Cross-Linguistic Acoustic Characteristics of Phonation: A Machine Learning Approach

Laura Maggia Panfili

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Richard Wright, Chair

Gina-Anne Levow

Sharon Hargus

Tanya Eadie

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Cross-Linguistic Acoustic Characteristics of Phonation:
A Machine Learning Approach

Laura Maggia Panfili

Chair of the Supervisory Committee:
Professor Richard Wright
Department of Linguistics

Phonation, the process of producing a quasi-periodic sound wave through vocal fold vibration, plays different roles in different languages. Phonation types, or voice qualities, are produced by adjusting the length, thickness, and separation of the vocal folds. In addition to being a complex physiological phenomenon, phonation is also a complex acoustic phenomenon; different phonation types are generally distinguished by a constellation of acoustic properties, and those properties vary from language to language. This dissertation presents a machine learning approach to investigating the role those acoustic properties play in phonation in different languages.

This study examines phonation in six languages from four families: English, Gujarati, Hmong, Mandarin, Mazatec, and Zapotec. These languages use phonation in a variety of ways, including contrastively, alongside tones, sociolinguistically, allophonically, and prosodically. Two machine learning algorithms – a Support Vector Machine and a Random Forest – are used to explore which acoustic properties best distinguish phonation types on vowels in each of those six languages. In addition to SVM weights and Random Forest importance, correlations and ablation studies are used in the analysis. Results reveal that while each of the six languages relies on a different subset of acoustic features to distinguish its phonation types, some features are consistently important. Phonation varies enough from

language to language that languages should be treated separately for the study of phonation. However, all six languages rely on at least one variant of Harmonics-to-Noise Ratio, as well as Variance of Pitch Tracks, a new measure that takes advantage of the pitch tracking errors commonly found in non-modal phonation.

Machine learning was also used to fine tune a classifier for English phonation types. Unlike other voice quality classifiers, this study focuses on just English and on the three-way breathy vs. modal vs. creaky contrast, rather than on a binary creaky vs. non-creaky distinction. The best performing classifier developed here achieves a weighted F1 score of 0.864, which is on par with state-of-the-art phonation classifiers but performs a more complex task. However, it still struggles with breathy voicing, largely a consequence of data sparsity.

This dissertation demonstrates that machine learning is a powerful tool for the study of phonation. It illuminates some of the previously unexamined similarities and differences between phonation types in different languages, and introduces a new measure, Variance of Pitch Tracks, which proves quite useful in machine classification of phonation. In addition to contributing to the understanding of phonation, this dissertation presents a new methodology for its study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, many thanks to Dr. Richard Wright for guiding me through the daunting process of writing a dissertation, showing me how to be a good scientist, and for teaching me early on to say "jitter and shimmer" and not "shimmer and jitter." Thanks as well to Dr. Gina Levow, who patiently and thoughtfully helped me wade through the nuances of machine learning methods and analysis; to Dr. Sharon Hargus, who reminded me to consider the big picture and always think like a linguist; and to Dr. Tanya Eadie, who helped me build a solid foundation in the anatomical and physiological sides of phonation. I feel very fortunate to have had an extra-large committee to help me throughout the process, with each member contributing a different piece of the puzzle.

This dissertation relies heavily on the ATAROS Corpus, which was built by a UW team lead by Dr. Gina Levow. Their thorough work and careful documentation made my work easier; Valerie Freeman's guidance in using the corpus was invaluable. Thanks as well to Nicole Chartier, who helped annotate phonation in the corpus, and to John Riebold, who patiently introduced me to Praat scripting, LaTeX, and various other technical skills. I also benefitted greatly from the support of Alli Germain, Molly FitzMorris, Andrea Kahn, and Esther Le Grézause. Thanks especially to Claire Jaja, who was instrumental in helping me get started with machine learning.

Thanks to my parents, Peter and Natalie Panfili, for fostering my scientific interest in language and for being unfailingly supportive of my long education. Finally, many thanks to Ethan Roday, who provided immense amounts of technical and moral support.

Portions of this work were supported by National Institutes of Health grant DC0060014 and The University of Washington Presidential Dissertation Fellowship.

DEDICATION

For Peter Panfili, a lifelong learner, and Natalie Panfili, a lifelong teacher.

Part I

BACKGROUND

Chapter 1

INTRODUCTION

Phonation, vocal fold vibration that produces the quasi-periodic sound wave used in voiced speech, is a complex and fascinating linguistic phenomenon. We have extremely fine control over the muscles involved in phonation and, as a result, we can produce many different types of vocal fold vibration. These variations, called *phonation types*, are used in various ways in different languages. Phonation can signal meaning, augment tones, convey sociolinguistic information, and more. Additionally, what's nominally the same phonation type can vary from language to language; the acoustic signature of a given phonation type may look different in different languages. An added complication is that phonation does not map to a single acoustic property. Rather, it's often best described by a constellation of acoustic properties that reflect different aspects of vocal fold vibration.

Linguists have tackled the complexity of phonation in various ways. Several studies have conducted in-depth analyses of the acoustic properties of phonation types for specific languages, including Gujarati (Khan, 2012) and Hmong (Esposito, 2012), usually by describing or statistically quantifying how acoustic properties vary between that language's phonation types. Many such studies are limited by scope – they generally consider between one and ten acoustic properties. By treating each of these properties individually, they also overlook potentially important interactions between properties.

The similarities and differences in the acoustic properties of phonation in different languages have the potential to shed light on the typology of phonation. But with different studies investigating different correlates to phonation and using different statistical techniques, it's difficult to compare findings across languages. Some studies, notably Keating et al. (2011), have conducted thorough cross-linguistic comparisons of phonation, but these

studies have also used small sets of acoustic properties and have largely not considered interactions.

Additionally, previous research on phonation – both language-specific and comparative studies – has focused on languages that use phonation contrastively, and to some extent on languages that use phonation alongside tones. More recently, however, linguists have taken an interest in sociolinguistic uses of phonation. Non-modal phonation in languages like English, for example, provides information about a speaker's identity and their role in the conversation, rather than providing semantic information. Previous studies of English phonation types have focused on the social meanings that they index but generally not on their acoustic properties (Andrus, 2011; Podesva, 2013; Yuasa, 2010).

In this dissertation, I present a machine learning approach to examining the acoustic properties that best describe phonation types in six languages from four families: English, Gujarati, Hmong, Mandarin, Mazatec, and Zapotec. The primary goal is to use machine learning to illuminate the set of acoustic properties that best distinguishes phonation types in these languages, and to explore the differences between them. The secondary goal is to build a high-performing phonation classifier for English that may be used in future sociolinguistic research.

In addressing the former – which I will refer to as the *linguistic question* – I aim to fill the gaps identified above in the existing phonation literature. I use machine learning to discover complex patterns in a large data set, identifying which acoustic properties can best distinguish phonation types as well as interactions between properties. Applying the same methodology to all six languages allows for meaningful cross-linguistic comparisons of these acoustic properties. In addition to looking at patterns between individual languages, I examine patterns across groups of languages that use phonation in different ways, as well as between related languages.

While machine learning is a powerful tool for studying the acoustic correlates of phonation types, its most common use is for classifying data. To address my secondary goal – the *classification question* – I fine-tune a machine learning model to automatically identify

English phonation types in previously unseen data. Throughout this dissertation, I hope to show that machine learning is an accessible and powerful tool for phonetics and typology.

This study consists of four main parts. Part One focuses on background information. Chapter 2 discusses the different phonation types included in this dissertation and their linguistic uses. In Chapter 3, I present the acoustic properties (called *features* in machine learning) that may correlate with phonation types and that will be used throughout this dissertation. Chapter 4 describes the corpora used and the data processing methodology. Finally, Chapter 5, provides a basic introduction to machine learning and describes evaluation metrics used to analyze the machine learning models. Part Two addresses the linguistic question. For each of the six languages, I report and analyze the results of the models, and Chapter 12 compares these findings across languages. Part Three addresses the classification question, in which I explore ways of fine-tuning an English phonation classifier to achieve state-of-the-art performance, and Part Four concludes the dissertation.

Chapter 2

PHONATION

Phonation is the process of using air pressure from the lungs to set the vocal folds into vibration, producing a quasi-periodic sound wave. We use the upper parts of the vocal tract – the tongue, lips, and nasal cavity – to amplify and manipulate that sound wave in order to produce different speech sounds. In addition, we also have fine control over the muscles around the vocal folds that allow us to change their thickness, length, and separation. Doing so can change the quality of the sound, producing different *voice qualities* or *phonation types*. Different languages use different phonation types and use them in different ways. This dissertation uses machine learning to examine the properties that distinguish phonation types in different languages and in languages that use phonation in different ways.

This chapter provides a broad overview of phonation. I first give an overview of the anatomy and physiology of phonation, describing the cartilages and muscles of the larynx that are involved in phonation and how they impact vocal fold vibration. I then describe the different types of phonation, how they're used in different languages, and how they have traditionally been measured.

2.1 Laryngeal Anatomy

Phonation occurs in the larynx, which is the upper end of the trachea and below the hyoid bone. The larynx itself is not made of bone but rather of cartilage, muscles, and ligaments. It houses the vocal folds, which vibrate to produce sound. Below, I briefly review the key cartilages and muscles of the larynx that are involved in phonation.

2.1.1 Cartilages of the Larynx

There are nine cartilages in the larynx: the cricoid cartilage, thyroid cartilage, epiglottis, two arytenoid cartilages, two corniculate cartilages, and two cuneiform cartilages. Four of these are particularly important to phonation: the cricoid, thyroid, and two arytenoid cartilages. Figure 2.1 illustrates these four cartilages, which are described below.

Figure 2.1: Laryngeal Cartilages

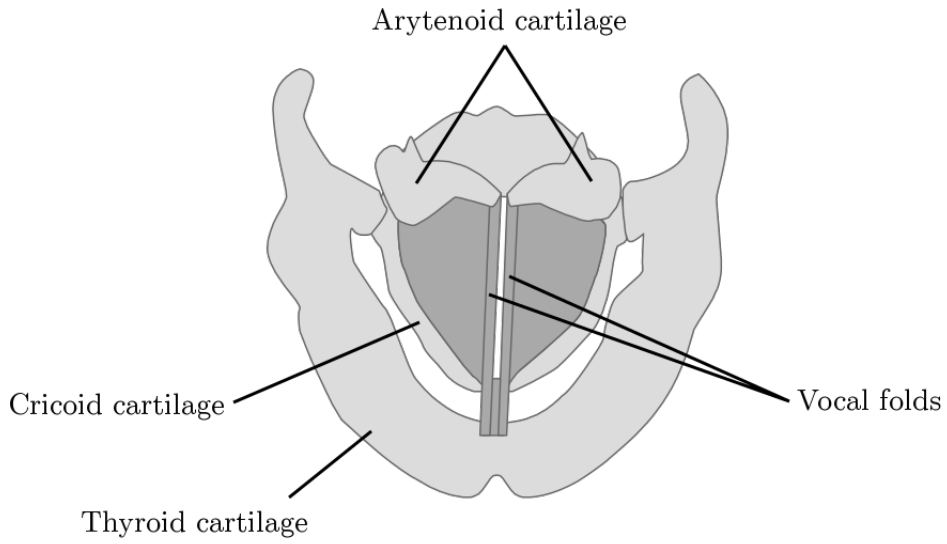


Image courtesy of Dan McCloy

The Cricoid Cartilage

The cricoid cartilage is a ring-shaped cartilage that forms the base of the larynx. The posterior end of the cricoid cartilage is flatter and wider than the anterior side (Gick et al., 2013; Zemlin, 1998).

The Thyroid Cartilage

Sitting on top of the cricoid cartilage is the thyroid cartilage. The thyroid cartilage, the largest of those in the larynx, is shaped like a shield. It consists of two plates that meet on the anterior side. The meeting point of those two plates, the thyroid angle, can be felt from the neck (more so for men than for women) and is often called the *Adam's apple* (Gick et al., 2013; Zemlin, 1998).

The Arytenoid Cartilages

The arytenoid cartilages are two pyramid-shaped cartilages that sit on the wider (posterior) side of the cricoid cartilage. At the anterior end of the base of each arytenoid cartilage is a small extension called the *vocal process*. The vocal processes connect to the *vocal ligaments*, which are part of the vocal folds (Gick et al., 2013; Zemlin, 1998).

2.1.2 Intrinsic Muscles of the Larynx

The muscles that support the larynx and control its location within the trachea are called *extrinsic* muscles, and those that control the movement of cartilages within the larynx (and in doing so, control sound production) are called *intrinsic muscles*. The extrinsic and intrinsic muscles aid in several non-phonatory functions, such as swallowing and breathing. As sound production is the focus of this dissertation, I'll focus here on the subset of the intrinsic muscles that are important in the control of phonation: the thyroarytenoid, lateral cricoarytenoid, interarytenoid, and cricothyroid muscles (Gick et al., 2013; Zemlin, 1998).

Thyroarytenoid Muscles

The thyroarytenoid muscles connect the interior of the point of the thyroid cartilage to the vocal processes in the arytenoid cartilages. Because of where they are connected, the thyroarytenoid muscles appear slightly twisted when adducted (brought together at the midline) but unwound when abducted (pulled away from the midline). The *vocalis muscle*,

the medial part of the thyroarytenoid muscle, is the main body of the vocal folds. When unopposed, contracting the thyroarytenoid muscles relaxes the vocal folds, shortening them. The thyroarytenoid muscles also help close the glottis by drawing the arytenoid cartilages forward, towards the thyroid prominence (Gick et al., 2013; Zemlin, 1998).

Lateral Cricoarytenoid Muscles

The lateral cricoarytenoid muscles connect the cricoid and arytenoid cartilages, running from the outer edge of the cricoid cartilage's ring (between the thyroid prominence and the arytenoid cartilages) to the base of the arytenoid cartilages. Their main function is to rotate the arytenoid cartilages, pivoting the vocal processes inward, which results in moving the vocal folds together (Gick et al., 2013; Zemlin, 1998).

Interarytenoid Muscle

The interarytenoid muscle (sometimes called the arytenoid muscle) is a muscle complex connecting the two arytenoid cartilages. It consists of the transverse interarytenoid muscle and the oblique interarytenoid muscle. The transverse interarytenoid muscle runs horizontally between the posterior surfaces of the two arytenoid cartilages, and the oblique interarytenoid muscles run diagonally from the bottom of one cartilage to the top of the other (the two of them essentially forming an X). This muscle complex serves to approximate (bring together) the arytenoid cartilages (Gick et al., 2013; Zemlin, 1998).

Cricothyroid Muscle

The cricothyroid muscle connects the cricoid cartilage and the thyroid cartilage. It inserts into the anterior end of the cricoid cartilage and to the inferior horn and lower surface of the thyroid cartilage. Its contraction can cause one of two movements. If the extrinsic laryngeal muscles keep the thyroid cartilage in place, contracting the cricothyroid muscle lifts the cricoid cartilage. If, instead, the cricoid cartilage remains in place, contracting the

cricothyroid muscle tilts the thyroid cartilage downward. Whether the cricoid or thyroid cartilage moves, contracting the cricothyroid muscle results in increased distance between the two cartilages, which increases tensions on the vocal folds (Gick et al., 2013; Zemlin, 1998).

2.2 Phonation Physiology: Three Laryngeal Parameters

Producing the quasi-periodic sound wave associated with phonation involves pushing air out of the lungs through the larynx. The glottis – the space between the vocal folds through which the air must pass – is a narrow point in the tube. Air flowing through this narrow opening causes the air pressure to decrease, which in turn pulls the vocal folds together. When the vocal folds are closed, subglottal pressure builds and eventually blows open the vocal folds. This begins the oscillation in the vocal folds that produces a sound wave, which is later modified in the upper portions of the vocal tract (Gick et al., 2013).

The specifics of how the vocal folds vibrate – how quickly, how much, which part, and so on – are controlled by the muscles described above. These muscles, in various combinations, adjust the larynx and therefore impact how the vocal folds vibrate. Laver (1980) describes three "laryngeal parameters" that, when combined in different ways, produce different types of phonation (described in the following section). These parameters – medial compression, adductive tension, and longitudinal tension – are described below and shown in Figure 2.2.

2.2.1 Medial Compression

Medial compression is the amount of force bringing the vocal folds together at the midline. This compression determines how much the vocal folds are approximated, and is controlled by various muscles. The lateral cricoarytenoid muscle rotates the arytenoid cartilages inward, bringing the vocal folds towards the midline. Medial compression also involves adductive tension, described below (Zemlin, 1998).

Figure 2.2: Laryngeal Parameters, after Laver (1980)

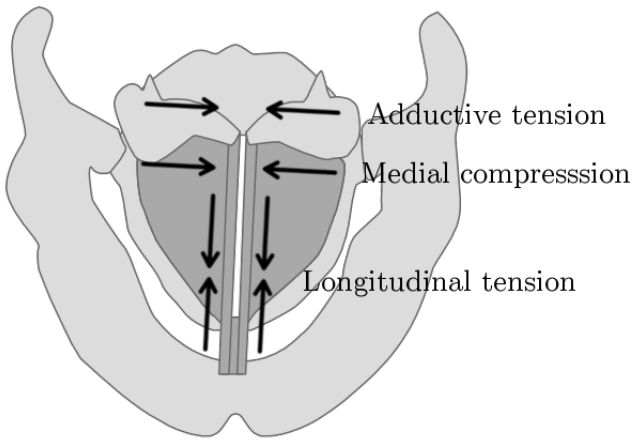


Image courtesy of Dan McCloy

2.2.2 Adductive Tension

Adductive tension is a force that aids medial compression; it is sometimes treated as part of medial compression rather than as a separate force. The interarytenoid muscle regulates adductive tension. Its contraction brings the arytenoid cartilages together, and because the vocal folds ultimately insert into the arytenoid cartilages, this helps increase medial compression (Laver, 1980; Zemlin, 1998).

2.2.3 Longitudinal Tension

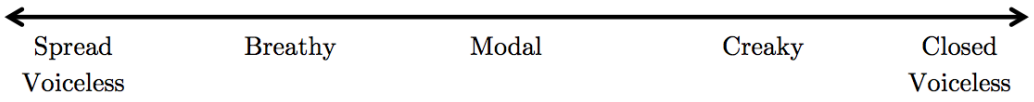
Longitudinal tension is "the degree of stretching force" on the vocal folds (Zemlin, 1998). It is regulated primarily by the thyroarytenoid muscle. When unopposed, its contraction reduces longitudinal tension by shortening the vocal folds, causing them to have more mass per unit length and therefore to vibrate more slowly, resulting in a lower fundamental frequency

(Raphael et al., 2007). (However, when the contraction of the thyroarytenoid muscle *is* opposed, vocal fold tension increases.) The cricothyroid muscle also impacts the length of (and therefore tension on) the vocal folds; its contraction increases the distance between the cricoid and thyroid cartilages, which increases longitudinal tension (Zemlin, 1998).

2.3 The Phonation Continuum

The three forces on the vocal folds described above – medial compression, adductive tension, and longitudinal tension – work together in various ways to produce the full range of phonation types. Phonation types are often depicted along a continuum, as in Figure 2.3. Note that the extremes of the continuum – spread voiceless and closed voiceless – are both voiceless, while the intermediate phonation types – breathy, modal, and creaky – are all types of voicing. It is therefore not a continuum of voicing but rather a continuum of glottal aperture; the vocal folds are completely apart for spread voiceless and completely together for closed voiceless, with the degree of constriction (as well as time spent closed) increasing from left to right. The following sections provide an overview of each phonation type.

Figure 2.3: The Phonation Continuum, after Gordon and Ladefoged (2001)



2.3.1 Spread Voiceless

When the vocal folds are spread, sounds are produced with no vocal fold vibration. Low medial compression and adductive tension keep the vocal folds separated such that they cannot be set into vibration as air passes between them (Johnson, 2011; Raphael et al., 2007).

2.3.2 Breathy Voice

When the vocal folds are held close enough together to vibrate, but not enough to vibrate effectively and regularly, breathy voicing occurs. Breathy voice involves an inefficient vibration of the vocal folds due to generally little muscular tension. Low medial compression and adductive tension means that subglottal pressure cannot build up, and low longitudinal tension results in turbulent airflow as the vocal folds vibrate without much contact, sometimes not closing completely (Gordon and Ladefoged, 2001; Johnson, 2011; Laver, 1980; Zemlin, 1998). Perceptually, breathy voiced sounds can sound whispery, a bit like sighing. Breathy voicing is sometimes called *murmur*.

2.3.3 Modal Voice

Modal phonation, in the middle of the continuum, is what most English speakers think of as typical voicing. It is produced when the vocal folds are positioned to produce maximum vibration, spending approximately equal amounts of time open and closed. The glottal cycle involves opening widely and closing tightly (Johnson, 2011). This posture is achieved with moderate medial compression and adductive tension, and little longitudinal tension (Laver, 1980). Vocal fold vibration occurs in a complex cycle, characterized by bottom-to-top and back-to-front opening and bottom-to-top and middle-outward closure (Gick et al., 2013). The approximately equal open and closed portions allow for maximum vocal fold vibration, which produces the loudest sound on the phonation continuum.

2.3.4 Creaky Voice

Creaky voice is produced by low longitudinal tension but strong medial compression and adductive tension. This results in vocal folds that are short and thick, but drawn together fairly tightly; they are "approximated tightly, but at the same time they appear flaccid along their free borders, and sub glottal air simply bubbles up between them" (Zemlin, 1998). Vibration only occurs on the anterior end (away from the arytenoid cartilages) (Ladefoged

and Johnson, 2015) and the vibration is slow, with the vocal folds spending more time together than apart (Johnson, 2011). It's often described as sounding froggy or like a "rapid series of taps, like a stick being run along a railing" (Catford, 1964). Creaky voiced sounds are sometimes described as *laryngealized* or *glottalized*.

English creaky voicing has recently been of great interest to sociolinguistics, and several studies have identified various subtypes of creaky voicing. Prototypical creaky voice is described above; it has a fairly high degree of glottal constriction and a low and irregular fundamental frequency. Other types of creaky voicing exhibit some but not all of these properties. *Non-constricted voice* or *Slifka voicing* involves less glottal constriction than prototypical creaky voicing. It sounds somewhat breathy (due to less glottal constriction) but still has a low and irregular fundamental frequency. Another type of creaky voicing, *vocal fry*, has glottal constriction and has a low fundamental frequency, but its fundamental frequency is regular (Keating and Garellek, 2015). These subtypes of creaky voice show some but not all signs of prototypical creaky voicing.

2.3.5 Closed Voiceless

Finally, sounds on the far right side of the phonation continuum are also voiceless, but for a different reason: when the glottis is completely closed, no air can pass through it and therefore no phonation can occur. The glottis can be closed by tight contraction of the interarytenoid muscles (high adductive tension) and cricoarytenoid muscles (high medial compression) (Gick et al., 2013). This produces a glottal stop, /ʔ/. Complete glottal closure often accompanies the stopping of air elsewhere in the vocal tract during other stop consonants.

2.3.6 Phonation Outside the Scope of this Dissertation

Constraining all phonation types to the continuum shown in Figure 2.3 is, of course, an oversimplification. The continuum can be broken down more finely and other voice qualities can be produced using various different combinations of laryngeal muscles. While this dissertation focuses on breathy, modal, and creaky voicing, I briefly review several other

phonation types. I do not include them in this study primarily because they are not typically used to describe phonation in the languages included in this dissertation (see Chapter 4). Additionally, definitions for some of these terms vary from study to study.

Tense vs. Lax Voice

Some languages contrast tense and lax voicing (not to be confused with tense and lax vowels). Kuang and Keating (2013) describe tense voice as falling between modal and creaky voice on the phonation continuum, and lax voice between modal and breathy. Essentially, tense and lax voice are less extreme versions of creaky and modal voice, respectively. Voicing contrasts and registers in many languages of Southeast Asia have been described as tense vs. lax, including Jingpho, Hani, Yi, and Wa (Maddieson and Ladefoged, 1985).

Stiff vs. Slack Voice

Similar to the tense vs. lax contrast, stiff vs. slack voice are less extreme versions of creaky and breathy voice. Ladefoged and Maddieson (1996) describe them in terms of how tightly the vocal folds vibrate and how much airflow is involved; slack voice has looser vibration and a higher rate of airflow than modal voice, and stiff voice has stiffer vocal folds and a lower rate of airflow than modal voice. Languages that employ stiff and/or slack voicing include Mpi (modal vs. stiff) (Ladefoged and Maddieson, 1996) and Javanese (slack vs. modal vs. stiff)[1] (Fagan, 1988). No known language contrasts creaky voice with stiff voice or breathy voice with slack voice.

Harsh or Pressed Voice

Edmondson and Esling (2006) describe harsh voice (also called pressed voice) as produced with irregular vibration (as in creaky voicing), but without the slackness associated with

[1]The phonation contrast in Javanese has also been described as tense vs. lax and *light* vs. *heavy*.

creaky voicing, so it has a higher pitch than creaky voicing. The ventricular folds[2] can impede vibration of the true vocal folds due to generally high tension in the larynx. They also note that harsh/pressed voice is sometimes conflated with creaky voicing in the literature. Gobl and Ní Chasaide (2003) caution that what's described as harsh often varies from study to study. For their purposes, they describe harsh voice as being a variant of tense voice but with more extreme settings as well as aperiodic vibration. Bai is described as contrasting breathy, modal, and harsh voice, and Somali has two registers, one of which is described as harsh (Edmondson and Esling, 2006).

Whisper Voice

Whisper (or whispery) voice, like breathy voice, involves inefficient vocal fold vibration due to low adductive tension and produces a turbulent sound. Unlike breathy voicing, however, it's characterized by fairly high medial compression (Gick et al., 2013; Gobl and Ní Chasaide, 2003). Whisper voice is often not distinguished from breathy voice.

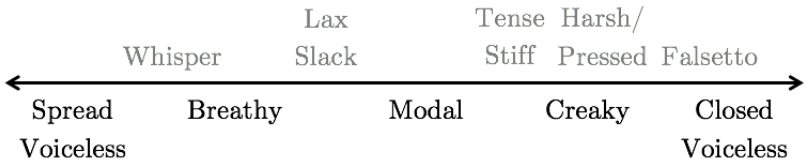
Falsetto

Falsetto, while often described by its pitch range, also exhibits a specific voice quality. The exact laryngeal mechanism involved in falsetto is not entirely clear, but it's generally understood to involve strong medial compression (though adductive tension is not required), such that only a small portion of the vocal folds can vibrate. The cricothyroid muscle is also used in falsetto to elongate the vocal folds. Unlike creaky voice, it has a high pitch (Zemlin, 1998).

Figure 2.4 shows an updated version of the phonation continuum, including the eight phonation types described above. Their locations on the continuum are approximate, as some of the phonation types don't fall neatly on a continuum of glottal aperture.

[2]The ventricular folds, also known as the false vocal folds or the vestibular folds, are mucous membranes that sit above the true vocal folds. They're generally not involved in phonation, but they can be set into vibration, as in Tuvan throat singing.

Figure 2.4: The Updated Phonation Continuum



Disordered Speech

As Ladefoged (1983) famously noted, "one person's voice disorder is another person's phoneme." While this dissertation focuses on the *linguistic* uses of phonation types, breathy and creaky voice (as well as other phonation types) can also be caused by speech disorders. Dysphonia is a general term for pathological non-modal voicing, though it is more often associated with breathiness than with creakiness. Dysphonia is a description of the symptoms (non-modal phonation) and the causes can vary. Vocal nodules, which are often caused by vocal abuse, can result in breathy voicing because the nodule prevents the vocal folds from closing completely. Another cause of dysphonia is laryngitis, the swelling of the vocal folds, which can be caused by colds, allergies, and various other ailments, as well as by overuse. Laryngitis results in thickened vocal fold mucosa which leads to aperiodic and inefficient vibration (Zemlin, 1998). As disordered speech is extralinguistic and essentially idiosyncratic, it will not be included in this dissertation.

2.4 The Function of Non-Modal Phonation in Languages

As noted above, this dissertation focuses on non-modal phonation that is used *linguistically*. Phonation can be used in various linguistically meaningful ways; I divide these into five categories, described below.

2.4.1 Contrastive Phonation

Phonation contrasts can appear as a property of consonants or vowels. When languages use phonation *contrastively* or *phonemically*, changing a phone's phonation type is enough to change the meaning of a word. For example, Gujarati, an Indo-European language spoken in India, contrasts breathy and modal vowels. The words [bar] *twelve* and [ba̤r] *outside* form a minimal pair, differing only in the phonation type of the vowel. While many languages contrast voiced and voiceless consonants, contrasting different types of phonation (on consonants or vowels) is much rarer and contrasting voicing on vowels is not clearly attested (Gordon and Ladefoged, 2001). Some languages employ a two-way contrast, contrasting modal voice with either breathy or creaky voice, while others employ a three-way phonation contrast.

2.4.2 Tonal Phonation

Phonation can play an important role in register or tonal systems. Tone languages use pitch contrastively; changing the fundamental frequency changes meaning. Some tone languages use phonation augmentively with tones, often employing creaky voicing to help mark a low tone. One example of this is Mandarin Chinese, which optionally but frequently marks the third tone (low dipping) with creaky voicing (see Chapter 4).

Register languages are similar to tone languages but instead of a pitch-based contrast, the contrast is in "a complex of phonetic qualities that includes pitch height, pitch contour, phonation mode, intensity, duration, and syllable structure"[3] (Gruber, 2011). Pitch alone cannot adequately describe the contrast in a register language, nor can phonation, but the combination of pitch and phonation often can describe the contrast.

---

[3]This particular set of phonetic qualities is specific to Burmese; other register languages employ different constellations of qualities.

2.4.3 Allophonic Phonation

Non-modal phonation can be allophonically induced by a neighboring phone. Vowels adjacent to /h/ are often allophonically breathy, and consonants can have breathy allophones (including [fi] as an allophone of /h/); vowels adjacent to glottal stops are often allophonically creaky, and glottal stops can be allophonically realized as creaky voicing (Redi and Shattuck-Hufnagel, 2001). Non-modal phonation can also spread to nearby segments (Gordon and Ladefoged, 2001).

2.4.4 Prosodic Phonation

Non-modal phonation can be used to mark prosodic boundaries. English speakers often mark the beginnings and ends of prosodic units, including words, with non-modal phonation (Pierrehumbert and Talkin, 1992; Redi and Shattuck-Hufnagel, 2001). In one of the few studies of prosodic phonation in languages besides English, Lehiste (1965) examined the properties of boundaries in Finnish, Czech, and Serbo-Croatian. All three languages use non-modal phonation to mark prosodic boundaries: Finnish speakers laryngealize sounds when prosodic and word boundaries occur between two vowels, Czech speakers mark prosodic boundaries with either a glottal stop or with "breathy and irregular phonation," and Serbo-Croatian speakers use laryngealization at the beginning of prosodic units. Some prosodically-induced non-modal phonation may actually be caused by prosodic intonation; as discussed above, low tones are often marked by creaky voicing.

2.4.5 Sociolinguistic Phonation

Finally, phonation can be used sociolinguistically. The vast majority of sociolinguistic studies of phonation have focused on various dialects of English; little is known about sociolinguistic uses of non-modal phonation in other languages. In English, non-modal phonation has been associated with gender (Podesva et al., 2015; Yuasa, 2010; Podesva, 2013), socioeconomic status (Yuasa, 2010), age (Podesva et al., 2015; Yuasa, 2010), and various emotions and

personality traits (Gobl and Ní Chasaide, 2003; Mendoza-Denton, 2011). More detail about sociolinguistic uses of phonation in English is provided in Chapter 4.

2.5 Measuring Phonation

Three general approaches are used in measuring phonation: imaging, electroglottography, and acoustic measurement. Imaging uses technology (such as a laryngoscope or an ultrasound machine) to directly view the larynx during phonation. Electroglottography uses an electrical current to track glottal aperture during phonation. Acoustic measurements extract information correlated with phonation from a recorded signal. These three techniques, described in more detail below, have various advantages and disadvantages.

2.5.1 Imaging

The most direct way of studying phonation is to see it in action. The glottis can be viewed using an instrument called a laryngoscope, a tube which is inserted into the upper vocal tract by way of the mouth or nose. A video camera on the tip of the tube records movement in the glottis during phonation (Zemlin, 1998). Ultrasound can also be used to visualize the glottis during phonation by pulsing high frequency sound waves into tissue and recording the echoes. Similarly, an MRI machine can show phonation by using magnetic fields to image the glottis.

While these techniques provide a clear picture (literally) of the glottis during phonation, they have some limitations. The technology involved in laryngeal imaging is expensive, not always transportable, and it requires a good deal of medical training to both conduct the imaging and interpret the results. (Laryngoscopy is also slightly invasive.) For these reasons, imaging is often not a practical route for linguists to study phonation.

2.5.2 Electroglottography

Electroglottography (EGG, also called laryngography) does not directly image the larynx, but instead uses an electrical current to measure the amount of contact between the vocal

folds. Two small electrodes are placed on either side of a speaker's neck such that a small electrical current sent between them must pass through the vocal folds. Electrical resistance is greater in air than in tissue, so the current is stronger when the vocal folds are closed than when they are open. The most common measures used in EGG studies are closed quotient and open quotient, which represent the percent of time the vocal folds spend open and closed during a glottal cycle, respectively. These two measures, along with various other EGG measurements, have been used to study phonation in White Hmong (Esposito, 2012), Takhian Thong Chong (DiCanio, 2009), Bo, Yi, and Hani (Kuang, 2013), and more, as well as in disordered speech (Hall, 1995).

EGG has several advantages over direct imaging of phonation. An electroglottograph is a much less expensive piece of equipment than a scope, an ultrasound machine, or an MRI. It's portable and is non-invasive (the electrical current is too small to be felt). It does require some training to use (correct placement of the electrodes is crucial) and some analysis to interpret (several software packages exist for this purpose). Speech involving large amounts of upward or downward movement of the larynx is problematic, as the electrodes remain fixed in place while the larynx moves. Finally, while EGG provides information about the degree of vocal fold contact, it does not provide information about where that contact is happening.

2.5.3 Acoustic Measurements

Finally, acoustic measures can be used to study phonation. Different aspects of phonation are reflected differently in the signal. For example, pitch is measured as the Fundamental Frequency of the signal (f_0); the cycle-to-cycle variation in period length as Jitter; and the speed of glottal closure as Spectral Tilt. These measurements can be extracted fairly straightforwardly (and often automatically) from a recording using acoustic analysis programs like Praat (Boersma and Weenink, 2016) and VoiceSauce (Shue et al., 2011).

Of all three ways of measuring phonation, acoustic measurements are, in many ways, the most straightforward to perform. All that's needed is a high quality recording and a computer

running a (free) acoustic analysis program. While a good quality recording requires both skill and good equipment, both are easier to come by in terms of effort and cost than what's required in using an MRI machine or an electroglottograph. Perhaps most importantly, acoustic recordings are the bread and butter of linguistics; the number of high quality recordings that are publicly available for research far outnumbers the available phonation imagery and electroglottographic data. This means that many already existing corpora can be used to study phonation.

In this dissertation, I use acoustic measurements to study phonation in six different languages. Large and varied corpora of high quality acoustic data sets are readily available; this allows me to use far more data than I would be able to collect myself, and include languages that would be otherwise difficult to find speakers for. For those six languages, I use machine learning to identify which acoustic properties distinguish phonation types. I describe the acoustic properties under consideration in Chapter 3 and the languages studied in Chapter 4. Chapters 6 through 11 report the results for each language.

Chapter 3

FEATURES

This chapter describes the acoustic and contextual measures that have been shown to correlate with phonation type and that will be considered for use as *features* in the machine learning models. Recall that in the context of machine learning, features are measures that help distinguish between the classes in question; in this dissertation, they are properties specific to breathy, modal, and creaky voice. How the features are used in machine learning is described in more detail in Chapter 5. Here, I describe each measure, previous findings linking it to phonation, and the various implementations or calculations that can be used. Many of the measures have several possible variants, each of which may contribute slightly different information to the model; here I outline thirteen categories of measures. Their many variants result in 39 features under consideration, some of which may ultimately prove to be redundant or not sufficiently discriminating. The full set of features is summarized at the end of the chapter in Table 3.8.

3.1 Spectral Tilt

Spectral Tilt is "the degree to which intensity [in the spectrum] drops as frequency increases" (Gordon and Ladefoged, 2001). This drop is quantified by comparing the amplitude of harmonics across frequencies. Typically, the dropoff is most extreme in breathy voice and flattest in creaky voice (Gordon and Ladefoged, 2001). The difference in tilt between phonation types is generally explained by glottal constriction and the speed of vocal fold closure during phonation (Keating and Garellek, 2015; Khan, 2012). The vocal folds come together more quickly in creaky voice than in modal voice, and more slowly in breathy voice than in modal voice. Their quick closure during creaky voice produces a high amplitude

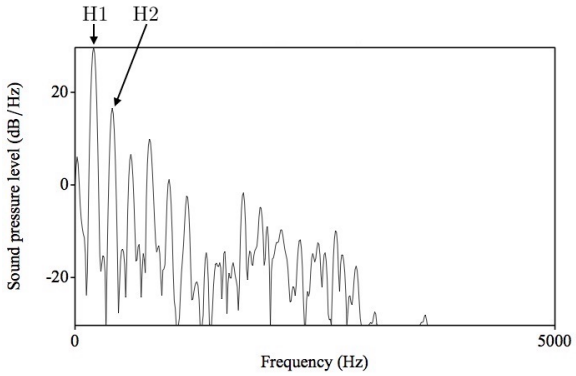
transient which, in turn, excites more harmonics throughout the spectrum; this results in a spectrum that is flatter than in modal voice. In contrast, the slow closure of breathy voice produces a weaker transient that does not excite higher harmonics, causing a steeper drop off in the harmonics than in modal voice (Kirk et al., 1993).

Figure 3.1 shows the spectral slices for breathy, modal, and creaky tokens of /aɪ/ as uttered by the same female speaker. The spectrum is drawn over the middle 80% of each. The drop between H1 and H2[1] is noticeably steeper in the breathy vowel (a) than in the modal vowel (b), and steeper in the modal vowel than in the creaky vowel (c). Additionally, the energy drops off at different frequencies in the different voice qualities – around 3000 Hz in the breathy vowel, 4000 Hz in the modal vowel, and 5000 Hz in the creaky vowel.

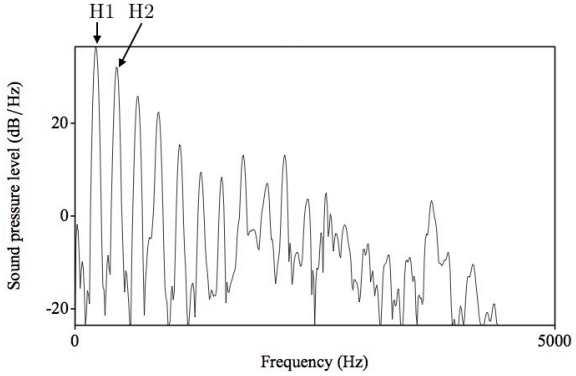
Fant (1980) was among the first to observe a relationship between spectral shape and voice quality. Jackson et al. (1985) went on to attempt to quantify the relevant spectral shapes. An initial strategy included many harmonics, but they found that "in many cases it is impossible to fit a statistically significantly straight line . . . through even the peaks of the first harmonics." At the same time (and, in fact, in the same volume), Ladefoged and Antoñanzas-Barroso (1985) explored possible acoustic correlates of breathy voice. They note the difference in the speed of vocal fold closure between breathy and modal voice, resulting in lower energy in higher harmonics for breathy voice. They employ two early measures of Spectral Tilt, f_0 – F1 and f_0 – H2, to study modal and breathy vowels in !Xóõ, finding that breathy vowels have a steeper drop between the measures than modal vowels do.

Today, Spectral Tilt is widely considered the best measure of phonation; it has successfully differentiated voice qualities in a wide variety of languages, though which pair of harmonics performs best varies between languages. Table 3.1 lists some of the languages for which Spectral Tilt is a useful measure in distinguishing phonation types. Jalapa Mazatec, for instance, has a three-way phonation contrast (breathy, modal, creaky). While the Spectral Tilt of each phonation type varies from speaker to speaker, the pattern remains consistent

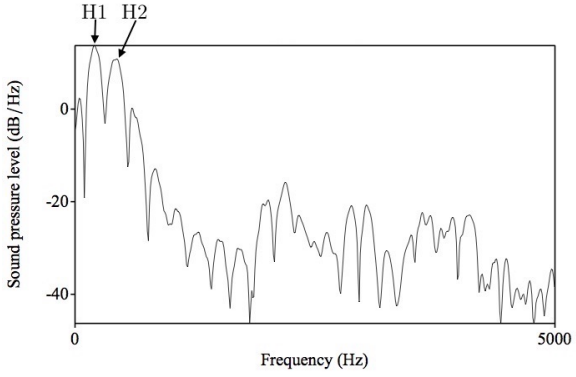
[1]I am following the linguistics convention of naming harmonics beginning with H1, rather than the engineering convention of beginning with H0.



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.1: Spectral tilt in breathy, modal, and creaky /aɪ/

across speakers: creaky voice has a lower Spectral Tilt than modal voice, and breathy voice has a higher Spectral Tilt than modal voice (Kirk et al., 1993). It is worth noting that the majority of the languages listed in Table 3.1 use phonation contrastively, though it successfully characterizes non-contrastive phonation in English as well.

Table 3.1: Languages With Phonation Types Distinguished By Spectral Tilt

Language	Phonation	Source(s)
Coatzospan Mixtec	M, C	Gerfen and Baker (2005)
Chong	B, M	Blankenship (2002)
English	M, C	Garellek and Keating (2015)
Green (H)Mong	B, M, C	Andruski and Ratliff (2000)
Gujarati	B, M	Keating et al. (2011); Khan (2012)
Jalapa Mazatec	B, M, C	Garellek and Keating (2011); Kirk et al. (1993)
San Lucas Quiaviní Zapotec	B, M, C	Gordon and Ladefoged (2001)
Santa Ana Del Valle Zapotec	B, M, C	Esposito (2010)
White Hmong	B, M, C	Esposito (2012); Keating et al. (2011)
!Xóõ	B, M	Ladefoged and Antoñanzas-Barroso (1985)

As a powerful measure of phonation, both contrastive and non-contrastive, Spectral Tilt may be a key feature in distinguishing phonation types using machine learning. VoiceSauce provides seven Spectral Tilt measures, all of which will be included as features in this study. Harmonics are calculated pitch-synchronously based on the default STRAIGHT f_0 algorithm.[2] Using that calculation, a maximum search algorithm finds harmonic magnitudes. Spectral Tilt can then be measured by calculating the difference between these various peaks. VoiceSauce outputs both uncorrected harmonic measurements and measurements that have been corrected for the effects of formant frequencies and bandwidths. Only corrected harmonics will be used in this study, as uncorrected harmonics will simply differentiate vowel qualities, rather than voice qualities. The seven Spectral Tilt measures that I will include as features are listed in Table 3.2.

[2]Because the calculations of the harmonics are dependent on the calculation of f_0, their measures are only as good as their pitch tracker. The STRAIGHT algorithm has proven to be fairly robust, even during non-modal phonation, but it is not perfect. (See Section 3.8 for more information about pitch tracking.)

Table 3.2: Spectral Tilt Measures in VoiceSauce (Shue, 2010)

Measure	Description
H1* – A1*	The amplitude of the first harmonic minus the amplitude of the first formant peak
H1* – A2*	The amplitude of the first harmonic minus the amplitude of the second formant peak
H1* – A3*	The amplitude of the first harmonic minus the amplitude of the third formant peak
H1* – H2*	The amplitude of the first harmonic minus the amplitude of the second harmonic
H2* – H4*	The amplitude of the second harmonic minus the amplitude of the fourth harmonic
H4* – 2k*	The amplitude of the fourth harmonic minus the amplitude of the harmonic closest to 2 kHz
2k* – 5k	The amplitude of the harmonic closest to 2 kHz minus the amplitude of the harmonic closest to 5 kHz

** Indicates measures that have been corrected for formant frequencies and bandwidths in VoiceSauce*

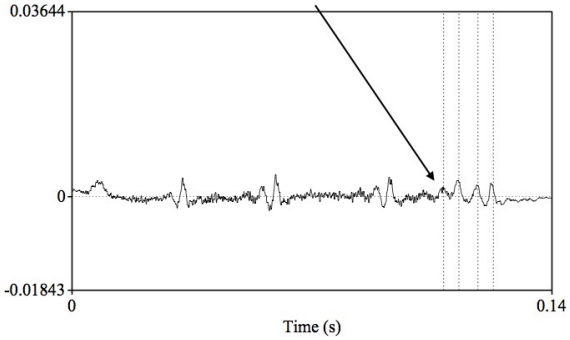
3.2 Jitter

Jitter is a measure of "cycle-to-cycle variations of fundamental frequency" (Farrús et al., 2007); specifically, it is the standard deviation of the time between glottal pulses. Non-modal phonation is often characterized by irregularly timed glottal pulses, resulting in a higher Jitter in non-modal signals than in modal signals (Gordon and Ladefoged, 2001). Figure 3.2 shows the waveforms and spectrograms for breathy, modal, and creaky tokens the vowel /æ/, as uttered by a female speaker of American English. Glottal pulses, drawn by Praat (Boersma and Weenink, 2016), are shown on the waveform as dotted vertical lines. In the modal vowel (b), the pulses are evenly spaced across the vowel and consistently correctly located by Praat. In contrast, the breathy (a) and creaky (c) vowels have irregularly spaced glottal pulses, and many are missed by Praat; this irregularity translates to higher Jitter.

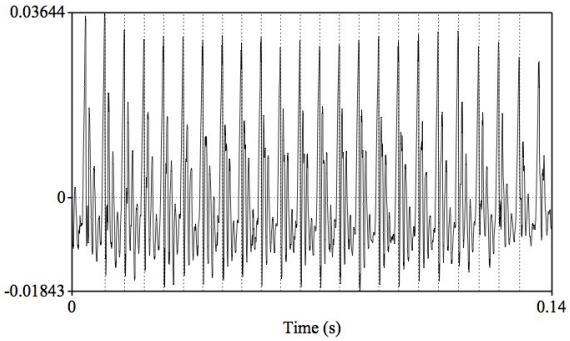
Measuring Jitter in speech signals began in the early 1960s, though the term *Jitter* did not appear in the acoustics literature until several years later.[3] In an early study, Lieberman (1961) measured the duration of pitch periods in sentences read with a variety of emotions, finding that the duration of consecutive periods was not consistent and that the difference between period duration varied by "emotional mode." Risberg (1961) provided another early description of Jitter in a pilot study involving "the range and the rate of change of the fundamental frequency, f_0" in English and Swedish, though his results were not analyzed.[4]

[3]The first studies of the characteristics of speech using waveforms occurred nearly 100 years earlier. Édouard-Léon Scott de Martinville filed a patent in 1857 for the *phonautograph*, which used sound vibrations in air to move a stylus that etched the signal onto a blackened glass plate (Brock-Nannestad and Fontaine, 2008). His letter to the Académie des Sciences provides an early account of Jitter: "Le timbre, considéré en géneral, se compose de trois éléments: 1° la forme des vibrations; 2°leur régularité ou leur irrégularité et leur isochronisme ou leur non-isochronisme; 3° le détail de la vibration" (Feaster, 2010).

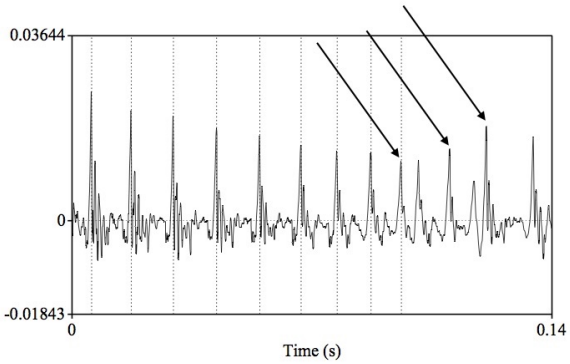
[4]Risberg (1961) describes the process of his acoustic analysis: "The fundamental frequency was recorded on a Mingograph oscillograph by means of a modified Grützmacher analyzer and the FO curve was sampled by hand at successive 25 msec intervals." I would like to note how much methods in this field changed over the 100 years between the phonautograph and Risberg's study, and again over the 57 years between his study and the present one.



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.2: Jitter in breathy, modal, and creaky /æ/

(Arrows indicate the approximate start of changes in glottal period length)

Not long after, Lieberman (1963) studied pitch perturbations in normal speakers and speakers with "pathologic growths on their vocal folds." He proposed a "perturbation factor," defined as the percent of a speaker's perturbations that were at least 0.5 ms, to help identify speakers with laryngeal pathologies.

More recently, Jitter has been used in the analysis of languages with contrastive non-modal phonation. Gerfen (2013) used Jitter in his analysis of Coatzospan Mixtec vowel glottalization, which is "characterized by a brief period of creaky voicing." Though he found variation in Jitter across speakers for both modal and glottalized vowels, he found no instances of a modal vowel having higher Jitter than a creaky vowel for a given speaker. This suggests that normalized Jitter may be indicative of phonation type. A study of Jalapa Mazatec found that the mean variance between pulses was .08 ms for modal vowels and 9.1 ms for creaky vowels. Similar to the Coatzospan Mixtec findings, the authors note that "although there is considerable variation in the degree of creaky voice among speakers, nevertheless for all speakers the [jitter] value for creaky voice is higher than that for modal voice for any speaker" (Kirk et al., 1984). As both Coatzospan Mixtec and Jalapa Mazatec use phonation contrastively, these two studies provide only one side of the story; they show that Jitter is an acoustic characteristic of glottalization or creak, but do not speak to how salient a cue Jitter is to listeners in judging phonation type.

Many of the studies that have employed Jitter to distinguish phonation types, including the two introduced above, have done so for languages with contrastive phonation. Jitter is used to describe the division between phonation types, rather than to determine where to draw that line; some studies have found that Jitter is *not* used by listeners to draw that line. Kreiman and Gerratt (2005) found that the association between Jitter and voice quality is "not sufficiently explanatory to justify continued reliance on Jitter . . . as [an index] of voice quality." Human perception of irregular voicing is instead based on spectral noise, a product of aperiodicity. For this reason, Jitter has fallen out of favor among phoneticians as a measure of phonation and has generally been replaced by measures of noise.

Despite Jitter's lack of salience to human listeners, it remains relevant to the present

study, as a computer rather than a human will be separating phonation types; Jitter is a direct quantification of aperiodicity, an important characteristic of non-modal phonation. This is a valuable acoustic measure and will be used as a feature in my models. Praat offers five different Jitter calculations, which are described along with their formulas in Table 3.3 (Boersma and Weenink, 2016). *Difference of Differences of Periods (DDP)* is the only measure that will be excluded from this study; its value is always three times the value of *Relative Average Perturbation (RAP)*, making it a redundant feature. The remaining four Jitter calculations (*Local Jitter*, *Local Absolute Jitter*, *RAP*, and *PPQ5*) will all be used as features in the models.

3.3 Intensity

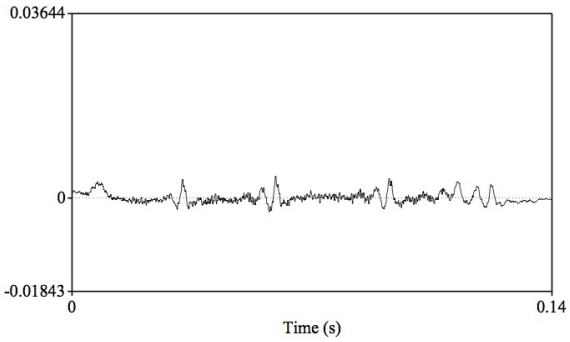
Intensity is the amount of energy in the acoustic signal, which listeners perceive as loudness. Creaky voice and breathy voice both have lower intensity than modal voice (Gordon and Ladefoged, 2001). Creaky voice has less energy than modal voice simply because the vocal folds spend more time closed in creaky voice than in modal voice (a lower open quotient), and air cannot pass through the vocal folds during closure. Breathy voice has less energy than modal voice because there is lower pressure buildup while the vocal folds are fairly open (Laver, 1980).

Figure 3.3 shows the same three tokens of /æ/ that were used to illustrate Jitter. The amplitude in the modal vowel (b) is much higher throughout the vowel than the amplitude of the breathy (a) or creaky (c) vowel. Note that the peak amplitude drops in the same portions of the vowel that have less periodic glottal pulses – intensity is lower towards the end of the breathy vowel (a), only briefly at the very end of the modal vowel (b), and near the middle of the creaky vowel (c).

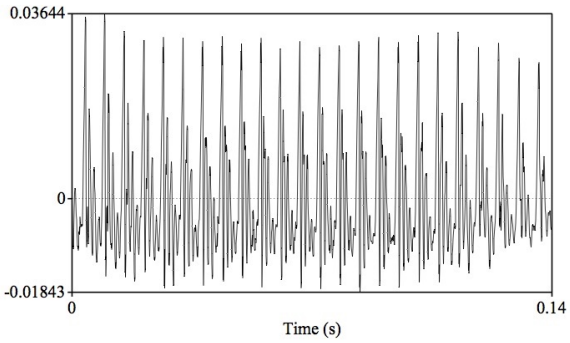
Khan (2012) found no significant effect of RMS Energy in distinguishing breathy vowels from modal vowels in Gujarati. In Kui, a register language with breathy and modal voice,

Table 3.3: Jitter Measures in Praat (Boersma and Weenink, 2016)

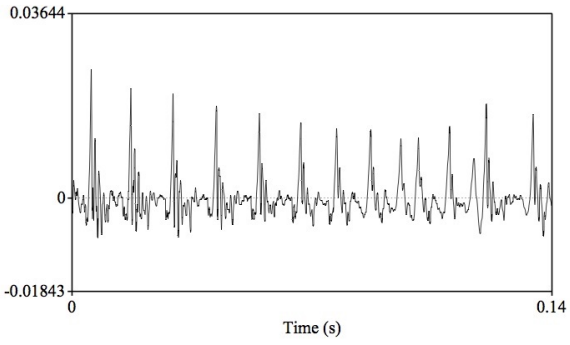
Measure	Description
Local Jitter	"the average absolute difference between consecutive periods, divided by the average period." $\frac{\Sigma_{i=2}^{N}\lvert T_i - T_{i-1}\rvert/(N-1)}{\Sigma_{i=1}^{N} T_i/N}$
Local Absolute Jitter	"the average absolute difference between consecutive periods, in seconds" $\Sigma_{i=2}^{N}\lvert T_i - T_{i-1}\rvert/(N-1)$
Relative Average Perturbation (RAP)	"the average absolute difference between a period and the average of it and its two neighbours, divided by the average period" $\frac{\Sigma_{i=2}^{N-1}\lvert T_i - (T_{i-1}+T_i+T_{i+1})/3\rvert/(N-2))}{\Sigma_{i=1}^{N} T_i/N}$
Five-Point Period Perturbation Quotient (PPQ5)	"the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period." $\frac{\Sigma_{i=3}^{N-2}\lvert T_i - (T_{i-2}+T_{i-1}+T_i+T_{i+1}+T_{i+2})/5\rvert/(N-4))}{\Sigma_{i=1}^{N} T_i/N}$
Difference of Differences of Periods (DDP)	*"the average absolute difference between consecutive differences between consecutive periods, divided by the average period."* $\frac{\Sigma_{i=2}^{N-1}\lvert (T_{i+1}-T_i)-(T_i-T_{i-1})\rvert/(N-2))}{\Sigma_{i=1}^{N} T_i/N}$



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.3: Intensity in breathy, modal, and creaky /æ/

Thongkum (1987) found that breathy vowels have a lower intensity than modal vowels, though these measures appear to have been based on amplitude height rather than RMS Energy. She also found that creaky voice has a lower intensity than modal or breathy voice in Chong.

Despite its inclusion in the above studies, intensity is often left out of phonation studies because it depends on a careful and consistent recording environment. If the source-to-microphone distance is inconsistent, intensity is no longer a meaningful measure. However, the largest corpus used in this study, ATAROS (Freeman, 2015) (described in Chapter 4), controlled for this potential issue by using a fixed microphone distance both between and within speakers. With this control, intensity may prove to be a useful feature in discriminating phonation types, at least for the English data, and will be included in the model.

Intensity can be calculated as the root mean square (RMS) energy of a sound. It is calculated as follows: at each window, the amplitude, (a), is squared and the square root of the average of those samples is calculated (Johnson, 2011).

$$RMSEnergy = \sqrt{\frac{a_1^2+a_2^2+a_3^2...+a_n^2}{n}}$$

RMS Energy is a particularly robust and widely used measure of energy because it is closely correlated with human perception of intensity and reflects the sound's overall energy, which is not necessarily the same as its peak amplitude, particularly in noisy and complex signals like speech (Johnson, 2011). In this study, I will use VoiceSauce's RMS Energy measure, which is "calculated at every frame over a variable window equal to five pitch pulses by default. The variable window effectively normalizes the energy measure with f_0 to reduce the correlation between them" (Shue et al., 2011).

3.4 Shimmer

Like Jitter, Shimmer is a measure of variation between cycles; while Jitter measures cycle-to-cycle variation in fundamental frequency, Shimmer measures variation in amplitude (Farrús

et al., 2007). Intensity is lower in non-modal phonation than in modal phonation (see Section 3.3), and more variation in amplitude may reflect changes in the duration of glottal periods (related to Jitter) or the amount of vocal fold closure. It can therefore be used to describe different voice qualities (Kreiman and Gerratt, 2005). Shimmer can be seen in Figure 3.4; the peak amplitude changes throughout the breathy (a)[5] and creaky (c) vowels, but remains relatively constant throughout the modal vowel (b).

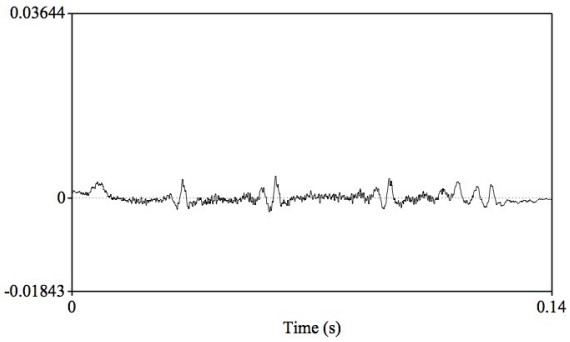
Horii (1980) notes that "In contrast to the data available on Jitter, there is a paucity of information on shimmer . . . in sustained phonation"; though this observation was made nearly thirty years ago, Shimmer remains significantly less studied than Jitter in the phonation literature. Kreiman and Gerratt (2005) found that Shimmer was not useful in indexing *perceived* voice quality, except for its "acoustic contributions to the overall pattern of spectrally shaped noise in a voice." The majority of studies using Shimmer to describe voice quality come from the clinical literature. Gramuglia et al. (2014) found that Shimmer was higher for children with vocal nodules than for those without. Eadie and Baylor (2006) found that Shimmer was predictive of listener judgments of roughness in dysphonic speech, and Hillenbrand (1988) found that increasing Shimmer on synthesized speech led to increased perceptions of roughness. Though perceived roughness is not necessarily the same as a linguistic voice quality, the two categories may overlap enough to warrant including Shimmer as a feature in my model.

Praat offers six Shimmer measurements, described in Table 3.4. Of these six, I will be using the first five as features; the last, DDA Shimmer, is always three times the value of APQ3 Shimmer, making it redundant.

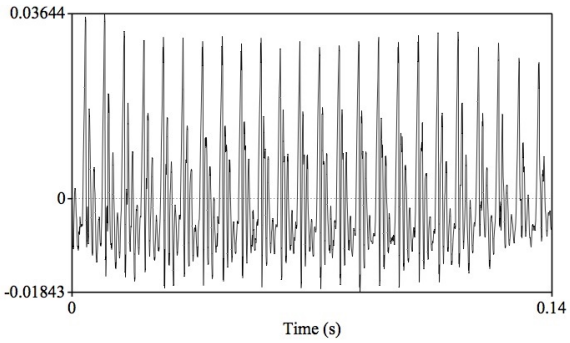
3.5 Harmonics-to-Noise Ratio

Harmonics-to-Noise Ratio (HNR) is, as the name suggests, the ratio of the amplitude of harmonics to the amplitude of noise in a given signal. Both breathy and creaky voice have

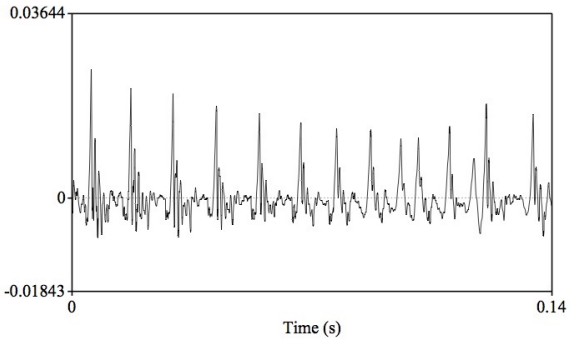
[5]The overall intensity of the breathy vowel is quite low; though the changes in intensity appear much less dramatic than in the creaky vowel, they are present relative to the overall intensity.



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.4: Shimmer in breathy, modal, and creaky /æ/

Table 3.4: Shimmer Measures in Praat (Boersma and Weenink, 2016)

Measure	Description
Local Shimmer	"the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude"
Local Shimmer, dB	"the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20"
Shimmer, Three-Point Amplitude Perturbation Quotient (APQ3)	"the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude"
Shimmer, Five-Point Amplitude Perturbation Quotient (APQ5)	"the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude"
Shimmer, 11-point Amplitude Perturbation Quotient (APQ11)	"the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude"
Shimmer, Difference of Differences of Amplitudes (DDA)	"the average absolute difference between consecutive differences between the amplitudes of consecutive periods"

more noise than modal voice; this can be caused by the characteristic decreased periodicity of both types of non-modal phonation, as well as the increased aspiration of breathy voice (Gordon and Ladefoged, 2001).

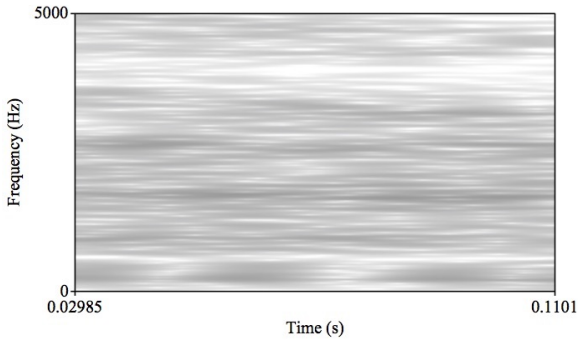
Harmonics are defined as peaks occurring at integer multiples of the fundamental frequency, while noise is everything else – "energy that does not appear at the frequency loci of the harmonics" (de Krom, 1993). In voiced signals, the harmonic peaks are generally better-defined (higher amplitude relative to the rest of the signal) than the noise, but this is not necessarily the case, particularly in non-modal phonation; HNR is therefore lower (reflecting more noise) in non-modal signals than in modal ones.

Because aperiodicity introduces noise, HNR correlates closely with Jitter. However, HNR appears to be more perceptually salient than Jitter. de Krom (1993) found that "HNR decreases almost linearly with increasing noise levels or increasing Jitter." He notes that while HNR may be useful in characterizing voice quality, "it cannot be directly interpreted in terms of underlying glottal events or perceptual characteristics" because it can be caused by various phenomena, including aperiodicity and incomplete vocal fold closure.

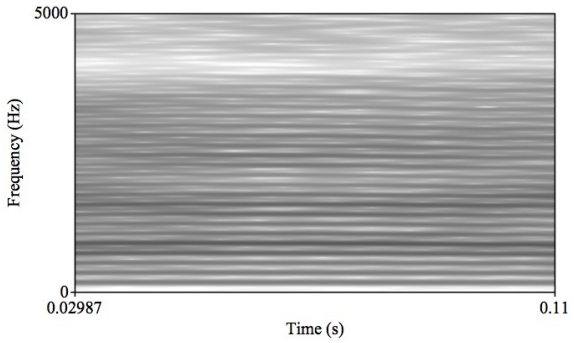
Figure 3.5 shows narrowband spectrograms for breathy, modal, and creaky /æ/. The harmonics, when visible, are shown as dark horizontal striations. They are much more clearly defined in the modal vowel than in either of the non-modal vowels, reflecting a higher ratio of harmonics to noise. The harmonics are less clear in the breathy and modal vowels, reflecting a lower ratio of harmonics to noise.

HNR has been useful in differentiating non-contrastive phonation types in English. Garellek and Keating (2015) found that HNR05 (see Table 3.5 for descriptions of the different HNR measures) was significantly lower for creaky vowels than for modal vowels in phrase-final position before /t/, and Khan et al. (2015) found that higher listener ratings of creak were strongly correlated with a lower HNR05, HNR15, and HNR25.

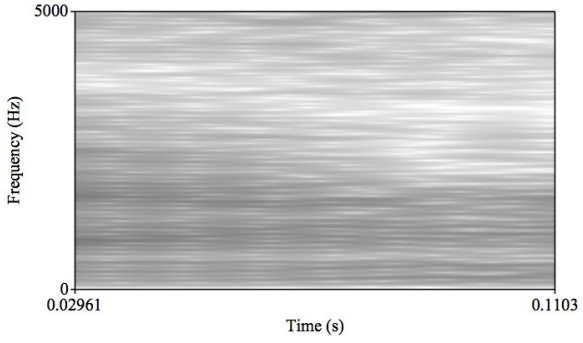
HNR has also been used to describe contrastive voice qualities. Miller (2007) found that breathy and epiglottalized vowels in Ju|'hoansi have a lower HNR than modal vowels throughout the duration of the vowel, but that glottalized vowels only have a lower HNR



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.5: Harmonics-to-Noise Ratio in breathy, modal, and creaky /æ/

during the first half of the vowel. Fulop and Golston (2009) used HNR to study modal, breathy, and whispery[6] phonation in White Hmong, though it was not able to differentiate breathy and modal releases. Khan (2012), however, did not find HNR to be particularly powerful in differentiating modal and breathy vowels in Gujarati; it was only marginally significant for HNR15, 25 and 35 when averaged across the entire duration of the vowel, and insignificant for HNR05.

As HNR is a salient cue to phonation in some languages, both subjectively (in non-contrastive uses) and objectively (in contrastive uses), it may be a useful feature in my model. VoiceSauce provides measures of HNR across four different bands, listed in Table 3.5, all of which will be included as features. In VoiceSauce, "the HNR measurements are found by liftering[7] the pitch component of the cepstrum and comparing the energy of the harmonics with the noise floor" (Shue, 2010), following the algorithm described by de Krom (1993).

Table 3.5: Harmonics-to-Noise Measures in VoiceSauce (Shue et al., 2011)

Measure	Bandwidth
HNR05	0 – 500 Hz
HNR15	0 – 1500 Hz
HNR25	0 – 2500 Hz
HNR35	0 – 3500 Hz

3.6 Subharmonic-to-Harmonic Ratio

Similar to HNR, the Subharmonic-to-Harmonic Ratio (SHR) compares the amplitude ratio of subharmonics to harmonics. Subharmonics are frequencies that are a fraction of the fundamental frequency (Sun, 2002). They can arise in the speech signal due to multiple

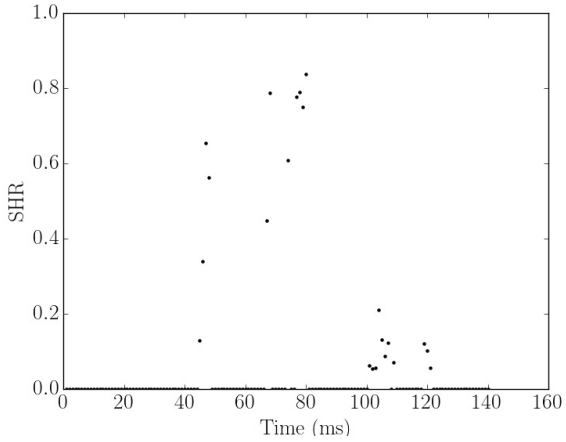
[6] "Whispery" voice is often considered a variety of breathy voice, though it has distinct enough articulatory and acoustic properties to "potentially be exploited in a linguistic sound system" (Fulop and Golston, 2009). While no language has been documented to contrast breathy and whispery voice, in White Hmong "the difference in phonetic implementation is important to prevent near-homophony between certain syllables" (Fulop and Golston, 2009).

[7] *Liftering* is filtering in the cepstral domain.

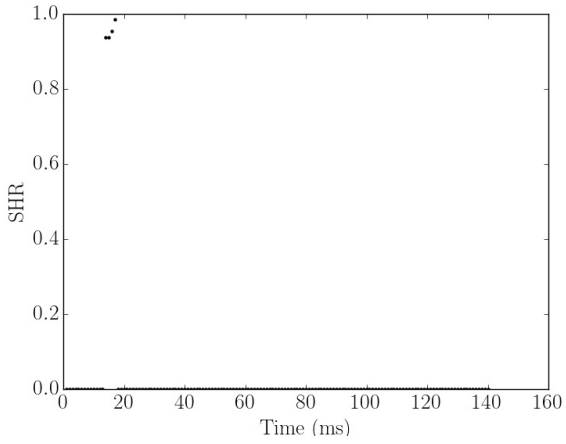
pulsing, which is characteristic of some types of creaky voice. Creaky voice with multiple pulses has two periodicities, resulting in two sets of harmonics; the stronger of the two sets determines the pitch, and the other set is the subharmonics (Garellek and Keating, 2015). Sun (2002) found that "the magnitude of subharmonics with respect to harmonics reflects the degree of deviation from modal voice," with creaky voice showing stronger subharmonics and therefore a larger SHR than modal voice. This can be seen in Figure 3.6; the breathy (a) and creaky (c) vowels have several points of very high SHR and generally many mid-range SHR values, while the modal vowel (b) has a brief period of high SHR (likely due to the transition from the preceding consonant) but remains at zero (no subharmonics) for the remainder of the vowel.

Few studies have so far used SHR to describe contrastive phonation. In a small survey of Cushillococha Ticuna, which uses non-modal phonation contrastively, Skilton (2016) found that higher SHR is one of the two most important correlates of non-modal phonation (along with steep Spectral Tilt) for two male speakers but was not important for the one female speaker. Lew and Gruber (2016) observed a "loose pattern" in which SHR is lower in the tense register than in the lax register in Louma Oeshi.

