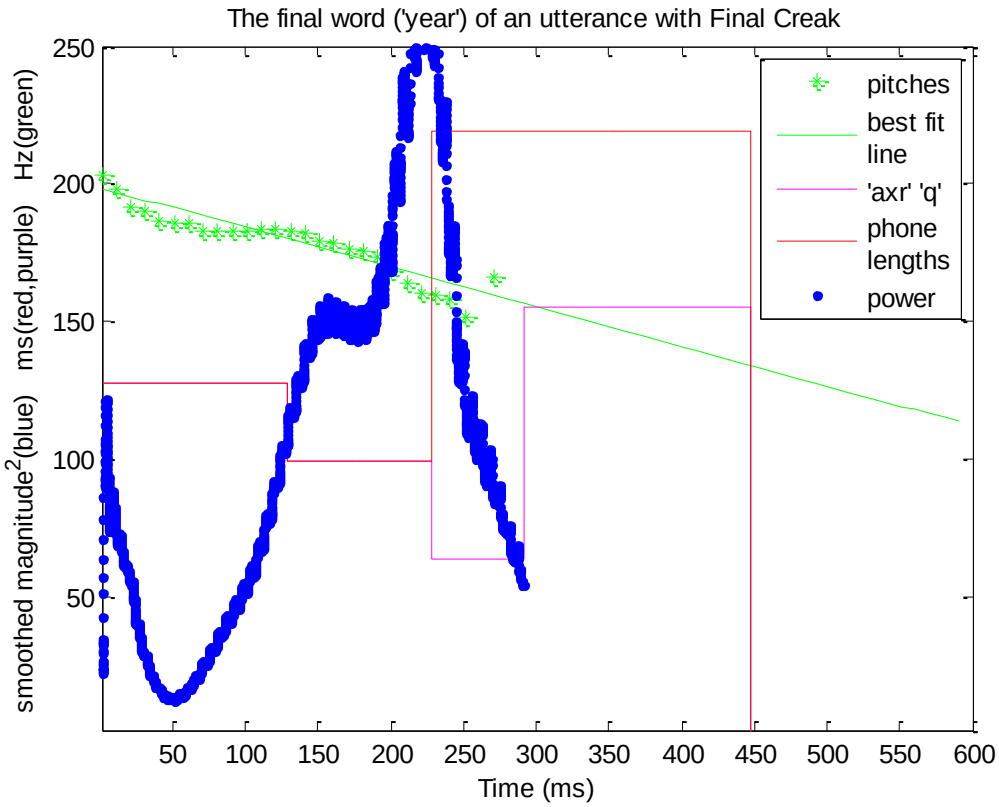

Figure 20. Illustration of pitch, smooth power, and phone duration in the word 'year' for a specific TIMIT case. 'axr' refers to the phonated portion of the 'r' in 'year' and 'q' refers to the irregularly phonated portion.

4.4 Summary

In this chapter, we have shown several relationships between irregular phonation and prosody. We have shown positive correlations between irregular phonation and mean pitch, pitch variance, mean power, and power variance. We have examined final creak specifically, looking in detail at the pitch tracks, power tracks, and phone durations within the last word of utterances which do and do not exhibit final creak. We conclude that greater variance in both pitch and power is related to the exhibition of irregular phonation.

Chapter 5

The Nature of Irregular Phonation

In Chapter 5, we examine the properties of individual instances of irregular phonation. We focus on the amount of spacing between individual pulses and the heights of those pulses. We investigate the relationships between these measures and the average and variance of pitch and power for our entire database, finding four several significant correlations. Finally, we conclude Chapter 5 with a look at how location within a word is related to the nature of the irregular phonation produced.

