EXPOSING THE HIDDEN VOCAL CHANNEL: ANALYSIS OF VOCAL EXPRESSION

BY

MARY B. PIETROWICZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Karrie Karahalios, Chair
Professor Mark Hasegawa-Johnson, Co-Chair
Associate Professor Julia Hockenmaier
Associate Professor Jerome McDonough
Professor Jennifer S. Cole
Associate Professor Gina-Anne Levow, University of Washington

ABSTRACT

This dissertation explored perception and modeling of human vocal expression, and began by asking what people heard in expressive speech. To address this fundamental question, clips from Shakespearian soliloquy and from the Library of Congress Veterans Oral History Collection were presented to Mechanical Turk workers (10 per clip); and the workers were asked to provide 1-3 keywords describing the vocal expression in the voice. The resulting keywords described prosody, voice quality, nonverbal quality, and emotion in the voice, along with the conversational style, and personal qualities attributed to the speaker. More than half of the keywords described emotion, and were wide-ranging and nuanced. In contrast, keywords describing prosody and voice quality reduced to a short list of frequently-repeating vocal elements.

Given this description of perceived vocal expression, a 3-step process was used to model vocal qualities which listeners most frequently perceived. This process included 1) an interactive analysis across each condition to discover its distinguishing characteristics, 2) feature selection and evaluation via unequal variance sensitivity measurements and examination of means and 2-sigma variances across conditions, and 3) iterative, incremental classifier training and validation. The resulting models performed at 2-3.5 times chance. More importantly, the analysis revealed a continuum relationship across whispering, breathiness, modal speech, and resonance, and revealed multiple spectral sub-types of breathiness, modal speech, resonance, and creaky voice.

Finally, latent semantic analysis (LSA) applied to the crowdsourced keyword descriptors enabled organic discovery of expressive dimensions present in each corpus, and revealed relationships among perceived voice qualities and emotions within each dimension and across the corpora. The resulting dimensional classifiers performed at up to 3 times chance, and a second study presented a dimensional analysis of laughter.

This research produced a new way of exploring emotion in the voice, and of examining relationships among emotion, prosody, voice quality, conversation quality, personal quality, and other expressive vocal elements. For future work, this perception-grounded fusion of crowdsourcing and LSA technique can be applied to anything humans can describe, in any research domain.

To Steve with love... The link is strong with this one.

To all my family, blessings to each one, for their love, encouragement, and support.

In memory of
John Robert McGregor (1936-2014)
John Charles McGregor (1904-1992)
My father and grandfather,
who walked this path before me.
Till we meet again.

ACKNOWLEDGEMENTS

Thank you to all who helped me during my years in graduate school. It is a debt I cannot repay, and words truly fail to express my gratitude. Thank you to the department for the opportunity. I wasn’t your traditional student. I did some unusual things and had some unusual responsibilities. In the darkest days I buried my dad. Thank you to those who had kind words, because your words meant the world to me. I was not able to talk about it very much, at the time. Thank you, for the opportunity and support.

My professional life is changed forever. I am humbled at the experiences I have had while in the program, and at the doors which are open to me now which were closed before. The process has changed me personally. I hope I can do the same for others; the opportunity was given for a reason.

Thank you to my advisors Karrie and Mark. I was blessed to have your guidance, example, and many enjoyable, interesting talks over the years. Thank you also to my committee members Jennifer Cole, Jerome McDonough, Gina Levow, and Julia Hockenmaier. Your perspectives have helped show me what it means to be a researcher.

Thank you, friends in my two research groups. It has been a pleasure to know you and work with you, and I’ll miss you! May our paths cross again.

Thank you most of all to my family, and especially to my husband Steve. How lucky I am to have you. Without you, this experience, and so many other wonderful things, would never have been possible. I am most proud to have stood with you to witness our daughters as they turned into beautiful, accomplished, strong, young women during the last six years.

Blessings, all.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Speech is a ubiquitous human mechanism for expressing thoughts, requests, commands and many other kinds of ideas. People often think about speech communication from the perspective of the word content of a language, but speech communicates much more than that. It telegraphs what we think about what we and others are saying, how we are emotionally, and how we feel physically. We often know intuitively when someone isn't feeling well, and when someone is traumatized, depressed, bored, or excited, just by the expressive gesture in the voice, the combination of variations in pitch, speed, emphasis, loudness, articulation, vocal timbre, and more. Sometimes, however, human ears miss subtle cues or simply ignore signals which we are unable or unwilling to address. What if, failing human ears, electronic ears could be trained to interpret vocal expression, and detect problems in mental or physical health and wellness? In the mental health realm, the impact could result in people with un-diagnosed or untreated PTSD, anxiety, or depression being referred for help and treatment, where treatment could prevent suicides, homelessness, and other suffering. In the realm of physical health, could speech patterns offer clues to wellness, with respect to conditions such as Alzheimer's or Parkinson's disease? The Nun Studies (Snowdon, 1997) discovered links between writing patterns from youth and the likelihood of developing Alzheimer's later in life. Could a longitudinal study using voices offer clues as well?

To quantify a fraction of the potential of this, an estimated 33,300 homeless veterans in the US have PTSD (American Psychological Association). What if their condition had been detected upon military discharge via a mobile voice analysis tool, and what if they had been offered the opportunity for treatment instead of being released to the street? A much larger number of Americans, about 5.5 million, have Alzheimer's disease (Alzheimer's Association). What if the millions of Americans currently affected by Alzheimer's had the opportunity to take preventative steps in their youth, or had the chance to receive early intervention for their condition? Could their difficulties have been avoided, or could the progression of disease been slowed down? An even larger segment of the population, about 40 million adults in the United States, have some form of anxiety disorder; and only 36.8% receive treatment (Anxiety and Depression Association of America). What if screening, diagnosis and treatment were more readily available for those who wished it?

From a more utilitarian point of view, consider TED talks, YouTube videos, movies, conference talks, and class lectures. Each of these kinds of resources has spoken content, and we have limited ability examine un-transcribed resources electronically. We typically use search engines to find them, and use simple controls like play, rewind, forward, and pause to review them. Using existing search engines to locate them is a mismatch from the beginning, because the user must express the request as *text keywords*; but the resource itself contains a *speech sound stream*, and *no text*. Even if the user manages to find an interesting resource, the only way to explore it is to play it. What if a sound resource could instead tell us something about itself, so that we could access the relevant parts directly, without having to play it from beginning to end?

My work seeks to enable such applications by understanding what people hear in human expression, understanding the relationships among expressive elements in speech, and developing processes and techniques for analyzing and modeling expressive speech which are better aligned with human perception. Such human-aligned analytics will be better suited to supporting application development, since applications are made to serve humans. In short, this work explores the question, "Can expressive vocal analytics be grounded in human perception?" It presents evidence from human perception studies, new analytic processes and methods, and validated baseline models which demonstrate analytics grounded in human expression.

1.1 Scope

At the highest level, this research explores human expressive speech by asking the following questions:

RQ1: How do people hear vocal expression; what elements do they perceive most strongly and consistently?

RQ2: What models can detect the elements and dimensions of vocal expression which people hear most strongly and consistently?

In more detail, it asks, with respected scripted (acted) voices:

RQ3: What elements of vocal expression do untrained listeners hear in male acted voices?

RQ4: What acoustic features can distinguish each of four levels of vocal effort, specifically whispering, breathiness, modal speech, and resonant/projected speech (these are expressive voice qualities) in male actor's voices?

RQ5: What elements of vocal expression do untrained listeners hear in female acted voices?
RQ6: What acoustic features can distinguish each of four levels of vocal effort (whispering, breathiness, modal speech, and resonant/projected speech) in female actor's voices?
RQ7: How does perception and the corresponding analytic techniques for perceived features differ between male and female voices?

Also in more detail, it asks with respect to semi-structured, unscripted voices:
RQ8: What elements of vocal expression do people hear in recorded, unscripted speech (and how do they compare with scripted speech)?
RQ9: What acoustic features map to these perceived elements of vocal expression (and how do they compare to the acoustic features of scripted speech)?

Again in more detail, it asks with respect to both scripted and semi-structured unscripted voices:
RQ10: What is the relationship between perception of selected emotions and voice qualities, particularly effort levels, in recorded, unscripted and unscripted speech?
RQ11: What models support detection of these perceived elements of vocal expression, and of the relationships between emotion and other elements of vocal expression?

This work crosses multiple disciplines by necessity. It draws from and contributes to 1) Human Computer Interaction and Psychology (RQ1, RQ3, RQ5, RQ7, RQ8, and RQ10), 2) Speech and Language Processing (RQ2, RQ4, RQ6, RQ7, RQ9, RQ10, and RQ11), 3) Computational Linguistics (RQ1-11), 4) Natural Language Processing (RQ10 and RQ11), 5) Information Science (RQ2 and RQ11), and 6) Vocal Performance and Acting (RQ3, RQ4, RQ5, RQ6, RQ7, RQ10, and RQ11). Less directly, this work is related to security (through whispering as covert activity), search and retrieval (through explorations of human perception and modeling of these perceived features), digital archival techniques (through curation of the corpora used in this research and the research methods used here), and software engineering (through the research methods applied here). Chapter 2 discusses prior work in detail.

1.2 Contributions

The contributions of this work include the following:

1) **An end-to-end, cross-disciplinary process for grounding human expression analytics in human perception.** Chapter 3 gives a high-level overview of the process, and chapters 4, 5, 6, and 7 describe specific steps leading to the development of perception-grounded analytics for vocal expression. All the research questions (RQ1-11) were critical for the outcome of this contribution.

   Here, I defined and piloted a new process incorporating perceived feature discovery, signal feature discovery and evaluation, machine modeling, and machine model validation. ***This process 1) bridges disciplines, and 2) can produce software artifacts which are better suited to support application development, because these services are aligned with human perception and human needs.*** Taken in its entirely, this process is new to speech researchers, linguists, natural language processing researchers, software engineering experts, and human-computer-interaction (HCI) scientists. For example, speech researchers (arguably armed with the best tools for signal processing and modeling) do not, as of the writing of this thesis, routinely incorporate human perception studies into the research process. In contrast, while HCI scientists routinely run user studies and employ crowdsourcing techniques, they typically do not model speech, and produce more documented research on visual interaction modalities than sonic experiences, including speech and language.

2) **The confirmation of the vocal progression through whispering, breathiness, modal speech, and resonance as a continuum, from both the human perception and acoustic analytic points of view.** Chapters 5 and 6 cover the human perception and acoustic analysis, respectively. The investigation of RQ1-RQ6, RQ8, and RQ9 were the most relevant questions for this discovery.

The treatment of whispering, breathiness, modal speech, and resonance as a continuum is a new idea for speech researchers, linguists, speech pathologists, security researchers, and vocal performance artists. Prior research typically treated these different phonation types, or effort levels, as discrete categories, not related entities. This could have been a result of the disjoint motivating factors for investigation from different disciplines. For example, many investigations of breathiness were driven by 1) speech pathology questions (patients who could not phonate normally), and by 2) speech researchers' observations that speech recognition techniques which worked well for modal speech did not work as well when applied to breathy phonation.

Prior investigations into whispering were also driven by specific questions from different disciplines, in this case, many investigations from speech pathology, security, and speech processing. Speech pathologists again wanted to understand problems which prevented normal phonation in their patients. From the perspective of security researchers, speakers typically only whispered over the phone when they did not want others to hear; therefore, whispering in phone speech was a signal for covert activity. From the perspective of speech processing, researchers again noticed difficulties in speech recognition during whispered phonation, and needed ways of detecting and processing this kind of speech.

Prior work in voice resonance was often driven by acting and speech performance technique. Speech researchers have not documented research for this voice quality to the degree that they have examined whispering and breathiness.

The continuum relationship provides new intuition for researchers who study even just one of these vocal qualities. For example, the new model design and feature selection described in Chapter 6 of this thesis reflects a single, direct impact of this insight for speech processing. Would speech pathologists and security researchers also make different diagnostic or design decisions, understanding the fuzzy boundaries between whispering and breathiness? Would vocal performers use their abilities to transition across effort levels differently, for maximal impact, with this insight?

3) **Baseline classifiers and feature sets for recognizing effort levels within male and female scripted and unscripted speech.** Chapter 6 covers this topic. Again, the investigation of RQ1-RQ6, RQ8, and RQ9 were the most relevant questions for this result.

   Use of entropy and entropy ratio features across the continuum from whispering, to breathiness, to modal speech, to resonance is new for speech processing. Application of entropy features, however, has been used to distinguish single voice qualities such as whispering from modal speech. The grounding of these models in human perception is also new, and is a result of the perception studies, which used techniques from HCI.

4) **Comparison between male and female effort levels in scripted and unscripted speech.** Chapter 6 and RQ7 cover this topic. Taken as a whole, gender comparisons across the continuum of effort levels, from whispering through resonance, are new. Gender comparisons within single voice qualities have been done before using different collections of signal features than were used here.

5) **A new, perception-grounded technique for discovering dimensions of expressive speech present in corpora.** Chapter 7 describes this process and presents the dimensions discovered in male and female scripted and unscripted speech corpora. RQ10 motivated the exploration.

   This technique is new. It combines crowdsourcing, latent semantic analysis, and data analysis techniques, and applies them to the human perception of expressive speech. While neither crowdsourcing nor latent semantic analysis are new by themselves, the combination of these techniques applied to the crowdsourced human perception of expressive speech is groundbreaking. ***It enabled an entirely new way of exploring the perception and recognition of human emotion, which potentially impacts, at a minimum, the fields of HCI, speech and language processing, computational linguistics, and psychology.*** This technique uses the information contained in nuanced descriptions of perceived emotions, instead of reducing all emotive perceptions into a short list of emotions considered "basic."

The technique also ***enables the organic discovery of expressive dimensions in a corpus***, instead of requiring mapping of perceived emotions onto predefined axes such as arousal, valence, and dominance. This thesis does discuss, however, methods for bridging between organically-discovered dimensions and predefined axes, so that relationships between this work and prior work can be leveraged and analyzed.

6) **A new technique for discovering relationships among perceived and measured emotion, voice quality, prosody, personal quality, and conversational quality.** The relationships between emotion and voice quality present in male and female scripted and unscripted corpora are described in Chapter 7 (and motivated by RQ10).

   The application of this technique is new, and again, crowdsourcing and latent semantic analysis are not. The potential impact of this technique, however, is possibly groundbreaking.

   This technique enables the systematic exploration of relationships among multiple categories of human description and perception. This thesis specifically explores relationships between emotion and voice quality, but this work could immediately be extended to explore relationships among emotion, voice quality, prosody, personal quality, and conversational quality.

   Furthermore, ***this is a generalized technique for exploring relationships among perceived qualities in any kind of multimodal human expression, not just speech, and in any observed entity which humans can describe***. For this research, the technique immediately extends to other modalities of human expression, including language text, whole body or localized physical gesture, eye focus, and any other expressive nuance. Beyond the space of human expression exploration, it extends to evaluation of creative artifacts, human interaction experiences, and consumer products – literally anything which humans can observe and describe. In addition to HCI and speech and language processing, the technique is potentially useful for augmenting research in psychology, computational linguistics, natural language processing, information science, digital archival, search and retrieval, education, and many other disciplines.

7) **An exploration of the dimensions of laughter present in unscripted speech**; this is described in Chapter 8. RQ1, RQ2, RQ8, RQ9, and RQ10 led to this contribution.

   The systematic discovery of multiple dimensions of laughter is new. It enables exploring the many sonic and functional modes of laughter. As the discussion in Chapter 8 confirms, not all laughter occurs in response to humor; and not all laughter reflects positive affect. This result demonstrates that dimensional discovery techniques can be applied to lower-level expressive elements such as laughter as effectively as they can be applied to the discovery of higher-level dimensions.

8) **Curation of suitable corpora for exploring male and female scripted and semi-structured unscripted speech**. The curation process is described in overview in Chapter 3 and in detail in Chapter 4.

   The curation methods explored here support the proposed end-to-end research process, and as a whole, they are new. These methods are suggested for digital archival of artifacts containing speech, as well as for speech analytics.

1.3 Overview

This dissertation explores human vocal expression and presents a perception-grounded process for exploring and modeling human expression.