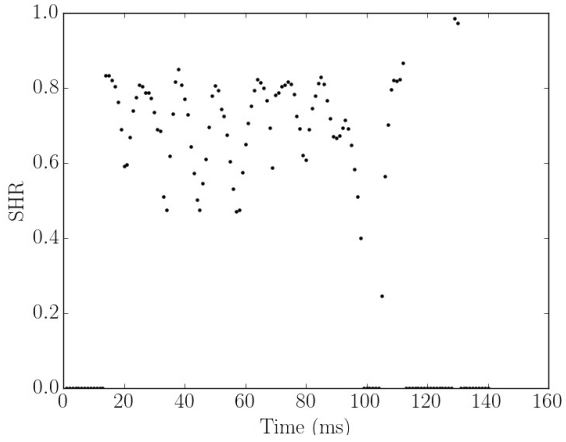
SHR has been used in several studies of English phonation; as phonation is non-contrastive in English, these studies revolve around measuring the differentiating qualities of what human listeners have judged as being different phonation types. Garellek and Keating (2015) found that SHR was higher for creaky vowels than for modal vowels before /t/ in phrase-final position, and Khan et al. (2015) found that SHR is correlated with listener judgments of creaky voice. However, Garellek and Seyfarth (2016) did not find SHR to correlate with listener perception of glottalization or phrasal creak in American English, though they note that "this result does not imply that glottalization and phrasal creak do not have multiply-pulsed voicing; rather, the SHR measure is not as well associated with a particular source of creaky voice compared to other noise or the spectral tilt measures." While SHR may not be a consistently salient cue to human listeners, it does reflect an acoustic reality of non-modal phonation types that a computer may nevertheless be able to



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.6: Subharmonic-to-Harmonic Ratio in breathy, modal, and creaky /æ/

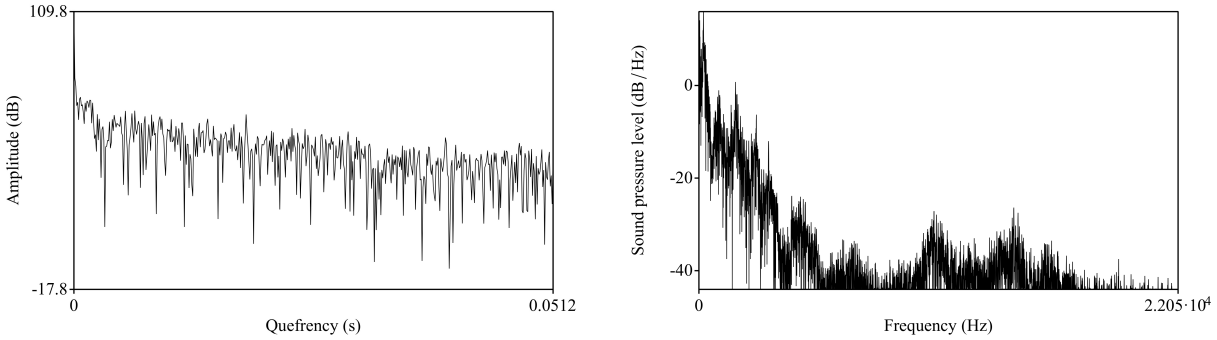
use in differentiating phonation types.

VoiceSauce measures SHR following the algorithm described by Sun (2002). In the presence of subharmonics, correctly identifying the harmonic peaks can be challenging; this algorithm approaches the issue by "decompos[ing] the effects of the harmonics and subharmonics, and examin[ing] whether the subharmonics are strong enough to be regarded as pitch candidates" (Sun, 2002). In VoiceSauce, SHR "is derived from the summed subharmonic and harmonic amplitudes calculated in the log domain using spectrum shifting" (Shue et al., 2011).

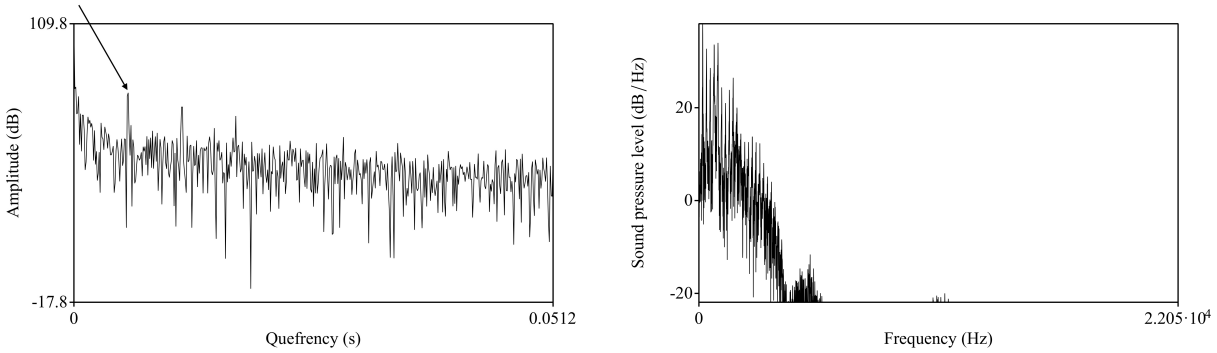
3.7 Cepstral Peak Prominence

Measuring the cepstrum, rather than the spectrum, of a sound can help separate the source and the filter, a useful distinction to make in studying phonation. A cepstrum is a spectrum created by performing a Fourier transformation on a power spectrum. It has a peak at the signal's fundamental frequency; this peak is more pronounced in more periodic and less noisy signals (Heman-Ackah et al., 2003; Blankenship, 2002). Cepstral Peak Prominence (CPP) is the measure of that peak relative to the rest of the cepstrum (de Krom, 1993), specifically "the difference in amplitude between the peak cepstral value and the mean of all cepstral values" (Blankenship, 2002). Because periodic signals have a better defined harmonic structure than aperiodic signals, the cepstral peak of modally voiced segments is more prominent than that of non-modally voiced segments (Hillenbrand et al., 1994). Figure 3.7 shows power cepstra for breathy, modal, and creaky /æ/, along with their corresponding spectra. Only the modal vowel (b) has a discernible peak.

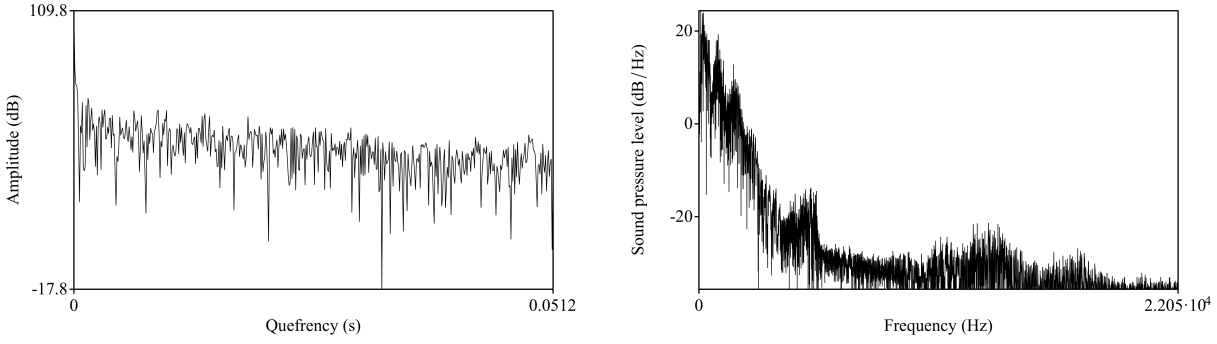
CPP has been useful in distinguishing contrastive phonation, though not with equal success for all phonation types or languages. Blankenship (2002) and Garellek and Keating (2011) found that it distinguished Mazatec breathy vowels from modal vowels, but laryngealized vowels patterned like modal vowels. Esposito (2012) successfully used CPP to differentiate between breathy, modal, and creaky tones in White Hmong. Blankenship (2002), however, found insignificant differences in CPP in Chong, which contrasts breathy



(a) Breathy



(b) Modal



(c) Creaky

Figure 3.7: Cepstral Peak Prominence in breathy, modal, and creaky /æ/

and modal voice; it appears that the direction of change, rather than any specific CPP value, may be relevant. CPP was not a useful measure in distinguishing Gujarati phonation types (Keating et al., 2011).

Few studies have investigated CPP in English's non-contrastive phonation. In those that have, CPP has been found to relate to voice quality. Garellek and Seyfarth (2016) found CPP to be useful in characterizing English glottal stops; CPP decreases throughout a creaky vowel followed by a glottal stop. Podesva et al. (2015) similarly found that CPP of English speakers decreased towards the end of a phrase, another location that often induces non-modal phonation.

Though CPP does not distinguish all phonation types across all languages, it has been successful enough to be included in my model. I will use VoiceSauce's CPP measure in this study. VoiceSauce implements the algorithm presented by Hillenbrand et al. (1994), which involves "fitting a linear regression line relating quefrency[8] to cepstral magnitude . . . the CPP measure is the difference in amplitude between the cepstral peak and the corresponding value on the regression line that is directly below the peak." VoiceSauce performs calculations over "a variable window length equal to 5 pitch periods," which is then multiplied by a Hamming window and "transformed into the real cepstral domain." The most prominent peak is found by searching around the pitch period's quefrency and then normalized "to the linear regression line which is calculated between 1 ms and the maximum quefrency" (Shue et al., 2011).

3.8 Fundamental Frequency

Breathy and creaky vowels often have a lower fundamental frequency (f_0) than their modal counterparts (Gordon and Ladefoged, 2001). As phonation and f_0 are primarily controlled with vocal fold tension, and both non-modal phonation and lower f_0 are achieved with decreased longitudinal tension, this correlation is logical.

---

[8] *Quefrency* is frequency in the cepstral domain.

Among the languages that have been shown to produce creaky vowels with a lower f_0 than modal vowels are English (utterance-finally) (Garellek and Keating, 2015), Jalapa Mazatec (Gordon and Ladefoged, 2001), Kwakw'ala (Gordon and Ladefoged, 2001), and Santa Ana del Valle Zapotec (Esposito, 2010); languages that produce breathy vowels with a lower f_0 than modal vowels include Jalapa Mazatec (Gordon and Ladefoged, 2001), Santa Ana del Valle Zapotec (Esposito, 2010), and Kedang (Samely, 1991).

In other languages, this pattern does not hold, and non-modal vowels do not have a lower fundamental frequency. The f_0 of Gujarati breathy vowels is not significantly different from the f_0 of their modal counterparts (Khan, 2012). Breathy vowels in Chanthaburi Khmer have a *higher* f_0 than modal vowels, though the authors note that this result is unexpected (Wayland and Jongman, 2003). Similarly, Andruski and Ratliff (2000) note that Green (H)Mong has the unusual combination of breathy phonation and high tone. I was unable to find any documentation of languages in which creaky voice has a higher f_0 than modal voice.

The relationship between phonation and tone can also shed light on the relationship between phonation and f_0, as many register languages associate non-modal phonation with low or falling tones rather than with high or rising tones. The traditional description of White Hmong's tone system, as outlined in Esposito (2012), includes low-falling creaky and mid-low breathy; non-modal phonation does not appear on the five other tones (high, mid, low, high-falling, and mid-rising). Vietnamese tones also have accompanying voice qualities. Kirby (2011) explains that "Glottalization plays an important role in the production and perception of the broken (C2) and glottalized (B2) tones." The falling tones are sometimes described as breathy, and the low falling tone as "accompanied by light final laryngealization." Mandarin uses creaky voice allophonically on its third tone, which is low-dipping; this creak "is assumed to be due to and coincide with reaching the bottom limit of the pitch range" (Davison, 1991).

Tonogenesis is yet another way to examine the relationship between phonation and f_0, as some tones developed from non-modal phonation. Breathy voiced consonants in Punjabi became devoiced and unaspirated, and resulted in a low tone on the following vowel (Gill

and Gleason Jr., 1969). On the other hand, *high* tones in some languages have developed due to the same glottal constriction seen in creaky voice – Vietnamese, Burmese, and Middle Chinese all have a high tone that originated from a glottal stop (Hombert et al., 1979). Athabaskan languages provide a particularly interesting example; the glottalic constriction associated with stem-final ejective consonants developed into a *high* tone in some languages (e.g., Chipewyan, Slave) and a *low* tone in others (e.g., Gwich'in, Dogrib) (Krauss, 2005).

As described above, the relationship between phonation and f_0 is complex and varied, though there is clearly an interesting and important link between the two. Measuring the f_0 of non-modal segments, however, is complicated by the fact that non-modal phonation is often characterized by a less periodic signal than modal phonation; pitch tracking algorithms often encounter difficulties under these circumstances. Many different algorithms exist to estimate f_0, some of which perform better than others in the face of aperiodicity. The present study considers the four pitch tracking algorithms available in VoiceSauce: Praat, SHR, Snack, and STRAIGHT.

Praat f_0 can be calculated using autocorrelation or crosscorrelation. In this study, I will use the cross-correlation method. This uses the algorithm described by Boersma (1993), with the time step, pitch floor, and pitch ceiling set by the user (see Section 4.2.2 for a discussion of the settings used in this study).[9]

SHR (Subharmonic-to-Harmonic Ratio) f_0 is "specifically designed to estimate a perceptual f_0 in the face of subharmonics" (Keating and Garellek, 2015), which makes it well-suited to track the pitch through some types of creaky voice (see Section 3.6 for more information about SHR). This algorithm was developed by Sun (2002).

The **Snack f_0** algorithm, from the Snack Sound Toolkit (Sjölander, 2004) implements Talkin's `Get_F0` algorithm (Talkin, 1995), estimating f_0 using normalized cross-correlation.

---

[9]Perhaps counterintuitively, I will be calculating Praat's f_0 measure in VoiceSauce, which runs Praat's exact algorithm. This is simply because of my ordering of events - I extracted data from VoiceSauce, which automatically includes Praat's f_0 algorithm, before extracting data from Praat, and thus already had the values when it came time to write the script. I confirmed that VoiceSauce's Praat f_0 output was identical to Praat's f_0 output.

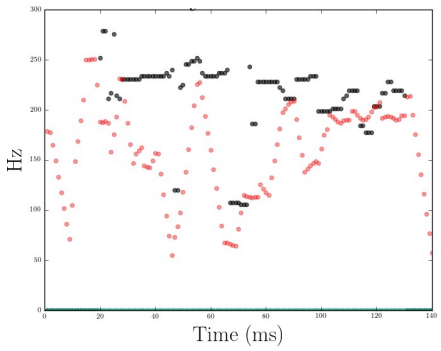
The **STRAIGHT (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram)** f_0 algorithm is particularly robust in the face of non-modal phonation and is widely used in the field. Developed by Kawahara et al. (1998), this algorithm uses a formal model of smoothness in its f_0 estimation.

3.9 Variance of Pitch Tracks

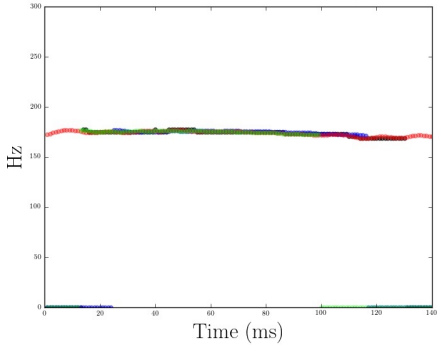
Pitch tracking algorithms do not always agree with each other, hence the value in including four of them. My observations and intuition suggest that the algorithms are more likely to agree during modal voicing than during non-modal voicing, as the pitch periods occur fairly regularly and without much noise. And when pitch tracking algorithms disagree, they generally disagree in different ways.

Figure 3.8 compares these four pitch tracks over the now-familiar tokens of /æ/. The four algorithms are Praat (blue), SHR (black), Snack (green), and STRAIGHT (red). The breathy vowel (a) shows significant inconsistencies between STRAIGHT and SHR. Both take several improbably large leaps and dives. The Snack and Praat pitch tracks fail completely; they both track the entire vowel's pitch at 0 Hz. The four pitch tracks for the modal vowel (b) are remarkably consistent. While Praat and SHR run into problems at the beginning and end of the vowel, tracking the pitch at 0 Hz, the variance between the four tracks during the majority of the vowel is generally under 2 Hz. Errors at vowel boundaries are common and expected during transitions between phones. In the creaky vowel (c), the pitch tracks are relatively consistent in the first third. Towards the middle, as the glottal pulses become less regular, the pitch tracks become less reliable. Snack and Praat both drop to 0 Hz while SHR jumps. STRAIGHT, living up to its reputation, gives a reasonable pitch track throughout the vowel.

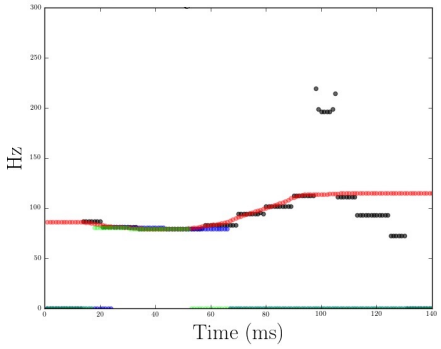
The observed inconsistencies between pitch tracking errors seen in Figure 3.8 lead me to propose a new measure that may be indicative of non-modal phonation: the Variance of Pitch Tracks (VoPT). This measure is a sum of the differences between each combination of the four pitch tracks described above. A higher VoPT indicates less consistency between pitch



(a) Breathy Pitch Tracks



(b) Modal Pitch Tracks



(c) Creaky Pitch Tracks

Figure 3.8: Pitch tracking errors in breathy, modal, and creaky /æ/; pitch tracks are Praat (blue), SHR (black), Snack (green), and STRAIGHT (red)

tracks, which may be due to difficulties in tracking aperiodic glottal pulses. In developing this measure, I tested three different calculations, described in Table 3.6.

The three formulas in Table 3.6 assume a set $|PT|$ of the four pitch tracking algorithms (Praat, SHR, Snack, and STRAIGHT). I represent each algorithm as a function that takes a vowel v and a time point t and returns that algorithm's pitch calculation for that vowel at that time point. Each vowel v has N_v time points.

Table 3.6: Variance of Pitch Track (VoPT) Measures

Measure	Calculation
Mean Absolute Difference	For each combination of pitch tracks, calculate the absolute difference of the mean f_0 of the vowel, then sum all the differences. $$VoPT(v) = \sum_{i=1}^{|PT|} \sum_{j=i+1}^{|PT|} \left\| \frac{\sum_{n=1}^{N_v} PT_i(v,n)}{N_v} - \frac{\sum_{n=1}^{N_v} PT_j(v,n)}{N_v} \right\|$$
RMSE3	For each combination of pitch tracks, calculate the Root Mean Square Error between the two tracks at *three* timepoints of the vowel, then sum all the RMSEs. $$VoPT(v) = \sum_{i=1}^{|PT|} \sum_{j=i+1}^{|PT|} \sqrt{\frac{\sum_{s=1}^{3} (PT_i(v, s\lfloor \frac{N_v}{4} \rfloor) - PT_i(v, s\lfloor \frac{N_v}{4} \rfloor))^2}{3}}$$
RMSE10	For each combination of pitch tracks, calculate the Root Mean Square Error between the two tracks at *ten* timepoints of the vowel, then sum all the RMSEs. $$VoPT(v) = \sum_{i=1}^{|PT|} \sum_{j=i+1}^{|PT|} \sqrt{\frac{\sum_{s=1}^{10} (PT_i(v, s\lfloor \frac{N_v}{11} \rfloor) - PT_i(v, s\lfloor \frac{N_v}{11} \rfloor))^2}{10}}$$

Figure 3.9 shows each of the three VoPT calculations for each phonation type for three

female English speakers from the Pacific Northwest.[10] The VoPT was calculated using 431 (22 B, 239 M, 170 C), 688 (31 B, 476 M, 181 C), and 839 (48 B, 598 M, 193 C) vowels, respectively.

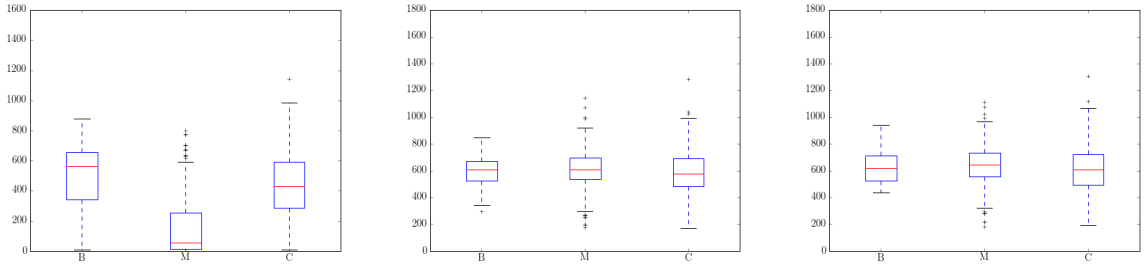
The Mean Absolute Difference (left column) shows by far the most interesting pattern: the mean VoPT for modal vowels is much lower than the mean VoPT for breathy and creaky vowels. While there is a long tail for the modal vowels, suggesting that some modal vowels have pitch tracking errors, the majority of modal VoPTs remain within a small range. Breathy vowels tend to have a slightly higher VoPT than creaky vowels, though the measure does not discriminate well between breathy and creaky vowels. These patterns are consistent for all three speakers.

When VoPT is calculated as the RMSE at three (middle column) or ten time points (right column), the above pattern does not hold. Perhaps a limited number of time points is not sufficient to capture errors; the errors still exist, but do not happen to be occurring at the three or ten time points at which measurements were taken.

Figure A.1 shows The VoPT for the 10,031 data points from the 22 English speakers (the corpus from which these recordings come and the steps taken to pare down the data set are described in detail in Chapter 4). The red lines indicate the median VoPT, and the red dots indicate the mean VoPT. The results for each of these 22 speakers are shown in Appendix A; the overall pattern generally holds for all speakers, with non-modal vowels having a higher average VoPT than modal vowels. The mean VoPT for creaky vowels is higher than the mean VoPT for modal vowels across all speakers. For three speakers, the mean VoPT for breathy vowels is slightly lower than the mean VoPT for modal vowels. However, in these cases, there are too few instances of breathy voice to draw conclusions from these data.

As the Mean Absolute Difference calculation of the Variance of Pitch Tracks is the most discriminating measure – at least in discriminating modal from non-modal phonation – this calculation will be included as a feature in the models. However, because both breathy and

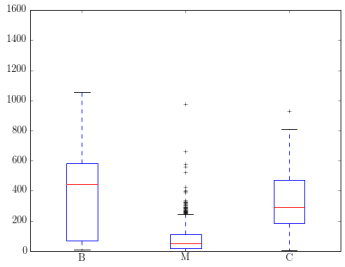
[10]The corpus from which these data are drawn is described in Section 4.1.1.



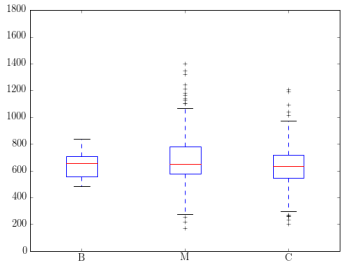
(a) NWF089, Mean

(b) NWF089, RMSE3

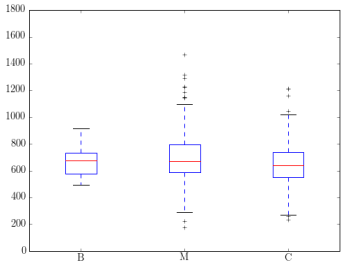
(c) NWF089, RMSE10



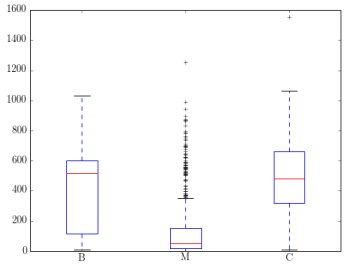
(d) NWF090, Mean



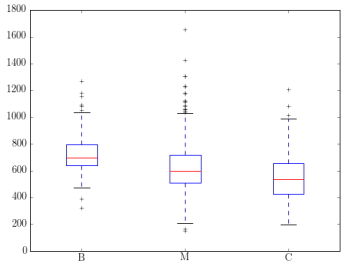
(e) NWF090, RMSE3



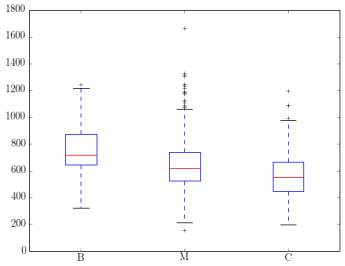
(f) NWF090, RMSE10



(g) NWF092, Mean



(h) NWF092, RMSE3



(i) NWF092, RMSE10

Figure 3.9: Variance of Pitch Tracks Calculations

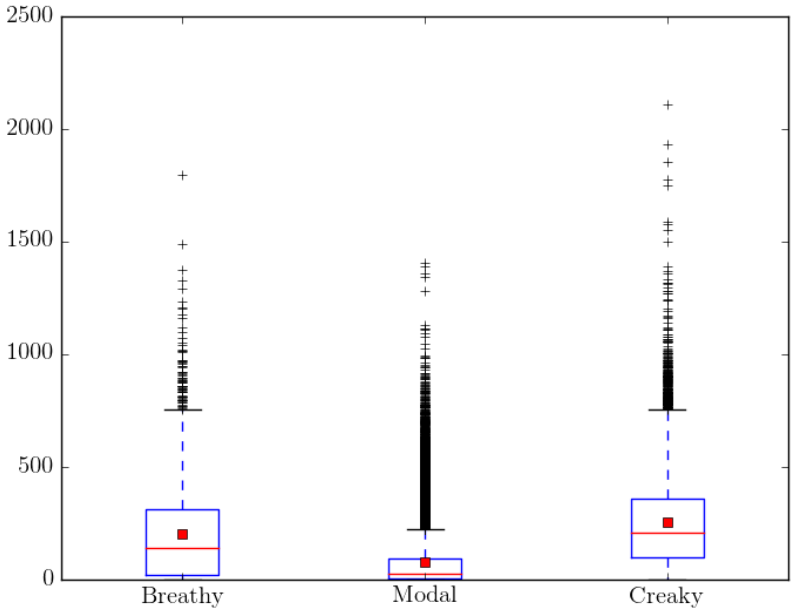


Figure 3.10: VoPT, All English Speakers

creaky voice are characterized by aperiodicity, and pitch tracks struggle with aperiodicity, this measure may be more powerful in differentiating modal from non-modal vowels than in separating the three phonation types. Figure 3.11 shows the mean VoPT for modal versus non-modal vowels for all English speakers in the data, illustrating that VoPT may be a powerful measure in modal versus non-modal classification.

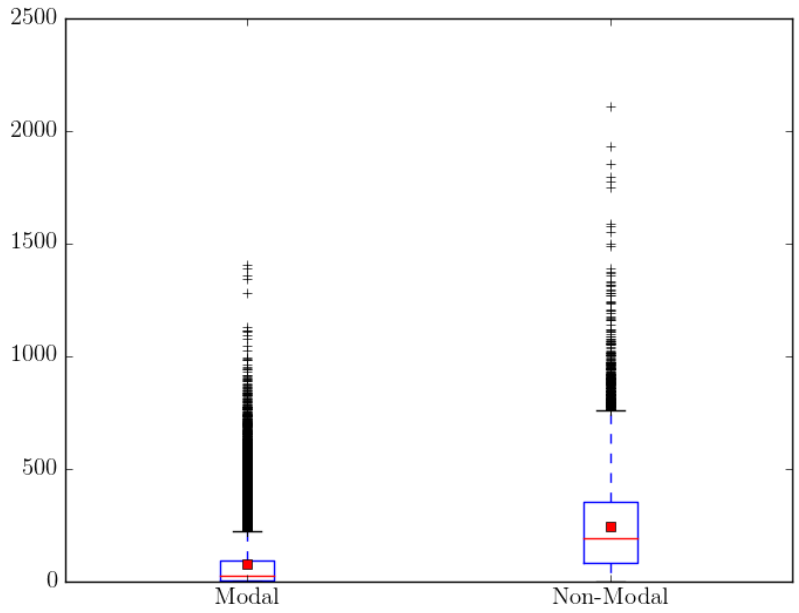


Figure 3.11: VoPT, All English Speakers

3.10 F1

Vowel qualities are often described by their first two formants, or resonating frequencies that are impacted by constrictions in the vocal tract (Ladefoged and Johnson, 2015). The first formant, F1, has also been linked with phonation type. F1 is often associated with vowel height. While it is typically described as a reflection of how high the tongue is in the oral cavity during articulation, F1 is actually the resonating frequency created by the location

of the constriction; low vowels (those with a high F1) have a pharyngeal constriction with a backed tongue root, while mid and high vowels have an oral cavity constriction with the tongue root displaced forward (Gick et al., 2013).

Many languages produce non-modal vowels with a higher F1 than their modal counterparts. Jalapa Mazatec creaky vowels have a higher F1 than modal or breathy vowels. (Kirk et al., 1993). Similarly, Haoni tense vowels have a higher F1 than lax vowels (Maddieson and Ladefoged, 1985). In Pacific Northwest English, which does not use phonation contrastively, vowels with a high F1 are more likely to be creaky than vowels with a low F1 (Panfili, 2015).

Not all languages exhibit this pattern. Nyah Kur, Kui (Thongkum, 1987), and Green (H)Mong (Andruski and Ratliff, 2000) do not show any systematic differences in formant frequencies between modal and breathy vowels. Chong non-modal vowels have a *lower* F1 than modal vowels (Thongkum, 1987), and all qualities of Kedang breathy vowels have a lower F1 than their modal counterparts (Samely, 1991).

The reason for the link between F1 and phonation type remains speculative. One explanation is that the tongue root displacement involved in articulating high vowels pulls the hyoid bone forward, which tilts the thyroid cartilage down and increases longitudinal tension, making creaky voice harder to achieve (Ladefoged, 1964; Lehiste, 1970). This explanation is often cited in reference to Intrinsic Fundamental Frequency (IF0), the observation that low vowels have a lower fundamental frequency than high vowels cross-linguistically.

Though some languages produce non-modal vowels with a lowered F1 while others produce non-modal vowels with an increased F1, many languages systematically vary F1 with phonation type. Therefore F1 will be included as a feature in the models. F1 is a particularly compelling feature as patterns have been found in languages that use phonation in different ways, including contrastively, as part of a register system, and socially. The F1 calculation included as a feature in this study will be measured in VoiceSauce based on the STRAIGHT pitch tracking algorithm.

3.11 Vowel Duration

Vowel duration is simply the length of the vowel, typically measured in milliseconds. With the onset and offset of a vowel marked, either by a human or a computer, the Vowel Duration is calculated by finding the amount of time between those two points. Vowel Duration is contrastive in some languages, such as Hungarian (Szende, 1999), and allophonic in others, such as English (Chen, 1970). Various factors can trigger changes in vowel length, including stress, voicing of the following phone (Ladefoged and Johnson, 2015), and nasalization (Styler, 2015).

In many languages, non-modal vowels have a longer duration than their modal counterparts. Kirk et al. (1993) found that both creaky and breathy vowels are longer than modal vowels in Jalapa Mazatec, and Samely (1991) found that breathy vowels are longer than modal ones in Kedang. Blankenship (2002) measured Vowel Duration in Mazatec and Chong, which both use phonation contrastively but not across the entire vowel. She found that "the duration of the interval where nonmodal values were observed is consistently longer in the contrastive languages, both in absolute time and as a percentage of the complete vowel." In Hupa, non-modal phonation occurs only on phonemic long vowels, and not on phonemic short vowels (Gordon, 2001). Quileute similarly only allows non-modal phonation on long vowels, and they must also be stressed (Powell and Woodruff Sr., 1976).

However, this correlation between vowel length and non-modal phonation is not universal. Green (H)mong and White Hmong vowels produced with a creaky tone are shorter than those with modal and breathy tones (Andruski and Ratliff, 2000; Esposito, 2012; Huffman, 1987), but Vowel Duration is otherwise not significant. Non-modal phonation is also not associated with a duration change in San Lucas Quiaviní Zapotec (Gordon and Ladefoged, 2001).

Though a change in Vowel Duration is not consistently associated with non-modal phonation, its calculation is trivial and may aid the model in distinguishing phonation types for some languages. It will be calculated in milliseconds based on the TextGrid boundaries

of each vowel.[11]

3.12 Surrounding Phones

Since Rousselot's kymographic research in the late 19th century, it has been well-documented that the phones flanking a given sound can influence that sound; speech is continuous and the boundaries between sounds are often unclear. Two primary aspects of surrounding phones may impact the phonation of the phone in between them – voicing and manner of articulation. The myoelastic factors caused by voicing and the aerodynamic factors caused by manner of articulation both influence the vocal folds in ways that can alter phonation. Additionally, voice quality can be influenced by the *absence* of surrounding phones when a phone falls utterance-initially or finally.

The relationship between a vowel's phonation type and the voicing and manner of its surrounding phones, if they exist, is not straightforward and may vary from language to language. Additionally, voicing and manner are often tangled together, making it difficult to tease out their separate influence. This section outlines some of the documented instances of these features impacting phonation.

The manner of articulation of a phone can impact the phonation type of an adjacent phone by changing airflow and pressure. Obstruents impede airflow, and speakers adjust laryngeal tension in order to maintain voicing despite the airflow impedance; these adjustments can result in changes in phonation. Specifically, transitions in and out of voiced stops are associated with lower f_0. During stop closure, longitudinal tension is decreased in order to maintain voicing despite pressure buildup; decreased longitudinal tension results in a decreased f_0, and can also result in non-modal phonation (Traill, 1987).

Stops whose voicelessness is achieved through a glottal stop can also result in changes in voice quality. In some dialects of English, utterance-final /t/, as well as syllable- and word-final /t/, can become a glottal stop. This is a form of debuccalization, in which an

[11]The procedures for marking boundaries vary from corpus to corpus and are discussed in Chapter 4.

oral consonant becomes laryngeal (O'Brien, 2012). The presence of the glottal stop can, in turn, trigger creaky voicing on a preceding sonorant. /p/ and /k/, the other two English voiceless stops, are also often accompanied by or achieved with a glottal gesture, though wide variation exists across dialects and across speakers (Roach, 1973; Huffman, 2005). German also optionally achieves closure for voiceless oral stops with a glottal stop, and this glottal stop can "be weakened to irregular pulsing" on an adjacent voiced sound (Kohler, 1994).

Allophonic breathy voice is also found near /h/ in some dialects of English. /h/ becomes breathy voiced /ɦ/ intervocalically, and the breathy voicing spreads to adjacent voiced sounds (Ladefoged and Johnson, 2015). This process can be seen in the words *head* [hɛd] and *ahead* [əɦɛ̤d].

Not all vowels are flanked by two phones; utterance-initial and utterance-final vowels are nevertheless influenced by their positions, as airflow and vocal fold vibration must begin or end. (Note the distinction between vowels at word boundaries and vowels at utterance boundaries. Vowels at word boundaries are of less interest here, as they are likely flanked by the final phone of the preceding word and the first phone of the following word. I focus here on vowels at utterance boundaries, as those are sure to be preceded or followed by a pause in speech.)

American English allophonically glottalizes word-initial vowels (Dilley et al., 1996), particularly when they are also at the beginning of an intonational phrase (Pierrehumbert and Talkin, 1992; Dilley et al., 1996). This is due to the insertion of a glottal stop, making it a similar process to the voiceless stop-induced glottalization described above. Though the reason for this phenomenon is unknown, Dilley et al. (1996) suggest that it may be in part "simply due to mechanical constraints of starting a vowel after a pause," or that "glottalization of word-initial vowels at prosodically significant locations may represent a strengthening of the articulatory gesture associated with the onset of the prosodic constituent or prominence." The latter hypothesis seems more plausible, as some languages contrast glottal stops and vowels word-initially, showing that it is mechanically feasible to begin voicing without a glottal stop (e.g., Tsou (Wright and Ladefoged, 1994), Mantauran (Zeitoun,

2007)).

Utterance-final creaky voice, often called phrase-final glottalization, has also been described in various dialects of English and other languages. For both male and female speakers of the Received Pronunciation and Modified Northern dialects of British English, creaky voice occurs more often on the final syllable of a sentence than on any other syllable (Henton and Bladon, 1988). In American English, it may also occur "at the ends of falling intonations for some speakers" (Ladefoged and Johnson, 2015); that is, creaky voice may be induced by way of utterance-final f_0 lowering and is also associated with non-modal phonation. In American English, listeners also use creaky voice, among other features, as a cue to sentence boundaries (Kreiman, 1982).

Unfortunately, the format of the data sets used in this study (see Chapter 4) leaves some gaps in information about surrounding phones. While most data sets do include surrounding phones, these are often based on the canonical pronunciation rather than actual utterances. For example, the phonetic transcriptions of the English data set are based on CMU's pronouncing dictionary (Weide, 2005), which does not capture when a voiceless stop is articulated with a glottal closure rather than an oral one, or when an intervocalic /h/ is realized as /ɦ/. These limitations mean that any attempt to capture how surrounding phones impact phonation will be imperfect, though perhaps still powerful enough to act as a predictor of voice quality, particularly in languages that use phonation allophonically.

Given the scope of observations that surrounding phones impact voice quality, the voicing, manner, and presence or absence of surrounding phones will be considered in the model. I encode this information in six binary features for preceding and following phones each: `pre_is_voiced`, `pre_is_obstruent`, `pre_phone_exists`, `fol_is_voiced`, `fol_is_obstruent`, and `fol_phone_exists`. Table 3.7 shows several examples of these features; the bolded vowel is the one in question.

Table 3.7: Example Surrounding Phones Feature Values

Phones	pre_is_voiced	pre_is_obstruent	pre_exists	fol_is_voiced	fol_is_obstruent	fol_exists
/l**æ**t/	1	0	1	0	1	1
/**ɑ**d/	0	0	0	1	1	1
/p**æ**θ/	0	1	1	0	1	1
/f**i**/	0	1	1	0	0	0

3.13 Prosodic Position

As mentioned in Section 3.12, the position of a vowel and its surrounding phones relative to the word and the utterance can be a factor in voice quality. In English, creaky voice can occur utterance-finally as well as in vowel-initial words and /t/-final words, though with great variation among speakers in both the occurrence and production of this creak (see Section 3.12). Because the majority of the research regarding this prosodic creak has been done on American English, it is generally unknown how much these findings apply to other languages.

In American English, creaky voice is often found towards the end of utterances, with glottalization occurring at a higher rate utterance-finally than utterance-medially. Additionally, Redi and Shattuck-Hufnagel (2001) found that it occurs more frequently "at the boundaries of full intonational phrases than at intermediate intonational phrases." They posit that this "boundary-related glottalization" is correlated with another boundary-related occurrence, such as a drop in f_0 or subglottal pressure, or it could be "independently planned." That is, it is a discrete event that is not caused by a separate event.

Similar uses of utterance-final creak have been found in a few other languages. Finnish

speakers use creaky voice as a turn-yielding device, to signal the end of their speaking turn during a conversation (Ogden, 2001). Swedish speakers use glottalization, among other cues, to identify prosodic boundaries (Carlson et al., 2005).

Various other phonetic phenomena are found in utterance-final position. Lengthening of the final vowel of a phrase, as well as of final consonants and syllables, has been widely described in English (e.g., Klatt (1976); Fougeron and Keating (1997); Shattuck-Hufnagel and Turk (1998)). Articulatory strengthening of English vowels in the form of reduced linguapalatal contact has been found to increase in final syllables; final strengthening is less consistent for consonants (Fougeron and Keating, 1997). Cross-linguistically, pitch is used to indicate the edges of syntactic units, typically using a falling pitch to mark the end of a unit (Ladefoged and Johnson, 2015). Given that pitch falls utterance-finally and non-modal phonation is associated with a lowered fundamental frequency, we might expect some degree of phonation change to occur at the end of an utterance.

Both edges of words can trigger voice quality changes in English. As mentioned in the previous section, word-initial vowels are allophonically and optionally glottalized (Dilley et al., 1996). The creaky voice accompanying allophonic glottal closure of voiceless stops is more likely to occur in word-final position, though this process can also occur in coda position more generally (Huffman, 2005).

I consider three measures of a vowel's prosodic position: milliseconds from the end of the utterance,[12] the percent of the way through the utterance, and the percent of the way through the word. All three calculations are made from the midpoint of the vowel. These measures may prove more useful in detecting non-modal phonation, perhaps specifically creaky voice, in languages that use phonation allophonically.

---

[12]Marking the end of an utterance is often not straightforward. Much of the data used in this study come from word lists, in which the utterance is well-defined; it is simply the word (if the task is just words) or the carrier phrase. ATAROS, the one conversational corpus used here (see Chapter 4), is annotated for *spurts*. A spurt is defined as a "stretch of speech said by one speaker between at least 500 ms of silence . . . manually marked during transcription" (Freeman, 2015). For the purposes of this study, I will consider spurt boundaries to be utterance boundaries.

3.14 Summary of Features

Table 3.8 lists the 39 features from the thirteen categories of measures described in this chapter. These measures have all been shown or hypothesized to relate to voice quality and will be considered as features in the machine learning model. In Chapter 4, I describe the corpora from which these measures will be extracted, how the measures will be extracted, and how they will be processed and prepared for use in the machine learning model.

Table 3.8: Summary of Features

Measure	Category	Implementation
H1* – A1*	Spectral Tilt	VoiceSauce
H1* – A2*	Spectral Tilt	VoiceSauce
H1* – A3*	Spectral Tilt	VoiceSauce
H1* – H2*	Spectral Tilt	VoiceSauce
H2* – H4*	Spectral Tilt	VoiceSauce
H4* – 2k*	Spectral Tilt	VoiceSauce
2k* – 5k	Spectral Tilt	VoiceSauce
Local_Jitter	Jitter	Praat
Local_Absolute_Jitter	Jitter	Praat
RAP_Jitter	Jitter	Praat
PPQ5_Jitter	Jitter	Praat
RMS_Energy	Intensity	VoiceSauce
Local_Shimmer	Shimmer	Praat
Local_Shimmer_dB	Shimmer	Praat
APQ3_Shimmer	Shimmer	Praat
APQ5_Shimmer	Shimmer	Praat
APQ11_Shimmer	Shimmer	Praat
HNR05	Harmonics-to-Noise	VoiceSauce
HNR15	Harmonics-to-Noise	VoiceSauce
HNR25	Harmonics-to-Noise	VoiceSauce
HNR35	Harmonics-to-Noise	VoiceSauce
SHR	Subharmonic-to-Harmonic	VoiceSauce
CPP	Cepstral Peak Prominence	VoiceSauce
Praat_f_0	Fundamental Frequency	VoiceSauce
SHR_f_0	Fundamental Frequency	VoiceSauce
Snack_f_0	Fundamental Frequency	VoiceSauce
STRAIGHT_f_0	Fundamental Frequency	VoiceSauce
Variance of Pitch Tracks	VoPT	VoiceSauce
F1	F1	VoiceSauce
Vowel_Duration	Duration	–†
Voicing_Preceding_Phone	Surrounding Phones	–†
Voicing_Following_Phone	Surrounding Phones	–†
Manner_Preceding_Phone	Surrounding Phones	–†
Manner_Following_Phone	Surrounding Phones	–†
Presence_Preceding_Phone	Surrounding Phones	–†
Presence_Following_Phone	Surrounding Phones	–†
Distance_from_Utterance_End_(ms)	Prosodic Position	–†
Distance_from_Utterance_End_(percent)	Prosodic Position	–†
Distance_from_Word End_(percent)	Prosodic Position	–†

** Indicates measures that have been corrected for formant frequencies and bandwidths in VoiceSauce.*

† *Indicates contextual measures whose values are determined based on TextGrids, but are not calculated in a specific program.*

Chapter 4

CORPORA AND DATA EXTRACTION

This chapter describes the corpora used in this study, the process of extracting the features described in Chapter 3 from those corpora, and the data processing conducted. Section 4.1 outlines the corpora containing recordings of the six languages under consideration; for each, I describe the languages included in the corpus and how they use phonation, as well as available information about the speakers, the recording procedure and conditions, and any pre-processing steps. Section 4.2 describes the methods used to extract the features proposed in Chapter 3 from those corpora. Finally, Section 4.3 discusses transforming the raw output of the extraction process into data that are ready for the machine learning model.

4.1 The Corpora

This dissertation relies on two corpora containing a total of six languages. These corpora – the ATAROS Corpus and the Production and Perception of Linguistic Voice Quality Corpus – are described in this section.

4.1.1 The ATAROS Corpus

The ATAROS corpus (Freeman, 2015) consists of naturalistic conversations in the Pacific Northwest dialect of American English. Because of the nature of the corpus, as well as the nature of phonation in English, ATAROS required the most complex set of steps to prepare it for use in this study.

English

English uses phonation allophonically, prosodically, and sociolinguistically. These uses are briefly described below.

English Allophonic Phonation

Both breathy and creaky voice can occur as allophonic variants of modal voice in English. Vowels adjacent to /h/ are often allophonically breathy, and vowels adjacent to glottal stop are often allophonically creaky (Gordon and Ladefoged, 2001).[1]

English Prosodic Phonation

English speakers often use non-modal phonation – specifically creaky voice – to mark prosodic boundaries. As discussed in Chapter 3, non-modal phonation is often found at the beginning and end of utterances and on word-initial vowels. English speakers are more likely to use creaky voice on word-initial vowels that are also intonational phrase-initial, and when the word is marked with pitch accent (Dilley et al., 1996). Similarly, they are more likely to glottalize utterance-final words that are also at the end of intonational phrases, rather than at the ends of utterance-medial intonational phrases (Redi and Shattuck-Hufnagel, 2001). However, Redi and Shattuck-Hufnagel (2001) note that there is "a wide range in the rates of glottalization and in preferred acoustic characteristics across individual speakers."

English Sociolinguistic Phonation

The sociolinguistic uses of non-modal phonation in English have recently piqued the interest of linguists. The linguistic community has investigated the use of creaky voice by different genders and the social meaning that it carries. Many studies report that women use creaky voice more frequently than men. Yuasa (2010) found that American women use creaky voice more than twice as frequently as American men, and almost twice as frequently as Japanese women. Podesva (2013) found that women use creaky voice significantly more than men in his study of white and African American speakers from Washington, D.C., though they use breathy voice at similar rates. He also reports similar rates of creaky voice

[1] In some dialects, glottal stop is an allophone of word-final voiceless stops /p/, /t/, and /k/, and that glottal stop triggers allophonic creaky voice (Roach, 1973; Huffman, 2005).

among women of different age groups, and points out that Yuasa (2010)'s claims that younger women are using creaky voice is problematic, as everyone in her study was 33 or younger. Panfili (2015), however, found that male and female Pacific Northwest English speakers of a wider variety of ages (21–70) use creaky voice at nearly identical rates.

Many of these studies ascribe meaning to the various voice qualities when they occur outside of the allophonically conditioned locations. Gobl and Ní Chasaide (2003) found that various synthesized voice qualities were associated with a constellation of "affective attributes" – breathy voice with being timid and intimate; creaky voice with being bored, content, and unafraid; and modal voice slightly with being interested and confident. Perhaps because creaky voice is associated with a lowered fundamental frequency, it has also been linked with toughness and authoritativeness (Mendoza-Denton, 2011). Yuasa (2010) found that college-age Americans living in California and Iowa "perceive female creaky voice as hesitant, nonaggressive, and informal but also educated, urban-oriented, and upwardly mobile," though this perception study was based on just one modal sample and one creaky sample from the same speaker.

Despite the many studies about sociolinguistic phonation, the social uses and meanings of English phonation remain poorly understood. The studies described above all involve a phonetically trained listener manually identifying periods of different phonation types; this means that findings are limited by scarcity of data as well as subjective and difficult to reproduce methods. All three of these limitations could be overcome with a tool to automatically classify phonation types; I aim to build said tool in Chapter 13.

These three uses of phonation in English – allophonic, prosodic, and sociolinguistic – are optional. Unlike contrastive phonation, these paralinguistic uses do not change lexical meaning. This allows for more variation in the use and production of voice qualities without disrupting communication; the environments in which different phonation types occur and their acoustic characteristics may prove too varied in English for a machine learning model to capture patterns.

The Corpus

The ATAROS *(Automatic Tagging and Recognition of Stance)* Corpus (Freeman, 2015) consists of dyadic conversations between adult native Pacific Northwest English speakers. This dialect region was defined as Washington, Oregon, and Idaho; speakers were required to have spent the majority of their childhood in this region, without a gap of two or more years, and consider English a native language. Speakers were matched roughly for age and either matched or crossed for gender.

Each dyad performed five collaborative tasks intended to elicit changes in stance[2] and increasing involvement in the topic. They were recorded in a sound attenuated booth at the University of Washington Phonetics Laboratory using head-mounted AKG C520 condenser microphones to create 16-bit stereo WAV-files at a 44.1 kHz sampling rate.

This study includes eleven dyads,[3] or 22 speakers (11 female, 11 male; ages 20-70; see Table 4.1 for dyad demographics) performing the Budget Task. The Budget Task, which is the highest stance level of the five ATAROS tasks, asks speakers to imagine that they are part of the county committee responsible for balancing the budget. Given a list of departments and the services they provide, speakers must agree on which services to cut from each department. This task was chosen for the present study because it was the final task, making for more naturalistic speech as the effects of observation wear off and the level of engagement increases.

Pre-Processing

Each dyad's recording was separated into stereo files for each of the five tasks. The conversations were manually transcribed, with each speaker's contributions on a separate tier of a Praat text grid. These transcriptions were automatically force-aligned using the Penn Phonetics Lab Forced Aligner (P2FA), which marks boundaries of words and phones

[2] *Stance* is defined as "attitudes and opinions about the topic of discussion" (Freeman, 2015).

[3] I retain the original dyad numbers from the ATAROS corpus. Some dyads were excluded from this study because they were performing strange voices, or to get even gender numbers and balance.

Table 4.1: Gender and Age of English Speakers from the ATAROS Corpus

Dyad 1	F, 21	M, 24
Dyad 2	F, 70	F, 68
Dyad 3	M, 26	F, 24
Dyad 4	M, 24	F, 23
Dyad 5	F, 21	F, 27
Dyad 6	F, 49	M, 49
Dyad 8	F, 39	M, 38
Dyad 9	F, 23	F, 19
Dyad 12	M, 43	M, 47
Dyad 15	M, 70	M, 67
Dyad 18	M, 20	M, 25

on separate tiers (Yuan and Liberman, 2008).[4] Human listeners also reviewed the relevant recordings, as described below.

Annotation

Phone boundaries were automatically determined during forced alignment and vowels were assigned stress based on CMU's pronouncing dictionary (Weide, 2005). All stressed vowels (primary and secondary stress) were then manually annotated for phonation type. Two phonetically-trained raters listened to stressed vowels from the eleven conversations. They were familiarized with the labeling system by listening to examples of various vowel qualities exhibiting classic breathy, modal, and creaky voice, as well as examples of problems that

[4]The ATAROS project went on to annotate the recordings for stance, though that information is not used in this dissertation; see Freeman (2015) for more information about stance, as well as more information about the recording procedure described above.

warrant exclusion. In order to accurately represent what we *hear* as different phonation types (as opposed to the acoustic properties phoneticians are trained to recognize in spectrograms and waveforms), the raters were instructed to rely on their ears to make a judgment; Praat's spectrogram, pitch track, intensity track, and formant tracks were not displayed during annotation, and raters were discouraged from looking at the waveform.

The raters gave each vowel one of five annotations, listed in Table 4.2. These annotations were recorded in a separate tier of the Praat text grid that was aligned with the tier containing phone boundaries. Breathy, modal and creaky vowels were marked as `B`, `M`, and `C`, respectively. Vowels containing a flaw in the recording were marked as `0`; flaws included clipping, misaligned boundaries resulting in the inclusion of part of a neighboring sound, and, most frequently, speech from their interlocutor that was picked up by their microphone. Vowels that were interesting, such as laugh-speech or a phonation type that did not fall into the three categories, were marked as `1`, to indicate that they are problematic for the present study but may be interesting to return to later. If a vowel contained more than one phonation type, raters were instructed to label it as the phonation type that is most prevalent during the vowel. The two raters had good inter-rater reliability (Cohen's Kappa 0.85 overlapping on 6.87% of the data points[5]).

Table 4.2: Phonation Annotations

Tag	Meaning
`B`	Breathy
`M`	Modal
`C`	Creaky
`0`	Flaw in recording or alignment
`1`	Something interesting but irrelevant or problematic

[5]In calculating Cohen's Kappa, I collapsed `0` and `1` into one category, therefore comparing ratings of `B`, `M`, `C`, and `unusable`. When `0` and `1` remained as their own categories, Cohen's Kappa was only slightly lower: 0.84.

Post-Processing

Two types of vowels were removed from the set before extracting features. First, I removed irrelevant vowels – those tagged as `0` (*flaw in recording or alignment*) or `1` (*something interesting but irrelevant or problematic*). These vowels do not represent usable audio or any of the phonation types. Because humans listened to each token, removing this set ensures that the tokens remaining for analysis do, in fact, include the expected vowel quality, and include no other sounds. Vowel boundaries were automatically marked during force alignment, which leaves room for error. However, if a vowel's boundaries included non-vowel sounds, that vowel received a tag of `0` and was ultimately excluded from analysis.

I also excluded vowels belonging to a set of phonetic stop words.[6] These words are function words such as *cause* and *let's* that tend to include reduced vowels that would potentially lead to unrepresentative data, or simply not enough data to extract accurate acoustic measures. Removing these stop words ensures that only stressed vowels were part of the analysis. A complete list of these phonetic stop words can be found in Appendix B.

After removing vowels belonging to stop words and vowels tagged as `0` or `1`, a total of 10031 English vowels remain: 576 breathy, 7631 modal, and 1824 creaky.

4.1.2 The Production and Perception of Linguistic Voice Quality Corpus

The other corpus used in this study is the Production and Perception of Linguistic Voice Quality Corpus (Keating, 2012).[7] It includes audio recordings for ten languages that use phonation in various ways. Each recording has a corresponding Praat text grid, most of which have the vowel boundaries hand-marked and annotated with the phonation type. Word lists and recording notes are also available, as well as electroglottographic data for some of the languages. For each language I will use from the Production and Perception of Linguistic

[6]Lists of *stop words*, which are words to be excluded from whatever the task is, are commonly used in computational linguistics. This is generally done to reduce statistical noise. *Stop* is unrelated to the phonetic sense of the word (plosive).

[7]The Production and Perception of Linguistic Voice Quality corpus is available at `http://www.phonetics.ucla.edu/voiceproject/voice.html`.

Voice Quality corpus (henceforth referred to as the Voice Project), I summarize, with as much detail as is available, how the language uses phonation, the recording procedures as described in the notes, and the steps I took to prepare the recordings and text grids for data extraction.

Gujarati

Gujarati (guj) is an Indo-European language spoken by 46 million people in India (Lewis et al., 2016). It contrasts modal and breathy vowels, sometimes called *clear* and *murmured* vowels. This contrast is seen in the minimal pair [baɾ] 'twelve' and [bə̤ɾ] 'outside.' Most of these breathy vowels are reflexes of historic [əɦiV] and [Vɦiə] sequences, though some may have developed independently (Khan, 2012).

Recording Procedure

The Gujarati recordings were made in a sound booth at UCLA from 2008 to 2009 by Sameer Khan. They consist of a word list read by three male speakers and seven female speakers. The list includes 75 words with twelve vowel qualities and two voice qualities (breathy and modal) on vowels. Gujarati also includes consonants that are aspirated and breathy (sometimes known as voiced aspirates), but these are excluded from the present study. The word list, excluding words in which the target vowel's phonation is determined by preceding aspirated and breathy consonants, can be found in Appendix D.

Annotation

Each word is a separate `.wav` file with information about its speaker and word contained in the file name. Corresponding text grids contain a single tier with vowel boundaries, vowel quality, and phonation. As Gujarati uses phonation contrastively, the phonation type listed is the *target* phonation type for that lexical item, which presumably but not necessarily corresponds with the actual phonation. I only include modal and breathy vowels in this study, and exclude the preceding aspirated and breathy consonants; this results in 2973 tokens, consisting of 1262 breathy vowels and 1711 modal vowels.

Hmong

Hmong (hmn) is a Hmong-Mien macrolanguage spoken by over 7.5 million people in China and various countries in Southeast Asia (Lewis et al., 2016). The two varieties included in the corpus are White and Green Hmong. Both dialects use complex contrastive tones that are distinguished by both pitch and phonation. These tones are described in Table 4.3 (Esposito, 2012). Hmong tones are orthographically represented by a word's final letter – tone *g* is breathy, tone *m* is creaky, and all others are modal. Note that for both non-modal tones, there is a modal tone with a similar or identical pitch (*s* for *m*, and *j* for *g*).

Table 4.3: Hmong Tones

Orthography	Tone	Example (White Hmong)
null	Mid (33)	[pɔ˧] po *spleen*
b	High (45)	[pɔ˥] pob *ball*
d	(214)*	[tɔ˨˩] tod *over there*
g	High-falling breathy (53)	[pɔ̤˥˩] pog *grandmother*
j	High-falling (53)	[pɔ˥˩] poj *female*
m	Low-falling creaky (21)	[pɔ̰˨˩] pom *to see*
s	Low (22)	[pɔ˨] pos *thorn*
v	Mid-rising (24)	[pɔ˧˥] pov *to throw*

**This tone is often excluded from descriptions of Hmong. I believe Esposito (2012) is referring to this tone in her brief note "There is also an eighth tone which is a syntactic variant of the low-falling creaky tone." As the phonation of this tone is not completely clear, and only one token of it appears in the word list, I exclude it from this study.*

Recording Procedure

The Hmong recordings were made by Christina Esposito and Sherrie Yang in Saint Paul, Minnesota in summer 2008. The recordings include 11 female and 21 male speakers of White Hmong, and two female and one male speaker of Green Hmong. Speakers read from a word list, with each word in the carrier phrase *say X again* [ɾo hai X duə]. The full word list is available in Appendix E.

Annotation

Each word, as well as its carrier phrase, is contained in a separate `.wav` file. The file names indicate the speaker, the word, the repetition number if applicable, and the variety of Hmong (White or Green). Accompanying text grids contain vowel boundaries marked with the vowel quality. I modified the text grids to contain a tier for the phonation type; the boundaries are the same as the vowel boundaries, and the phonation type was determined by the final letter of the word, which corresponds to the tone. Tone *g* (53) is breathy, *m* (21) is creaky, and all others are modal.[8] Due to incomplete annotation or inconsistencies in naming conventions, a significant number of Hmong files had to be excluded from this study; a total of 2717 tokens – 535 breathy, 1494 modal, and 688 creaky – are used.

Mandarin

Mandarin Chinese (cmn) is a Sino-Tibetan language spoken by over one billion people in China (Lewis et al., 2016). It is typically described as having four phonemic citation tones: high level (˥), mid to high rising (˧˥), mid to low to mid dipping (˨˩˧), and high to low falling (˥˩) (Lee and Zee, 2003). It is not described as using phonation contrastively, though voice quality and tone interact. In the Beijing dialect of Mandarin Chinese, as in other dialects, creaky voice optionally but frequently occurs with low tones, including the low dipping tone (˨˩˧) (Chao, 1968). In Tianjin Mandarin, a dialect closely related to the Beijing dialect, only the low dipping tone, as opposed to low pitches more generally, co-occurs systematically with a specific phonation type; Davison (1991) found that tone 214 has the lowest spectral tilt values, while tone 51 has the highest.

Mandarin differs from many of the languages in this corpus in that its non-modal phonation is optional and non-contrastive; regardless of the phonation of a tone, its pitch contour is enough to identify it. For the purposes of this dissertation, I will treat all tokens of the third tone (˨˩˧) as creaky. Though this is a potentially dangerous assumption, I listened to all the Mandarin recordings to manually adjust boundaries (see below) and found that

[8]The one word containing tone *d* is excluded from the study.

the third tone was reliably creaky; while this may not be the case in conversational speech, the citation form used in the recordings likely led to hyperarticulation and an extreme low pitch. If this is not the case and a significant proportion of third tone vowels are modal, the machine learning model will likely fail to find any patterns useful in distinguishing phonation types. The fourth tone (\) can be creaky towards the end, during the low portion. Some but not all of the speakers in this corpus produced a very brief period of creaky voice; because of the inconsistency in the presence of creaky voice on this tone, I will not include it in this study.

By coding all low dipping tones as creaky and high level and mid rising as modal, a machine learning model would conflate tone and phonation and simply rely on f_0, a known cue to tone differentiation, to separate creaky from modal vowels. For this reason, I will exclude the four f_0 measures from the model for Mandarin.

Recording Procedure

The Mandarin recordings include 12 speakers (six male and six female) saying the syllable [ma] with four different tones - 55, 35, 214, and 51. Each word (/ma/ plus a given tone) was read five times by each speaker, for a total of 240 tokens. The recordings were made by Jianjing Kuang in Beijing, China in 2011.

Table 4.4: Mandarin Tones on /ma/

Chao	Letter	Name	Gloss
55	˥	High level	Mother
35	˧˥	Mid rising	Hemp
214	˨˩˧	Low dipping	Horse
51	\	High falling	Scold

Annotation

Unlike other recordings from this corpus, the Mandarin text grids include boundaries that mark the word, but not the vowel. I manually adjusted these boundaries to capture the vowel and exclude the preceding consonant. The recordings appear to be readings of isolated

words. Treating all low dipping tones as creaky and all others as modal results in 120 modal tokens and 60 creaky tokens, for a total of 180 tokens. This is a very small sample size for machine learning, so caution will be needed in interpreting results.

Mazatec

Jalapa Mazatec (maj, sometimes called Jalapa de Díaz Mazatec) is an Otomanguean language spoken by 17,500 people in southern Mexico (Lewis et al., 2016). It has five vowel qualities, [a o u i[9] æ], though its vowel inventory is greatly expanded by three tones, nasalization, and length contrasts (Silverman, 1997). Additionally, Jalapa Mazatec uses breathy, modal, and creaky voice, with the "non-modal phonation realised primarily in the first portion of the vowel, actually beginning toward the end of any prevocalic sonorant" (Silverman, 1997). Non-modal phonation in the second half of the vowel is typically weakened to the point of being nearly modal.

Recording Procedure

The Mazatec data come from the Production and Perception of Linguistic Voice Quality Corpus by way of the UCLA Phonetics Archive. The recordings were made in Mexico City and Jalapa de Díaz, Tuxtepec District, Oaxaca, Mexico in 1982 and 1993. Six female and seven male speakers were recorded in 1982, and three additional males were recorded in 1993. Speakers read words in isolation from a word list; different lists were used during the two years of recording, though with some overlap. The full word list from each year can be found in Appendix F.

Annotation

Each token is its own `.wav` file whose name contains the word and speaker. Corresponding text grids contain an aligned transcription of the vowel quality and its phonation type and tone, each on separate tiers. However, as Mazatec phonation occurs primarily on the first half of the vowel, the vowel boundaries are not necessarily the same as the phonation

[9]This vowel is described as [i] by Silverman (1997) and Silverman et al. (1995), though the notes associated with the present corpus data describe it as [ɨ].

boundaries. Changes in acoustic measures due to non-modal phonation may be washed out when calculations are made across the entire vowel, but should be captured by those calculated over thirds of the vowel (see Section 4.2.1). In total, 482 vowels are included: 70 breathy vowels, 195 modal vowels, and 217 creaky vowels.

Zapotec

Zapotec (zap) is an Otomanguean macrolanguage spoken by about 441,000 people in Mexico (Lewis et al., 2016). This corpus includes two dialects: Santiago Matatlan Zapotec (smz, a dialect of Mitla Zapotec (zaw)) and San Juan Guelavía Zapotec (sjg, also known as Western Tlacolula Valley Zapotec (zab)) (Lewis et al., 2016). Few phonetic descriptions of these particular dialects are available; I will rely here on descriptions of dialects that belong to the same subgroups as the two dialects in the corpus.

Santa Ana Del Valle Zapotec, part of the San Juan Guelavía subgroup, contrasts breathy, modal, and creaky voice on its six vowel qualities [i e ɨ a u o]. Non-modal vowels all have a falling tone, and modal vowels contrast high or rising tone (Esposito, 2010). Mitla Zapotec, the larger group to which Santiago Matatlan Zapotec belongs, also contrasts breathy, modal, and creaky voice on its seven vowel qualities [i y e æ a u o]. Examples (1), (2), and (3) show this three-way phonation contrast (Stubblefield and Miller Stubblefield, 1991).

(1) /sa̤/ *manteca*[10]

(2) /sa/ *anda*

(3) /sa̰/ *bueno*

Recording Procedure

The Zapotec recordings are of six speakers of Santiago Matatlan Zapotec (four speakers; two male and two female) and two speakers of San Juan Guelavia Zapotec (both male).

[10]Stubblefield and Miller Stubblefield (1991) provide glosses in Spanish. I provide these same glosses rather than risk mistranslation, as there is some semantic ambiguity.

The recordings were made by Christina Esposito in Los Angeles in 2010. Speakers read from a list of 48 words including 13 modal vowels, 17 breathy vowels, and 18 creaky vowels. This list can be found in Appendix G. Some speakers recorded multiple tokens of the same word. Most vowels are flanked by consonants, though several vowels are word-final. These surrounding phones are not included in the text grids, and manually adding them would be prohibitively time consuming. Each recording contains the vowel and ten milliseconds of the surrounding phones. However, because little beyond the boundaries of the word is included, it is impossible to determine the vowel's distance from the end of the word or utterance.

Annotation

Vowel boundaries are indicated in text grids, which indicate the vowel quality, phonation type, and the word. Both phones and voice qualities are based on the canonical Zapotec pronunciation, as provided in the word list accompanying the recordings. The recordings contain 117 breathy vowels, 101 modal vowels, and 126 creaky vowels, for a total of 344 tokens.

4.1.3 Summary of Data Sets

Table 4.5 below summarizes the data sets used in this study – the languages, the types of phonation they use and how they use them, and the number of speakers and tokens ultimately included. Throughout this dissertation, I will often refer to the languages by their ISO 639-3 codes, listed alongside the language names.

4.2 Data Extraction

I extracted the features described in Chapter 3 from all vowels annotated for phonation in the above corpora[11] using two programs – Praat (Boersma and Weenink, 2016) and VoiceSauce (Shue et al., 2011). I used a Praat script to extract jitter, shimmer, vowel

[11] This excludes English vowels tagged as `0` or `1` and those belonging to phonetic stop words. It also excludes a significant number of Hmong vowels that I was unable to automatically process due to inconsistencies in annotation or naming convention.

Table 4.5: Summary of Data Sets

Language	Tokens				Speakers	Phonation Use(s)
	B	M	C	Total		
English *eng*	576 *5.742%*	7631 *76.074%*	1824 *18.184%*	10031	22	Allophonic, Prosodic, Sociolinguistic
Gujarati *guj*	1262 *42.449%*	1711 *57.551%*	–	2973	10	Contrastive
Hmong *hmn*	535 *19.691%*	1494 *54.987%*	688 *25.322%*	2717	35	Alongside Tones
Mandarin *cmn*	–	120 *66.667%*	60 *33.333%*	180	12	Alongside Tones
Mazatec *maj*	70 *14.523%*	195 *40.456%*	217 *45.021%*	482	16	Contrastive
Zapotec *zap*	117 *34.012%*	101 *29.36%*	126 *36.628%*	344	6	Alonside Tones

duration, surrounding phones, and prosodic position. The English text grids from the ATAROS corpus contain more information than the text grids from the Voice Project, such as spurt[12] and word boundaries, and this different format required a slightly modified Praat script. The scripts (one for each language) are available on GitHub[13]. The remaining features – spectral tilt, RMS energy, HNR, SHR, CPP, f_0, variance of pitch tracks,[14] and F1 – were extracted from VoiceSauce.

Two significant decisions were made during data extraction – over what time period to make calculations, and the pitch settings. These decisions are described below.

[12]A *spurt* is defined as a "stretch of speech said by one speaker between at least 500 ms of silence." Spurts were manually annotated in the ATAROS data (Freeman, 2015).

[13]`https://github.com/lpanfili/dissertation/tree/master/extract-voice`

[14]Variance of pitch tracks was not directly extracted using VoiceSauce, but calculated based on f_0 data extracted by VoiceSauce.

4.2.1 Temporal Calculations

The boundaries of voice qualities do not necessarily coincide perfectly with the boundaries of a vowel; conflating the two could result in washing out the acoustic effects of non-modal phonation when looking at averages over the entire vowel. Mazatec, for example, generally realizes non-modal phonation at the beginning of the vowel, and phonation approaches modal by the end of the vowel (Silverman, 1997); looking at the vowel as a whole may not capture the fact that its onset was non-modal. For this reason, all possible measures[15] will be calculated as averages over each third of the vowel, as well as over the entire vowel.

4.2.2 Pitch Settings

Praat and VoiceSauce allow users to enter the minimum and maximum frequencies considered in finding pitch candidates; the algorithm will only search for candidates within this range. To select pitch settings to use in this study, I tried six different combinations of minimum and maximum pitch settings on the 22 English speakers from that ATAROS Corpus. My diagnostic for the best combination of pitch settings was the number of jitter and f_0 measures with undefined calculations in Praat, as these are the most frequently undefined, particularly in English, but potentially extremely useful measures.[16] Praat outputs "`--undefined--`" when it cannot calculate the requested value; for instance, it cannot calculate the jitter of a segment if it cannot locate the glottal pulses. A reduced number of undefined values suggests that the settings are better-suited to accurately detecting (or at least, to detecting in the first place) the relevant information about the source.

Table 4.6 lists the various pitch settings I tried on the English data and the resulting count of undefined results for jitter and f_0. I show raw counts of undefined results and the percentage of all results that were undefined for the entire data set. As men and women

[15]Measures that will be calculated over thirds of the vowel, as well as over the entire vowel, are spectral tilt, jitter, RMS energy, shimmer, HNR, SHR, CPP, f_0, and F1.

[16]This includes local jitter, local absolute jitter, RAP jitter, and PPQ5 jitter. I used Praat's f_0 algorithm. All jitter measures and the f_0 measure were computed across the entire vowel.

have different sized vocal folds, resulting in different pitch ranges, I anticipated that different pitch settings may impact f_0 and jitter calculations in different ways for male and female speakers, so I also separate counts and percentages for each sex.

Table 4.6: Testing Pitch Ranges in Praat

	Pitch Range (Hz)		Undefined Counts		
	Male	Female	Overall *(percent)*	Male *(percent)*	Female *(percent)*
Praat Defaults	75–600	75–600	15209 *(12.63%)*	7835 *(13.43%)*	7374 *(11.88%)*
VoiceSauce Defaults	40–500	40-500	14773 *(13.43%)*	7439 *(15.52%)*	7334 *(11.82%)*
Low Min	50–600	50–600	15183 *(12.61%)*	7850 *(13.46%)*	7333 *(11.81%)*
Very Low Min	25–600	25–600	15546 *(12.91%)*	8190 *(14.04%)*	7356 *(11.85%)*
Sex-Specific	75–300	100–500	15159 *(12.91%)*	7791 *(13.36%)*	7748 *(12.48%)*
Sex-Specific Max, Low Min	50–300	50-500	15561 *(12.59%)*	7829 *(13.42%)*	7330 *(11.82%)*

Surprisingly, the six settings have little impact on the number of undefined results; the percentage ranged from 12.59% to 13.43%. I was additionally surprised that all six settings performed better for female speakers than for male speakers. Given how little variation exists in undefined counts,[17] I will use Praat's default settings of 75–600 Hz for both male and female speakers for calculations in Praat and VoiceSauce.

[17]Praat pitch tracking errors can also present themselves as doubling or halving errors, which would not have been noted using these methods.