5.1 The Characteristic Spacing

We first examined the 'usual' spacing found in irregular phonation. We determined the average pitch, variance of pitch, average power, and variance of power as described previously. We detected the location of pulses (e.g., glottal closings) using two methods. One was an implementation of the Minimum Entropy Deconvolution (MED) method. Another was a fusion method. The fusion technique is currently being developed at MIT Lincoln Laboratory. The method combines a variety of different state-of-the-art pulse-time detectors, such as those based on inverse-filtering (e.g., MED [4]) and linear-phase-estimation [27]. We concluded that use of the fusion method provided more accurate pulse locations and therefore more reliable correlation results. **Figure 21** shows through scatter plots the relationship between speakers' average pitch and average spacing of pulses in irregularly phonated regions of speech, as calculated with the MED and fusion methods, respectively. Regions 1 and 2 of the TIMIT database were used. By switching to the more accurate fusion method, we were able to improve our onset detection and correspondingly our Pearson correlation coefficient (from 0.049 to -0.230).Using this method, we found the spacing between individual glottal pulses and calculated the maximum inter-pulse spacing for each episode of irregular phonation.

Figure 22 shows the distribution of maximum spacing over all episodes of irregular phonation labeled in our database. They are color-coded by gender and region, and displayed along a pitch scale. Maximum spacing is a useful characteristic because it is has been posited that a pitch of below 83 Hz is a signal to listeners that irregular phonation is occurring. According to this hypothesis, maximum pulse spacing is one of the most important characteristics of irregular phonation. The majority of the incidents measured do in fact live up to that definition of an episode of irregular phonation. 66% of the sentences had a maximum spacing above 12 ms, and therefore 'pitch' (difficult to define without regular pitch periods) of less than 83 Hz. It may be that errors in pulse detection account for the rest. It could also be that irregular phonation under some circumstances does not need to make use of this signal – perhaps in some cases it is already obvious enough for other reasons that the 12 ms threshold need not be crossed.