Chapter 2 presents relevant background and prior work. It also highlights important areas which I have leveraged or extended. It discusses relevant areas in prosody, vocal quality, nonverbal quality, emotion, gender differences in speech, grounding in perception, oral history datasets and practices, the dimensional analysis of expressive speech, and (briefly) common machine learning techniques which have been applied to vocal expression and I have used or considered for use.

Chapter 3 presents an overview of the interdisciplinary, end-to-end research process and methods for my work. This process is new (defined and piloted in this work) and is intended to ground the resulting analytic machine models in human perception. It covers corpora selection, perception studies, interactive analysis of perceived expressive features, modeling of distinct, frequently-perceived features, discovery of expressive dimensions, modeling the discovered expressive dimensions, and mapping organically-discovered dimensions to other predefined dimensional sets.

Chapter 4 discusses the selection and curation of the datasets I used for both scripted and unscripted speech. The resulting scripted and unscripted corpora source (Shakespearian acted speech from YouTube and veterans' oral histories from the Library of Congress, respectively) are publically accessible.

Chapter 5 presents the vocal expression perception studies, including an overview, the methods, the results, and their statistical significance (RQ1, RQ3, and RQ8). It also provides potential directions for future work, given the results, in the summary. These studies employ crowdsourcing techniques to answer the research questions concerning perception.

Chapter 6 describes the methods used to model frequently-perceived voice qualities in scripted and unscripted speech, specifically effort levels and creaky voice, along with the results (RQ4, RQ6, RQ9, and RQ11). It leverages the results of the perception studies in Chapter 5, and provides comparisons between scripted and unscripted speech, and between male and female speech.

Chapter 7 discusses the analysis and organic discovery of repeating, expressive dimensions in our scripted and unscripted speech corpora. These methods enable leverage of nuanced description from the perception studies and avoid the reduction of nuanced description into large, super-categories (especially useful for addressing emotion). They also enable analysis of the relationships among emotion, voice quality, and other expressive elements in the voice. The discussion includes 1) analysis from within a discovered expressive dimension by examining co-occurrence of strongly-associated descriptors with the dimension, and 2) analysis from a higher level by examining correlation of descriptors across an entire corpus, not just within a dimension. Both are necessary for a complete picture of relationships among descriptor categories (such as emotion and voice quality).

Chapter 8 presents a brief discussion of the dimensions of laughter discovered in the unscripted speech corpora. Dimensions of laughter are not covered for the scripted speech because, while laughter was an important, and frequently-perceived element of the unscripted corpora, it was not a frequently-occurring element of the scripted speech corpora.

Chapter 9 discusses the results and the potential for future explorations.

Finally, Chapter 10 summarizes the conclusions.

CHAPTER 2: BACKGROUND AND PRIOR WORK

This chapter reviews the related prior work, specifically work in analysis and modeling of expressive speech, male-female differences in speech, oral history datasets and practices, selected dimensional analysis techniques (particularly Latent Semantic Analysis), selected machine learning techniques, processing of speech streams, and crowdsourcing techniques. Note that a portion of this information has been referenced as a part of our prior work (Pietrowicz et al., 2015; Pietrowicz et al. 1, 2017; and Pietrowicz et al. 2, 2017).

2.1 Analysis and Modeling of Expressive Speech

This body of work covers the relevant prior work in analysis and modeling of prosody, emotion, voice quality (VQ), and nonverbal quality (NQ). Its primary sources are the fields of linguistics (with emphasis on perception, prosody, and formant analysis), speech and language processing (with emphasis on signal processing, voice quality, nonverbal quality, emotion, and pitch and loudness tracking), psychology (the study of emotion and basic emotions), speech pathology (voice quality, particularly breathiness, whispering, and creakiness in speakers who could not phonate normally), acting and vocal performance (to achieve and measure resonant voice quality), security (to detect whispered voice quality, which is indicative of covert activity), and natural language processing (detection of emotion in text). The references are presented in the sub-sections below, along with the associated research disciplines and objectives.

2.1.1 Analysis of Prosody

Speech prosody covers variation in pitch, loudness, speaking rate, and segment duration. These components are highly inter-related, and global effects should be considered separately from local prosodic effects. For example, at the global level, speaking rate influences segment duration. At the local level, however, segment duration is a prosodic marker of semantic focus. Not only do speakers vary prosody to mark segment boundaries in speech, but they also use it to create emphasis, signal information status, and transmit emotion. Here again, the prosodic function at the global level is different from the prosodic function at local levels. For example, local emphasis functions as a marker for semantic focus. In contrast, global emphasis is related to affect. The relationships among prosodic elements are not universal to all languages; this work considers English only.

Duration and Speaking Rate: Our work has revealed that listeners consciously (explicitly) sensed and described time-based prosody features more frequently than they described other features, but why, and what caused them to perceive speech as fast or slow? And how should speech rate be measured to reflect perception? Furthermore, what should be measured? Pfitzinger asked these questions and demonstrated that people perceive speaking rate to be a linear combination of the syllable rate and the phone rate (Pfitzinger 1998). Specifically, he found the correlations between syllable rate alone, phone rate alone, and a linear combination of syllable and phone rate to be r=0.81, r=0.73, and r=0.88, respectively. He also identified three problems with tracking speaking rate. First, a system had to deal with pauses (his system detected pauses and excluded them from the analysis of speaking rate, instead of including pauses as an element of speaking rate). Next, a system had to select an appropriate window size for the analysis; and finally, a system had to use an appropriate windowing function. If a fixed window size were to be used, it should be greater than the maximum syllable duration, so that it would be possible to find the syllable nucleus. On the other hand, the window size should be small enough so that the speaking rate could remain (ideally) constant within it. If the window were too long, the speaking rate could change with it, which is undesirable for analysis. Pfitzinger observed that the PhonDat corpus revealed that syllable distances were less than 700 ms; and the longest phone duration was 444 ms. Further analysis revealed that the system gave bad results with window sizes < 500 ms and bad results with window sizes > 700 ms; so he selected a window size of 625 ms and a hop size of 100 msec. The hop size was not very important to the analysis, but the Hanning window was found to minimize outliers in comparison to other window types.

Pfitzinger also asked what the sources of variance were for mean phone duration. Phone duration varies across a corpus, across the different kinds of phones, within the same kinds or classes of phones, and across speakers. Prior to Pfitzinger's work, speech models rejected the idea of, or did not consider, speaker-dependent phone duration. Other models (Rao, 2012; van Santen, 1993) were either sequential (expert-defined) rule systems (Klatt, 1979; Kumar, 1990), additive or multiplicative systems (which multiplied an intrinsic duration by or added to a series of context-dependent parameters) (van Santen, 1993; van Santen 1994), tree-based models (which followed a path through a tree to calculate duration) (Mobius and van Santen, 1996; Riley, 1992), or stochastic models based on ANNs or HMMs (Campbell, 1990; Campbell, 1992; Wang, 1997). Still others were hybrids of the classes Van Santen identified (Corrigan, et al., 1997). Pfitzinger found that by normalizing out the variation in phone duration, the resulting variance from the speaker source would be greatly reduced, showing that the speaker did indeed contribute to the speaking rate variance (Pfitzinger, 2002a).

Pfitzinger asked the more general question of whether it was possible to reduce the amount of unexplained duration variability by normalizing the speech rate. To address this question, two things are needed: 1) a reliable local rate estimation procedure, and 2) a reliable normalization procedure. Within the window, he took the average of the durations, and took the reciprocal to get the rate per window of syllables and phone. This method introduced discontinuities in the rate curve, over time; so he proposed an alternate, but more computationally-expensive process. To get the perceived local speech rate (PLSR), he applied the linear combination of local syllable and local phone rate which has been shown to provide a good approximation of rate perception. Doing a normalization on the data, then, means calculating the local speech rate curve, taking the inverse of it, and applying the inverse to the data.

Further examination of the data, however, revealed that the phone duration distribution and changes to phone duration distribution (caused by normalization) were not Gaussian, but skewed. Using a general linear model (GLM) on such data would not be correct. The author addressed the problem by taking the log of the original and normalized phone durations, the result of which more closely approximated Gaussian curves, and then applied GLM analysis. Some deviation from normal, however, remained; and some of this could be explained by the utterance-initial accelerando and pre-final segment lengthening. These segments could be removed from analysis as desired to remove variation due to changes at the start and ends of utterances.

The final analysis of duration variation showed that about 26.77% of normalized vowel durations depended on the vowel type, 1.48% of vowel durations depended on the speaker, and 2.57% depended on the interaction between speaker and vowel type (the last two are very small effects). The interactive component means that different speakers assign different durations to phone types. He found similar results for consonants. And, when he factored in stress, he discovered 33% of duration variance came from stress, 3.7% from the speaker, 10.1% from the vowel, and 4.4% from the interaction of stress x vowel. He also found interactions among stress and the other variables, but these not statistically significant, and neither was the interaction between speaker x vowel (Pfitzinger, 2002b). However, a similar experiment (Pfitzinger, 1998) showed that the interaction between speaker x vowel was statistically significant. The Pfitzinger papers collectively provided insight for the analytic segment of my work.

How might a listener interpret changes in segment duration or speaking rate? Segment duration has a significant effect on prominence perception. Kochanski, Grabe, Coleman, and Rosner investigated the impact of five factors on prominence, including perceptual loudness, phone duration, aperiodicity, spectral slope, and F0. Previous work assumed that F0 was strongly correlated with prominence, but this study found only a weak correlation. Instead, they found prominence to have the strongest correlation between loudness and segment duration. Between these two factors, loudness played a stronger role; but loudness and duration were correlated strongly enough that separating the effects of the two might not be reasonable to do. They considered segment duration (D(t)) to be the instantaneous duration of the current phone over the time series. To find phone boundaries they looked for regions with stable spectra, which were approximately coterminous with the phone boundaries. Longer regions of stability, for example, corresponded to sonorants. The analysis process for tracking phone boundaries suggests options for phone segmentation (without the use of a forced aligner) and prominence detection for my future work (Kochanski et al., 2005).

Cole, Mo, and Hasegawa-Johnson (Cole, et al., 2010) found that perceived prominence is strongly correlated with acoustic measures from the stressed vowels, especially the duration of the vowel, and is therefore, at least partly "signal-driven." They also found that perceived prominence correlates with the word frequency and repetition in discourse, and is, therefore, also "expectation-driven." The categories of word frequency and perceived prominence, however, do not share the same acoustic correlates. Low-frequency words have increased high-frequency spectral content, and prominent words have increased duration. This difference may occur because the listener perceives words as prominent when something in the speech attracts attention and makes it stand out from other words in the utterance. If the word has acoustic properties which cause it to stand out (e.g., louder, longer, or greater effort level), it attracts the listener's attention, and may be perceived as prominent. Likewise, if the word or idea has not been introduced into the discourse yet, the human has to expend more effort and resources to process it. More effort may imply increased attention, which then can increase perceived prominence. Both concepts, therefore, appear to be linked by attention; and the speakers themselves are also linked by attention, both to produce and to interpret the utterance.

Not only does segment duration contribute to the perception of emphasis, but it also provides cues for phrase boundaries. Wightman (Wightman et al., 1992) examined this in detail, and found that segment lengthening, when used as a phrase boundary cue, occurs only in the syllable preceding the phrase boundary, and that humans can perceive at least four different levels of boundaries based on this segment lengthening. Interestingly, the study also found other perceptual levels of boundaries which were *not* distinguishable by segment lengthening, but by some other prosodic cue. Klatt (Klatt, 1979) showed that segment lengthening occurs before boundaries, even if pauses are missing. Many researchers have found that when pauses are present, however, the length of the pauses can help encode the boundary levels (Fant and Krukenberg, 1996).

Pitch and Loudness: How does pitch function, and what might listeners be hearing and describing as a result of prosodic pitch changes (if they aren't commenting on pitch directly)? Research suggests that pitch variation and emotion perception are linked. Specifically, the F0 mean and range vary according to the degree of emotional activation in acted speech, but the pitch contour does not change shape according to emotion. F0 tends to be higher, and the F0 range tends to be wider for high arousal than for low arousal (Banziger and Scherer, 2005). In addition, the seven basic emotions each have a typical mean and F0 range; but by itself, the F0 mean and range does not differentiate across all of the emotions very well. Surprise, fear, and anger have the clearest differentiation here (Pell et al., 2009).

Some researchers have observed that pitch changes also help denote the "given-ness" of an item in discourse. Items which are new to the discourse are indicated by pitch accent, and tend to have higher pitch because of the pitch accent (Pierrehumbert and Hirschberg, 1990). Others noticed that these pitch changes (within the context of pitch accents) provided an important disambiguation function (Shafer et al., 2000). Still others have observed that pitch inflections denote the end of a phrase. Typically, in declarative statements, the pitch falls, and in questions, it rises, but not always (Gordon, 2014). The perception of overall prominence can include duration, pitch, and loudness together. However, Kochanski found that loudness has a greater influence on the perception of prominence than pitch.

Pitch tracking methods themselves fall into three categories: time domain (e.g., zero crossing), autocorrelation (Schroeder and Atal, 1962), maximum likelihood, adaptive filter methods), frequency domain (e.g., harmonic product spectrum, cepstral, and maximum likelihood methods), and human hearing models, which involve modeling the cochlea. Each of these methods have advantages and disadvantages, usually involving overall accuracy, accuracy within a specific band, and cost (Gerhard, 2003).

2.1.2 Analysis of Vocal Quality (VQ) and Nonverbal Quality (NQ) Features

The National Center for Voice and Speech defines voice quality as a combination of vocal tract configuration, vocal tract anatomy, and the application of learned voice production techniques (National Center for Voice and Speech Tutorial) and presents a list of voice qualities with a corresponding mapping to human perception and to the physiology of production. They acknowledge that perceived voice qualities are currently not described very well, and that researchers to not agree universally on the definitions of various voice qualities.

Prior work exists in the automated detection of whispered, breathy, creaky, and to a much lesser extent, resonant voice. Until this research, none of it 1) addressed the range of phonation and effort across whispered, breathy, creaky, modal, and resonant speech, 2) examined the transitions across the continuum of effort levels from whispered through resonant voice, and 3) compared differences in effort level detection between males and females. A majority of prior work focuses on male voices. This research begins to fill in these gaps.

Whispered voice, which we investigate in this work, is distinct from voiced speech. Previous work examining the difference between voiced and unvoiced speech has found that normalized autocorrelation in the F0 range produces a strong maximum at the fundamental period, and components at regular intervals, which are both lacking in whispered speech (Atal, 1962). Whispered speech is noise-like and aperiodic in comparison to voiced speech, and measures of spectral entropy in various bands reflect this difference. Entropy ratios, particularly ratios of high to low frequency spectral entropy (e.g., 2800-3000 vs 450-650 Hz), show distinguishing voicing-dependent differences; while the use of MFCC features, standard for speech processing, yields reduced voice correlation when compared with spectral entropy and spectral tilt (Zhang 2012). Other measures which can reveal the aperiodicity of whispered speech and the spectral tilt differences include the first and second reflection coefficients (RC1 and RC2) and noncausal pitch prediction gain (Campbell and Tremain, 1986). Reduced spectral tilt is a frequent observation in unvoiced speech (Campbell and Tremain, 1986; Lim 2010), along with shifts in formant frequencies (Kallail and Emanuel, 1984), differences in the ratios of high-frequency to low-frequency energy (which captures tilt), and zero crossing rate (ZCR). The glottal component in the voice is useful, too. The residual signal, extracted via LPC analysis, models the glottal excitation, and its maximum autocorrelation is smaller for unvoiced speech than for voiced speech (Carlin et al., 2006; and Morris 2003).

Previous work has also addressed breathy vs. modal voice, and found that the difference between the first two harmonics (H1-H2), the difference between the first formant and the first harmonic (H1–A1), and the difference between the third formant and the first harmonic (H1-A3) may provide separation between breathy and modal vowels (Helen Hanson, 1995; Wayland and Jongman, 2003). The H1-H2 cue was stronger than the other cues in a study of clear vs. breathy vowels in the Khmer dialect, but the authors also say that the contrast may be between a tense vs. lax voice, and not a breathy vs. modal voice (Wayland and Jongman, 2003). They also observed that the H1-H2 difference between the breathy and modal voice was measurable within speaker but not across all speakers; the H1-H2 value for one speaker's breathiness could be the value for another speaker's modal speech. This finding raises questions about the un-normalized application of these kinds of features across a set of voices with significant variance across speakers, as well as questions about whether the relationships hold in other languages, particularly English. Other studies found that pitch and amplitude perturbations are higher for breathy voices in comparison to modal voices, and that glottal excitation features (abruptness of glottal closure, glottal pulse width and skewness, and the turbulent noise component) distinguish breathy and modal voices (Childers and Lee, 1991).