4.3 Data Processing

The data extraction described in Section 4.2 provides all the data that will be used in the machine learning model, though three modifications are still necessary – normalizing the data, converting undefined measures into interpretable and meaningful measures, and balancing the data set. These three processes are described below.

4.3.1 Normalization

Many machine learning models require normalized data to work properly. Additionally, some of the features must be normalized by speaker to account for variation between speakers. All data in this study are Z-normalized, either by speaker or over an entire language's data. Table 4.7 lists the features and whether they are normalized by speaker or overall; these decisions are explained below.

Measures that are normalized by speaker are RMS Energy, HNR, f_0, F1, and vowel duration. RMS Energy must be normalized by speaker to account for differences in talker loudness, as well as for variability in microphone distance and recording settings. Similarly, HNR is impacted by the recording procedure and can also vary inherently between speakers. f_0 must be normalized by speaker to account for variation in speaker pitch. While we may expect to see a small vowel effect due to Intrinsic Fundamental Frequency (If_0, the observation that low vowels have a slightly lower f_0 than high vowels cross-linguistically), this effect is likely too small to warrant normalization by vowel. Vocal tract differences between speakers will result in different F1 values for different speakers, so they must also be normalized by speaker. Finally, vowel duration will be normalized by speaker, as rate of speech varies. While the voicing of the following phone can impact a vowel's duration, normalizing by following phone would be problematic because the following phone's voicing is considered separately as a feature in the model.

Measures that do not require by-speaker normalization are jitter, shimmer, SHR, CPP, VoPT, voicing, place and manner of surrounding phones, and distance from the end of the

word and utterance. Many of these measures are not expected to vary by speaker or by vowel quality, and others cannot be normalized. Jitter is a fairly direct measure of phonation, and is therefore dangerous to normalize. As a standard deviation measure, it is more robust against variation between speakers. Additionally, I do not anticipate vowel quality-induced jitter differences and by-speaker normalization would wash out variation in the rate of non-modal phonation between speakers. Though shimmer is related to RMS Energy, it is, like jitter, a standard deviation and therefore it is not necessary to normalize by speaker, and vowel quality effects are unlikely. SHR and CPP are also not expected to vary by speaker or vowel quality. VoPT, as a measure of the agreement between pitch tracks, does not need to be normalized by vowel or by speaker. The surrounding phones measures cannot be normalized, but measures of prosodic position will be normalized overall.

Spectral tilt is a slightly different case. These measures are already corrected for vowel quality in VoiceSauce, as they are significantly impacted by the formant frequencies. As spectral tilt is a difference measure, speaker normalization is not necessary.

Table 4.7: Normalization of Features

Measure	Category	Normalization
H1* – A1*	Spectral Tilt	Overall; already by vowel
H1* – A2*	Spectral Tilt	Overall; already by vowel
H1* – A3*	Spectral Tilt	Overall already by vowel
H1* – H2*	Spectral Tilt	Overall; already by vowel
H2* – H4*	Spectral Tilt	Overall; already by vowel
H4* – 2k*	Spectral Tilt	Overall; already by vowel
2k* – 5k	Spectral Tilt	Overall; already by vowel
Local Jitter	Jitter	Overall
Local Absolute Jitter	Jitter	Overall
RAP Jitter	Jitter	Overall
PPQ5 Jitter	Jitter	Overall
RMS Energy	Intensity	By speaker
Local Shimmer	Shimmer	Overall
Local Shimmer, dB	Shimmer	Overall
APQ3 Shimmer	Shimmer	Overall
APQ5 Shimmer	Shimmer	Overall
APQ11 Shimmer	Shimmer	Overall
HNR05	Harmonics-to-Noise	By speaker
HNR15	Harmonics-to-Noise	By speaker
HNR25	Harmonics-to-Noise	By speaker
HNR35	Harmonics-to-Noise	By speaker
SHR	Subharmonic-to-Harmonic	Overall
CPP	Cepstral Peak Prominence	Overall
Praat f_0	Fundamental Frequency	By speaker
SHR f_0	Fundamental Frequency	By speaker
Snack f_0	Fundamental Frequency	By speaker
STRAIGHT f_0	Fundamental Frequency	By speaker
Variance of Pitch Tracks	VoPT	Overall
F1	F1	By speaker
Vowel Duration	Duration	By speaker
Voicing of Preceding Phone	Surrounding Phones	None
Voicing of Following Phone	Surrounding Phones	None
Manner of Preceding Phone	Surrounding Phones	None
Manner of Following Phone	Surrounding Phones	None
Presence of Preceding Phone	Surrounding Phones	None
Presence of Following Phone	Surrounding Phones	None
Distance from Utterance End (ms)	Prosodic Position	Overall
Distance from Utterance End (percent)	Prosodic Position	Overall
Distance from Word End (percent)	Prosodic Position	Overall

* *Indicates measures that have been corrected for formant frequencies and bandwidths in VoiceSauce.*

4.3.2 Handling Undefined Measures

Despite attempts to reduce the number of undefined measures by tailoring pitch settings, as described in Section 4.2.2, some remain. Many of these are jitter and shimmer measures in Praat that are output as `--undefined--`. VoiceSauce also encounters calculation errors for f_0, as well as SHR, which are output as `0`.[18] These two types of undefined measures must be converted to something that is both interpretable and meaningful for the models.

Several approaches to this problem are possible. Perhaps most obviously, I could exclude all vowels that have an undefined measure. This would greatly reduce the number of data points and, more importantly, likely skew the data set by removing particularly aperiodic vowels, as aperiodicity is often the reason a measure is undefined. A second option is to reduce the number of undefined measures by picking the calculations of a feature that result in the fewest undefined measures. Another option is to replace undefined values with another value, often the mean of that measure over other similar data points. Ideally, this value would be the mean of that feature for a given speaker's production of that vowel quality with that phonation type. Unfortunately, in this study, that mean would be based on far too few data points to be meaningful. Finally, I could make the measurements by hand, though the number of undefined measures in this particular data set is far too large to make this a feasible option.

I opt for a twofold approach to handling undefined measures – replacing undefined measures with the mean of that measure by class (phonation type) in the training data and the overall mean in the testing data,[19] and excluding features with too many undefined measures. Training data is divided into subsets called folds (see Chapter 5 for more information); the mean of each class will be based on the data in each fold.

Deciding how many undefined measures is too many is not a straightforward question,

---

[18]VoiceSauce's f_0 errors are generally washed out because they are output for a single time point, but I only consider f_0 as averaged over a timespan of at least one third of the vowel.

[19]Replacing undefined measured in the testing data with the class mean would unfairly boost performance, as it requires knowing the class, which is the ultimate goal.

and no industry standard exists for this. I determined a cutoff point by running a single feature SVM for all features and all languages. The classifier uses the information from one feature at a time to separate phonation types. Replacing a large percentage of the data points with the mean often boosts the classifiers accuracy. For these features, high accuracy does not reflect their ability to distinguish phonation types; rather, it reflects the fact that nearly all the values are identical for a given class, as undefined values are replaced with the same mean on a fold-by-fold basis. When too many values are identical, the classification task becomes artificially easy, resulting in inflated accuracy. Figure 4.1 shows how accuracy increases as the percentage of undefined values increases. This plot includes all features for all six languages, along with a line of best fit.

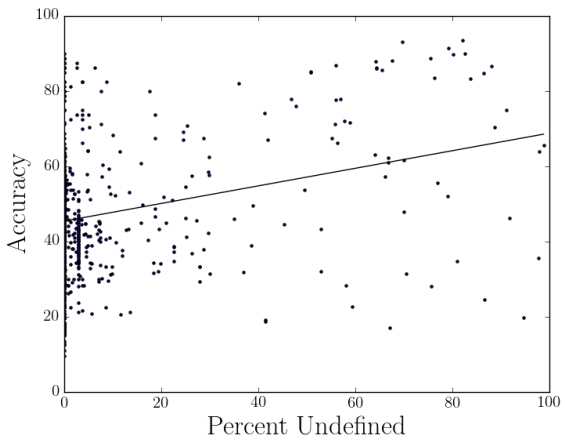


Figure 4.1: Percent Undefined vs. Accuracy (All Features, All Languages)

Figure 4.1 shows a trend towards higher accuracy as the percent undefined increases. Though no point is immediately obvious at which a high percent of undefined measures artificially inflates accuracy, the upward trend emerges after a cluster with fewer than 15% undefined measures. Based on this information, I will exclude all features that have 15% or more undefined measures on a language-by-language basis.

4.3.3 Balancing the Data Set

Many of the corpora described above suffer from imbalanced classes; they have very different numbers of tokens of different phonation types. Imbalanced data sets can be problematic for machine learning models. The two primary ways of balancing a data set are described below.

The first option is *undersampling*. In undersampling, each class is reduced to the number of instances of the smallest class. That is, given 10 B, 70 M, and 20 C, the data set would be reduced 10 B, 10 M, and 10 C. Of course, the drawback to undersampling is that data are thrown out, especially in extremely unbalanced data sets like some of those included in this dissertation. Table 4.8 lists the number of tokens for each language, as well as this figure after undersampling has occurred. For most of the languages, undersampling reduces the token count too much to be useful.

Table 4.8: Undersampling Token Counts

Language	Actual Size	Undersampled Size
English (eng)	10031	576
Gujarati (guj)	2973	1262
Hmong (hmn)	2717	535
Mandarin (cmn)	180	60
Mazatec (maj)	482	70
Zapotec (zap)	344	101

The second option is *oversampling*. This involves making up the difference between the majority class (in this case, modal voice) and the minority class (in this case, both breathy and creaky voice) by adding more instances of the minority class. Those instances can be duplicates of pre-existing ones, or they can be synthesized.

Given the extreme loss of data that undersampling would cause, I opt to oversample using SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002), a widely used oversampling algorithm that creates synthetic instances of the minority class based on two

similar instances with small changes made based on neighboring instances. See Chapter 5 for more information about resampling procedure.

With measures normalized, undefined measures replaced and reduced, and the data set balanced, I now have a set of features for each language that is ready for the machine learning model. The following chapter describes the basics of machine learning and how I will use it to examine the properties of phonation types in different languages.

Chapter 5
MACHINE LEARNING

This chapter provides an overview of machine learning, the primary statistical tool used in this dissertation. I first describe the basic concept of machine learning and the two algorithms I use. I then briefly discuss some of the uses of machine learning in linguistics and its relevance to the present question. Finally, I describe the metrics used to evaluate a machine learning model's performance to evaluate each feature's contribution to that performance.

Before doing that, I'd like to briefly review the primary and secondary goals of this dissertation, as they are both important in my decision to use machine learning as the tool to answer these questions, and in my decisions about which types of machine learning models to use. The primary goal of my dissertation is to answer the *linguistic* questions: which acoustic properties best distinguish phonation types in a given language, how do those properties vary from language to language, and how do they vary between languages that use phonation in different ways? The secondary goal is the *classification* question, specifically for English: I aim to build a machine learning model (also called a *classifier*) that predicts the phonation type of a given English vowel with high accuracy. These two goals require slightly different pieces of information and methods, and are important to keep in mind throughout this chapter's overview of machine learning.

5.1 Machine Learning Basics

The basic concept behind classification in machine learning[1] is that an algorithm observes patterns in a data set and then applies those patterns to make predictions about new data.

[1]Regression is another type of machine learning that predicts a continuous variable rather than a categorical variable.

More specifically, the algorithm aims to group the new data into *classes* based on the classes in the original data set. The classes in the original data set can be either pre-determined by the researcher in *supervised* learning, or discovered by the algorithm in *unsupervised* learning. I'll be using supervised learning in this dissertation, so I will focus on that in this section.

Before delving into how machine learning applies to the question of phonation classification, I'll review a different example. A classic example of supervised machine classification is tumor diagnosis. In this example, the goal is predicting whether a previously unseen tumor is benign or malignant given information about it. We'll walk through this example to see the different pieces of the machine learning process.

First, the machine learning algorithm is given a data set as input. This *training data* consists of *instances* of tumors. For each tumor in the training data, we know its *class* (also called a *label*) (`benign` or `malignant`) and some potentially useful information about it. Ideally, this information is about properties of the tumor that help determine whether it is benign or malignant, such as its size and its rate of growth. These predictors are called *features*. The algorithm uses the features and the labels in the training data to extrapolate patterns that separate the data by class. This process – the algorithm searching for patterns – is called *training*. With the classifier trained, it can now be shown the *test set*, which consists of previously unseen tumours. The model will classify the data from the test set by applying the patterns it learned in the training set; given a tumor's size and its rate of growth, the classifier will predict whether each tumor is benign or malignant.

In order to know how well the classifier did in determining whether a tumor is benign or malignant, we need to know the actual class for the tumors in the test set. We can then compare the actual class of each tumor in the test set to the class predicted by the model. I describe the metrics for evaluating a classifier's performance in Section 5.4. If the classifier performs well enough, it can be used to classify real-world data: tumors whose actual class is unknown.

In this dissertation, the goal is to classify a vowel as breathy, modal, or creaky. The training data consists of vowels whose phonation type is known, either because phonation

is predictable (contrastive or tonal phonation) or because the data have been manually annotated (sociolinguistic, prosodic, and allophonic phonation). For each vowel, the features consist of a variety of acoustic and contextual measures that have been associated with phonation, as described in Chapter 3, and the process of extracting the features is described in Chapter 4. The patterns observed in the training data are then applied to test data and used to classify the phonation type of unlabeled vowels.

Classification, however, is the secondary goal of this dissertation. The primary goal is linguistic; this is a study of the acoustic properties of phonation types in different languages and how they vary in languages that use phonation differently. Machine learning allows me to answer these questions by telling me which features the algorithm relies most on in making its classification decisions. This information is available for many machine learning models in the form of a number representing each feature's importance to the classifier. This information – feature weights and importance (discussed in Sections 5.5.2 and 5.5.3) – sheds light on which acoustic properties describe phonation types.

5.2 Machine Learning in Linguistics

Machine learning has a wide range of linguistic applications, though it has most often been used for its classification power rather than as a tool for gaining linguistic knowledge. Email spam filters and machine translation are powered by machine learning. It can be used to predict whether an Amazon review is positive or negative, when a Facebook post describes a major life event, and the political leaning of a tweet's author. The goal of machine learning in all these examples is high-accuracy classification, as is the goal for the classification part of my dissertation.

Machine learning has also been used to gain linguistic information, as I aim to do, by examining *why* a classifier works. In her dissertation about *um* and *uh* in stance, Le Grézause (2017) shed light on the discourse roles of those two disfluencies by comparing a machine learning model's accuracy in classifying stance with and without *um* and *uh* as features. She found that they are, in fact, useful features in predicting stance, indicating that they are

more than simply disfluencies. Styler (2015) used machine learning to study nasalization. He used two different models – a Support Vector Machine and a Random Forest, described in Section 5.3 – to study contrastive nasality in French and coarticulatory nasality in English. Using these machine learning models, he was able to identify the most salient acoustic correlates of these two types of nasality.

While many studies have examined the acoustic properties of phonation types, machine learning brings a new methodology to the field. Most studies have focused on voice quality in a given language, and different statistical tools and methodologies have been employed for different languages. In this dissertation, I apply the same methodology across six languages, making the results straightforward to compare.

Machine learning offers several advantages over more traditional statistical tools. First, it *discovers* patterns rather than validating them. An algorithm can discover interactions between features without being explicitly told to do so. Additionally, the patterns that it discovers are easily generalizable to new data, which is particularly important when building a classifier is part of the goal. Many models handle large sets of features well; given a large set of features, the algorithm will determine what's important and what's not, and then will focus on the important features. While not all models are transparent about this, some output a ranking of which features are most important. Machine learning models can quickly and efficiently analyze large amounts of data and juggle large numbers of features, making them well suited to tackle complex problems, such as phonation. Finally, machine learning allows me to use the same tool to address both the linguistic and classification questions.

In this dissertation, I use two machine learning models: a Support Vector Machine and a Random Forest. In Section 5.3 below, I describe why I picked these two models and how they work.

5.3 Two Machine Learning Models

Though many machine learning models exist, I focus here on two: the Support Vector Machine (SVM) and the Random Forest. These two models have several advantages. First

and foremost, both provide a ranking of which features were most important in drawing the decision boundaries; this information is critical in answering the linguistic questions but is not available from all models. Additionally, both are very powerful, widely-used, and robust to relatively small data sets. I describe how these two algorithms work in Sections 5.3.1 and 5.3.2 and some logistical considerations for machine learning in Section 5.3.3.

5.3.1 Support Vector Machines

The Support Vector Machine is a widely used machine learning model that works by essentially finding lines that separate classes. Though the idea of the SVM has been around for a while, it was first described close to its current state by Boser et al. (1992).

Returning to the tumor example, imagine the training data plotted on a plane. With two features – the tumor's size and the patient's age – we have just two dimensions; in a problem with more features, the plot has many dimensions. The SVM's goal is to find the line (or hyperplane, when more dimensions are involved – this is more generally called the *decision boundary*) that best separates the data points by class and gives the widest margin between the two classes. The data points that dictate the line's path are called the *support vectors*. Figure 5.1 shows a hypothetical set of training data. Each point is a tumor, and the line is the separation drawn by the SVM. The data points that lie along the edge of the line, marked with stars rather than circles, are the support vectors. This dividing line can then be used to predict whether a previously unseen tumor is benign or malignant.

This tumor example uses just two features, so the classes can be separated in one dimension. In general, given a data set with n features, the SVM builds an n - *1*-dimensional hyperplane. Higher dimensional problems are harder to visualize, but the same concept applies – the SVM tries to separate the classes by the largest possible margin.

The tumor example has two classes: benign and malignant. Four of the six languages used in this study have three classes: breathy, modal, and creaky voice. A multiclass problem is slightly more complicated than the binary classification task used in the tumor example. The SVM can handle this in one of two ways, both of which involve training multiple classifiers

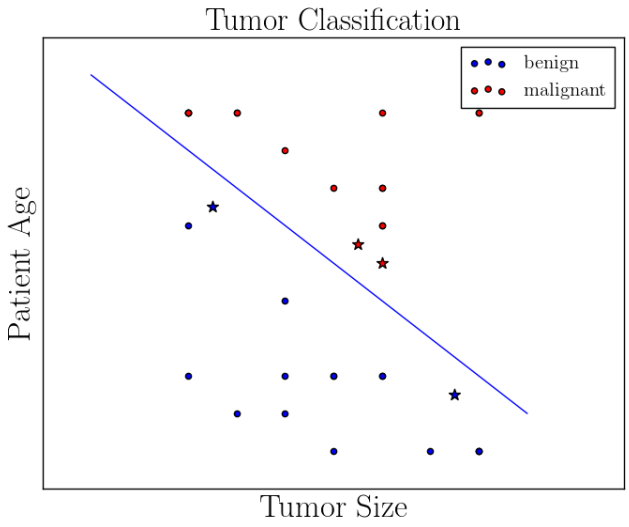


Figure 5.1: SVM Separating Hypothetical Tumor Data

and then synthesizing their results. The *one-vs.-rest* strategy trains models that classify instances as belonging to class X or not belonging to class X. In the case of three-way phonation classification, that would mean training the following models:

1. Breathy vs. Non-Breathy (Modal + Creaky)
2. Modal vs. Non-Modal (Breathy + Creaky)
3. Creaky vs. Non-Creaky (Breathy + Modal)

The other option is to train *one-vs.-one* models. This involves a separate classifier for each combination of two classes. For the three phonation classes, the following three models would be trained:

1. Breathy vs. Modal
2. Breathy vs. Creaky
3. Creaky vs. Modal

I opt for the one-vs.-one option for this study, largely because I am interested in the breathy vs. creaky classification. Breathy voice and creaky voice are often presented as

opposite ends of the phonation continuum, and yet they share many characteristics, such as decreased periodicity and energy relative to modal voicing (see Chapter 3). I am interested in exploring the details that distinguish the two non-modal classes. I use `scikit-learn`'s `svm.SVC` implementation, which uses the one-vs.-one strategy (Pedregosa et al., 2011).

5.3.2 Random Forests

Random Forests are ensemble classifiers, meaning that they construct and combine multiple classifiers. The classifiers used in Random Forests are called *Decision Trees*. To understand Random Forests, we must first understand Decision Trees.

Let's go back to the tumor example. When presented with a tumor, a doctor seeks pieces of information about it to determine whether it's benign or malignant. How fast is it growing? How big is it? Does it light up in a PET scan? How old is the patient? After asking enough questions and the relevant questions, the doctor arrives at a diagnosis, classifying the tumor as either benign or malignant. The doctor has essentially worked through a Decision Tree by asking a series of questions about features relevant to the classification and ultimately classifying a tumor.

In machine learning, a Decision Tree uses labeled data to determine the sequence of feature-based questions that leads to the highest classification accuracy. In addition to ordering the questions, the Decision Tree must pick the criteria for answering the questions. For discrete features, this can be a yes/no question, such as whether or not a tumor lights up in a PET scan. For continuous data, like patient age, a cutoff point is needed. If the training data shows that a certain type of cancer is more likely in patients 65 or older, 65 would be the number the decision tree picks to determine which branch to continue on; if the patient is under 65, we follow one branch, and we follow the other if they're 65 or older. When presented with unclassified data, the Decision Tree starts with the *root node* and works through the series of *decision nodes* (those questions, representing decision points) to predict its class in the tree's *leaves*. An example of a basic Decision Tree is shown in Figure 5.2.

Decision Trees are fairly transparent – the ordering of the nodes reflects which features are

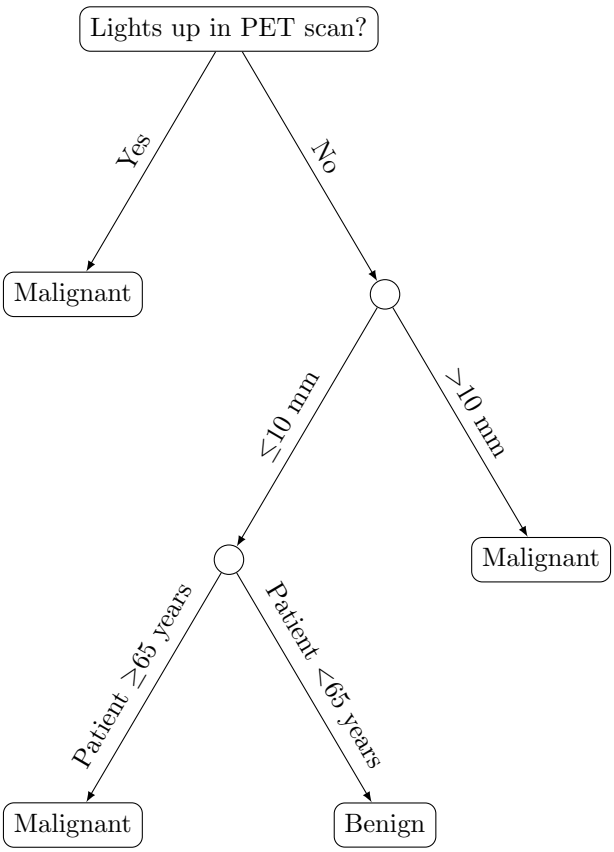


Figure 5.2: A Hypothetical Decision Tree for Tumor Classification

most distinguishing at each step and we know the criteria used for each (the cutoff point for continuous data). They also weed out unimportant features. If the doctor asked a question that is ultimately irrelevant to the tumor's diagnosis (such as whether it's on the right or left side of the body), that feature would be omitted from the Decision Tree because it does not provide any useful information.

The phonation classification problem involves many more features than the tumor classification example. If the tumor classification had many more features, it might be hard for a single doctor to optimize their ordering. Additionally, deeper Decision Trees (those with a longer path to the decision due to more features) are prone to overfitting, or building a model that fits one specific data set but will not reliably fit another. If that doctor presented the case to a group of doctors at a conference, each of those doctors might come up with their own subset of questions to ask and their own ordering of those questions. They could then look across all of those Decision Trees to find the features most accurately separate the classes, and avoid overfitting the model. This is, essentially, a Random Forest.

To classify phonation, the Random Forest will construct a series of Decision Trees using subsets of the features to determine the best set and ordering of features. When presented with the test data, the classifier works through those features to determine each instance's class.

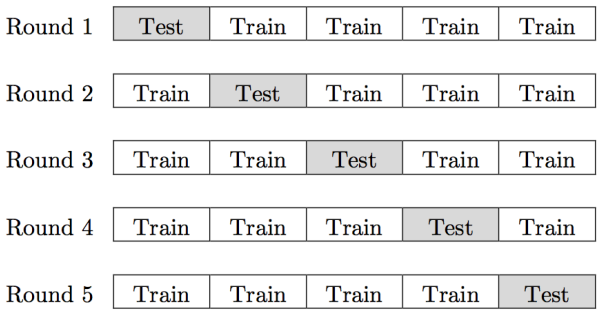
5.3.3 Some Logistical Considerations in Machine Learning

Running a machine learning model often requires making a daunting number of decisions that can impact its performance. For the linguistic questions, in which high classification accuracy is not the goal, I keep things simple by generally keeping default parameters. However, there are two important pieces of my methods that warrant description – cross-validation and kernelization.

Cross-Validation

To split the data set into training data and testing data, I use five-fold *cross-validation*. Cross-validation is a way of splitting the data into testing and training sets in an iterative manner. It's particularly useful when working with small data sets to avoid making either set too small and to reduce variability. The full data set is first divided into k subsamples (*folds*) of equal size; in the case of five-fold cross-validation, a standard way of performing cross-validation, $k = 5$. The model is then trained on $k - 1$ folds, and tested on that one fold that was excluded from the training set. This process is repeated until each of the k folds has been used as the test set. Figure 5.3 shows a schematic example of how a data set is split up during five-fold cross-validation.

Figure 5.3: Five-Fold Cross-Validation



Using k-fold cross-validation has the advantage that each data point is, at some point, used in both the training and testing data. Often, ten folds are used in cross-validation. However, due to the small size of some of my data sets, I opt for five-fold cross-validation to ensure that no training or testing set is too small. The metrics used to evaluate the model's performance (see Section 5.4) are available on a by-fold basis, though combining them across the five folds provides the best picture. I report all metrics calculated cumulatively over the folds, rather than averaged across the five folds; in other words, the accuracy is not calculated for each fold and then averaged, but the correctly and incorrectly identified instances are

summed over the folds and the accuracy calculated based on the total.

Recall from Chapter 4 that I resample the data, using the SMOTE technique to add synthetic instances of the minority class(es). I resample only the training data and resample based on only those folds. This means that the synthetic training data are not based on the testing data, and that the testing data contains no synthetic instances.

Kernelization

Kernelization is a technique specific to Support Vector Machines. The tumor example shown in Figure 5.1 used a straight line to separate the classes. Whether the decision boundary is a straight line or a more complex shape is determined by the *kernel* used by the SVM. The kernel that leads to straight lines is called a *linear* kernel. Realistically, classes often can't be separated by a straight line. This is where a *kernel trick* comes in.

Kernels other than linear ones essentially transform the data to a high-dimensional feature space that allows the model to better separate the classes. However, this improved classification comes with a significant drawback: information about how the model used each feature is more difficult to interpret when the kernel is anything but linear. This means that I *must* use a linear kernel, which is not `scikit-learn`'s default kernel, in order to answer the linguistic questions. Other kernels are options in answering the classification question, as classification accuracy, rather than feature importance, is the goal; I explore other kernels in Chapter 13.

This section has focused on how the two classifiers – the SVM and the Random Forest – go about finding patterns in the training data and applying them to the test data. The following section focuses on ways to evaluate how well the classifier performs.

5.4 Evaluating Model Performance

Before delving into the details of *why* a machine learning model works, it's important to evaluate *how well* it's working. Working well is also the primary goal of building an English classifier in Chapter 13. Working well can be defined in various ways. I use five standard

metrics to describe a classifier's performance – accuracy, precision, recall, F1 score, and weighted F1 score.

I'll use an example data set to illustrate these metrics. This data set, shown in Table 5.1, consists of ten instances of tumors. The lefthand column indicates each tumor's *actual* class and the righthand column indicates what a machine learning model predicted each tumor's class to be. Correct guesses are shaded green and incorrect guesses are shaded red.

Table 5.1: Example Data

Actual	**Predicted**
Benign	Benign
Benign	Benign
Benign	Malignant
Benign	Malignant
Malignant	Malignant
Malignant	Malignant
Malignant	Malignant
Malignant	Malignant
Malignant	Malignant
Malignant	Malignant

5.4.1 Accuracy

Accuracy is simply what percent of the test data the model classified correctly. In this case, it's the percentage of vowels whose phonation type was correctly identified. For the data set in Table 5.1, the accuracy is 80%, as the classifier correctly identified the class for eight of the ten tumors. While this metric seems very straightforward, it is potentially problematic for an imbalanced data set. The example data set is not balanced; there are more (actually) malignant tumors than benign tumors. If the classifier were to simply label everything as

malignant, it would correctly classify six of the ten instances, achieving 60% accuracy, and this baseline must be considered in interpreting accuracy. Here, 80% accuracy is quite a bit better than 60% accuracy. This gets more problematic the more skewed the data set is.

Take the English data, described in Chapter 4, as an example. In its imbalanced form, 78.9% of the instances are modal. If the classifier labels *all* tokens as modal, it achieves a very solid 78.9% accuracy despite making extremely naive predictions. Treating this as the baseline, we see that 78.9% accuracy is, in fact, low. For a balanced data set, the accuracy baseline is the percentage of the data set represented by each phonation type – 50% for languages with two phonation types and 33.3% for languages with three.

5.4.2 Precision

While accuracy is reported over the entire data set, precision, along with recall and F1 score, is reported by class. *Precision* measures the positive predictive ability of the classifier. It is the number of true positives (T_P) over the number of true positives and false positives (F_P). Precision is calculated as $P = \frac{T_P}{T_P+F_P}$. In the example data from Table 5.1, the classifier has benign precision of 1.0 (2/2) and malignant precision of 0.75 (6/8). For classification of phonation types, precision is the number of vowels that are actually creaky (or breathy, or modal) divided by the total number of vowels the classifier labelled as creaky (or breathy, or modal). It does not include vowels that are actually creaky but were not identified as such by the classifier – that's a job for recall.

5.4.3 Recall

Recall measures how complete the classifier is. It is the number of true positives over the number of true positives and false negatives (F_N), effectively measuring how many of the members that actually belong that class were correctly identified. Recall is calculated as $R = \frac{T_P}{T_P+F_N}$. In the example data, benign recall is 0.5 (2/4) and malignant recall is 1.0 (6/6). For phonation, it is how many tokens were identified as creaky (or breathy, or modal) from among the set of *actually* creaky (or breathy, or modal) tokens.

5.4.4 F1 Score

F1 score is the harmonic mean of precision and recall. In considering both precision and recall, F1 score is an important measure for imbalanced data sets and is often considered the most representative single metric *if* precision and recall are equally important. F1 score is calculated as $F = 2\frac{P*R}{P+R}$. In the tumor example data, the benign F1 score is 0.67 and the malignant F1 score is 0.86.

5.4.5 Weighted F1 Score

Finally, weighted F1 score is a single number (not broken down by class) that considers the precision and recall for each class as well as the balance between the classes. The example data from Table 5.1 has a weighted F1 score of 0.78. I treat weighted F1 score as the best representation of a classifier's overall performance. Weighted F1 scores range from 0 to 1, with 1 being the best. A weighted F1 score of 1.0 would mean perfect precision and recall; this is not a realistic expectation for this classifier.

I begin each language's chapter by presenting the accuracy, weighted F1 score, precision, recall, and F1 score for four classifiers: an SVM and a Random Forest, each based on imbalanced and resampled data sets. These five metrics are also key in fine-tuning the English classifier in Chapter 13.

5.5 Evaluating Feature Importance

In order to answer the linguistic questions, I need to find out which features are most useful to the classifiers in distinguishing phonation types. I examine feature contributions through four lenses: correlations, feature weights, feature importance, and ablation, each outlined below.

5.5.1 Correlations

Correlations, which do not involve machine learning, are a more traditional statistical test that measures the strength of the relationship between two variables. In this case, that's the relationship between a feature and its class. I use Pearson correlations, a widely used statistical measure, using normalized data[2] to calculate coefficients that represent this relationship.

Correlation coefficients range from -1 to 1, though values near the extremes are rare. -1 means that a feature is perfectly negatively correlated with a phonation type, while 1 means that a feature is perfectly positively correlated with a phonation type. Because both positive and negative ends of the spectrum indicate strong correlation, it's important to consider the *magnitude* (absolute value) of the correlation. Typically, ± 0.5 to ± 1 is considered a strong correlation, ± 0.3 to ± 0.49 is considered a medium correlation, and anything below that is considered a low correlation.

Correlations are usually calculated over an entire data set and effectively measure the strength of a one-vs.-rest relationship. However, because I opt to run the SVMs as one-vs.-one, I opt to calculate the correlations as one-vs.-one to facilitate comparison between the metrics. To do so, I remove one phonation type at a time from the data set and then calculate the correlations based on the other two types. For example, to calculate the correlations for features that distinguish creaky voice from modal voice, I include only the creaky and modal vowels when I calculate the correlations. If HNR05_Mean has a correlation of -0.73 in the creaky vs. modal contrast, that means that having a *greater* value for HNR05_Mean is correlated with modal voicing; having a *smaller* value for HNR05_Mean is correlated with creaky voicing.

Features that are strongly correlated with a phonation type are likely to be important to the machine learning models, while weakly correlated features are unlikely to prove useful. Given that previous research has not found a single feature that consistently and accurately

[2]Pearson correlations assume a normally distributed data set.

distinguishes phonation types, I expect correlation coefficients to be, overall, low.

5.5.2 Feature Weights

Support Vector Machines report how heavily they rely on different features by assigning each feature a *weight* representing its importance. Recall that sk-learn's `svm.SVC` package handles multi-class problems by running multiple one-versus-one models. That is, given a three-way distinction between breathy, modal, and creaky voice, it would run three separate models: breathy vs. creaky, breathy vs. modal, and creaky vs. modal. For two-way phonation contrasts, this can be handled with just one model (either breathy vs. modal or creaky vs. modal). Feature weights are output for each pairwise comparison, representing how important that feature is in separating those two phonation types. Like correlations, both extremely negative and extremely positive feature weights indicate features that are important to the decision function. There is no industry-standard interpretation of weights, but ± 1 or more is generally considered important.

When multiple features provide very similar information, such as HNR05 and HNR15, caution must be used in interpreting weights. This similarity, called *collinearity*,[3] can impact weights in a few different ways. One possibility is that low weights are assigned to features whose work has already been done by a collinear feature. Another possibility is that redundant features can be assigned weights that cancel each other out; once one feature is found to provide a certain piece of information, other features that provide the same information can be assigned weights of -1 and +1 and together, not impacting the model. My feature set is full of potentially collinear features – many features have several variants in their calculation (e.g., HNR15 vs. HNR25) and are calculated over several time spans (e.g., HNR15_1, HNR15_2, HNR15_3, HNR15_Mean). For each language, I will pare down the set of features used by the SVM to avoid collinearity; the methods for this vary by language and will be discussed in each chapter.

[3]More formally, two features are collinear when one can be predicted from the other.

I am to pare down the set of features to include just one feature from each category.[4] I perform several sanity checks to confirm that each category's feature is, in fact, representative of that category. First, I consider whether or not the sign of the feature's category, indicating its associated with a phonation type, is consistent with previous studies of that feature and phonation; if a weight suggests that increased Jitter is strongly associated with modal voicing compared to creaky voicing, that raises some suspicion. Second, I compare the weights to the correlations; again an inconsistency may indicate that the wrong feature has been chosen. Finally, I check whether that feature weight's sign (positive or negative) matches the sign of the category's average weight. For example, if I picked HNR05_3 and its weight was -0.5, I want to make sure that the mean weight of all HNR features is also negative. I'm essentially trying to make sure I didn't accidentally pick one of the odd and extreme weights caused by multicollinearity. While I expect the signs of the larger weights to match the average, very small weights (indicating unimportant features) are likely to not match.

5.5.3 Feature Importance

Random Forest *importance* is analogous to SVM weights. Importance represents how much a feature contributes to the decision function based on how much the impurity decreases when that feature's node is reached. The sum of all feature weights is 1.0 and higher feature importance indicates greater reliance on the feature; all values are positive. Unlike correlations and SVM feature weights, Random Forest importance is reported overall, not for each class comparison. That is, each feature is assigned a single value representing its overall importance to the model.

5.5.4 Ablation

Ablation is the final way I evaluate each feature's contribution to distinguishing phonation types. Ablation testing involves systematically excluding features from a model and seeing

[4] I include all features from the Surrounding Phones category, as they each provide different information.

how the classifier's performance changes. Intuitively, removing a crucial feature should result in worse model performance. However, this may not be the case when collinear features exist; when one of those collinear features is ablated, another will be able to do the same job and the classifier's performance won't be impacted. I avoid this issue by initially ablating entire categories of features, thereby ablating collinear features at the same time.

I perform two types of ablation. I first ablate one category of features at a time, train and test the classifiers, and then put that category back in the data. This type of ablation looks at how individual categories impact the classifiers. The second type of ablation, which I call *iterative ablation*, also involves ablating one category at a time. I then exclude again the category whose ablation caused the largest drop in weighted F1 score and ablate the remaining categories, one at a time. The category whose ablation caused the largest drop is then also excluded (along with the first category to be excluded), and this process continues until just one category remains.

The following six chapters are each devoted to one language. In these chapters, I report the performance of the two machine learning models and the importance of features for the languages using the metrics described above. Chapter 12 synthesizes this information, comparing how the models performed on the six languages and which features proved most important. In Chapter 13, I return to the English data to address the classification question and build a model that automatically identifies English phonation types as accurately as possible.

Part II

THE LINGUISTIC QUESTION

Chapter 6

ENGLISH

This chapter is the first of six that explore the acoustic properties of different phonation types in specific languages; here, I focus on English. After briefly reviewing some logistics of the English data, I report the performance of the two classifiers. I then examine the role of different features through four lenses – correlations, SVM weights, Random Forest importance, and ablation – and summarize my findings.

The English data set is described in detail in Chapter 4. English uses three types of phonation: breathy, modal, and creaky voice. These three voice qualities are used in several ways, all of which are non-contrastive: sociolinguistically, allophonically, and prosodically. Because of its non-contrastive nature, the English data set is particularly imbalanced between the three phonation classes; the distributions in the data are shown in Table 6.1.

Table 6.1: English Phonation Distribution

Type	Count	*Percent*
B	576	*5.742%*
M	7631	*76.074%*
C	1824	*18.184%*
Total	10031	

Recall that the data have already undergone several pre-processing steps (see Chapter 4 for more detail). Only vowels with primary or secondary stress that are in content words are included. All data have been normalized, by speaker, by vowel, or over the whole data set. All features that have 15% or more undefined measures have been excluded from the data, and any remaining undefined measures will be replaced on fold-by-fold basis as the models

are run.[1] This leaves a total of 80 features from ten categories, as listed in Table 6.2. These are the features used by the SVM and the Random Forest throughout this chapter.

Table 6.2: English Features

Feature Category	Number of Features
Spectral Tilt	28
Harmonics-to-Noise Ratio	16
f_0	12
Surrounding Phones	6
Cepstral Peak Prominence	4
Prosodic Position	4
F1	4
RMS Energy	4
Variance of Pitch Tracks	1
Vowel Duration	1
Jitter	0
Subharmonic-to-Harmonic Ratio	0
Shimmer	0
Total	80

6.1 Model Performance

Table 6.3 reports the accuracy, weighted F1 score, precision, recall, and F1 score[2] for a Support Vector Machine and a Random Forest based on imbalanced and resampled data sets. I review the model performance first based on the imbalanced data set and then on the resampled data set.

[1]Undefined measures will be replaced with the class mean in the training data and the overall mean in the testing data.

[2]See Chapter 5 for a description of these metrics.

Table 6.3: English Classifier Performance

Balance	clf	Accuracy	Weighted F1	Precision			Recall			F1 Score		
				B	M	C	B	M	C	B	M	C
Imbalanced	SVM	83.77	0.81831	0.49	0.86	0.74	0.17	0.96	0.54	0.25	0.91	0.63
	RF	81.408	0.79152	0.45	0.84	0.71	0.28	0.96	0.38	0.35	0.9	0.5
Resampled	SVM	75.207	0.78767	0.21	0.94	0.66	0.71	0.79	0.62	0.32	0.86	0.63
	RF	78.317	0.78895	0.27	0.88	0.64	0.51	0.88	0.47	0.35	0.88	0.54

6.1.1 Imbalanced Data

The SVM and Random Forest trained on the imbalanced data perform reasonably well, with weighted F1 scores of 0.818 and 0.792, respectively. Weighted F1 scores in this range indicate that the classifiers are finding patterns and there is some degree of confidence in those patterns.

Both the imbalanced SVM and Random Forest handle modal vowels quite well. 84 – 86% of the tokens identified as modal are actually modal (precision), and 96% of the tokens that are actually modal are identified as such. Precision and recall are lower for creaky voice. While the SVM's creaky recall is 0.54, the Random Forest's creaky recall is just 0.38; the two models have similar precision (0.74 and 0.71), but the SVM is much more complete than the Random Forest for creaky voice. Finally, both classifiers perform worse for breathy voice than for modal or creaky voice. The SVM and the Random Forest again have similar precision (0.49 and 0.45), but the Random Forest this time has higher recall (0.28) than the SVM (0.17). This overall pattern – the classifiers are best at identifying modal voicing, followed by creaky voicing, followed by breathy voicing – approximately reflects each phonation type's distribution in the data set; the more tokens of a given class the algorithm has seen, the better it's going to do at identifying other members of that class.

6.1.2 Resampled Data

Given the significant class imbalance in the English data, I expect the imbalanced and resampled data set to provide somewhat different results. In the resampled data, additional synthesized instances of the two minority classes are added to provide equal representation from each of the three classes (see Chapter 4). The two classifiers both experience a small drop in weighted F1 score when the data are resampled, down to 0.788 for the SVM and 0.79 for the Random Forest.

As the data are no longer skewed towards modal voice, the pattern seen in the imbalanced data is no longer as clear. But adding in synthetic data does not perfectly solve the problem of imbalance; if the synthetic data are based on a small number of actual instances, they will not necessarily be good representatives of their class.

Both classifiers continue to handle modal voicing the best. Now, their recall is lower than (or equal to) their precision; they're finding fewer modal vowels but they're more accurate about what they do find. Unlike modal voicing, breathy and creaky precision *decrease* with resampling and breathy and creaky recall *increase* with resampling. In other words, resampling leads the classifiers to find more instances of the minority classes but they become a little sloppier.

The four classifiers' weighted F1 scores show that they are successfully finding patterns among the features that distinguish breathy, modal, and creaky voicing. To explore what those features are, I use correlations, weights, importance, and ablation in the following sections. For the three metrics involving machine learning, I use the resampled data, as that gives the classifiers as much information as possible about what the instances of the minority classes look like.

6.2 Correlations

Before looking at how each feature contributes to the SVM and Random Forest classifications, I look at the features though a more traditional statistical lens: correlations. Correlations

describe the strength of the relationship between two variables. In this case, they're describing the strength of the relationship between a feature and a phonation type.

I am specifically interested in the features that best distinguish each combination of phonation types (B vs. C, B vs. M, and C vs. M). In each contrast, a positive value indicates that a larger value for that feature is associated with the alphabetically first phonation type; a negative value indicates that a larger value for that feature is associated with the alphabetically second phonation type. Strong correlations are values from ±0.5 to ±1, medium correlations from ±0.3 to ±0.49, and low correlations ±0.29 to 0.

The ten most strongly correlated features for each of the three contrasts are listed in Table 6.4. Because both very positive and very negative values indicate strong correlations, Table 6.4 reports the ten largest *magnitudes* of correlations. The correlations for all eighty features can be found in Table H in Appendix H. The sections below review the strongest correlations for each of the three contrasts.

Table 6.4: English Top Feature Correlations

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
Feature	Correlation	Feature	Correlation	Feature	Correlation
HNR15_Mean	0.447	CPP_Mean	-0.363	Snack_f_0_Mean	-0.544
Snack_f_0_Mean	0.430	CPP_3	-0.362	VoPT	0.511
HNR25_Mean	0.424	CPP_2	-0.360	Snack_f_0_2	-0.457
HNR15_2	0.411	VoPT	0.274	CPP_3	-0.410
HNR05_Mean	0.410	CPP_1	-0.252	CPP_2	-0.385
Snack_f_0_2	0.400	Snack_f_0_Mean	-0.150	CPP_Mean	-0.384
HNR35_Mean	0.397	RMS_Energy_Mean	-0.147	Snack_f_0_3	-0.360
HNR15_1	0.396	RMS_Energy_2	-0.144	Snack_f_0_1	-0.333
HNR25_2	0.394	HNR35_1	0.137	HNR05_3	-0.303
HNR15_3	0.382	RMS_Energy_3	-0.135	HNR05_Mean	-0.286

6.2.1 Breathy vs. Creaky Correlations

Figure 6.1 shows each feature's correlation with the breathy vs. creaky contrast. Each bar represents an individual feature, color coded by category. The magnitude of the bar

represents that feature's correlation with a phonation type; highly negative values are more strongly associated with creaky voice, and highly positive values with breathy voice. The x-axis spans from -1 to 1, which would mean a perfect correlation with creaky voice and a perfect correlation with breathy voice, respectively. Note that no feature has a particularly strong correlation with either phonation type; the strongest correlation is 0.447, which is at the upper range of medium correlations (± 0.3 – ± 0.49).

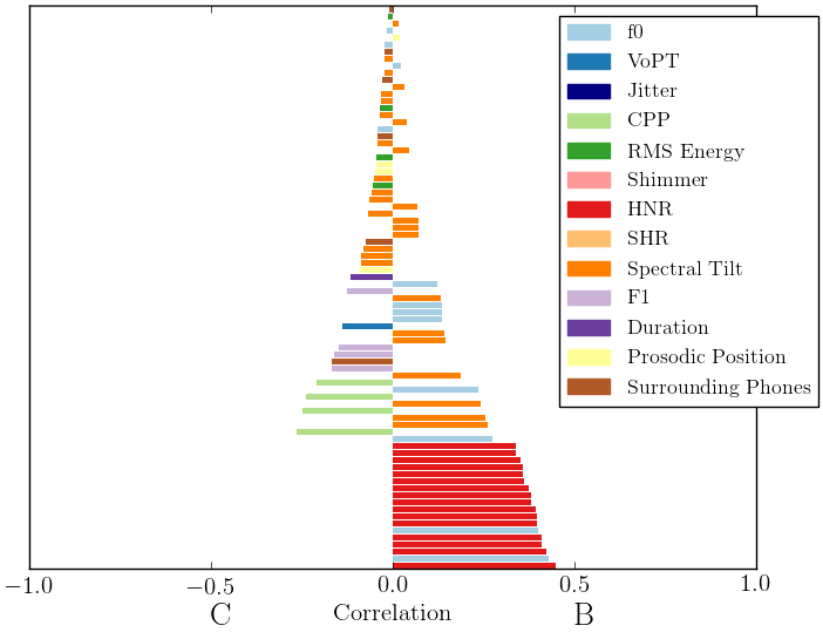


Figure 6.1: English Feature Correlations, B vs. C

HNR (red) accounts for most of the strongest correlated features for this contrast; higher HNR indicates breathy voice. HNR quantifies periodicity by comparing the amplitude of the harmonics to the amplitude of the noise in the signal. The higher the ratio of harmonics to noise, the more periodic the signal; HNR is typically lower in non-modal signals than in modal ones. The fact that HNR is correlated with the breathy vs. creaky contrasts suggests that these two phonation types, despite both being non-modal, exhibit markedly different

degrees of periodicity in English, with breathy voice being more periodic than creaky voice.

Looking back at Table 6.4, we see that eight of the ten features most correlated with the breathy vs. creaky distinction are HNR measures. Four of those ten are HNR15, which is HNR measured from 0 to 1500 Hz; the remaining four are split among HNR05, HNR25, and HNR35. Similarly, four of the eight are calculated over the entire vowel; the remaining four are calculated over each third of the vowel. While HNR appears to be rather important in distinguishing English breathy and creaky vowels, only rough patterns emerge regarding the most informative bandwidths and time periods.

Several f_0 measures (light blue) are sprinkled among the HNR measures; a higher f_0 is also associated with breathy voice. Like HNR, this feature is rather unexpected for the breathy vs. creaky contrast, as both breathy voice and creaky voice are often associated with a lower f_0 than modal voice in English (Hombert, 1978; Garellek and Keating, 2015). The correlation results do not necessarily contradict these findings, but do suggest that creaky voice has a lower f_0 than breathy voice in English.

While the STRAIGHT algorithm is generally thought of as the most robust to aperiodicity, the Snack algorithm accounts for the four most correlated f_0 measures. However, Snack being the most strongly correlated feature does not necessarily mean that it is *correctly* measuring f_0 – it's possible that the algorithm fails in different ways during breathy and creaky voiced segments, exaggerating or even creating this statistical relationship.

Cepstral Peak Prominence (light green) is the feature most strongly correlated with creaky voice in this contrast, though the correlation for all four CPP measures is weak. CPP is expected to be lower in non-modal segments than in modal ones; its association with creaky voice suggests that CPP is higher in creaky vowels than in breathy ones, meaning that creaky vowels are more periodic than breathy ones. This contradicts my (hesitant) interpretation of HNR, though the correlations are much stronger for HNR than for CPP.

The remaining correlations are too low to provide much information, but I'd like to point out a few trends. First, the correlations suggest that F1 (light purple) is higher for creaky vowels than for breathy vowels. In several languages, non-modal vowels have a higher F1

than their modal counterparts (see Chapter 3); these data suggest that there may also be a difference in F1 between breathy and creaky voice. Second, higher Variance of Pitch Tracks (bright blue) is weakly correlated with creaky voice. This is consistent with my findings in Chapter 3, which compared VoPT for the three English voice qualities on a subset of the data. In that subset, creaky vowels were had a slightly higher VoPT than breathy vowels. The fact that this small difference was reflected in correlations gives me hope that VoPT will prove particularly useful in contrasts involving modal voice. Finally, I turn to Spectral Tilt (dark orange). Spectral Tilt measures fall on both sides of the dividing line – some are negative, and some are positive. These seem to contradict each other; how could one increased Spectral Tilt measure be correlated with breathy voice, and another with creaky voice? I believe the answer is that they're not. Most of the Spectral Tilt measures have very low correlations and are likely uninformative features in distinguishing breathy voice from creaky voice. Examining Spectral Tilt's role in this contrast using other metrics – SVM weights and Random Forest importance – may help shed light on this.

6.2.2 Breathy vs. Modal Correlations

Figure 6.2 shows each feature's correlation with breathy vs. modal voicing; features strongly correlated with breathy voice are positive and features strongly correlated with modal voice are negative. Again, none of the correlations are strong, and they're in fact a bit weaker overall than in the breathy vs. creaky contrast.

Cepstral Peak Prominence (light green) accounts for the top three strongest correlations for the breathy vs. modal contrast. As expected, a higher CPP is correlated with modal voice; modal signals have a better defined f_0 peak than breathy signals because of their relatively low amounts of noise, leading to a more pronounced peak in the cepstrum. CPP calculated over the mean as well as the middle and final thirds have extremely similar correlations, with CPP_1 somewhat lower. However, even the three strongest correlated CPP measures still only provide medium correlations.

Variance of Pitch Tracks (bright blue) is the fourth most strongly correlated feature for

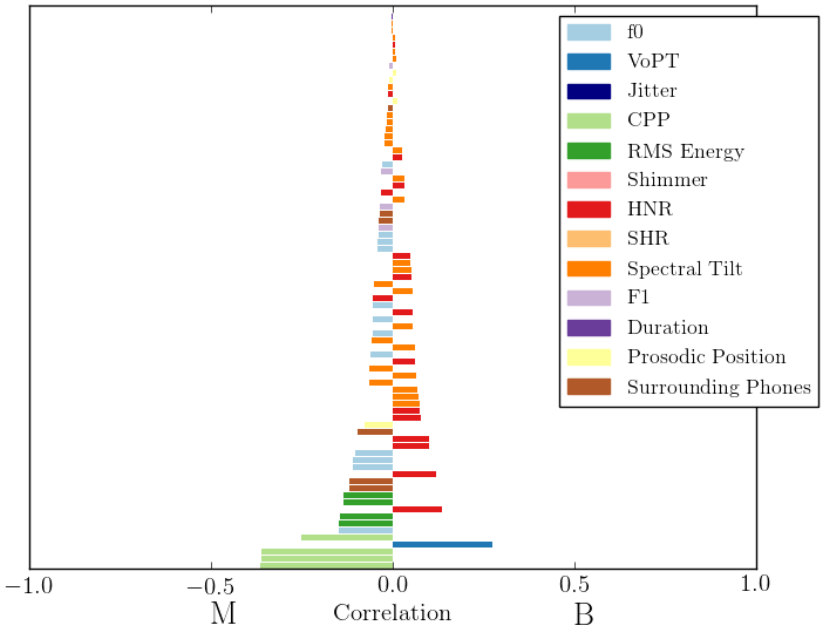


Figure 6.2: English Feature Correlations, B vs. M

this contrast; a higher VoPT is associated with breathy voicing. VoPT indirectly measures voice quality by measuring pitch tracking errors, which are more common in non-modal segments than in modal ones. However, even in fourth place, VoPT's correlation is weak.

The remaining correlations are too weak to warrant much speculation. Before moving on to the final contrast, it's worth noting that HNR has a positive value in this contrast, suggesting that an increased Harmonics-to-Noise ratio is indicative of breathy voice. This is counterintuitive, and given that the correlation is so weak, likely not actually meaningful.

Only three of the top ten features for this contrast have even moderate correlations, suggesting that the difference between English breathy and modal voicing is not well captured by these features. This could be for a variety of reasons: the difference between the two voice qualities could be less significant than the differences between other pairs, the production of one or both voice qualities could vary between speakers, or the differences could simply be captured better by features that are not included in this study.

6.2.3 Creaky vs. Modal Correlations

Figure 6.3 shows the magnitude of each feature's correlation with English creaky voice (positive values) and modal voice (negative values). The features most correlated with the creaky vs. modal contrast in English are stronger than for the other two contrasts, with two strongly correlated features.

Several f_0 measures (light blue) are among the top correlations. Creaky voicing is typically produced with a lower f_0 than modal voicing (Garellek and Keating, 2015), and these correlations support this observation.

Variance of Pitch Tracks (bright blue) is just strong enough to be considered strongly correlated with creaky voice. In my initial testing of VoPT, I found that the difference in VoPT was greatest between creaky and modal segments; this is reflected in the correlations in that the strongest correlation for VoPT is in the creaky vs. modal contrast.

CPP (light green) and HNR (red) are also correlated with modal voicing; both measure the amount of noise in the signal relative to the harmonic structure and are expected to be

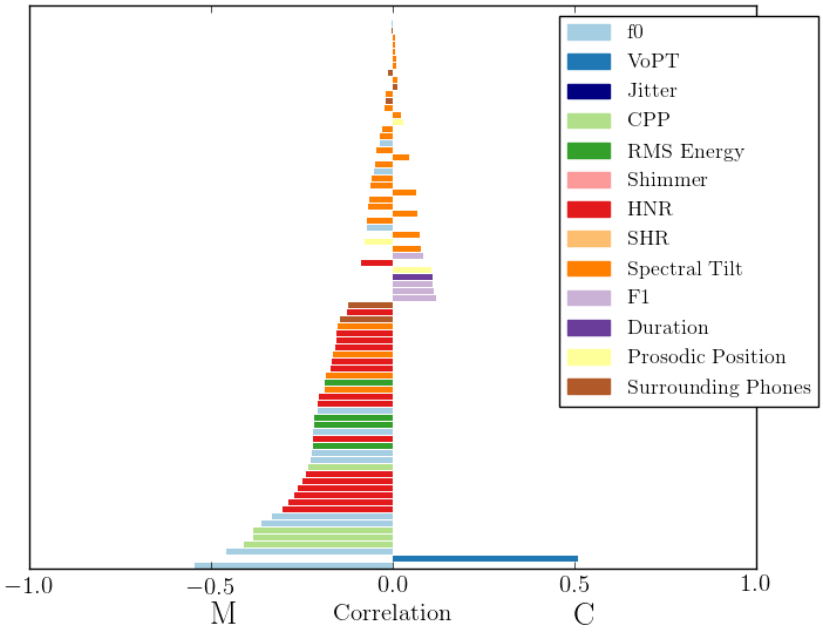


Figure 6.3: English Feature Correlations, C vs. M

higher in modal voicing than in non-modal voicing. CPP has a stronger correlation when measured towards the end of the vowel, as does HNR.

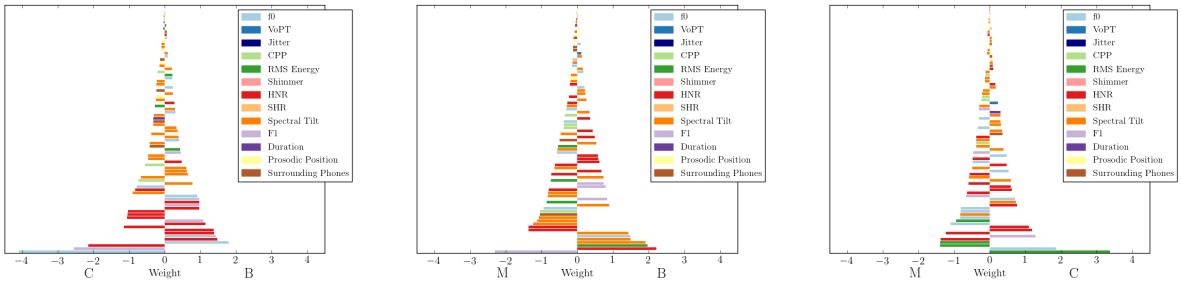
The correlations presented in this section paint a picture of how the features are statistically related to the phonation contrasts. A sort of short list of potentially important features emerges from these correlations: HNR, CPP, f_0, and VoPT; these features will likely also be important to the two machine learning models. Next, I turn to the feature weights output by the Support Vector Machine.

6.3 SVM Weights

I now look at the first of the two classifiers: the Support Vector Machine. The SVM ranks each feature's contribution to the classification process in the form of *weights*. Like correlations, weights can be positive or negative and larger values in either direction indicate that a feature is important. Weights are also calculated for each contrast, representing the feature's importance in distinguishing between two phonation types at a time. A larger magnitude weight essentially means that an increased value for that feature is associated with one of the two phonation types – a more strongly positive weight with one type and a more strongly negative weight with the other.

Figure 6.4 shows the feature weights for English data. They follow the same format as the correlations; strongly positive features are correlated with one phonation type, and strongly negative features with another. The figures are intentionally too small to provide much detail because I'd first like to point out a general pattern – in all three figures, features of the same category (the same color bars) have weights of opposite signs. For example, increased HNR (red) in Figure 6.4(a) appears to signal both breathy voice (positive) and creaky voice (negative).

For contradictions in the correlations, I dismiss them as being too small to be significant. The contradictory features in the weights, on the other hand, are among the top weighted features, many of which were also among the top correlated features. These contradictions are instead caused by collinear features, or features that provide the model with the same



(a) B vs. C (b) B vs. M (c) C vs. M

Figure 6.4: English SVM Weights

information. This can result in collinear features being assigned weights that cancel each other out, like +1 and -1. As it's difficult to interpret SVM weights in the face of multicollinearity, I try to eliminate collinear features by selecting just one from each category.

How to pick that one feature from each category quickly becomes a very messy process. There are many questions – which time span is best, which variation on a measurement is best, which of the three phonation contrasts to use in picking the 'best' – and no clear answers to those questions. So I tried the simplest solution I could come up with: training the model using the single feature from each category that had the largest magnitude weight in the SVM that included all features. I evaluated how representative this subset of features is in three ways. First, I looked at the performance of the model. Its weighted F1 score is 0.77153, down from 0.78767 when all features are included. A drop is expected, as I'm providing the model with less information, and this drop is relatively small considering that I've reduced the number of features from eighty to fifteen. The second way I evaluated this subset is by looking at the feature weights. I considered whether they make sense based on general wisdom about phonation, and whether they correspond relatively well to the correlations; both looked good. Finally, I check whether the feature representing each category has the same sign as the mean of the category; matching signs shows that that feature is *not* one of the features assigned a counterintuitive weight due to multicollinearity. With the exception

of features with extremely small weights (which are likely not meaningful anyway), the signs of these individual features match the mean sign of their category. Given the relatively small drop in weighted F1 score and reasonable results, I will use this subset of features to examine SVM weights.

Before continuing, however, I'd like to point out some of the flaws in this technique. No matter how I opt to reduce collinearity, I inevitably lose data. In losing data, I potentially lose interesting interactions between features. Additionally, there are countless ways of selecting the subset of features to include. I do not claim that my subset of features is the best subset, but simply that it is a reasonable one that will allow me to accurately investigate the role of these features in distinguishing English's three phonation types.

The subset I use here includes fifteen features, chosen because they have the largest magnitude weight of the category for across the three phonation contrasts.[3] These features and their weights are listed in Table 6.5, sorted by magnitude for each contrast. In the sections below, I discuss the results for each of those contrasts.

6.3.1 Breathy vs. Creaky Weights

Figure 6.5 shows the feature weights for English's breathy vs. creaky contrast. Positive weights indicate features that, when their value is larger, are associated with breathy voicing, and negative weights with creaky voicing.

HNR has the largest magnitude weight and is positive, which indicates that an increased HNR is associated with breathy voice rather than creaky voice. This matches the results of the correlations, which also identified HNR as being higher in breathy vowels than in creaky vowels.

CPP has the next largest weight, and its negative weight means that increased CPP is associated with creaky voicing. This is again consistent with the correlations. I pointed out in the previous section that HNR and CPP both measure aperiodicity by way of noise. It's

[3]Note that I include all six surrounding phones features. They are independent of each other and will therefore not be collinear.

Table 6.5: English Feature Weights

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
HNR15_Mean	1.982	CPP_3	-1.477	HNR15_Mean	-0.845
CPP_3	-1.104	Presence_Pre	-1.087	VoPT	0.83
Presence_Pre	-0.887	Voicing_Fol	-0.395	CPP_3	-0.713
VoPT	-0.721	HNR15_Mean	0.394	STRAIGHT_f_0_Mean	-0.613
Vowel_Duration	-0.622	Manner_Fol	-0.37	RMS_Energy_Mean	-0.479
Voicing_Fol	-0.516	STRAIGHT_f_0_Mean	-0.342	Presence_Fol	0.442
RMS_Energy_Mean	0.486	RMS_Energy_Mean	-0.218	Vowel_Duration	0.391
F1_Mean	0.407	Vowel_Duration	-0.152	Manner_Pre	-0.187
Manner_Fol	-0.266	Presence_Fol	-0.134	Voicing_Fol	-0.1
Dist_Word_(%)	-0.2	Dist_Word_(%)	-0.118	2k* – 5k_Mean	-0.069
Voicing_Pre	-0.187	VoPT	0.111	Manner_Fol	-0.066
STRAIGHT_f_0_Mean	0.165	2k* – 5k_Mean	-0.057	Presence_Pre	0.05
Presence_Fol	-0.155	Manner_Pre	-0.056	F1_Mean	0.041
Manner_Pre	-0.071	F1_Mean	0.04	Voicing_Pre	0.006
2k* – 5k_Mean	-0.046	Voicing_Pre	-0.029	Dist_Word_(%)	-0.005

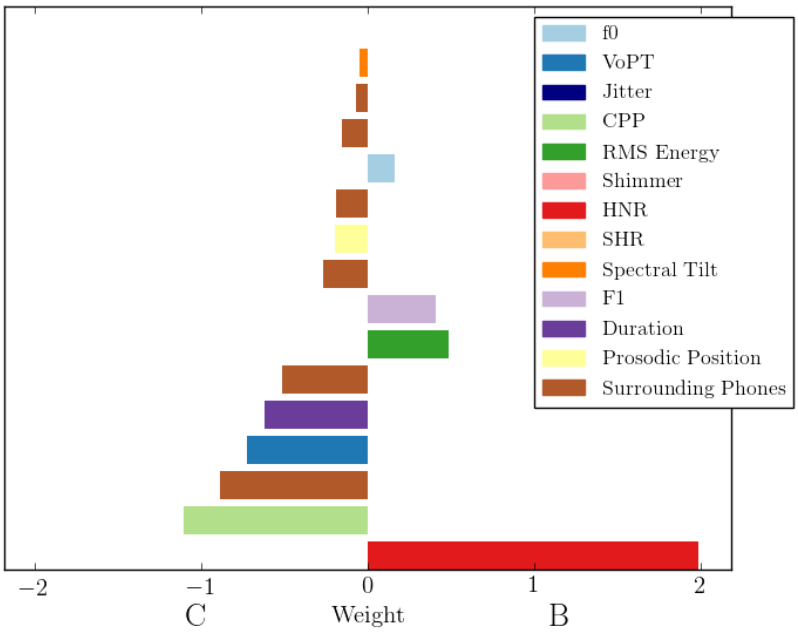


Figure 6.5: English Feature Weights, B vs. C

therefore somewhat unexpected that they would have opposite signs, and also unexpected that they would be among the strongest correlates with the breathy vs. creaky contrast, as both non-modal phonation types are characterized by noise. However, CPP and HNR measure noise in different ways.

Next, we see the first feature from the Surrounding Phones category make an appearance – the *absence* of a preceding phone is associated with creaky voicing. English speakers often glottalize word-initial vowels, particularly at the beginning of intonational phrases (Dilley et al., 1996). Though its weight is relatively high, this feature has a very low correlation in this contrast (-0.167).

6.3.2 Breathy vs. Modal Weights

The feature weights for the breathy vs. modal contrast are shown in Figure 6.6. They are, overall, quite a bit lower than the weights for the breathy vs. creaky contrast. This is the same pattern observed in the correlations, providing further support for my observation that these features do not distinguish breathy voice from modal voice as well as they do other combinations of voice qualities.

The strongest weight is CPP; increased CPP is associated with modal voicing rather than breathy voicing. This is expected, as the cepstrum's peak is more prominent when there is less noise in the signal.

Not far behind CPP is the presence of a preceding phone, which is associated with breathy voicing. English breathy voicing can be caused by an adjacent /h/ phone, but no other phones are typically associated with breathy voicing. It seems somewhat implausible that /h/ alone resulted in such a high weight for this feature. The Surrounding Phones features could get a boost from the absence of (statistical) noise; all other features are likely to suffer from some noise in their measurements, but these features do not run that risk. However, recall that Surrounding Phones features have very low correlations. This may be because these features are categorical – their value is always `0` or `1`. The correlations may have failed to capture a relationship because that is not the ideal statistical tool for describing the relationship

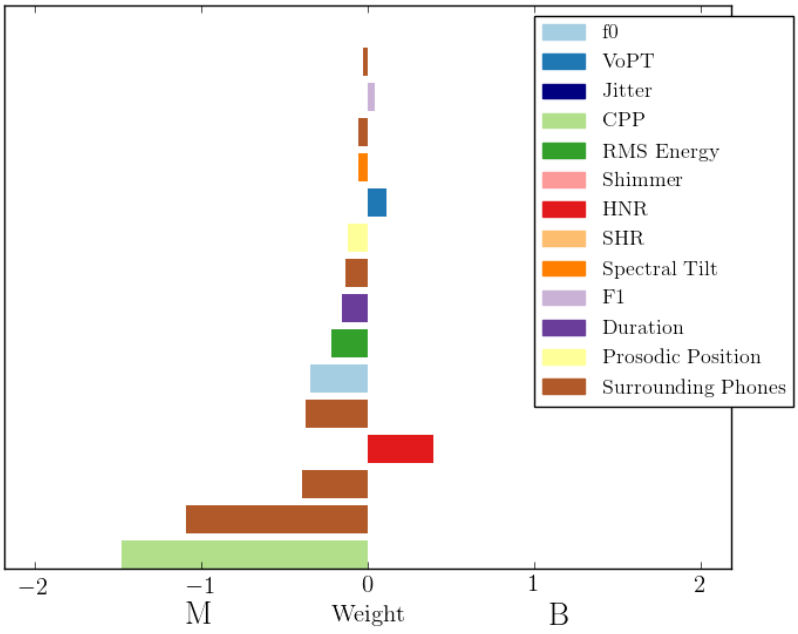


Figure 6.6: English Feature Weights, B vs. M

between a categorical feature and a categorical class.

6.3.3 Creaky vs. Modal Weights

Finally, Figure 6.7 shows the weights for the creaky vs. modal contrast. The largest weights are smaller than any of the other two contrasts, which does not match the pattern seen among the correlations. That said, the features with the largest weights and the features with the strongest correlations are fairly consistent for this contrast.

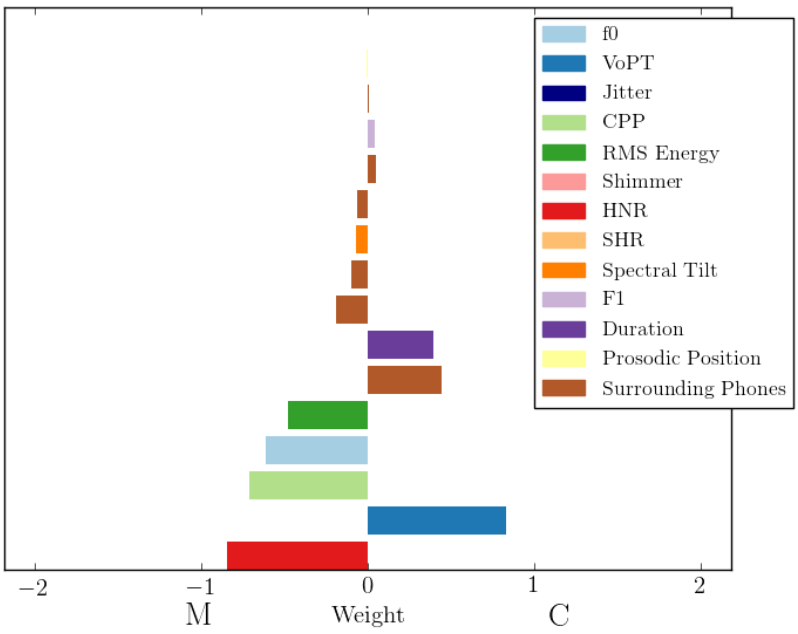


Figure 6.7: English Feature Weights, C vs. M

HNR has the largest weight and is, as expected, associated with modal voicing. VoPT has the second largest weight for this contrast, as well as the second strongest correlation. The other two categories of features ranked above HNR in the correlations are the next two most heavily weighted features – CPP and f_0, both associated with modal voicing.

Overall, the correlations and weights paint a similar picture of which features best

distinguish English phonation types. The same shortlist of features that I identified based on the correlations comes up again in the SVM weights – HNR, CPP, VoPT, and f_0 – with the addition of Presence of the Preceding Phone. Next, I explore each feature's role in the other classifier, the Random Forest.