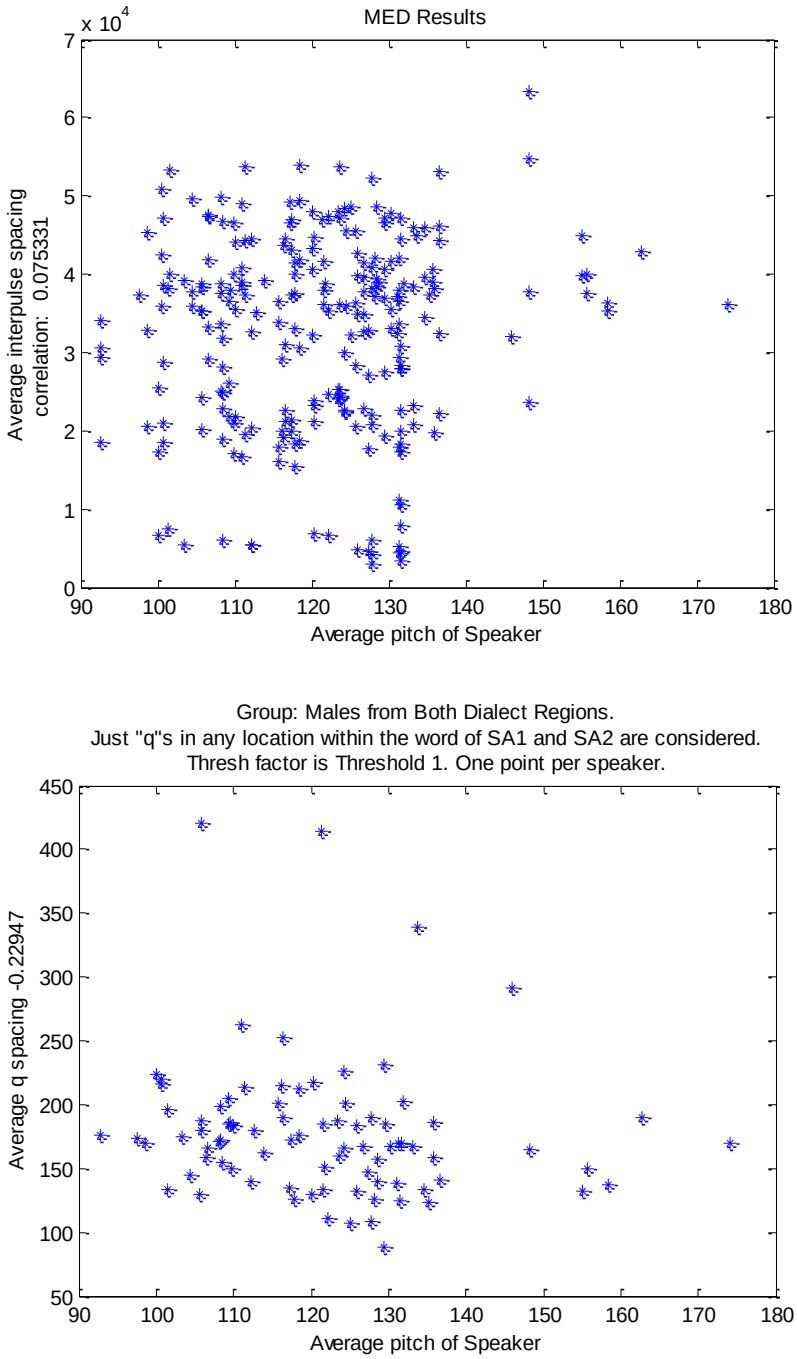

Figure 21. Average pitch versus average spacing in irregular spacing obtained using the MED method (top) and fusion method (bottom). Data shown here is for males from dialect regions 1 and 2. The data over two sentences was combined to produce one data point per speaker. Each speaker spoke the same two sentences.

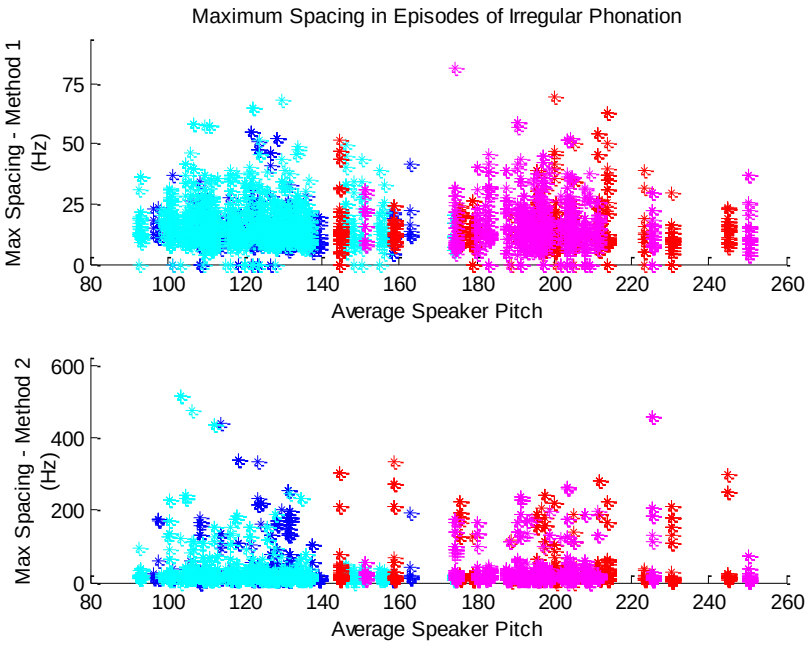

Figure 22. Distribution of maximum spacing during irregular phonation. (Colors differentiate gender and region of speaker.)

5.2 Prosodic Dynamics and Spacing

In seeking to understand what attributes of a speaker and what type of speech utterance might impact the realization of irregular speech, we examined the effect of the mean and the variance of pitch and power. The prosodic variables utilized were mean pitch, pitch variance, mean power, and power variance. These were determined as described previously. We compared these to three properties of irregular phonation: maximum inter-pulse spacing, median inter-pulse spacing, and mean height. Maximum inter-pulse spacing is robust to the errors of the pulse-detection system, since these errors usually involve detection of an 'extra' pulse. It is also important because of the proposition that maximum irregular pulse spacing above a certain threshold may clue observers in to irregular phonation. Median inter-pulse spacing is also fairly robust to insertions or deletions of pulses and comes closer to representing the 'typical' irregular pulse. Mean pulse height is measured because we suspect that it will be particularly associated with changes in mean power. Variables were determined as described previously, with the exception of mean height. Mean height was calculated by smoothing the original speech signal and measuring it at points determined by the output of the fusion pulse detection system

		Max Spacing	Median Spacing	Mean Height
Mean Pitch	r-value	-0.2382	-0.2966	-0.1157
	p-value	0.0200	0.0035	0.2639
Pitch Variance	r-value	-0.0040	-0.1101	-0.1393
	p-value	0.9688	0.2880	0.1781
Mean Power	r-value	-0.2808	-0.3145	.0004
	p-value	0.0058	0.0019	0.9971
Power Variance	r-value	-0.0349	-0.1618	0.0074
	p-value	0.7370	0.1171	0.9431

Table 5. Relationship between acoustics and attributes of irregular phonation. The three most prominent correlations are highlighted.