Studies comparing resonant with modal voice production suggest that speakers produce a resonant tone via "first formant alignment," which produces a higher harmonic component in the portion of the spectrum corresponding to the first formant (4-7 dB stronger). Also, resonant voice has stronger harmonics in the 2.0-3.5 kHz band (Smith et al., 2005). Actors work very hard to learn to produce resonant voice. Researchers studying the difference between actors' non-resonant and resonant voices (via the Lessac Y-Buzz technique) find a reduction in the difference between the first formant and second harmonic in men (Barrichelo-Lindstrom and Behlau, 2007).

Research which examines differences in phonation types (breathy/modal/pressed) uses features characterizing glottal function (Bozkurt et al., 2004; Gowda and Kurimo, 2013), and finds low-frequency spectral density (LFSD) to reflect the differences in open quotient and the corresponding increase in low frequency energy in breathy voices (Gowda and Kurmo, 2013). Amplitude quotient (AQ) and normalized amplitude quotient (NAQ) of the glottal pulse are superior separators, along with harmonic difference H1-H2 (Gowda and Kurimo, 2013; Airas and Alku, 2007; Kane and Gobi, 2011), closing quotient, quasi open quotient, and brightness (Airas and Alku, 2007).

Formants, especially the first two formants F1 and F2, are typically used to distinguish one vowel from another, because typical formant frequencies vary according to the vowel being spoken. The frequency of the first formant F1 is dependent on the height of the tongue body, and the tongue height has an inverse relationship with F1. For American English-speaking speaking males, F1 in modal speech typically ranges from around 300 Hz to around 800 Hz. For females, the F1 range for modal speech is around 400 Hz to around 900 Hz (Peterson and Barney, 1952; Hillenbrand et al., 1995). The second formant F2 is dependent on the frontness or backness of the tongue body. Front vowels have a higher F2 than back vowels. Typical ranges of F2 for modal speech for males and females respectively are about 900 Hz – 2300 Hz and 1000 Hz – 2800 Hz (Peterson and Barney, 1952; Hillenbrand et al., 1995). Formant differences may correlate with differences in phonation type as well. As an example, when two similar vowels are spoken by the same person in breathy voice vs. modal voice, the modal voice sample has been shown to have a stronger F1 than the breathy sample (Ladefoged, 2005). Some types of creaky voice have more narrow formant bandwidths than modal speech (Keating, Garellek, and Kreiman, 2015). Furthermore, singers achieve a more resonant voice by aligning formants slightly above a harmonic (Titze 2003). This research further explores the relationship between phonation types and formant strength and leverages these differences in the selection of classifier features.

Prior work in the detection of creaky voice has explored many features as potential markers for its presence. The difference in amplitude of the first two harmonics (H2-H1) is mentioned in multiple sources (Drugman et al., 2014; Gobl et al., 2004; Yoon et al. 2008). Multiple sources also mention difference between the first harmonic amplitude and one or more formant amplitudes, H1-A1, H1-A2, and H1-A3 and a lowered F0 (Drugman, et al., 2014; Gobl et al., 2004; Yoon et al. 2008). Overall spectral slope is another commonly-cited feature (Gordon and Ladefoged, 2001; Yoon et al., 2007). One of Drugman's methods noted the secondary peaks in the LP residual signal, passed it through a resonator (expecting more harmonics because of the secondary peaks), and measured the resulting H2-H1 (Drugman, et al. 1 and 2, 2012; Drugman et al., 2014). Researchers frequently noticed irregularity of the glottal pulses and leveraged numerous features to capture and characterize this irregularity, including jitter, shimmer, and examination of the autocorrelation

(AC) result itself. Ishi's work yielded multiple features based on the examination of the first two peaks of the AC function, including Peak magnitude, Peak position, Peak width ratio, Maximum peak magnitude, Maximum peak position, and Maximum peak width (Ishi, 2004). Yoon used the mean AC ratio, along with jitter and shimmer (2007). Drugman used the residual peak prominence (Peak-Prom) as a stand-in for other spectral and periodicity measurements, along with inter-pulse similarity (IPS) and intra-frame periodicity (IFP).

Several researchers have noted multiple spectral patterns associated with perceived creaky voice, and some of them have noticed and acknowledged that some features and resulting methods work better on some kinds of creaky voice than others. (Keating, et al. 2015; Drugman et al., 2014; Ishi, 2004; Ishi 2008). The research described in this document extends prior work in this area and confirms at least four different spectral patterns, all perceived as creaky voice, which were present in the unscripted speech of oral history interviews. It also explores distinguishing creaky voice from both modal and nonmodal phonation types, when most of the current work is focused on the distinction between creaky and modal phonation.

Previous studies of voice quality are often motivated by considerations of speech pathology (Gerratt and Kreiman, 2001; Hillenbrand and Houde, 1996), phonology (Gowda and Kurimo, 2013), or speaker identity in speech synthesis (Hanson, 1995); and therefore, no previous study considers a continuum of expressive speech that includes within-speaker and across-speaker distinctions among whispered, breathy, modal, and resonant voice qualities. There are significant, practical difficulties in the analysis of real-world expressive, acted speech. First, acted speech is characterized by greater than usual variation both within speaker and across speakers. In comparison to spontaneous or read speech, it has exaggerated extremes of pitch, volume, speaking rate, phoneme duration, phrasing, and vocal quality. Second, production of quality acted speech requires expertise. Existing corpora do not contain representative examples of expressive, acted speech; and it is not reasonable to create a suitable corpus from untrained voices.

Laughter detection is a frequently-perceived quality in this work, and research has explored its function, modeling, and detection. Some types of laughter (but not all) have clear, rhythmic pulses. Laughter with clearly-defined pulses is similar to syllables, and therefore, some prior research has used "pseudo-syllable" features for detection (An, Brizan, and Rosenberg, 2013; Oh, Cho, and Slaney 2013). Specifically, An's research uses AuToBI (Rosenberg, 2010) for extracting these syllable-like regions (frames) and for extracting features, including normalized intensity, normalized pitch, mean spectral tilt, pulse durations, silence durations, and their deltas. Oh, Cho, and Slaney's work proposes that modal speech and laughter "syllables" differ in intensity contour, pitch contour, timbral contour, and rhythmic patterns. They focused on the maximums, minimums, means, ranges, selective slopes within frames, and deltas of the results. Flux was used to represent timbre, and the frequency of intensity pulses to represent rhythmic patterns.

Other approaches in laughter detection use a simple feature set (e.g. mel-filterbank and pitch features), optimized deep learning techniques, and background models which contain many examples of unscripted, interactive laughter (Kaushik et al., 2015). Much of the prior work uses a straightforward feature set which includes some combination of MFCCs, PLP features, pitch, intensity, formants, long-term averaged spectrum (LTAS) features (spectral profiles within selected frequency bands), jitter/shimmer, and the deltas, delta-deltas, means, and standard deviations of these. They also use a variety of machine learning models, most commonly SVM, GMM, HMM, Neural Networks, Decision Trees, and CNNs (Kaushik et al., 2015; Kennedy and Ellis, 2004; Knox and Mirghafori, 2007; Krikke and Truong, 2013; Neuberger and Beke, 2013). Feature and machine learning selections seem to depend on the application, and the specific corpora under investigation.

While some of these prior investigations acknowledge different kinds and functions of laughter, the majority of prior work does not attempt to distinguish among the different kinds of laughter, perceived or acoustic. This work extends prior work by addressing multiple modalities of laughter, and laughter models.

2.1.3 Analysis of Emotion

Work in emotion detection is often limited to acted speech, which has been shown to differ from unprompted speech (Erickson et al., 2004). Some prior work focuses on categorical detection of one or more basic emotions identified by a given emotion theory. Many emotion theories assume that a small set of "basic" emotions exist which are rooted in human biology and psychology, and that these "basic" emotions are the building blocks of emotions which are not considered basic. Elkman's theory of basic emotions is rooted in the idea that facial expressions corresponding to anger, distrust, fear, joy, sadness, and surprise are universal (Elkman, et al. 1982). Plutchik relates adaptive biological processes to acceptance, anger, anticipation, disgust, joy, fear, sadness, and surprise (Plutchik, 1980). Tomkins cites a relationship of the density of neural firing to anger, interest, contempt, disgust, distress, fear, joy, shame and surprise (Tomkins, 1984). Mowrer notes that pain and pleasure are both unlearned responses (Mowrer, 1960). Several researchers believe that basic emotions responses are “hardwired” in human biological systems (Gray, 1982; Izard, 1971; Panksepp, 1982; Watson, 1930). Gray’s hardwired emotions include rage and terror, anxiety, and joy; while Izard’s hardwired basic emotion list included anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise. Panksepp’s list included expectancy, fear, rage, and panic; while Watson’s very early list of hardwired emotions included fear, love, and rage. In contrast, Arnold noted relationships between anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, and sadness to action tendencies (Arnold, 1960). Emotion researchers clearly do not agree on either the list of basic emotions, the descriptors used to name emotions, or the physiological basis for including these emotions in the list. Researchers also tend to emphasize one of the two following views over the other: 1) that emotions are primarily based in evolutionary physiology, or 2) that emotions are primarily based in psychology. A researcher who believes that emotions are first based in physiology might look for psychological connections to biology. In contrast, the researcher who believes that emotions are first based in the mind could then find relationships to human physiology. While agreement on a theory of basic emotions could make analysis more tractable, this disagreement in the field over basic emotions has led some researchers to question the view of basic emotions entirely, and refute the idea that basic emotion theory has a sound theoretical or empirical basis (Ortony and Turner, 1990).

Typical approaches to exploring speech and emotion are deep explorations into single emotions (such as anger or depression) (Bozkurt et al., 2014; Cummins et al., 2014; Honig et al., 2014; Polzehl et al., 2011), focus on variance of an acoustic parameter across a discrete set of emotions (Busso, Lee, and Narayanan, 2009; Gangamohan et al., 2014), or exploration into the recognition of the list of basic emotions supporting a given theory (Koolagudi, Nandy, and Rao, 2009; Lee et al., 2011). We found, however, that human listeners provide nuanced description of the emotions they hear in unscripted speech, which go far beyond the 5-7 emotions which are considered basic. Synonym reduction to basic emotions results in loss of information: it nullifies the expressive perceptual capability of the human listener, and also discards information about the relationships which emotion may have to other expressive elements in the voice, such as voice quality (VQ) or prosody. An alternative approach is n-dimensional representation of emotion along other axes, such as affect, arousal, and dominance (Eyben, Wollmer, and Schuller, 2012). This approach captures a greater range of emotional expression, but typically does not leverage the average human’s description of what they hear. Listeners, for example, will say that they hear laughter and embarrassment, or that speech is hesitant, sarcastic, and flat. They do not say that an angry speaker has high arousal, low affect, and high dominance. Our approach leverages the nuanced description of the human, and preserves the relationships among emotion, prosody, VQ, and nonverbal vocalization, which are embedded in the description. Furthermore, this approach encourages the development of software analytics which are aligned with human perception and are thus better able to support application development.

Text analysis along predefined dimensions such as affect, arousal, and dominance has also contributed to the analysis of emotion. Two main approaches to this are typically explored: 1) the degree of association of a given vocabulary word (in isolation) with the dimension of interest, or 2) the interpretation of a word’s syntactic and linguistic function within a natural language text, along with each word’s binary association with predefined dimensions. The Warriner et al. database of normalized arousal, affect, and dominance scores (Warriner et al., 2013) is an example of the first approach, and LIWC (Tausczik and Pennebaker, 2010) is an example of the second. LIWC is a more comprehensive approach, which considers words in their natural language context, and provides counts of the number of words (and overall percentages) associated with a given category within a text. In absence of a natural language text to analyze (we considered just crowdsourced keywords), and given the desire to understand the relationship strength, positive or negative, between words and affect/arousal/dominance, the first approach was a better fit to the problem at hand.

2.2 Male-Female Differences in Speech

Gender classification for speech processing motivated most of this cited work. Male and female voices require different analytic processing and modeling, and therefore, identifying the speaker gender is a crucial step in any speech processing system.

Male and female talkers have physiological differences which manifest in their speech signals, altering the processing required for expressive speech characteristics. Although the largest changes in a person’s voice occur during childhood and puberty, the voice continues to change slightly throughout a person’s adult life. A male talker’s H1-H2 will drop by about 5dB prior to age 16, and the H1-A3 will drop by about 10dB between the ages 8-39. Male voices’ F0 also drops during puberty. In comparison, a female talker’s H1-H2 doesn’t change very much. The H1-A3 drops by only about 4dB from age 8-39, but F0 changes near the age of puberty. Because of the difference in voice changes between gender, and because of the basic differences in the vocal tracts of men and women (for different reasons), adult females have higher F0, H1-H2, and H1-A3 than adult males (Shue and Iseli, 2008). Also, adult males have lower formant frequencies than females, because of the differences in vocal tract length (Peterson and Barney, 1952); therefore, Hanson and Chuang recommended correcting for vocal tract resonances. They also pointed out that spectral tilt affected perception of voice quality, and suggested that tilt (summarized by the H1-A3 measure) could be a particularly strong influence on perception of gender (Hanson and Chuang, 1999).

Gender classification motivated much of the prior work, beginning with gender differences in voice intensity. In an example voice intensity method, a polynomial of degree 3 was fitted through signal amplitude peaks (determined using 20 msec frames), and scaled for a better fit. Then, Simpson's rule was used to calculate the area under the peak-fitted curve. The result of this calculation, when compared to a threshold, determined gender. Differences in intensity yielded 96-98% accuracy in gender detection (Alsulaiman et al., 2011; Alsulaiman et al., 2012). Other studies show that it is possible to group age, gender analysis, and regional accent classification via a feature set of zero crossing rate (ZCR), RMS energy, F0, harmonics-to-noise ratio (HNR), and Mel Frequency Cepstral Coefficients (MFCC) 1-2, along with vector quantization, GMM modeling, and SVM techniques. This approach yielded 97-98% accuracy (Nguyen, 2010).

2.3 Grounding in Perception

This body of work considers what people hear, employs methods of qualitative and quantitative measurement of human perception, and seeks to enable discovery of acoustic correlates for frequently-perceived elements of expressive speech. Its primary sources are human computer interaction research and psychology.

Adequate models of paralingual, acted speech (i.e., the components of speech which are nonphonemic, or are not words) require a better understanding of perceived vocal quality. The basic question, "What do listeners hear?" should be addressed. Then, ideally, paralingual models should be expressed in these qualitative terms matching what typical people hear. Finally, to complete the model, the relationships between what people hear (qualitative human descriptors) and what can be measured (quantitative acoustic features) should be explored and discovered. This approach better connects acoustic analysis with what the general population hears, and avoids basing work on the perception of a single researcher or the capabilities of a single existing analytic library. Assuming that a single person's perception is the norm is the fallacy of generalization (Damer 2009). A goal is discovery of mappings between human-perceived qualities and measurable acoustic features. Thinking this way could ultimately influence what is considered baseline vocal quality analysis for expressive speech.

The approach grounded in perception also better supports application development. For example, an application which searches speech for paralingual expression could be designed to function in terms which the user would know and understand. Typical users do not use voice quality terms from speech processing such as "jitter" and "shimmer," and may not know what these qualities are, or even be able to hear these qualities or distinguish between them. They would be more likely to hear and express a perceptual term such as "rough," (Macmillan Dictionary; Tumblr Writing Helpers Blog) or express the quality in emotional terms, which then might have a relationship to jitter, shimmer, or other measurable qualities in the voice. Figure 1 summarizes the desired relationships among what people hear, what can be measured in the voice, and how systems using these features should respond. Note that this approach leverages standard human-computer interaction practices, particularly those involving intuitive interactions and the use of perceptible information (Dix et al., 2004).