6.4 Random Forest Importance

Information about each feature's usefulness in the Random Forest's classification process is quantified in feature *importance*. Unlike correlations and weights, feature importance is reported overall, not broken down by phonation contrast. Additionally, all importance values are positive (they sum to 1.0) and do not indicate a relationship with a specific class; higher importance simply means that a feature contributed more to the model's overall performance. (Luckily, we can infer a feature's relationship with each class based on the correlation and SVM weights). The ten most important features are listed in Table 6.6 (the full set is available in Table H of Appendix H) and the importance of all eighty features is plotted in Figure 6.8.

Table 6.6: English Top Feature Importance

Feature	Importance
Snack_f_0_2	0.107639
CPP_Mean	0.101649
Snack_f_0_3	0.047351
CPP_3	0.043654
Snack_f_0_Mean	0.042618
HNR05_Mean	0.035638
Snack_f_0_1	0.033504
RMS_Energy_2	0.024234
VoPT	0.023703
RMS_Energy_3	0.023091

Two features are doing far more work than the others: Snack f_0 in the vowel's middle third and CPP calculated over the entire vowel. The importance of these features is 0.108 and 0.102 respectively. The third most important feature, CPP_3, has an importance of 0.047,

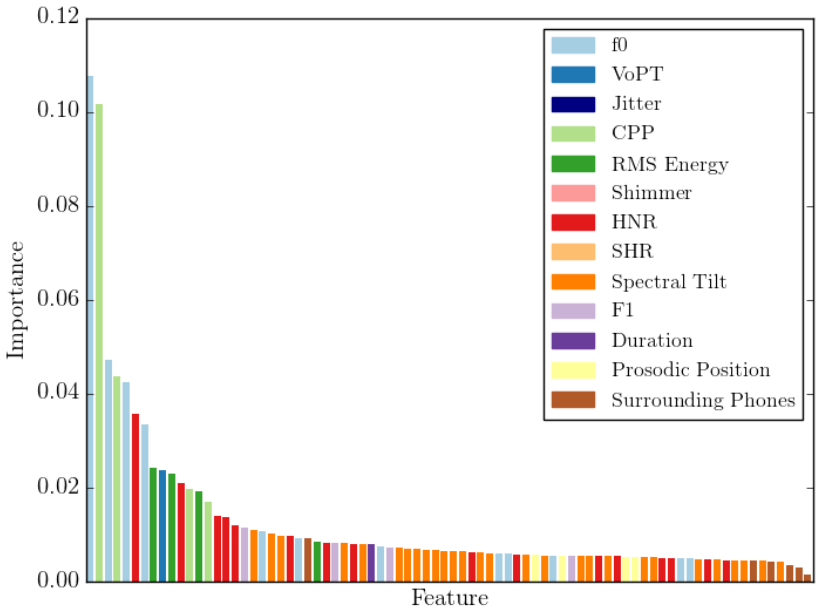


Figure 6.8: English Feature Importance

less than half the value of the top two. This has the intuitive interpretation that Snack_f_0_2 and CPP_Mean are each over twice as important to the Random Forest as CPP_3.

The most important features are dominated by CPP and f_0 measures. HNR, RMS Energy, and VoPT are also among the top ten, but with a significant drop in importance from the features at the top of the list. HNR and VoPT features have also been identified as important in distinguishing English phonation by the correlations and SVM weights; larger HNR is associated with *not* being creaky (i.e., higher in both breathy and modal voicing when compared to creaky voicing) and larger VoPT with non-modal phonation, particularly creaky voicing. RMS Energy has not yet been discussed. In the correlations and weights, RMSE features are consistently, though not particularly strongly, associated with modal voicing over either type of non-modal voicing.

After these five categories of features – f_0, CPP, HNR, RMS Energy, and VoPT – the importance values drop to very small numbers, starting at 0.012. This makes for another similar short list of important categories: f_0, CPP, HNR, RMS Energy, and VoPT.

6.5 Ablation

A fourth and final way I evaluate a feature's contribution to a classifier's decision making process is *ablation*. Ablation involves systematically removing features from the model and observing how this impacts the classifier's performance.

As a first pass at ablation, I ablate entire feature categories; I re-run the SVM and Random Forest, each time excluding all the HNR features, then replacing them and excluding all the Spectral Tilt features, and so on. As many of the features within a given category are collinear, ablation testing of individual features may overlook their contribution; when one feature is ablated, a similar feature can pick up the slack and result in little or no change to the model's performance. Ablating entire categories instead provides a better overview of the types of features that are important to the model.

The features used in the English classifiers come from ten categories. Table 6.7 lists how a resampled SVM and Random Forest perform when the features from each category are

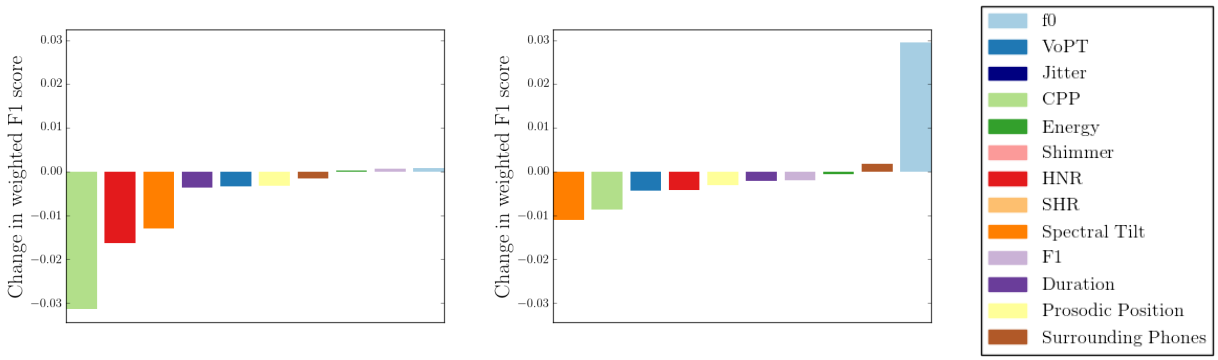
excluded; I report the model's weighted F1 score, as well as the change in weighted F1 score from the model containing all the features. The same information is plotted in Figure 6.9.

Table 6.7: English Category Ablation

Feature	SVM		Random Forest	
	Weighted F1	Change	Weighted F1	Change
Spectral Tilt	0.77464	-0.01303	0.77791	-0.01104
CPP	0.75636	-0.03131	0.78035	-0.0086
RMS Energy	0.78787	0.0002	0.78843	-0.00052
HNR	0.77138	-0.01629	0.78474	-0.00421
f_0	0.78854	0.00087	0.81844	0.02949
F1	0.78839	0.00072	0.78704	-0.00191
Vowel Duration	0.784	-0.00367	0.78683	-0.00212
Prosodic Position	0.78439	-0.00328	0.78591	-0.00304
VoPT	0.78431	-0.00336	0.78463	-0.00432
Surrounding Phones	0.78618	-0.00149	0.79078	0.00183

No single category's exclusion causes much of a drop in weighted F1 for either classifier. For the SVM, ablating all CPP features causes the largest drop in weighted F1 score. Ablating HNR leads to the next largest drop, followed by Spectral Tilt. CPP and HNR are both familiar measures by now, as they've been identified as important by the various other metrics. Spectral Tilt, however, has been noticeably absent so far, despite being generally considered to be the most reliable measure of phonation types in other languages. Its presence here does not necessarily mean that it's a good measure of English phonation, as ablating it does very little to impact the model. Spectral Tilt's ablation also causes the largest drop for the Random Forest, but even the largest drop is quite small. Ablating f_0 from the Random Forest causes an *increase* in weighted F1. (Random Forests inherently involve more randomness than SVMs, so an increase in performance does not necessarily mean that ablating f_0 actually helped the model.)

In the above ablation, I replace each category after removing it. Another form of ablation testing involves iteratively excluding feature categories. Iterative ablation begins with normal ablation (removing and then replacing) to determine which feature category causes the largest



(a) SVM (b) Random Forest

Figure 6.9: English Category Ablation

drop. This category is then excluded, and another round of normal ablation determines which category to exclude next. The results of this iterative category ablation for the English data are reported in Table 6.8.

Table 6.8: English Iterative Category Ablation (SVM)

Category	Weighted F1
CPP	0.75636
Spectral Tilt	0.73805
VoPT	0.71747
f_0	0.6903
HNR	0.61715
RMS Energy	0.56825
Surrounding Phones	0.35959
Prosodic Position	0.31459
Vowel Duration	–

As seen in the categorical ablation, ablating CPP causes the largest drop in weighted F1 for the SVM. With CPP excluded, the next largest drop comes from Spectral Tilt. In the non-iterative ablation, Spectral Tilt came in third, after HNR. Spectral Tilt appears

to do more when CPP isn't in the picture; perhaps an interaction between HNR and CPP disappears. After that, we see some of the usual features – VoPT, f_0, and HNR, as well as RMS Energy.

These two types of ablation – ablating categories one at a time and iteratively – suggest that CPP, HNR, and Spectral Tilt are the most important feature categories for English phonation classification.

6.6 Summary

The above sections have reported the features that best distinguish English breathy, modal, and creaky voice according to four consideration – correlations, SVM weights, Random Forest importance, and ablation. While there is no neat way of synthesizing this information into one number for each feature, several trends have emerged. Cepstral Peak Prominence and Harmonics-to-Noise Ratio are among the top features for all four metrics. f_0 and Variance of Pitch Tracks are also fairly consistently important, but less so than CPP and HNR. RMS Energy and Surrounding Phones both come up more than once, but appear to do even less than f_0 and VoPT. These six features are reviewed below.

Cepstral Peak Prominence is a top-performing feature across the four metrics. Increased CPP is associated with modal voicing. It's moderately correlated with modal voicing compared to both breathy and creaky voicing and among the most significant weights for both contrasts. In both the correlations and weights, CPP is higher in creaky voice than breathy voice, though this relationship is less significant than in contrasts involving modal voice; this suggests that English creaky voice is more periodic than breathy voice. Two CPP features are the second and fourth most important features to the Random Forest and ablating CPP causes the largest drop in weighted F1 score for the SVM. The two CPP features that appear to be the most important are CPP_Mean and CPP_3.

CPP has been shown to vary by phonation type in many other languages, including Gujarati (Khan, 2012), Hmong, Mazatec, and Yi (Keating et al., 2011). Few studies have set out to describe the acoustic properties of English non-modal phonation, so I was unable

to find an analogous study describing CPP's role in English phonation. That said, CPP has been used in a few studies of American English. Podesva et al. (2015) examines creaky voice in male and female Californians according to various acoustic measures, including CPP, which they used to measure periodicity. They found that both genders get less periodic (lower CPP) as a phrase progresses. A closer look using Spectral Tilt revealed that that aperiodicity is different: men get breathier while women get creakier. Garellek and Seyfarth (2016) found that CPP can distinguish between phrase-final creak and /t/ glottalization. This points to a potential source of noise in my data: I did not distinguish between these two types of creaky voice. If they differ in CPP, putting them in the same category will make the classifier's job harder. Nonetheless, the two classifiers do rely significantly on Cepstral Peak Prominence to distinguish English phonation types.

Harmonics-to-Noise Ratio, another way to describe the amount of noise in the signal, is also important according to all four metrics; larger HNR indicates less noise relative to harmonics and is associated with modal voicing. In the correlations, HNR is useful for distinguishing modal voice from creaky voice, but not for distinguishing modal voice from breathy voice. However, it dominates the top correlations for the breathy vs. creaky contrast, suggesting that English breathy voice is more periodic than creaky voice. This same trend appears in the SVM weights. For the Random Forest, HNR05_Mean is the sixth most important feature, though ablating all HNR features barely impacts the Random Forest's performance. Ablating the category from the SVM causes the second largest drop in weighted F1 score and it's ablated fifth during iterative ablation. The HNR measurements that come up most are HNR05 and HNR15, suggesting that the noise of non-modal phonation is concentrated in the lower frequencies. Many of the features are measured over the vowel's middle third or the entire vowel, though this pattern is a bit less clear. Overall, the pattern that emerges for HNR is that low HNR is associated with creaky voicing; modal voicing appears to have the highest HNR, followed by breathy voicing and then by creaky voicing. Like CPP, finding previous studies of HNR in English phonation types is challenging. The two studies that used CPP as a measure (Podesva et al., 2015; Garellek and Seyfarth, 2016)

also looked at HNR, but do not report the results of those measures.[4]

***f*$_0$** is identified as somewhat important in distinguishing English phonation, particularly according to correlations and importance. Larger f_0 is correlated with modal voice; the correlation is stronger when compared with creaky voice and weaker with breathy voice. In the breathy vs. creaky contrast, it's moderately correlated with breathy voicing. The strongest correlated f_0 feature for all three contrasts is Snack_f_0_Mean. This same feature is fifth most important for the Random Forest; Snack's f_0 algorithm calculated over the middle and final thirds of the vowel are the first and third most important features, respectively. f_0 has the fourth largest SVM weight in the creaky vs. modal contrast, but is otherwise unimportant, and it plays a small role in ablation. These results suggest that f_0 is generally highest in modal voicing and lowest in creaky voicing, with breathy voicing between those two, and that the differences are best captured by Snack's algorithm over the entire vowel. While this pattern makes sense, I'm surprised that it's so strong; f_0 is essentially a side effect of the different voice qualities, and for this reason is often excluded from consideration in studies of voice quality. These results suggest that it plays a more important role in our perception of phonation. Lower f_0 has been previous associated with English creaky voicing. Podesva et al. (2015) and Garellek and Seyfarth (2016) both found that creaky voice is more prevalent in segments with a lower f_0. However, I don't believe that f_0 has been previously found to differ between English breathy vowels and creaky vowels.

Variance of Pitch Tracks, the new measure that exploits pitch tracking errors to identify periods of non-modal phonation, appears several times in the English metrics. It's weakly correlated with breathy voicing in the breathy vs. modal contrast (though it's the fourth most strongly correlated feature for that contrast) and strongly correlated with creaky voicing in the creaky vs. modal contrast. VoPT has the second largest weight in the creaky vs. modal contrast, the fourth in the breathy vs. creaky contrast and the 11th in the breathy vs. modal contrast. It's not particularly impressive in the Random Forest importance

[4]I believe the lack of discussion in Garellek and Seyfarth (2016) suggests that HNR measures were not valuable to their analysis.

rankings, in ninth place with a rather low importance. Ablating it does little to either classifier's accuracy, though it's the third category to go in the iterative SVM ablation.

Two categories of features have a weaker track record but are nonetheless worth discussing, as they've come up in at least one of the four metrics. **RMS Energy** measures the intensity of the signal, which is usually higher in modal signals than in non-modal signals. The correlations reflect this, and while RMS Energy features are generally weakly correlated with phonation type, they're among the top correlations for the breathy vs. modal contrast. The SVM weights paint a different picture: they show that RMS Energy is useful in the creaky vs. modal contrast, but not in the breathy vs. modal contrast. RMS Energy features take eighth and tenth place in Random Forest importance, but ablating the entire category does not greatly impact either classifier. **Surrounding Phones** are also spottily important. All six features in this category have weak correlations, low importance, and barely impact accuracy when ablated. However, they show up towards the top of the list in the SVM weights. Two features in particular have fairly high weights: presence of the preceding phone and voicing of the following phone. The presence of a preceding phone and a voiceless following phone are associated with breathy voicing. This suggests that breathy voicing tends to occur when the vowel is not utterance-initial and when it's followed by a voiceless phone in English.

The six categories of features summarized here – CPP, f_0, HNR, VoPT, RMSE, and Surrounding Phones – seem to contribute different amounts of discriminatory power to the classification task. CPP and HNR appear to do the most work, followed by f_0 and VoPT, and then by RMSE and Surrounding Phones. It's difficult to synthesize all of the information from the four metrics, so my final look at these categories of features is to see how the SVM and Random Forest perform when those categories of features are the only ones included. Table 6.9 shows the accuracy of a resampled SVM and Random Forest when only those categories of features are included.

Table 6.9 roughly confirms my impression of these proposed three groups of feature categories. Using categories from these six features causes just a small drop in the weighted

Table 6.9: English Accuracy Using Subsets of Features

Feature Categories	Weighted F1	
	SVM	Random Forest
Baseline (all features)	*0.78767*	*0.78895*
CPP, HNR, f_0, VoPT, RMSE, Surrounding Phones	0.77007	0.78383
CPP, HNR	0.70294	0.74513
f_0, VoPT	0.7292	0.7357
RMSE, Surrounding Phones	0.60767	0.68221

F1 score of both classifiers. Using just CPP and HNR features causes a larger drop (larger for the SVM than for the Random Forest). Using just f_0 and HNR results in a very slightly lower weighted F1 score for the Random Forest but a *higher* weighted F1 score for the SVM. Using just RMS Energy and Surrounding Phones results in lower scores for both classifiers. While the exact ranking of f_0 and VoPT is still a bit unclear, these six feature categories do appear to the most useful in distinguishing English phonation types.

Taken as a whole, these results suggest that noise is the most important cue to identify English voice qualities. Pitch, as well as pitch tracking errors signaling a periodicity, are both useful, and intensity and phonetic environment provide useful information as well. Recall that because English phonation is not predictable, all tokens are labeled with their phonation types by human listeners. This means that the results speak to what listeners *perceive* as different phonation types, rather than how people *produce* something that is nominally a given voice quality.

The next five chapters provide similar discussions of the results for the other languages. In Chapter 13, I return to English with the goal of fine-tuning a classifier for English's three voice qualities.

Chapter 7

GUJARATI

This chapter presents and discusses the models trained on Gujarati data. I first review the performance of the classifiers and then explore which features best distinguish Gujarati phonation types.

Gujarati phonation is phonemic; breathy and modal voicing are contrastive on vowels. This means that phonation type can be identified by word meaning, without a human annotator. While this removes the risk of bias and human error, it adds the risk of incorrectly assuming voice quality; we know each token's nominal voice quality, but without listening, we don't know how it was *actually* produced. Though the original Gujarati corpus contains breathy consonants as well as vowels, I exclude all words with such phones to avoid an interaction between them and adjacent vowels. The distribution of phonation types in the Gujarati data, described in more detail in Chapter 4, is listed in Table 7.1.

Table 7.1: Gujarati Phonation Distribution

Type	Count	*Percent*
B	1262	*42.449%*
M	1711	*57.551%*
C	0	*0.0%*
Total	2973	

The Gujarati data have been normalized (either by speaker, by vowel, or overall – see Chapter 4 for more detail), all features with 15% or more undefined measures have been excluded, and the remaining undefined measures will be replaced on a fold-by-fold basis with the class mean in the training data and the overall mean in the testing data. A total

of 101 features from ten categories will be used for the Gujarati models; Table 7.2 lists the number of features per category.

Table 7.2: Gujarati Features

Feature Category	Number of Features
Spectral Tilt	28
Harmonics-to-Noise Ratio	16
f_0	16
Surrounding Phones	0
Cepstral Peak Prominence	4
Prosodic Position	0
F1	4
RMS Energy	4
Variance of Pitch Tracks	1
Vowel Duration	1
Jitter	13
Subharmonic-to-Harmonic Ratio	0
Shimmer	14
Total	101

The 101 features will be used to train four classifiers on Gujarati data: two Support Vector Machines and two Random Forests, with one of each trained on imbalanced data and the other on resampled data.

7.1 Model Performance

I begin by reporting the accuracy, weighted F1 score, precision, recall, and F-score for an SVM and a Random Forest, each trained once on an imbalanced data set and once on a resampled data set, in Table 7.3. As seen in Table 7.1, the Gujarati data is not hugely skewed but does contain more modal tokens than breathy tokens.

Overall, the four classifiers perform similarly. Breathy voice and modal voice are not represented evenly in the imbalanced data set, but the skew is not massive, so resampling does not cause a large change. That said, there are more modal vowels than breathy vowels

Table 7.3: Gujarati SVM and RF Performance

Balance	clf	Accuracy	Weighted F1	Precision		Recall		F-Score	
				B	M	B	M	B	M
Imbalanced	SVM	73.831	0.73065	0.76	0.73	0.57	0.86	0.65	0.79
	RF	72.519	0.72572	0.67	0.77	0.69	0.75	0.68	0.76
Resampled	SVM	72.149	0.72079	0.68	0.75	0.65	0.77	0.67	0.76
	RF	72.587	0.72694	0.66	0.78	0.72	0.73	0.69	0.75

in the data set, which may contribute to precision and recall being slightly higher for modal vowels than for breathy vowels in most of the models. Weighted F1 scores range from 0.72 to 0.73. While this is not particularly impressive (especially considering that the Gujarati classifiers have the easier task of distinguishing between two classes rather than three), it's solid enough to show that the classifiers are finding patterns in the data. In the following sections, I use correlations, weights, importance, and ablation to examine what those patterns are, and then synthesize this information and compare it to previous studies of Gujarati phonation.

7.2 Correlations

My first pass at examining the relationship between Gujarati features and its two phonation types is correlations. Unlike English, Gujarati has just two voice qualities, meaning that there is only one possible contrast. Features with positive correlations have larger values in breathy voicing and features with negative correlations have larger values in modal voicing; the larger the value, the stronger the relationship between that feature and that phonation type. Table 7.4 reports the ten most strongly correlated features for Gujarati, and the full set of correlations for all 101 features can be found in Appendix I. Figure 7.1 plots the full set of correlations.

The strongest Gujarati correlation is 0.333, which is barely a medium correlation; the next strongest correlation, -0.254, is already considered to be a low correlation. The fact that individual features are not strongly correlated with Gujarati voice qualities, yet the

Table 7.4: Gujarati Top Feature Correlations

Breathy vs. Modal	
Feature	Correlation
Vowel_Duration	0.333
CPP_2	-0.254
Local_Shimmer_dB_2	0.217
Local_Shimmer_2	0.210
H1* – A2*_2	0.208
H1* – A1*_2	0.198
HNR35_2	-0.196
APQ11_Shimmer_Mean	0.182
HNR25_2	-0.180
HNR05_2	-0.166

classifiers reach a weighted F1 score of 0.73, suggests that perhaps an interaction between features is important to distinguishing phonation types.

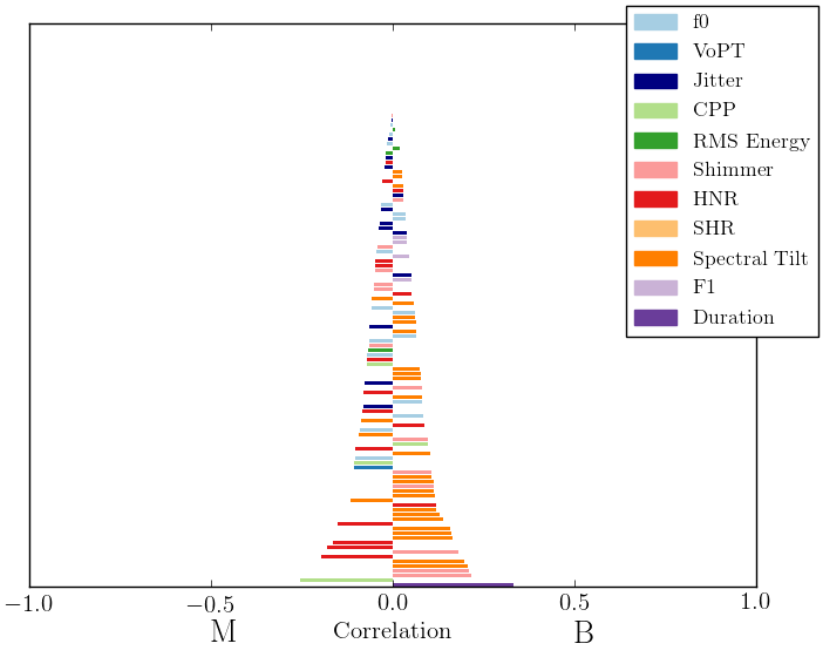


Figure 7.1: Gujarati Feature Correlations

Vowel Duration (dark purple) has the strongest correlation with Gujarati phonation types; a longer duration is associated with breathy voicing. The remaining features are all weakly correlated with Gujarati phonation. CPP (light green), a measure of noise in the signal, is the strongest after Vowel Duration and is correlated, as expected, with modal voicing. Several Shimmer measures (pink) and Spectral Tilt measures (dark orange) are correlated with breathy voicing; increased shimmer indicates variation in intensity between periods, and increased Spectral Tilt indicates slower vocal fold closure and reduced glottal constriction. Finally, HNR (red) is weakly correlated with modal voicing. Like CPP, HNR indicates a greater degree of periodicity by way of reduced noise.

Though the top features come from various categories, there is one clear trend: eight of the ten strongest correlated features are measured over the middle third of the vowel. This follows the timing of phonation that Khan (2012) observes. His measures most efficiently distinguish breathy voice from modal voice when calculated around the vowel's midpoint. Figure 7.2 again plots each feature's correlation, with colors representing the time span over which the measurement is calculated rather than feature categories. Features are measured either over the entire vowel or its first, second, or final third. For two features – Vowel Duration and VoPT – reporting the time period does not make sense; these are labeled as 'n/a' (dark blue).

Figure 7.2 reveals, as expected, a cluster of features measured over the middle third of the vowel (pale yellow) with the some of the strongest correlations. Of course, not all of the top features are calculated over the middle of the vowel; diagnostic information is likely available from other parts of the vowel as well. In particular, several top features are calculated over the mean of the vowel (red). These features could have such clearly distinguishing values in the middle third that the mean is impacted by that middle third, or the beginning and final thirds could be providing useful information as well. Looking towards the middle and top of the figure, where the correlations are weakest, the four time spans seem fairly evenly distributed; no time span emerges as providing the least diagnostic information.

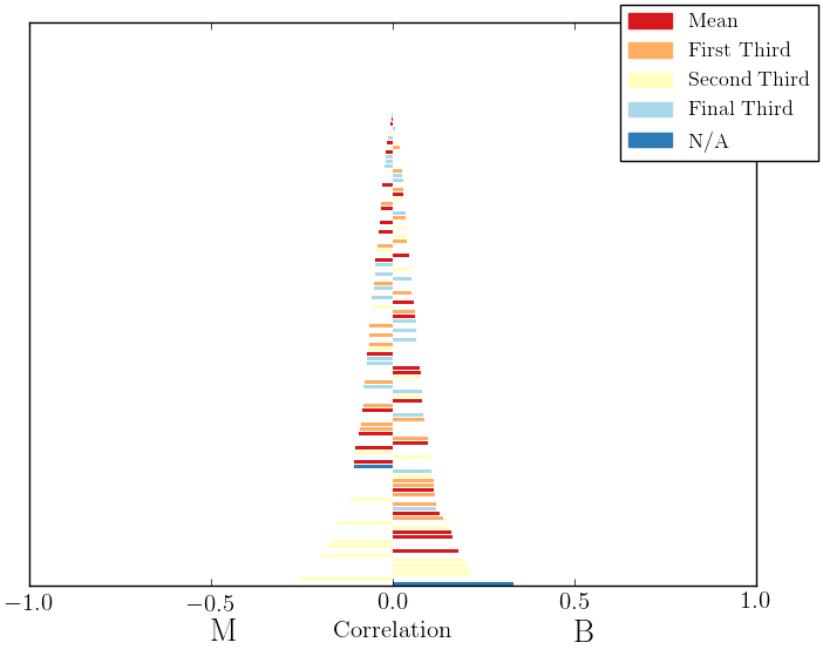


Figure 7.2: Gujarati Feature Correlations, By Time Period

7.3 SVM Weights

I turn now to the features most important to the Support Vector Machine. Figure 7.3 plots the weights for each feature. As in the English data, multicollinearity leads to contradictory weights. In order to get meaningful weights, I'll again need to narrow down my set of features by eliminating redundant ones.

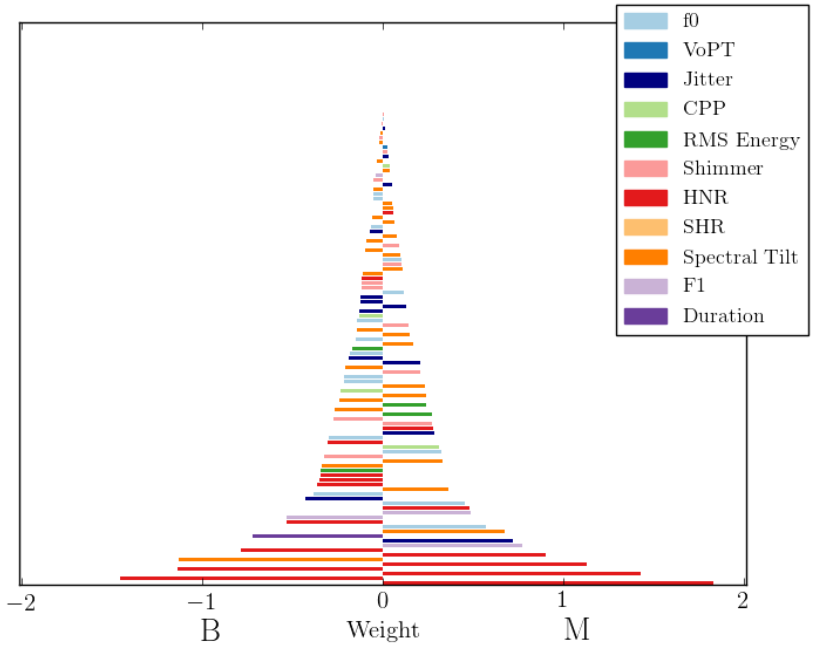


Figure 7.3: Gujarati Feature Weights

As with the English data, I first try the simplest solution: using a subset of the features based on each category's feature that received the highest weight when all features were in included. This subset of features achieves a weighted F1 score of 0.67536, which I consider to be a tolerable drop from the original 0.72079. These weights look plausible given the correlations and previous studies on Gujarati voice quality. Additionally, I checked that the directionality of the mean weight of each feature category matches the directionality of the

feature I'm using in the subset.[1] This was the case for the five largest weights but not the five smallest; given just how small the five smallest weights are, I do not consider this to be a problem. The weights of this subset of features are reported in Table 7.5 and plotted in Figure 7.4.

Table 7.5: Gujarati Top Feature Weights

Feature	Weight
Vowel Duration	-0.816
F1_2	0.738
H1* – A2*_2	-0.566
HNR35_2	0.492
VoPT	0.401
RMS_Energy_2	0.161
SHR_f_0_2	0.073
CPP_2	0.049
Local_Jitter_2	0.039
APQ3_Shimmer_2	0.036

As in the correlations, Vowel Duration is the most important of the features. Not far behind is F1, which is associated with modal voicing but has a very low correlation. Spectral Tilt and HNR have the third and fourth strongest weights (Spectral Tilt associated with breathy voicing and HNR with modal voicing), which is consistent with the correlations.

The weights show the same timing trend as the correlations: eight of the ten top weighted features are calculated over the vowel's middle third. The two that are not are Vowel Duration and VoPT, which both inherently require the entire vowel to calculate.

7.4 Random Forest Importance

The next way of evaluating how features contribute to the classification task is *importance*. Importance is the Random Forest's equivalent of SVM weights. Unlike weights, importance

[1]Multicollinearity can lead to odd weights of individual features, but those odd weights should cancel each other out. Comparing an individual feature's weight to the category's average will confirm that it's representative of what the feature is actually doing.

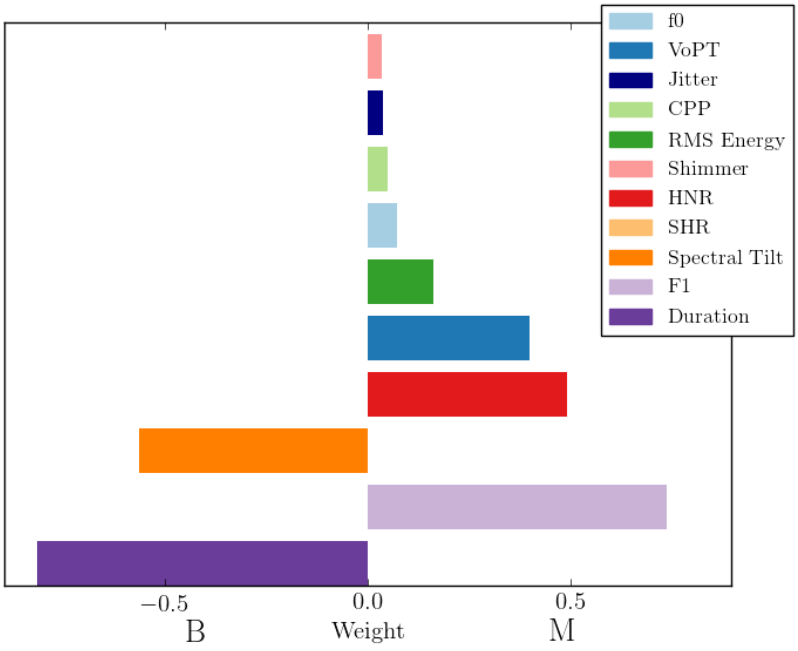


Figure 7.4: Gujarati SVM Weights

values are always positive. This means that importance does not indicate an association with a specific phonation type, though that can generally be inferred from the correlations and weights. The ten most important features are listed in Table 7.6. The importance values for all 101 features are plotted in Figure 7.5 and listed in Appendix I. Overall, the Gujarati importances are more consistent with the correlations than with the weights.

Table 7.6: Gujarati Top Feature Importance

Feature	Importance
Vowel_Duration	0.032483
APQ3_Shimmer_1	0.029940
Local_Shimmer_2	0.025131
APQ3_Shimmer_2	0.024441
Local_Shimmer_dB_2	0.023512
VoPT	0.020127
CPP_2	0.020112
HNR35_2	0.018165
Local_Shimmer_dB_1	0.016286
H1* – A1*_2	0.015990

Vowel Duration (dark purple) is the most important feature. This is also the feature with the strongest correlation and the strongest weight in both subsets of features for the SVM. Increased vowel duration is associated with breathy voicing.

A cluster of four Shimmer features (pink) are the most important features after Vowel Duration. Two of those four features are ranked third and fourth in terms of correlations, but Shimmer was not heavily weighted by the SVM. Following Shimmer is Variance of Pitch Tracks (bright blue). VoPT has so far been unremarkable, but it has been on the upper end of the middle of the pack by both correlations and weights. However, it's weakly associated with modal voicing according to both correlations and weights, which makes me suspicious that it's not actually doing much work.

As seen in Figure 7.5, there's no clear drop off after which features importance is much lower; the slope is fairly gradual. After VoPT, there are a few familiar colors – light green (CPP), red (HNR), more pink (Shimmer), and dark orange (Spectral Tilt). I will refrain

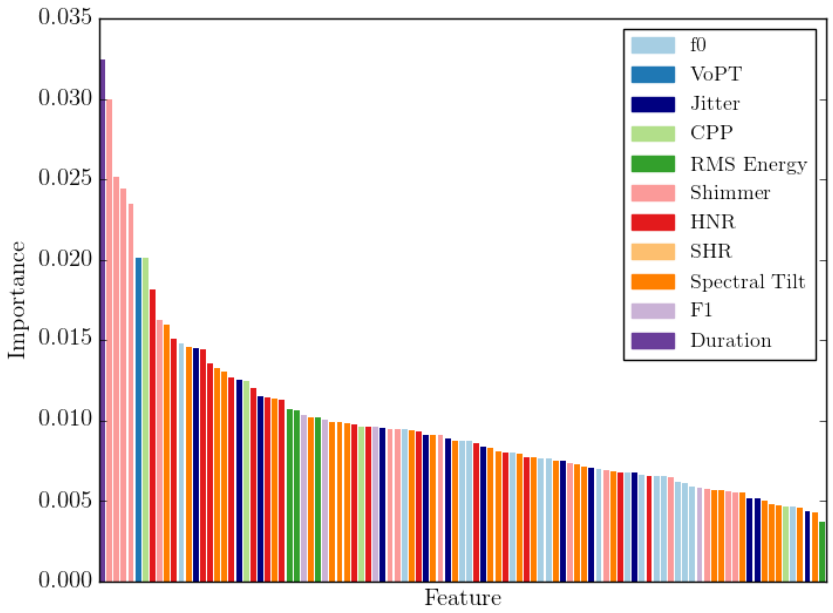


Figure 7.5: Gujarati Feature Importance

from speculating about the remaining features due to their relatively low importance.

The same trend regarding timing appears for the importance values – many of the most important features are calculated over the middle third of the vowel. Figure 7.6 shows feature importance color coded by time span of the calculation rather than by feature category. The cluster of pale yellow features on the left side are those measured in the vowel's middle third.

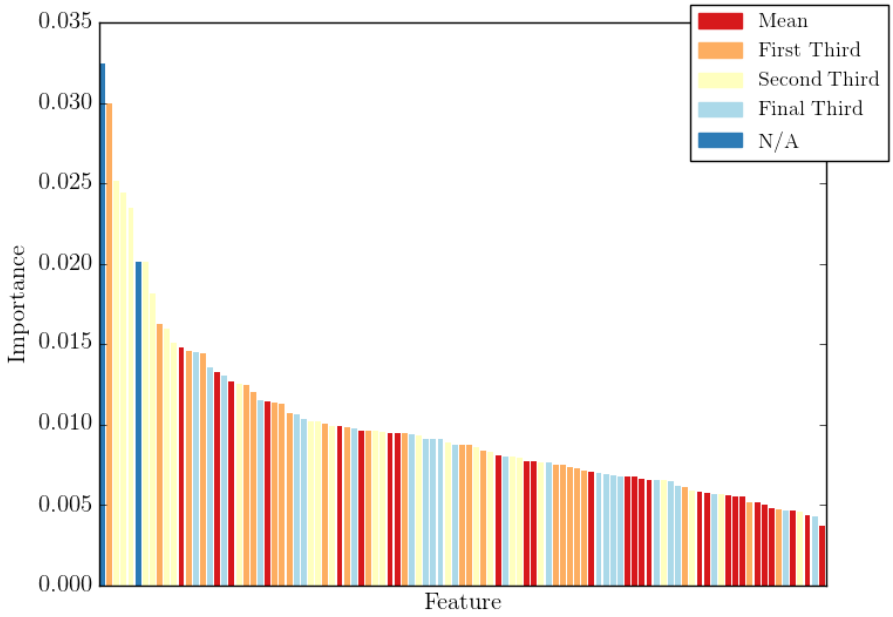


Figure 7.6: Gujarati Feature Importance, by Time Period

7.5 Ablation

Ablation is the final way I'll evaluate which features contribute most to distinguishing Gujarati breathy voice from modal voice. I first ablate entire categories of features, and then ablate features by time span.

7.5.1 Ablation by Category

Gujarati's 101 features fall into ten categories. For each feature category, I run an SVM and a Random Forest excluding all features within that category. The resulting weighted F1 score, as well as the change in weighted F1 from the original model, is listed in Table 7.7. Figure 7.7 plots the change in accuracy for each category.

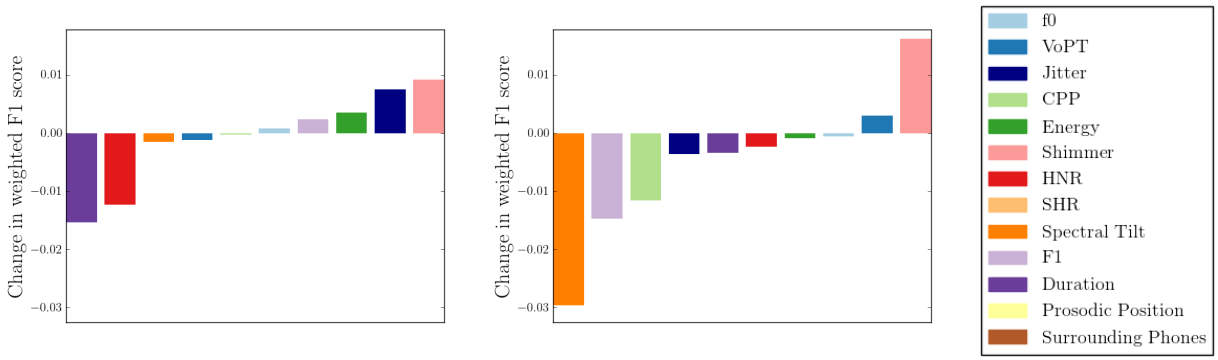
Table 7.7: Gujarati Category Ablation

Feature	SVM		Random Forest	
	Weighted F1	Change	Weighted F1	Change
Spectral Tilt	0.71936	-0.00143	0.69728	-0.02966
CPP	0.72057	-0.00022	0.71534	-0.0116
Energy	0.72441	0.00362	0.72614	-0.0008
HNR	0.70855	-0.01224	0.72466	-0.00228
f_0	0.72171	0.00092	0.72643	-0.00051
F1	0.72328	0.00249	0.71224	-0.0147
Duration	0.70546	-0.01533	0.72357	-0.00337
Jitter	0.72846	0.00767	0.72341	-0.00353
Shimmer	0.7301	0.00931	0.74326	0.01632
VoPT	0.71965	-0.00114	0.73004	0.0031

The two algorithms paint rather different pictures for categorical ablation. However, each category's ablation causes such a small drop in weighted F1 that I hesitate to read much into these results.

Ablating Vowel Duration (dark purple) causes the largest drop in accuracy for the SVM. Vowel Duration has the strongest correlation, weight, and importance, so it's no surprise that its ablation causes the largest drop. What is surprising is that ablating Vowel Duration makes almost no difference to the Random Forest.

HNR (red) causes the second largest drop in weighted F1 score for the SVM. Several HNR features are weakly correlated with modal voicing (but still among the most strongly correlated features), one has the fourth largest weight, and the eighth most important feature is also HNR. Ablating HNR causes a negligible change in the Random Forest. No feature



(a) SVM (b) Random Forest

Figure 7.7: Gujarati Category Ablation

categories besides Vowel Duration and HNR cause much of a drop in the SVM's weighted F1 score.

For the Random Forest, ablating Spectral Tilt (dark orange) causes the largest drop in weighted F1 score. Several Spectral Tilt features are among the most correlated features (though only weakly correlated with breathy voice). The one Spectral Tilt feature used in determining weights has the third largest weight, and the most important Spectral Tilt feature to the Random Forest is ranked tenth.

After Spectral Tilt, ablating F1 features (light purple) causes the second largest drop in weighted F1 score. All F1 features have low correlations and low importance, but F1_2 has the second largest weight. Ablating CPP (light green) causes a slightly smaller drop. The second strongest correlated feature is CPP_2, which is also the seventh most important feature. CPP is ranked very low among the weights.

A few features cause the weighted F1 score to *increase*. Recall that there's some randomness involved in Random Forests, so small increases are par for the course; this is less so for SVMs. Both models improve the most when all Shimmer features (pink) are ablated. This directly contradicts the correlations and importance rankings: Shimmer features make up three of the ten strongest correlated features and five of the ten most important features.

AQP3_Shimmer_2, the one feature included in weight calculations, has the smallest weight of all ten features. Many of the Shimmer features just barely make it past my 15% cutoff point for undefined measures; that is, they have fewer than 15% undefined measures but not *that* many fewer. I suspect that this plays into why Shimmer's role is different according to different metrics.

I also tried iteratively ablating categories; instead of ablating one category at a time and then replacing it, I pick the category whose ablation resulted in the largest drop in weighted F1 score and continue to exclude it. I repeat the process, ablating whichever category causes the largest drop, until there's one category left. Table 7.8 reports the weighted F1 score when the category of features causing the greatest drop is removed and not replaced for an SVM.

Table 7.8: Gujarati Iterative Category Ablation (SVM)

Category	Weighted F1
Vowel Duration	0.70546
HNR	0.70124
Spectral Tilt	0.68195
CPP	0.64357
Shimmer	0.6207
f_0	0.61005
RMS Energy	0.52883
Jitter	0.45527
F1	0.36262
VoPT	–

Following the trend seen in the non-iterative categorical ablation, Vowel Duration, HNR, and Spectral Tilt cause the three largest drops in weighted F1 when ablated iteratively. At the bottom of the list – the features whose ablation makes the least difference – are F1 and VoPT, which are fairly highly ranked by weights and importance, respectively. (Note that there's no value for VoPT, because ablating it would mean that no features are left.)

7.5.2 Ablation by Time Span

So far, features calculated over the vowel's middle third have generally been more useful to the classifiers. Here, I ablate features grouped by time span rather than by feature category. Recall that there are four time spans – each third of the vowel, plus the entire vowel – as well as two features[2] whose calculations don't quite fit into this paradigm ('n/a').

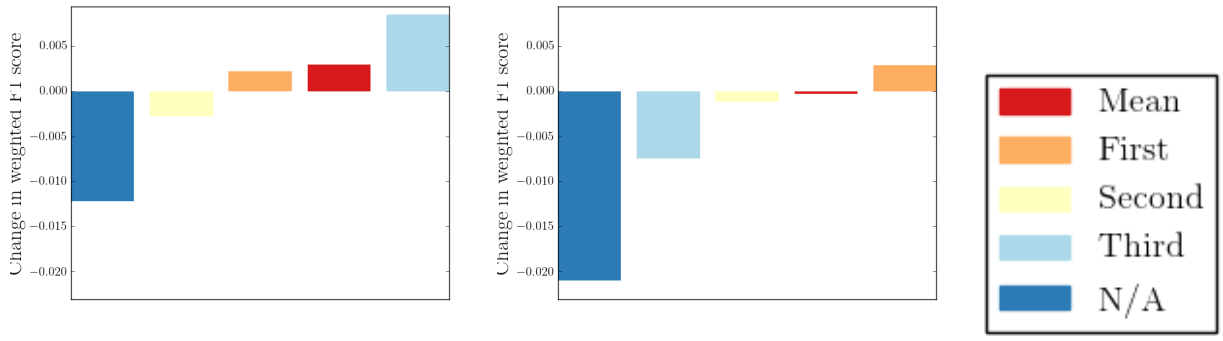
Table 7.9: Gujarati Time Span Ablation

Feature	SVM		Random Forest	
	Weighted F1	Change	Weighted F1	Change
1st Third	0.72298	0.00219	0.72983	0.00289
2nd Third	0.71801	-0.00278	0.72576	-0.00118
3rd Third	0.72931	0.00852	0.71944	-0.0075
Mean	0.72377	0.00298	0.72661	-0.00033
n/a	0.70858	-0.01221	0.70595	-0.02099

Ablating the 'N/A' features (dark blue) causes the largest drop in weighted F1 score for both classifiers, despite there being only two of these features. Ablating the middle third of the vowel causes the second largest drop for the SVM but barely changes the Random Forest.

Though features calculated over the vowel's middle third are fairly consistently ranked higher than those calculated over other parts of the vowel, excluding those features from the model doesn't change much. This doesn't necessarily mean that those features aren't important; redundant features could just be doing the same work. Of course, ablating groups of features based on time spans does reduce one form of redundancy, though each third of the vowel is represented in the mean and parts of the mean are represented by the other three time spans. Unlike features in the other four time spans, the two 'n/a' features do not have potentially redundant features, so it makes sense that ablating them impacts the classifiers the most.

[2]Those two features are Vowel Duration and VoPT.



(a) SVM (b) Random Forest

Figure 7.8: Gujarati Time Span Ablation

7.6 Summary

In the above sections, I examined how Gujarati's 101 features contribute to distinguishing breathy and modal vowels by reporting correlations, SVM weights, Random Forest importance, and changes caused by ablation. Several categories of features come up as important again and again in these four metrics. I'll review these features to summarize the Gujarati findings. I see three approximate tiers of feature categories within the subset of important ones: Vowel Duration and Shimmer in the top tier, followed by CPP and Spectral Tilt, and finally HNR and F1.

Vowel Duration is the top feature according to correlations, weights, importance, and SVM ablation (both iteratively and not); the only time it's not in first place is in Random Forest ablation. Khan (2012) reports that while an increased vowel duration is associated with breathiness in several languages (including Gujarati), but he does not include it as a measure in his study of Gujarati phonation types. Its consistent importance in the present study suggests that it is, perhaps, worth including as a measure.

Shimmer isn't as consistently highly ranked as Vowel Duration. In fact, it has the lowest weight and ablating the entire category causes the greatest *increase* in weighted F1 score.

That said, many Shimmer features are among the strongest correlations (correlated with breathy voicing) and most important features. In the iterative ablation, Shimmer is ranked fifth, solidly at the middle of the pack. I remain wary of the Shimmer features, as many of them have quite a few undefined measures (though all have fewer than 15% undefined measures). Shimmer has not been previously associated with Gujarati phonation, though I wasn't able to find any studies that tested it.

Down a tier from Vowel Duration and Shimmer is **Cepstral Peak Prominence**. Increased CPP is associated with modal voicing, as it indicates less noise in the signal. CPP_2 is the second most strongly correlated feature and the seventh most important feature, though its weight is low and its impact when ablated is small. Khan (2012) found a significant interaction (using two-way repeated measures ANOVAs) between CPP and Gujarati phonation at the midpoint of the vowel; this is consistent with my methods identifying CPP_2 as the most important of the four CPP features.

Spectral Tilt is more consistently important than CPP but its level of importance is lower. Increased Spectral Tilt is associated with breathy voicing, reflecting slower glottal closure. Spectral Tilt features hold third and fourth place among the correlations, third place among the weights, tenth among the importance values, and ablating the category causes the largest drop in weighted F1 score for the Random Forest. Khan (2012) used five measures of Spectral Tilt: H1* – H2*, H1* – H4*, H1* – A1*, H1* – A2*, and H1* – A3* in his study of Gujarati phonation, all of which are included in this study. He found that Gujarati breathy voicing has a larger Spectral Tilt than modal voicing according to all five measures. Two of the measures, H1* – H2* and H1* – H4*, were more significantly different at the vowel's midpoint than elsewhere in the vowel. My methods identify H1* – A1* and H1* – A2* as the most important overall, particularly when calculated using the middle third of the vowel.

I now move on to the final tier of features that stand out. **Harmonics-to-Noise Ratio** appears on all four lists, but is generally not especially high on those lists. A larger HNR is associated with modal voicing, as it's a fairly direct measure of periodicity. HNR35_2,

25_2, and 05_2 are the 7th, 9th, and 10th strongest correlated features. HNR35_2 has the 4th largest weight and the 8th largest importance. Ablating all HNR features leads to the second largest drop in weighted F1 score for the SVM, but does little to change the Random Forest. Khan (2012) tested the same HNR ranges as I did using ANOVA. He found that while the three highest frequency ranges of HNR were significantly different in breathy and modal vowels, HNR05 was not. From this, he concludes that the noise associated with breathy voicing is concentrated above 500 Hz. While HNR05 is the tenth most strongly correlated feature in this study, HNR35 is more strongly correlated, so I consider my results to be consistent with Khan's observation that Gujarati breathy voicing is noisier in higher frequencies.

Finally, **Variance of Pitch Tracks** comes up twice in the Gujarati results. VoPT has the fifth largest weight and is the sixth most important feature. However, its correlation with Gujarati phonation types is very weak and ablating it does little to change either model. I hesitate somewhat to include it here because of the correlations and ablation results, but both machine learning models identify it as useful.

These six categories of features – Vowel Duration, Shimmer, CPP, Spectral Tilt, HNR, and VoPT – contribute various amounts of information to the classifiers. These features have been identified by correlations, weights, importance, and ablation as being useful in distinguishing Gujarati breathy voice from modal voice. One final way of looking at the contributions of these features is to see how a classifier performs with *just* this subset of features. Table 7.10 reports the weighted F1 score of an SVM and a Random Forest trained using the subset of six feature categories, as well as using the features from each tier described above.

Using features from only six categories slightly improves the SVM's weighted F1 score and slightly decreases the Random Forest's. The fact that features from just these categories can do nearly the same job as all features suggests that I have correctly identified a subset of important features; their relative ranking, however, I may not have gotten right. Using just Vowel Duration and Shimmer, the two features I identified as the most important, lead

Table 7.10: Gujarati Weighted F1 Using Subsets of Features

Feature Categories	Weighted F1	
	SVM	Random Forest
Baseline (all features)	*0.72079*	*0.72694*
Duration, Shimmer, CPP, Spectral Tilt, HNR, VoPT	0.72281	0.71873
Duration, Shimmer	0.65969	0.58765
CPP, Spectral Tilt	0.67879	0.73204
HNR, VoPT	0.66924	0.6624

to worse weighted F1 scores than either of the other tiers. This doesn't mean that Vowel Duration and Shimmer aren't important; more likely, they just don't provide a complete enough picture on their own. CPP and Spectral Tilt together lead to a better weighted F1 than Duration and Shimmer for both classifiers, and much more so for the Random Forest. HNR and VoPT lead to a better classifier than Duration and Shimmer, but a slightly worse classifier than CPP and Spectral Tilt.

While the exact ranking of feature categories for Gujarati is a little murky, these results suggest that Gujarati breathy voice and modal voice can be distinguished by periodicity (CPP, HNR), spectral characteristics (Spectral Tilt), various irregularities in the signal (Shimmer, VoPT), and the duration of the vowel. The following chapter explores Hmong phonation through the lense of machine learning.

Chapter 8

HMONG

This chapter focuses on Hmong, discussing how the classifiers perform and the features most associated with Hmong's three phonation types according to correlations, SVM weights, Random Forest importance, and ablation.

Hmong uses complex contrastive tones that are distinguished by both pitch and phonation. Three phonation types are represented in Hmong's tones – breathy, modal, and creaky voice. The Hmong data are rather imbalanced regarding phonation; as seen in Table 8.1, over half of the tokens are modal. Chapter 4 describes Hmong phonation and the data set used here in more detail.

Table 8.1: Hmong Phonation Distribution

Type	Count	*Percent*
B	535	*19.691%*
M	1494	*54.987%*
C	688	*25.322%*
Total	2717	

Before running the two classifiers, the Hmong data has undergone several pre-processing steps. All data have been normalized (either by speaker, by vowel, or overall – see Chapter 4 for more detail). All features with 15% or more undefined measures are removed, and any remaining undefined measures will be replaced[1] within each fold. Finally, because Hmong tone and phonation are inherently intertangled, I exclude all f_0 measures. This leaves a total

[1]Undefined values in the training data are replaced with the mean of the class; undefined values in the testing data are replaced with the overall mean. See Chapter 4 for more information on handling missing values.

of 62 features from nine categories, as listed in Table 8.2.

Table 8.2: Hmong Features

Feature Category	Number of Features
Spectral Tilt	28
Harmonics-to-Noise Ratio	16
f_0	0
Surrounding Phones	0
Cepstral Peak Prominence	4
Prosodic Position	0
F1	4
RMS Energy	4
Variance of Pitch Tracks	1
Vowel Duration	1
Jitter	2
Subharmonic-to-Harmonic Ratio	2
Shimmer	0
Total	62

8.1 Model Performance

I start by training four classifiers on the Hmong data: a Support Vector Machine and a Random Forest, each based on imbalanced and resampled data. Each model's performance is listed in Table 8.3.

Table 8.3: Hmong Classifier Performance

Balance	clf	Accuracy	Weighted F1	Precision			Recall			F1 Score		
				B	M	C	B	M	C	B	M	C
Imbalanced	SVM	69.709	0.68747	0.62	0.72	0.69	0.47	0.84	0.55	0.54	0.77	0.62
	RF	67.832	0.67407	0.55	0.73	0.63	0.46	0.78	0.62	0.51	0.76	0.63
Resampled	SVM	63.416	0.6444	0.45	0.82	0.57	0.69	0.61	0.64	0.54	0.7	0.6
	RF	66.434	0.66723	0.52	0.76	0.6	0.58	0.71	0.63	0.54	0.73	0.62

8.1.1 Imbalanced Data

The SVM and Random Forest trained on imbalanced data follow roughly the same patterns. They have similar weighted F1 scores: 0.68747 and 0.67407, respectively. Unsurprisingly, precision and recall are highest for modal voicing (the majority class), followed by creaky voicing (the larger of the two minority classes), and finally (not far behind) by breathy voicing (the smaller of the two minority classes). Precision is higher than recall for both minority classes and recall is higher than precision for the majority class.

8.1.2 Resampled Data

Resampling the data evens out the distribution of phonation classes, so modal voicing no longer dominates the data. This causes a drop in weighted F1 score, though not a particularly large drop: the SVM drops from 0.68747 to 0.6444 and the Random Forest from 0.67407 to 0.66723. For both classifiers, precision is still higher for modal voice than for breathy or creaky voice. SVM recall, however, is higher for both breathy and creaky voice than for modal voice. This is a result of resampling; the SVM is now more complete for the minority classes. Because resampling levels the playing field for the three phonation types, I opt to use resampled data in the following sections.

Across the board, the weighted F1 scores for the four Hmong classifiers are lower than for English or Gujarati. While acheiving a weighted F1 score in the range of 0.6 to 0.7 requires some patterns in the data, it does not signal strong patterns or confidence in those patterns. This relatively low performance could have to do with how Hmong uses phonation. Unlike in English or Gujarati, phonation alone is not contrastive; pitch provides another cue to the phonemic tone. Phonation does not need to do all the work and may therefore be less clearly encoded in the speech signal. While human listeners can use both pitch and phonation to figure out the tone, I've provided my classifiers with only the phonation piece of the puzzle, which may negatively impact their ability to classify phonation types. Nonetheless, they perform above chance and I expect that patterns will emerge in which features are used

most.

8.2 Correlations

The first way I evaluate the relationship between features and phonation types is correlations. This section presents each feature's correlation with Hmong's three phonation contrasts. The ten strongest correlated features for each contrast are listed in Table 8.4, and the full set is listed in Appendix J. I review these correlations for the three contrasts below.

Table 8.4: Hmong Top Feature Correlations

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
Feature	Correlation	Feature	Correlation	Feature	Correlation
SHR_Mean	-0.379	SHR_Mean	-0.334	HNR05_Mean	-0.479
H1* – H2*_2	0.371	H1* – H2*_2	0.327	HNR05_2	-0.451
H1* – H2*_Mean	0.366	SHR_1	-0.310	HNR05_3	-0.421
H1* – H2*_1	0.338	H1* – H2*_Mean	0.309	RMS_Energy_3	-0.341
H1* – A1*_2	0.337	H1* – H2*_1	0.305	HNR05_1	-0.337
HNR05_3	0.336	H1* – A1*_2	0.279	RMS_Energy_2	-0.291
SHR_1	-0.329	H1* – A1*_1	0.274	RMS_Energy_Mean	-0.275
H1* – A1*_Mean	0.322	H1* – A1*_Mean	0.264	HNR15_2	-0.245
HNR05_Mean	0.315	HNR05_2	-0.260	HNR35_2	-0.231
H1* – H2*_3	0.308	H1* – A2*_2	0.252	HNR25_2	-0.229

8.2.1 Breathy vs. Creaky Correlations

Figure 8.1 plots feature correlations for the breathy vs. creaky contrast. Each bar represents a feature, its color represents its category, and its magnitude represents its correlation. Features with positive correlations have larger values in breathy voicing and features with negative correlations have larger values in creaky voicing. No correlations for this contrast are strong, but many are moderate.

Subharmonic-to-Harmonic Ratio (light orange) provides the strongest and seventh strongest correlation for this contrast. SHR measures the subharmonics in the signal, which often appear in multiply pulsed signals and is typically used to distinguish creaky

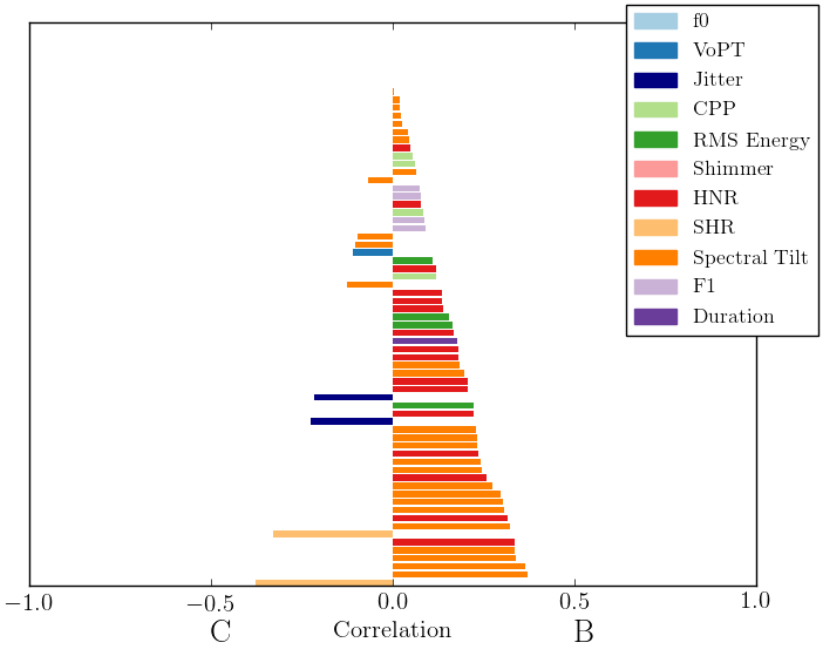


Figure 8.1: Hmong Feature Correlations, B vs. C

voice from modal voice. Here, greater SHR is correlated with creaky voicing compared to breathy voicing. SHR has not been previously associated with Hmong phonation, and I am particularly surprised to see it correlated with this contrast.

Spectral Tilt features (dark orange) account for many of the features most strongly correlated with breathy voicing in this contrast. Spectral Tilt is often considered to be the most reliable measure of phonation cross-linguistically. Larger Spectral Tilt represents a larger drop in energy as frequency increases, which is caused by reduced glottal constriction and slower vocal fold closure, both characteristic of breathy voicing. Spectral Tilt is generally highest in breathy voicing and lowest in creaky voicing, so this contrast should reflect the largest difference in Spectral Tilt. The three most strongly correlated Spectral Tilt measures are all H1* – H2*, calculated over different parts of the vowel.

Sprinkled among the Spectral Tilt features are several Harmonics-to-Noise Ratio features (red). Larger HNR is typically indicative of modal voicing, though here it's correlated with breathy voicing rather than creaky voicing. This suggests that Hmong breathy voice is more periodic than creaky voicing.

Finally, though the correlations are weak, I'd like to point out Jitter (dark blue), which is correlated with creaky voice. Only two Jitter measures – Local Jitter and Local Absolute Jitter, both calculated over the entire vowel – have few enough undefined measures to be included. But those two are among the features most correlated with creaky voicing, suggesting that Hmong creaky voice may include cycle-to-cycle variation in f_0.

8.2.2 Breathy vs. Modal Correlations

The top correlations for Hmong's breathy vs. modal contrast, plotted in Figure 8.2, look rather similar to the top correlations for the breathy vs. creaky contrast. The correlations are overall slightly weaker, which comes as a bit of a surprise, as I generally expect a non-modal voice quality to have less in common with modal voice than with another non-modal voice quality.

Once again, SHR (light orange) is the strongest correlated feature. This time, it's

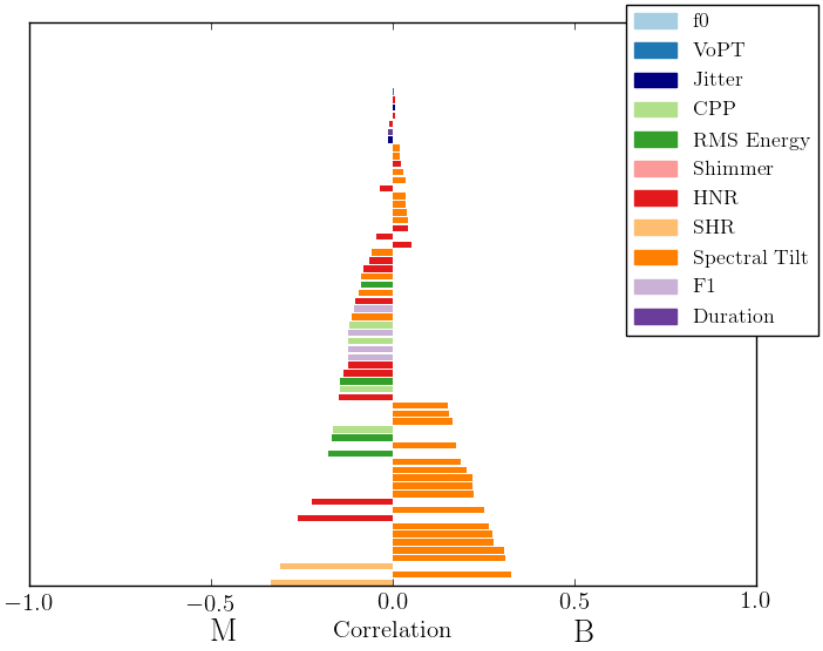


Figure 8.2: Hmong Feature Correlations, B vs. M

correlated with modal voicing. In other words, in both contrasts, larger SHR is *not* correlated with breathy voicing; Hmong breathy voicing likely does not contain subharmonics.

Spectral Tilt (dark orange) is again associated with breathy voicing, as it is in the breathy vs. creaky contrast. Its correlations are slightly weaker compared to the breathy vs. creaky contrasts, which is consistent with the idea that Spectral Tilt is largest for breathy voicing and smallest for creaky voicing. And again, H1* – H2* accounts for the three most strongly correlated Spectral Tilt measures.

An HNR feature (red) is the ninth most strongly correlated feature for this contrast, though it is only weakly correlated with modal voice. Looking a little farther down the list, all of the strongest correlated HNR measures are HNR05, suggesting that the noise that distinguishes Hmong breathy and modal voice qualities is concentrated in the lower frequencies.

8.2.3 Modal vs. Creaky Correlations

Finally, Figure 8.3 plots each feature's correlation for the creaky vs. modal contrast. The top correlations for this contrast are stronger than for either of the other two Hmong contrasts, though still just below the range considered to be strong. Additionally, this figure is perhaps the most color-separated plot yet – features within a category are performing very similarly.

HNR (red) is correlated with modal voicing and accounts for the three strongest correlated features. As in the breathy vs. modal contrast, the strongest correlated HNR measures are all HNR05. These features are more strongly correlated with modal voicing compared to creaky voicing than compared to breathy voicing, suggesting that creaky voicing has more noise than breathy voicing in Hmong.

RMS Energy (dark green) is also correlated with modal voicing. Modal voicing typically has a higher intensity than both types of non-modal voicing. These correlations suggest that Energy also differs between the two types of non-modal phonation; creaky voicing has less energy than breathy voicing.

Delving into weak correlations, Cepstral Peak Prominence (light green) and F1 (light

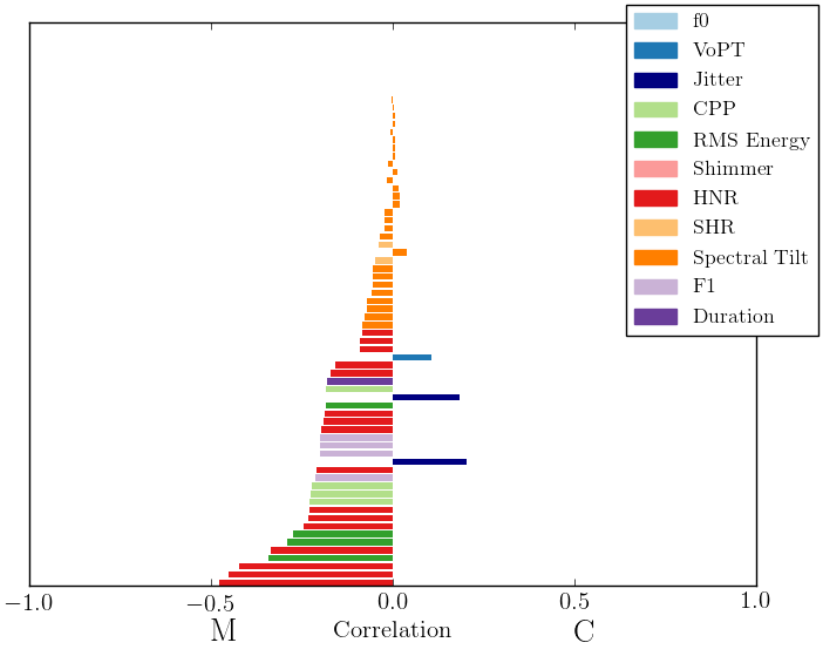


Figure 8.3: Hmong Feature Correlations, C vs. M

purple) are weakly correlated with modal voicing. The two Jitter measures (dark blue) again appear as two of the few measures correlated with creaky voice. Finally, the weakest correlations are all orange – SHR (light orange) and Spectral Tilt (dark orange). These two categories provide many of the features most correlated with breathy voicing, but clearly are much less relevant to creaky and modal voicing.

Several categories of features clearly emerge as the key players among the correlations: SHR, Spectral Tilt, and HNR, with Jitter and RMS Energy also important, but less so. The next sections use machine learning to examine the role of the various features in the classification task.

8.3 SVM Weights

The next lens through which to evaluate each feature's contribution to the classification task is *weights*. Weights are output by the Support Vector Machine and indicate how much the classifier uses each feature. As seen in the English and Gujarati chapters, weights are negatively impacted by multicollinearity. Redundant features – variations on calculations or time spans – are often assigned contradictory weights that are not meaningful. In order to get meaningful weights, I must pare down the set of features to include important ones but not redundant ones.

As a first pass at a subset of features, I use the feature from each category that was assigned the largest weight (for any contrast) in the SVM that used all 62 features. An SVM using these nine features has a weighted F1 score of 0.59289. This is fairly low, though not a particularly large drop from the SVM with 62 features, whose weighted F1 score is 0.6444. Moreover, the weights assigned to this subset of nine features are consistent with the correlations and their signs, for the most part, match each category's mean sign, indicating that they are representative of what that feature is actually doing and were not assigned an opposite weight due to multicollinearity.[2] The nine features that I'll use for the SVM

[2]Features whose signs don't match have extremely small weights and are unimportant.

are listed in Table 8.5, along with their weights for each of the contrasts. Like correlations, the magnitude of the weights indicates their importance; features with extremely negative or extremely positive weights are more useful to the SVM.

Table 8.5: Hmong Feature Weights

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
Local_Jitter_Mean	-0.947	SHR_1	-0.969	HNR05_Mean	-0.794
SHR_1	-0.945	RMS_Energy_3	-0.308	RMS_Energy_3	-0.642
HNR05_Mean	0.396	HNR05_Mean	-0.304	Local_Jitter_Mean	0.261
RMS_Energy_3	0.315	VoPT	-0.224	VoPT	-0.172
Vowel_Duration	0.131	H2* – H4*_Mean	0.141	F1_Mean	0.106
VoPT	-0.089	F1_Mean	-0.132	Vowel_Duration	-0.101
F1_Mean	-0.035	Local_Jitter_Mean	-0.054	H2* – H4*_Mean	-0.065
CPP_1	-0.017	Vowel_Duration	0.044	SHR_1	-0.055
H2* – H4*_Mean	0.009	CPP_1	-0.019	CPP_1	-0.005

8.3.1 Breathy vs. Creaky Weights

Figure 8.4 reports weights for the subset of features in the breathy vs. creaky contrast. Features with positive weights tend to have higher values in breathy voicing; those with negative weights tend to have higher values in creaky voicing.