We found that mean pitch and mean power were both significantly correlated to maximum inter-pulse spacing during irregular phonation and median inter-pulse spacing during irregular phonation. The three strongest correlations are highlighted in the results table. The fact that the coefficients on these more significant relationships are negative indicates that higher pitch and greater amounts of power are associated with wider pulse spacing during irregular phonation. We did not, however, find significant results involving pitch and power variance or mean pulse height in this experiment. This may be due to a lack of a relationship. It may also be due to a more complex relationship. As we show in the next section, other factors such as gender, dialect, and location may influence these correlations.

5.3 Location as a Correlate to Characteristics of Irregular Phonation

We now examine each of the above highlighted relationships in more detail according to the location within a word of the occurrence of the irregular phonation. The results of these more detailed examinations are shown below.

We define "beginning" as onset of speech in a word. This means that irregular phonation is heard before any part of any other phone. "end" refers to the very last event that occurs in the word".

"Middle," of course, falls between the two. We label each episode of irregular phonation from each sentence by which part of the word it occurred in, as well as the speaker's gender.

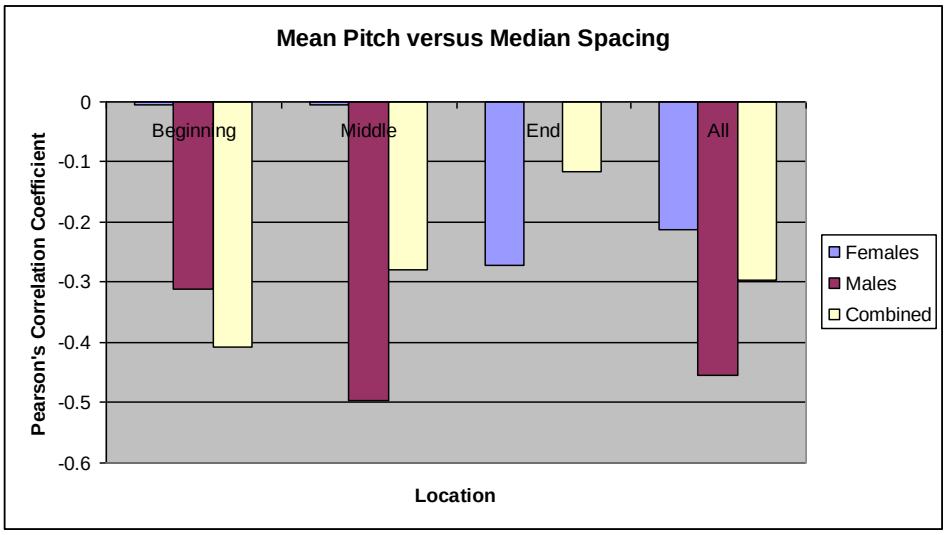

Figure 23. Mean pitch versus median spacing. Subsets not shown did not occur in data.