User study methods can be augmented with the power of human-in-the-loop computing, or specifically, crowdsourcing. These techniques help reach a range of subjects from diverse geographical regions, with diverse demographic backgrounds; this diversity is difficult to achieve with in-lab studies using only subjects from a local or campus community. Furthermore, crowdsourcing enables these studies at scales which would also be difficult to achieve in the typical in-lab study, for reasons of both cost and time (Amazon Mechanical Turk; Crowdflower; Ipeirotis, 2010).

Emotion intensifies perception, in that people remember emotional experiences better than non-emotional ones. In other words, paralingual speech is perceived more acutely when it stands out from its context in some way (Lewis and Critchley, 2003; Erk, et al., 2003). Paralingual speech also interacts with the spoken text, functions as a powerful disambiguation function, encodes emotions, and helps transmit special discourse techniques such as humor and sarcasm (Shafer et al., 2000; Forsell 2007; Mehrabian 1971). When the text and paralingual channels conflict, research shows that the paralingual elements typically win the conflict in the perception of the listener (Mehrabian 1971).

2.4 Oral History Datasets and Practices

This body of work collects spoken accounts of historical events, contexts, perspectives, and times, usually from original sources who provide first-person accounts. The primary sources in this space come from human computer interaction, history, digital archival. The HCI perspective is primarily concerned with research methods, while the historians are primarily interested in collecting and preserving facts. Digital archivists are motivated with preserving the data and enabling or simplifying future access via search and other methods.

Oral histories are a special type of qualitative interview. They involve a mutual process of discovery between an interviewee and a researcher, and they cover a person's experiences, memories of events, opinions and viewpoints, and attitudes and beliefs. The uncovering of truth puts the two parties in a collaborative relationship, where authority is shared. The researcher is not the "knowing" or "controlling" party during data collection, and the interviewee functions as a witness offering their observations and opinions. A person's thoughts, opinions, ideas, and stories are important here, and may differ from another person's perspective on the same topic or event (that is expected). These interviews are also less structured than the related in-depth interview, which favors a fixed, focused, depth-favoring format.

People conduct oral history interviews for many reasons, which commonly include 1) adding to recorded historical knowledge, 2) surveying or understanding a person's individual (subjective) experience, particularly with respect to a historical event, historical period (particularly one of social change), or current event 3) understanding the link between a person's individual story and culture, or 4) studying the shared experiences within a community (Leavy, 2011; Thompson, 2000).

With digitization, oral history practices are shifting to better meet the needs of users who want to find and access the resources online. Metadata is key to finding and accessing information, and the oral history community seeks best practices (Maze, 2006). The Audio Engineering Society suggests standards for both audio preservation and for core metadata (AES57-2011; AES60-2011), and the Dublin Core suggests a common set of core metadata for most resources (Dublin Core Metadata Initiative, 2011). While these practices may be standard for curator and archivist, however, do they serve the needs of the end user who wants to find and access the resource? I suspect that either these standards are missing user needs, or the state of the art tools have not provided the capabilities within the standards, or the state of the art tools are lacking in usability, or all the above. The Oral History Association's website has a section devoted to Oral History in the Digital Age (Oral History Online), another section devoted to best practices for Oral History (OHDA Essay Collection). Furthermore, a recent survey (Cohen, et al., 2012) suggests the need for updated best practices for digitized resources. Standards and practices appear to be in flux.

Some tools for exploring oral histories are beginning to appear, such as OHMS (Boyd, 2012), which support text following (text scrolls in sync with the playing of the interview), and search/access via text. While these capabilities greatly help, they require transcriptions which run about $200 per interview hour; and that is unaffordable at scale. ASR is not yet accurate enough to do the job. Indexing is a lower-cost alternative, that allows a curator to mark points of interest within the interview, and summarize or transcribe these specific points. Ding (Ding, 2012) provides another example of summarization. The proprietary HistoryMaker's (History Maker's Collection; Christel and Frisch, 2008; Christel et al., 2006; Christel et al., 2010; Christel, 2007; Christel and Yan, 2007) system has similar text following, search, summarization, faceted search, and indexing functions; but it is behind a pay wall and not generally accessible. Again, transcriptions, along with tagging and summarization, are bottlenecks to access.

Some of the best quality digital oral history collections include 1) the HistoryMakers, 2) the Library of Congress (LOC) civil rights collection (Library of Congress Civil Rights Collection), 3) the LOC veteran's history collection (Library of Congress Veterans History Project), and the StoryCorps collection (StoryCorps collection). Of these, the HistoryMakers collection has consistently the highest quality recordings, full transcriptions, segmentation and summarization, with text following and keyword search. Furthermore, each speaker has a rich profile, which offers opportunities for exploring relationships among vocal patterns, profession, background, and other demographic information. Each of these corpora clearly have advantages and disadvantages, which will be discussed further in Chapter 4.

2.5 Dimensional Analysis of Expressive Speech

This body of work is concerned with finding repeating patterns of expressivity in a corpus. Typically, these are expressed in terms of either perceived features or measured signal features which co-occur again and again. The primary contributors to this body of prior work include speech processing, psychology, and natural language processing (NLP). The speech processing field typically explores signal features and their co-occurrences; and the psychology and NLP fields typically explore language text.

Work in VQ, NQ, prosody, and emotion tends to examine qualities individually, such as whispering, creakiness, resonance, laughter, or depression (An, et al., 2014; Bozkurt et al., 2014; Gowda and Kurimo, 2013; Hillendbrand and Houde, 1996; Ishi, et al., 2008; Knox and Mirghafori, 2007; Smith, et al., 2005; Wayland and Jongman, 2003; Zhang, 2012). Alternatively, it focuses on specific acoustic measurements such as jitter and shimmer and observes their variance across different voice qualities, prosodic elements, or emotions of interest. A smaller set of research examines relationships among perceived emotion, prosody, VQ, and NQ in corpora (Cullen, et al., 2013; Gobl and Chasaide, 2003; Scherer et al., 2013). These repeating relationships, or co-occurrences, among various elements of prosody, emotion, etc., could be interpreted as dimensions. Other research focuses on the relationships within acoustic measurements and applies factor analysis methods to discover dimensions within an acoustic feature set (Moriyama et al., 1997; Song et al., 2015; Song et al., 2016). Typically, this dimensional analysis is done to improve emotion recognition in speech and for dimensionality reduction (which emphasizes the important information, and discards less important information). In these cases, the meaning of the discovered dimensions is not known. Principal components analysis (PCA) (Abdi, 2010), independent component analysis (ICA) (Comon, 1994), and non-negative matrix factorization (NMF) (Lee and Seung, 2000) are common factoring techniques in signal processing.

In the text realm, latent semantic analysis (LSA) has been used for similar purposes (Landauer et al., 1996; Foltz, 1996). It is typically used to support search, and is useful for discerning word (synonym) and document similarity by considering the contexts in which each word appears with other words. Our work extends these approaches by first exploring what people hear with respect to vocal expression in a corpus, and then uses the text from human description to reveal patterns of expressivity across a corpus via Latent Semantic Analysis (LSA). Typically, the meaning of the factors from an LSA process is not known, but the most positively and negatively-associated keywords can be interpreted as indicating the meaning of each hidden dimension. By examining co-occurring emotion, prosodic, VQ, NQ, and personal quality keywords within dimensions, and by finding strong positively and negatively-correlated descriptors, we can discover relationships among these components of vocal expression. This fusion of technique forms the basis for the discovery of emotion-VQ relationships described in Chapter 7.

2.6 Summary

This chapter first provided an overview of the prior work related to the analysis and modeling of speech. Specifically, it covered methods for analyzing prosody, voice quality, nonverbal quality, and emotion in speech, which are leveraged and extended in this work. Next, it covered male-female differences in speech which were relevant to the research questions explored in this thesis, especially for human perception, analytic methods, and gender difference analysis. Then, it discussed reasons and methods for grounding analytics in human perception. The primary reason for grounding is to achieve alignment of the resulting software with human perception to better support application development. For example, in an acoustic search engine application, people would be more likely to search for expressive characteristics they know, remember, and can articulate, such as "roughness," than characteristics they do not know or name name, such as "jitter". Next it addressed oral history data sets and the current state of the art in curating and archiving them, since this work utilizes oral histories for unscripted speech analysis. Finally, a summary of dimensional analysis techniques are presented which aided in the development of a new approach for discovering expressive dimensions present in a corpus.

The next chapter connects the overview of prior work given here with an overview of the (new) investigative process in this thesis. It flows through the following stages: 1) ***corpora selection*** (leverages section 2.4 of prior work) 1) ***perception studies*** (leverages sections 2.1, 2.2, 2.3, 2.4 of prior work), 2) ***interactive analysis of perceived expressive features*** (leverages sections 2.1 and 2.2 of prior work), 3) ***modeling distinct, frequently-perceived features*** (leverages sections 2.1 and 2.2 of prior work), 4) ***discovery of expressive dimensions*** (leverages section 2.5 or prior work), 5) ***modeling discovered dimensions*** (leverages section 2.5 of prior work), and 6) ***mapping the organically-discovered dimensions onto predefined dimensions*** (here, the axes of affect, arousal, and dominance, which leverage sections 2.1.3 and 2.5 of prior work).

CHAPTER 3: EXPERIMENTAL METHODS OVERVIEW

This chapter gives an overview of the research process from beginning to end. It is intended to be a “roadmap” or “outline,” which provides the big picture without the burden of too many details. Note that the details will be discussed in later chapters. This chapter also highlights the interdisciplinary nature of the work. It presents the investigative process which enabled this interdisciplinary exploration across human-computer interfaces, speech and language processing, linguistics, and digital curation. This process is one of the contributions of this research.

The purpose of this process is to allow human perception to drive the investigation of vocal expression. If this is done correctly, the resulting machine models will better align with human perception, and will therefore be better positioned to fit the needs of application development discussed in the introduction. In addition, this process encourages the discovery of mappings between the realm of qualitative human description of vocal expression (that which humans can perceive and describe), and the domain of quantifiable, computable acoustic features (that which can be measured and extracted from signals). It provides appropriate and different methods for exploring small numbers of frequently-articulated perceived features (such as prosodic and selected voice qualities), and large numbers of diverse, nuanced features (such as emotion and personal quality). The method is also extensible to other modalities of human expression, including multimodal expression, and scalable across multiple levels of detail.

The process steps include 1) corpora selection and curation to collect representative expressive speech samples suitable for analysis, 2) perception studies to learn what people hear in expressive speech, 3) interactive analysis of perceived features to understand the relationships between perceived features and measurable acoustic features, 4) modeling distinct, frequently-perceived features and validating the models, 5) organic discovery of expressive dimensions, 6) modeling organically-discovered dimensions and validating the models, and 7) exploring mappings of organically-discovered dimensions to predefined dimensions. Each of these steps are described at a high level in this chapter and in detail in later chapters. Note that a portion of this information has been published as part of our prior work (Pietrowicz et al. 1, 2017; Pietrowicz et al. 2, 2017).

3.1 Corpora Selection

Suitable corpora had to be identified and curated. The research questions, corpora content, cost, accessibility, and privacy/IRB concerns drove corpora selection. Development of a new corpus was out of scope for this work. This process (and the resulting corpora) is described in detail in Chapter 4. Shakespearian scripted speech from movies and the stage, and selections from the Library of Congress Veterans' Oral History Collections (semi-structured, unscripted speech) were curated for this research.

3.2 Perception Studies

The process began with user studies to understand what everyday listeners heard in expressive speech. In this step, Mechanical Turk workers were asked to describe the vocal expression in three separate studies, including Shakespearian soliloquy (scripted speech), oral history interviews (unscripted, semi-structured, conversational speech), and laughter in the context of oral history interviews (a specialized dimension of unscripted speech). Each Mechanical Turk worker was presented with an audio clip, one per task, and asked to provide a minimum of 1-3 words describing what they heard in the vocal expression, not the word content. A minimum of 10 workers evaluated each single clip. For the convenience of relating this work to prior work, and to highlight potential relationships among groups of keywords (for example, between emotion and voice quality), the data analysis included some clustering of the resulting descriptive terms into groups from prior literature (specifically, voice quality, prosody, emotion, conversation quality, and personal quality). Descriptors were also "synonym-reduced," that is, clustered into groups of close synonyms, as defined by Thesaurus (Online Thesaurus of English), and tagged by the most frequent term in the cluster. Note that the clustering and data analysis was the work of researchers, not the Mechanical Turk workers. Also note that the purpose of this step was understanding what people heard, not labeling audio clips for training machine models; this is an important distinction. Chapter 5 describes the detailed process, analysis, and results for this step.

3.3 Interactive Analysis of Perceived Expressive Features

Frequently-perceived features were selected for in-depth investigation. It is important to note that the items selected for exploration here (e.g., breathiness, whispering, resonance, creakiness, etc.) were taken directly from the descriptors provided by the listeners in the user studies, in those terms exactly, and their close synonyms. Chapters 5 and 6 provide the detailed rationale for selected specific features for investigation. A quality was selected for detailed investigation if it was perceived at a disproportionately high rate across all the speakers in the study, and at least once in each speaker. It is interesting to observe that the terms which the users heard at disproportionately high rates, such as “breathiness,” “whispering,” “resonance,” and “creaky,” have been grouped and described in prior literature as “effort levels,” “phonation types,” and less frequently “voice qualities.”

Next, multiple expert listeners coded voice samples with the labels given by the listeners such that classifiers could be trained to recognize the selected qualities. The listeners were in high agreement with kappa values all above 0.75. Then, acoustic analysis was conducted to discover acoustic features useful for recognizing and distinguishing each of these features. The goal was to be able to distinguish each of these qualities from each other, as well as from conversational, “modal” speech. This step mapped the qualitative features which listeners reported hearing with acoustic features which could be measured quantitatively. Then, each of the acoustic features identified were evaluated with respect to their ability to distinguish each perceived feature. Chapter 6 describes this process in detail, and the results.

3.4 Modeling Distinct, Frequently-Perceived Features

Based on the results of interactive analysis, classifiers were trained using groups of acoustic features identified in the prior step and cross-validated. Both binary and n-way classifiers were explored, where binary classifiers separated a single perceived feature from both modal speech and the other non-modal categories under investigation, and n-way classifiers classified speech samples into n categories. Groups of the strongest separators were selected, used to train machine learning models, and then cross validated. Depending on the precision, recall, and overall accuracy of each resulting model, the feature sets were selectively tuned to optimize performance of the resulting model. About 40 different groups of models were examined for both males and females, in both scripted and unscripted speech. Chapter 6 describes this process and the results in more detail.

3.5 Discovery of Expressive Dimensions

The process described up to this point works well for small numbers of concise, frequently-occurring prosodic and voice quality features. Listeners did not perceive or describe emotion and personal qualities in concise terms, however. They used nuance, especially, to describe emotion. Happiness was different from joy, peace, humor, contentment, laughter, ecstasy, and the many other subtle labels listeners used. These distinctions contained information, and were used in different expressive contexts. Trying to reduce all the related terms to one single term (or basic emotion) such as “happy” so that the prior technique could be applied would result in the loss of information and the richness of expressive nuance. Instead, we used the Latent Semantic Analysis (LSA) technique for the organic discovery of dimensions of expression present in each corpus we explored. LSA both preserved and leveraged listener nuance, but allowed abstraction via contextual similarity of nuanced terms. It also enabled analysis of relationships among emotion, prosody, voice quality, conversation quality, and personal quality, which supported our research questions. Chapter 7 describes this process, including the discovery of dimensions, extraction of the descriptions of each dimension, and the exploration of the relationships between emotion and voice quality present in each of the scripted, unscripted, and laughter corpora.

3.6 Modeling Discovered Dimensions

When the dimensions present in a corpus had been discovered, the LSA technique was used again to discover the audio clips which were most strongly representative of each dimension. These sets of clips would be used to train models to recognize expression from each of the top approximately 20 dimensions discovered. Clips were not necessarily strong representations of a single dimension; and, some clips were not strongly representative of any of the top dimensions. However, simple separation of the clips into two groups - those which were strongly representative of a dimension and those which were not – was sufficient for training the models. The descriptors given in each dimension, prior literature, and results described in 7.3 guided feature selection for the models, and a similar process of iterative cross validation and model tuning described in 7.4 guided evaluation of model performance. Chapter 7 describes the process of dimensional modeling, and the results.