Jitter (dark blue) has the largest weight for this contrast. Increased Jitter is associated with creaky voicing. Jitter is also consistently but not strongly correlated with creaky voicing. Though Jitter is arguably one of the more direct ways of quantifying phonation, it is often excluded from studies of phonation because human perception of aperiodicity is better captured by other measures (Kreiman and Gerratt, 2005). Nonetheless, it appears to capture a phonetic reality of Hmong phonation fairly well.

SHR (light orange) has nearly as strong a weight as Jitter and is also associated with creaky voicing. In the correlations, SHR is associated with *not* being breathy; SHR's weight here is consistent with this observation.

The next two features have weights that are quite a bit smaller than Jitter and SHR.

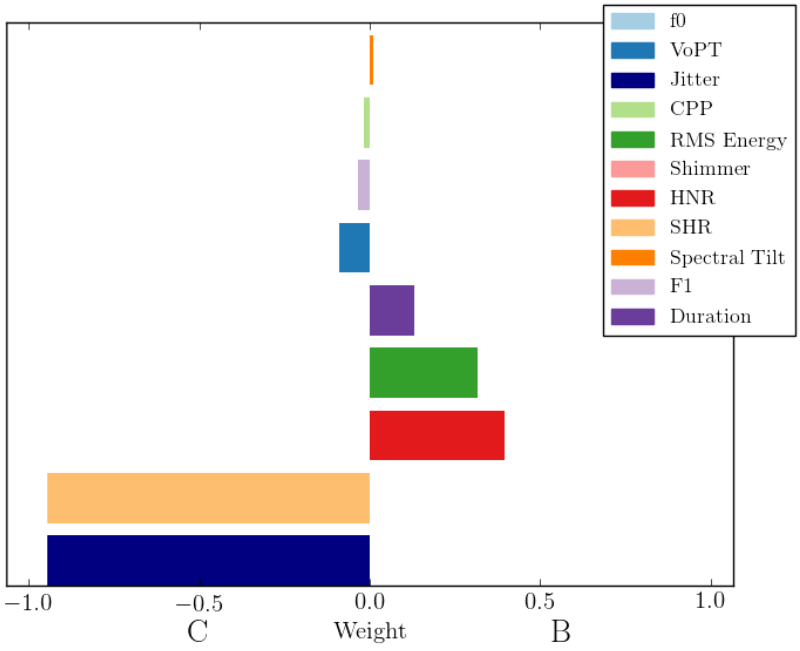


Figure 8.4: Hmong SVM Weights, B vs. C

HNR (red) and RMS Energy (dark green) are both associated with breathy voicing rather than modal voicing. Their usefulness in distinguishing the two types of non-modal phonation is unexpected but consistent with the correlations.

8.3.2 Breathy vs. Modal Weights

Feature weights for the breathy vs. modal contrast are shown in Figure 8.5. They are generally consistent with the correlations, with one exception.

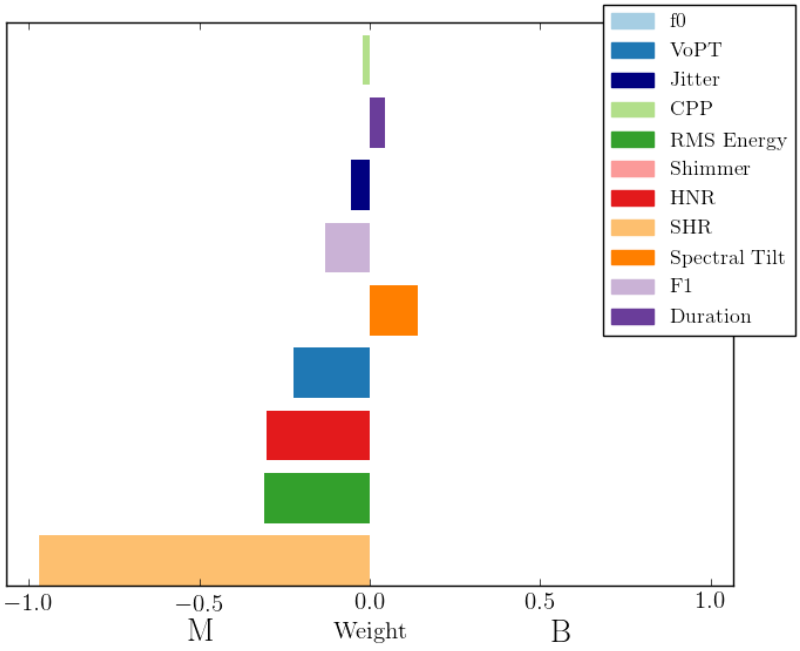


Figure 8.5: Hmong SVM Weights, B vs. M

Like the correlations, SHR is the most strongly weighted feature and is associated with modal voicing. After SHR, three features have similar weights, all quite a bit lower than SHR's weight: larger RMS Energy (dark green), HNR (red), and Variance of Pitch Tracks (bright blue) are all associated with modal voicing. RMSE has a small weight, but its weight is the second largest for this contrast; in the correlations, RMSE features are only weakly

correlated with modal voicing. HNR is just behind RMSE, though features from this category have larger correlations than RMSE in this contrast. VoPT, which has the fourth largest weight, has the weakest correlation of all 62 features for the breathy vs. modal contrast. However, its weight is quite small; importance and ablation may help shed light on VoPT's role in distinguishing Hmong phonation types.

8.3.3 Creaky vs. Modal Weights

Finally, feature weights for the creaky vs. modal contrast are listed in Figure 8.6. These features are again consistent with those identified as important based on correlations for this contrast.

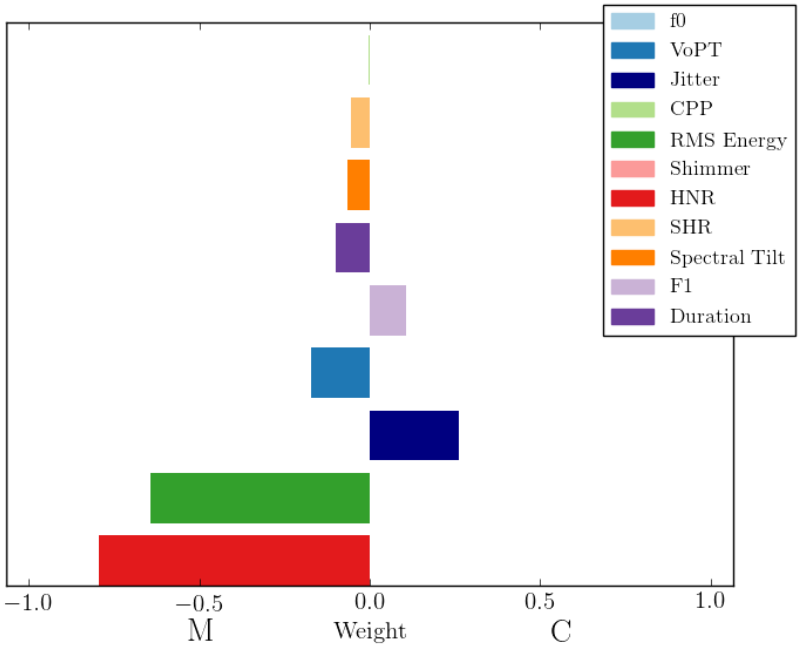


Figure 8.6: Hmong SVM Weights, C vs. M

HNR (red) has the largest weight. Greater HNR is associated with modal voicing according to both weights and correlations; this makes sense, as modal voicing should be

more periodic than creaky voicing, resulting in less noise relative to harmonics.

RMS Energy (dark green) is also associated with modal voicing with a slightly smaller weight than HNR. Energy is expected to be higher in modal voicing than in either non-modal voice quality, and that appears to be the case in Hmong according to the weights. The weights from the three contrasts taken together suggest that RMS Energy is greatest in modal voicing and lowest in creaky voicing, with breathy voicing somewhere in between.

Again, a short list of key feature categories comes out of the SVM weights – SHR, Jitter, HNR, and RMS Energy.

8.4 Random Forest Importance

The second classifier is the Random Forest, which assigns value to features using *importance*. Importance doesn't indicate a feature's association with a particular phonation type, but rather its overall contribution to the model. Table 8.6 lists the ten features with the greatest importance. The full set of 62 features is plotted in Figure 8.7 and their values are listed in Appendix J. Overall, the importance values are rather low, which is consistent with rather low weighted F1 score.

Table 8.6: Hmong Top Feature Importance

Feature	Importance
Local_Abs._Jitter_Mean	0.087574
Local_Jitter_Mean	0.071645
HNR05_2	0.044905
SHR_1	0.038391
SHR_Mean	0.035642
VoPT	0.034468
HNR05_3	0.032611
H1* – H2*_2	0.030976
HNR05_Mean	0.025285
RMS_Energy_3	0.024221

Two Jitter features (dark blue) are nearly twice as important as any other feature, indicating that they are doing nearly twice as much work as any other feature. Both are

measured over the entire vowel, one using Local Absolute Jitter and one using Local Jitter. These two measures are very similar, the only difference being that Local Jitter is averaged over the period while Local Absolute Jitter is not.

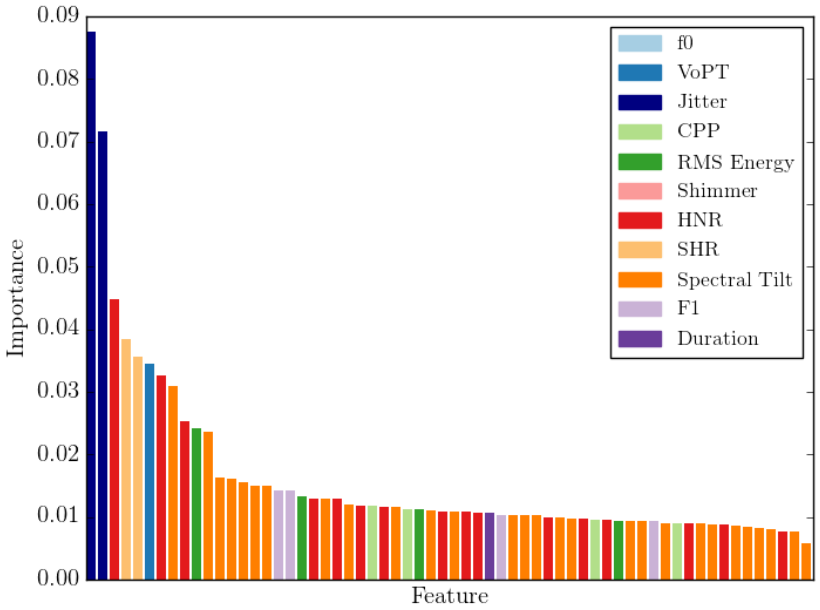


Figure 8.7: Hmong Feature Importance

HNR (red) is the third most important feature, though it is approximately half as important as the most important feature. Three HNR measures are among the top ten most important, and all three measure noise from 0 to 500 Hz.

The two SHR features are the fourth and fifth most important features for the Random Forest. SHR has so far come up when breathy voice is involved, and it's associated with whatever the other voice quality is, suggesting that Hmong breathy voicing has very low SHR values.

Variance of Pitch Tracks (bright blue) is the sixth most important feature. VoPT doesn't play a major role in distinguishing Hmong phonation types according to the correlations

and the weights, so I'm somewhat surprised to find it here. It is, however, amongst the few features positively correlated with creaky voicing, so I suspect its role in the Random Forest is to distinguish creaky voicing from modal and breathy voicing.

Spectral Tilt (dark orange) is the eighth most important feature, and a cluster of six more is found just below the top ten. H1* – H2*_2 is the most important; the cluster contains H1* – H2* and H1* – A1* measures over various time spans. Though spectral tilt was notably absent from the top weights, this set of features is very similar to the Spectral Tilt features that are most correlated with breathy voicing.

Finally, RMS Energy (dark green) is the tenth most important feature. It is by far most important when calculated over the final third of the vowel – the three RMS Energy measures taken over other time points are much less important. It has so far been associated with modal voicing, particularly as compared to creaky voicing.

Several tiers of important features are visible in Figure 8.7. The clear top performers are Jitter features. In the next tier is HNR, SHR, VoPT and Spectral Tilt. There are some repeats in the next tier, and RMS Energy. The remaining features have very low importance.

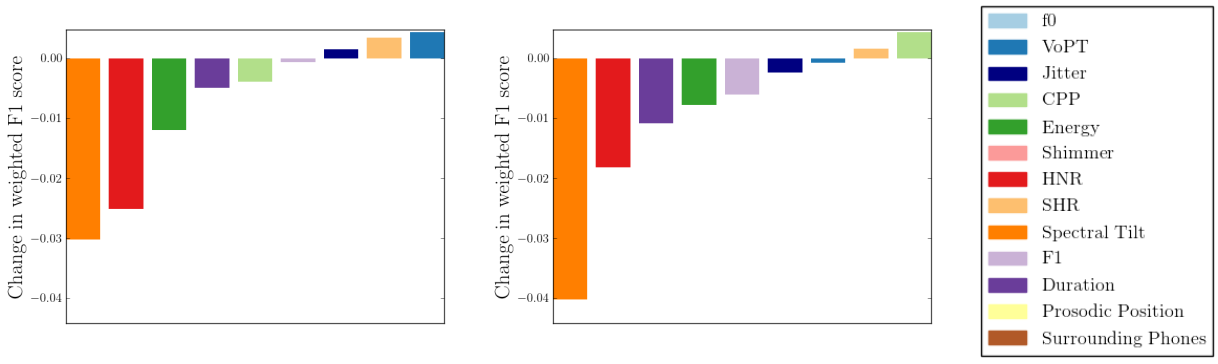
8.5 Ablation

This chapter has so far identified features deemed important to distinguishing Hmong phonation based on correlations, SVM weights, and Random Forest importance. A final way to examine a feature's contribution to the models is to train the classifier without it. Ablating a single feature, however, is not very informative when the set of features contains multiple collinear features; when one is removed, a similar feature will take over the job of the ablated feature. To avoid this problem, I start by ablating entire categories of features. Hmong's 62 features come from nine categories. Table 8.7 lists how the SVM and Random Forest perform when each category's features are ablated. The same information is presented graphically in Figure 8.8.

Ablating Spectral Tilt (dark orange) causes the largest drop in weighted F1 score for both the SVM and the Random Forest. Though Spectral Tilt features have come up in

Table 8.7: Hmong Category Ablation

Feature	SVM		Random Forest	
	Weighted F1	Change	Weighted F1	Change
Spectral Tilt	0.6142	-0.0302	0.6271	-0.04013
CPP	0.64056	-0.00384	0.67162	0.00439
Energy	0.63247	-0.01193	0.6595	-0.00773
HNR	0.61932	-0.02508	0.64914	-0.01809
SHR	0.64792	0.00352	0.66894	0.00171
F1	0.64382	-0.00058	0.6613	-0.00593
Duration	0.63953	-0.00487	0.65643	-0.0108
Jitter	0.64601	0.00161	0.66497	-0.00226
VoPT	0.6488	0.0044	0.66659	-0.00064



(a) SVM (b) Random Forest

Figure 8.8: Hmong Category Ablation

correlations and importance, this is the first time they are at the front of the pack. Here, their absence impacts the two classifiers more than any other category of features.

Ablating HNR (red) features causes the next largest drop in performance for both classifiers. These features have come up again and again, so their importance according to ablation is unsurprising. After HNR, the two classifiers suffer most from ablating Energy and Duration features, though in opposite orders. While Energy features have been somewhat important, this is the first time Duration has come up. That said, it causes such a small drop in weighted F1 for both classifiers that it's probably not important.

The feature categories whose ablation causes the largest drop in accuracy are fairly consistent with the categories identified by other metrics. However, several feature categories that have been important according to correlations, weights, and importance have little impact when ablated.

In addition to ablating categories one at a time, I ablate categories iteratively; I identify the category whose exclusion causes the largest drop in weighted F1, exclude it permanently, and then again look for the category whose exclusion causes the next largest drop. The results of this iterative ablation, conducted using just an SVM, are presented in Table 8.8.

Table 8.8: Hmong Iterative Category Ablation (SVM)

Category	Weighted F1
Spectral Tilt	0.6142
HNR	0.58719
Energy	0.55821
Jitter	0.51803
F1	0.50044
CPP	0.48669
SHR	0.33073
VoPT	0.31682
Duration	–

Iterative ablation, for the most part, shows that excluding the same feature categories we've been seeing over and over again hurts the classifiers the most. Spectral Tilt, HNR,

Energy, and Jitter features are important, while F1, CPP, and Duration are less important. Two features, however, are near the bottom of the iterative ablation list but identified by other metrics as important: SHR and VoPT. Though SHR is among the last features to be ablated, its ablation causes a large drop in weighted F1, from 0.48669 to 0.33073. The same can't be said for VoPT.

Overall, the two types of ablation consistently identify Spectral Tilt, HNR, and Energy as the three most important categories of features for distinguishing Hmong phonation types.

8.6 Summary

In each section – correlations, weights, importances, and ablation – I report a short list of feature categories identified as important in classifying Hmong phonation types. These four lists are rather consistent in their contents, but less so in their ordering. The one feature category that appears on all four lists is HNR. Several others appear on three of the four lists: SHR, Spectral Tilt, and Energy. Jitter appears twice, and VoPT once.

Harmonics-to-Noise Ratio, the only category that was among the top features for all four metrics, measures noise in the signal. A higher HNR value indicates less noise and is therefore associated with modal voicing. In both the correlations and the weights, larger HNR is associated with modal voicing compared to either type of non-modal voicing, but more strongly compared to creaky voice than compared to breathy voice. Larger HNR is also associated with breathy voice in the breathy vs. creaky contrast, but less so than for the other two contrasts. This suggests that not only is Hmong modal voice characterized by periodicity, but that breathy voice has a higher degree of periodicity than creaky voice. HNR05 (HNR measured from 0 to 500 Hz) is consistently the top feature within its category; the noise that distinguishes Hmong voice qualities is likely concentrated in the lower frequencies.

Fulop and Golston (2009) measured HNR (called *harmonicity*) in two speakers of White Hmong. Their goal was to distinguish Hmong breathy voice from *whispery* voice, which they demonstrated has a much lower HNR than breathy voice. I follow the tradition of most linguists in grouping breathy voice and whispery voice together; they are non-contrastive

and simply occupy different ranges along the phonation continuum. As such, the HNR distinction between breathy and whispery is generally irrelevant to my findings, though if two types of breathy voicing exist, that adds noise to the data and could worsen the classifiers' performance. Nonetheless, they demonstrated that the whispery/breathy type of phonation has a lower HNR than modal phonation in White Hmong. Garellek (2012) also used HNR (specifically HNR05) to study Hmong voice qualities, comparing the same three phonation types I do. He found, as I did, that creaky voice has a lower HNR than breathy voice, and that both non-modal phonation types have a lower hNR than modal voice.

Subharmonic-to-Harmonic Ratio measures subharmonics in the signal, which can arise in some types of creaky voicing due to multiple pulsing. In both the correlations and the weights, SHR is associated with creaky voicing over breathy voicing and with modal voicing over breathy voicing. However, it's not remarkable for creaky voicing compared to modal voicing. My interpretation of this is that SHR is noteworthy because Hmong breathy voicing lacks subharmonics. SHR features are also the fourth and fifth most important to the Random Forest. Ablating the entire category does little to either classifier's performance and while it's ablated late in the iterative ablation, it causes a large drop in weighted F1 score when it is ablated. Despite how important SHR is here, I was unable to find other studies that have investigated its role in describing Hmong phonation.

Spectral Tilt, which is widely considered to be the most robust measurement of phonation, comes up in three of my four short lists. Typically, breathy voicing has a large Spectral Tilt because of slow and incomplete glottal closure, while creaky voicing typically has a small Spectral Tilt because of faster and more complete closure; modal voicing is in the middle. The correlations suggest that Spectral Tilt is larger in breathy voice than in modal or creaky voice. The eighth most important feature is a measure of Spectral Tilt, and ablating the entire category causes the largest drop in weighted F1 score for both classifiers. However, the one Spectral Tilt feature used in the weights has a very low weight for all three contrasts. The correlations and importance results both suggest that H1* – H2* (as calculated over various time spans) is the most useful of the Spectral Tilt features; the one

used in the weights is H2* – H4*. I suspect the weights would paint a different picture if H1* – H2* were used instead. Many other studies have also identified H1* – H2* as being useful in distinguishing Hmong phonation types, including Andruski and Ratliff (2000), Esposito (2012), and Keating et al. (2011).

Energy, measured here as Root Mean Squared Energy (RMSE), is typically lower in non-modal segments than in modal ones. RMSE comes up several times in the Hmong data. In both the correlations and the weights, the relationship is stronger between phonation and RMSE in the creaky vs. modal contrast than in the breathy vs. modal contrast (or the breathy vs. creaky contrast), suggesting that Hmong breathy voice has more energy than creaky voice, and modal voice has the most. One RMSE feature comes in tenth place in the importance rankings, and ablating all four members of the category causes the third largest drop in weighted F1 score for the SVM and the fourth largest for the Random Forest. RMSE in the final third of the vowel appears to be the most useful in distinguishing phonation types. I was only able to find one instance of Energy measures used to study Hmong phonation, but unsuccessfully so (Keating et al., 2011).

Jitter, which measures cycle-to-cycle variation in f_0, is often excluded from studies of phonation, but its inclusion for Hmong appears fairly valuable. Jitter features are among the few features associated with creaky voicing compared to both breathy and modal voicing, though their correlations are weak. The same pattern emerges in the weights. This all suggests that Jitter is useful for distinguishing Hmong's two types of non-modal phonation, which perhaps explains why the two Jitter features are ranked first and second by the Random Forest. However, Jitter is unremarkable in the ablation testing. Andruski and Ratliff (2000) also found Jitter to be greatest in creaky voicing in Hmong.

Finally, **Variance of Pitch Tracks** is ranked sixth in terms of Random Forest importance but does not stand out in the correlations, weights or ablation. In the correlations, VoPT provides the third most strongly correlated feature in the creaky vs. modal contrast (associated with creaky voicing), but this does not hold for the weights. Its ablation barely impacts the Random Forest but causes a small *increase* in the weighted

F1 score of the SVM. Though VoPT has been an important feature for both English and Gujarati, it's usefulness in Hmong is less clear.

In previous chapters, I identify three tiers of categories and examine how the classifiers perform using just each tier. With a less clear sense of their relative importance, I'll treat the frequency of each category in the lists as my groupings: HNR, followed by SHR, Spectral Tilt and Energy, followed by Jitter and VoPT.

Table 8.9: Hmong Weighted F1 Using Subsets of Features

Feature Categories	Weighted F1	
	SVM	Random Forest
Baseline (all features)	*0.6444*	*0.66723*
HNR, SHR, Spectral Tilt, RMSE, Jitter, VoPT	0.64403	0.65649
HNR	0.57753	0.56453
SHR, Spectral Tilt, RMSE	0.57568	0.60916
Jitter, VoPT	0.50049	0.48962

Training both classifiers on features from only these six categories causes only very small drops in the weighted F1 scores, showing that this subset is (relatively) effective in distinguishing Hmong's three phonation types. Even though HNR features are identified as important according to all four diagnostics, HNR does not stand alone; both classifiers suffer from a large drop in weighted F1 score when HNR features are the only features. That said, the classifiers perform similarly when trained on three feature categories: SHR, Spectral Tilt, and RMSE; the fact that one single category can do almost as much work as three suggests that HNR is, in fact, quite important. Finally, another large drop comes from training the classifiers on just Jitter features and VoPT. These six feature categories provide different pieces of information and suggest that Hmong phonation types are differentiated by their periodicity, their intensity, and the speed and completeness of glottal closure.

Chapter 9

MANDARIN

This chapter focuses on the results of an SVM and a Random Forest trained on the Mandarin data. Mandarin has a bit of a different status than the other languages in this dissertation. First, its non-modal phonation is induced by its third (low dipping) tone. I treat all third tones as creaky, exclude all fourth tones (which inconsistently introduce creaky voicing), and treat all first and second tones as modal. Second, the data consist of just four words, all of which consist of the same phonemes. Finally, the data set is quite small – only 180 tokens (before resampling). The distribution of these 180 tokens among the phonation classes is listed in Table 9.1.

Table 9.1: Mandarin Phonation Distribution

Type	Count	*Percent*
B	0	*0.0%*
M	120	*66.667%*
C	60	*33.333%*
Total	180	

Before analysis, the Mandarin data, described in more detail in Chapter 4, have undergone several pre-processing steps. Features with 15% or more undefined measures have been excluded, and remaining undefined measures are replaced with the mean of the class in the training data and the overall mean in the testing data. The data have been normalized and f_0 features have been excluded, as Mandarin tones would likely be enough to classify voice quality on their own. These pre-processing steps leave 92 features from ten categories, listed in Table 9.2.

Table 9.2: Mandarin Features

Feature Category	Number of Features
Spectral Tilt	28
CPP	4
Energy	4
HNR	16
SHR	2
F1	4
Duration	1
Jitter	16
Shimmer	16
VoPT	1
Total	92

In this section, I report the performance of a Support Vector Machine and a Random Forest, each trained on an imbalanced and a resampled data set. Table 9.3 lists the accuracy, weighted F1 score, precision, recall, and F-score for each model. The classifiers perform extremely well.

Table 9.3: Mandarin Classifier Performance

Balance	clf	Accuracy	Weighted F1	Precision		Recall		F-Score	
				M	C	M	C	M	C
Imbalanced	SVM	96.111	0.96119	0.97	0.93	0.97	0.95	0.97	0.94
	RF	95.556	0.95515	0.95	0.96	0.98	0.9	0.97	0.93
Resampled	SVM	96.111	0.96119	0.97	0.93	0.97	0.95	0.97	0.94
	RF	96.111	0.96134	0.98	0.92	0.96	0.97	0.97	0.94

Despite the imbalanced data set, resampling has little to no impact on the model's performance. Adding synthesized creaky tokens provides the classifiers with more information, but that additional information doesn't add much, if anything. This suggests a high degree of consistency between instances of a given class, and enough information for the classifiers to discover clear divisions between them.

The four classifiers perform very similarly. The imbalanced Random Forest has the lowest weighted F1 score (0.95515) and the resampled Random Forest has the highest (0.96134). Precision and recall are higher for modal vowels than for creaky vowels in three of the four classifiers. A weighted F1 score of about 0.96 is quite impressive, and the breakdown by class is equally good.

I'm somewhat surprised by the overall excellent peformance of the Mandarin classifiers for two reasons. First, the data set is extremely small: 180 tokens before resampling, and 240 tokens after resampling. Using a small data set in machine learning makes outliers dangerous and can introduce a lot of noise. Training on small data sets also runs the risk of overfitting the model. Second, I conflate tone and phonation, treating all third tones as creaky. My impression, having listened to all 180 tokens, is that third tones are typically creaky, and this is corroborated by the literature (Davison, 1991); however, it is still possible that not all third tone vowels are creaky and not all first and second tone vowels are modal.

That said, there are some things working in the classifiers' favor. Both SVMs and Random Forests are relatively robust when given small data sets. Additionally, this data set has relatively little noise because all tokens are the same vowel in the same phonetic context, /ma/ with different tones. These factors seem to balance themselves out and result in high performing classifiers.

9.1 Correlations

Correlations provide the first look at the relationship between features and Mandarin phonation. Table 9.4 lists the ten features most strongly correlated with creaky vs. modal voice in Mandarin; correlations for the full set of 94 features is listed in Appendix K and plotted in Figure 9.1.

As expected given the high accuracy of the classifiers, the Mandarin correlations are, overall, much stronger than the correlations for other other languages. Thirty nine of the 94 features are strongly correlated with the creaky vs. modal contrast in Mandarin.

HNR (red) features are strongly correlated with modal voicing in Mandarin, indicating,

Table 9.4: Mandarin Top Feature Correlations

Creaky vs. Modal	
Feature	Correlation
HNR05_Mean	-0.845
HNR15_Mean	-0.837
HNR05_2	-0.828
HNR15_2	-0.820
HNR25_Mean	-0.818
HNR35_Mean	-0.798
HNR25_2	-0.797
HNR35_2	-0.774
Local_Jitter_Mean	0.707
RMS_Energy_2	-0.704

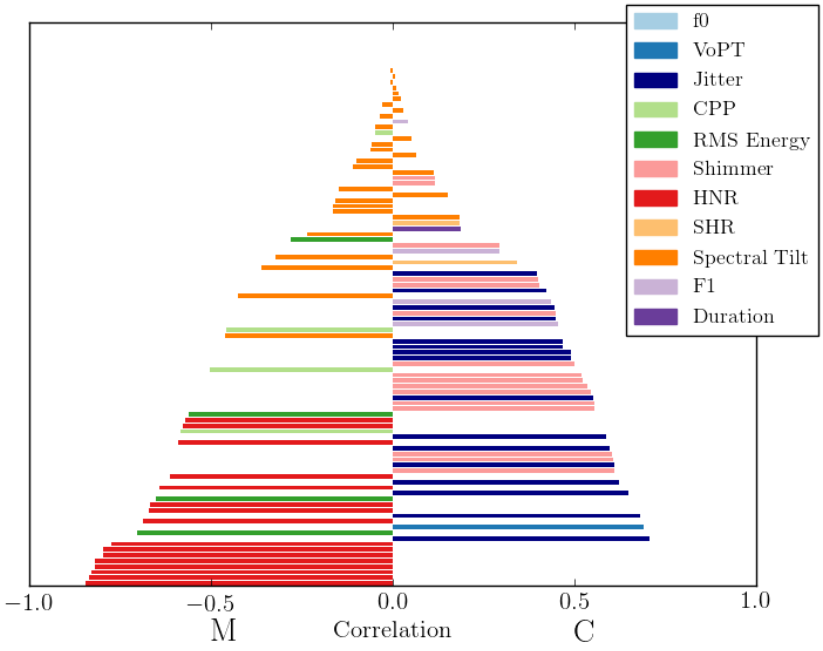


Figure 9.1: Mandarin Feature Correlations

as expected, that modal segments are more periodic than creaky segments. The top HNR measures are measured over either the middle third of the vowel or the entire vowel, suggesting that differences in Mandarin phonation are strongest around the middle of the vowel; this is consistent my observations in listening to the tokens. The frequency bands of the top HNR measures vary, but the lower bands seem to have a higher correlation with creaky voice than the higher bands.

RMS Energy (dark green) is also strongly correlated with modal voicing. Intensity is typically higher in modal segments than in creaky ones because the vocal folds spend more time open, increasing airflow. Like HNR, increased RMS Energy is most strongly correlated with phonation types during the middle portion of the vowel.

Jitter (dark blue) is strongly correlated with Mandarin creaky voice. Local Jitter and Local Absolute Jitter (two very similar measurements) calculated over the entire vowel and the first third account for the four strongest correlated Jitter features; this is somewhat different from HNR and RMS Energy, which have the strongest correlations when calculated over the middle of the vowel. While the middle third of the vowel does not emerge as the most diagnostic in terms of Jitter, the final third emerges as the least diagnostic – the four least correlated Jitter features are each of the four calculations taken over the final third of the vowel.

Variance of Pitch Tracks (bright blue) is not in the top ten most correlated features but is nonetheless strongly correlated with Mandarin creaky voicing. Note that VoPT does not measure f_0, so its strong correlation cannot be attributed to the inherent link between Mandarin tone and phonation. Instead, VoPT is higher when pitch tracking algorithms disagree, which is often the case in aperiodic signals.

Shimmer (pink) also provides many measures that are strongly correlated with creaky voice. Like Jitter, Shimmer is a measure of variability between cycles; Shimmer measures cycle-to-cycle intensity changes. Shimmer is rarely used in studies of phonation, so I am rather surprised to see how strongly correlated it is with Mandarin creaky voice.

A few other feature categories are moderately correlated with Mandarin phonation – CPP

and Spectral Tilt with modal voicing, and F1 with creaky voicing – though their correlations are quite a bit lower than those at the top of the list. Quite a few categories have strong correlations with Mandarin phonation: HNR, Jitter, RMS Energy, VoPT, and Shimmer.

9.2 SVM Weights

As seen in previous chapters, the set of features must be pared down to eliminate redundancy before investigating SVM weights. Using the single feature from each category that received the largest weight in the SVM using all features has a lower but still excellent weighted F1 score of 0.9336. The weights for each feature go in a direction and magnitude consistent with, the correlations, with previous studies phonation, and with the sign of the mean weight per category, showing that they were not erroneously assigned an opposite weight due to multicollinearity. The weights, based on an SVM trained on just these ten features, are shown in Figure 9.2 and listed in Table 9.5.

Table 9.5: Mandarin Feature Weights

Feature	Weight
VoPT	-1.677
RMSE_3	1.263
H1* — H2*_3	1.037
APQ5_Shimmer_2	-0.933
F1_2	-0.87
SHR_Mean	-0.614
PPQ5_Jitter_3	-0.539
HNR35_3	0.335
Duration	-0.257
CPP_2	0.187

Variance of Pitch Tracks (bright blue) has the largest weight and is associated with modal voicing. That's followed by RMS Energy (dark green) and Spectral Tilt (dark orange), both higher in modal vowels. Larger Shimmer (pink) and F1 (light purple) are associated with creaky voicing. Notably low on the list are HNR and Jitter, both of which ranked highly in the correlations.

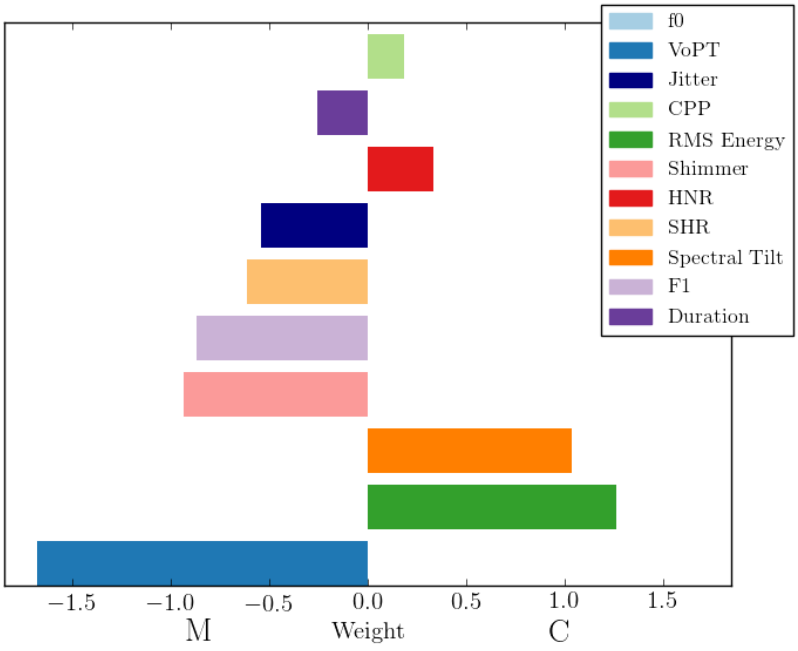


Figure 9.2: Mandarin Feature Weights

The weights identify a somewhat different set of most important feature categories from the correlations: VoPT, RMS Energy, and Spectral Tilt.

9.3 Random Forest Importance

This section reports the Random Forest Importance for Mandarin. The ten most important features are listed in Table 9.6. The weights of all features are plotted in Figure 9.3 and are listed in Appendix K. Recall that unlike correlations and weights, importance doesn't tell us if a feature's value is typically higher or lower for a given phonation type; it simply tells us how important each feature is to the classification task.

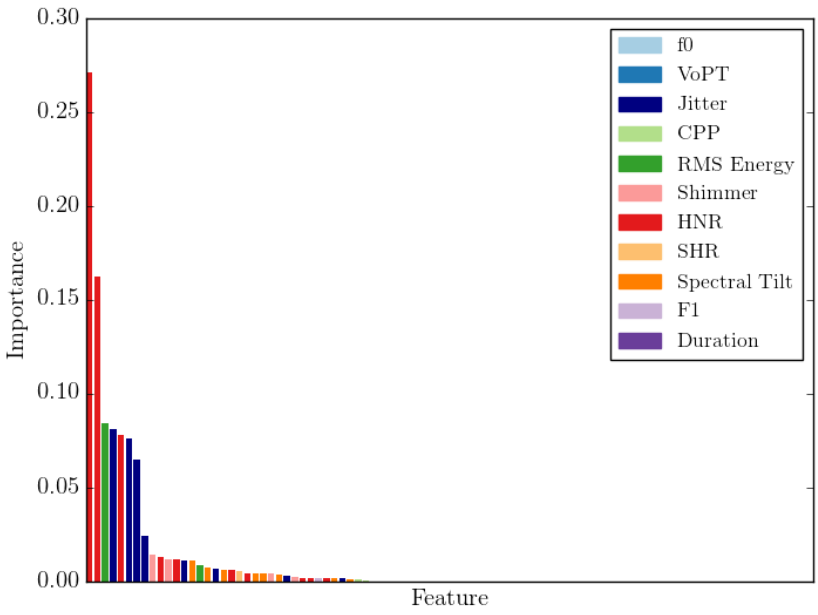


Figure 9.3: Mandarin Feature Importance

The shape of Figure 9.3 is strikingly different from the importance figures we've seen so far – there are two clear peaks followed by two clusters of similar importance, and the remaining features have an importance of nearly zero.

Table 9.6: Mandarin Top Feature Importance

Feature	Importance
HNR05_Mean	0.271200
HNR05_2	0.162460
RMS_Energy_2	0.084177
Local_Jitter_2	0.081498
HNR25_Mean	0.078378
Local_Abs._Jitter_Mean	0.076215
Local_Abs._Jitter_1	0.065285
RAP_Jitter_2	0.024645
APQ5_Shimmer_Mean	0.014522
HNR15_Mean	0.012843

Those two peaks are both measures of HNR (red) – HNR05 calculated over the mean and the vowel's middle third. The measurement made over the mean is nearly twice as important as the one made over the middle third. HNR05_Mean is also the feature most strongly correlated with Mandarin's phonation contrast. Based on the correlations, we know that HNR is larger in Mandarin modal voicing than in creaky voicing.

The cluster of similarly important features just below those two peaks consists of RMS Energy (dark green), Jitter (dark blue), and another HNR measure (red). This is, overall, consistent with the picture painted by the correlations. However, the overall shape of the importance plot does not match the overall shape of the correlations plot – the importances plot shows that most of the work is being done by just a few features, while the correlations plot shows that a large number of features are strongly correlated with phonation. The shortlist of feature categories that are highly ranked according to the Random Forest is HNR, RMS Energy, and Jitter.

9.4 Ablation

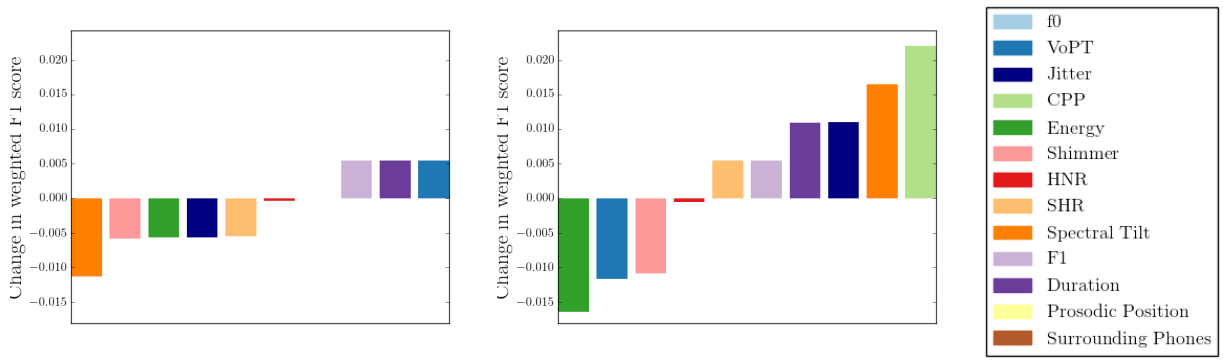
The final way to examine how features contribute to the classification task is by removing them and retraining the classifier. I first ablate categories one at a time, and then iteratively. Recall that I ablate entire categories because they contain many collinear features that, when

ablated individually, won't impact the model much.

Table 9.7 lists the change in weighted F1 score for both classifiers when all features from a category are removed; Figure 9.4 plots the same information.

Table 9.7: Mandarin Category Ablation

Feature	SVM		Random Forest	
	Accuracy	Change	Accuracy	Change
Spectral Tilt	0.94989	-0.0113	0.97787	0.01653
CPP	0.96119	0	0.98343	0.02209
RMS Energy	0.95556	-0.00563	0.94487	-0.01647
HNR	0.96085	-0.00034	0.96085	-0.00049
SHR	0.95573	-0.00546	0.9668	0.00546
F1	0.96667	0.00548	0.9668	0.00546
Duration	0.96667	0.00548	0.97228	0.01094
Jitter	0.95556	-0.00563	0.97239	0.01105
Shimmer	0.95536	-0.00583	0.95047	-0.01087
VoPT	0.96667	0.00548	0.94967	-0.01167



(a) SVM (b) Random Forest

Figure 9.4: Mandarin Category Ablation

Here, we see only very small changes in the weighted F1 score. This, along with the many strong correlations, suggests that there are multiple robust cues to Mandarin phonation that work fairly independently of one another. The SVM is most negatively impacted by ablating

Spectral Tilt features (dark orange), followed by Shimmer (pink), RMS Energy (dark green), Jitter (dark blue), and SHR (light orange). The Random Forest is most negatively impacted by ablating RMS Energy (dark green), followed by VoPT (dark blue) and Shimmer (pink). Random Forests inherently include some degree of randomness, so the fact that five features boost the weighted F1 score here is not necessarily meaningful.

Iterative ablation involves permanently excluding whichever feature causes the largest drop in weighted F1 score each round. The results of this ablation are listed in Table 9.8.

Table 9.8: Mandarin Iterative Category Ablation (SVM)

Category	Weighted F1
Spectral Tilt	0.94989
HNR	0.93333
VoPT	0.92762
Jitter	0.91073
RMSE	0.87195
SHR	0.81256
Duration	0.82994
Shimmer	0.83087
CPP	0.77206
F1	–

The first categories to be ablated are unsurprising: Spectral Tilt, HNR, VoPT, Jitter, and RMS Energy. No category's ablation causes a particularly large drop, consistent with the idea that Mandarin phonation is well described by multiple acoustic properties. I am surprised to see Shimmer so low on the list, based on its usefulness in correlations, weights, and previous ablation.

These two forms of ablation present a similar short list of important feature categories to those we've seen so far: Spectral Tilt, HNR, VoPT, Jitter, and RMSE – are familiar.

9.5 Summary

The Mandarin classifiers perform extremely well, as many of the features appear to distinguishing Mandarin modal and creaky voice. Phonation in Mandarin is optional and secondary to tone contours, so f_0 features are not considered. A consistent group of other features is identified as important to distinguishing phonation through correlations, SVM weights, Random Forest importance, and ablation. These six features, described below, are HNR, Jitter, RMS Energy, VoPT, Shimmer, and Spectral Tilt.

Harmonics-to-Noise Ratio features are among the top performing features according to correlations, importance, and ablation. The aperiodic nature of non-modal voicing means that it contains more noise relative to the harmonic structure than modal voicing does. All HNR features are strongly correlated with modal voicing, two are extremely important and one is moderately important, and it's ablated second in iterative ablation. The highest ranked HNR features are in the lower frequencies (0-500 Hz and 0-1500 Hz) and measured in the middle third or mean of the vowel; this suggests that noise is concentrated in lower frequencies and in the middle of the vowel.

Jitter measures cycle-to-cycle variation in f_0, with a higher Jitter value representing irregularly spaced pitch periods. As such, larger Jitter indicates creaky voicing. Many Jitter features are strongly correlated with creaky voicing and are in the second tier of important features. While Jitter is the fourth category to be ablated in iterative ablation, ablating it individually causes a rather small drop in performance for the SVM and an improvement in performance for the Random Forest. It also receives a low weight according to the SVM. Local and Local Absolute Jitter calculated over the entire vowel and its middle third are generally the highest ranked Jitter features. Cao et al. (2012) examined how voice quality impacts Mandarin tone perception. They used synthesized speech and manipulated Jitter (as well as intensity) to make creaky tokens. They found that this creakiness is "not a direct perceptual cue to Tone 3 identification" but that native speakers of both English and Mandarin identified tokens as Tone 2 when the creaky voicing was closer to the pitch drop.

RMS Energy measures intensity, which is associated with modal voicing. Three of the four RMS Energy features are strongly correlated with modal voicing; the exception is RMSE_3. RMSE_3, however, is the one feature from its category included in the weights, and it has the second largest of all the weights. RMSE_2 is the third most important feature (the two most important are HNR), and remaining features are not important. Ablating all four RMSE features leads to the largest drop in weighted F1 for the Random Forest and the third largest for the SVM; it's removed fifth during iterative ablation. All this suggests that the intensity difference in Mandarin modal and creaky voicing is rather salient, particularly around the middle of the vowel. While intensity is generally not needed to describe Mandarin tones, it's widely acknowledged that Tone 3 has a drop in energy (Cao et al., 2012; Davison, 1991). Additionally, Mixdorff et al. (2005) found that intensity likely aided native Mandarin speakers in identifying tones that are devoiced (removing f_0 information).

Variance of Pitch Tracks, which measures pitch tracking errors that are often caused by irregular f_0, is associated with non-modal voicing. VoPT is ranked eleventh in the correlations and its correlation is strong. It has the largest weight, but its importance is zero; I interpret this as meaning that while it's important, the Random Forest had already encountered features that provided similar information by the time it got to VoPT. This is corroborated by ablation: ablating VoPT causes the second largest drop in weighted F1 score for the Random Forest.

Shimmer, like Jitter, measures irregularity, but of intensity rather than f_0. Many of the Shimmer features are strongly correlated with creaky voice, though they're weaker than HNR, Jitter, RMSE, and VoPT. Shimmer has the fourth largest weight but is barely important to the Random Forest. However, ablating it causes the second largest drop in performance for the SVM and third for the Random Forest. Like Jitter, Shimmer best captures differences in phonation type during the middle and mean of the vowel.

Spectral Tilt is often used to describe phonation types, as the drop off in harmonics depends on the speed and completeness of glottal closure; it's higher in modal voice than in creaky voice. Here, its success in distinguishing Mandarin phonation types is less

clear. A handful of Spectral Tilt features have strong correlations with phonation type and the remaining correlations are low. It's ranked third by SVM weights and very low by Random Forest importance. Ablating it causes the largest drop for the SVM but the second largest increase for the Random Forest. Davison (1991) found that Spectral Tilt "serves to distinguish between tonal categories" in Tianjin Mandarin (a different dialect from the one used here), particularly for distinguishing tones one and two (considered here to be modal) from three (considered here to be creaky). Belotel-Grenie and Grenie (1994) also found Spectral Tilt to be the most reliable way to distinguish Mandarin phonation types.

Features from many of these categories appear to fairly effectively distinguish Mandarin modal voice from creaky voice. My final look at these six feature categories is to first train a classifier using just the features from these categories, and then to train a classifier on a single category at a time. The results are presented in Table 9.9.

Table 9.9: Mandarin Weighted F1 Using Subsets of Features

Feature Categories	Weighted F1	
	SVM	Random Forest
Baseline (all features)	*0.96119*	*0.96134*
HNR, Jitter, RMSE, VoPT, Shimmer, Tilt	0.96103	0.96134
HNR	0.94487	0.92282
Jitter	0.93848	0.89466
RMSE	0.91209	0.91716
VoPT	0.83126	0.83652
Shimmer	0.84968	0.79702
Tilt	0.78723	0.80799

Training the classifiers on features from just these six categories barely changes the SVM's performance and does not impact the Random Forest's performance. When the classifier are trained on just one of those categories at a time, HNR does best. I didn't claim to have ordered these six features because I didn't feel confident in their ordering, but I did list them in order of my roughly estimated importance. Looking at Table 9.9, my ordering was more or less correct - training on just HNR is best, and training on just Spectral Tilt is worst.

Overall, these six categories effectively distinguish Mandarin voice qualities.

Chapter 10

MAZATEC

Here, I examine the features that the classifiers rely on to identify Mazatec phonation types. Mazatec, described in more detail in Chapter 4, contrasts breathy, modal, and creaky voice. The data set contains similar numbers of modal and creaky tokens, but many fewer breathy tokens; the distribution of the data set is listed in Table 10.1.

Table 10.1: Mazatec Phonation Distribution

Type	Count	*Percent*
B	70	*14.523%*
M	195	*40.456%*
C	217	*45.021%*
Total	482	

Like the other data sets, the Mazatec data have been normalized. Features with 15% or more undefined measures have been removed, and remaining missing values will be replaced. This leaves 105 features from ten categories, as shown in Table 10.2. These 105 features will be used to train an SVM and a Random Forest; the performance of these two classifiers is reported in the following section.

10.1 Model Performance

Table 10.3 presents the accuracy, weighted F1 score, precision, recall, and F1 score of four classifiers: an SVM and a Random Forest, trained once on an imbalanced data set and once on a resampled data set.

The four classifiers have similar weighted F1 scores, ranging from 0.66891 (resampled

Table 10.2: Mazatec Features

Feature Category	Number of Features
Spectral Tilt	28
CPP	4
Energy	4
HNR	14
f_0	16
F1	4
Duration	1
Jitter	16
Shimmer	15
VoPT	1
Total	105

Table 10.3: Mazatec Classifier Performance

Balance	clf	Accuracy	Weighted F1	Precision			Recall			F1 Score		
				B	M	C	B	M	C	B	M	C
Imbalanced	SVM	68.88	0.69378	0.45	0.71	0.78	0.6	0.66	0.75	0.51	0.68	0.76
	RF	70.747	0.69786	0.44	0.68	0.79	0.29	0.76	0.8	0.35	0.71	0.8
Resampled	SVM	69.502	0.6994	0.45	0.69	0.8	0.57	0.67	0.76	0.51	0.68	0.78
	RF	66.598	0.66891	0.35	0.65	0.79	0.41	0.61	0.8	0.38	0.63	0.8

Random Forest) to 0.6994 (resampled SVM). All four perform best according to both precision and recall on creaky voice, followed by modal voice, and then by breathy voice; this matches the makeup of the data set shown in Table 10.1.

The most striking difference between classifiers is between the imbalanced SVM and Random Forest. The SVM has much higher breathy recall (0.6) than the Random Forest (0.29). For the two classifiers using resampled data, the SVM still performs better, though with a smaller difference (0.57 vs. 0.41). I interpret this to mean that the SVMs are capturing something (perhaps an interaction between features) that helps them recognize breathy vowels better than the Random Forests.

The weighted F1 scores are below 0.7 for the four Mazatec classifiers. While they're large

enough to indicate some patterns in the data, those patterns are likely not strong. Mazatec uses phonation types contrastively, so it's important that *listeners* can find patterns that help identify voice quality. Those patterns may not be coming through strongly here if they're not well captured by the set of features I'm using. It's also possible that patterns would emerge more clearly in a larger data set. However, they have a high enough weighted F1 score to more forward and examine what role the features play in distinguishing Mazatec phonation types.

10.2 Correlations

This section reviews each feature's correlation with Mazatec phonation types. Table 10.4 lists the ten top correlated features for those three contrasts; the correlations for all 105 features are listed in Appendix L. The following three sections review patterns in correlations for the three contrasts.

Table 10.4: Mazatec Top Feature Correlations

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
Feature	Correlation	Feature	Correlation	Feature	Correlation
H1* – A1*_1	0.547	SHR_f_0_1	-0.33	H1* – H2*_2	-0.505
H1* – A2*_1	0.429	Praat_f_0_1	-0.323	H1* – H2*_Mean	-0.484
H1* – A2*_2	0.421	H1* – H2*_3	-0.294	H1* – A2*_2	-0.474
SHR_f_0_1	-0.41	SHR_f_0_Mean	-0.294	H1* – A2*_Mean	-0.446
H1* – A1*_Mean	0.387	Praat_f_0_Mean	-0.285	H1* – A1*_Mean	-0.436
H1* – A2*_Mean	0.385	STRAIGHT_f_0_1	-0.279	H1* – A1*_2	-0.426
H1* – A3*_1	0.385	SHR_f_0_3	-0.255	VoPT	0.422
H1* – H2*_2	0.383	CPP_3	0.243	H1* – A1*_1	-0.406
H1* – A1*_2	0.363	Snack_f_0_1	-0.239	H1* – H2*_3	-0.392
H1* – H2*_1	0.357	SHR_f_0_2	-0.233	H1* – A3*_2	-0.389

10.2.1 Breathy vs. Creaky Correlations

Figure 10.1 shows the strength of each feature's correlation with either breathy or creaky voice in Mazatec. Each bar represents a feature, color coded by category. Bars with positive

values are correlated with breathy voicing, and bars with negative values are correlated with creaky voicing; the magnitude of the correlation indicates the strength of the relationship between the feature and the phonation type.

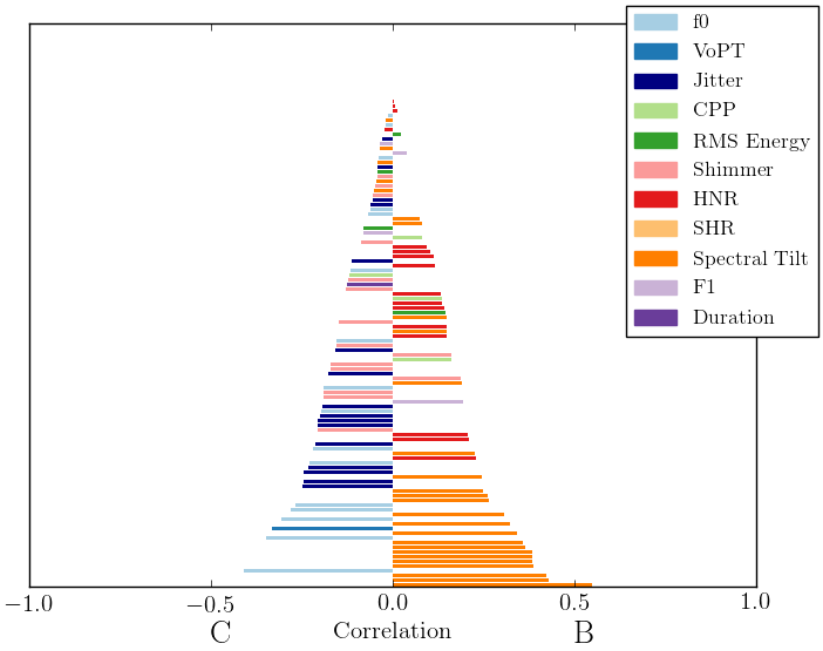


Figure 10.1: Mazatec Feature Correlations, B vs. C

Spectral Tilt (dark orange) accounts for nine of the top ten features, including the only feature whose correlation is strong. Larger Spectral Tilt is associated with breathy voicing, as the energy drops off more steeply in higher frequencies for breathy voicing than for other phonation types due to slower and less complete glottal closure. H1* – A1* provides the four strongest correlated Spectral Tilt measures. Features calculated over the first or second third of the vowel are much more strongly correlated with phonation than those calculated over the vowel's final third.

Increased f_0 (light blue) is correlated with creaky voicing; several of the f_0 measurements provide a moderate correlation with phonation. While Jalapa Mazatec has three tones – ˩,

˧, and ˥ – those tones are not inherently linked with phonation types. Imbalance in the tones included in the data set could account for this, or there could be a subtle difference in how the tones are realized depending on the voice quality. The SHR f_0 algorithm accounts for the three strongest f_0 correlations. Unlike Spectral Tilt, the f_0 correlations do not reveal that Mazatec's phonation tends to occur vowel-initially; the third strongest f_0 correlation is measured over the final third of the vowel.

Variance of Pitch Tracks (bright blue) is moderately correlated with creaky voicing in Mazatec. Its moderate correlation here suggests that the aperiodicity of Mazatec creaky voicing poses more problems for pitch tracking algorithms than breathy voicing.

Delving into the weaker correlations, Jitter (dark blue) and Shimmer (pink) are associated with creaky voicing and HNR (red) with breathy voicing. All three of these measures are more often used to distinguish modal voicing from non-modal voicing, rather than to distinguish between the two types of non-modal phonation. But the fact that they appear here – though not strongly – suggests they they differ between breathy and creaky voicing as well. Specifically, they suggest that breathy voicing is more periodic than creaky voicing, as it has stronger harmonics relative to noise and less cycle-to-cycle variation.

10.2.2 Breathy vs. Modal Correlations

Figure 10.2 shows each feature's correlation with the breathy vs. modal contrast in Mazatec. These correlations are, overall, lower than for the previous contrast; no correlations are strong and only two are moderate.

f_0 measures (light blue) account for the strongest correlations, including the two moderate correlations. Increased f_0 is correlated with modal voice. Seven of the top ten correlations are f_0, and four of those seven are measured in the vowel's first third.

Spectral Tilt (dark orange) provides the third most strongly correlated feature, H1* – H2*_3, which is correlated with modal voicing. However, another Spectral Tilt feature, H1* – A1*_1, has a similar magnitude of correlation, but is correlated instead with breathy voicing. Similarly surprising is CPP, which is counterintuitively correlated with breathy

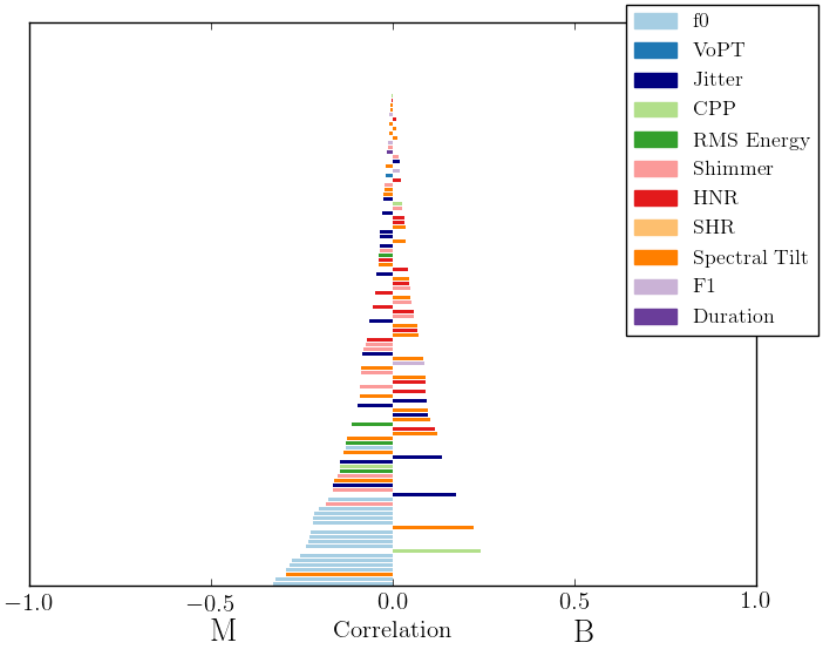


Figure 10.2: Mazatec Feature Correlations, B vs. M

voicing. These correlations are likely too weak to be meaningful.

10.2.3 Creaky vs. Modal Correlations

Finally, the creaky vs. modal contrast has just one strongly correlated feature but several others in the upper range of medium correlations.

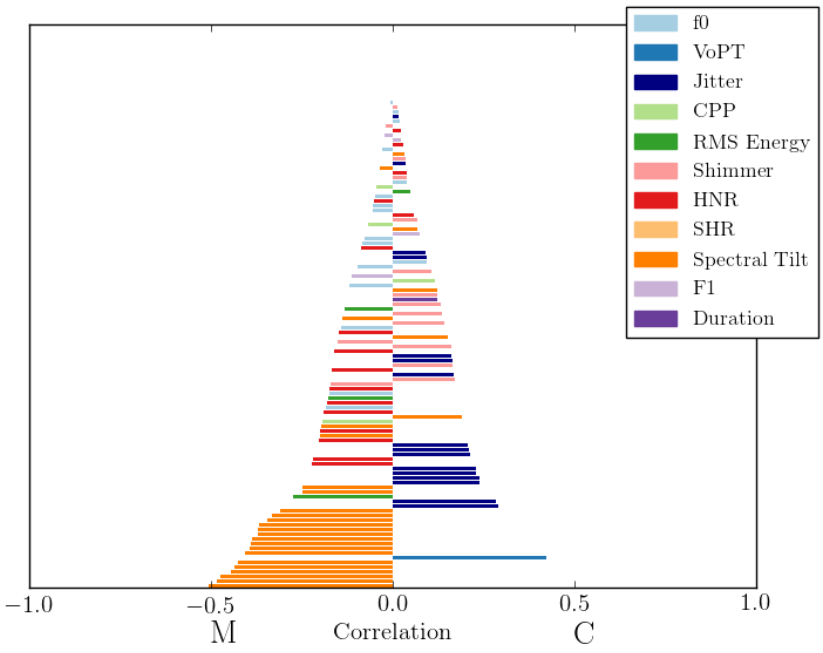


Figure 10.3: Mazatec Feature Correlations, C vs. M

Spectral Tilt (dark orange) again dominates the most strongly correlated features for this contrast; here, it's correlated with modal voicing. Taken together with the other two contrasts, we see that Mazatec phonation types follow the expected pattern of Spectral Tilt: it's largest in breathy voicing, followed by modal voicing and then by creaky voicing. Several Spectral Tilt calculations are among the top features (H1* – H2*, H1* – A2*, and H1* – A1*), and their correlations are strongest when measured in the vowel's middle third or entirety.

Variance of Pitch Tracks (bright blue) is moderately correlated with creaky voicing. This is VoPT's first appearance for Mazatec, suggesting that it captures aperiodicity better in creaky vowels than in breathy vowels.

Jitter (dark blue) is weakly correlated with creaky voicing. The strongest correlations are taken over the vowel's middle third and mean. Other features with weak correlations but noticeable patterns include HNR (red, correlated with modal voicing) and Shimmer (pink, correlated with creaky voicing).

Through these three contrasts, two feature categories emerge as more strongly correlated with phonation types than the others: Spectral Tilt and f_0.

10.3 SVM Weights

I next examine the relationship between features and phonation types using SVM weights. As seen in previous chapters, features that provide very similar information, such as different Jitter calculations or a features measured over the mean and middle third of a vowel, can lead to some odd weights. I avoid this problem by paring down the set of features to just one per category. In previous chapters, I've selected the feature from each category with the strongest weight (for any contrast). For Mazatec, however, that subset of features reduces the SVM's weighted F1 score from 0.6994 to 0.60152, which I consider to be too large a drop. I tried instead picking the feature from each category with the strongest correlation. An SVM trained on this subset has a weighted F1 score of 0.66303, which is much closer to the original. Moreover, nearly all of the weights have the same sign as the category average, indiciating that the features are representative of their categories and were not assigned an opposite weight due to multicollinearity. These weights are listed below and discussed for each contrast.

10.3.1 Breathy vs. Creaky Weights

Figure 10.4 shows the weights of the ten features for the breathy vs. creaky contrast. As in the correlations, each bar represents a feature, its color represents the feature's category, and

Table 10.5: Mazatec Feature Weights

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
SHR_f_0_1	-1.832	SHR_f_0_1	-1.85	VoPT	0.984
H1* – A1*_1	1.666	VoPT	1.077	RMS_Energy_2	-0.927
Local_Jitter_Mean	-1.36	Local_Jitter_Mean	-0.861	CPP_3	0.676
RMS_Energy_2	0.883	H1* – A1*_1	0.742	H1* – A1*_1	-0.568
F1_1	-0.732	CPP_3	0.584	Duration	0.385
Duration	-0.399	Duration	-0.501	HNR25_3	-0.346
APQ3_Shimmer_2	-0.368	APQ3_Shimmer_2	-0.454	SHR_f_0_1	0.271
HNR25_3	0.295	F1_1	-0.253	Local_Jitter_Mean	0.079
VoPT	0.177	RMS_Energy_2	0.189	APQ3_Shimmer_2	0.056
CPP_3	0.028	HNR25_3	-0.035	F1_1	-0.004

the magnitude and sign of each bar represents the feature's weights. Features with positive weights have an increased value in breathy voice, and features with negative weights have an increased value in creaky voice.

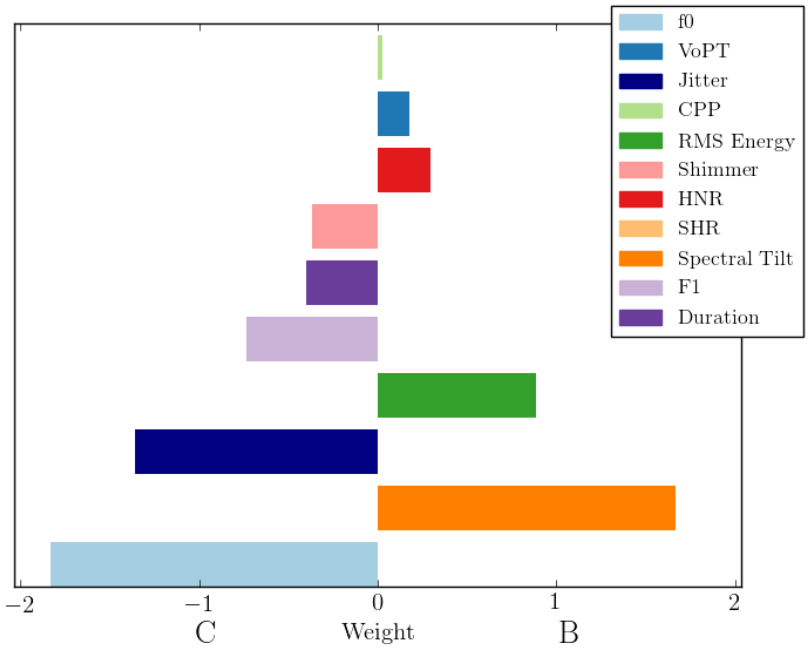


Figure 10.4: Mazatec Feature Weights, B vs. C

Three features have rather large weights: f_0 (light blue) and Jitter (dark blue) are associated with creaky voicing, while Spectral Tilt (dark orange) is associated with breathy voicing. This matches the correlations; though the correlations determined which features are used here, the weights are calculated in an entirely different manner, so the consistency here is meaningful.

10.3.2 Breathy vs. Modal Weights

The weights for the breathy vs. modal contrast are shown in Figure 10.5. These weights are less consistent with the correlations than those in the previous contrast.

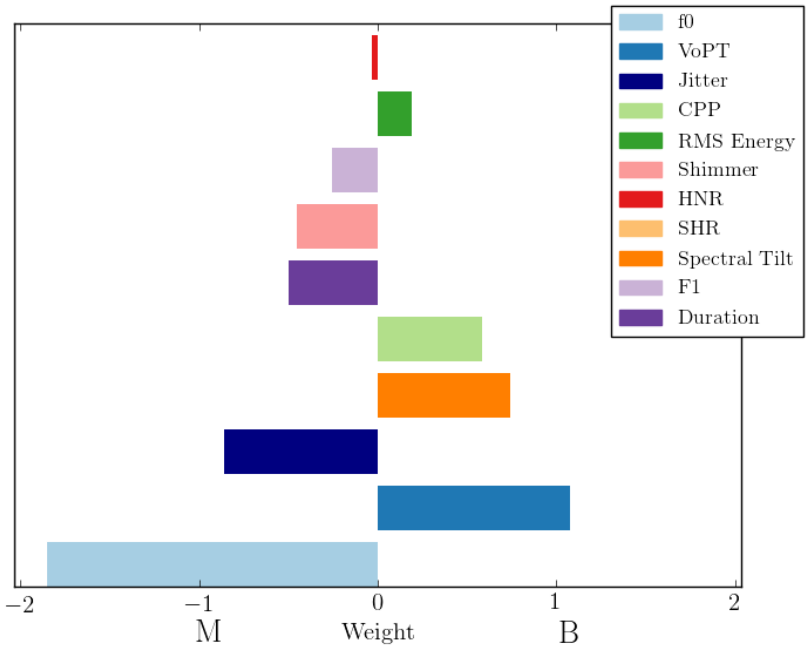


Figure 10.5: Mazatec Feature Weights, B vs. M

f_0 (light blue) again has the largest weight (associated with modal voicing). The next strongest weight, however, comes from VoPT (bright blue), which has an extremely low correlation for this contrast. After VoPT is Jitter (dark blue), associated with modal voicing.

I am skeptical of this, as Jitter is (weakly) correlated with breathy voicing. This may be a fluke of the feature selection process or it may be too small a weight to consider.

10.3.3 Creaky vs. Modal Weights

Finally, Figure 10.6 shows the weights of the ten features for the creaky vs. modal contrast.

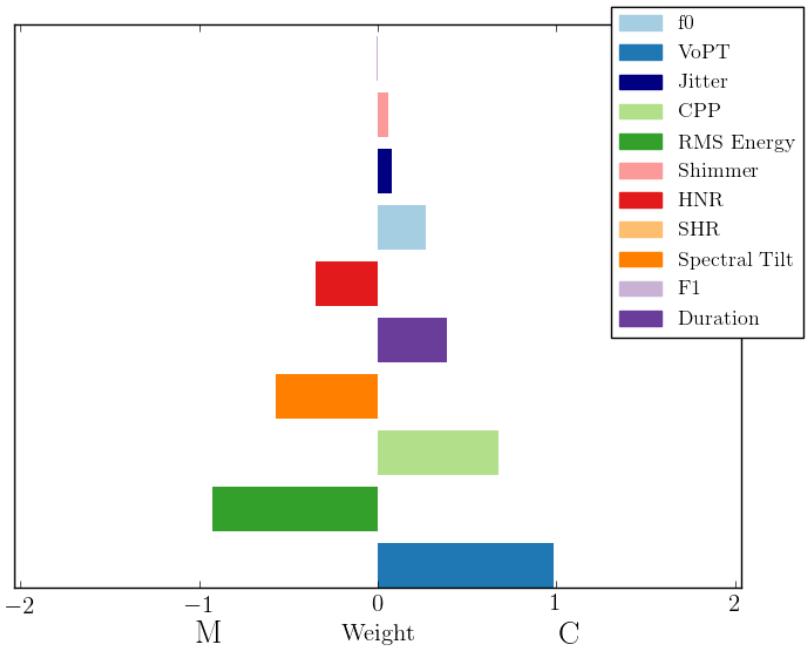


Figure 10.6: Mazatec Feature Weights, C vs. M

The largest weight for this contrast, belonging to VoPT (bright blue), is quite a bit smaller than the largest weight for the other two contrasts; in the previous section, this contrast has similar strength correlations to the breathy vs. creaky contrast and stronger correlations than in the breathy vs. modal contrast. VoPT is associated with creaky voicing, followed by RMS Energy (dark green), which is associated with modal voicing. However, that's followed by CPP (light green), associated with creaky voicing; increased CPP is typically a good indicator of modal voicing, as it represents increased periodicity.

Though a few of the weights seem odd (perhaps they're unrepresentative of their category for some contrasts), the weights of the top tier of features make sense. This top tier of features consists of f_0, Spectral Tilt, Jitter, VoPT, and RMS Energy.

10.4 Random Forest Importance

In this section, I present the Mazatec feature importance from the Random Forest. Importance, unlike correlations and weights, importance does not tell us which phonation type a feature is identifying. Table 10.6 lists the importance values for Mazatec's ten most important features; Figure 10.7 plots the importance of each one, and the importances for each feature are listed in Appendix L.