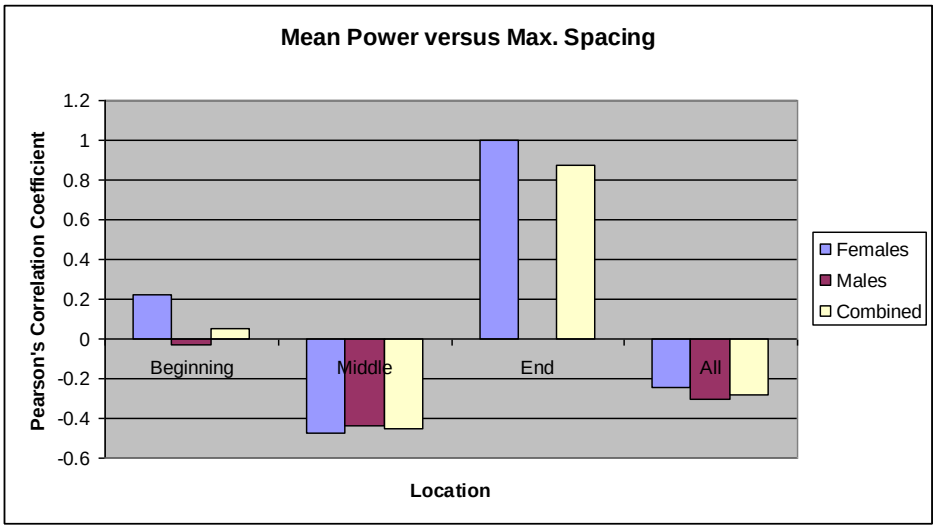

Figure 24. Mean power versus maximum spacing. Subsets not shown did not occur in data.

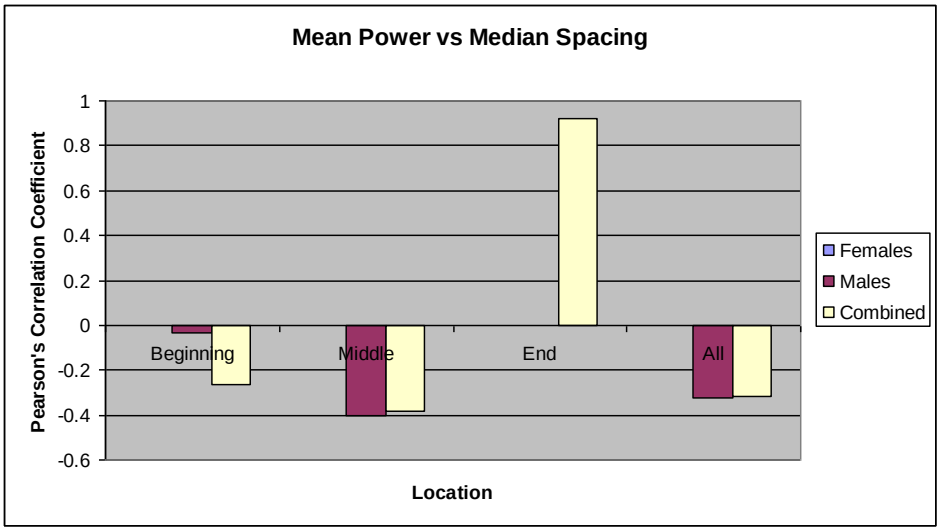

Figure 25. Mean power versus median spacing. Subsets not shown did not occur in data.

Figures 23 through **25** show the correlation between the highlighted variables of **Table 5**. The most striking of these is **Figure 24**. We see that when irregular phonation occurs at the beginning or end of a word, increased power is associated with increased speaking, whereas the opposite is true in the middle of the word. This is the effect we described above, when location-specific factors wash out an otherwise higher correlation. We see that, when examined in more detail, areas of **Figure 24** which were not previously significant may hold promising correlations within subsets of the population.

5.4 Summary

This chapter has found several key relationships:

- Between the mean pitch of the utterance and median spacing of the irregular phonation
- Between the mean power of the utterance and maximum spacing of irregular phonation
- Between the mean power of the utterance and median spacing of irregular phonation
- Between the mean pitch of the utterance and the maximum spacing of irregular phonation.

We have also found that the level of correlation of various factors varies with the location within the word in which there is irregular phonation. This is likely due to the fact that, in general, the vocal tract has similar patterns of configuration and movement at the beginnings of words, during the middles of word, and at the ends of words. The results are promising because using location-dependence in the future may reveal stronger correlations.

Chapter 6

Multivariate Correlations

In this chapter, we present the results of a model which takes into account multiple variables that we have shown relate to the existence of and spacing within irregular phonation. Dialect, gender, pitch ratio, and pitch variance are standout features, in the sense that they have high significance and are weighted heavily in the model.

6.1 Model fitting

We have used Matlab's stepwise function as well as Excel's linest to predict irregular phonation and its properties using acoustics and demographics. Both these functions are used in multivariate modeling of data. Stepwise is an interactive function which, step-by-step, adds and removes potential predictive factors to a model of prediction for the output factors dependent on their probability of explaining variation in the data. Our initial work was with stepwise, giving us intuition for which variables to include in the model later generated using linest.