3.7 Mapping Organically-Discovered Dimensions to Predefined Dimensions

Much work has been done in the realm of investigating emotion along predefined axes. It may be possible to leverage this prior work, and contribute to it, by understanding the relationship between organically-discovered dimensions and these predefined dimensions, specifically affect, arousal, and dominance. By measuring the affect, arousal, and dominance of keyword descriptors according to results of prior studies, we can profile each discovered dimension along these axes, and develop models which translate among dimensions. This discovery process and the potential of affect, arousal, and dominance to represent organically-discovered dimensions, are described in detail in Chapter 7.

3.8 Summary

This chapter presented an overview of the different phases of the research process followed here. The process began with corpora discovery and curation for the support of specific research questions, and the user studies presented representative clips to diverse listeners who described what people heard in the selected corpora. The frequently-heard, concise qualities were explored interactively, and examined for representative spectral and waveform patterns. Then, acoustic (or higher-level) features were evaluated for their ability to distinguish among the target qualities. Combinations of these features were selected and incorporated into machine models; and the models were validated and iteratively improved.

LSA was used to leverage nuanced listener description, particularly that of perceived emotion and personal quality in the voice. The LSA technique was used in the organic discovery of expressive dimensions, the mapping of clips onto these dimensions for training models, and in exploring relationships between voice quality and emotion. As before, acoustic features were evaluated for their ability to distinguish among the discovered expressive dimensions; and models were iteratively explored and improved. Finally, mappings between the organically-discovered and predefined dimensions were explored to evaluate the possibility of leveraging and contributing to prior work.

Future work will involve scaling up the models, and involve leveraging lower-level dimensional models (such as those for creaky voice or laughter) in the analysis of higher-level, organically-discovered dimensions. Human-in-the-loop crowdsourced systems could help with scalability via a bootstrapped system for training models on corpus subsets, using the resulting models to label other data in the corpus, and then using crowdsourcing to confirm or correct these labels. The crowdsourced corrections could be used in the iterative tuning of the resulting models.

Finally, the human perception-grounded models could be deployed and evaluated in the context of applications, particularly those in search and human health and wellness. Studying the models in the context of applications will provide important data to close the loop and further improve the models.

The next chapter focuses in on the details of dataset curation (mentioned very briefly in section 3.1 of this chapter). It discusses in detail the dataset selection criteria and process, the curation process, preprocessing, support for perception studies, support for analytics, and support for dimensional discovery. It covers the curation of both scripted and unscripted corpora, and discusses the scale of dataset.

CHAPTER 4: DATASET CURATION

This chapter describes the selection and curation of the datasets for this research, both for scripted and unscripted speech. It first gives the desired characteristics for scripted and unscripted corpora, the rationale for selection, and the reasons for not selecting other popular alternatives. Then, it gives detailed information about curating for perception studies, voice quality and nonverbal quality analysis, and expressive dimension discovery. Note that a portion of this information has been published as a part of our prior work (Pietrowicz et al., 2015; Pietrowicz et al. 1, 2017; and Pietrowicz et al. 2, 2017).

As previously mentioned, we required suitable corpora for exploring research questions concerning vocal expression of both scripted and unscripted speech. Scripted speech is speech in which the text is fixed, and the speaker either reads it or speaks it from memory. Scripted speech can exhibit a range in purpose, structure, and preparation on the part of the speaker. On one extreme, the speaker could be seeing the text for the first time, and reading it out loud, as in a speaker opening and reading a letter to friends. The speaker could also be a TV newscaster who is still reading text, but with different purpose, more preparation, and professional training which modulates and limits the expressivity of the presentation. On the other extreme, as in the stage literature, an actor has studied and memorized a script, studied the character he or she is playing, understood the context of the story, studied the character's interactions with other characters in the story, studied expert prior performances of the part, and practiced many times to the point of internalizing every nuance of the performance.

Unscripted speech, in contrast, is spontaneous, and the speaker reacts within the context as the conversation or situation unfolds; and it, too, ranges in purpose and structure. Examples of different kinds of unscripted speech include familiar discussions among friends, professional meeting discourse, teaching lectures, and interviews. Discussion among friends is an example of relatively unstructured, unplanned discourse, which is in great contrast to teaching lectures, which usually have a predefined structure, expected range of expressivity, and constrained (if any) interaction with listeners. Both of these examples contrast with interviews, which have expected participant roles, expectation for question-response discourse with turn taking, and entirely different social expectations on expressivity, depending on the type of interview. Because of the obvious differences, separate corpora were required for scripted and unscripted speech.

4.1 Selection of Scripted Speech Corpora

The characteristics of an ideal scripted corpus suitable for exploring the research questions included the following: 1) wide-ranging expression, 2) social context which allowed and encouraged expressivity, 3) expressivity in a context comparable in complexity and utterance duration to the selected unscripted speech corpus to enable some comparison between the two, 4) expert talkers adept in producing the gamut of human expression, 5) a range of diverse L1 English Speakers to enable evaluation of model performance across speakers, 6) multiple performances of the same text to enable comparison of expressive interpretation across speakers, 7) existing recordings which could be freely referenced and shared across the research community, and 8) high enough quality recordings for analysis, with critical sections free of music and noise. Use of existing recordings was strongly preferred to building a corpus from scratch, because of the costs in time and money to collect and validate suitable recordings. Developing a new corpus would require recording equipment, studio space, time to develop or select appropriate text, time to make the recordings, time to find and evaluate suitable performers and their performances, time to resolve IRB issues involving recording and sharing human subjects' voices, and funds to pay for the costs of making the recordings and the performers' time. Corpora development is a major project by itself. Clearly, recording entirely new data was out of scope for the scale of this work, and was not necessary to answer the targeted research questions. Using existing, or "found" recordings, from the internet had the added benefit of forcing development of analytics which did not require controlled studio conditions. Most of the recordings available on the internet did not come from controlled conditions. They were recorded with variable equipment in varying environments across varying levels of noise and acoustic conditions. Software analytics which could tolerate a range of conditions would be more applicable to the range of recorded sound artifacts available on the internet.

Many existing speech corpora were not ideal. Recordings of news readings (Minard et al., 2016, and Graff 20050), for example, had limited expressive range, and typically did not have multiple recordings of the same text by different speakers in the same language. Emotion corpora (Wu et al. 2006, Ververidis and Kotropoulos 2003, and Krothapalli and Koolagudi 2013) also typically supported only a small number of basic emotions, did not provide a larger context for evaluating the expressive results, did not provide suitable range in complexity and duration, and often did not use L1 English speakers. Recordings of preachers and political speakers provided good expressive range, but came with many complex social influences which were outside the scope of the research questions. Verifying whether the speech was read, memorized, or unscripted was not possible for most of these speakers. Also, the words of famous preachers and politicians are not typically performed by a range of speakers. Many existing corpora did not provide multiple talkers performing the same script. Some of the corpora also presented financial barriers to use, including expensive memberships with per-corpus access fees (Linguistic Data Consortium), and distribution restrictions.

In contrast, many recordings of expert actors performing well-known scenes from the literature were widely accessible on the internet. They were free, or at most, they were accessible for the cost of a DVD or legal download. Recordings of Shakespearean plays met the desired qualifications. These performances had wide-ranging expression. Many of the best actors performed them, so finding multiple instances and varying expressive interpretations by expert performers was not a problem. Most of the speakers followed the script exactly, and this removed the variability of the underlying text so that the expression could be studied apart from confounding factors of text variability. No two interpretations of a part were alike. The performances had a context, with utterance complexity and length comparable to that of the unscripted corpus selected for study. Finally, sharing references to the recordings along with limited segments of the recordings themselves with the research community was legal and possible with no IRB restrictions.

Because the Shakespearean soliloquy fulfilled all the major data set requirements, corpora were created of male actors all speaking a common male soliloquy, and of female actresses all speaking a common female soliloquy. The corpora featured expressive experts at their craft, in roles typical of their gender. They also enabled exploration of analytic techniques applicable to recordings which are typically available on the web (not necessarily made in controlled studio conditions). Therefore, mp4 or wav recordings of predominantly movie and some stage performances were collected because these formats were readily available. Speakers were selected for their professional acting ability, diversity of expressive speaking style (which occurred naturally), and diversity of origin (which included L1 British, American, and Australian actors). Male and female corpora were created which had the same speaking style (Shakespearean acting), and similar topic content. The Act II Scene I Hamlet Soliloquy (Shakespeare, Signet Classic 1998) ("to be or not to be…") served for the men, and featured a king contemplating suicide. The Lady

Macbeth soliloquy, from Act I Scene V of Macbeth (Shakespeare, Simon & Schuster 2013) served for the women, and featured Lady Macbeth contemplating the murder of a king.

4.2 Curation of Scripted Speech Corpora

The goal of the initial curation step was selection of specific samples of the Hamlet Act II Scene I and Macbeth Act I Scene V soliloquy, according to the criteria mentioned above. The Hamlet corpus of male actors included expert performances of the Hamlet soliloquy (Act III, Scene I) by Mel Gibson, Derek Jacobi, Richard Burton, David Tennant, and Kenneth Branagh (Hamlet Soliloquy Performances). These speakers were selected for their collective difference in expressive style across speaker and for their professional acting and speaking ability. This small number of speakers provided a large range of expression for analysis. For example, in just the first sentence of the soliloquy, Jacobi's voice ranges from breathy to resonant, soft to loud, and ranges in pitch over almost an octave. Tennant's voice is breathy, soft, and gently inflected, while Burton's voice is modal and flat in comparison. Branagh's voice is all angst, and ranges from breathy-modal in the first phrase, to a chilling whisper in the second phrase. Gibson's speech is rapid, his pauses, minimal, and tone, almost businesslike. Each speaker's pitch and volume variation, accent points, and phrasing were different, and that is just a high-level observation over just the first sentence. This range is typical of Shakespearean actors' speech. Recordings for the Macbeth corpus of female actors were selected in similar ways, and included expert performances by Judi Dench, Harriet Walter, Joanne Whalley, Kate Fleetwood, and Allison Jean White (Lady Macbeth Soliloquy Performances). Most suitable samples were found on YouTube. The next steps included general preprocessing of the recorded signal, followed by curation to support perception studies, voice quality analysis, and discovery of expressive dimensions in the corpora.

4.2.1 General Preprocessing in Scripted Corpora

To prepare the corpus for analysis, we first downsampled it to 16 kHz, normalized the signal within each speaker, and excluded portions with music, excessive noise, sound effects, interfering voices or background sound, or significant reverb (with noticeable echo or delay). We also converted it to single-channel WAV format because of the availability of analysis tools for the WAV format, and relative ease of a single channel vs. multiple channel analysis. SoX (SoX), Switch (Switch), and Wavepad (Wavepad) provided the necessary preprocessing tools, and we wrote simple scripts to do the sound processing.

4.2.2 Support for Perception Studies in Scripted Corpora

The perception studies (to be described in detail in chapter 5) asked L1 US English speakers on Mechanical Turk to listen to a clip, and provide keywords describing the expressive qualities they perceived in the voice. The purpose of the perception studies was to provide insight into the expressive vocal gestures which untrained listeners would hear, and then use this information to guide selection of consistently-perceived vocal features for detailed analysis grounded in human perception. Curation for this task meant selecting and extracting meaningful clips for the Mechanical Turk workers to describe. In order to learn what untrained listeners would hear, expressively speaking, in scripted, Shakespearean speech, the soliloquy were first segmented into hierarchical phrases and subphrases. At the highest level, the phrase boundaries were long pauses for at least one speaker. At the lower levels, the phrases were bounded by short pauses and breath groups in at least one speaker. Elan (ELAN) was a convenient tool for tiered, hierarchical corpus markup, and the segmentation boundaries were output in the TextGrid format for import into Praat (Praat). Custom Praat scripts automated the extraction of phrase clips at the marked boundaries.

Specific clips were selected for the studies with the goals of maximizing both coverage of the range of expressivity within speaker, and coverage of the range of expressivity across speakers. The same phrases were extracted across all speakers to remove text as a variable in the study, and allow focus on variability resulting from differences in vocal expression. Because the soliloquy were natural hierarchies of phrases and subphrases, and because perception might differ between longer, higher level phrases and the shorter subphrases, we curated both longer clips and their subphrases for the perception studies. For example, a higher-level phrase from the Hamlet corpus was, "To be or not to be. That is the question." This phrase had three sub-phrases, including "To be," "or not to be," and "That is the question."

4.2.3 Support for Voice Quality Analytics in Scripted Corpora

Listeners consistently reported hearing whispering, breathiness, and resonance in speech for both male and female speakers in the perception studies; therefore, we focused on these frequently-perceived voice qualities and grounded our analysis in human perception. See Chapter 6 for detailed discussion of voice quality analysis. Curation for this task meant labeling and extracting samples of whispering, breathiness, modal speech, and resonance to support analysis and modeling of these voice qualities. Note that listeners did not explicitly report hearing modal speech (i.e., standard conversational quality), but we included it in the analysis as a default, "baseline" quality to stand in contrast to the exceptional qualities of whispering, breathiness, and resonance which listeners reported. As Section 6 explains, listeners commented on expressivity which stood out from the baseline, not on the baseline itself.

Vowels vary more than consonants do across whispering, breathiness, and resonance (imagine whispered vs. voiced utterances of the word "sassy"), so we extracted all the vowel sounds in the corpora which were at least 60 msec long. We selected 60 msec for a duration lower bound because the analysis frames were 60 msec, and because shorter frames were too difficult for listeners to discern the differences in vocal quality reliably. A forced aligner (FAVE-align) was helpful in this process, but we overrode it manually when it made errors.

One expert listener hand-coded each performance (audio recording) in the corpus to the syllable level for four commonly-perceived voice quality conditions (whispered, breathy, modal, and resonant). Expert listeners for voice quality labeling tasks were defined as those who had formal instruction in speech processing, linguistics or music. Musicians in their domain are trained to hear and produce nuanced variations in tempo, dynamics, accenting, timbre, etc.; and this translated well to the specific task at hand. By our definition, whispered speech had no voicing, breathy speech had weak voicing with an airy quality, modal speech had an average voiced conversational quality, and resonant voice had a ringing, or projected quality in comparison to modal voice. To validate the coding, we randomly selected 20 samples from each condition across all the speakers, and asked a second expert to classify the samples as whispered, breathy, modal, or resonant speech. Before running the experiment, we gave our experts the definition of each type of speech, and demonstrated it with example recordings. We reached 95%, 85%, 65%, and 90% classification agreement (chance was 25%) over the whispered, breathy, modal, and resonant conditions, respectively, with 85% agreement overall, and a Cohen's kappa of 0.8. The Hamlet corpus final result included **83 whispered** (63 Branagh, 4 Burton, 8 Gibson, 4 Jacobi, 4 Tennant); **329 breathy** (86 Branagh, 13 Burton, 60 Gibson, 86 Jacobi, 86 Tennant); **353 modal** (30 Branagh, 85 Burton, 68 Gibson, 85 Jacobi, 85 Tennant); and **276 resonant** (4 Branagh, 80 Burton, 20 Gibson, 160 Jacobi, 4 Tennant) utterances (1041 total). Each actor's soliloquy contributed some of each condition, although not all voices had the same distribution of each type of condition.

A similar process was used to curate and code the Lady Macbeth corpus, resulting in inter-rater agreement of 95%, 83%, 74%, and 83% across whispered, breathy, modal, and resonant conditions, respectively (kappa = 0.83). The corpus final result included a comparable **80**

whispered (16 Dench, 41 Fleetwood, 7 Walter, 9 Whalley, 7 White), **385 breathy** (42 Dench, 77 Fleetwood, 41 Walter, 166 Whalley, 59 White), **316 modal** (20 Dench, 56 Fleetwood, 56 Walter, 99 Whalley, 122 White), and **158 resonant utterances** (21 Dench, 15 Fleetwood, 16 Walter, 43 Whalley, 63 White), (939 total). Again, each actress's soliloquy contributed some of each condition, although not all voices had the same amount of each kind of expressive speech. The overall distribution of effort levels for both males and females were similar, with females having slightly greater use of breathy voice and slightly lesser use of resonant voice than males.