Table 10.6: Mazatec Top Feature Importance

Feature	Importance
Praat_f_0_1	0.055872856
H1* – A2*_2	0.041961545
SHR_f_0_1	0.035970057
H1* – H2*_1	0.0310014
PPQ5_Jitter_3	0.028933285
H1* – H2*_2	0.027132014
VoPT	0.024038735
H1* – A2*_1	0.02226283
H1* – A1*_2	0.021144679
PPQ5_Jitter_2	0.021050482

The top-performing features are mostly blue and orange: f_0 (light blue), Spectral Tilt (dark orange), Jitter (dark blue), and VoPT (bright blue).

The first and third features, according to importance, are both f_0 measured in the vowel's first third. We've so far seen that increased f_0 is associated with modal voice compared to breathy, and with creaky voice compared to breathy; in other words, breathy voicing has a low f_0 in Mazatec.

Five of the top ten features are measures of Spectral Tilt. In particular, Spectral Tilt calculated in the first and second third of the vowel and using H1* – A2* and H1* – H2*

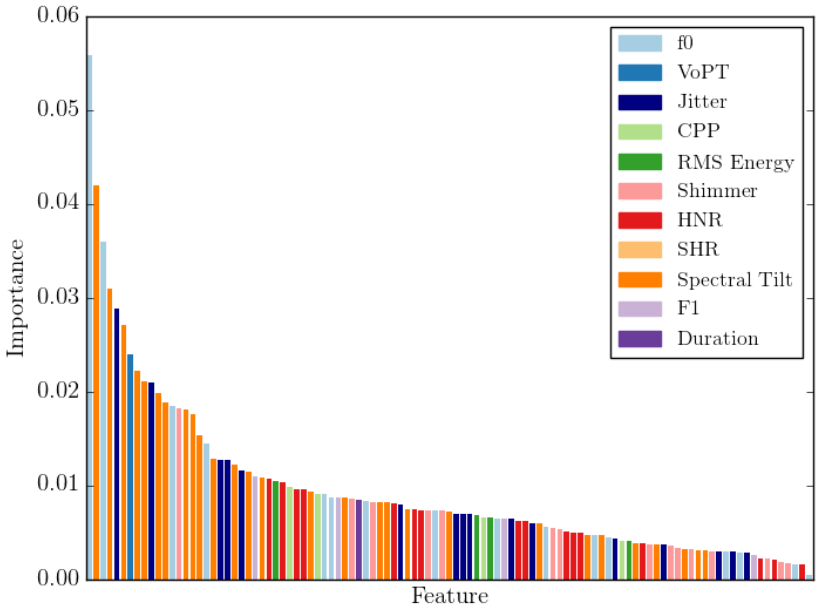


Figure 10.7: Mazatec Feature Importance

are most important. In the correlations and weights, we saw that Spectral Tilt follows the expected pattern: it's greatest in breathy voicing and smallest in creaky voicing.

Two Jitter features and VoPT are also in the top ten. Both have been so far associated with non-modal phonation, particularly with creaky voice. The Random Forest relies on f_0, Spectral Tilt, Jitter, and VoPT most in distinguishing phonation types in Mazatec.

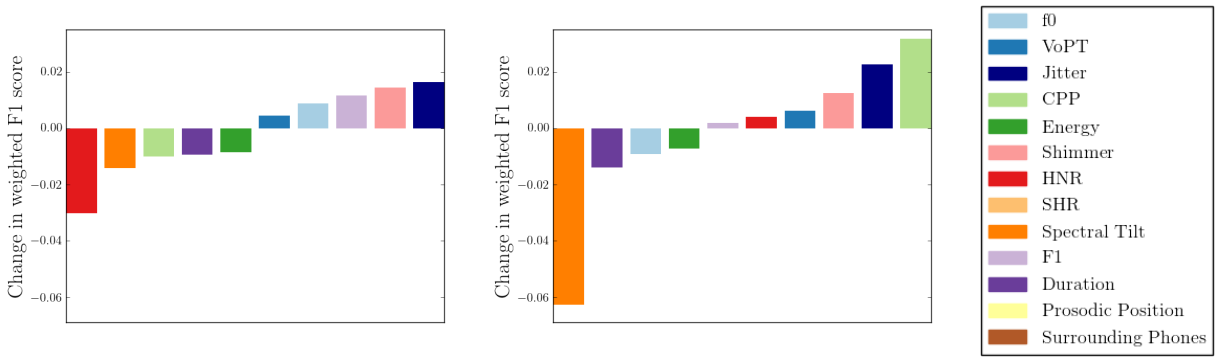
10.5 Ablation

The final way of evaluating how features contribute to distinguishing phonation types is ablation. I first ablate categories one at a time and then iteratively. Recall that I ablate entire categories due to the effects of mulitcollinearity; when multiple features can do the same job, ablating any one of them won't impact the model and therefore won't be informative. Table 10.7 reports weighted F1 score of an SVM and a Random Forest trained without each category, as well as the change in weighted F1 score when each category is excluded. The same information is shown in Figure 10.8.

Table 10.7: Mazatec Category Ablation

Feature	SVM		Random Forest	
	Weighted F1	Change	Weighted F1	Change
Spectral Tilt	0.68527	-0.01413	0.60623	-0.06268
CPP	0.68943	-0.00997	0.70079	0.03188
Energy	0.69092	-0.00848	0.66175	-0.00716
HNR	0.6693	-0.0301	0.67306	0.00415
f_0	0.7083	0.0089	0.65986	-0.00905
F1	0.7111	0.0117	0.67088	0.00197
Duration	0.69004	-0.00936	0.65492	-0.01399
Jitter	0.71592	0.01652	0.69158	0.02267
Shimmer	0.71382	0.01442	0.68153	0.01262
VoPT	0.70402	0.00462	0.67516	0.00625

Like the classifiers for the other languages, neither the SVM nor the Random Forest experiences a large drop from ablating any single category. The SVM suffers most from ablating HNR (red) followed by Spectral Tilt (dark orange). While Spectral Tilt has



(a) SVM (b) Random Forest

Figure 10.8: Mazatec Category Ablation

consistently been identified as important, HNR has not yet come up for Mazatec. The next two are also unexpected: CPP (light green) and Duration (dark purple). The Random Forest experiences the largest drop when Spectral Tilt (dark orange) is ablated, and that drop is much larger than any other.

The second type of ablation tries ablating each category but continues to exclude the category whose exclusion caused the largest drop. This process continues until there's only one category of features left. The results of this ablation, conducted only using an SVM, are reported in Table 10.8.

As HNR causes the largest drop in the categorical ablation, it's inherently ablated first. It's followed by three features that we've seen already for Mazatec: Energy, Spectral Tilt, and f_0. Surprisingly, two major players are at the bottom of the list: VoPT and Jitter.

These two forms of ablation highlight that HNR and Spectral Tilt play a major role in distinguishing Mazatec phonation types. Energy and f_0 also have a role, but less strongly and less consistently than the top two.

Table 10.8: Mazatec Iterative Category Ablation (SVM)

Category	Weighted F1
HNR	0.6693
Energy	0.64282
Spectral Tilt	0.61629
f_0	0.57085
CPP	0.5319
Shimmer	0.48176
Duration	0.46891
VoPT	0.46595
Jitter	0.35683
F1	–

10.6 Summary

This chapter used correlations, SVM weights, Random Forest importance, and two types of ablation to identify a set of feature categories that contribute more to Mazatec voice quality classification than others. These categories, ranked roughly by how often they came up and how important they are, are Spectral Tilt, f_0, Jitter, VoPT, RMS Energy, and HNR.

Spectral Tilt is living up to its reputation as the best measure of phonation types. For Mazatec, it dominates the correlations; it's strongly correlated with breathy compared to creaky voice, and with modal compared to creaky voice. But its correlation is much weaker in the breathy vs. modal contrast. This suggests that the difference in Spectral Tilt is much greater from modal to creaky than from breathy to modal, but the weights don't agree. The SVM weights also rank Spectral Tilt fairly highly; its weight is largest in the breathy vs. creaky contrast, followed by breathy vs. modal, and then by creaky vs. modal. Spectral Tilt features account for five of the ten most important features. Its ablation causes the largest drop for the Random Forest and the second largest for the SVM.

Previous studies (Silverman, 1997; Blankenship, 2002) have found that Mazatec phonation is more prominent early in the vowel. Here, most of the Spectral Tilt features that

perform well are measured over the first or second third of the vowel. Garellek and Keating (2011), considering just the first third of the vowel, used linear discriminant analysis to examine the acoustic properties of Mazatec's three phonation types. They found that three of the five Spectral Tilt features they tested are significant: H1* – H2*, H1* – A1*, H1* – A2*; this is consistent with the Spectral Tilt measures that I found to be most useful to the classifiers.

Lower $\boldsymbol{f}_0$, typically associated with both types of non-modal phonation, appears to be more associated with Mazatec breathy voice than creaky voice. Both correlations and weights show that f_0 is lower in breathy voicing than in modal or creaky voicing. f_0 has the largest weight for both contrasts involving breathy voicing, and accounts for many of the top correlations for those to contrasts. The first and third most important features are f_0 features, and while the category's ablation doesn't much impact either classifier, it's the third to go in iterative ablation. Many but not all of the most important f_0 features are calculated over the first third of the vowel. This follows previous observations that Mazatec's breathy voicing lowers f_0 relative to modal voicing (Gordon and Ladefoged, 2001). Garellek and Keating (2011) also found f_0 to vary significantly between phonation types in Mazatec. Their only f_0 measure was STRAIGHT_f_0_1, which is among the better performing f_0 features here, but not at the top. Interestingly, they found that f_0 was less important than CPP, which was not identified by the classifiers or the correlations as important.

Jitter, though rarely used nowadays in studies of phonation, again proves itself useful, though seemingly less so than Spectral Tilt and f_0. Several Jitter features are weakly to moderately (but consistently) correlated with creaky voicing compared to both breathy and modal voicing. In the weights, Jitter is ranked third in the breathy vs. creaky contrast and the breathy vs. modal contrast (associated with breathy voicing in the latter), suggesting that it's highest in creaky voicing, followed by breathy voicing, and then by modal voicing. The most important Jitter feature is ranked fifth by the Random Forest. However, ablating all Jitter features causes both the SVM and the Random Forest to improve, and it's the second to last category to be ablated in the iterative ablation. Despite its usefulness

according to weights and importance, Jitter is not typically used to measure phonation; in fact, Blankenship (2002) reports that Jitter "did not differentiate the Mazatec phonation types well in the pilot study."

Variance of Pitch Tracks relies on f_0 measures, but is a separate measurement. Rather than directly measuring f_0, it describes how problematic it was to measure f_0 over a segment; it's higher when pitch tracks disagree with one another. It's moderately correlated with creaky voice compared to modal voice but otherwise not remarkable in the correlations. According to the weights, it's quite important for distinguishing modal from non-modal phonation; it has the second largest weight for the breathy vs. modal contrast and the highest weight for the creaky vs. modal contrast. And while it's also ranked seventh according to importance, it's ranked low by both types of ablation.

RMS Energy is in the lowest tier of important features. Typically higher in modal voicing than in breathy or creaky voicing, Energy also differentiates breathy from creaky voice according to the weights; breathy voicing has more Energy than creaky voicing. Its Random Forest importance is low but it's the second feature category to go in the iterative ablation. Though only two of the four tests highlight Energy's role, Keating et al. (2011) found that it's significant; Garellek (2012), however, also measured Mazatec Energy but didn't find its role in distinguishing phonation types to be significant.

Finally, **Harmonics-to-Noise Ratio** only appears once for Mazatec: its ablation causes the largest drop in weighted F1 score for the SVM. Otherwise, HNR is unimportant. The correlations are weak but suggest that it's lower in creaky voice than in breathy or modal. It's at or near the bottom of the list for SVM weights and around the middle of the list for Random Forest importance. Studies of Mazatec phonation generally rely on Cepstral Peak Prominence rather than HNR to examine noise in the signal. Though CPP was generally unremarkable here, several studies have identified it as significant (Garellek, 2012; Keating et al., 2011). Perhaps I have unfairly determined that CPP is unimportant; while HNR shows up once but strongly, CPP shows up several times but towards the middle of the pack. However, both classifiers are impacted more by ablating HNR than CPP.

Features from these six categories – Spectral Tilt, f_0, Jitter, VoPT, RMSE, and HNR – all contribute to Mazatec voice quality classification, though they contribute different amounts of information and different pieces of the puzzle. I've presented them in what I believe to be their ranking, with Spectral Tilt providing the most information and HNR the least.[1] To check my intuitions, I trained a few more classifiers on subsets of feature categories and report their weighted F1 score in Table 10.9.

Table 10.9: Mazatec Weighted F1 Using Subsets of Features

Feature Categories	Weighted F1	
	SVM	Random Forest
Baseline (all features)	*0.6994*	*0.66891*
Spectral Tilt, f_0, Jitter, VoPT, RMSE, HNR	0.70112	0.66118
Spectral Tilt, f_0	0.65413	0.63812
Jitter, VoPT	0.47855	0.51955
RMSE, HNR	0.56133	0.54064

Training on features from just the six categories identified here causes trivial changes to the weighted F1 score for both classifiers; this suggests that this subset of features includes those crucial to the classification task. On their own, Spectral Tilt and f_0 features do remarkably well compared to the baseline. The weighted F1 score decreases considerably with the other two pairs. Though I thought Jitter and VoPT both contributed more than RMSE or HNR, the opposite appears to be true when they're the only features considered by a model.

Overall, these machine learning models suggest that Mazatec's three phonation types differ primarily in speed and completeness of glottal closure and pitch, and to a lesser degree irregularity in glottal pulses, intensity, and noise.

[1] While HNR is last, it's last on the list of *important* feature categories.

Chapter 11

ZAPOTEC

The last of the six languages is Zapotec. This chapter examines which features distinguish its three phonation types. Zapotec, described in more detail in Chapter 4, contrasts breathy, modal, and creaky voice. Tone is also employed; high or rising tone is contrastive on modal vowels and non-modal vowels have falling tone (Esposito, 2010). The Zapotec data set is fairly small but rather balanced, as shown in Table 11.1.

Table 11.1: Zapotec Phonation Distribution

Type	Count	*Percent*
B	117	*34.012%*
M	101	*29.36%*
C	126	*36.628%*
Total	344	

Because of the relationship between phonation and tone, I exclude all f_0 features from consideration for Zapotec. I also exclude features with 15% or more undefined measures; remaining undefined measures are imputed for each fold as the models are trained and tested. Additionally, the data have been normalized (see Chapter 4.) These pre-processing steps leave 74 features from ten categories, shown in Table 11.2, that the classifiers will use to distinguish Zapotec's three phonation types.

11.1 Model Performance

The performance of the two classifiers – the SVM and the Random Forest – based on imbalanced and resampled data sets is listed in Table 11.3. Zapotec's three phonation classes

Table 11.2: Zapotec Features

Feature Category	Number of Features
Spectral Tilt	28
CPP	4
Energy	4
HNR	16
F1	4
SHR	1
Jitter	9
Shimmer	6
VoPT	1
Duration	1
Total	74

are fairly evenly distributed in the data set, so only very small changes appear between the imbalanced and resampled sets.

Table 11.3: Zapotec Classifier Performance

Balance	clf	Accuracy	Weighted F1	Precision			Recall			F1 Score		
				B	M	C	B	M	C	B	M	C
Imbalanced	SVM	69.767	0.6962	0.69	0.7	0.7	0.77	0.71	0.62	0.73	0.71	0.66
	RF	63.663	0.63405	0.7	0.59	0.61	0.77	0.56	0.57	0.73	0.58	0.59
Resampled	SVM	69.767	0.6959	0.69	0.68	0.73	0.79	0.7	0.61	0.73	0.69	0.67
	RF	63.081	0.62859	0.66	0.59	0.63	0.74	0.57	0.57	0.7	0.58	0.6

Using both the imbalanced and resampled data sets, the SVM performs quite a bit better than the Random Forest; the SVMs have a weighted F1 score of about 0.7 and the Random Forests about 0.63. Their precision and recall is similar for breathy voicing, but the SVM handles creaky voicing and particularly modal voicing better than the Random Forest. All four classifiers have higher recall than precision for breathy voicing, suggesting that they're guessing it too often. The opposite is true for creaky voicing; all four have higher precision than recall. In other words, they're doing a better job of accurately identifying creaky voice,

but in doing so are missing some tokens.

I focus here on classifiers trained on resampled data. With a weighted F1 score of just under 0.7, the SVM is finding some patterns in the data, though the confidence in those patterns is questionable. The Random Forest, on the other hand, has a weighted F1 score of 0.63, which is not particularly confidence-inspiring. As the classifiers perform rather poorly for Zapotec, I explored what mistakes they are making. Tables 11.4 and 11.5 are confusion matrices for the SVM and the Random Forest, respectively.

Table 11.4: Zapotec Confusion Matrix, SVM

		Predicted		
		B	M	C
True	B	**92**	13	12
	M	27	**77**	22
	C	15	15	**71**

Table 11.5: Zapotec Confusion Matrix, Random Forest

		Predicted		
		B	M	C
True	B	**88**	19	10
	M	25	**79**	22
	C	12	25	**64**

The confusion matrices do not reveal any systematic patterns in the classifiers' mistakes; for the most part, incorrectly labeled instances are roughly split between the two incorrect classes. This apparent randomness suggests that the classifiers are struggling to find meaningful patterns, rather than finding misleading patterns. This difficulty may be because the features do not successfully describe Zapotec phonation, or because the data set is too small for patterns to emerge. The following four sections explore what those patterns are by examining correlations, SVM weights, Random Forest importance, and ablation.

11.2 Correlations

Table 11.6 lists the ten features most strongly correlated with each of Zapotec's three phonation contrasts. The correlations for all 74 features are listed in Appendix M. In the following three sections, I review the trends in top correlations for the three contrasts.

Table 11.6: Zapotec Top Feature Correlations

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
Feature	Correlation	Feature	Correlation	Feature	Correlation
H1* – H2*_Mean	0.476	H1* – H2*_3	0.61	VoPT	0.408
H1* – H2*_2	0.461	H1* – H2*_Mean	0.586	HNR05_Mean	-0.358
H1* – H2*_3	0.445	H1* – A1*_3	0.547	HNR05_3	-0.346
H1* – A1*_3	0.402	CPP_3	-0.544	HNR05_2	-0.325
H1* – A1*_Mean	0.4	H1* – H2*_2	0.531	CPP_3	-0.312
RMS_Energy_2	-0.39	H1* – A1*_Mean	0.503	HNR15_Mean	-0.283
H1* – A1*_2	0.38	HNR05_2	-0.484	HNR15_2	-0.275
RMS_Energy_Mean	-0.366	CPP_Mean	-0.467	HNR25_Mean	-0.274
SHR_Mean	-0.284	CPP_2	-0.461	HNR35_Mean	-0.272
RMS_Energy_1	-0.283	SHR_Mean	-0.46	F1_3	0.272

11.2.1 Breathy vs. Creaky Correlations

Figure 11.1 plots the correlations for Zapotec's breathy vs. creaky contrast. Each bar represents a feature, with the feature's category indicated by the bar's color. The bar's value indicates the feature's correlation with a voice quality; positive features are correlated with breathy voicing and negative features with creaky voicing, and the magnitude of the value represents the strength of the relationship.

Spectral Tilt (dark orange) accounts for the five strongest correlations in this contrast, though no correlations are strong. Increased Spectral Tilt is correlated with breathy voicing, which reflects slower and less complete glottal closure. Spectral Tilt is typically largest in breathy voice and smallest in creaky voice, so it should be particularly useful in this contrast. The two measures that do best are H1* – H2* and H1* – A1*.

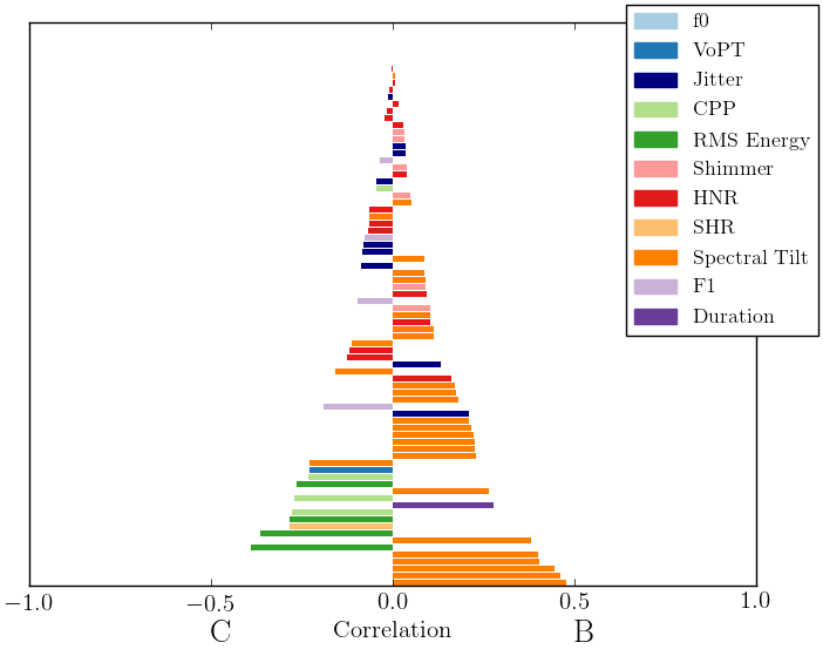


Figure 11.1: Zapotec Feature Correlations, B vs. C

RMS Energy (dark green) is moderately correlated with creaky voicing and accounts for three of the top ten features for this contrast. Both voice qualities included in this contrast – breathy and creaky – typically have a lower intensity than modal voicing. The vocal folds spend too much time open during breathy voicing to build up enough air pressure to achieve the same intensity as modal voicing, and too much time closed during creaky voicing to expel enough air to achieve the same intensity as modal voicing. The moderate correlation between RMS Energy and creaky voicing in Zapotec suggests that breathy voice may have particularly low intensity or that creaky voice has more energy than the prototypical creaky voice, or some combination of the two.

Several features have weak correlations with this contrast, but are still near the top of the list. SHR (light orange) is weakly correlated with creaky voicing, Duration (dark purple) with breathy voicing, and VoPT (bright blue) with creaky voicing.

11.2.2 Breathy vs. Modal Correlations

Figure 11.2 plots each feature's correlation with the breathy vs. modal contrast. Six of these features have strong correlations.

Spectral Tilt (dark orange) accounts for five of the six strongly correlated features, all correlated with breathy voicing. This is consistent with Spectral Tilt's correlation with breathy voice in the breathy vs. creaky contrast. However, I would have expected Spectral Tilt to correlate more strongly with breathy voice compared to creaky voice than compared to modal voice, as the difference in Spectral Tilt is expected to be larger between breathy and creaky vowels than between breathy and modal vowels.

CPP (light green) is strongly correlated with modal voicing; two other CPP features are among the top ten for this contrast and have medium correlations. One HNR feature (red) is among the top ten as well. Both CPP and HNR reflect the amount of noise in the signal and are typically higher for modal voice than for either type of non-modal phonation; this is consistent with what we see here.

One SHR feature (light orange) makes it into the top ten, also correlated with modal

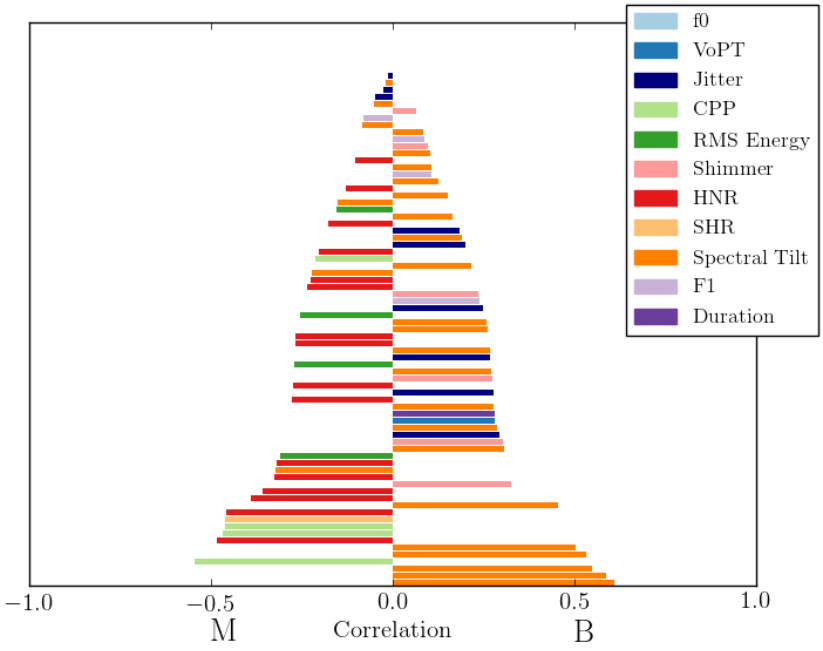


Figure 11.2: Zapotec Feature Correlations, B vs. M

voicing. SHR typically describes multiply pulsed creaky voicing. I am puzzled by its correlation with modal voicing here, which implies that modal voicing has more subharmonics than breathy voicing; perhaps it simply means that breathy voicing has less clear harmonics than modal voicing. Weights, importance, and ablation will help shed light on whether SHR is actually a useful feature.

11.2.3 Creaky vs. Modal Correlations

Finally, Figure 11.3 plots the correlations of each feature in the creaky vs. modal contrast. No features have strong correlations for this contrast.

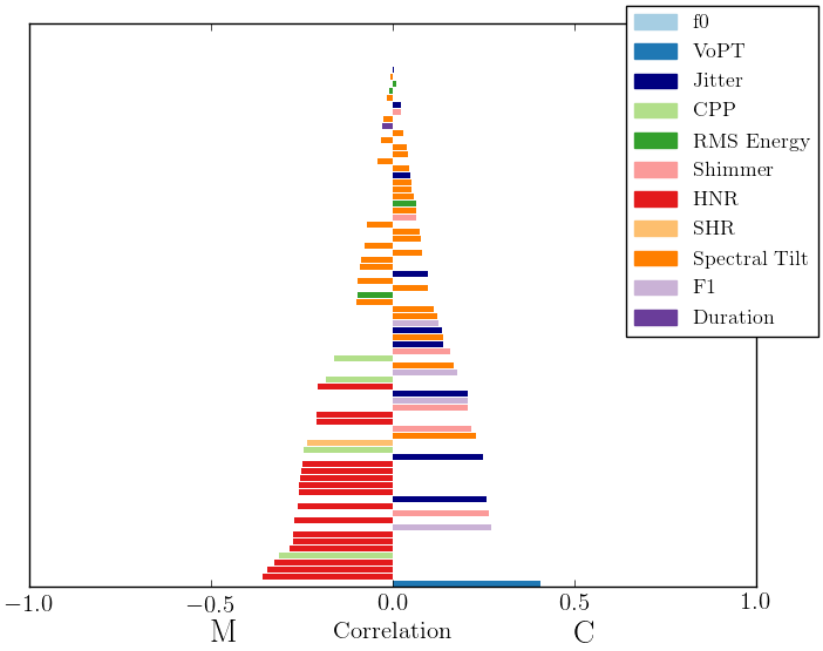


Figure 11.3: Zapotec Feature Correlations, C vs. M

Variance of Pitch Tracks (bright blue) is the strongest correlation for this contrast, though with a coefficient of 0.408, it's only moderately correlated with creaky voice. VoPT seems to be characterizing non-modal phonation nicely in Zapotec – its correlation is strongest for

creaky vs. modal, followed by breathy vs. modal, and then breathy vs. creaky; this suggests that VoPT is highest for creaky voice, followed by breathy voice, and then modal voice.

HNR (red) accounts for seven of the top ten features for this contrast and is correlated with modal voicing. The three strongest are all HNR05, suggesting that creaky voicing's noise is concentrated from 0 to 500 Hz. Several CPP features (light green), which are similar to HNR, are also correlated with modal voicing, though generally not as strongly.

F1 (light purple), Shimmer (pink), and Jitter (dark blue) features are weakly but noticeably correlated with creaky voicing. Several languages have been found to produce non-modal vowels with a higher F1 than their modal counterparts (see Chapter 3), and Jitter and Shimmer are both associated with the irregularity characteristic of non-modal phonation.

The correlations for Zapotec's three phonation contrasts highlight several categories of features that are useful in distinguishing phonation type: Spectral Tilt, HNR, CPP, VoPT, and RMS Energy.

11.3 SVM Weights

The next way to examine the relationship between features and Zapotec phonation is through weights, which represent how useful each feature is to the SVM's classification process. As seen in previous chapters, the multicollinearity present in all of my feature sets poses problems for weights. In order to get meaningful weights, I must pare down the set of features to remove collinear features. My strategy in several other chapters, using the single feature from each category with the largest weight, causes too large a drop in weighted F1 score, from 0.6959 to 0.55724. Using the single feature from each category with the largest *correlation* results in a much more reasonable weighted F1 score relative to the baseline: 0.63948. And, importantly, the signs of the weights in the subset generally match the signs of the category's average weight, indicating that they were not assigned one of the erroneous weights caused by multicollinearity.

The weights of each feature from this subset are listed in Table 11.7. Like correlations,

weights are by-contrast (each contrast is discussed below), their sign indicates association with one phonation type, and their magnitude indicates the strength of that relationship.

Table 11.7: Zapotec Feature Weights

Breathy vs. Creaky		Breathy vs. Modal		Creaky vs. Modal	
RMS_Energy_2	-1.03	H1* – H2*_3	1.333	VoPT	0.77
H1* – H2*_3	0.658	HNR015_2	-1.194	RAP_Jitter_Mean	0.556
RAP_Jitter_Mean	-0.536	RAP_Jitter_Mean	-0.963	Local_Shimmer_dB_Mean	0.547
SHR_Mean	-0.488	Duration	0.831	Duration	0.525
VoPT	-0.473	RMS_Energy_2	-0.693	H1* – H2*_3	0.42
HNR015_2	-0.424	F1_3	0.665	F1_3	0.401
Local_Shimmer_dB_Mean	-0.181	CPP_3	0.523	HNR015_2	-0.099
Duration	0.133	SHR_Mean	-0.507	CPP_3	0.092
CPP_3	0.111	Local_Shimmer_dB_Mean	0.242	RMS_Energy_2	0.081
F1_3	0.081	VoPT	-0.103	SHR_Mean	0.006

11.3.1 Breathy vs. Creaky Weights

Figure 11.4 shows each feature's weight in the breathy vs. creaky contrast. The weights are, overall, not particularly large.

RMS Energy (dark green) has the largest weight for this contrast and is associated with creaky voicing. This is consistent with the correlations. While Energy is typically used to describe modal vs. non-modal phonation, it appears to also provide a distinction between Zapotec's two types of non-modal phonation.

After Energy is Spectral Tilt (dark orange), which is higher in breathy vowels than in creaky vowels. Spectral Tilt is also highly ranked in the correlations – in fact, several Spectral Tilt features have stronger correlations than Energy features.

11.3.2 Breathy vs. Modal Weights

Features weights in the breathy vs. creaky contrast, shown in Figure 11.5, are overall much larger than in the breathy vs. creaky contrast.

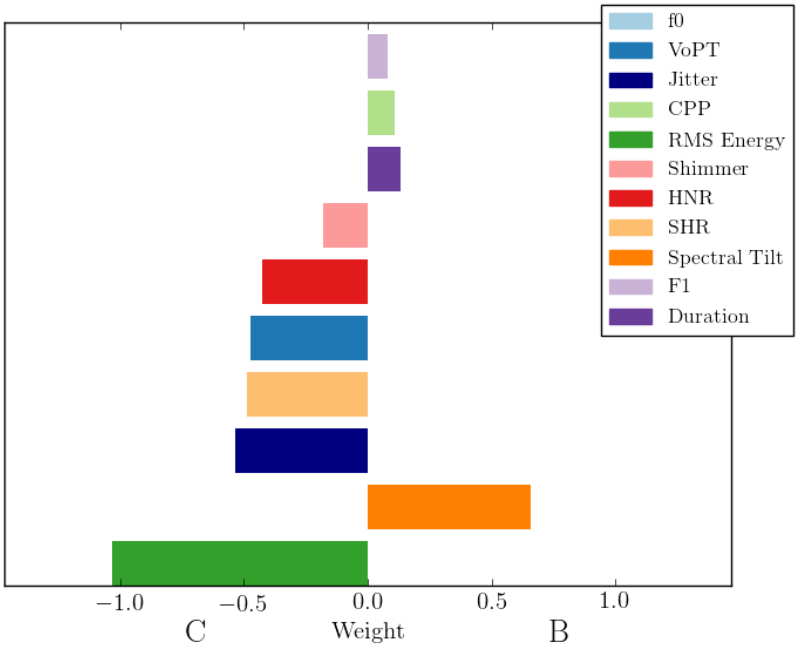


Figure 11.4: Zapotec Feature Weights, B vs. C

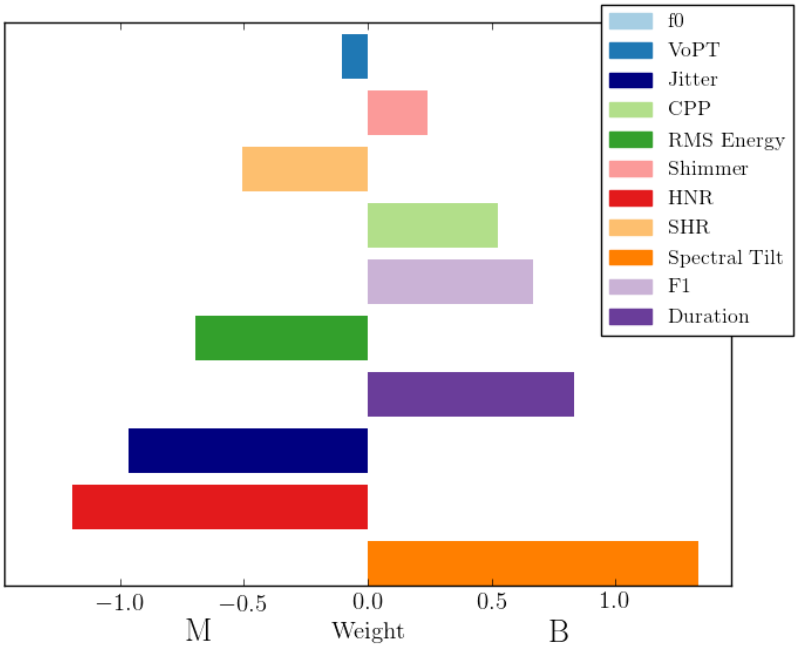


Figure 11.5: Zapotec Feature Weights, B vs. M

The largest weight comes from Spectral Tilt (dark orange). I'm again surprised to see that the relationship between Spectral Tilt and phonation is stronger in the breathy vs. modal contrast than in the breathy vs. creaky contrast.

After Spectral Tilt is HNR (red), which is larger in modal voicing than in breathy voicing. But not far behind is Jitter (dark blue), also associated with modal voicing; this counterintuitive association makes me suspect that this is, perhaps, the wrong Jitter feature to include.

11.3.3 Creaky vs. Modal Weights

Finally, the weights for the creaky vs. modal contrast are shown in Figure 11.6. Two things about this figure stand out: the weights are quite a bit smaller than for either of the other contrasts, and the figure is lopsided; nearly everything is associated with creaky voicing.

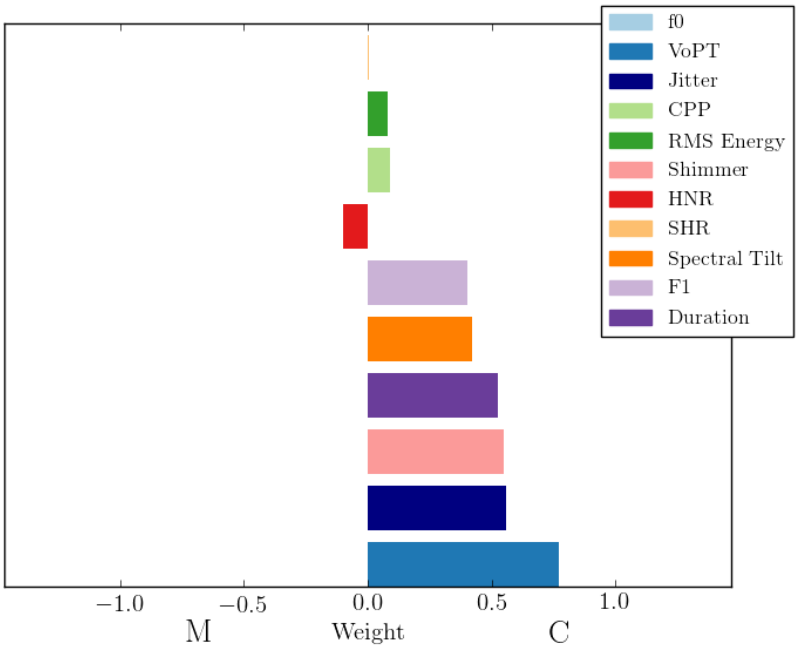


Figure 11.6: Zapotec Feature Weights, C vs. M

VoPT (bright blue) has the largest weight for this contrast, which is consistent with the correlations. This is also consistent with what VoPT has done in other languages; it seems to distinguish creaky voice from modal voice fairly well. VoPT is followed by Jitter (dark blue) and Shimmer (pink), which indicate cycle-to-cycle variation in f_0 and intensity, respectively.

Though a handful of the weights are unexpected, the features with the largest weights are generally consistent with the correlations: Spectral Tilt, HNR, Energy, VoPT, and Jitter.

11.4 Random Forest Importance

Next, I turn to Random Forest feature importance. Importance values are all positive; they represent a feature's overall use to the model, rather than a link to a specific side of a contrast. Table 11.8 lists those values for the ten most important features, and Figure 11.7 plots the importance for the full set; the values for the full set can be found in Appendix M.

Table 11.8: Zapotec Top Feature Importance

Feature	Importance
H1* – A1*_3	0.0427
H1* – H2*_3	0.03659
RAP_Jitter_Mean	0.03401
CPP_2	0.03104
H1 * – A1*_Mean	0.02989
Duration	0.02822
H1* – H2*_2	0.02807
VoPT	0.02802
SHR_Mean	0.02768
HNR05_1	0.0262

The ten most important features come from seven different categories. This diversity can also be seen in Figure 11.7, which includes very few clusters of a given color.

Four of those ten, including the top two, are Spectral Tilt (dark orange). After Spectral Tilt, few patterns emerge. Several feature categories have one or two fairly important features, but other features from that category are found much lower down on the list. While nearly all categories make an appearance in the top ten, Spectral Tilt is the only

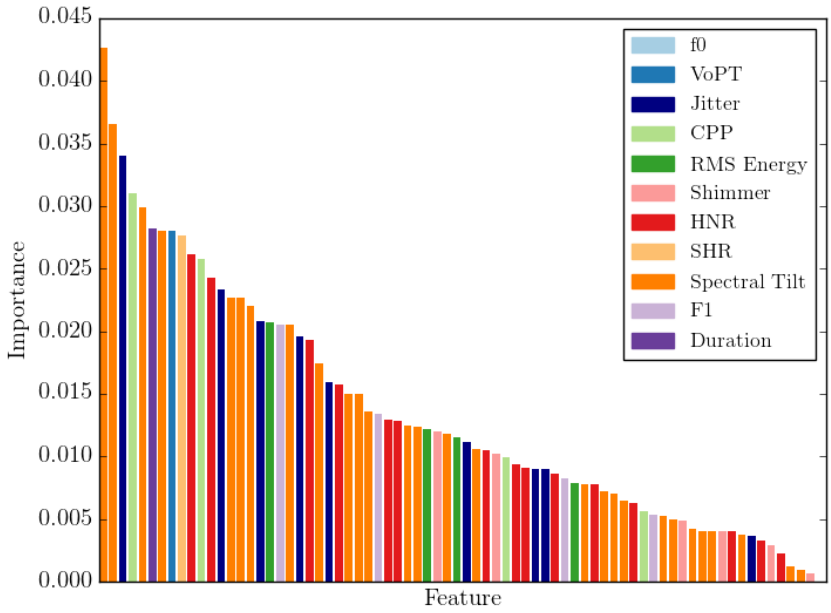


Figure 11.7: Zapotec Feature Importance

feature category that stands out according to importance, though Jitter and CPP are also near the top of the list.

11.5 Ablation

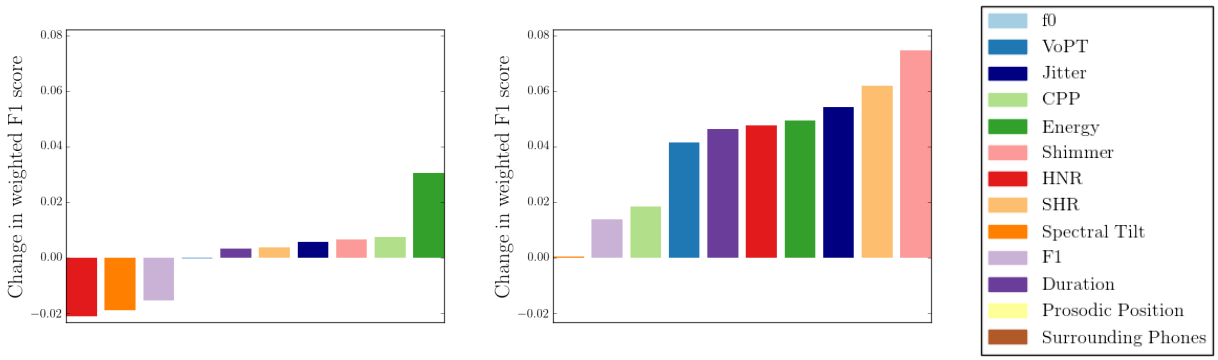
Ablation provides the final of the four lenses through which to view the features' usefulness in Zapotec phonation classification. I use two types of ablation: categorical ablation and iterative ablation.

Categorical ablation involves training the classifiers on all but one category of features and seeing how each category's exclusion impacts the classifier's performance. Table 11.9 and Figure 11.8 report the results of categorical ablation.

Table 11.9: Zapotec Category Ablation

Feature	SVM		Random Forest	
	Weighted F1	Change	Weighted F1	Change
Spectral Tilt	0.67701	-0.01889	0.62889	0.0003
CPP	0.70334	0.00744	0.64697	0.01838
Energy	0.72632	0.03042	0.67807	0.04948
HNR	0.67464	-0.02126	0.6763	0.04771
SHR	0.69956	0.00366	0.69066	0.06207
F1	0.68047	-0.01543	0.64243	0.01384
Duration	0.69903	0.00313	0.67504	0.04645
Jitter	0.70161	0.00571	0.68294	0.05435
Shimmer	0.70241	0.00651	0.70333	0.07474
VoPT	0.69562	-0.00028	0.67012	0.04153

As was in case for the other languages, ablating any given category doesn't hurt the classifiers much. Unlike in the other languages, the Random Forest *improves* when each category is ablated. Some improvement is unsurprising, as Random Forests are inherently random and one classifier's ordering might be better than the next, but such consistent improvement is unexpected. To see whether or not this was a fluke, I conducted categorical ablation again two more times for the Random Forest. In the first of those two tests, every single category again caused improvement. In the second, one feature category, CPP, caused



(a) SVM (b) Random Forest

Figure 11.8: Zapotec Category Ablation

a very small decrease in weighted F1 score, and the rest again caused improvement. This is certainly odd and perhaps reflects the fact the the Random Forest struggled to find patterns in the first place; its weighted F1 score is the lowest of the entire dissertation.

In iterative ablation, I pick the category whose ablation causes the largest drop in weighted F1 score and exclude it. I then ablate each remaining category and exclude whichever led to the next largest drop, and continue until there's only one category remaining. I use an SVM for iterative ablation and report the results in Table 11.10.

As seen in categorical ablation, HNR's exclusion negatively impacts the SVM the most. This is followed by Spectral Tilt, which has been consistently important for Zapotec, and CPP, which is similar to HNR but has generally been less useful than HNR. Energy and Jitter, which have been identified as important in previous sections, are near the bottom of the list.

The results of the two types of ablation are a bit odd but not contradictory to the correlations, weights, or ablation. HNR and Spectral Tilt are the feature categories that most impact classifier performance when they're ablated.

Table 11.10: Zapotec Iterative Category Ablation (SVM)

Category	Weighted F1
HNR	0.67464
Spectral Tilt	0.63033
CPP	0.59131
SHR	0.53907
F1	0.52395
VoPT	0.48244
Duration	0.48073
Energy	0.42838
Jitter	0.35405
Shimmer	–

11.6 Summary

This chapter has examined the features that relate to Zapotec's three-way phonation contrast using correlations, weights, importance, and ablation. There have been several surprises along the way, which may be caused by the classifiers (particularly the Random Forest) struggling to find meaningful patterns in the data. Nonetheless, a subset of particularly useful feature categories has emerged: Spectral Tilt and HNR appear to be the most and most consistently crucial in distinguishing Zapotec voice qualities, followed by Energy and VoPT, and finally by Jitter and CPP.

Spectral Tilt is the only category identified in all four sections as important. Spectral Tilt reflects the speed and completeness of glottal closure and is therefore generally the most distinct between breathy and creaky voice, with modal voice in the middle. In the Zapotec data, according to both weights and correlations, Spectral Tilt's relationship with phonation is stronger in the breathy vs. modal contrast than in the breathy vs. creaky contrast. Additionally, it's generally ranked low in the creaky vs. modal contrast. This may suggest that Zapotec creaky and modal voice are produced similarly in terms of the speed and completeness of glottal closure. The two most important features for the Random

Forest are Spectral Tilt features, ablating Spectral Tilt causes the second largest drop in performance for the SVM and the smallest increase in performance for the Random Forest, and it's the second to go in iterative ablation. H1* – H2* and H1* – A1* come up the most, and end of the vowel appears (and, to a lesser degree, the middle and mean) to be the most meaningful time span.

Few studies have examined the acoustic properties that differentiate Zapotec phonation types, but those that do rely on Spectral Tilt. Esposito (2004) used six Spectral Tilt measures to study phonation types in two speakers of Santa Ana del Valle Zapotec. She found that H1 – F3 best described phonation types for her male speakers and H1 – H2 for her (one) female speaker. Avelino Becerra (2004), working with a related dialect called Yalálag Zapotec, found that H1 – H2, H1 – A1, and H1 – A3 all differ between modal and creaky voicing.[1]

Harmonics-to-Noise Ratio appears to be diagnostic of Zapotec's modal vs. non-modal phonation; as it measures strength of noise in the signal relative to harmonics, it's higher in modal voicing than in breathy or creaky voicing. The correlations indicate that HNR is higher in modal voicing than in either breathy or creaky voicing, but the weights only indicate its importance in the breathy vs. modal contrast. An HNR feature is ranked tenth by the Random Forest, though the category's ablation improves the Random Forest's performance. For the SVM, its ablation causes the largest decrease in performance. The top HNR features are consistently measured in the lowest frequency range (0 – 500 Hz; HNR05), though no one time span stands out. I am not aware of other studies that investigate HNR in Zapotec phonation types.

Energy generally seems less useful than Spectral Tilt or HNR for Zapotec. In both the correlations and the weights, it's most associated with creaky voice in the breathy vs. creaky contrast, less so with modal in the modal vs. breathy contrast, and not important in the creaky vs. modal contrast. In other words, Energy appears to be lowest in breathy voicing in Zapotec. While correlations and weights identify Energy as useful, it receives a

[1]Unlike the dialects studied here, Yalála Zapotec does not contrast breathy voice.

low importance value from the Random Forest, causes an increase in performance for both classifiers, and is ablated near the end of the list in iterative ablation. Previous studies have not relied on Energy to distinguish Zapotec phonation types, but Avelino Becerra (2004) notes that Yalálag Zapotec's creaky voice is characterized by a period of reduced intensity, among other things.

Variance of Pitch Tracks relies on f_0 measures but measures their disagreement, rather than their actual values. It is therefore safe to use in Zapotec, despite the relationship between phonation and pitch. VoPT has the strongest relationship with the creaky vs. modal contrast (associated with creaky voicing) according to both correlations and weights. It's weakly correlated with breathy voicing in the breathy vs. modal contrast but has a low weight. It's ranked eighth in Random Forest importance but is generally around the middle of the pack in all ablation tests. While it seems to do a solid job of distinguishing modal voice from non-modal voice, its role seems quite a bit weaker than Spectral Tilt or HNR's role.

In what I believe to be the final tier of important features is **Cepstral Peak Prominence**. CPP provides similar information to HNR but seems generally less useful in describing Mazatec phonation. It's strongly correlated with modal voicing in the breathy vs. modal contrast and moderately correlated with creaky in breathy vs. creaky and with modal in creaky vs. modal; taken together, this suggests that CPP is greatest in modal voicing and lowest in breathy voicing, with creaky voicing somewhere in between. However, this pattern isn't really corroborated by the weights, which rank CPP rather low. It's ranked fourth in Random Forest importance and is the third category to be ablated in the iterative ablation. Esposito (2006) found that CPP is significantly higher in Zapotec modal vowels than breathy vowels, but did not investigate creaky vowels.

Finally, **Jitter** features are not at the top of any list, but they're near the top in weights and importance. RAP_Jitter_Mean has the third largest weight for breathy vs. creaky and breathy vs. modal (associated with creaky and modal, respectively), and the second largest weight for creaky vs. modal (associated with creaky). This same feature has the

third largest importance as well. The weights' suggestion that modal voicing has higher Jitter than breathy voicing is odd, but its association with creaky voicing in the other two contrasts makes sense.

Jitter is not often used to study phonation, and the existing literature on Zapotec is no exception. Avelino Becerra (2004) observes of Yalálag Zapotec creaky vowels that "the middle portions presents [sic] the most irregular F0 values (often showing pitch doubling)," but does not specifically use Jitter in his analysis.

Features from these six categories all contribute to differentiating Zapotec phonation types. I've presented them in what I believe to be their order of importance, based on the results of the correlations, weights, importance, and ablation. To conclude this chapter, I check that order by training classifiers on subsets of these features; these results are reported in Table 11.11.

Table 11.11: Zapotec Weighted F1 Using Subsets of Features

Feature Categories	Weighted F1	
	SVM	Random Forest
Baseline (all features)	*0.6959*	*0.62859*
Spectral Tilt, HNR, RMSE, VoPT, Jitter, CPP	0.68915	0.69429
Spectral Tilt, HNR	0.57953	0.644
RMSE, VoPT	0.54517	0.57365
Jitter, CPP	0.63126	0.60016

Training on features from these six categories causes a small drop in the SVM's weighted F1 score and a significant increase in the Random Forest's; perhaps some of the other features were presenting contradictory information and it can now perform better in their absence. Training on pairs of feature categories, Spectral Tilt and HNR lead to the highest weighted F1 scores for both classifiers, as expected. Jitter and CPP, however, perform slightly better together than Energy and VoPT.

These results suggest that Zapotec's three voice qualities are best described by the abruptness and completeness of vocal fold closure and by periodicity. Energy and irregularity

also play a role in distinguishing them.

Chapter 12

A CROSS-LINGUISTIC COMPARISON OF PHONATION

The previous six chapters examined which acoustic properties best distinguish phonation types in six different languages. I used machine learning as well as more traditional statistical methods to examine the relationship between acoustic features (properties) and phonation: correlations, Support Vector Machine weights, Random Forest importance, and ablation. This chapter examines how these features and the overall classifier performance vary between the six languages included in this dissertation.

12.1 Comparing Performance

I'll start by comparing the performance of the Support Vector Machine and the Random Forest for each of the six languages. I report the weighted F1 score of each classifier, trained on resampled data, in Table 12.1. I also include the improvement over baseline for each classifier.[1] Recall that there are several important differences between the corpora and how the six languages use phonation. The corpora range in size from 180 tokens (Mandarin) to 10,031 tokens (English). Four of the languages use breathy, modal, and creaky voice; Mandarin uses only modal and creaky, while Gujarati uses only breathy and modal. Finally, Gujarati and Mazatec use phonation phonemically, Hmong, Mandarin, and Zapotec use it alongside tones, and English uses it sociolinguistically, allophonically, and prosodically. (See the respective language-specific chapters for more information.)

The Mandarin classifiers perform by far the best of the set; the SVM has a weighted F1 score of 0.96119 and the Random Forest 0.96134. Recall, however, that the Mandarin

[1] I calculated the baseline weighted F1 score for each language based on what that number would be if every instance were labeled as the majority class.

Table 12.1: Classifier Performance

Language	SVM		Random Forest	
	F1	Improvement	F1	Improvement
English	0.78767	0.13063	0.78895	0.13163
Gujarati	0.72079	0.30055	0.72694	0.30655
Hmong	0.6444	0.25383	0.66723	0.27683
Mandarin	0.96119	0.42767	0.96134	0.42767
Mazatec	0.6994	0.41947	0.66891	0.38947
Zapotec	0.6959	0.49961	0.62859	0.43261

data set is a bit different from the others: it's the smallest, all words consist of the same phonemes, and I treat all third tones as creaky, first and second tones as modal, and exclude all fourth tones. Additionally, several acoustic measures that are included in the feature set, such as Energy, have been shown to strongly correlate with tone; while f_0 itself is excluded, several other features may be doing essentially the same job as f_0. Because of these caveats, it's unclear how this classifier would scale up. Nonetheless, the two Mandarin classifiers have overcome several sources of potential noise and challenges and perform extremely well on this data set.

At the other end of the spectrum is Zapotec. The SVM's weighted F1 score is 0.6959 and the Random Forest's is 0.62859; however, the Zapotec classifiers have the largest improvement over the baseline. This indicates that while the classifiers are finding some patterns, they struggle to find enough patterns to successfully differentiate Zapotec's three phonation types. They could struggle for various reasons, or for a combination of reasons. Perhaps the features didn't capture phonation differences effectively, perhaps there was too much variation between or within speakers, or perhaps the data set was too small for patterns to emerge. Either way, an out-of-the-box classifier was unable to confidently distinguish Zapotec's phonation types.

The remaining four languages fall in between the extremes. The Mazatec and Hmong

classifiers perform not much better than Zapotec's (in fact, Hmong's SVM doesn't do as well as Zapotec's). The English and Gujarati classifiers both perform quite a bit better, though not nearly as well as Mandarin's. This range of classifier performance may help shed light on a typological question regarding phonation use, discussed below.

12.1.1 Performance by Phonation Use

The six languages fall into three categories of phonation use. Gujarati and Mazatec use phonation contrastively. Hmong, Mandarin, and Zapotec phonation accompanies tone in some way (as part of a complex register system in Hmong, allophonically in Mandarin, and all non-modal vowels have a falling tone in Zapotec), and English phonation is sociolinguistic, prosodic, and allophonic. Including languages that use phonation differently allows me to examine similarities and differences across these categories. While I do not expect languages in the same category to use the same acoustic properties to distinguish phonation types, I am interested in how well the classifiers perform between categories.

Phonetic contrast limits variability in production for the intuitive reason that variability of a meaningful segment leads to more possibility for confusion. Manuel (1990), for example, using a sample of three languages, found less anticipatory vowel-to-vowel coarticulation for the language with the most crowded vowel space. A more crowded vowel space means that coarticulation may move a vowel's realization too close to a neighboring vowel, while this is less of a risk in a language with a less crowded vowel space.

Following this idea, it seems plausible that, within a language, there is less variability in how phonation is produced and that the differences between phonation types are more extreme when recognizing phonation type is crucial to understanding. Both of these factors would result in increased accuracy by way of reducing noise[2] and increasing the difference between classes. In languages that use phonation alongside tone, phonation is less important because secondary cues can do some of the work. In languages that use

[2]Noise here refers to noise in the data, not acoustic noise.

phonation sociolinguistically, allophonically, and prosodically, differentiating phonation types not important to understanding lexical meaning; this could result in more variation in its production as well as less clear boundaries between classes. In other words, I am looking at the classifier accuracy to see if it is highest in languages that use contrastive phonation, followed by tonal phonation, and then non-contrastive phonation. Table 12.2 reports the same information as Table 12.1, but grouped by how phonation is used.

Table 12.2: Classifier Performance

Phonation Use	Language	SVM		Random Forest	
		F1	Improvement	F1	Improvement
Contrastive	Gujarati	0.72079	0.30055	0.72694	0.30655
	Mazatec	0.6994	0.41947	0.66891	0.38947
Tonal	Hmong	0.6444	0.25383	0.66723	0.27683
	Mandarin	0.96119	0.42767	0.96134	0.42767
	Zapotec	0.6959	0.49961	0.62859	0.43261
Socio., Allo., Prosod.	English	0.78767	0.13063	0.78895	0.13163

Though the sample is small – there are between one and three languages in each category – no patterns resembling my prediction emerge. In fact, the language with the best performing classifier (Mandarin) and the language with the worst performing classifier (Zapotec) are in the same category. However, it's more meaningful to look at improvement over baseline rather than at the actual weighted F1 score. Zapotec, which uses phonation alongside tones, has the greatest improvement in weighted F1 score for both classifiers. English, which uses phonation sociolinguistically, allophonically, and prosodically, has the smallest improvement in weighted F1 score over baseline. This does not follow the expected pattern that classifiers would generally perform best for languages that use contrastive phonation, though it does follow the expected pattern that classifiers would generally perform worst for languages that use phonation in less lexically important ways.

While the English classifiers show the least improvement over baseline, I see no gradient

between contrastive and tonal languages; languages that use phonation contrastively do not experience a larger improvement in weighted F1 score than tonal languages. Of course, weighted F1 score can be increased or decreased by many factors, including noise in the data and sample size, so it is not a direct measure of how consistently produced or well differentiated phonation types are. Examining the between and within speaker variation, as well as quantifying phonation types, would be a more direct way to approach this question. While these strategies, along with including more languages in each category, may paint a different picture, the current six languages do not show any patterns linking accuracy with phonation use.

Recall that these twelve classifiers are essentially out-of-the-box; they use all default parameters, with the exception of a linear kernel for the SVMs. But machine learning models have many hyperparameters that can be fine-tuned and lead to greatly improved performance; I do just this for English in the following chapter. The weighted F1 scores reported here are fairly good, considering that the classifiers use (mostly) default settings. There is still room for improvement, but weighted F1 scores, even of just around 0.7, indicate that the classifiers are finding meaningful patterns; this suggests that automatic classification of phonation types is possible using machine learning.

12.2 Comparing Key Feature Categories

In each of the previous six chapters, correlations, feature weights, feature importance, and ablation highlighted a subset of feature categories that do more of the work than others in distinguishing phonation types. I rank them, approximately, based on how often those categories come up according to the different tests and how important they are relative to other categories. For each language, I identified six categories of features that are useful in distinguishing phonation types.[3] These categories are listed in Table 12.3 below. They are ordered; those at the top of the list are doing more work than those at the bottom of the list.

[3]I did not set out to identify six categories; it just so happened that six categories emerged for each language.

Note that they differ from many other studies of phonation in that they represent what an algorithm found useful in distinguishing phonation types, not what a human listener found useful; the results therefore do not necessarily reflect human perception of phonation.

Table 12.3: Useful Features, Ranked

English	Gujarati	Hmong	Mandarin	Mazatec	Zapotec
CPP	Duration	HNR	HNR	Spectral Tilt	Spectral Tilt
HNR	Shimmer	SHR	Jitter	f_0	HNR
f_0	CPP	Spectral Tilt	RMS Energy	Jitter	RMS Energy
VoPT	Spectral Tilt	RMS Energy	VoPT	VoPT	VoPT
RMS Energy	HNR	Jitter	Shimmer	RMS Energy	Jitter
Surrounding Phones	VoPT	VoPT	Spectral Tilt	HNR	CPP

The list of important feature categories varies quite a bit from language to language. HNR and Spectral Tilt are both ranked first for two languages. Most feature categories included in these lists appear more than once. Table 12.4 makes these lists a bit easier to digest by counting how many languages include each category in their top six, but excluding rankings.

Two feature categories are used by classifiers for all languages: Harmonics-to-Noise Ratio and Variance of Pitch Tracks. HNR is in the top two features for four of the six languages, while VoPT is generally ranked lower. Nonetheless, they appear to both be consistently useful (though to varying degrees) in phonation type classification in different languages. Slightly less consistently useful are RMS Energy and Spectral Tilt, which are useful in five of the six languages. CPP is useful in four languages, Shimmer and f_0 in two, and remaining features are useful to only one or no languages.

The above comparisons provide a nice overview but are limited in that not all six languages contrast the same phonation types. If a feature is relevant to the breathy vs. creaky contrast, it'll only come up in languages that use both breathy and creaky voice. Additionally, features that come up for specific contrasts shed light on the nature of those contrasts. In the following three sections, I review the key features for each contrast. I

Table 12.4: Useful Features, Counts

	eng	guj	hmn	cmn	maj	zap	**Total**
HNR	x	x	x	x	x	x	**6**
VoPT	x	x	x	x	x	x	**6**
RMS Energy	x		x	x	x	x	**5**
Spectral Tilt		x	x	x	x	x	**5**
Jitter			x	x	x	x	**4**
CPP	x	x				x	**3**
f_0	x				x		**2**
Shimmer		x		x			**2**
Duration		x					**1**
SHR			x				**1**
Surrounding Phones	x						**1**
F1							**0**
Prosodic Position							**0**

select these features based on the two metrics that operate on a contrast-by-contrast basis: correlations and weights.

12.2.1 Breathy vs. Creaky

Four of the six languages employ breathy and creaky voice. This contrast is of particular interest for me; breathy and creaky voice are often described as existing on opposite ends of the voicing continuum, yet we saw in Chapter 3 that many acoustic properties are shared by the two types of non-modal phonation. Table 12.5 lists the features that are useful to each language's classifier in distinguishing breathy voice from creaky voice; they are unranked.

All four languages use Spectral Tilt to distinguish breathy voice from creaky voice. Spectral Tilt is one of the few features whose values match the voicing continuum: it's greatest in breathy voicing and smallest in creaky voicing. Thus, of all the contrasts, it should be particularly useful in this contrast, and it is.

HNR and f_0 are each used by two languages, and the remaining categories by one each. Four feature categories are not represented at all here: Shimmer, Prosodic Position, F1, and

Table 12.5: Top Feature Categories: Breathy vs. Creaky

	eng	hmn	maj	zap	Total
Spectral Tilt	x	x	x	x	4
HNR	x	x			2
f_0	x		x		2
Jitter		x	x		2
CPP	x				1
Surrounding Phones	x				1
SHR		x			1
VoPT			x		1
RMS Energy				x	1

Duration; these categories generally do not contribute to the distinction between breathy and creaky voicing in these languages.

12.2.2 Breathy vs. Modal

Table 12.6 shows which features are used to distinguish breathy voice from modal voice in the five languages that use those two phonation types.

Table 12.6: Top Feature Categories: Breathy vs. Modal

	eng	guj	hmn	maj	zap	Total
CPP	x	x			x	3
Spectral Tilt			x	x	x	3
VoPT	x			x		2
Surrounding Phones	x					1
SHR			x			1
Duration		x				1
Shimmer		x				1
F1		x				1
f_0				x		1
Jitter					x	1
HNR					x	1

The most-used feature categories are only used by three of the five languages: CPP and Spectral Tilt. CPP appears to capture differences in noise between breathy and modal vowels. As discussed above, differences in Spectral Tilt are most extreme between breathy and creaky phonation, but it also appears effective at distinguishing breathy from modal phonation. Variance of Pitch tracks is the only other feature used by more than one language to contrast breathy and modal voice. Absent from the list are prosodic position (unsurprising) and RMS Energy. I'm surprised that Energy is not among these feature categories, as breathy voicing is often described as having a lower intensity than modal voicing. While its absence doesn't contradict that, it implies that other cues are stronger than Energy for this contrast.

12.2.3 Creaky vs. Modal

Finally, Table 12.7 lists the categories that are used to distinguish creaky voice from modal voice in the five languages that use both. This list is notably shorter than for either of the other two contrasts, suggesting that what distinguishes creaky voice from modal voice is more consistent from language to language than what describes the other contrasts.

Table 12.7: Top Feature Categories: Creaky vs. Modal

	eng	hmn	cmn	maj	zap	Total
VoPT	x		x	x	x	4
HNR	x	x	x		x	4
RMS Energy		x	x	x		3
Jitter			x	x	x	3
Spectral Tilt			x	x		2
f_0	x					1
CPP	x					1

In four of the five languages, VoPT is useful in distinguishing creaky voice from modal voice. In originally developing VoPT, this particular algorithm also seemed best suited for distinguishing creaky voice from modal voice, though it was only tested on English; its presence here suggests that this may extend to other languages. However, in effectively

measuring pitch tracking errors, VoPT doesn't directly measure any particular acoustic property, it just measures how difficult it was to track the pitch.

HNR is also useful in four of the five languages and RMS Energy in three of the five. Though HNR and RMSE are both typically lower in non-modal phonation than in modal phonation, several of these languages suggest that they differ in breathy and creaky voice in a meaningful way.

12.2.4 Summary

Breaking down the set of useful features by contrast shows that what describes a phonation type in one language is not the same as what describes the same phonation type in another. This is not exactly news, as a handful of other studies (Keating et al., 2011) have made this comparison. What sets these results apart is that a machine learning algorithm identified the patterns; rather than a human checking for statistical relationships, the algorithms sifted through a very large set of potential relationships to find ones that helped them classify phonation types.

Comparing across languages has also highlighted the fact that some of the measures traditionally used to distinguish modal from non-modal voicing can also distinguish between the two types of non-modal voicing. Four of the six languages employ both breathy and creaky voicing; in all four of those languages, at least one feature that's useful in the breathy vs. creaky contrast are generally associated broadly with non-modal phonation. English uses HNR and f_0, Hmong uses HNR, Mazatec uses f_0, and Zapotec uses RMS Energy. This suggests that a deeper dive into how these acoustic properties vary between types of non-modal phonation, rather than between modal and non-modal phonation, is warranted.

The number of feature categories in each of the contrast subsets above is also noteworthy. Between the six languages, eleven feature categories are involved in the breathy vs. modal contrast, nine in the breathy vs. creaky contrast, and seven in the creaky vs. modal contrast. This suggests that, cross-linguistically, what describes creaky voice may be more consistent that what describes breathy voice.

12.2.5 Related Languages

Intuitively, related languages often share linguistic traits, though wide diversity also exists within families. Linguistic similarities can include meaningful use of phonation types; many Otomanguean languages, for example, employ contrastive or tonal phonation. On a finer level, it's also possible that phonation types in related languages share properties. Two pairs of languages used in this data set are related: English and Gujarati (Indo-European), and Mazatec and Zapotec (Otomanguean). Though this sample is quite small, it's a good first pass at examining how the properties of phonation types vary within a language family.

English and Gujarati

English and Gujarati are both Indo-European languages; their relationship is shown in Figure 12.1. They exhibit some marked differences in how they use phonation. First, English voice quality is sociolinguistic, prosodic, and allophonic, while Gujarati voice quality is contrastive. Second, English uses breathy, modal, and creaky voice, while Gujarati only contrasts breathy and modal.

Chapters 6 and 7 explored which features best distinguish English and Gujarati phonation types, respectively. Overall, the two languages both use three of the same feature categories relatively heavily: HNR, VoPT, and CPP. Those three features have fairly different rankings, as seen in Table 12.3. However, creaky voicing is included in the English classifier but not the Gujarati classifier; it's therefore a fairer comparison to focus on the breathy vs. modal contrast for both languages. The breathy vs. modal contrast in English, according to correlations and weights, is best described by CPP, VoPT, and Surrounding Phones. In Gujarati, it's best described by CPP, Duration, Shimmer, and F1. With only CPP in common, English and Gujarati breathy and modal voice are rather different.

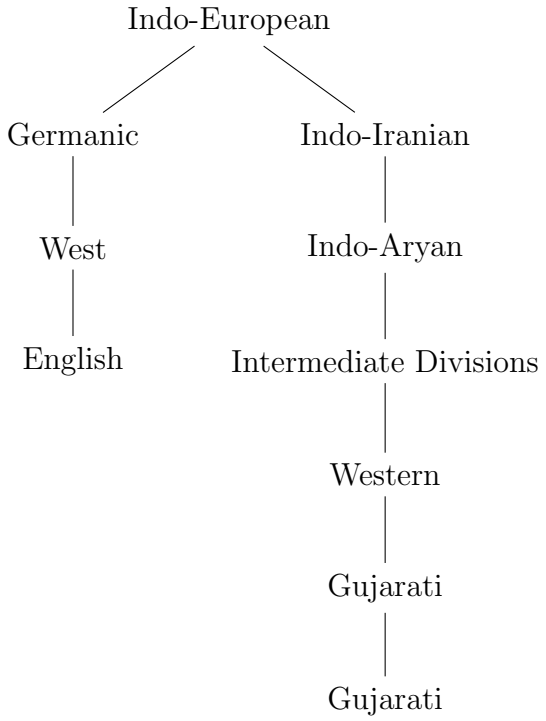


Figure 12.1: English and Gujarati within the Indo-European Family (Lewis et al., 2016)

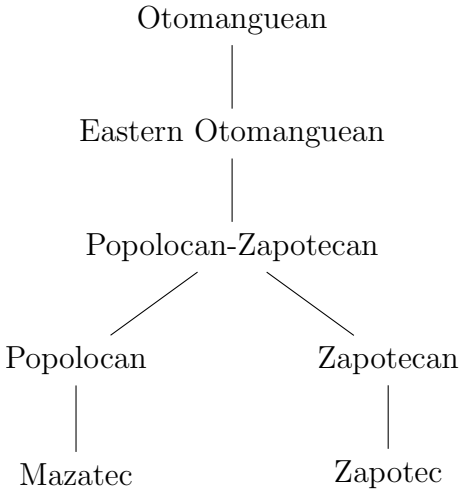


Figure 12.2: Mazatec and Zapotec within the Otomanguean Family (Lewis et al., 2016)

Mazatec and Zapotec

Mazatec and Zapotec are also related. Both are Otomanguean languages; their relationship is shown in Figure 12.2. Like English and Gujarati, they use phonation differently: it's contrastive in Mazatec and tonal in Zapotec. However, both languages use the same three voice qualities.

The machine learning takes on the types of features most important to phonation classification are described in Chapter 10 for Mazatec and Chapter 11. They share five of the top six categories that best distinguish phonation for the languages. The one feature category important to Mazatec but not Zapotec is f_0, which wasn't even considered for Zapotec because of the inherent link between tone and phonation. Both languages rank Spectral Tilt first, though the order of remaining features varies. However, looking at the by-contrast breakdowns shown above, *when* those categories are useful varies. There's some agreement – particularly on Spectral Tilt – but also some differences. Overall, Mazatec and Zapotec voice qualities seem more similar to each other than English and Gujarati voice qualities.