Linest stands for linear estimation, and it also performs a multivariate linear regression using ordinary least squares. This is the program we ultimately used for this work. The generated model is linear in the sense that it is linear in the dependent variable (number of instances of irregular phonation), although we used nonlinear independent variables in some cases. It was necessary to take natural logarithms of the mean power and power variance methods such that their magnitudes would more closely align with those of other variables. We also used the variables which compared pitch and power estimates for one sentence to the speaker's 10-sentence average. Factors whose inclusion in the model we considered include Dialect, Gender, Mean Pitch 1 (mean pitch over one sentence), Mean Pitch 10 (over ten sentences), Mean Pitch Ratio (mean pitch of one sentence divided by ten-sentence mean), Pitch Variance 1 (over one sentence), Pitch Variance 10 (over ten sentences), Pitch Variance Ratio (pitch variance of one sentence divided by ten-sentence pitch variance), Mean Power (natural logarithm of mean power over one sentence), and Power Variance (natural logarithm of mean power over ten sentences).

The results of our fits are shown in **Table 6** and **Table 7**.

	Dialect	Gender	Mean Pitch 1	Mean Pitch 10	Mean Pitch Ratio	Pitch Var 1	Pitch Var 10	Pitch Var Ratio	Log Mean Power	Log Power Var	α
β	0.358	0.503	0.038	-0.04	-8.75	0.0002	0.0001	-.069	1.518	-1.671	44
p	0.000	0.000	0.015	0.006	0.000	0.000	0.340	0.341	0.000	0.000	0.00

Table 6. Coefficients resulting from autoregressive prediction of number of instances of irregular phonation. β is the coefficient of each term and *p* indicates the significance of the result. Lower values of *p* are preferable. α is the constant term of the linear equation.

	Dialect	Gender	Mean Pitch 1	Mean Pitch 10	Mean Pitch Ratio	Pitch Var 1	Pitch Var 10	Pitch Var Ratio	Log Mean Power	Log Power Var	Numq	α
B	-13.10	38.75	-0.82	0.055	53.56	0.011	0.002	-6.16	-35.54	40.44	-2.870	-601
p	0.001	0.000	0.420	0.957	0.730	0.003	0.690	0.184	0.092	0.005	0.047	0.02

Table 7. Coefficients resulting from autoregressive prediction of median inter-pulse spacing of irregular phonation. β is the coefficient of each term and *p* indicates the significance of the result. Lower values of *p* are preferable. α is the constant term of the linear equation.

6.2 Discussion of Model Fits

We first sought to predict whether irregular phonation would occur, given our measurements and demographics. We then sought to predict the median inter-pulse spacing given that we know irregular phonation exists in the sentence.

In our prediction of whether irregular phonation would occur, eight of the ten variables we used contributed meaningfully to prediction at the 95% confidence level. These were Dialect, Gender, Mean Pitch 1, Mean Pitch 10, Mean Pitch Ratio, Pitch Variance 1, Mean Power, and Power Variance. Furthermore, the F statistic of the regression was 29, meaning that there is no question that we are explaining something. The R-squared statistic was only 0.128, meaning that our model explains nearly 13% of the variance in numbers of instances of irregular phonation. The F statistic and R-squared statistic, taken together, tell us that we have definitely captured something that explains the number of instances of irregular phonation in the sentence. However, they also tell us that there are other variables, not included in this model, which we would need to find in order to fully explain the variation.

In our prediction of the median spacing of the irregular phonation, we used the same variables as in the previous regression, with the addition of the knowledge of how many instances of irregular phonation had occurred in the sentence. The fit of this model was weaker – only five of the variables contributed significantly. These were Dialect, Gender, Pitch Variance 1, Power Variance, and NumQ (our newly available variable, the number of instances of irregular phonation in the sentence). The F statistic was 11; though not as strong as the F statistic in the previous model, this still indicates that the model does capture some features of the behavior of spacing, but not all of them. The R-squared statistic is 0.058. This is less than half as large as in the first model, so we know that we have explained significantly less variation in the median spacing than we did in the number of instances of irregular phonation, despite the addition of information in the form of the number of instances of irregular phonation in the sentence. It may be that the factors affecting spacing differ from those affecting the number of instances of irregular phonation.