4.2.4 Support for Dimensional Discovery in Scripted Corpora

Curation for this task meant supporting discovery of perceived expressive dimensions in the voice. An expressive dimension is represented by simultaneous, repeating clusters of perceived emotions, prosodic inflections, and voice qualities. For example, a dimension might include the cluster of perceived emotional frustration, high prosodic variation (particularly in pitch and speaking rate), and resonant voice quality. Another contrasting dimension might include sadness or low affect, slow speaking rate, breathy voice, lowered pitch, and quiet speech. Curation for this task meant first segmenting and extracting phrases representative of a speaker's range of vocal expression at multiple levels in the phrase hierarchy. Then, given the representative phrase hierarchy for a speaker, the next curation steps required collecting descriptive keywords from untrained listeners which described the perceived vocal expression for each curated phrase. The phrases and corresponding descriptive keywords would support a latent semantic analysis (LSA) (Landauer et al., 1998) as described in detail in Chapters 5 and 7.

4.2.5 Scripted Corpora Scale

To summarize the overall scale, the scripted corpora included 10 speakers total, 5 male and 5 female, with the males performing the Hamlet Act II Scene I Soliloquy, and the females performing the Lady Macbeth Act I Scene V Soliloquy. The male scripted corpora performances ranged from 131 seconds to 207 seconds. Each speaker's usable soliloquy length, including vocal pauses, was as follows: 1) Branagh, 164 seconds, 2) Burton, 131 seconds, 3) Gibson, 207 seconds, 4) Jacobi, 174 seconds, 5) Tennant, 167 seconds, for a total Hamlet corpus length of 843 seconds.

The female scripted corpora performances ranged from 88 seconds to 204 seconds of usable data, including vocal pauses, with each speaker's contribution as follows: 1) Dench, 113 seconds, 2) Fleetwood, 158 seconds, 3) Walter, 88 seconds, 4) Whalley, 204 seconds, and 5) White, 185 seconds. The total length of the Lady Macbeth corpus was 748 seconds. The total scripted corpus length was 1591 seconds, or just under 30 minutes of speech. Approximately 8.5% of the scripted corpus was sampled for the perception studies (see Chapter 5 for a discussion of the scripted speech perception studies). All the scripted corpus was used for training and validating voice quality models (see Chapter 6 for a discussion voice quality modeling in scripted speech).

4.2.6 Potential Biases Within the Scripted Corpora

The Hamlet and Lady Macbeth corpora have a few inherent biases. First, the speech is entirely Shakespearean, acted speech. The speakers are performing in a specific style, with expressive exaggerations not typically seen in conversational speech. The range of expression is limited to what is called for in the given soliloquy; this differs from everyday conversational speech. Beyond this, the relatively small number of speakers and small number of speech samples could allow patterns which repeat within one speaker to be over-emphasized in analysis. Sampling across the range of speaker expressivity will help guard against single-speaker bias. The speakers did have individual biases (for example, one speaker might use more resonant speech, overall, and another, more breathy speech overall). No speaker, however, stayed within a single expressive style across the entire soliloquy. When possible, using larger numbers of speakers, sampling across the range of expressivity of each speaker, and including a larger number of speech samples will help avoid single-speaker bias.

4.3 The Selection of Unscripted Speech Corpora

The remaining sections of this chapter describe curation of our unscripted corpora. The ideal unscripted corpus included the following characteristics: 1) wide-ranging expression, 2) social context which allowed and encouraged expressivity, 3) expressivity in a context comparable in complexity and utterance duration to the selected scripted speech corpora, 4) talkers with no noticeable speech pathologies, with no special training expected, 5) a range of diverse L1 English speakers to enable modeling performance across speakers, 6) a common context which placed the same constraints and structure on the discourse or flow across each recording , 7) a common topic and purpose across all speakers (since we do not fix the text, we constrain the topic to remove that variable), 8) high enough quality recordings for analysis, with critical sections free of music and noise, 9) common speaker roles across the corpora, 10) large size to support ongoing analysis at large scale, 11) representation of both male and female voices, and 12) open access for us and for others in the research community. In general, we desired qualities similar to the scripted corpora, but with common constraints across all recordings. The text varied, by definition, across speakers in unscripted speech; therefore, we wanted constraints and structure within the corpus which would hold other factors constant to enable some comparison across speakers. As in the scripted corpora, use of existing recordings was strongly preferred to building a corpus from scratch for the same reasons discussed in scripted speech.

Again, many existing speech corpora were not ideal. Recordings of meetings often had limited expressive range because of social constraints due to the business setting, not to mention the wide variety of topics (AMI Corpus; and Carletta 2006). TED talks (TED) had many of the desired characteristics of expressive range, and some constraints placed on the discourse from the TED talk format. TED talks, however, were diverse in topic and flow, and would not allow elimination of those variables in analysis. Furthermore, these talks required significant preparation and practice. While the TED talk text was not usually fixed, some of the speakers might have memorized their talks; and none of the speakers spoke spontaneously. Interviews, in contrast, had the potential for high expressivity, with constraints on the discourse and flow (question-answer format), topic limitations, and speaker roles (interviewer/interviewee). Interviewers usually had an interview plan (but not fixed text), and interviewees spoke spontaneously. We sought a corpus of high-quality interviews, on high-quality media, with a common topic, which was easily-accessible in the public domain. StoryCorps (StoryCorps) interviews were a possibility, given enough of them focused on a specific topic. Many of their interviews, however, were not publically available for use. Furthermore, StoryCorps added background sound (e.g., music) to many of their publically available interviews, making them unusable for our purposes. Also, many of the publically-available interviews were curated for promotion of their organization and social agenda. Some other collections, such as the HistoryMakers (The History Makers Video Oral History Collection), offered high-quality, high-expressivity, topic-limited content, but placed limitations and costs for access, use, and sharing of content, even for research and teaching purposes. Many oral history interview collections were low quality or were recorded in inconvenient formats (e.g., non-digitized aging tape media), low in expressive range (e.g., unskilled, unengaged speakers), and provided barriers to access (such as having to be physically present at specific, far-ranging locations at specific times to use the media).

In contrast, the library of congress Veterans' Oral History Project (Library of Congress Veterans History Project) provided an open collection of oral history interviews which met the requirement for semi-structured, unscripted speech on the part of the interviewee. This collection contains thousands of interviews covering military activity beginning with WWI, continuing through the most recent conflicts in Iraq and Afghanistan. At this writing, almost 100,000 interviews in the collection are digitally available online, and only an estimated 10% of the collection has been digitized. Interview materials included variable combinations of written interview transcripts, photographs, other written artifacts, and audio or video recordings. Each interview lasted about 0.5-2.0 hours. While male speakers were in the majority in this corpus, it had sufficient female speakers. In addition, the structure of the interviews had similar format and many common questions across the corpus. Almost all interviews, for example, asked subjects to state their names and basic demographic information at the start of the interview; and almost all interviewees responded to these questions with neutral expression (modal voice quality, neutral emotion, and neutral prosody). Many interviewers asked why and how their subjects joined the military, and about their experiences with basic training. Most also asked subjects to relate one or more stories about their individual personal experiences. These characteristics conveniently enabled some comparison of vocal expression across answers to similar questions. The corpus was unprompted, natural, and sparse in non-neutral expression, with islands of expressive bursts. It provided a wide range of speakers exhibiting a wide range of expression. The interview/storytelling format encouraged expressivity, and most speakers freely communicated throughout the interview.

Quality of the recordings varied, and most were made in public or home environments with non-professional equipment. Collection artifacts were donated items, and quality overall depended on the personal dedication, skill, and resources of the person who created them. In general, media quality was superior for more recent artifacts because they were recorded on digital media; while older accounts were recorded on tape. Furthermore, tape recordings degrade over time, so a recording made in the 1990's on tape (and only recently digitized) would have a lower baseline quality than recent digital recordings, and would have an expected 20 years of degradation. Quality of supporting documents such as transcripts also varied in accuracy and completeness.

Interviewers ranged in skill from completely untrained high school students to professionally trained interviewers, to interviews conducted as part of government-funded research (complete with the text of the IRB on the recording). We preferred the more skilled interviewers, but even the untrained interviewers tended to ask similar questions in similar order. Furthermore, the interviewees did most of the talking. Many of them were irrepressible and uninhibited in their storytelling; and an untrained interviewer did not detract from the quality of the expressive stories of the veterans being interviewed.

The remainder of this chapter describes the curation of the unscripted speech corpus, and includes selection of interviews for analysis, general preprocessing, and support for perception studies, voice quality analytics, and dimensional discovery.

4.4 Curation of Unscripted Speech Corpora

The initial goal of unscripted speech curation was selection of specific interviews for detailed analysis, according to the criteria given above. The resulting dataset included a balanced amount of male and female speech (about 5 hours of speech for males and 5 for females). Our corpus sub-sample used recent interviews collected during the last 10 years on digital recording equipment, with all subjects representing the Iraq and Afghanistan conflicts. We preferred interviews containing video recordings for future multimodal analytic work; and we excluded from analysis speech which contained significant background interference (e.g., other voices, street noise, music, reverb, or high levels of buzz/hum/hiss). We also preferred interviews with complete and accurate transcripts. No transcript which we examined was perfect, however; and all required correction to be suitable for text alignment. Finally, we preferred speakers with individual and collective expressive diversity. The speakers represented multiple dialects, and note that dialect was not included in the analysis, due to the limited number of speakers overall and the limited number of speakers representing each dialect. The five female interviews selected from the collection included Elida Trinidad Sluss (nee Fernandez) (LOC Interview, Elida Trinidad Fernandez Sluss collection), Nicole Cabral Ferretti (LOC Interview, Nicole Cabal Ferretti collection), Amanda R. Kean (nee Fichera) (LOC Interview, Amanda R. Fichera Kean collection), Ingrid C. Lim (LOC Interview, Ingrid C. Lim Collection), and Teresa Michelle Little (LOC Interview Teresa Michelle Little collection). These speakers covered a range of military experiences, personal difficulties encountered during service, and long-term impact on their lives. Their vocal expressions reflect their experiences and emotions as they tell their stories. As an example of this experiential and expressive diversity, Elida Sluss recalls her career military experience as overwhelmingly positive, becomes emotional when describing the positive impact on her life (her voice becomes breathy and shaky, and she cries) and she speaks of the military and her colleagues with great respect (her voice softens and slows, and varies in rate to emphasize key thoughts). When she describes her personal and professional relationships with others in the military, even to being godmother to many of their children, her voice becomes animated in pitch and tempo, and becomes more resonant. Overall, her voice radiates the most passion, humor, affection, warmth, and positivity of all the selected female speakers; but she laughs the least of the selected speakers. In contrast to this speaker, Nicole Ferretti joined the military in part for financial reasons, was assigned duty as a truck driver in a mixed but primarily male unit. She had to assist with a wide variety of tasks, some of them horrific, at the various destinations. She suffers from long-term back injuries due to riding many hours in trucks with insufficient shocks, and from having to lift more weight than she was safely able to carry. She also suffers from self-reported, long-term emotional distress (possible PTSD), and had not been able to return to college at the point in time of her interview. She has had many challenges returning to life as a civilian. She is an articulate, master storyteller, and overall, soft and constrained in her expression. Her voice becomes more expressively compressed when relating her most difficult experiences, but emotional and shaky when she speaks of her family being proud of her. She laughs frequently. Amanda Kean, also a truck driver, disliked her experience in the military greatly, but related her experiences to another older veteran with genuine humor. Amanda did not report any traumatic experiences which could have had caused long-term negative emotional impact. She sustained a back injury from her service, for which she continued to seek treatment, and reported attending college on the GI Bill. She seems to have resumed civilian life successfully. Her voice, overall, is animated with high prosodic variation; and overall, her affect is positive. She laughed sincerely and frequently. In contrast to the other speakers, Ingrid Lim was a psychologist and an officer who looked after the mental health of service men and women. She suffered compassion fatigue and a divorce due to her service. She was effusive, articulate, and emotional in her speaking style, with long responses and few pauses. She laughed frequently, and punctuated her stories with sarcasm. Her voice had significant negative affect with accompanying creakiness when she voiced a negative opinion or related a negative experience. Her voice had high prosodic variation, and frequent points of resonance. As a final contrast, Teresa Little was sent into combat zones without a working gun, and her commanders knew it. She relates the story of a superior officer who got their entire unit lost in the desert and put them at risk by navigating into combat zones which they were supposed to avoid. She was in a mixed male-female unit, and tells the story of becoming pregnant during her service. She relates the story of telling her commanding officers and her mom, and their respective responses ("no you aren't/not another one," and "how did this happen," respectively). She relates her stories in a direct, matter-of-fact style, with liberal doses of humor. She communicates both positive and negative emotions, in a controlled way. When she spoke of negative experience, her voice became softer and more constrained, had moments of creakiness and breathiness, and the pitch dropped. She seems to have transitioned back into civilian life well.

The selected male speakers were just as diverse, and included Joseph Daniel Ancona (LOC Interview, Joseph Daniel Ancona collection), Dax Ashlee Carpenter (LOC Interview, Dax Ashlee Carpenter collection), Andrew James Chier (LOC Interview, Andrew James Chier collection), Christopher M. Gamblin (LOC Interview, Christopher M. Gamblin collection), and Jeremy Brandon Hurtt (LOC Interview, Jeremy Brandon Hurtt collection). Ancona joined while very young, and related his military experience in a very positive way. Before he joined, he was undirected, undisciplined, and flunked out of college. The military gave him direction, discipline, skills training, and new perspectives via new experiences and exposure to diverse cultures. He was an entertaining, engaging speaker and skilled storyteller. His casual, "surfer" style speech was punctuated with many "ums," "ahs," and non-word vocalizations. His speaking style was animated with high prosodic variation, and his affect was overwhelmingly positive. Carpenter, in contrast, spoke of his experiences with reverance, and related his completion of basic training as an accomplishment greater than graduating college in compressed time. He spent his active duty off-base, on missions, in the field, and in combat. One of his stories relates his experience of getting blown up in his truck, surviving, and returning to combat while still injured, at the request of his commanding officer and his own insistence. His soft, expressively-compressed, yet emotional and passionate voice reflects his reverence of the military and pride in his military service, and the pain of the experiences. When he spoke of the ceremony which proclaimed him a marine, his voice became trembly and emotional. In contrast, when he spoke of getting blown up, his voice become expressively compressed and more monotonic. His voice also had periods of creakiness corresponding with negative emotion. He was passionate about his service, and felt responsible for the outcomes of others. He did not laugh often, and had a solemn affect. He seemed to have continuing trauma from his experience, and reported re-enlisting for another tour. He did not seem to want to rejoin civilian life. Chier's experience, in contrast, again, was as an on-base vehicle service personnel. He relates his stories with humor and laughter, and reported what it was like to be gay in the military before this was accepted. He was an articulate speaker who sometimes seemed hesitant and constrained. Hurtt represented another perspective, and joined because of and shortly after the 9/11 attack on the World Trade Center; he was extremely angry at the attackers, and directed his anger into the fight against the forces which attacked the US. He was assigned to the cavalry in Iraq, and sent overseas after a lengthy delay. The military would not release him when his service time was complete. He relates being in dangerous situations during his "overtime" with understated affect, claiming "I'm not even supposed to be here…" Compared to the other speakers, Hurtt was calm, understated, and expressively controlled. Gamblin, another soft-spoken speaker, seemed to have low energy and sounded depressed. His use of creaky voice coincided with relating negative experiences, or possibly boredom. Gamblin reported having some difficulty transitioning back to civilian life. He described feeling annoyed when others discussed mundane events of everyday life, which seemed frivolous to a person who had just returned from active duty. Clearly, we achieved diversity of expression within and across speakers.

4.4.1 General Preprocessing in Unscripted Corpora

Preprocessing was nearly identical to scripted speech. We again downsampled it to 16 kHz, normalized the signal within speaker, excluded sections which had noticeable noise or other sound interfering with the speakers, and converted it to single-channel WAV format. In addition, since the text was not fixed, we had to either create transcriptions or correct the existing Library of Congress transcriptions. Even the best transcriptions in the Veterans History Project required correction, since we intended to force align the text with the signal.