We see here that phonation types in different languages are described by different acoustic properties. Only two feature categories are used by all six languages: Harmonics-to-Noise Ratio and Variance of Pitch Tracks. HNR is generally skipped over in studies of phonation in favor of CPP; perhaps its ability to examine specific frequencies makes HNR more powerful than CPP in this study. VoPT is a new measure, described in Chapter 3, that may significantly aid voice quality identification in the future.

12.3 Training on One Language and Testing on Another

The previous section highlights the similarities and differences in the acoustic properties that distinguish voice qualities in different languages. While there's quite a bit of overlap between the lists, no language relies on the same set of features. This essentially means that the classifiers are language-specific; the patterns that they learn for one language likely won't transfer to another. Here, I examine just how language-specific the classifiers are by training on one language's data and testing on another.

For these purposes, I use just an SVM. I train the SVM on the entire data set from one language (resampled to be balanced) and test on the entire data set from another language (imbalanced). As in the other classifiers, all data have been normalized and missing values replaced with the class mean in the training data and the overall mean in the testing data. Features with 15% or more undefined values in the *training* data are excluded, but if they have 15% or more undefined values in the *testing* data, they are included, as the classifier expects them. Several of the features used in the English classifier are not available for the other languages; in order to test the English-trained classifier on other languages, these English-specific features must also be excluded. One final logistical note is that two of the classifiers – those trained on Gujarati and Mandarin – are trained only to identify two phonation types. A Gujarati-trained classifier identifies breathy and modal voice; a Gujarati-trained classifier tested on Mandarin will encounter no instances of breathy voice. This poses an extra challenge.

Table 12.8 reports the weighted F1 scores of classifiers trained on one language and

tested on another. The rows represent a classifier trained on the same language and tested on different languages. Values are missing along the diagonal, as they represent a classifier trained and tested on the same language (see Chapters 6 through 11 for that information).

Table 12.8: Mismatch Train/Test Matrix of Weighted F1 Scores

		Testing Lang					
		eng	guj	hmn	cmn	maj	zap
Training Lang	eng	–	0.54219	0.49616	0.78107	0.38923	0.42958
	guj	0.4725	–	0.41888	0.34891	0.18168	0.33264
	hmn	0.54571	0.5919	–	0.64957	0.32391	0.47165
	cmn	0.69509	0.36585	0.48458	–	0.51486	0.29454
	maj	0.54222	0.39352	0.3529	0.69711	–	0.35158
	zap	0.34809	0.42548	0.38659	0.35036	0.33048	–

Overall, the classifiers perform quite poorly when trained on one language and tested on another. These scores provide a way of quantifying just how different one language's phonation types are from another's. Training and testing on languages that rely on the same types of features will likely lead to better weighted F1 scores.

The best results come from training an SVM on English data and testing on Mandarin data; the weighted F1 score is 0.78107. These two languages had the highest performing classifiers in the first place (when tested on the correct language), indicating that the training phase results in clearer patterns than for the other language. The fact that the English classifier also performs well on the Mandarin data indicates that those patterns are applicable to Mandarin phonation as well. This likely boils down to f_0, which was included in the English classifier and is inherently linked with Mandarin tone. If English creaky voice has a lower f_0 than modal voice (as is the case in Mandarin, but quite extremely), this pattern could account for the success of the English-trained and Mandarin-tested classifier.

Two other classifiers perform fairly well: trained on Mandarin and tested on English (0.69509) and trained on Mazatec and tested on Mandarin (0.69711). The relationship

between tone and phonation in Mandarin likely contributes to both of these; like English, f_0 is useful in distinguishing Mazatec voice qualities. Additionally, Mandarin and Mazatec overlap on five of the six important feature categories.

The worst classifier is the one trained on Gujarati and tested on Mazatec, with a weighted F1 score of 0.18168. Gujarati does not use creaky voice, so it has no chance at correctly identifying Mazatec creaky vowels, which make up approximately 40% of the data set. Nonetheless, it appears to do quite poorly on its other vowels as well. Gujarati and Mazatec use features from three of the same categories: HNR, VoPT, and Spectral Tilt. But this overlap doesn't appear to do much for the classifier. Looking back at Chapters 7 and 10, the SVM weights highlight very different sets of features for the breathy vs. modal contrasts. This likely explains the extremely poor performance of the Gujarati classifier on the Mazatec data.

The six languages included here do not all use the same types of phonation; as mentioned above, Gujarati does not use creaky voice. Additionally, a classifier trained on Gujarati data expects to see roughly equal numbers of breathy and modal vowels, which is not the actual distribution in any data set. Similarly, a classifier trained on Mandarin expects to see approximately equal parts modal and creaky voice, and a classifier trained on English, Hmong, Mazatec, and Zapotec expects to see all three phonation types. This mismatch between the distribution of the training and testing sets can also contribute to low performance.

The vast majority of the weighted F1 scores reported in Table 12.8 are solidly in between the extremes; the average weighted F1 score for classifiers trained on one language and tested on another is 0.45029. This indicates that the patterns that the SVM finds for one language are generally not particularly applicable to another, again showing that what describes one phonation type in one language does not necessarily describe that phonation type in another.

12.4 Training on Five Languages and Testing on the Sixth

A twist on the above study is a sort of leave-one-out study: I train a classifier on five of the six languages and test on the one that was left out. This makes for a much larger and more varied training set; not only does the classifier see *more* instances of each class, but it sees *different* instances of each class. For example, the classifier is trained to identify modal vowels based on what modal vowels are in five languages. While this may make for a better classifier, if what's modal in the sixth language is different from what's modal in the other five, the classifier will struggle.

In order to train and test these classifiers, the methods again require slightly modification. I focus just on a Support Vector Machine, using a linear kernel and imbalanced data for both the training and testing data. When the testing language uses phonation tonally, I exclude f_0 from the features. However, I include features with 15% or more undefined measures, as excluding features that are problematic to the testing language would be artificially simplifying the problem. In the training set, those missing values are replaced with the class mean calculated across all five languages; in the testing set, they are replaced with the overall mean. Table 12.9 reports the performance of each classifier.

Table 12.9: Train on Five Languages, Test on One

Testing Language	Weighted F1 Score	Change
English	0.63464	-0.02273
Gujarati	0.46279	0.04234
Hmong	0.41847	0.0283
Mandarin	0.81066	0.27733
Mazatec	0.33669	0.05716
Zapotec	0.37456	0.17817

Despite the potential benefits of a larger training set, most of the classifiers perform badly and experience only a small improvement over the baseline weighted F1 score. This is further support for my previous findings that what describes voice quality in one language does not

necessarily describe it in another.

The classifier trained on all languages but English and tested on English performs below baseline. This may indicate that English distinguishes phonation types differently from other languages; the findings presented in Table 12.4 show that English uses some commonly used feature categories (HNR, VoPT, Energy) and some less commonly or rarely used categories (CPP, f_0, Surrounding Phones). English may also suffer because of numbers: the English data set is by far the largest, so the training set is smallest when English is excluded from it.

The classifier tested on Zapotec experiences a substantial boost over the baseline weighted F1 score, but overall still performs very poorly; while its predictions are more informed than random ones, training on the other five languages does not allow the classifier to successfully differentiate Zapotec's phonation types.

Finally, the Mandarin classifier both experiences a large boost in performance over baseline and has a very reasonable weighted F1 score. I suspect that this again comes from the various features that have been shown to correlate with Mandarin tones, such as Energy. We've seen that non-modal phonation in many languages is characterized in part by a drop in Energy, and this pattern is seen strongly in the Mandarin data as well.

Five of the six classifiers trained on five languages and tested on the sixth perform badly. A five-language training set, though larger and more varied than in previous experiments, does not do well when tested on a sixth language. The patterns may not match; what's creaky in the testing language may not be creaky in the training languages. Additionally, the combination of languages in the training data may cause noise; if there's overlap between, for instance, breathy voice in one language and creaky voice in another, the classifier will encounter difficulty separating these classes. The poor performance of the classifiers on a left-out language again shows that phonation in one language (or in five) may not match phonation in another.

12.5 Conclusions

This chapter compares how phonation classifiers perform for six different languages and what information those classifiers rely on. Generally good performance of the Support Vector Machines and Random Forests indicate that machine classification of voice qualities is possible and may be improved by tuning hyperparameters. (I try this out for English in the following chapter.) Classifier performance did not vary between languages that use phonation in different ways, though more directed methodology and a larger set of languages would provide a clearer answer. The set of important feature categories varies quite a bit from language to language. Two types of features – HNR and VoPT – are used by all six languages, though they're more important to some languages than to others. The classifiers are language-specific enough that they generally perform quite poorly when trained on one language and tested on another, or even when trained on five of the languages and tested on a sixth; what describes a breathy or creaky or modal vowel in one language does not necessarily describe it in another.

Part III

THE CLASSIFICATION QUESTION

Chapter 13

FINE-TUNING AN ENGLISH CLASSIFIER

The previous seven chapters have addressed the *linguistic* questions of which acoustic properties are most indicative of phonation types in different languages. I now shift gears to the *classification* question in which I aim to build a classifier for English voice qualities. Rather than examining *why* the classifiers work, I focus in this chapter on *how well* the classifiers work.

Sociolinguists have recently taken an interest in English phonation. Non-modal voice qualities have been associated with a wide variety of social meanings. Creaky voice has been linked with being bored, content, and unafraid (Gobl and Ní Chasaide, 2003), tough (Mendoza-Denton, 2011), and hesitant, non-aggressive, educated, urban-oriented, and upwardly mobile (Yuasa, 2010). Though most sociolinguistic work on phonation has focused on creaky voice, some have studied breathy voice. Breathy voice is associated with being timid and intimate (Gobl and Ní Chasaide, 2003).

These studies have relied on human annotators to identify voice qualities; someone must listen to the data and decide whether each segment is breathy, modal, or creaky. This process is both time-consuming and subjective; a classifier that can automate this process with some degree of reliability would allow sociolinguistic studies of English voice qualities to significantly increase both the amount of data as well as the objectivity of the annotation. Here, I aim to develop such a tool. Several other classification tools have been developed to perform similar tasks, with varying degrees of success. These tools are described in more detail in Section 13.4.

It is worth noting that the data used in this dissertation – used to develop this classification tool – suffers from the same drawbacks mentioned above: phonation type

was manually identified and is subjective. The first part is not a drawback, per se, in that the time has already been invested; over 10,000 English vowels have already been tagged for voice quality. Though the primary purpose of these vowels was to answer the linguistic question, the same vowels, along with their labels and features, can be leveraged to build a classifier. The subjectivity of the annotations, however, is potentially problematic. This is somewhat mitigated by using two phonetically-trained annotators with good inter-rater reliability (Cohen's Kappa 0.85). Still, training a classifier on somewhat subjectively annotated data does not make for a truly objective classifier, but it's a good place to start. In the future, comparing the annotations made by this classifier to those made by other linguists (not the two annotators who labeled the training data) would help shed light on the classifier's generalizability.

I separate this chapter into two primary goals for an English classifier. First, in Section 13.1, I aim for the highest performing classifier. In Section 13.2, I try to build a classifier that performs well but uses fewer features, making it less computationally expensive and more practical for use in sociolinguistic studies. Finally, I explore whether male and female speakers produce phonation types differently in English.

Both goals, the best possible classifier and the practical classifier, rely on English phonation types being produced consistently. If what's breathy is different from speaker to speaker, or if what's breathy varies from instance to instance within a speaker, even the best classifier will struggle to detect enough of a pattern. In Chapter 12, I hypothesized that when phonation type is not semantically important (as is the case in English), producing it consistently is also less important. This pattern did not emerge among the six languages in this dissertation, so I have hope that English voice quality is produced systematically enough to build a useable classifier.

I focus here on one of the two machine learning models used in this dissertation: the Support Vector Machine. SVMs are widely used, quite robust, and have a number of *hyperparameters*, essentially implementation options, that can be tweaked to improve the classifier. In Chapter X, I trained two SVMs to identify English voice quality: one using

an imbalanced data set and one using a resampled data set. Both classifiers used almost entirely default settings, with the exception of a linear kernel rather than the default RBF kernel. The data are normalized and all features containing 15% or more undefined measures are excluded.[1] The results of these two classifiers are reproduced in Table 13.1.

Table 13.1: Baseline Classification Metrics

Balance	Accuracy	Weighted F1	Precision			Recall			F-Score		
			B	M	C	B	M	C	B	M	C
Imbalanced	83.77	0.81831	0.49	0.86	0.74	0.17	0.96	0.54	0.25	0.91	0.63
Resampled	75.207	0.78767	0.21	0.94	0.66	0.71	0.79	0.62	0.32	0.86	0.63

I treat weighted F1 score as the single most informative metric; it captures both precision and recall while also considering imbalance between classes. In answering the linguistic questions, I have focused on classifiers trained on resampled data, as resampling avoids any bias due to imbalanced classes. However, we expect an English classifier to be encountering phonation types in an imbalanced way – modal voice is much more common in English than either breathy or creaky voice. Imbalance can therefore be useful when the goal is correct classification; if breathy voice occurs rarely, breathy voice should be guessed rarely. I will therefore rely on imbalanced data in building my classifiers. With this in mind, the baseline weighted F1 score going forward is 0.818.

13.1 Goal 1: A Highly Accurate Classifier

The first classifier I train in this chapter is intended to achieve the highest possible performance. I go about this in three ways: kernelizing, optimizing hyperparameters, and adding back in features with many undefined measures.

[1]See Chapter 4 for more information about methodology.

13.1.1 Kernelization

The baseline Weighted F1 Score of 0.818 comes from an SVM using nearly all default settings. The one exception is that I used a linear kernel. Intuitively, a linear kernel separates the classes linearly, essentially drawing straight lines between classes. The data, however, are rarely truly linearly separable. Two other kernels, *polynomial* and *Radial Basis Function (RBF)* kernels, can separate the data in more dimensions than the linear kernel, often making for a better classifier. However, there is a trade-off in using these often more accurate kernels: feature weights are no longer available. While giving up features weights would be a dealbreaker in answering the linguistic questions, it's a worthwhile sacrifice to make for the classification question. Table 13.2 lists the performance of three classifiers: one each using a linear, polynomial, and RBF kernel.

Table 13.2: Kernelizing

Kernel	Accuracy	Weighted F1	Precision			Recall			F-Score		
			B	M	C	B	M	C	B	M	C
Linear	83.77	0.81831	0.49	0.86	0.74	0.17	0.96	0.54	0.25	0.91	0.63
Polynomial	84.199	0.82202	0.59	0.86	0.79	0.22	0.97	0.51	0.32	0.91	0.62
RBF	85.046	0.83084	0.67	0.86	0.78	0.17	0.97	0.58	0.27	0.91	0.67

While both the polynomial and RBF kernels provide some boost over the linear kernel, the improvement is rather small; weighted F1 is only boosted from 0.818 to 0.831. These two kernels have hyperparameters – numbers that configure how the algorithm learns from the training data – that can be optimized to improve model performance. These hyperparameters are explored in the following section.

13.1.2 C and γ

C and γ (*gamma*) are two hyperparameters in polynomial and RBF kernels[2] relating to the cost of misclassified instances and the impact of individual instances, respectively.

C controls *regularization*, or the balance between complex decision boundaries[3] and misclassification. When C is small, the cost of misclassification is low; reducing the cost of misclassified instances can make the decision boundary smoother and less prone to overfitting. When C is large, the cost of misclassification is high. Increasing the cost of misclassified instances increases the risk of overfitting the decision boundaries to the data; the decision boundary basically wiggles around to group as many instances as possible with the correct class. In other words, the model can correctly classify more instances if it's more data-specific, and C regulates the tradeoff between misclassification and smoothness.

`Gamma` controls how much individual instances influence the model. When γ is large, instances near the decision boundary have more influence over the boundary's exact path than instances further from the boundary. When γ is small, instances that are far from the decision boundary also impact the boundary's path. If γ is high, instances near the boundary can have so much influence that overfitting is again a problem. Gamma regulates how much instances impact the decision boundary depending on their distance from that boundary.

Optimizing C and γ – picking the values that result in the best classifier – is performed by conducting a *grid search*. A grid search exhaustively searches through a range of values[4] for C and γ, training an SVM using each combination of the two parameters and evaluating the model's performance. The optimal values for C and γ can then be determined based on which values led to the best performing classifier. I determined the optimal C and γ based on the classifier with the highest Weighted F1 Score. Table 13.3 lists the performance of four

---

[2] C also applies to linear kernels, though I'll stick with polynomial and RBF kernels as they perform better without tuning than the linear classifier.

[3] Decision boundaries are the separations between classes – essentially the lines drawn to separate classes.

[4] For my grid search, I used six values of C and γ spaced logarithmically from 10^{-3} to 10^{2}: 0.001, 0.01, 0.1, 1.0, 10, and 100.

classifiers: SVMs using polynomial and RBF kernels with default and optimized values for C and γ.

Optimizing C and γ makes for more trivial improvements. The optimal values for both polynomial and RBF kernels are not on the extremes of the ranges used in the search, suggesting that the search range was inclusive enough to find the appropriate values. These optimized hyperparameters, however, only boost the classifier from a baseline weighted F1 score of 0.818 to 0.84.

13.1.3 Missing Values

Recall that my pre-processing steps, overviewed in Chapter 4, involved handling missing (*undefined*) values. Several of the features I use are difficult or impossible to calculate for extremely non-modal vowels, so I had many missing values to handle. In all previous chapters of this dissertation, I excluded all features with 15% or more missing values and replaced remaining missing values with the mean of the class on a fold-by-fold basis. Not only did this process lead to the exclusion of 44 features for English, it also excluded or replaced values for instances on the far ends of the phonation continuum, as extremely non-modal vowels are often the most problematic.

Here, I *include* features with 15% or more undefined values, increasing the number of features from 80 to 124. In the training set, missing values are replaced with the class mean on a fold-by-fold basis; in the testing set, they're replaced with the overall mean, as the by-class mean is unknown. I first try both polynomial and RBF kernels with default hyperparameters. The first two rows of Table 13.4 report the performance of the two classifiers that use the default hyperparameters and include features with 15% or more undefined values, with missing values in the testing data replaced by the overall mean.

Adding back in the features with many missing values *decreases* the performance of the models compared to the polynomial and RBF kernel SVMs with default parameters. In these models, missing values in the training data are replaced with the by-class mean, but missing values in the testing data are replaced with the overall mean, in order to simulate

Table 13.3: Optimizing Hyperparameters

Kernel	C	γ	Accuracy	Weighted F1	Precision			Recall			F-Score		
					B	M	C	B	M	C	B	M	C
Polynomial	*default*		84.199	0.82202	0.59	0.86	0.79	0.22	0.97	0.51	0.32	0.91	0.62
	0.01	0.1	84.448	0.831	0.49	0.87	0.78	0.31	0.96	0.54	0.38	0.91	0.64
RBF	*default*		85.046	0.83084	0.67	0.86	0.78	0.17	0.97	0.58	0.27	0.91	0.67
	10	0.01	84.997	0.84	0.5	0.88	0.77	0.34	0.95	0.58	0.41	0.91	0.66

not knowing their actual class. I also tried replacing missing values with the overall mean in both the training and testing sets. This led to some improvement, shown in the third and fourth rows of Table 13.4. Using an RBF kernel, this produces a weighted F1 score of 0.858, the highest so far.

Finally, I conduct a grid search for the RBF kernel to see how optimizing hyperparameters can improve performance. The optimal hyperparameters ($C = 100$ and $\gamma = 0.001$) bring the weighted F1 score up to 0.864.

Separately, I tried two other ways of handling missing values, though both were unsuccessful. Assuming that missing values are often caused by instances at the extremes of the phonation continuum, I tried replacing missing values in the test set with extreme values. The six values I tried (-3 to +3) did not improve the classifier. The second technique was to convert features with 15% or more missing values into binary measures; a missing value is converted to `0` and an actual number is converted to `1`.[5] Neither technique boosted the weighted F1 score of about 0.864.

Three strategies – kernelizing, optimizing hyperparameters, and including features with 15% or more missing values – impacted the classifier in various ways. Kernelization boosted the weighted F1 score from 0.818 to 0.83 (RBF kernel). Optimizing hyperparameters boosted it further to 0.84. Finally, adding back in features with 15% or more undefined measures and replaced those missing values with the overall mean, as well as optimizing hyperparameters, increased the weighted F1 score to 0.864.

The performance of this classifier, which I'll call the "best" classifier, is listed in Table 13.5 alongside the baseline classifier's performance.

Table 13.5 shows some improvement over the baseline classifier in almost all metrics. The one exception is in modal recall, which drops from 0.96 to 0.95. However, recall for the two minority classes is much improved. This classifier overall does a much better job handling modal and creaky voice than breathy voice. Even the "best" classifier looks more like a

[5]Remaining undefined measures from features with fewer than 15% undefined measures were replaced with the class mean in the training data and overall mean in the testing data.

Table 13.4: Including Missing Values

Missing Values	Kernel	Accuracy	Weighted F1	Precision			Recall			F-Score		
				B	M	C	B	M	C	B	M	C
Mean,	Polynomial	77.071	0.69545	0.4	0.78	0.68	0.09	0.99	0.08	0.14	0.87	0.14
by class	RBF	76.573	0.684	0.31	0.77	0.63	0.05	0.99	0.06	0.09	0.87	0.1
Mean,	Polynomial	85.355	0.83233	0.6	0.86	0.81	0.14	0.97	0.58	0.23	0.92	0.67
overall	RBF	87.389	0.8572	0.67	0.9	0.78	0.13	0.96	0.75	0.22	0.93	0.77
Optimized	RBF	87.13	0.864	0.51	0.91	0.77	0.3	0.95	0.73	0.38	0.93	0.75

Table 13.5: Classification Metrics: Baseline vs. Best Classifier

Classifier	Accuracy	Weighted F1	Precision			Recall			F-Score		
			B	M	C	B	M	C	B	M	C
Baseline	83.77	0.81831	0.49	0.86	0.74	0.17	0.96	0.54	0.25	0.91	0.63
Best	87.13	0.864	0.51	0.91	0.77	0.3	0.95	0.73	0.38	0.93	0.75

modal vs. creaky classifier than a true English voice quality classifier.

13.2 Goal 2: A Practical Classifier

The classifier built in Section 13.1 above achieves a weighted F1 score of 0.864. While this score indicates some confidence in the patterns the SVM is finding, it requires 124 features. Extracting 124 features is computationally expensive; to extract all the features from the English data set, VoiceSauce ran for several days and the Praat script took several hours on a computing cluster (it required too much computing power to run on a personal computer). We have reason to believe that not all 124 are necessary – we saw in Chapter 6 that some features are more useful than others, and that many of the features are collinear. In this section, I try to pare down the set of features while still maintaining a comparable weighted F1 score. I use an RBF kernel, keeping C and γ at 100 and 0.001 respectively, and including features with 15% or more undefined values[6]. I eliminate features in three steps: removing impractical features, redundant features, and finally unimportant features.

13.2.1 Step 1: Exclude Impractical Features

A good place to start for making the classifier more practical is to exclude features that are specific to this data set: the two features that rely on utterance boundaries. Recall that the ATAROS Corpus is annotated for *spurts*, which I treat as utterances. Requiring a data set to be annotated at the utterance-level seems like a high bar to set in order to use this classifier. Fortunately, excluding the two features that involve utterance boundaries – the

[6]I replace missing values in both the training and the testing sets with the overall mean.

vowel's distance from the end of the utterance in milliseconds and as a percent – does not cause a large drop in the weighted F1 score, even though English non-modal phonation can occur at prosodic boundaries. When the model does not consider the vowel's distance from the end of the utterance, using a total of 122 features, the weighted F1 score is 0.865. This is a small increase from the SVM with all 124 features!

Many corpora, ATAROS included, have both word- and phone-level annotations. However, the fewer requirements of the corpus, the more widely useable the classifier will be. Most of the features depend only on the boundaries of the vowel in question, but eight features require more information than that. Those eight are the vowel's distance from the end of the word (as a percent and in milliseconds) and the voicing, manner, and presence of the surrounding phones. Again, these features are apparently not doing much of the legwork; the classifier has a Weighted F1 Score of 0.864 without them.

13.2.2 Step 2: Exclude Redundant Features

With ten impractical features excluded from the list, 114 features still remain. Recall from Chapter 6 that many of the features are collinear; they provide essentially the same information. Collinearity is introduced by including variations on a calculation (like APQ3 Shimmer and APQ5 Shimmer) as well as the same measurement made over different time periods. When two or more features are redundant, their presence does not add value to the model. Here, I try to reduce the set of features by eliminating redundant features.

The 114 remaining features fall into eleven categories. These categories are listed in Table 13.6, along with the number of features in those categories. Nine of these eleven categories have multiple measures, and these measures may be redundant with each other.

I aim to narrow down the feature set to one feature per category. To do this, I ablate features within a category iteratively. For CPP, for instance, I train an SVM without CPP_1, then add CPP_1 back in and train another without CPP_2, and so on, until each CPP feature has been excluded once. Whichever feature resulted in the *highest* weighted F1 score (that is, caused the smallest drop), I remove permanently. I then ablate the remaining three features

Table 13.6: Features Per Category

Category	Number of Features
Spectral Tilt	28
CPP	4
RMS Energy	4
HNR	16
SHR	4
f_0	16
F1	4
Vowel Duration	1
Jitter	16
Shimmer	20
VoPT	1

and remove the feature whose ablation caused the smallest drop in F1. I repeat this process until either there is one feature left in the category, or I reach a weighted F1 score that is too low for a reasonable classifier. I complete this process separately for each category; that is, after determining which CPP feature(s) to keep, I put all CPP features back in the mix before going through the same process for RMS Energy.

This process reduces the number of features within each category significantly. For each category, I was satisfied with the Weighted F1 Score from the final feature, meaning that each category has been pared down to just one feature. This process allowed me to identify a set of eleven features that should not contain any redundancy. These 11 features are listed in Table 13.7; two feature categories have only one feature in the first place, so there's no paring down to be done.

The weighted F1 scores in Table 13.7 use one feature from each category and *all* features from the other categories. An SVM trained using just those eleven features will not necessarily have as much success. The weighted F1 score of an SVM trained on these eleven features is 0.821, a drop from the "best" classifier's score of 0.864. This is already a bit lower than I'd like.

Table 13.7: Single Category Models

Category	Best Feature(s)	F1
Spectral Tilt	H1* – A1*_3	0.854
CPP	CPP_Mean	0.863
RMS Energy	RMSE_1	0.864
HNR	HNR15_2	0.862
SHR	SHR_Mean	0.865
f_0	STRAIGHT_f_0_Mean	0.857
F1	F1_3	0.863
Vowel Duration	Vowel Duration	–
Jitter	Local_Jitter_3	0.866
Shimmer	APQ11_Shimmer_3	0.869
VoPT	VoPT	–

Two of the eleven categories in Table 13.7 have slightly lower weighted F1 scores than the others: Spectral Tilt and f_0. Replacing all features in those two categories while keeping just one from the others (53 total features) brings the weighted F1 score up to 0.856. This reintroduces quite a bit of redundancy from those two categories. To reduce this redundancy (but not eliminate it, as eliminating it causes the weighted F1 to drop too much), I treat Spectral Tilt and f_0 as a single category and perform the same ablation process.

For now, I aim to keep the weighted F1 score at or above 0.85. This happens with 29 features. A few combinations of 29 features keeps the weighted F1 score at this point, but the best combination gets a score of 0.85162.

13.2.3 Step 3: Exclude Unimportant Features

We saw in Chapter 6 that the features are not all equally important in distinguishing English's three phonation types. The strategy I used above to reduce redundancy above also eliminated many features that were doing less of the work. Nevertheless, it seems unlikely that the remaining twenty nine features are all doing the same amount of work. Though twenty nine features is a reasonable number for a classifier, removing features that are not contributing

Table 13.8: 29 Features for 0.85162 Weighted F1 Score

Category	Feature
Spectral Tilt	H1* – A1*_3
	H1* – A2*_Mean
	H1* – A2*_3
	H1* – A3*_1
	H1* – H2*_2
	H1* – H2*_3
	H2* – H4*_Mean
	H2* – H4*_2
	H2* – H4*_3
	2k* – 5k_1
	2k* – 5k_2
	2k* – 5k_3
	H4* – 2k*_1
CPP	CPP_Mean
RMS Energy	RMSE_1
HNR	HNR15_2
SHR	SHR_Mean
f_0	Praat_f_0_2
	Snack_f_0_Mean
	Snack_f_0_1
	Snack_f_0_3
	SHR_fo_1
	STRAIGHT_f_0_Mean
	STRAIGHT_f_0_3
F1	F1_3
Vowel Duration	Vowel Duration
Jitter	Local_Jitter_3
Shimmer	APQ11_Shimmer_3
VoPT	VoPT

much, if anything, to the classification task will make the classifier more widely useable.

I follow the same methodology used in Step 2 to further pare down the remaining features. I remove one feature (no longer working on a single category at a time), train an SVM, and then replace that feature. I then permanently exclude the feature whose ablation resulted in the smallest drop in weighted F1 score, and repeat. The goal is to exclude as many unimportant features as possible but keep the weighted F1 score fairly close to its current value: 0.85162. I don't have a specific number of features in mind, but I would like to keep the weighted F1 score around 0.84.

Table 13.9 lists the drop in weighted F1 score as I ablate features. The list is ordered; features ablated first cause the lowest drop in weighted F1. Once ablated, a feature is not replaced.

Table 13.9: Ablating Unimportant Features

Ablated Feature	Weighted F1
H1* – A2*_Mean	0.84993
STRAIGHT_f_0_3	0.84919
Snack_f_0_3	0.84948
2k* – 5k_3	0.8488
Local_Jitter_3	0.84829
H1* – H2*_2	0.84682
H1* – A1*_3	0.84508
Praat_f_0_2	0.84559
Snack_f_0_1	0.84494
APQ11_Shimmer_3	0.84351
RMSE_1	0.84428
H2* – H4*_2	0.84309
H1* – A3*_1	0.84174
H1* – H2*_3	0.8405
SHR_Mean	0.8401

Fifteen features can be ablated without the weighted F1 score dropping below 0.84. In other words, a weighted F1 score of 0.84 can be achieved with fourteen features, listed in Table 13.10. These fourteen features come from six categories: Spectral Tilt, CPP, HNR, f_0, Vowel Duration, and Variance of Pitch Tracks.

Table 13.11 compares the performance of the three classifiers that I've focused on in this chapter: the baseline classifier, "best" classifier, and this one, the "practical" classifier.

Compared to the baseline classifier, the "practical" classifier is better by almost all metrics. However, compared to the "best" classifier, the "practical" classifier generally does not perform as well. The biggest exception to this is breathy Precision, which *increases* from 51% to 62%. This increase, however, comes at the expense of breathy Recall, which drops from 38% to 18%; the classifier is finding fewer instances of breathy voicing, but it's right

Table 13.10: 14 Features for 0.8401 Weighted F1 Score

Category	Feature
Spectral Tilt	H1* – A2*_3
	H2* – H4*_Mean
	H2* – H4*_3
	2k* – 5k_1
	2k* – 5k_2
	H4* – 2K*_1
CPP	CPP_Mean
HNR	HNR15_2
f_0	Snack_f_0_Mean
	SHR_f_0_1
	STRAIGHT_f_0_Mean
F1	F1_3
Vowel Duration	Vowel Duration
VoPT	VoPT

Table 13.11: Classification Metrics: Baseline vs. Best Classifier

Classifier	Accuracy	Weighted F1	Precision			Recall			F-Score		
			B	M	C	B	M	C	B	M	C
Baseline	83.77	0.81831	0.49	0.86	0.74	0.17	0.96	0.54	0.25	0.91	0.63
Best	87.13	0.864	0.51	0.91	0.77	0.3	0.95	0.73	0.38	0.93	0.75
Practical	85.954	0.8401	0.62	0.88	0.75	0.1	0.96	0.68	0.18	0.92	0.72

about them slightly more often.

This classifier handles modal vowels well; it finds 96% of them and is correct about 88% of what it labels as such. These figures are substantially lower for creaky voicing, and so bad for breathy voicing that it's not a useable classifier.

While I feel that this classifier is ready for use on modal vowels and perhaps creaky vowels, I would not use it to automatically identify breathy vowels. Before discussing potential reasons that this classifier was unsuccessful, I try one other take in the classification question: gender-specific classifiers.

13.3 An Aside: Gender

I have so far trained and tested all classifiers on vowels produced by both male and female speakers. It's possible, however, that male and female speakers do not produce voice qualities in the same way. Biological sex impacts the larynx, which is where phonation happens; for example, men have longer vocal folds and a thicker vocalis muscle than women, which can lead to a lower oscillatory rate, and have some acoustic characteristics, such as Spectral Tilt, that lie on the creakier end of the spectrum (Stevens, 2000). Additionally, sociolinguistic studies have suggested that voice quality indexes different social characteristics for men and women (Mendoza-Denton, 2011; Yuasa, 2010). It's therefore not a given that the characteristics of different phonation types are the same for male and female speakers.[7]

If male speakers produce phonation types differently from female speakers, combining their data, as I have been doing so far, introduces noise. This noise would make the classifier's task more challenging, as a vowel of a given phonation type may look different depending on whether it's produced by a male or a female speaker. Separating the data by speaker gender and testing and training SVMs separately for male and female speakers provides a way to test this. If there's a gender-specific difference, the classifiers will likely perform better when trained and tested on a single gender. If there's not a gender-specific difference, I expect to see a drop in classifier performance, as it's being trained on fewer data points.

I tried both classifiers – the "best" classifier using all features and the "practical" one using fourteen features – on gender-specific data. In other words, I trained and tested on just male data, and then on just female data. Before reviewing how these classifiers perform, it's important to consider how the data set has changed by splitting it by speaker gender. Table 13.12 breaks down the two new data sets by phonation type.

Given that about 82% of the vowels produced by male speakers are modal, the classifier could be 82% accurate by guessing modal for all instances. Of course, guessing that all

[7]I've identified reasons to expect that phonation may vary by both sex and gender. All speakers in the ATAROS Corpus (Freeman, 2015) are cisgender; that is, their biological sex matches their gender identity.

Table 13.12: Phonation Distribution by Gender

Gender	B	M	C	Total
Male	234	3960	664	4858
	4.817%	*81.515%*	*13.668%*	
Female	342	3671	1160	5173
	6.611%	*70.964%*	*22.424%*	

instances are modal doesn't mean that the classifier has learned anything. In interpreting the results of the gender-specific classifiers, listed in Table 13.13, we need to treat baseline accuracy as 81.515 for male speakers and 70.964 for female speakers.

Table 13.13: Gender-Specific Classifiers

	Gender	Accuracy	Weighted F1	Precision			Recall			F-Score		
				B	M	C	B	M	C	B	M	C
Best	All	87.13	0.864	0.51	0.91	0.77	0.3	0.95	0.73	0.38	0.93	0.75
	Male	82.091	0.82077	0.37	0.9	0.49	0.27	0.9	0.56	0.31	0.9	0.53
	Female	85.54	0.85207	0.5	0.9	0.79	0.41	0.93	0.76	0.45	0.91	0.78
Practical	All	85.954	0.8401	0.62	0.88	0.75	0.1	0.96	0.68	0.18	0.92	0.72
	Male	83.882	0.8179	0.46	0.88	0.56	0.05	0.94	0.5	0.09	0.91	0.53
	Female	85.502	0.84325	0.61	0.89	0.78	0.25	0.95	0.74	0.35	0.91	0.76

For both classifiers, handling the two genders separately leads to a drop in nearly every single metric compared to the classifiers that use data from both genders. Because the classes are imbalanced, and imbalanced in different ways in the two data sets, the weighted F1 score is the best metric to compare the classifiers. For both classifiers (the 124 feature "best" classifier and the 14 feature "practical" classifier), the female classifier performs slightly better than the male classifier. Unsurprisingly, both the male and female "best" classifiers perform slightly better than the male and female "practical" classifiers.

Let's consider the fact that the female classifiers perform better than the male classifiers. Table 13.12 shows that the male data set contains more modal instances than the female data set, both in raw numbers and as a percent. That said, the two female classifiers have equal

or slightly higher modal precision and recall than the two male classifiers. This suggests that the patterns in female modal voicing are more consistent than in male modal voicing, as the classifier performs better despite fewer examples. While the female data contains fewer modal tokens than the male data, it contains more of the two minority classes than the male data; as expected, the female classifiers handle breathy and modal vowels better than the male classifiers.

I interpret the overall similarity between the male and female classifiers as suggesting that there are no strong gender-specific trends in how voice qualities are produced. It's possible, however, that gender-specific patterns do exist but that their boost is overshadowed by the drop in weighted F1 score due to effectively dividing the training data in half. Evidence to this end comes from training a classifier on one gender's data and testing it on another's.

If the different phonation types produced by male versus female speakers differ greatly in the features that the classifiers rely on, we can expect a classifier trained on one gender and tested on the other to perform poorly. If phonation types are produced similarly, a classifier trained on one gender and tested on the other will not look much different from the original classifier. Table 13.14 shows what happens when each classifier is trained on one gender's data and tested on the other's.

Table 13.14: Gender-Specific Confusion Matrix: Weighted F1 Score

		Test			
		Best Classifier		Practical Classifier	
		M	F	M	F
Train	M	–	0.75348	–	0.74251
	F	0.82983	–	0.83276	–

All four classifiers perform slightly worse when trained on one gender's data and tested on the other gender's data. I see two potential explanations for these small drops in weighted F1 score. One possibility is that some of characteristics of phonation types for one gender are not applicable to phonation types for the other gender. Another possibility is that the

class imbalance causes the change in performance; a male classifier expects to find relatively more instances of modal voicing than exist in the female data.

Gender-specific classifiers do not improve performance, but training on one gender and testing on another does worsen performance. This section is inconclusive as to whether male and female speakers of American English produce phonation types differently.

13.4 Classification Conclusions

In this chapter, I built two classifiers with slightly different goals. The "best" classifier aimed to achieve the best possible performance. To do so, I kernelized the SVM, optimized hyperparameters C and γ, and added back in features with 15% or more undefined measures, replacing missing values with the overall mean in both the training and testing set. The "practical" classifier aimed to achieve high performance with few features. I reduced the feature set by strategically excluding impractical, redundant, and unimportant features. The performance of these two classifiers is listed in Table 13.15.

Table 13.15: Best and Practical Classifiers

Classifier	Accuracy	Weighted F1	Precision			Recall			F-Score		
			B	M	C	B	M	C	B	M	C
Best	87.13	0.864	0.51	0.91	0.77	0.3	0.95	0.73	0.38	0.93	0.75
Practical	85.954	0.8401	0.62	0.88	0.75	0.1	0.96	0.68	0.18	0.92	0.72

The "best" classifier handles modal voicing quite well, handles creaky voice decently, and does quite poorly with breathy voicing. However, it requires a whopping 124 features that are computationally expensive to extract, limiting the classifier's practicality. The "practical" classifier sacrifices some performance for all phonation types, particularly for breathy voice, but requires only fourteen features. Ultimately, neither classifier seems ready for use.

Several other studies have aimed to do some variation of this phonation classification task. These studies, some of which I describe below, differ from the present study in three primary ways. First, the vast majority of voice quality classifiers focus only on modal and creaky

voice, to the exclusion of breathy voice. Second, they have generally not been language-specific; as seen in Chapter 12, a given phonation type may have very different properties from language to language, making a universal voice quality classifier a lofty goal. Finally, other classifiers consider larger segments, often not mapping directly to individual phones, while I only consider vowels.

Ishi et al. (2008) developed a method for detecting creaky voice that uses three acoustic features. First, "power peaks" are identified over very short frames; looking across frames, the amplitude of the peak is compared to the peak amplitude of surrounding frames. The power peak is the largest difference between peaks. Second, "intraframe periodicity" helps identify modal segments. Finally, an "interpulse similarity measure" helps separate "impulsive noises" from creaky glottal pulses by seeing how similar they are; glottal pulses are expected to be more similar to one another than pulses due to noise. On a small data set of Japanese conversational speech, this method achieved 74% accuracy; my "best" classifier, which distinguishes breathy voicing as well as modal and creaky voicing, achieves 87.13% accuracy.[8]

The Resonator-based Creaky Voice Detection (RCVD) technique (Drugman et al., 2012) relies on secondary peaks characteristic of creaky voicing caused by "secondary laryngeal excitations and also from sharp discontinuities at glottal opening, following a long glottal closed phase" – essentially subharmonics. It uses H2 – H1[9] to measure the importance of these secondary peaks. The RCVD technique was developed using a small set of male and female speakers of Finnish, American English, and Scottish English, with human-annotated creaky voicing. For each of three speakers, the best F1 score ranges from just under 0.7 to just over 0.8. Like the classifier developed by Ishi et al. (2008), the RCVD technique classifies creaky vs. non-creaky but does not specifically look at breathy voice.

The two methods described above are combined in Drugman et al. (2014); they train a

[8]They do not report their weighted F1 score. Drugman et al. (2012) tested this classifier on their three speakers; resulting F1 scores range from about 0.25 to 0.55.

[9]Note that while this the opposite of H1 – H2, it conveys the same information.

Binary Decision Tree and an Artificial Neural Network to identify creaky voicing. Their features are H2 – H1, Peak-Prom ("characterises the prominence of LP-residual peaks relative to its immediate neighbourhood"), F0Creak (the fundamental frequency of one of the resonators), as well as those described by Ishi et al. (2008). Three additional features were used to detect "silences and sudden bursts" (as they were not looking at individual vowels, as I was). Their corpus include Swedish, American English, and Japanese data, which have been manually annotated for phonation type (creaky and non-creaky). The ANN trained on all features performed best for ten of eleven speakers, with F1 scores between about 0.6 and 0.85.

Both my "best" and "practical" classifiers perform similarly to, if not better than, the state-of-the-art voice quality classifiers. Additionally, they perform the more complex task of making a three-way phonation distinction. However, both of my classifiers rely on a larger set of features than those described above. Additionally, I focus on just one dialect of one language; how my classifier performs on other dialects of American English remains unclear.

In Chapter 12, I hypothesize that when voice quality is not lexically important, as in English, there will be more variation in how different phonation types are produced. Both inter- and intra-speaker variation is possible; what's breathy for one speaker may not be the same as what's breathy for another speaker, and what's breathy may vary from instance to instance for a given speaker. Increased variation in how phonation types are produced requires more data for a classifier to find patterns. I suspect that the amount of data I used was not sufficient to find these complex patterns. This is supported by the fact that the classifier performs better on classes for which there's more data. It's also possible that the features I use do not capture what best distinguishes English voice qualities. That said, weighted F1 scores of 0.864 and 0.8401 show that there is some systematicity to English voice qualities.

This chapter has focused on the *classification question* of how good of a classifier I can build for English phonation types. I split this question into two parts, first working towards a very high performing classifier and second working towards a classifier that would require

fewer features without sacrificing much performance. The "best" classifier has a weighted F1 score of 0.864, and while I was able to reduce the number of features significantly without losing much in the weighted F1 score, both struggle too much with breathy voice to be ready for use.

Part IV

CONCLUSIONS

Chapter 14

CONCLUSION

This dissertation has examined the acoustic properties of phonation types in different languages through the lens of machine learning. To conclude, I will briefly review the major findings, first for the linguistic question and then for the classification question. Finally, I will discuss the contributions this study has made and explore directions for possible future work on phonation.

14.1 Summary of Results

The primary goal of this dissertation was to understand which acoustic properties best describe phonation types in different languages, and how those properties vary from language to language. To answer this question, I trained and tested Support Vector Machines and Random Forests to classify phonation types in six languages, outlined in Table 14.1. I examined feature correlations, SVM weights, Random Forest importance, and the results of ablation testing to evaluate which features were most important in distinguishing phonation types.

Table 14.1: Summary of Data Sets

Language	Phonation Types	Tokens	Phonation Use(s)	Weighted F1 Score	
				SVM	RF
English	B, M, C	10031	Allophonic, Prosodic, Sociolinguistic	0.78767	0.78895
Gujarati	B, M	2873	Contrastive	0.72079	0.72694
Hmong	B, M, C	2717	Alongside Tones	0.6444	0.66723
Mandarin	M, C	180	Alongside Tones	0.96119	0.96134
Mazatec	B, M, C	482	Contrastive	0.6994	0.66891
Zapotec	B, M, C	344	Alongside Tones	0.6959	0.62859

The English classifiers performed fairly well: the SVM and Random Forest, based on resampled data, both had weighted F1 scores of about 0.79. Cepstral Peak Prominence and Harmonics-to-Noise Ratio proved the most useful features, suggesting that noise is key to differentiating English phonation types. f_0 (particularly as calculated by the Snack algorithm) and Variance of Pitch Tracks were also useful features, and Energy and Surrounding Phones were useful but to a lesser degree.

The SVM and Random Forest trained on Gujarati, which contrasts breathy and modal voice, performed reasonably well but not as well as English. Their weighted F1 scores were both around 0.72. Duration and Shimmer were identified as the best cues to Gujarati phonation. Despite the fact that both of these measures have been associated with Gujarati breathy voicing, neither has typically been used in studies of the subject. Cepstral Peak Prominence and Spectral Tilt were also useful, as were Harmonics-to-Noise Ratio and Variance of Pitch Tracks. Consistent with previous findings that Gujarati phonation is most salient around the vowel's midpoint, many of the features that best distinguished Gujarati voice qualities were measured in the vowel's middle third.

The Hmong classifiers were a step down from the Gujarati classifiers; the SVM's weighted F1 score was 0.6444 and the Random Forest's was 0.66723. They relied primarily on Harmonics-to-Noise Ratio (particularly from 0 to 500 Hz), with Subharmonic-to-Harmonic Ratio, Spectral Tilt, and Energy playing a smaller role and Jitter and Variance of Pitch Tracks an even smaller one.

The SVM and Random Forest were quite successful on the Mandarin data, both achieving weighted F1 scores of about 0.96. Many features contributed to this high performance, particularly Harmonics-to-Noise Ratio, Jitter, and Energy. Also important were Variance of Pitch Tracks, Shimmer, and Spectral Tilt. These features generally were more useful when measured in the middle of the vowel.

Mazatec's SVM and Random Forest were less impressive than Mandarin's, with weighted F1 scores of 0.6994 and 0.66891, respectively. Spectral Tilt and f_0 were the most crucial features for Mazatec, followed by Jitter and Variance of Pitch Tracks, and then by Energy and

Harmonics-to-Noise Ratio. Many of the most important Mazatec features were calculated over the first or middle third of the vowel, consistent with previous findings that Mazatec phonation is realized at the beginning of the vowel (Silverman, 1997).

Finally, the Zapotec classifiers were the lowest-performing classifiers in the dissertation. The SVM's weighted F1 score was 0.6959 and the Random Forest's was 0.62859. The Random Forest in particular struggled to find meaningful patterns in the data. What these classifiers did find was that Spectral Tilt and Harmonics-to-Noise Ratio were the best predictors of Zapotec phonation type, followed by Energy and Variance of Pitch Tracks, with Jitter and Cepstral Peak Prominence also useful but playing a smaller role.

Looking across these six languages, it is clear that the set of acoustic properties that best distinguishes different voice qualities varies significantly across languages. That said, some trends appear. HNR and VoPT were useful in all six languages, though to varying degrees. Looking at specific contrasts, Spectral Tilt was consistently useful in distinguishing breathy voice from creaky voice, and VoPT often useful in distinguishing creaky voice from modal voice.

The six languages in this dissertation use phonation in different ways: contrastively, alongside tones, and sociolinguistically, prosodically, and allophonically. Given that contrastiveness of segments often limits variability, I looked for patterns in classifier performance between languages that use phonation in different ways. In this small sample, classifiers did not perform better on languages that use phonation contrastively than on languages that use phonation in other ways.

Finally, I addressed the classification question, in which the goal was to build a high performing English voice quality classifier. A tool to automatically identify phonation types in English would allow sociolinguists to use larger data sets and more objective phonation annotations. By kernelizing, optimizing hyperparameters γ and C, and reconsidering how I handled missing values, I was able to boost the SVM's weighted F1 score to 0.864. I also pared down the feature set to include just fourteen features, sacrificing some performance (a weighted F1 score of 0.8401) but making a more practical classifier. However, both classifiers

handled breathy voicing rather poorly. This is likely a data sparsity problem, though perhaps a sign that breathy voicing may be more difficult to distinguish from modal voicing than creaky voicing is. Ultimately, though, these classifiers' difficulty with breathy voice means that they are not yet ready for use.

14.2 Contributions and Future Directions

This dissertation has shown that machine learning can be an effective way to tackle the complex problem of phonation. It can be used not only to build an automatic classifier, but also to illuminate which acoustic properties are crucial to the classification task. I have shown that those properties can vary widely across languages, highlighting the importance of treating languages separately in studies of phonation. As phoneticians rely more and more on large data sets and automatic measures, machine learning can be a powerful tool to discover complex patterns.

The English voice quality classification tool developed in Chapter 13 shows promise for automating phonation labeling. I believe that its good performance indicates that reliable automatic phonation type identification is possible for English. It would benefit from more training data – especially instances of breathy voice – as well as more fine-tuning. With these improvements, sociolinguistic studies would be able to include much larger data sets and increase the power of their findings. Even in its current state, the classifier could be used as a preliminary tool to aid researchers in annotation.

In using machine learning, this dissertation differs from most other studies of phonation. Most only consider the acoustic properties that human listeners rely on in distinguishing phonation types, but I relied instead on properties that an algorithm finds useful. While many of these overlap, two features stand out. First, Jitter has fallen out of favor in studies of phonation because, while it's a direct representation of aperiodicity, humans perceive that aperiodicity better as Harmonics-to-Noise Ratio or Cepstral Peak Prominence (Kreiman and Gerratt, 2005). However, when an algorithm performs the classification, Jitter is a useful feature; it was relatively important to distinguishing phonation types in four of the six

languages studied here. Second, Variance of Pitch Tracks has no direct perceptual correlate but was useful in all six languages.

Variance of Pitch Tracks is a new measure presented in this dissertation. In Chapter 3, I described several variants of this measure that I tested on the English data before settling on which calculation to use. Considering how well VoPT performed across languages in this study, I believe it warrants more investigation. Different calculations may perform better in different languages, and it may be a useful feature for studying acoustic phenomena besides phonation.

Finally, this dissertation has produced a large data set that can be used for future research on phonation. I have extracted a total of 114 acoustic measurements relating to phonation for a total of 16,727 vowels, uttered by 101 speakers of six languages. This data set can be used to answer a variety of questions, including questions of inter- and intra-speaker variation in the production of different phonation types.

This dissertation used machine learning to study the complex phenomenon of phonation in different languages. Applying a new methodology to this problem has confirmed several smaller-scale previous findings and led to several new insights into the cross-linguistic nature of phonation.

BIBLIOGRAPHY

Andrus, J. N. (2011). A sociophonetic investigation of creaky voice in Pacific Northwest English. Master's thesis, University of Washington.

Andruski, J. E. and Ratliff, M. (2000). Phonation types in production of phonological tone: the case of Green Mong. *Journal of the International Phonetic Association*, 30(1/2):37–61.

Avelino Becerra, H. (2004). *Topics in Yalálag Zapotec, with particular reference to its phonetic structures*. PhD thesis, University of California, Los Angeles.

Belotel-Grenie, A. and Grenie, M. (1994). Phonation types analysis in Standard Chinese. In *3rd International Conference on Spoken Language Processing*.

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics*, 30:163–191.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Science*, volume 17, pages 97–110.

Boersma, P. and Weenink, D. (2016). Praat: doing phonetics by computer.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92.

Brock-Nannestad, G. and Fontaine, J.-M. (2008). Early use of the Scott-Koenig phonautograph for documenting performance. *Journal of the Acoustical Society of America*, 123:6240–6244.

Cao, R., Wayland, R., and Kaan, E. (2012). The role of creaky voice in Mandarin Tone 2 and Tone 3 perception. In *Interspeech*.

Carlson, R., Hirschberg, J., and Swerts, M. (2005). Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46:326–333.

Catford, J. (1964). Phonation types: the classification of some laryngeal components of speech production. In Abercrombie, D., Fry, D. B., MacCarthy, P. A. D., Scott, N. C., and Trim, J. L. M., editors, *In honour of Daniel Jones*. Longmans.

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. University of California Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22:129–159.

Davison, D. S. (1991). An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics*, 78:50–57.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research*, 36:254–266.

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association*, 39(2).

Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24:423–444.

Drugman, T., Kane, J., and Gobl, C. (2012). Resonator-based creaky voice detection. In *Interspeech 2012*.

Drugman, T., Kane, J., and Gobl, C. (2014). Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech and Language*, 28.

Eadie, T. L. and Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice*, 20(4):527–544.

Edmondson, J. A. and Esling, J. H. (2006). The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. *Phonology*.

Esposito, C. (2010). Variation in contrastive phonation in Santa Ana Del Valle Zapotec. *Journal of the International Phonetic Association*, 40(2):181 – 198.

Esposito, C. (2012). An acoustic and electroglottographic study of of White Hmong tone and phonation. *Journal of Phonetics*, 40:466–476.

Esposito, C. M. (2004). Santa Ana Del Valle Zapotec phonation. *UCLA Working Papers in Phonetics*, (103):71 – 105.

Esposito, C. M. (2006). *The Effects of Linguistic Experience on the Perception of Phonation*. PhD thesis, University of California at Los Angeles.

Fagan, J. L. (1988). Javanese intervocalic stop phonemes: the light/heavy distinction. In McGinn, R., editor, *Studies in Austronesian Linguistics*. Ohio University Center for International Studies.

Fant, G. (1980). Voice source dynamics. In *Department for Speech, Music and Hearing Quarterly Progress Report*, volume 21, pages 017–037. KTH Royal Institute of Technology.

Farrús, M., Hernando, J., and Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. In *The 8th Annual Conference of the International Speech Communication Association*.

Feaster, P. (2010). The phonautographic manuscripts of Édouard-Léon Scott de Martinville.

Fougeron, C. and Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 101(6).

Freeman, V. (2015). *The Phonetics of Stance-Taking*. PhD thesis, University of Washington.

Fulop, S. A. and Golston, C. (2009). Breathy and whispery voicing in White Hmong. *Journal of the Acoustical Society of America*, 123(5):3883–3883.

Garellek, M. (2012). The timing and sequence of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics*, 40:152 – 161.

Garellek, M. and Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, 41(2):185–205.

Garellek, M. and Keating, P. (2015). Phrase-final creak: Articulation, acoustics, and distribution. In *The Linguistic Society of America Annual Meeting*, Portland, OR.

Garellek, M. and Seyfarth, S. (2016). Acoustic differences between English /t/ glottalization and phrasal creak. In *Interspeech 2016*, pages 1054–1058, San Francisco, CA.

Gerfen, C. (2013). *Phonology and Phonetics in Coatzospan Mixtec*. Springer.

Gerfen, C. and Baker, K. (2005). The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics*, 33:311–334.

Gick, B., Wilson, I., and Derrick, D. (2013). *Articulatory Phonetics*. Blackwell.

Gill, H. S. and Gleason Jr., H. A. (1969). *A Reference Grammar of Punjabi*. Department of Linguistics, Punjabi University, Patiala.

Gobl, C. and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood, and attitude. *Speech Communication*, 40:189–212.

Gordon, M. (2001). Laryngeal timing and correspondence in Hupa. *UCLA Working Papers in Linguistics*, 7.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29:283–406.

Gramuglia, A. C. J., Tavares, E. L., Rodrigues, S. A., and Martins, R. H. (2014). Perceptual and acoustic parameters of vocal nodules in children. *International Journal of Pediatric Otorhinolaryngology*, 78:312–316.

Gruber, J. F. (2011). *An articulatory, acoustic, and auditory study of Burmese tone*. PhD thesis, Georgetown University.

Hall, K. D. (1995). Variations across time in acoustic and electroglottographic measures of phonatory function in women with and without vocal nodules. *Journal of Speech and Hearing Research*, 38:783–793.

Heman-Ackah, Y. D., Heuer, R. J., Michael, D. D., Ostrowski, R., Horman, M., Baroody, M. M., Hillenbrand, J., and Sataloff, R. T. (2003). Cepstral peak prominence: A more reliable measure of dysphonia. *The Annals of Otology, Rhinology and Laryngology*, 112(4):324–333.

Henton, C. and Bladon, A. (1988). Creak as a sociophonetic marker. In *Language, Speech, and Mind: Studies in Honor of Victoria A. Fromkin*. London: Routledge.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America*, 83(6):2361–2371.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, 37:769–1994.

Hombert, J.-M. (1978). Consonant types, vowel quality, and tone. In Fromkin, V., editor, *Tone: A linguistic survey*. Academic Press.

Hombert, J.-M., Ohala, J. J., and Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, 55(1):37–58.

Horii, Y. (1980). Vocal shimmer in sustained phonation. *Journal of Speech, Language, and Hearing Research*, 23:202–209.

Huffman, M. K. (1987). Measure of phonation type in Hmong. *Journal of the Acoustical Society of America*, 81:495–504.

Huffman, M. K. (2005). Segmental and prosodic effects on coda glottalization. *Journal of Phonetics*, 33:335–362.

Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., and Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 16(1).

Jackson, M., Ladefoged, P., Huffman, M., and Antoñanzas-Barroso, N. (1985). Measures of spectral tilt. *UCLA Working Papers in Phonetics*, 61:72–78.

Johnson, K. (2011). *Acoustic and Auditory Phonetics*. Wiley-Blackwell, 3rd edition.

Kawahara, H., de Cheveign, A., and Patterson, R. D. (1998). An instantaneous-frequency-based pitch extraction method for high quality speech transformation: Revised TEMPO in the STRAIGHT-suite. In *Proceedings of ICSLP*.

Keating, P. (2012). Production and perception of linguistic voice quality.

Keating, P., Esposito, C., Garellek, M., ud Dowla Khan, S., and Kuang, J. (2011). Phonation contrasts across languages. In *Proceedings of the 17th International Congress of Phonetics Sciences*, pages 1046–1049.

Keating, P. and Garellek, M. (2015). Acoustic analysis of creaky voice. Poster presented at the 89th Annual Meeting of the Linguistic Society of America.

Khan, S., Becker, K., and Zimman, L. (2015). The acoustics of perceived creaky voice in American English. Paper presented at the 170th Meeting of the Acoustical Society of America.

Khan, S. u. D. (2012). The phonetics of contrastive phonation in Gujarati. *Journal of Phonetics*, 40:780–795.

Kirby, J. P. (2011). Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association*, 41(3):381–392.

Kirk, P., Ladefoged, J., and Ladefoged, P. (1993). Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In Mattina, A. and Montler, T., editors, *American Indian linguistics and ethnography in honor of Laurence C. Thompson*. University of Montana Press.

Kirk, P., Ladefoged, P., and Ladefoged, J. (1984). Using a spectrograph for measures of phonation types in a natural language. *UCLA Working Papers in Phonetics*, 59.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59(5):1208–1221.

Kohler, K. J. (1994). Glottal stops and glottalization in German. *Phonetica*, 51(1-3):38–51.

Krauss, M. (2005). Athabaskan tone. In Hargus, S. and Rice, K., editors, *Athabaskan Prosody*, pages 55–136. John Benjamins Publishing Company.

Kreiman, J. (1982). Perceptual sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10:163–175.

Kreiman, J. and Gerratt, B. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America*, 117(4):2201–2211.

Kuang, J. (2013). *Phonation in Tonal Contrasts*. PhD thesis, University of California, Los Angeles.

Kuang, J. and Keating, P. (2013). Glottal articulations in tense vs lax phonation contrasts. *Journal of the Acoustical Society of America*, 134(4069).

Ladefoged, P. (1964). Some possibilities in speech synthesis. *Language & Speech*, 7(4):205–214.

Ladefoged, P. (1983). The linguistic use of different phonation types. In Bless, D. and Abbs, J., editors, *Vocal fold physiology: Contemporary research and clinical issues*. College Hill Press.

Ladefoged, P. and Antoñanzas-Barroso, N. (1985). Computer measures of breathy voice quality. *UCLA Working Papers in Phonetics*, 61:79–86.

Ladefoged, P. and Johnson, K. (2015). *A Course in Phonetics*. Cengage Learning.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the world's languages*. Blackwell.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press, 1st edition.

Le Grézause, E. (2017). *Um and Uh, and the Expression of Stance in Conversational Speech*. PhD thesis, University of Washington.

Lee, W.-S. and Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1):109–112.

Lehiste, I. (1965). Juncture. In *Proceedings of the 5th International Congress of Phonetic Sciences*.

Lehiste, I. (1970). *Suprasegmentals*. MIT Press, 1st edition.

Lew, S. and Gruber, J. (2016). An acoustic analysis of tone and register in Louma Oeshi. In *Proceedings of the Linguistic Society of America*, volume 1, pages 1–14.

Lewis, M. P., Simons, G. F., and Fennig, C. D., editors (2016). *Ethnologue: Languages of the World*. SIL International, 19th edition.

Lieberman, P. (1961). Perturbations in vocal pitch. *Journal of the Acoustical Society of America*, 33(4):597–603.

Lieberman, P. (1963). Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *Journal of the Acoustical Society of America*, 35.

Maddieson, I. and Ladefoged, P. (1985). "tense" and "lax" in four minority languages of China. *UCLA Working Papers in Phonetics*, 60.

Manuel, S. Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America*, 88(3).

Mendoza-Denton, N. (2011). The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropoloy*, 2(261–280).

Miller, A. L. (2007). Gutteral vowels and gutteral co-articulation in Ju|'hoansi. *Journal of Phonetics*, 35:56–84.

Mixdorff, H., Hu, Y., and Burnham, D. (2005). Visual cues in Mandarin tone perception. In *Interspeech 2005*.

O'Brien, J. (2012). *An Experimental Approach to Debuccalization and Supplementary Gestures*. PhD thesis, University of California, Santa Cruz.

Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1):139–152.

Panfili, L. (2015). Linking vowel height and creaky voice. Generals Paper Defended at the University of Washington.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pierrehumbert, J. and Talkin, D. (1992). Lenition of /h/ and glottal stop. In Doherty, G. and Ladd, D., editors, *Papers in laboratory phonology II: gesture segment prosody*, pages 91–117. Cambridge University Press.

Podesva, R. J. (2013). Gender and the social meaning of non-modal phonation types. *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*.

Podesva, R. J., Callier, P., and Szakay, A. (2015). Gender differences in the acoustic realization of creaky voice: Evidence from conversational data collected in Inland California. Paper presented at the 89th Annual Meeting of the Linguistic Society of America.

Powell, J. V. and Woodruff Sr., F. (1976). *Quileute Dictionary*. Northwest Anthropological Research Notes.

Raphael, L. J., Borden, G. J., and Harris, K. S. (2007). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Lippincott Williams & Wilkins, 5 edition.

Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29:407–429.

Risberg, A. (1961). Statistical studies of fundamental frequency range and rate of change. *Department for Speech, Music and Hearing Quarterly Progress Report*, 2(4):007–008.

Roach, P. J. (1973). Glottalization of English /p/, /t/, /k/, and /tʃ/ - a re-examination. *Journal of the International Phonetic Association*, 3(1):10–28.

Samely, U. (1991). *Kedang (Eastern Indonesia): Some aspects of its Grammar*. Forum phoneticum. H. Buske.

Shattuck-Hufnagel, S. and Turk, A. (1998). The domain of phrase-final lengthening in English. In *The Sound of the Future: A Global View of Acoustics in the 21st Century, Proceedings of the 16th International Congress on Acoustics and 135th Meeting Acoustical Society of America*.

Shue, Y.-L. (2010). *The voice source in speech production: Data, analysis, and models*. PhD thesis, The University of California at Los Angeles.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2011). VoiceSauce: A program for voice analysis. In Lee, W.-S. and Zee, E., editors, *Proceedings of the 17th International Congress of Phonetics Sciences*, volume 3, pages 1846–1849.

Silverman, D. (1997). Laryngeal complexity in Otomanguean vowels. *Phonology*, 14(2):235–261.

Silverman, D., Blankenship, B., Kirk, P., and Ladefoged, P. (1995). Phonetic structures in Jalapa Mazatec. *Anthropological Linguistics*, 37(1):70–88.

Sjölander, K. (2004). Snack sound toolkit. http://www.speech.kth.se/snack/.

Skilton, A. (2016). Contrastive voice quality in Cushillococha Ticuna. Master's thesis, The University of California, Berkeley.

Stevens, K. N. (2000). *Acoustic Phonetics*. The MIT Press.

Stubblefield, M. and Miller Stubblefield, C. (1991). *Diccionario Zapoteco de Mitla*. Summer Institute of Linguistics.

Styler, W. (2015). *On the Acoustical and Perceptual Features of Vowel Nasality*. PhD thesis, The University of Colorado.

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 333–336.

Szende, T. (1999). Hungarian. In *Handbook of the International Phonetic Alphabet: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Talkin, D. (1995). A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn, W. B. and Paliwal, K. K., editors, *Speech Coding and Synthesis*. Elsevier Science Inc.

Thongkum, T. L. (1987). Phonation types in Mon-Khmer languages. *UCLA Working Papers in Phonetics*, 67.

Traill, A. (1987). Depressing facts about Zulu. *African Studies*, 46(2):255–274.

Wayland, R. and Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics*, 31:181–201.

Weide, R. (2005). The Carnegie Mellon pronouncing dictionary [cmudict v. 0.6].

Wright, R. and Ladefoged, P. (1994). A phonetic study of Tsou. *UCLA Working Papers in Phonetics*, 87.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5).

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, 85(3):315–337.

Zeitoun, E. (2007). *A Grammar of Mantauran (Rukai)*. Language and Linguistics Monograph A4-2. Institute of Linguistics, Academia Sinica.

Zemlin, W. R. (1998). *Speech and Hearing Science: Anatomy and Physiology*. Allyn and Bacon, 4 edition.

Appendix A

VARIANCE OF PITCH TRACKS

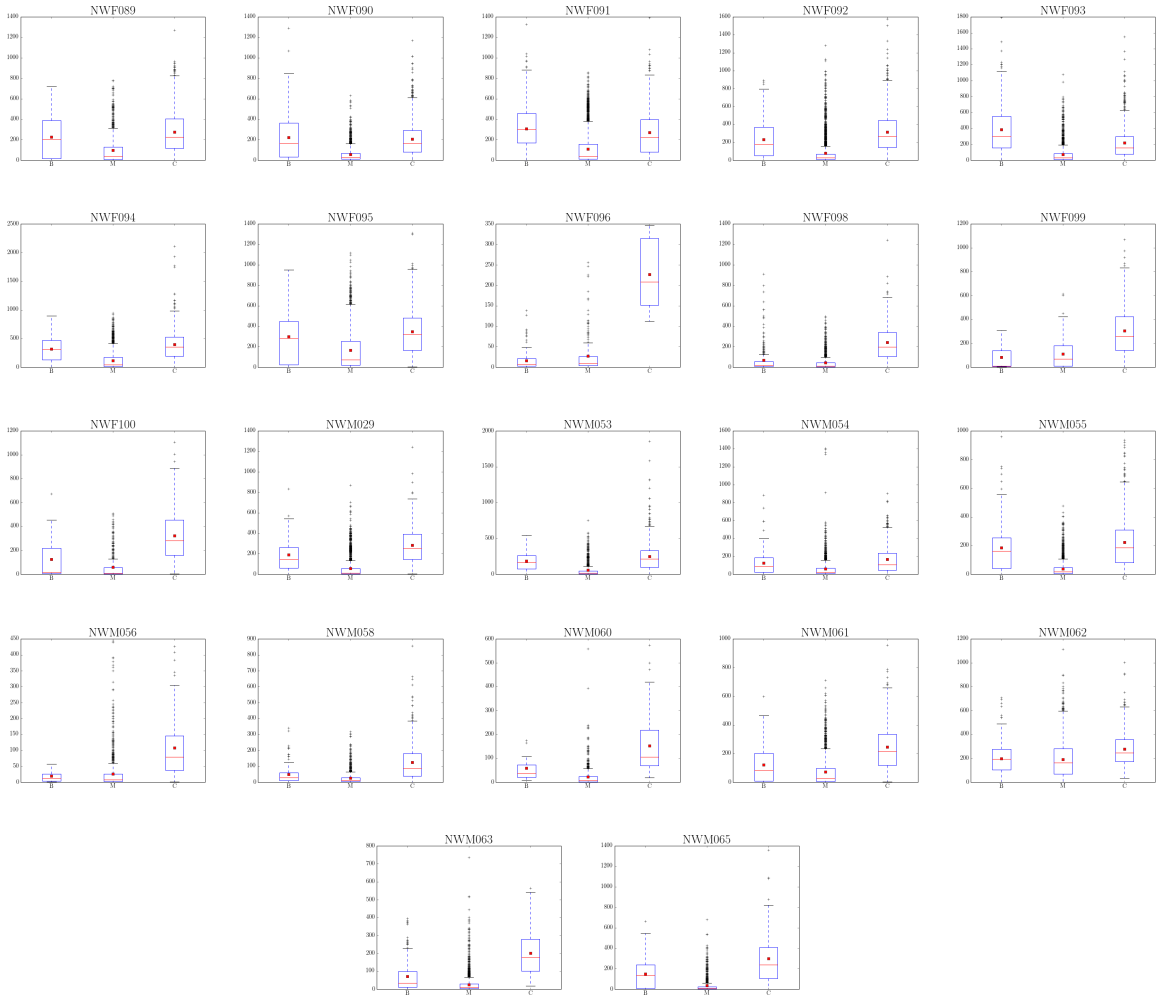


Figure A.1: Variance of Pitch Tracks for all English speakers

Appendix B

PHONETIC STOP WORDS

ABOUT	ELSE	IF	SO	WANNA
ALL	EM	IM	SOME	WANT
AM	FEW	IN	STILL	WANTS
AN	FOR	INTO	THA	WAS
AND	FROM	IS	THAT	WE
ANY	GET	ISN'T	THATS	WE'D
ARE	GETS	IT	THAT'S	WE'LL
AS	GOING	IT'D	THE	WELL
AT	GONNA	IT'S	THEIR	WE'RE
BE	GOT	JUST	THEIRS	WE'VE
BEEN	GOTTA	K	THEM	WENT
BEFORE	GOTTEN	KAY	THEN	WERE
BEING	HAD	LET	THERE	WHAT
BUT	HAS	LET'S	THERE'S	WHEN
BY	HAVE	LIKE	THESE	WHERE
CAN	HAVEN'T	MAY	THEY	WHERE'S
CAN'T	HAVIN	ME	THEY'D	WHICH
CEPT	HE	MY	THEY'LL	WHILE
CAUSE	HER	ND	THEY'RE	WHO
COULD	HERS	OF	THEY'VE	WITH
CUZ	HIM	ON	THIS	WOULD
DID	HIS	OR	THOSE	YOU
DIDN'T	HOW	OUR	TIL	YOUR
DO	I	OURS	TILL	YOURS
DOES	I'D	OUT	TO	YOU'D
DOING	I'LL	OWN	UH	YOU'RE
DON'T	I'M	SHE	UM	
DUNNO	I'VE	SHOULD	US	

Appendix C

200 MOST FREQUENT WORDS IN ATAROS

YEAH (356)
OKAY (205)
IMPORTANT (137)
THINK (124)
KNOW (106)
CUT (64)
ONE (61)
FOOTBALL (59)
NOT (57)
MAYBE (54)
EDUCATION (52)
MEAN (50)
PUBLIC (46)
PROBABLY (46)
RID (45)
OH (45)
KEEP (43)
REALLY (41)
RIGHT (40)
ACCESS (40)
INVASIVE (39)
NEED (37)
PEOPLE (35)
LICENSES (34)
GO (34)
CONTROL (34)
REPRODUCTIVE (32)
WEED (31)
UPKEEP (30)
SEE (30)
STOPS (29)
MORE (29)
HOSPITAL (29)
NO (28)
THING (27)
SUBWAY (27)
FEEL (27)
BASKETBALL (27)
STATION (25)
NEWS (25)
COMMUNITY (25)
ALRIGHT (25)
THINGS (24)
TOO (21)
TATTOO (21)
STADIUM (21)
MUCH (21)
TAXI (20)
TAKE (20)
SURE (20)
SERVICES (20)
SAY (20)
NEIGHBORHOOD (20)
MASSAGE (20)
BUS (20)
THREE (19)
MONEY (19)
GUESS (19)
CERTIFICATES (19)
HUNTING (18)
DOG (18)
DATABASE (18)
CLASSES (18)
SEX (17)
PROGRAMS (17)
MANY (17)
ADS (17)
WHATEVER (16)
UP (16)
SOCCER (16)
POTHOLE (16)
FISHING (16)
COOKING (16)
ACTING (16)
VETERINARY (15)
SOMEONE (15)
RECREATION (15)
PROGRAM (15)
KINDA (15)
JUICE (15)
INFRASTRUCTURE (15)
HEALTH (15)
DEFINITELY (15)
BECAUSE (15)
TWO (14)
TRUE (14)
SOMETHING (14)
NOW (14)
DONE (14)
D (14)[1]
YES (13)
SUGAR (13)
SPECIES (13)
FREE (13)
FOUR (13)
CATCHER (13)
ALSO (13)
AGREE (13)
ADDITIONS (13)
WITHOUT (12)
WATCH (12)
THINKING (12)
SYSTEM (12)
SAME (12)
S (12)
PRETTY (12)
LIMIT (12)
ALREADY (12)
ACTUALLY (12)
TOXIC (11)
THROUGH (11)
T (11)
SUPPORT (11)
STUFF (11)
STRAY (11)
PROBLEM (11)
PEST (11)
MACHINES (11)
LEAGUE (11)
LAW (11)

[1]Letters in this list were pronounced in acronyms. One of the topics up for discussion in the ATAROS budget task was STD education, hence many occurrences of the letters S, T, and D.