The F statistics for both of these models indicate that we have explained part of the variance involved in the prediction of irregular phonation and its median spacing. It is not surprising that the R-squared statistics are not higher; after all, we have not included any information about the linguistic content of the sentences. Much variance may be introduced by the fact that the set of sentences spoken by the various speakers are not identical (only two of the ten sentences in each set are held constant).

6.3 Discussion of Prominent Features

Two of the features we included are binary: Region (Dialect) and Gender. Therefore, the coefficient in the linear model calculated associated with each of these variables is equal to the mean difference between the expected value of the predicted variable in the two cases.

Gender is modeled as binary between male (0) and female (1). Its coefficient in our first model is 0.503. This means that on average, a female speaker is expected to produce 0.503 more instances of irregular phonation than a male speaker per utterance. The coefficient on Gender in our second model is 38.7, meaning that a female's median spacing during irregular phonation is, on average, 38.7 samples longer than a male's. Given that our sampling rate is 16 kHz, this amount to difference of 2.4 ms.

Similarly, in the first regression, the dialect coefficient of 0.358 tells us that, on average, we expect 0.358 more instances of irregular phonation from members of the New England Dialect Region than those of the Northern Dialect Region. In the second regression, the coefficient of -13.10 tells us that we expect people of the New England Dialect Region to have slightly shorter median spacing during irregular phonation – in expectation 0.8 ms shorter.

Other prominent features (high significance and large weight) include Mean Pitch Ratio (in the first regression) and Power Variance (in the second regression). Mean Pitch Ratio is the ratio between the mean pitch of the sentence in question and the overall mean pitch of the speaker, measured over ten sentences. Its coefficient of -8.75 means that for a given individual, speaking lower in their range is predictive of a tendency towards greatly increase levels of irregular phonation in their speech. Power variance is the natural logarithm of the variance of power measured over the sentence in question. The natural logarithm of the variance of power ranges from 33.22 to 35.02 for our subjects. Power Variance's coefficient of 40.44 indicates that

increased power therefore tells us that we expect speakers with the highest amount of variance in their power to have longer spacing during irregular phonation by about 4.5 ms.

6.4 Summary

We used multivariate regressions to explain 12.8 percent of the variance inherent in the number of instances of irregular phonation in sentences and 5.8 percent of the variance inherent in the median spacing of episodes of irregular phonation. Key variables involved were gender, dialect, sentence pitch relative to a speaker's average pitch, and the natural logarithm of variance in power within the sentence. Though our results are significant, they also indicate that other variables are necessary to better explain the variance in these two aspects of irregular phonation. These may include indicators of the linguistic structure of the sentence.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have described the pitch, pitch variance, and irregular phonation distributions inherent in a large and publicly available corpus of read speech. We then use those and other features to provide information about which factors are correlated with the occurrence and properties of irregular speech. Overall pitch variance, steep pitch trajectories, and swinging power were characteristic of sentences with irregular phonation.

These findings fit well with our hypothesis of change leading to potential instability in the larynx. We also found significant gender and dialect effects at play, as seen in Chapter 6. In investigations to date, we have fitted models which use these factors to predict the number of instances of irregular phonation in a given utterance and their median spacing. Gender, Dialect, and Mean Pitch Ratio, Pitch Variance were found to have the most significant effect.

As reported by Slifka, irregular phonation a marker of real and intended silence [1]. Based on our results, we suggest that it is also a probabilistically marker of "change" or variance.

7.2 Future Work

Our investigations suggest that this work is applicable to the prediction of the occurrence of irregular phonation and its characteristics. To further study dialect differences, it will be necessary to label the six remaining dialect regions of TIMIT for irregular phonation, or to find another fully labeled corpus. One drawback of TIMIT is that it is read speech as opposed to spontaneously generated speech. Using spontaneous speech would further add to this work. While we have found correlations between individual predictive aspects of speech and properties of irregular phonation, their interaction may be extremely important. We have begun to use the predictive factors jointly, in our multivariate estimation of Chapter 6. This work should be expanded to include other factors, such as linguistic indicators. We should also attempt to find non-linear relationships and patterns among the data. The ability to detect the occurrence of irregular phonation will be useful to the analysis and modification of speech. Currently, the majority of errors in pitch detection are caused by irregular phonation [22]. The ability to predict irregular phonation would lead to more robust functioning of these systems. Modification of speech would be improved not only with better analysis of the original signal, but also the ability to insert or delete irregular phonation in a manner which increases the 'natural' quality of the synthesized signal.