4.4.2 Support for Perception Studies in Unscripted Corpora

The curation process for perception studies of unscripted speech was similar to that of scripted corpora. The studies themselves (see Chapter 5 for details about the perception studies), were also nearly identical to that of scripted speech. Listeners on Mechanical Turk were asked to listen to a clip and provide keywords describing the expressive qualities they perceived in the voice. The purpose of the perception studies was to provide insight into the expressive vocal gestures which untrained listeners would hear, and then use this information to guide selection of consistently-perceived vocal features for detailed analysis grounded in human perception. We were curious if listeners would hear different elements in scripted and unscripted speech, and curious whether the same analytic processes would serve both kinds of speech. Curation for this task meant selecting and extracting meaningful clips for the Mechanical Turk workers to describe. We sought to cover the range of expressivity within and across speakers, and if possible, curate data from similar sections across all interviews. Fortunately, the interviews had similar structure, and we noticed than nearly all the interviews had the following sections: 1) introduction of interviewee, in which the speaker gave their name and demographic data, 2) the interviewee's explanation of the reason for joining the service, 3) the interviewee's description of experiences from basic training, 4) specialized training experiences (e.g., truck driving, auto mechanics, flight school, paratrooping training, etc.), 5) and stories specific to their service. We also noticed that by extracting clips from each of these areas, we could cover the range of vocal expression for each speaker. The introduction served well to model a speaker's natural "baseline," or "neutral," way of speaking. The interviewee's reasons for joining the service provided insight into each speaker's personal circumstances, and provided a common, "baseline" question with opportunities for expressing their individual reasons vocally. Some of the reasons for joining included 1) finding focus, discipline, and training if not personally disciplined or motivated, 2) desire to experience new things and see the world, 3) need of money for college via the GI bill, and 4) desire to serve. The basic training provided a "baseline" experience for each speaker to describe. While each person's response to basic training would be unique, the required elements of basic training were consistent within males and within females. Specialized training experiences and stories specific to service allowed each speaker to tell their unique stories and revealed each speakers' personal expressive style. We found that by selecting 10-15 clips across these areas, we could cover the range of each person's spoken expressivity.

The recordings were segmented into hierarchical phrases and subphrases by the same process as scripted speech. The highest level segments were speaker changes, usually questions from the interviewer and answers from the interviewees. If one of the speakers spoke over the top of another, but floor control did not change, we did not create a new segment for that brief interruption. Segmentation below this level followed the hierarchy of phrases. The highest-level phrases were separated by long pauses, and lower-levels phrases were separated by short pauses. The smallest phrase separations occurred along breath boundaries. We marked the segments in Elan, with separate tiers for each hierarchical level, exported a TextGrid file, and used a custom Praat script to extract the desired clips from the hierarchy. Clips ranged in length from about 2 to 50 seconds. Each exported clip had a name which encoded the speaker, the clip, and its position in the hierarchy. Please see Appendix A for a description of the naming convention.

Because listeners reported hearing laughter so frequently in unscripted speech, and because laughter seemed to have a wide range of expressive character and function, we decided to do a laughter perception study as well, to learn how untrained listeners heard and interpreted laughter. We coded laughter segments on a laughter tier in Elan, and named them with an identification code specifying the speaker, clip number (at the lowest segmentation in the hierarchy), laughter event number within the segment, and finally, and a laughter co-occurrence indicator. This indicator identified whether the laughter was 1) shared between both parties, 2) single-person laughter, 3) the speaker talking and laughing simultaneously, 4) the speaker laughing with the listener talking over the top of him, or 5) the speaker talking and the listener laughing in response. Table 4.1 summarizes the occurrences of these laughter events across the speakers. Mechanical Turk workers listened to these laughter clips and provided descriptors for each laughter segment. Please refer to Chapter 5 for a detailed discussion of the perception study results.

Table 4.1: Laughter Events Per Speaker. This table shows the total number of laughter events per speaker, and the raw numbers and percentage of different kinds of single-person laughter events and interactive laughter events. Some speakers clearly laughed more than others, and the distribution of the different kinds of laughter across speakers varied as well.

	Single-Person Events		Interactive Events			
Speaker	Laughter	Laughing & Talking	Shared Laughter	Speaker Talking/ Listener Laughing	Speaker Laughing/ Listener Talking	Total
Ferretti	7 (78%)	1 (11%)	1 (11%)	0	0	9
Kean	5 (46%)	4 (36%)	2 (18%)	0	0	11
Lim	14 (44%)	3 (9%)	10 (31%)	5 (16%)	0	32
Little	40 (70%)	5 (9%)	4 (7%)	5 (9%)	3 (5%)	57
Sluss	4	6	0	0	0	7
Total	70 (60.3%)	16 (13.8%)	17 (14.7%)	10 (8.6%)	3 (2.6%)	116

4.4.3 Support for Voice Quality and Nonverbal Quality Analytics in Unscripted Corpora

The curation process for scripted voice quality analytics was very similar to that of scripted corpora. Please refer to section 4.2 in this chapter for details of the voice quality curation process, and to Chapter 6 for details about the unscripted voice quality analysis. An important difference between curation for scripted and unscripted voice quality analytics was that listeners reported hearing creaky voice in the unscripted corpora, but did not report hearing whispered voice.

Listeners also reported hearing a wider range of emotion in unscripted voices. These findings are reasonable, since a speaker in an interview could use creaky voice to express a wider range of spontaneous emotion than the emotional range which is present in the Shakespearean soliloquy samples; and creaky voice is not typical in Shakespearean acted speech. Furthermore, interviewers and interviewees typically do not whisper. They are speaking to each other in "public," recorded dialogue, and want to be clearly and audibly heard. Curation for unscripted voice quality analysis required labeling and extracting samples of breathiness, creaky voice, modal speech, and resonance to support modeling of these voice qualities. As before, we included modal speech as a baseline comparison for the exceptional qualities which listeners reported hearing (breathiness, creakiness, and resonance).

We again forced-aligned the text, extracted all the vowel sounds in the corpus which were at least 60 msec long, and asked one expert listener to code them per syllable according to breathiness, modal speech, creakiness, and resonance. As with the scripted corpora, expert listeners for this activity were defined as those who had formal training in linguistics, speech processing or music. Musicians in their domain are trained to hear and produce nuanced variations in tempo, dynamics, accenting, timbre, etc.; and this translated well to the specific task at hand. A second expert listener coded a random sampling of 20 samples from each condition across the corpus. For females, we reached 85%, 75%, 90%, and 90% agreement over breathy, modal, creaky, and resonant speech respectively, with a Cohen's kappa of 0.80. The corpus final result included a comparable **199 breathy** (15 Ferretti, 28 Kean, 38 Lim, 92 Little, 26 Sluss), **1120 modal** (220 Ferretti, 208 Kean, 112 Lim, 294 Little, 286 Sluss), **1318 creaky** (36 Ferretti, 14 Kean, 46 Lim, 195 Little, 52 Sluss), and **703 resonant vowel utterances** (42 Ferretti, 20 Kean, 20 Lim, 20 Little 47 Sluss), (1848 vowels total). This corpus sample provided a total of 423, 3548, 1318, and 703 individual data frames for analysis for breathy, modal, creaky, and resonant speech, respectively. For males, we reached 85%, 65%, 85%, and 90% agreement over breathy, modal, creaky, and resonant speech respectively, with a Cohen's kappa of 0.78. The corpus final result included a comparable **152 breathy** (5 Ancona, 55 Carpenter, 36 Chier, 31 Gamblin, 25 Hurtt), **769 modal** (174 Ancona, 134 Carpenter, 176 Chier, 138 Gamblin, 147 Hurtt), **126 creaky** (14 Ancona, 21 Carpenter, 32 Chier, 26 Gamblin, 33 Hurtt), and 163 **resonant vowel utterances** (135 Ancona, 5 Carpenter, 6 Chier, 0 Gamblin, 17 Hurtt), (1210 vowels total). This corpus sample provided a total of 278, 2749, 613, and 827 individual data frames for analysis for breathy, modal, creaky, and resonant speech, respectively. Each speaker contributed some of each condition, although not all of the voices had the same amount of each kind of expressive speech.

A second difference in the scripted and unscripted speech was the presence of laughter. As before, a primary expert listener coded the laughter events described in section 4.4.2 of this chapter; and a secondary expert listener coded a random 20% of the laughter samples, with a Cohen's kappa of 0.9 for agreement on the presence of laughter and agreement of 95%, 87%, 90%, 95%, and 90% for cases 1-5 of the laughter co-occurrences described above.

4.4.4 Support for Dimensional Discovery in Unscripted Corpora

We ran two different dimensional discovery analyses for unscripted speech. The first analysis sought to identify general expressive dimensions in unscripted speech, using the descriptors from the perception study and an LSA analysis process. The second analysis sought to identify different dimensions of laughter, again using the descriptors (this time from the laughter perception study) and a separate LSA analysis process. Chapters 7 and 8 describes the dimensional discovery work in detail, including general expressive dimensional discovery and laughter dimension discovery.

4.4.5 Unscripted Corpora Scale

To summarize the overall scale of the unscripted corpora, it included 10 speakers total, 5 male and 5 female, with the total lengths of the female interviews given as follows: 1) Ferretti, 6567 seconds, 2) Kean, 2751 seconds, 3) Lin, 2685 seconds, 4) Little, 4773 seconds, and 5) Sluss, 1791 seconds. The total amount of available female veteran speech was, therefore, 18,867 seconds, or about 5.24 hours. Of this total available speech, 614 seconds (about 3.25% of the available data, ie, around 10 minutes) was sampled and used for perception studies, voice quality modeling, and dimensional analysis (Feretti, 149 seconds; Kean, 97 seconds; Lin, 78 seconds; Little, 150 seconds; and Sluss, 140 seconds).

The scale of the male unscripted corpora was similar, with the lengths of the male interviews given as follows: 1) Ancona, 1896 seconds; 2) Carpenter, 4567 seconds, 3) Chier, 5048 seconds, 4) Gamblin, 2145, and 5) Hurtt, 3441 seconds. The total length of the available male veteran speech was, therefore 17,097 seconds, or about 4.75 hours. Of this total available speech, 559 seconds (about 3.3% of the available data, ie, around 9.5 minutes) was sampled and used for perception studies, voice quality modeling, and dimensional analysis (Ancona, 78 seconds; Carpenter, 77 seconds; Chier, 122 seconds; Gamblin, 92 seconds, and Hurtt, 90 seconds).

4.4.6 Potential Biases Within LOC Oral History Data

The Veterans Oral History collection, taken in its entirely, has the potential for a few inherent biases, including age, gender, demographics, technology, and biases caused by the relatively small number of speakers. First, the corpus contains many more male speakers than female speakers. This is expected, given the gender biases for war in our culture. Also, male recordings are more likely to contain accounts of direct combat than female recordings, simply because of restrictions for women participating in combat. This could result in differences in expressivity due to the nature of the experiences under discussion. On the other hand, many of the male speakers' duties were limited to military bases; and they did not experience violence. Conversely, female veterans, although not permitted to be combat soldiers, often experienced violence because their duties (e.g., truck driving) put them into a combat zone. Many of these biases can be overcome by exploring a balanced number of male and female speakers, and by selecting for a range of experiences (across the range of violence, and many other things). The selections made for this research succeeded in overcoming these gender and topic biases.

The corpus sub-sample studied for this research does have potential for biases due to the small number of speakers examined. For example, a single speaker with a distinctive expressive style could cause discovery of an expressive dimension which is specific to that speaker. This isn't necessarily a bad thing, if the discovered dimension is weighted in proportion to its occurrence, with respect to the other expressive dimensions in the corpus. One way to compensate for this bias type is selecting a diverse range of speakers with different demographic backgrounds, while still limiting the selections to L1 speakers of American English. Also, when possible, sampling across each speaker's expressive range helps guard against speaking style biases from individual speakers. Speakers usually vary their expressive speaking styles across a 30-120 minute interview; therefore, getting diverse samples from each speaker is natural and is not difficult. This research purposely used a diverse selection of male and female speakers, with diverse demographics. I also sampled each speaker's range of expressivity, and sampled across common discussion points in the interviews, particularly self-introduction, discussion of basic and specialized training, reasons for joining the military, and descriptions of 1-3 personal experiences from their service. Although the best solution to the single-speaker bias problem is using a large number of speakers and a large number of samples from each speaker, the techniques mentioned here help overcome single-speaker biases when large-scale analysis is not possible.

Another potential bias is the result of the evolution of technology. The highest-quality digital recordings are generally the most recent interviews. Older recordings made on tape usually are noisier, and depending on the age of the media, might have degraded beyond the base quality of tape, if many years elapsed between the original recording and digitization. In general, older, noisier media (especially cassette tape) were used to record veterans of the WWII, Korean, and often, Vietnam conflicts; and newer digital media were used to record veterans of the Iraq and Afghanistan conflicts. This difference in quality of media over time introduces a technological bias which, in turn, introduces a potential age bias in the sampled data. Seeking higher-quality recordings for analysis will overwhelmingly select younger speakers. WWII veteran speakers recorded on high-quality digital media are relatively rare in this collection. The WWII speakers' voices which are present on high-quality media will reflect the characteristics of old age in addition to general expressivity. Overall, however, the corpus is biased toward an older speaker age, simply because it contains more collectively more interviews with the WWII, Korean, and Vietnam veterans.

The corpus contains interviews with veterans from diverse cultural regions across the country, and the recordings will reflect the speakers' regional dialects and cultures. This characteristic is an advantage for avoiding regional bias, but it also introduces challenges for analysis to consider these factors. The scale of the data examined in this research did not enable close examination of dialect-dependent expression, but an exploration of data at larger scale could. Furthermore, at larger scale, listeners speaking a given dialect could be paired with speakers of a given dialect to examine cross-cultural perceptive biases.

The specific sub-sample examined in this research avoided many of the biases present in the larger data set. It contained a diverse sample of dialects (noted in the corpus metadata) and the listeners coding the vocal expression also came from different dialect-speaking regions of the US. These listener demographics were noted, but as mentioned, a larger sample of speakers and listeners would be required either to control for or explore any listener and speaker biases. The research sub-sample also selected a balanced number of males and females to gender differences; so the research results did not reflect gender bias. Furthermore, the speaker samples contained males and females exposed to both peaceful and violent experiences. The speaker samples were, however, taken from the most recent recordings in digital media, and therefore are biased to younger speakers. This sub-sample has the opposite age bias in comparison to the overall bias in the data set.

4.5 Privacy and Copyright Restrictions

The Shakespearean soliloquy recordings (Lady Macbeth and Hamlet corpora) examined here are available on YouTube; and these YouTube recordings are excerpts from movies or stage performances, which another party clearly owns. They should not be used in ways which will violate copyright laws. This means that using samples of the recordings are allowed only if it complies with the "fair use" doctrine (YouTube Fair Use Policy). Fair use considerations include 1) the purpose of the use, e.g., whether it is being used for commercial or educational reasons, 2) the nature of the work (e.g., whether fictional or factual), 3) the specific portion of the work selected for use, and the amount with respect to the size of the whole, and 4) whether the use will affect the market value of the object in question. It is best to secure permission for specific uses from the copyright owner.

The oral history interviews are available from the Library of Congress, and their use policies are discussed in the Veterans History Project FAQ (Library of Congress Veterans History Project, Use of Collection Materials). The LOC does not claim ownership of the interviews themselves because they are a federal agency, and therefore, their publications are in the public domain. The interviewers and interviewees, however, hold the copyright over their recordings. This means that, in general, securing permission is required to use materials in exhibition, publication, etc. The LOC does say that individuals may use materials in ways "permitted by copyright law," which means anything which is allowed by the fair use doctrine. This means that the materials can be used for research, and the objects can be examined and studied, and the results research shared; but, activities such as 1) quoting text from the interviews, 2) playing or exhibiting the recordings publically, or 3) publishing the recorded or printed material directly requires permission from the copyright holders.

4.6 Summary

In this chapter, I covered the ideal characteristics of scripted and unscripted corpora from the perspective of answering the research questions. Then, I discussed the corpora selections, and reasons for rejecting other corpora. Next, I discussed curation and preprocessing necessary for the support of perception studies, analysis of expressive vocal and nonverbal qualities, and for discovery of expressive dimensions. Finally, I discussed the potential biases present in the scripted and unscripted corpora (the Hamlet, Lady Macbeth, and Veterans Oral History datasets), discussed the scale of the corpora examined in this research (and the potential for scaling up), and reviewed the restrictions for use of each corpus.