KIND (11)
FOOD (11)
EVEN (11)
EACH (11)
DISPOSAL (11)
CUTTING (11)
CLUB (11)
ANYWAY (11)
WHY (10)
VOLUNTEER (10)
SPEED (10)
OTHER (10)
ONLY (10)
OFFENDER (10)
NEEDLE (10)
LOOK (10)
LIST (10)
JUNIOR (10)
GETTING (10)
ED (10)
DRAINAGE (10)
DISABILITY (10)
COUNTY (10)
COACHES (10)
ARTIST (10)
ADDITIONAL (10)
WOULDN'T (9)
VERY (9)
TEACHING (9)
TEACHERS (9)
TAGS (9)
SOUP (9)
SAFETY (9)
MEANS (9)
MATH (9)
MARIJUANA (9)
MAKE (9)
LICENSING (9)
LIBRARY (9)
HUH (9)
HIGH (9)
GOVERNMENT (9)
BOOKS (9)
WAY (8)
VACCINATIONS (8)
TOWING (8)
SPECIFIC (8)
SPAYING (8)
REMOVAL (8)
PAY (8)
LICENSE (8)
KIDS (8)
JOB (8)
HOUSING (8)
HERE (8)
EXCHANGE (8)
ENOUGH (8)
EITHER (8)
DOWN (8)
CITY (8)
BOOKKEEPING (8)
BOATING (8)
BANK (8)
ARTISTS (8)
ANYTHING (8)
ALTHOUGH (8)
TUTORS (7)
THOUGH (7)
STATE (7)
START (7)
SCHOOL (7)
SAID (7)
REVENUE (7)
PICK (7)
OVER (7)
NOTETAKERS (7)
MUSIC (7)
MAINTENANCE (7)
INSPECTIONS (7)
COOKS (7)

Appendix D

GUJARATI WORD LIST

Table D.1: Gujarati Word List

Target Pronunciation	Target Vowel	Gloss
kan	a	*ear*
kạ̈n	ạ̈	*Krishna*
paɖʊ̃	a	*to draw, to cause to fall*
pạ̈ɖ	ạ̈	*mountain*
baɾ	a	*twelve*
bạ̈ɾ	ạ̈	*outside*
baɳ	a	*arrow*
bạ̈nũ	ạ̈	*excuse*
malik	a	*boss, god*
mạ̈ɾaɟʒ	ạ̈	*priest, emperor*
ɾab	a	*gruel, baby food*
ɾạ̈bəɾ	ạ̈	*guide*
cʃə̱tʊɾai	ai	*cleverness*
sərʋəgɾa̤i̤	a̤i̤	*comprehensive*
cʃalaki	a	*cleverness, trick*
səlạ̈kaɾ	ạ̈	*adviser*
ʋan	a	*complexion*
ʋạ̈n	ạ̈	*vehicle*
ləʋaɾo	a	*prattle, gossip*
ʊjəʋạ̈ɾũ	ạ̈	*feasible*
səməgɾə	ə	*whole, entire*
sə̱tjagɾə̤	ə̤	*insistence on truth*
səb^{ɦ}asə̱ɖ	ə	*member*
sə̤b^{ɦ}agi	ə̤	*partner*
ɖ̱ərə̱ɖ	ə	*disease, pain*
sərə̤̈ɖ̱	ə̤	*frontier, border*
kevũ	e	*of what kind*
kɛ̤̈ʋũ	ɛ̤	*to say*
ɖ̱eʃi	e	*native, villager*

Target Pronunciation	Target Vowel	Gloss
d̪ɛ̤ʃət̪	ɛ̤	*apprehension*
pelũ	e	*that*
pɛ̤lũ	ɛ̤	*first*
benəmun	e	*unparalleled*
bɛ̤n	ɛ̤	*sister*
beɾoɟʒgaɾ	e	*unemployed*
bɛ̤ɾaʃ	ɛ̤	*deafness*
mɛ̃	ɛ	*I*
mɛ̤man	ɛ̤	*guest*
mɛl	ɛ	*dirt, filth*
mɛ̤l	ɛ̤	*palace*
ʋeɾ	e	*revenge, enmity*
ʋɛ̤ɾ	ɛ̤	*sawdust*
ʋeʋəi	e	*father-in-law of one's child*
ʋɛ̤ʋaɾ	ɛ̤	*daily interaction*
ʃeɾɖi	e	*sugarcane*
ʃɛ̤ɾ	ɛ̤	*city*
cʃint̪a	i	*worry*
cʃi̤nə	i̤	*mark, trait*
ʋiʋɾəɳ	i	*criticism, commentary*
ʋi̤ʋəɭ	i̤	*puzzled and upset*
kɔgɭo	ɔ	*mouthful of liquid, gargle*
kɔ̤ʋũ	ɔ̤	*to rot*
ɖoɭo	o	*eyeball*
ɖoɭʋũ	o	*to swing*
ɖɔ̤ɭũ	ɔ̤	*dirty, polluted*
ɖɔ̤ɭʋũ	ɔ̤	*to stir in*
d̪oɾo	o	*thread, necklace*
d̪ɔ̤ʋũ	ɔ̤	*to milk*
mokɭũ	o	*spacious, open*
mɔ̃	ɔ	*mouth, face*
mɔko	ɔ	*favorable opportunity*
mɔ̤k	ɔ̤	*charming*
locʃən	o	*eye*
lo̤cʃumbək	o̤	*magnet*
soɖo	o	*trade*
sɔ	ɔ	*hundred*
sɔ̤d̪əɾ	ɔ̤	*born of the same mother*
nəiʋed̪	əi	*food offered to god*
nə̤i̤ʋət̪	ə̤i̤	*slightly*

Target Pronunciation	Target Vowel	Gloss
bəud̪ʱːik	əu	*intelligent*
bə̤ṳ	ə̤ṳ	*many, much*

Appendix E

HMONG WORD LISTS

Table E.1: White Hmong Word List

Target Pronunciation	Target Vowel(s)	Gloss
ca33	a	*log*
ca45	a	*haul or tow*
ca53	a̤	*root, origin, what*
ca53	a	*ridge*
ca22	a	*why, how*
ca24	a	*argue, disagree, praise (can also be a tree that has fallen)*
ci53	i̤	*light up*
ci21	ḭ	*mark, sign, symbol*
cu53	ṳ	*collect*
ka45	a	*insect*
ka53	a	*dawn, bright*
ka21	a̰	*allow, permit*
ka22	ã	*maggots*
ka24	a	*stem, stalk*
ki21	ḭ	*expensive; boy's name*
ki22	i	*to infect*
ko22	o	*stem*
ku53	ṳ	*very poor*
ku21	ṵ	*boy's name*
pa53	a	*flower*
pa21	a̰	*bridge, blanket*
pa21	a̰	*blanket*
pa22	a	*pond, lake, staff*
pa22	a	*stick*
pi21	ḭ	*vagina*
po33	o	*pancreas*
po45	o	*lump*
po53	o̤	*grandmother*

Target Pronunciation	Target Vowel(s)	Gloss
po53	o	*female*
po21	o̰	*see*
po22	o	*thorn/cover*
po24	o	*throw*
pu21	ṵ	*to see*
pu22	u	*escape with all or part of the trap/arrow still attached*
qa53	a̤	*axle, swivel*
qa22	a	*disgust, sicken*
ʈau33/ʈʌu33	au/ʌu	*six*
ʈau53/ʈʌu53	a̤ṳ/ʌ̤ṳ	*to suffer (something)*
ʈau53/ʈʌu53	au/ʌu	*hammer*
ʈau21/ʈʌu21	a̰ṵ/ʌ̰ṵ	*kidney*
ʈau22/ʈʌu22	au/ʌu	*to be full (of food)*
ʈau24/ʈʌu24	au/ʌu	*to light a fire*
ʈʰẽ45/ʈʰẽŋ45	ẽ	*to scratch the ground*
ta53	a̤	*done*
ta21	a̰	*represent, sharpen (knife)*
ta22	a	*finished/end, done*
tau33/tʌu33	au/ʌu	*to be able*
tau45/tʌu45	au/ʌu	*pumpkin*
tau53/tʌu53	a̤ṳ/ʌ̤ṳ	*to follow*
tau53/tʌu53	au/ʌu	*species of grass*
tau21/tʌu21	a̰ṵ/ʌ̰ṵ	*bean*
tau24/tʌu24	au/ʌu	*to dam up (water)*
taɨ53/taɯ53	a̤ɨ̤/a̤ɯ̤	*explode*
taɨ21/taɯ21	a̰ɨ̰/a̰ɯ̰	*to go out*
taɨ22/taɯ22	aɨ/aɯ	*to emit light*
ti53	i̤	*turn, squirm, steer*
ti21	ḭ	*because of it*
ti22	i	*wing, level, layer, to sew*
to33	o	*puncture*
to45	o	*very deep*
to53	ö	*chair*
to53	o	*ahead*
to21	o̰	*to bite*
to22	o	*to wait*
to24	o	*to mix*
ʈʂuɔ53 ʈʂuɔ53 / ʈsuɔ53 ʈʂuɔ53	uɔ	*hurry*

Target Pronunciation	Target Vowel(s)	Gloss
tu53	u̬	*whose, properly*
tu21	ṵ	*to stack*
tu22	u	*open-minded, stable, peaceful*
vu22	u	*post-verbal intensifier*
z̬a53/ʒa53	a̬	*a sentence*
z̬a21/ʒa21	a̰	*avoid, to step out of the way*
z̬a22/ʒa22	a	*color, paint*

Table E.2: Green Hmong Word List

Target Pronunciation	Target Vowel(s)	Gloss
tau24 dle53 / tʌu24 dle53	au e	*to dam up (water)*
ʈau33/ʈʌu33	au/ʌu	*six*
ʈau45/ʈʌu45	au/ʌu	*to scratch the ground*
pã53/pãŋ53	ã̬	*lake, pond*
kã45/kãŋ45	ã	*insect*
cu21	ṵ	–
ku21	ṵ	*a boy's name*
tu22	u	*to bind, fold*
ki22	i	*corner, to contaminate*
c^{h}o53	o̬	*blanket*
vu22	u	*alternative name for Vaaj*
tu53	ü	*flat place; long objects; to float; classifier*
pi53	i̬	*tomorrow*
qa53	a̬	*have space left; to shine (moon); belching*
cu53	u̬	*treadmill; to collect in a vessel*
ʈau21/ʈʌu21	a̰ṵ/ʌ̰ṵ	*kidney*
ʈau53/ʈʌu53	a̬u̬/ʌ̬u̬	*hammer*
pu22	u	*to go together; to close up an opening; to bake*
ti53	i̬	*to turn; to move*
cã53/cãŋ53	ã̬	*ridge*
tau45/tʌu45	au/ʌu	*pumpkin*
pu21	ṵ	*see*
to53	o̬	*to wait*
pi22	i	*how much*
qa22	a	*dirty*
cu22 mbuɔ33	u uɔ	*a piece of good looking meat along spine of animal*
kã21/kãŋ21	ã	*allow, permit*
to53	o̬	*ahead*
pã53/pãŋ53	ã̬	*flower*
z̬ã22/ʒɲ̃22	ã	*to dye; to face off*

Target Pronunciation	Target Vowel(s)	Gloss
ku22	u	*a stem from leaf*
ta21	a̰	*to invite to a meal; generation; first (son)*
ku53	ṵ	*very poor*
cã33/cãŋ33	ã	*log*
ʈ̟ʂã53/ʈ̟ʂãŋ 53/tʃã53/tʃãŋ53	ã̰	*chair*
qa22	a	*dirty, ugly, bad*
kã24/kãŋ24	ã	*stem, stalk*
ʈau24/ʈʌu24	au/ʌu	*to light a fire*
cã53/cãŋ53	ã̰	*root, origin, what*
ci21	ḭ	*to remember, remind, remark; mark, sign; season*
cã24/cãŋ24	ã	*argue, disagree, praise, log*
co22	o	*lump*
pu53	ṵ	*to fall (of dew)*
taɨ53/taɯ53	a̰ɨ̰/a̰ɯ̰	*explode*
ʈau22/ʈʌu22	au/ʌu	*to be full (of food)*
ʈau53/ʈʌu53	a̰ṵ/ʌ̰ṵ	*to follow*
ki21	ḭ	*expensive*
ci22	i	–
ci53	ḭ	*to burn, to set fire to*
tau33/tʌu33	au/ʌu	*to be able*
pã22/pãŋ22	ã	*stick*
taɨ21/taɯ21	a̰ɨ̰/a̰ɯ̰	*to go out*
tau21/tʌu21	a̰ṵ21/ʌ̰ṵ	*bean*
ti22	i	*wing; to name; to sew*
cã53/cãŋ53	ã̰	*venereal disease*
vu53	ṵ	*spear*
po33	o	*pancreas*
kã22/kãŋ22	ã	*maggots*
ʐ̟a21/ʒa21	a̰	*to avoid forgive; dress up; color; in order*
ta22	a	*to say, subordinator 'that'; birthmark*
ti21	ḭ	*over there; because of, depend on*
pu53	u	*grandmother*
pã21/pãŋ21	ã̰	*to prepare*
ka53 ka53	a̰ a̰	*hurry*
po24	o	*throw*
to33	o	*puncture*
cã21/cãŋ21	ã	*to stick, glue; to argue, debate*
pã22/pãŋ22	ã	*stick, a swallow (of liquid)*
pu53	ṵ	*female*
cu22	u	*(shake)*
ʈau53/ʈʌu53	a̰ṵ/ʌ̰ṵ	*to suffer (something)*
cã45/cãŋ45	ã	*haul or tow*
tau53/tʌu53	a̰ṵ/ʌ̰ṵ	*species of grass*
ʐ̟ã53/ʒãŋ53	ã̰/ã̰	*a given name*
to24	o	*to mix*
tu21	ṵ	*to bite, to fight; to stack*

Target Pronunciation	Target Vowel(s)	Gloss
tu21	ṵ	*to bite*
taɨ22/taɯ22	aɨ/aɯ	*to emit light*
pu53 cu22	ṵ u	*earring*
tu45	u	*very deep*
ta53	a̰	*true, real, certain; to hold*
vu21	u	*to steam, to warm*
ki53	ḭ	*morning*
pu21	ṵ	*to see*
kã53/kãŋ53	ã̰	*dawn, bright*
pi21	ḭ	*vagina*
suɔ24 pau53/suɔ24 pʌu53	uɔ a̰ṵ/uɔ ʌ̰ṵ	*thorn, cover*

Appendix F

MAZATEC WORD LISTS

Table F.1: 1993 Recordings

Target Pronunciation	Target Vowel(s)	Gloss	Index
ⁿdæ̤11	æ̤	*horse*	10
næ̰33	æ̰	*becomes*	48
βa̰33	a̰	*becomes*	55
βæ̰33	æ̰	*hits*	57
ja̰33	a̰	*brings, transports*	60
jæ̰33	æ̰	*excrement*	62
ⁿka̰33	a̰	*high*	66
ʍa̰33	a̰	*passes*	74
ha̰33	a̰	*he passed*	76
ⁿdæ̰33	æ̰	*companion, man*	83
βa̰11 *or* ʍa̰11	a̰	*thus*	84
tʃa̰55	a̰	*loads*	87
na33m̥i33tʃa̰33	a, i, a̰	*nobody*	88
sæ̰33	æ̰	*to exist*	89
t͡ʃæ̰11	a̰e	*his, hers, theirs*	90
jæ̤33	a̤e	*boil (noun)*	123
ⁿda̤11	a̤	*animal horn*	126
tʰæ̰33	æ̰	*sorcery*	138
superndæ̰11	æ̰	*buttock*	142
ɸi33kʰa̰55	i, a̰	*is going to bring*	149
tʃa̰11	a̰	*spoon*	150
tʰæ33	æ	*itch*	151
ta33tʰa33	a, a	*sticky*	152
tsʰæ33	æ	*spotted*	153
t͡sʰa55	a	*gives*	154
tʃʰa11tæ11	a, æ	*wasp*	156
ti55fi33kʰæ33	i, i, æ	*is finished*	157
tʃu11kʰa55	u, a	*skunk*	158
kʍhæ11	æ	*file*	159

Target Pronunciation	Target Vowel(s)	Gloss	Index
kʍha11	a	*will happen*	160
tæ33	æ	*ten*	161
ka33ma33ta33	a, a, a	*it will become thick*	162
tsa33	a	*moral*	164
tʃa55	a	*old*	166
ki33kæ33	i, æ	*I saw him*	167
ka33	a	*bald*	168
ntæ55	æ	*shoes*	171
nta33	a	*soft*	172
ntsæ̃33	æ̃	*brother*	173
$^{n}\widehat{ts}^{h}$a11	a	*hair*	176
$^{n}\widehat{ts}^{h}$ae11	ae	*kind of gourd*	181
^{n}dæ33	æ	*deceased*	188
^{n}da33	a	*good*	189
na11	a	*woman*	194
jæ11	æ	*snake*	219
ja55	a	*tree, wood*	220
st^{h}æ55	æ	*garbage*	226
^{n}dæ̈11	æ̈	*horse*	239
ją̈33	ą̈	*clothes*	248
ŋgʍą̈33	ą̈	*I will put on*	250
tsæ33	æ	*full*	278

Table F.2: 1984 Recordings

Target Pronunciation	Target Vowel(s)	Gloss	Index
ʔa33tʃa11ndæ̈11	a, a, æ̈	*my horse*	22
ʔa33tʃa11ndæ̃11	a, a, æ̃	*my buttocks*	23
ʔa33tʃa11nt^{h}æ11	a, a, æ	*my seeds*	24
hæ̃33	æ̃	*green ear of corn*	51
ʔiɨɐʔja11	i, a	*big leafcutter ants*	32
ʔja11	a	*leafcutter ants*	31
jæ11	æ	*snake*	30
jæ̃11	æ̃	*manure*	29
ʔjũ55nda33	ũ , a	*very good*	10
kæ̃33	æ̃	*dead*	49
k^{w}ą̃11	ą̃	*it will happen*	75

Target Pronunciation	Target Vowel(s)	Gloss	Index
mæ̃̈33	æ̃̈	*he wants*	48
^{n}dæ̈11	æ̈	*horse*	17
^{n}dæ̃11	æ̃	*buttocks*	18
ndʒä̤55ʃu55	ä̤, u	*chocolate drink*	62
ng^{w}ä̤11	ä̤	*he puts on*	74
(n)t^{h}æ11	æ	*seed*	19
t^{h}æ33	æ	*itch*	54
t$^{hæ̃}$33	æ̃	*sorcery*	20
ti55mã33^{n}dzæ̈33	i, ã, æ̈	*visible*	47
ti55βä̤55ʔa11	i, ä̤, a	*weave*	108
ti55βä̤33	i, ä̤	*he hits*	68
tʃa55	a	*old*	65
tʃã̤55	ã̤	*load*	8
tʃæ̃33	æ̃	*lazy*	50
tʃu11jæ̈11	u, æ̈	*turtle*	28

Appendix G

ZAPOTEC WORD LIST

Table G.1: Zapotec Word List

Target Pronunciation	Target Vowel(s)	Gloss
pap	a	*potato*
bag	a	*cow*
bed	e	*Peter*
bub ∼ bob ∼ bab	u ∼ o ∼ a	*drool*
bot	o	*vote*
dad	a	*dice* or *dad*
lad	a	*side*
pag	a	*pay*
tap	a	*lid*
diʒ	i	*language*
ʃop	o	*six*
ʃuk	u	*upper arm*
lat	a	*(tin) can*
t̤ap	a̤	*four*
b(d)a̤g	a̤	*leaf*
b(d)i̤g	i̤	*ant*
be̤l	e̤	*fish*
be̤ts	e̤	*brother (of a man)*
da̤ ∼ de̤	a̤ ∼ e̤	*powder*
de̤ts	e̤	*back*
la̤	a̤	*name*
la̤ʒ	a̤	*town, village*
la̤d	a̤	*clothings*
ʃ'*ig	i̤	*jicara*
ɾṳ ∼ ɾo̤	ṳ ∼ o̤	*cough*
lo̤	o̤	*face*
ʃṳn	ṳ	*eight*
be̤ ∼ bi̤	e̤ ∼ i̤	*wind, air*
kṳd	ṳ	*thigh*

Target Pronunciation	Target Vowel(s)	Gloss
gə̰̈t	ə̰̈	*tortilla*
dą̰(ʔą̰)	ą̰, ą̰	*a type of straw mat*
lą̰(ʔą̰)	ą̰, ą̰	*huaje bean*
lą̰ts	ą̰	*field*
dʒą̰̂p	â̰	*girl*
bą̰(ʔą̰)	ą̰, ą̰	*tomb*
bą̰d	ą̰	*scabies*
gą̰ʔą̰	ą̰, ą̰	*nine*
dą̰gts	ą̰	*empty*
bə̰̃ʔ	ə̰̃	*mushroom*
bə̰̃l:	ə̰̃	*meat*
ʒõ̰b	õ̰	*corn*
gṵ̃n	ṵ̃	*bull*
guʒą̰̃d ∼ guʃą̰̃d	u, ą̰̃	*grasshopper*
bą̰̃	ą̰̃	*eyeball*
bə̰̃ld	ə̰̃	*snake*
guną̰̃	u, ą̰̃	*woman*
są̰̃	ą̰̃	*cloud*
bõ̰	õ̰	*charcoal*

Appendix H

ENGLISH FEATURE METRICS

Table H.1: English Feature Correlations

Feature	B vs. C	B vs. M	C vs. M
H1* – H2*_Mean	0.255	0.065	-0.185
H1* – H2*_1	0.187	0.033	-0.153
H1* – H2*_2	0.242	0.07	-0.164
H1* – H2*_3	0.261	0.073	-0.188
H2* – H4*_Mean	-0.066	-0.022	0.024
H2* – H4*_1	-0.087	-0.02	0.046
H2* – H4*_2	-0.053	-0.022	0.013
H2* – H4*_3	-0.036	-0.017	0.006
H1* – A1*_Mean	0.067	0.049	0.007
H1* – A1*_1	0.04	0.031	0.009
H1* – A1*_2	0.071	0.054	0.011
H1* – A1*_3	0.07	0.05	0.001
H1* – A2*_Mean	0.132	0.056	-0.045
H1* – A2*_1	0.072	0.026	-0.03
H1* – A2*_2	0.142	0.069	-0.035
H1* – A2*_3	0.146	0.06	-0.06
H1* – A3*_Mean	0.016	-0.004	-0.022
H1* – A3*_1	-0.033	-0.016	0.007
H1* – A3*_2	0.031	0.008	-0.018
H1* – A3*_3	0.044	-0.003	-0.05
H4* – 2k*_Mean	-0.08	0.001	0.077
H4* – 2k*_1	-0.057	0.009	0.069
H4* – 2k*_2	-0.068	0.006	0.075
H4* – 2k*_3	-0.088	-0.012	0.063
2k* – 5k_Mean	-0.033	-0.064	-0.07
2k* – 5k_1	-0.043	-0.066	-0.063
2k* – 5k_2	-0.024	-0.058	-0.069
2k* – 5k_3	-0.023	-0.051	-0.058
CPP_Mean	-0.266	-0.363	-0.384
CPP_1	-0.25	-0.252	-0.231
CPP_2	-0.24	-0.36	-0.385
CPP_3	-0.211	-0.362	-0.41
Dist_Word_(%)	-0.046	-0.011	0.028
Dist_Utt_(%)	-0.089	0.013	0.105
Dist_Word_(ms)	0.019	0.011	-0.001
Dist_Utt_(ms)	-0.047	-0.077	-0.076

Feature	B vs. C	B vs. M	C vs. M
VoPT	-0.138	0.274	0.511
RMS_Energy_Mean	-0.044	-0.147	-0.219
RMS_Energy_1	-0.056	-0.135	-0.188
RMS_Energy_2	-0.035	-0.144	-0.217
RMS_Energy_3	-0.013	-0.135	-0.217
HNR05_Mean	0.41	-0.013	-0.286
HNR05_1	0.357	0.056	-0.171
HNR05_2	0.35	-0.033	-0.271
HNR05_3	0.34	-0.054	-0.303
HNR15_Mean	0.447	0.049	-0.248
HNR15_1	0.396	0.099	-0.156
HNR15_2	0.411	0.026	-0.24
HNR15_3	0.382	0.007	-0.262
HNR25_Mean	0.424	0.074	-0.205
HNR25_1	0.38	0.118	-0.125
HNR25_2	0.394	0.05	-0.203
HNR25_3	0.362	0.032	-0.219
HNR35_Mean	0.397	0.101	-0.154
HNR35_1	0.358	0.137	-0.087
HNR35_2	0.374	0.077	-0.158
HNR35_3	0.339	0.06	-0.167
STRAIGHT_f_0_Mean	0.137	-0.055	-0.226
STRAIGHT_f_0_1	0.134	-0.041	-0.205
STRAIGHT_f_0_2	0.134	-0.056	-0.224
STRAIGHT_f_0_3	0.123	-0.06	-0.219
Snack_f_0_Mean	0.43	-0.15	-0.544
Snack_f_0_1	0.236	-0.111	-0.333
Snack_f_0_2	0.4	-0.11	-0.457
Snack_f_0_3	0.274	-0.102	-0.36
SHR_f_0_Mean	-0.021	-0.043	-0.036
SHR_f_0_1	0.024	-0.03	-0.071
SHR_f_0_2	-0.017	-0.056	-0.052
SHR_f_0_3	-0.041	-0.04	-0.003
F1_Mean	-0.161	-0.031	0.112
F1_1	-0.147	-0.036	0.084
F1_2	-0.167	-0.039	0.11
F1_3	-0.125	-0.01	0.118
Vowel_Duration	-0.116	-0.002	0.109
Voicing_Pre	-0.075	-0.036	0.014
Voicing_Fol	-0.029	-0.097	-0.122
Manner_Pre	-0.041	-0.038	-0.02
Manner_Fol	-0.01	-0.014	-0.012
Presence_Pre	-0.167	-0.12	-0.004
Presence_Fol	-0.022	-0.12	-0.146

Table H.2: English Random Forest Feature Importance, Resampled

Feature	Importance
H1* – H2*_Mean	0.005785545
H1* – H2*_1	0.007227143
H1* – H2*_2	0.006892261
H1* – H2*_3	0.009865814
H2* – H4*_Mean	0.007097403
H2* – H4*_1	0.004528491
H2* – H4*_2	0.00683486
H2* – H4*_3	0.006732433
H1* – A1*_Mean	0.006229393
H1* – A1*_1	0.006387231
H1* – A1*_2	0.00820653
H1* – A1*_3	0.006554134
H1* – A2*_Mean	0.010215961
H1* – A2*_1	0.004836765
H1* – A2*_2	0.008104124
H1* – A2*_3	0.010998286
H1* – A3*_Mean	0.004169069
H1* – A3*_1	0.004499445
H1* – A3*_2	0.004411501
H1* – A3*_3	0.004691808
H4* – 2k*_Mean	0.005138585
H4* – 2k*_1	0.005532094
H4* – 2k*_2	0.005475572
H4* – 2k*_3	0.006609096
2k* – 5k_Mean	0.005566926
2k* – 5k_1	0.005127914
2k* – 5k_2	0.005519396
2k* – 5k_3	0.006109796
CPP_Mean	0.101648833
CPP_1	0.0198241
CPP_2	0.017064511
CPP_3	0.043653864
RMS_Energy_Mean	0.008422795
RMS_Energy_1	0.019306991
RMS_Energy_2	0.02423388
RMS_Energy_3	0.023091008
HNR05_Mean	0.035637848
HNR05_1	0.020915932
HNR05_2	0.013878406
HNR05_3	0.013777355
HNR15_Mean	0.009843313
HNR15_1	0.005873881
HNR15_2	0.008304805
HNR15_3	0.005409024
HNR25_Mean	0.005482441

Feature	Importance
HNR25_1	0.008107572
HNR25_2	0.006253655
HNR25_3	0.004533757
HNR35_Mean	0.01209546
HNR35_1	0.004784497
HNR35_2	0.005056569
HNR35_3	0.005025682
STRAIGHT_f_0_Mean	0.010633557
STRAIGHT_f_0_1	0.007521275
STRAIGHT_f_0_2	0.005541462
STRAIGHT_f_0_3	0.00932834
Snack_f_0_Mean	0.042618402
Snack_f_0_1	0.033503634
Snack_f_0_2	0.107638919
Snack_f_0_3	0.047351055
SHR_f_0_Mean	0.004901247
SHR_f_0_1	0.006034058
SHR_f_0_2	0.006022004
SHR_f_0_3	0.004948322
F1_Mean	0.007278461
F1_1	0.005537767
F1_2	0.011557777
F1_3	0.008262268
Vowel_Duration	0.008044536
Dist_Word_(percent)	0.005540984
Dist_Utt_(%)	0.005244758
Dist_Word_(ms)	0.005749131
Dist_Utt_(ms)	0.005365188
VoPT	0.023702715
Voicing_Pre	0.004197829
Voicing_Fol	0.00931376
Manner_Pre	0.004453642
Manner_Fol	0.003594753
Presence_Pre	0.001585741
Presence_Fol	0.002954664

Appendix I

GUJARATI FEATURE METRICS

Table I.1: Gujarati Feature Correlations

Feature	B vs. M
H1* – H2*_Mean	0.058
H1* – H2*_1	0.025
H1* – H2*_2	0.104
H1* – H2*_3	0.027
H2* – H4*_Mean	0.074
H2* – H4*_1	0.062
H2* – H4*_2	0.078
H2* – H4*_3	0.063
H1* – A1*_Mean	0.163
H1* – A1*_1	0.115
H1* – A1*_2	0.198
H1* – A1*_3	0.12
H1* – A2*_Mean	0.162
H1* – A2*_1	0.112
H1* – A2*_2	0.208
H1* – A2*_3	0.107
H1* – A3*_Mean	0.13
H1* – A3*_1	0.138
H1* – A3*_2	0.157
H1* – A3*_3	0.064
H4* – 2k*_Mean	0.078
H4* – 2k*_1	0.112
H4* – 2k*_2	0.08
H4* – 2k*_3	0.028
2k* – 5k_Mean	-0.094
2k* – 5k_1	-0.088
2k* – 5k_2	-0.116
2k* – 5k_3	-0.057
CPP_Mean	-0.105
CPP_1	0.098
CPP_2	-0.254
CPP_3	-0.072
Local_Jitter_Mean	-0.037
Local_Jitter_1	-0.082
Local_Jitter_2	0.05
Local_Jitter_3	-0.023

Feature	B vs. M
Local_Abs._Jitter_Mean	-0.033
Local_Abs._Jitter_1	-0.064
Local_Abs._Jitter_2	0.039
Local_Abs._Jitter_3	-0.014
RAP_Jitter_Mean	-0.039
RAP_Jitter_1	-0.079
RAP_Jitter_2	0.029
RAP_Jitter_3	-0.02
PPQ5_Jitter_Mean	-0.004
Local_Shimmer_Mean	0.08
Local_Shimmer_1	-0.066
Local_Shimmer_2	0.21
Local_Shimmer_3	-0.05
Local_Shimmer_dB_Mean	0.098
Local_Shimmer_dB_1	-0.051
Local_Shimmer_dB_2	0.217
Local_Shimmer_dB_3	-0.052
APQ3_Shimmer_Mean	0.029
APQ3_Shimmer_1	-0.042
APQ3_Shimmer_2	0.107
APQ3_Shimmer_3	-0.004
APQ5_Shimmer_Mean	0.112
APQ11_Shimmer_Mean	0.182
VoPT	-0.106
RMS_Energy_Mean	-0.019
RMS_Energy_1	0.018
RMS_Energy_2	-0.069
RMS_Energy_3	0.008
HNR05_Mean	-0.028
HNR05_1	0.118
HNR05_2	-0.166
HNR05_3	-0.02
HNR15_Mean	-0.047
HNR15_1	0.088
HNR15_2	-0.151
HNR15_3	-0.049
HNR25_Mean	-0.083
HNR25_1	0.053
HNR25_2	-0.18
HNR25_3	-0.071
HNR35_Mean	-0.103
HNR35_1	0.029
HNR35_2	-0.196
HNR35_3	-0.08
STRAIGHT_f_0_Mean	-0.016
STRAIGHT_f_0_1	-0.065
STRAIGHT_f_0_2	-0.059
STRAIGHT_f_0_3	0.08

Feature	B vs. M
Snack_f_0_Mean	0.062
Snack_f_0_1	0.036
Snack_f_0_2	-0.009
Snack_f_0_3	0.085
Praat_f_0_Mean	-0.005
Praat_f_0_1	-0.032
Praat_f_0_2	-0.044
Praat_f_0_3	0.065
SHR_f_0_Mean	-0.07
SHR_f_0_1	-0.091
SHR_f_0_2	-0.104
SHR_f_0_3	0.036
F1_Mean	0.045
F1_1	0.04
F1_2	0.04
F1_3	0.051
Vowel_Duration	0.333

Table I.2: Gujarati Random Forest Feature Importance, Resampled

Feature	Importance
H1* – H2*_Mean	0.009902581
H1* – H2*_1	0.007151668
H1* – H2*_2	0.007981768
H1* – H2*_3	0.005719122
H2* – H4*_Mean	0.005544291
H2* – H4*_1	0.011370803
H2* – H4*_2	0.005682015
H2* – H4*_3	0.009373164
H1* – A1*_Mean	0.013265164
H1* – A1*_1	0.00753109
H1* – A1*_2	0.015989798
H1* – A1*_3	0.008771955
H1* – A2*_Mean	0.007718348
H1* – A2*_1	0.007298814
H1* – A2*_2	0.009934049
H1* – A2*_3	0.006860007
H1* – A3*_Mean	0.004792625
H1* – A3*_1	0.004759316
H1* – A3*_2	0.010237798
H1* – A3*_3	0.004284176
H4* – 2k*_Mean	0.00806589
H4* – 2k*_1	0.014577556
H4* – 2k*_2	0.008279371
H4* – 2k*_3	0.013076184
2k* – 5k_Mean	0.005039658
2k* – 5k_1	0.009857459
2k* – 5k_2	0.004558627
2k* – 5k_3	0.009102354
CPP_Mean	0.009660661
CPP_1	0.012492754
CPP_2	0.020111977
CPP_3	0.004646969
RMS_Energy_Mean	0.003734903
RMS_Energy_1	0.010752567
RMS_Energy_2	0.010218725
RMS_Energy_3	0.01068075
HNR05_Mean	0.011429116
HNR05_1	0.012042918
HNR05_2	0.015100498
HNR05_3	0.008031202
HNR15_Mean	0.006596959
HNR15_1	0.014457646
HNR15_2	0.00930629
HNR15_3	0.006804511
HNR25_Mean	0.007728319

Feature	Importance
HNR25_1	0.01129103
HNR25_2	0.008605563
HNR25_3	0.009746437
HNR35_Mean	0.012718619
HNR35_1	0.009653174
HNR35_2	0.018165219
HNR35_3	0.013561705
STRAIGHT_f_0_Mean	0.006612716
STRAIGHT_f_0_1	0.006133051
STRAIGHT_f_0_2	0.008026493
STRAIGHT_f_0_3	0.007635947
Snack_f_0_Mean	0.014820134
Snack_f_0_1	0.009458521
Snack_f_0_2	0.006528429
Snack_f_0_3	0.006966575
Praat_f_0_Mean	0.004631962
Praat_f_0_1	0.008721182
Praat_f_0_2	0.007689909
Praat_f_0_3	0.006223075
SHR_f_0_Mean	0.006802013
SHR_f_0_1	0.008755427
SHR_f_0_2	0.005912127
SHR_f_0_3	0.006582929
F1_Mean	0.005821631
F1_1	0.010082044
F1_2	0.009641937
F1_3	0.010380723
Vowel_Duration	0.032483215
Local_Jitter_Mean	0.004399683
Local_Jitter_1	0.007528213
Local_Jitter_2	0.008917735
Local_Jitter_3	0.011494249
Local_Abs._Jitter_Mean	0.007079892
Local_Abs._Jitter_1	0.005153653
Local_Abs._Jitter_2	0.012524777
Local_Abs._Jitter_3	0.014511059
RAP_Jitter_Mean	0.005150072
RAP_Jitter_1	0.008397362
RAP_Jitter_2	0.009531513
RAP_Jitter_3	0.009141375
PPQ5_Jitter_Mean	0.006774673
Local_Shimmer_Mean	0.005730774
Local_Shimmer_1	0.007344899
Local_Shimmer_2	0.025131206
Local_Shimmer_3	0.006904096
Local_Shimmer_dB_Mean	0.009502634
Local_Shimmer_dB_1	0.016286461
Local_Shimmer_dB_2	0.023512276

Feature	Importance
Local_Shimmer_dB_3	0.00649054
APQ3_Shimmer_Mean	0.005577359
APQ3_Shimmer_1	0.02993967
APQ3_Shimmer_2	0.02444057
APQ3_Shimmer_3	0.009096089
APQ5_Shimmer_Mean	0.005649453
APQ11_Shimmer_Mean	0.009488724
VoPT	0.020126793

Appendix J

HMONG FEATURE METRICS

Table J.1: Hmong Feature Correlations

Feature	B vs. C	B vs. M	C vs. M
H1* – H2*_Mean	0.366	0.309	-0.035
H1* – H2*_1	0.338	0.305	0.004
H1* – H2*_2	0.371	0.327	-0.022
H1* – H2*_3	0.308	0.219	-0.078
H2* – H4*_Mean	0.042	0.028	-0.012
H2* – H4*_1	0.019	0.018	0
H2* – H4*_2	0.065	0.036	-0.023
H2* – H4*_3	0.025	0.019	-0.004
H1* – A1*_Mean	0.322	0.264	-0.057
H1* – A1*_1	0.303	0.274	-0.023
H1* – A1*_2	0.337	0.279	-0.056
H1* – A1*_3	0.241	0.165	-0.072
H1* – A2*_Mean	0.274	0.22	-0.071
H1* – A2*_1	0.231	0.186	-0.055
H1* – A2*_2	0.297	0.252	-0.055
H1* – A2*_3	0.228	0.15	-0.083
H1* – A3*_Mean	0.233	0.203	0.006
H1* – A3*_1	0.185	0.173	0.013
H1* – A3*_2	0.244	0.222	0.017
H1* – A3*_3	0.198	0.155	-0.015
H4* – 2k*_Mean	0.024	0.041	0.02
H4* – 2k*_1	0.046	0.036	-0.005
H4* – 2k*_2	0.002	0.039	0.038
H4* – 2k*_3	0.018	0.034	0.019
2k* – 5k_Mean	-0.103	-0.092	0.007
2k* – 5k_1	-0.125	-0.113	0.006
2k* – 5k_2	-0.098	-0.088	0.005
2k* – 5k_3	-0.068	-0.057	0.005
CPP_Mean	0.083	-0.145	-0.227
CPP_1	0.06	-0.121	-0.183
CPP_2	0.054	-0.166	-0.223
CPP_3	0.118	-0.118	-0.228
SHR_Mean	-0.379	-0.334	-0.038
SHR_1	-0.329	-0.31	-0.049
Local_Jitter_Mean	-0.215	0.005	0.202
Local_Abs._Jitter_Mean	-0.226	-0.013	0.184

Feature	B vs. C	B vs. M	C vs. M
VoPT	-0.109	0.003	0.108
RMS_Energy_Mean	0.163	-0.144	-0.275
RMS_Energy_1	0.11	-0.088	-0.185
RMS_Energy_2	0.156	-0.169	-0.291
RMS_Energy_3	0.222	-0.177	-0.341
HNR05_Mean	0.315	-0.224	-0.479
HNR05_1	0.208	-0.149	-0.337
HNR05_2	0.234	-0.26	-0.451
HNR05_3	0.336	-0.135	-0.421
HNR15_Mean	0.224	0.005	-0.211
HNR15_1	0.135	0.042	-0.091
HNR15_2	0.18	-0.081	-0.245
HNR15_3	0.257	0.053	-0.196
HNR25_Mean	0.167	-0.035	-0.191
HNR25_1	0.079	-0.009	-0.085
HNR25_2	0.14	-0.103	-0.229
HNR25_3	0.207	0.022	-0.172
HNR35_Mean	0.137	-0.063	-0.188
HNR35_1	0.047	-0.045	-0.09
HNR35_2	0.118	-0.124	-0.231
HNR35_3	0.18	0.006	-0.159
F1_Mean	0.088	-0.122	-0.213
F1_1	0.074	-0.121	-0.201
F1_2	0.089	-0.106	-0.199
F1_3	0.076	-0.123	-0.2
Vowel_Duration	0.176	-0.012	-0.18

Table J.2: Hmong Random Forest Feature Importance, Resampled

Feature	Importance
H1* – H2*_Mean	0.016185212
H1* – H2*_1	0.023664894
H1* – H2*_2	0.030976219
H1* – H2*_3	0.014954687
H2* – H4*_Mean	0.00941201
H2* – H4*_1	0.008516977
H2* – H4*_2	0.009743312
H2* – H4*_3	0.011607372
H1* – A1*_Mean	0.015625343
H1* – A1*_1	0.01622215
H1* – A1*_2	0.015025199
H1* – A1*_3	0.009017821
H1* – A2*_Mean	0.012873382
H1* – A2*_1	0.010316776
H1* – A2*_2	0.011109601
H1* – A2*_3	0.008207743
H1* – A3*_Mean	0.01088469
H1* – A3*_1	0.010381409
H1* – A3*_2	0.010293037
H1* – A3*_3	0.009918174
H4* – 2k*_Mean	0.009389753
H4* – 2k*_1	0.008958652
H4* – 2k*_2	0.008154207
H4* – 2k*_3	0.008695091
2k* – 5k_Mean	0.011973696
2k* – 5k_1	0.005838931
2k* – 5k_2	0.007684676
2k* – 5k_3	0.008775119
CPP_Mean	0.009536461
CPP_1	0.009014271
CPP_2	0.011294557
CPP_3	0.011726939
RMS_Energy_Mean	0.011269333
RMS_Energy_1	0.00944552
RMS_Energy_2	0.013384546
RMS_Energy_3	0.024221391
HNR05_Mean	0.025284644
HNR05_1	0.009672617
HNR05_2	0.044904721
HNR05_3	0.032610813
HNR15_Mean	0.01073416
HNR15_1	0.010861238
HNR15_2	0.01089019
HNR15_3	0.011769106
HNR25_Mean	0.008721569

Feature	Importance
HNR25_1	0.008989773
HNR25_2	0.012890187
HNR25_3	0.00949995
HNR35_Mean	0.011695168
HNR35_1	0.009978814
HNR35_2	0.012853642
HNR35_3	0.007755432
SHR_Mean	0.035641631
SHR_1	0.038390971
F1_Mean	0.0142398
F1_1	0.009365257
F1_2	0.010388027
F1_3	0.01426018
Vowel_Duration	0.010615674
Local_Jitter_Mean	0.071644718
Local_Abs._Jitter_Mean	0.0875743
VoPT	0.034468264

Appendix K

MANDARIN FEATURE METRICS

Table K.1: Mandarin Feature Correlations

Feature	M vs. C
H1* – A1*_Mean	-0.195
H1* – A1*_1	-0.103
H1* – A1*_2	-0.152
H1* – A1*_3	-0.195
H1* – A2*_Mean	-0.239
H1* – A2*_1	-0.196
H1* – A2*_2	-0.193
H1* – A2*_3	-0.193
H1* – A3*_Mean	-0.071
H1* – A3*_1	-0.038
H1* – A3*_2	-0.066
H1* – A3*_3	-0.069
H1* – H2*_Mean	-0.406
H1* – H2*_1	-0.301
H1* – H2*_2	-0.349
H1* – H2*_3	-0.472
H2* – H4*_Mean	0.03
H2* – H4*_1	0.056
H2* – H4*_2	0.007
H2* – H4*_3	0.019
H4* – 2k*_Mean	0.014
H4* – 2k*_1	0.013
H4* – 2k*_2	0.078
H4* – 2k*_3	-0.069
2k* – 5k_Mean	-0.068
2k* – 5k_1	-0.132
2k* – 5k_2	-0.183
2k* – 5k_3	0.159
Local_Jitter_Mean	0.733

Feature	M vs. C
Local_Jitter_1	0.689
Local_Jitter_2	0.534
Local_Jitter_3	0.464
Local_Abs._Jitter_Mean	0.716
Local_Abs._Jitter_1	0.657
Local_Abs._Jitter_2	0.614
Local_Abs._Jitter_3	0.496
RAP_Jitter_Mean	0.635
RAP_Jitter_1	0.556
RAP_Jitter_2	0.441
RAP_Jitter_3	0.396
PPQ5_Jitter_Mean	0.701
PPQ5_Jitter_1	0.603
PPQ5_Jitter_2	0.473
PPQ5_Jitter_3	0.468
Local_Shimmer_Mean	0.728
Local_Shimmer_1	0.725
Local_Shimmer_2	0.405
Local_Shimmer_3	0.263
Local_Shimmer_dB_Mean	0.611
Local_Shimmer_dB_1	0.729
Local_Shimmer_dB_2	0.343
Local_Shimmer_dB_3	0.295
APQ3_Shimmer_Mean	0.621
APQ3_Shimmer_1	0.561
APQ3_Shimmer_2	0.26
APQ3_Shimmer_3	0.322
APQ5_Shimmer_Mean	0.64
APQ5_Shimmer_1	0.626
APQ5_Shimmer_2	0.648
APQ11_Shimmer_Mean	0.645
SHR_Mean	0.168
SHR_3	0.365
CPP_Mean	-0.468
CPP_1	-0.38
CPP_2	-0.536
CPP_3	-0.067
VoPT	0.644
RMS_Energy_Mean	-0.65
RMS_Energy_1	-0.598

Feature	M vs. C
RMS_Energy_2	-0.667
RMS_Energy_3	-0.104
HNR05_Mean	-0.755
HNR05_1	-0.599
HNR05_2	-0.756
HNR05_3	-0.444
HNR15_Mean	-0.745
HNR15_1	-0.57
HNR15_2	-0.75
HNR15_3	-0.508
HNR25_Mean	-0.733
HNR25_1	-0.57
HNR25_2	-0.737
HNR25_3	-0.512
HNR35_Mean	-0.719
HNR35_1	-0.549
HNR35_2	-0.722
HNR35_3	-0.527
F1_Mean	0.259
F1_1	0.602
F1_2	0.51
F1_3	-0.047
Vowel_Duration	0.233

Table K.2: Mandarin Random Forest Feature Importance, Resampled

Feature	Importance
H1* – H2*_Mean	0
H1* – H2*_1	0
H1* – H2*_2	0.004176446
H1* – H2*_3	0.010999425
H2* – H4*_Mean	0
H2* – H4*_1	0
H2* – H4*_2	0
H2* – H4*_3	0.003913043
H1* – A1*_Mean	0
H1* – A1*_1	0

Feature	Importance
H1* – A1*_2	0.001992754
H1* – A1*_3	0.006500057
H1* – A2*_Mean	0
H1* – A2*_1	0
H1* – A2*_2	0.00417074
H1* – A2*_3	0.001389492
H1* – A3*_Mean	0
H1* – A3*_1	0
H1* – A3*_2	0
H1* – A3*_3	0
H4* – 2k*_Mean	0.007590746
H4* – 2k*_1	0
H4* – 2k*_2	0
H4* – 2k*_3	0
2k* – 5k_Mean	0
2k* – 5k_1	0
2k* – 5k_2	0
2k* – 5k_3	0
CPP_Mean	0.000437508
CPP_1	0.001051402
CPP_2	0
CPP_3	0
RMS_Energy_Mean	0
RMS_Energy_1	0
RMS_Energy_2	0.084176887
RMS_Energy_3	0.008799802
HNR05_Mean	0.271200432
HNR05_1	0.011741794
HNR05_2	0.162459998
HNR05_3	0.002072318
HNR15_Mean	0.012842688
HNR15_1	0
HNR15_2	0
HNR15_3	0
HNR25_Mean	0.078378342
HNR25_1	0.00626087
HNR25_2	0
HNR25_3	0
HNR35_Mean	0.004662005
HNR35_1	0

Feature	Importance
HNR35_2	0.002060836
HNR35_3	0.002083333
SHR_Mean	0.005783157
SHR_3	0
F1_Mean	0
F1_1	0.002064989
F1_2	0
F1_3	0
Vowel_Duration	0
Local_Jitter_Mean	0.001947826
Local_Jitter_1	0.006581821
Local_Jitter_2	0.081497896
Local_Jitter_3	0
Local_Abs._Jitter_Mean	0.07621498
Local_Abs._Jitter_1	0.065285182
Local_Abs._Jitter_2	0
Local_Abs._Jitter_3	0
RAP_Jitter_Mean	0
RAP_Jitter_1	0.00292328
RAP_Jitter_2	0.024644547
RAP_Jitter_3	0
PPQ5_Jitter_Mean	0.0110616
PPQ5_Jitter_1	0
PPQ5_Jitter_2	0
PPQ5_Jitter_3	0
Local_Shimmer_Mean	0
Local_Shimmer_1	0.004077015
Local_Shimmer_2	0
Local_Shimmer_3	0
Local_Shimmer_dB_Mean	0
Local_Shimmer_dB_1	0
Local_Shimmer_dB_2	0.011748018
Local_Shimmer_dB_3	0
APQ3_Shimmer_Mean	0
APQ3_Shimmer_1	0
APQ3_Shimmer_2	0
APQ3_Shimmer_3	0
APQ5_Shimmer_Mean	0.01452242
APQ5_Shimmer_1	0
APQ5_Shimmer_2	0.002686351

Feature	Importance
APQ11_Shimmer_Mean	0
VoPT	0

Appendix L

MAZATEC FEATURE METRICS

Table L.1: Mazatec Feature Correlations

Feature	B vs. C	B vs. M	C vs. M
H1* – H2*_Mean	0.306	-0.135	-0.484
H1* – H2*_1	0.357	0.049	-0.368
H1* – H2*_2	0.383	-0.088	-0.505
H1* – H2*_3	0.08	-0.294	-0.392
H2* – H4*_Mean	0.073	0.034	-0.037
H2* – H4*_1	0.26	0.121	-0.139
H2* – H4*_2	-0.018	0.012	0.033
H2* – H4*_3	-0.041	-0.038	0.001
H1* – A1*_Mean	0.387	-0.006	-0.436
H1* – A1*_1	0.547	0.223	-0.406
H1* – A1*_2	0.363	-0.019	-0.426
H1* – A1*_3	0.147	-0.162	-0.332
H1* – A2*_Mean	0.385	-0.025	-0.446
H1* – A2*_1	0.429	0.089	-0.387
H1* – A2*_2	0.421	-0.01	-0.474
H1* – A2*_3	0.226	-0.127	-0.371
H1* – A3*_Mean	0.322	-0.011	-0.372
H1* – A3*_1	0.385	0.07	-0.346
H1* – A3*_2	0.343	-0.005	-0.389
H1* – A3*_3	0.19	-0.091	-0.309
H4* – 2k*_Mean	0.249	0.036	-0.248
H4* – 2k*_1	0.148	-0.023	-0.201
H4* – 2k*_2	0.266	0.044	-0.247
H4* – 2k*_3	0.245	0.068	-0.198
2k* – 5k_Mean	-0.036	0.083	0.151
2k* – 5k_1	-0.052	0.104	0.19
2k* – 5k_2	-0.046	0.011	0.068
2k* – 5k_3	0.001	0.096	0.122
CPP_Mean	0.081	0.026	-0.067

Feature	B vs. C	B vs. M	C vs. M
CPP_1	-0.118	-0.146	-0.045
CPP_2	0.162	-0.002	-0.193
CPP_3	0.136	0.243	0.116
Local_Jitter_Mean	-0.247	-0.035	0.291
Local_Jitter_1	-0.112	0.093	0.212
Local_Jitter_2	-0.245	-0.028	0.284
Local_Jitter_3	-0.177	-0.096	0.091
Local_Abs._Jitter_Mean	-0.158	0.018	0.211
Local_Abs._Jitter_1	-0.041	0.175	0.162
Local_Abs._Jitter_2	-0.214	-0.026	0.239
Local_Abs._Jitter_3	-0.056	-0.037	0.017
RAP_Jitter_Mean	-0.233	-0.066	0.229
RAP_Jitter_1	-0.029	0.135	0.165
RAP_Jitter_2	-0.194	-0.044	0.206
RAP_Jitter_3	-0.201	-0.164	0.036
PPQ5_Jitter_Mean	-0.246	-0.083	0.24
PPQ5_Jitter_1	-0.06	0.097	0.169
PPQ5_Jitter_2	-0.206	-0.035	0.228
PPQ5_Jitter_3	-0.205	-0.146	0.092
Local_Shimmer_Mean	-0.19	-0.074	0.166
Local_Shimmer_1	-0.048	0.052	0.124
Local_Shimmer_2	-0.17	-0.164	0.035
Local_Shimmer_3	0.161	0.016	-0.17
Local_Shimmer_dB_Mean	-0.13	-0.091	0.067
Local_Shimmer_dB_1	-0.088	0.027	0.141
Local_Shimmer_dB_2	-0.155	-0.183	-0.02
Local_Shimmer_dB_3	0.187	0.059	-0.153
APQ3_Shimmer_Mean	-0.17	-0.037	0.17
APQ3_Shimmer_1	-0.041	0.047	0.105
APQ3_Shimmer_2	-0.207	-0.08	0.161
APQ3_Shimmer_3	-0.147	-0.152	0.013
APQ5_Shimmer_Mean	-0.191	-0.088	0.132
APQ5_Shimmer_1	-0.121	-0.014	0.134
APQ11_Shimmer_Mean	-0.054	-0.023	0.039
VoPT	-0.332	-0.02	0.422
RMS_Energy_Mean	0.023	-0.128	-0.179
RMS_Energy_1	-0.041	-0.146	-0.133
RMS_Energy_2	0.146	-0.113	-0.275
RMS_Energy_3	-0.081	-0.038	0.047
HNR05_Mean	0.102	0.042	-0.087

Feature	B vs. C	B vs. M	C vs. M
HNR05_1	0.003	0.057	0.058
HNR05_2	0.117	-0.038	-0.18
HNR05_3	0.136	0.116	-0.052
HNR15_Mean	0.148	0.033	-0.161
HNR15_1	0.013	0.046	0.028
HNR15_2	0.141	-0.049	-0.223
HNR15_3	0.207	0.091	-0.174
HNR25_Mean	0.148	0.023	-0.168
HNR25_1	0.008	0.033	0.022
HNR25_2	0.131	-0.056	-0.22
HNR25_3	0.23	0.091	-0.202
HNR35_Mean	0.112	-0.004	-0.147
HNR35_1	-0.022	0.01	0.038
HNR35_2	0.094	-0.071	-0.19
HNR35_3	0.211	0.068	-0.199
STRAIGHT_f_0_Mean	-0.198	-0.229	-0.055
STRAIGHT_f_0_1	-0.268	-0.279	-0.047
STRAIGHT_f_0_2	-0.069	-0.177	-0.12
STRAIGHT_f_0_3	-0.219	-0.202	0.017
Snack_f_0_Mean	-0.02	-0.219	-0.185
Snack_f_0_1	-0.117	-0.239	-0.096
Snack_f_0_2	-0.013	-0.225	-0.175
Snack_f_0_3	-0.04	-0.128	-0.077
Praat_f_0_Mean	-0.23	-0.285	-0.056
Praat_f_0_1	-0.282	-0.323	-0.03
Praat_f_0_2	-0.062	-0.22	-0.143
Praat_f_0_3	-0.19	-0.215	-0.006
SHR_f_0_Mean	-0.35	-0.294	0.019
SHR_f_0_1	-0.41	-0.33	0.095
SHR_f_0_2	-0.155	-0.233	-0.084
SHR_f_0_3	-0.307	-0.255	0.039
F1_Mean	0.04	0.019	-0.022
F1_1	0.192	0.087	-0.112
F1_2	-0.081	-0.009	0.074
F1_3	-0.034	-0.013	0.022
Vowel_Duration	-0.127	-0.015	0.124

Table L.2: Mazatec Random Forest Feature Importance, Resampled

Feature	Importance
H1* – H2*_Mean	0.008281687
H1* – H2*_1	0.0310014
H1* – H2*_2	0.027132014
H1* – H2*_3	0.018168442
H2* – H4*_Mean	0.00479351
H2* – H4*_1	0.010877809
H2* – H4*_2	0.008694827
H2* – H4*_3	0.011482704
H1* – A1*_Mean	0.007248607
H1* – A1*_1	0.015386202
H1* – A1*_2	0.021144679
H1* – A1*_3	0.009371554
H1* – A2*_Mean	0.017671912
H1* – A2*_1	0.02226283
H1* – A2*_2	0.041961545
H1* – A2*_3	0.012885674
H1* – A3*_Mean	0.018832762
H1* – A3*_1	0.012252393
H1* – A3*_2	0.0198475
H1* – A3*_3	0.004706697
H4* – 2k*_Mean	0.005987912
H4* – 2k*_1	0.003928731
H4* – 2k*_2	0.003150852
H4* – 2k*_3	0.008239407
2k* – 5k_Mean	0.003205263
2k* – 5k_1	0.003154252
2k* – 5k_2	0.003790995
2k* – 5k_3	0.007487181
CPP_Mean	0.006622022
CPP_1	0.004110744
CPP_2	0.009926879
CPP_3	0.009082322
RMS_Energy_Mean	0.010524442
RMS_Energy_1	0.006590252
RMS_Energy_2	0.006847569
RMS_Energy_3	0.004097979
HNR05_Mean	0.006289457

Feature	Importance
HNR05_1	0.006229986
HNR05_2	0.007415684
HNR05_3	0.009653772
HNR15_Mean	0.004957903
HNR15_1	0.003865907
HNR15_2	0.002080269
HNR15_3	0.001564256
HNR25_Mean	0.005117859
HNR25_1	0.004939688
HNR25_2	0.007479684
HNR25_3	0.0080756
HNR35_Mean	0.009667618
HNR35_1	0.010361906
HNR35_2	0.002269791
HNR35_3	0.010696707
STRAIGHT_f_0_Mean	0.001602831
STRAIGHT_f_0_1	0.003020306
STRAIGHT_f_0_2	0.007351701
STRAIGHT_f_0_3	0.008795467
Snack_f_0_Mean	0.006558817
Snack_f_0_1	0.014522376
Snack_f_0_2	0.002926746
Snack_f_0_3	0.0090716
Praat_f_0_Mean	0.018448164
Praat_f_0_1	0.055872856
Praat_f_0_2	0.00833534
Praat_f_0_3	0.004746875
SHR_f_0_Mean	0.005572844
SHR_f_0_1	0.035970057
SHR_f_0_2	0.000462029
SHR_f_0_3	0.004558665
F1_Mean	0.008724106
F1_1	0.011047747
F1_2	0.002646672
F1_3	0.006504928
Vowel_Duration	0.008448146
Local_Jitter_Mean	0.008018647
Local_Jitter_1	0.006460396
Local_Jitter_2	0.012761668
Local_Jitter_3	0.004430848

Feature	Importance
Local_Abs._Jitter_Mean	0.007008563
Local_Abs._Jitter_1	0.002887186
Local_Abs._Jitter_2	0.0060336
Local_Abs._Jitter_3	0.002945755
RAP_Jitter_Mean	0.006947529
RAP_Jitter_1	0.003025828
RAP_Jitter_2	0.003707944
RAP_Jitter_3	0.006991352
PPQ5_Jitter_Mean	0.012771888
PPQ5_Jitter_1	0.011635128
PPQ5_Jitter_2	0.021050482
PPQ5_Jitter_3	0.028933285
Local_Shimmer_Mean	0.003796137
Local_Shimmer_1	0.001861653
Local_Shimmer_2	0.018218512
Local_Shimmer_3	0.008290507
Local_Shimmer_dB_Mean	0.007408468
Local_Shimmer_dB_1	0.003199426
Local_Shimmer_dB_2	0.003674637
Local_Shimmer_dB_3	0.00861074
APQ3_Shimmer_Mean	0.003060624
APQ3_Shimmer_1	0.00218909
APQ3_Shimmer_2	0.005428157
APQ3_Shimmer_3	0.001810875
APQ5_Shimmer_Mean	0.005452977
APQ5_Shimmer_1	0.00339702
APQ11_Shimmer_Mean	0.007346434
VoPT	0.024038735

Appendix M

ZAPOTEC FEATURE METRICS

Table M.1: Zapotec Feature Correlations

Feature	B vs. C	B vs. M	C vs. M
H1* – A1*_Mean	0.397	0.504	0.146
H1* – A1*_1	0.204	0.272	0.078
H1* – A1*_2	0.38	0.459	0.12
H1* – A1*_3	0.4	0.545	0.17
H1* – A2*_Mean	0.22	0.273	0.056
H1* – A2*_1	0.108	0.105	-0.009
H1* – A2*_2	0.217	0.263	0.058
H1* – A2*_3	0.222	0.307	0.079
H1* – A3*_Mean	0.229	0.292	0.051
H1* – A3*_1	0.17	0.161	-0.021
H1* – A3*_2	0.234	0.276	0.039
H1* – A3*_3	0.186	0.294	0.087
H1* – H2*_Mean	0.47	0.584	0.129
H1* – H2*_1	0.256	0.217	-0.066
H1* – H2*_2	0.456	0.531	0.079
H1* – H2*_3	0.441	0.607	0.23
H2* – H4*_Mean	0.101	0.185	0.061
H2* – H4*_1	0.013	0.108	0.089
H2* – H4*_2	0.101	0.137	0.031
H2* – H4*_3	0.11	0.222	0.064
H4* – 2k*_Mean	0.116	0.019	-0.083
H4* – 2k*_1	0.172	0.102	-0.074
H4* – 2k*_2	0.055	-0.025	-0.067
H4* – 2k*_3	0.092	-0.011	-0.095
2k* – 5k_Mean	-0.149	-0.219	-0.049
2k* – 5k_1	-0.063	-0.085	-0.008
2k* – 5k_2	-0.105	-0.151	-0.036
2k* – 5k_3	-0.219	-0.318	-0.08
Local_Jitter_Mean	-0.088	0.182	0.206

Feature	B vs. C	B vs. M	C vs. M
Local_Jitter_1	-0.053	-0.026	0.03
Local_Jitter_2	0.135	0.244	0.128
Local_Abs._Jitter_Mean	0.034	0.205	0.143
Local_Abs._Jitter_1	-0.019	-0.009	0.011
Local_Abs._Jitter_2	0.211	0.273	0.089
RAP_Jitter_Mean	-0.078	0.296	0.247
RAP_Jitter_1	-0.082	-0.043	0.051
PPQ5_Jitter_Mean	0.036	0.268	0.255
Local_Shimmer_Mean	0.037	0.253	0.204
Local_Shimmer_1	0.03	0.044	0.014
Local_Shimmer_dB_Mean	0.025	0.307	0.251
Local_Shimmer_dB_1	0.021	0.076	0.054
APQ3_Shimmer_Mean	0.086	0.292	0.2
APQ5_Shimmer_Mean	0.095	0.218	0.149
SHR_Mean	-0.275	-0.446	-0.225
CPP_Mean	-0.225	-0.465	-0.252
CPP_1	-0.037	-0.214	-0.169
CPP_2	-0.273	-0.458	-0.19
CPP_3	-0.263	-0.541	-0.316
VoPT	-0.229	0.267	0.398
RMS_Energy_Mean	-0.366	-0.253	0.011
RMS_Energy_1	-0.283	-0.155	0.064
RMS_Energy_2	-0.39	-0.308	-0.01
RMS_Energy_3	-0.262	-0.269	-0.098
HNR05_Mean	-0.059	-0.45	-0.355
HNR05_1	0.005	-0.265	-0.257
HNR05_2	-0.122	-0.476	-0.32
HNR05_3	0.005	-0.386	-0.344
HNR15_Mean	0.099	-0.198	-0.282
HNR15_1	0.115	-0.094	-0.205
HNR15_2	0.024	-0.259	-0.272
HNR15_3	0.161	-0.128	-0.256
HNR25_Mean	-0.004	-0.274	-0.273
HNR25_1	0.04	-0.169	-0.209
HNR25_2	-0.062	-0.309	-0.253
HNR25_3	0.039	-0.226	-0.256
HNR35_Mean	-0.059	-0.32	-0.269
HNR35_1	-0.01	-0.224	-0.21
HNR35_2	-0.111	-0.346	-0.246
HNR35_3	-0.017	-0.265	-0.25

Feature	B vs. C	B vs. M	C vs. M
Praat_f_0_Mean	-0.159	-0.114	0.102
Praat_f_0_1	-0.001	-0.032	-0.022
Praat_f_0_2	-0.155	-0.171	0.075
Praat_f_0_3	-0.177	-0.095	0.134
SHR_f_0_Mean	-0.372	-0.125	0.294
SHR_f_0_1	-0.235	-0.184	0.114
SHR_f_0_2	-0.383	-0.232	0.218
SHR_f_0_3	-0.29	0.04	0.332
Snack_f_0_Mean	-0.215	-0.45	-0.159
Snack_f_0_1	-0.038	-0.169	-0.1
Snack_f_0_2	-0.201	-0.428	-0.113
STRAIGHT_f_0_Mean	-0.251	-0.334	0.048
STRAIGHT_f_0_1	-0.102	-0.178	0.002
STRAIGHT_f_0_2	-0.203	-0.393	-0.048
STRAIGHT_f_0_3	-0.312	-0.308	0.104
F1_Mean	-0.097	0.096	0.196
F1_1	-0.188	-0.093	0.111
F1_2	-0.078	0.075	0.166
F1_3	-0.038	0.229	0.265
Vowel_Duration	0.282	0.291	-0.022

Table M.2: Zapotec Random Forest Feature Importance, Resampled

Feature	Importance
H1* – H2*_Mean	0.011773561
H1* – H2*_1	0.014956813
H1* – H2*_2	0.028067977
H1* – H2*_3	0.036588549
H2* – H4*_Mean	0.006452245
H2* – H4*_1	0.000906079
H2* – H4*_2	0.012501926
H2* – H4*_3	0.007818901
H1* – A1*_Mean	0.029892699
H1* – A1*_1	0.001181032
H1* – A1*_2	0.00403561
H1* – A1*_3	0.042699329
H1* – A2*_Mean	0.017401658

Feature	Importance
H1* – A2*_1	0.00525501
H1* – A2*_2	0.022644742
H1* – A2*_3	0.021990012
H1* – A3*_Mean	0.013567517
H1* – A3*_1	0.003784096
H1* – A3*_2	0.010561528
H1* – A3*_3	0.007076423
H4* – 2k*_Mean	0.014979572
H4* – 2k*_1	0.012411601
H4* – 2k*_2	0.004219554
H4* – 2k*_3	0.007227069
2k* – 5k_Mean	0.004969219
2k* – 5k_1	0.004075038
2k* – 5k_2	0.02050916
2k* – 5k_3	0.022684273
CPP_Mean	0.009951385
CPP_1	0.005579004
CPP_2	0.031042934
CPP_3	0.025770216
RMS_Energy_Mean	0.012201166
RMS_Energy_1	0.011510422
RMS_Energy_2	0.020711579
RMS_Energy_3	0.007921085
HNR05_Mean	0.024257972
HNR05_1	0.026198279
HNR05_2	0.015750052
HNR05_3	0.003987361
HNR15_Mean	0.01932987
HNR15_1	0.008634317
HNR15_2	0.003285422
HNR15_3	0.00907092
HNR25_Mean	0.002245693
HNR25_1	0.007789353
HNR25_2	0.012876732
HNR25_3	0.012960846
HNR35_Mean	0.009352697
HNR35_1	0.006316465
HNR35_2	0
HNR35_3	0.010462334
SHR_Mean	0.027680993

Feature	Importance
F1_Mean	0.013416121
F1_1	0.008284978
F1_2	0.0053492
F1_3	0.020544261
Vowel_Duration	0.028222538
Local_Jitter_Mean	0.011163993
Local_Jitter_1	0.020801447
Local_Jitter_2	0.019578107
Local_Abs._Jitter_Mean	0.009033411
Local_Abs._Jitter_1	0.015946316
Local_Abs._Jitter_2	0.009007298
RAP_Jitter_Mean	0.034012768
RAP_Jitter_1	0.003624695
PPQ5_Jitter_Mean	0.023317037
Local_Shimmer_Mean	0.000663797
Local_Shimmer_1	0.004001564
Local_Shimmer_dB_Mean	0.010206468
Local_Shimmer_dB_1	0.004834494
APQ3_Shimmer_Mean	0.002860358
APQ5_Shimmer_Mean	0.011993361
VoPT	0.0280195