Given the high level of variability observed in speakers' tendencies to use irregular phonation, and preliminary results with speaker identification [8], it is quite possible that individual profiles of irregular phonation could be built and used to a variety of purposes. Changes in an individual's profile may provide information about physical or mental wellbeing. After all, humans have the ability to tell mood via speech - therefore, the data exists to be found. Finally, another application is that of dialect ID where dialect-dependence of irregular phonation can be exploited.

References

‘

[1] J. Slifka, "Irregular Phonation and Its Preferred Role as Cue to Silence in Phonological Systems," presented at the ICPhS, Saarbrücken, Germany, 2007.

[2] K. Surana and J. Slifka, "Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English?," presented at the Speech Prosody, Dresden, Germany, 2006.

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

[4] R. A. Wiggins, "Minimum Entropy Convolution," *Geoexploration*, vol. 16, pp. 21-35, 1978.

[5] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 2000.

[6] C. H. Coker, *et al.*, *Speech Synthesis*. Baltimore, Md.: Waverly Press, Inc., 1963.

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

[8] N. Malyska, "Analysis of Nonmodal Glottal Event Patterns with Application to Automatic Speaker Recognition," Ph.D., Health Science and Technology, M.I.T., Cambridge, MA, 2008.

[9] J. Greenberg, "Course Notes," unpublished.

[10] H. Hollien and R. W. Wendahl, "Perceptual Study of Vocal Fry," *Journal of the Acoustical Society of America*, vol. 43, pp. 506-509, 1968.

[11] H. Hollien and J. F. Michel, "Vocal Fry as a Phonational Register," *Journal of the Acoustical Society of America*, pp. 600-604, 1968.

[12] T. Bohm, *et al.*, "Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles," *Acoustics '08 Paris*, pp. 6141-6146, 2008.

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

[14] "Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)." ASHA Special Interest Division 3, Voice and Voice Disorders, 2002.

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

[16] I. R. Titze, *Definitions and Nomenclature related to voice quality*. Vocal Fold Physiology: voice quality control. San Diego, CA: Singular Publishing Group, 1995.

[17] K. Surana, "Classification of vocal fold vibration as regular or irregular in normal, voiced speech," M.Eng., Computer Science and Engineering, M.I.T., Cambridge, MA, 2006.

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

[19] T. Bohm and S. Shattuck-Hufnagel, "Utterance-final glottalization as a cue for familiar speaker recognition," presented at the Interspeech, Antwerp, Belgium, 2007.

[20] T.-J. Yoon, *et al.*, "Detecting Non-modal Phonation in Telephone Speech," presented at the Speech Prosody, Campinas, Brazil, 2008.

[21] C. Henton and A. Bladon, "Creak as a sociophonetic marker," *Language, Speech, and Mind: Studies in Honour of Victoria Fromkin*, pp. 3-29, 1988.

[22] P. Hedelin and D. Huber, "Pitch Period Determination of Aperiodic Speech Signals," presented at the Acoustics, Speech, and Signal Processing, Alburquerque, NM, 1990.

[23] L. Dilly, *et al.*, "Glottalization of word-initial vowels as a function of prosodic structure," *Phonetics*, vol. 24, pp. 423-444, 1996.

[24] J. Pierrehumbert and D. Talkin, *Lenition of /h/ and glottal stop* vol. II. Cambridge: University Press, 1992.

[25] T. Grivičić and C. Nilep, "The Use and Function of *yeah* and Creaky Voice," *Colorado Research in Linguistics*, vol. 17, 2004.

[26] "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus: NIST Speech Disc CD 1-1.1," ed: National Institute of Standards and Technology, 1990.

[27] M. Brookes, *et al.*, "A quantitative assessment of group delay methods for identifying glottal closures in voice speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 456-466, 2006.