The selection and curation of the datasets were critical for entire body of analytic work in this thesis, beginning with the perception studies, which answered questions about what people heard in scripted and unscripted speech. The next chapter discusses the perception study processes and results for both the scripted and unscripted corpora.

CHAPTER 5: PERCEPTION OF VOCAL EXPRESSION

This chapter describes the methods used to address research questions RQ1, RQ3, and RQ8, along with the experimental results. Specifically, it addresses what untrained listeners hear, expressively speaking, in scripted and unscripted speech. Of all possible expressive characteristics listeners could perceive and describe, which ones would win in their perception? Would they notice whispering or breathiness, and could they notice the difference between the two? Which prosodic features would they perceive and articulate? How would listeners handle emotive expression? What differences manifested between the perception of male and female expressivity, and what differences could be found between the perception of scripted and unscripted perception in our corpora? These exploratory studies (approved by institutional board review, or IRB, procedures at the University of Illinois) were used to gain insight into the perception of vocal expression, not to draw conclusions yet about the relationships among the elements which listeners perceived, and not to provide ground truth coding for acoustic modeling. We emphasize that developing acoustic models of vocal quality, and the ground truth coding to train the models, was a separate activity (described in detail in Chapter 6). These studies provided guiding insight into human perception of vocal expression. They revealed which expressive qualities listeners perceived repeatedly and consistently, provided rationale for selecting specific voice qualities and nonverbal qualities for detailed analysis, and provided the raw data which could be used in the discovery of expressive dimensions (described in detail in Chapter 7). They revealed a qualitative difference in perception of emotion vs. prosody and voice quality, which guided the selection of analysis techniques to fit the differences. These studies were a fundamental and critical step in grounding the analysis in human perception.

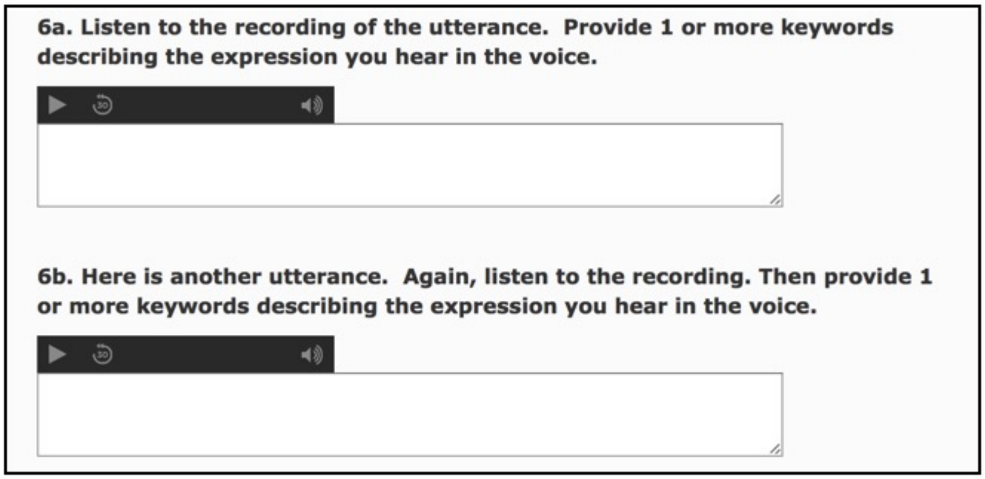
This chapter presents the methods and results for the following three studies: 1) vocal expression perception in Shakespearian scripted speech, 2) vocal expression perception in semi-structured, unscripted, oral history interviews, and 3) laughter perception in semi-structured unscripted oral history interviews. More specifically, the chapter describes Mechanical Turk studies which elicit open-response keyword description of expressive scripted and unscripted speech and laughter from the listeners. These study results serve the purposes of 1) discovering what people hear regarding human expression, 2) grounding the analytic work in human perception (described in later chapters), and 3) enabling a latent semantic analysis driven discovery of the expressive dimensions present in speech and laughter. This chapter closes with a summary of the results across all three studies. Much of this perception work has been published (Pietrowicz et al. 1, 2017; and Pietrowicz et al. 2, 2017).

5.1 Methods

We used similar methods for each of the three perception studies. In each study, Mechanical Turk workers were presented with a task containing the task description, the IRB boilerplate, an option to accept or decline the task (collected for IRB purposes), a short demographic survey (designed to determine whether the listener was an L1 American English speaker and reveal listener biases), task instructions, a qualifying task (to help ensure that participants understood the task, took it seriously, were capable of doing the task, and had no technical problems which would interfere with completing the task), and a single listening task designed to elicit description of vocal expression present in an audio clip. The listening task was intentionally simple. It asked workers to listen to a single audio clip, then describe the vocal expression present in the task by giving keyword descriptors in open-response format.

For each of the three studies, we iteratively piloted the task on clips taken from the corpora, but not selected for use in the perception studies. Please refer to Chapter 4 for detailed discussion about selection and curation of corpora to support the perception studies. The pilot tests allowed us to experiment with different presentations of the task, payment amounts, clip sizes, etc. with the goals of minimizing worker error, maximizing the quality of the data collected, setting an appropriate price, and minimizing the time required to complete the study. In general, simplifying the task, presenting the task description in light, playful language, and shortening/simplifying text increased the quality of the listener description and decreased task completion time. Specific factors found to decrease the quality of the results included 1) asking for more keywords than listeners could provide, 2) asking workers to evaluate multiple clips in a single task, 3) unnecessary complexity in the directions, 4) too many demographic questions, 5) excessive text in general, and 6) audio clips longer than about 40 seconds. When these studies were conducted, Mechanical Turk workers were accustomed to short, simple tasks with quick rewards, not 45-minute in-lab-style studies. Many workers would not accept tasks which appeared too complex. Even inclusion of the IRB boilerplate increased study time completion because workers were slower to accept tasks which required them to view and navigate “legal speak”. Encapsulating the IRB text in a box, introducing it as “the fine print,” and reducing its font size in comparison to the actual study text font size helped direct workers to the task and resulted in increased task acceptance rates.

The first and last perception studies were separated in time by about two years; therefore, piloting each study was helpful for tuning the tasks to adapt to the changing skill, experience, and demographic of Mechanical Turk workers. The overall task format remained the same from beginning to end, but we had to increase payment for the tasks by about 20 cents per task to get the workers to accept the work for the last study. We were, however, able to ask for more keywords in later studies without decreasing the quality of the keywords collected. We suspect that either 1) the worker base had become more experienced, and was more willing and able to handle slightly more complex tasks, 2) the nature of unscripted speech and laughter was more complex and invited more detailed description naturally, 3) the emergence of worker search tools allowed them to prioritize available tasks by reward and exclude from viewing tasks and task sources which did not meet a minimum payment standard, and 4) emerging standards for payment on the Mechanical Turk platform. The following sections describe the methods and results for each study in detail. Figure 5.1 shows a sample Mechanical Turk task which presents the audio clip and solicits keyword descriptors.



6a. Listen to the recording of the utterance. Provide 1 or more keywords describing the expression you hear in the voice.

6b. Here is another utterance. Again, listen to the recording. Then provide 1 or more keywords describing the expression you hear in the voice.

Figure 5.1: This Mechanical Turk task excerpt shows the format of the presentation of clips and request for keyword descriptors. 6a is the qualification task, and 6b is our data collection task. The format is a simple open-response prompt.

The study intentionally asked listeners to describe their perceptions in their own words because understanding human perception was a goal. The demographics collected of the listeners revealed a range of backgrounds, listening and speaking experiences, and wide regional distribution across the US, both in their current residency and in the regions where they grew up and acquired language skills. The sample size encouraged and accomplished diversity in background and origin of native English-speaking Mechanical Turk workers, which was a desired goal. However, also note that a larger number of listeners would be required to discover any regional, cultural, or experiential biases which might cause a listener to describe expressive content in a specific way using specific vocabulary. While the current design does control for selection of L1 speakers of American English, it does not control for regional biases. Latent semantic analysis applied to the keywords helps mitigate any biases present in the listener base by considering the contexts (audio clips) in which specific terms occur (refer to Chapter 7 for a detailed description of the technique). Exploring perceptive biases based on demographic background variation is an interesting question for future work, which can be accomplished by sampling more listeners and ensuring sufficient numbers of listeners in the demographic categories of interest.

5.1.1 Perception of Vocal Expression in Scripted Speech

Mechanical Turk workers (limited to native speakers of American English) were invited to play a single Hamlet or Lady Macbeth soliloquy excerpt, and then provide one or more keywords describing what they heard in the vocal expression (not in the speech text). Many workers provided multiple keywords, but this was not required. Prior studies have shown that simplifying the task, and reducing cognitive load in Mechanical Turk tasks, improves the quality of the results on this crowdsourcing platform, especially when dealing with sound (Pietrowicz et al., 2013). Workers were not given a list of keywords to choose from, but were simply asked to describe the characteristics of the speakers' vocal expression. Each Turk task contained only a single audio clip to avoid priming effects and fit the study into the Mechanical Turk paradigm.

For males, the survey included sound excerpts from the five professional actors described in Chapter 4 (Branagh, Burton, Gibson, Jacobi, and Tennant), and focused on the opening phrase of the Hamlet soliloquy, "To be or not to be, that is the question." Listeners heard this clip either in its entirety, or heard one of the three sub-phases: "To be," "Or not to be," or "That is the question." This exploratory study included a total of forty Mechanical Turk tasks for each speaker, with ten listeners evaluating each clip. This way, listeners (as a collective) had the opportunity to hear and comment on an excerpt in its larger context, or comment on the isolated, specific, expressive nuances featured in one of the smaller sub-phrases. For example, in Jacobi's utterance, the initial "To be" was soft, and breathy or whispery, but the middle phrase, "Or not to be," was loud and resonant in quality (an impressive contrast already). The final phrase, "That is the question," had a modal quality with a large variation in pitch, and multiple accents. By taking this approach, the study incorporated differences in perception at multiple scales.

For this exploratory survey of male acted speech, the sample (approximately 4% of the Hamlet corpus) was representative of the range of vocal expression across the entire Hamlet corpus, captured the essence of each speaker's performance in the soliloquy, provided vastly different interpretations of the character Hamlet and represented a variety of native-English speaking countries and backgrounds (Great Britain, America, and Australia).

For the female speakers, a similar exploratory study used excerpts from the Act I Scene V soliloquy (Lady Macbeth speaking) described in detail in Chapter 4. The survey included five professional actors (Dench, Fleetwood, Walter, Whalley, and White) and included the following four phrases: 1) Phrase 1: "Unsex me here and fill me from the crown to the top full of direst cruelty,", 2) Phrase 2: "Come thick night and pall thee in the dunnest smoke of hell," 3) Phrase 3: "Nor heaven peep through the blanket of the dark to cry, Hold, hold'" and 4) an excerpt included the full context of phrases 1, 2, and 3. The selections were representative of the expressive range across and within speaker, and to attain this representation, about 27% of the corpus was surveyed. Again, forty mechanical Turk tasks were presented per speaker, ten distinct listeners per each of the four phrases. As with Hamlet, the female actors presented different interpretations of the character, speaking the same text, with varying vocal expression. The format of the Mechanical Turk task for female speech was identical to that of male speech.

This study design had several advantages over the traditional approach of bringing subjects into the lab, and presenting every clip to every research subject. First, the Mechanical Turk platform provided access to a wide range of workers from a wide range of backgrounds. About 400 different listeners collectively provided description, which avoided the problems of within-subject bias in small numbers of listeners and of a limited range of subjects. By using larger numbers of listeners, the collective response approached a population normal. Next, the Turk platform allowed limiting listeners to native US speakers of English, which the study accomplished by using only US workers and then collecting demographic data designed to identify qualified listeners. Also, because the tasks contained only one audio clip, and not the entire set of audio clips, the listeners were not subject to hearing both a long phrase and one of its sub-phrases. Individual listeners, therefore, did not hear any part of a phrase more than once, and did not experience priming effects. Next, inclusion of a long phrase and its sub-phrases allowed incorporating perception of short and longer phrases, with different amounts of expressive variance, in the same study. Finally, perception of the long phrase did not equal the sum of the perception of the 3 sub-phrases; listeners provided different sets of keywords. For this reason, and because each Turk worker analyzed only one clip, the analytic approach was simplified to treat each phrase as a single, independent entity. This design supported the study goals (to provide a general understanding of what listeners heard in expressive speech, and to provide information necessary for guiding selection of perceived features for detailed acoustic analysis).

For both male and female studies, we designed a qualification task to verify the ability and willingness of a worker to provide valid input. This task was effective for screening out trolls (participants purposely providing offensive or contrary responses), people who were not able to hear differences in vocal expression, or people who misunderstood the study directions. Examples of answers which failed the qualification task included profanity, obvious summarization of the text instead of description of the expression, statements indicating that a listener could not hear the recording, or non-serious responses which did not provide any description. The qualification task was nearly identical to the target task. It presented a Hamlet or Macbeth audio clip which was not used in our study and asked listeners to provide keywords describing vocal expression. About 5% of responses were excluded for the reasons cited above.

The research question asks what people consciously perceive in acted voices and addresses the question in the manner that listeners naturally hear language, with the semantic and paralingual content intact. It does not attempt to obfuscate the semantic content. If listeners misunderstood the study directions and instead described the semantic content, the qualification task provided justification to exclude their responses. Less than 2% of responses fell into this category.

5.1.2 Perception of Vocal Expression in Semi-structured, Unscripted Speech

The methods for exploring perception of unscripted speech were nearly identical to the methods used for exploring scripted speech. Again, we presented representative audio samples covering the range of expression for each of 10 speakers (5 male, 5 female), and asked Mechanical

Turk workers to provide three or more keywords describing the vocal expression in the speakers' voices. Chapter 4 describes corpora curation criteria and processes used to select the representative set of speakers and clips for each speaker. The survey included 10-15 representative speech segments (4-45 seconds each) for each speaker, and 10 Mechanical Turk workers evaluated each clip for a total of over 1000 Turk tasks and over 3000+ keywords describing the range of vocal expression across the speakers. The main methodological differences between the scripted and unscripted studies were 1) small updates in the required IRB boilerplate, 2) the ability of workers to respond to a request of three or more descriptors instead of just 1 or more descriptor, 3) the ability of workers to respond to longer sound clips, and 4) clips presented at a single hierarchical level for the unscripted speech and at multiple levels for the scripted speech.

5.1.3 Perception of Laughter in Semi-structured, Unscripted Speech

When laughter was present in a clip, listeners almost always commented on it. Laughter, like speech, differs in its expressive quality. It accompanies a range of emotions, not always humorous ones, and often not positive ones. It, like speech, seemed to have its distinct set of voice qualities (or timbres), emotions, prosodic inflections, and conversational elements associated with it. The results of the semi-structured, unscripted perception study were so provocative regarding laughter that we ran a low-level perception study focused just on laughter to investigate. Furthermore, some of the expressive dimensions discovered in Chapter 7 included an element of laughter; therefore, detailed analytics of laughter would possibly be useful in support of the detection of expressive dimensions in speech. The methods for exploring laughter were almost identical to the methods for exploring general expression in scripted and unscripted speech. Chapter 4 describes the curation criteria and processes used to select and extract laughter clips from unscripted speech. All laughter events from the selected oral history speakers were extracted and presented to Mechanical Turk workers, with the same open prompt requesting three or more descriptors. The study included 116 clips total, and 10 Mechanical Turk workers evaluated each clip, for a total 1160+ keywords describing the range of expression listeners perceived in laughter. The main methodological differences included 1) the ability of workers to respond to requests of three or more keywords, and 2) all laughter clips were selected for analysis (not just a representative sample from each speaker), and 3) laughter clips were presented by themselves, not in the larger context, to allow listeners to focus on the quality of the laughter itself. Laughter studied in context requires additional considerations and should be the focus of future studies.

5.2 Results

5.2.1 Perception of Vocal Expression in Scripted Speech

The keywords were collected for each speaker, consolidated by close synonym as defined by a thesaurus (Online thesaurus of English), and sorted by frequency. All words defined as "close synonyms" by the referenced thesaurus were grouped together under the most frequently used tag within the synonym group. For example, "resonant," "sonorous," "projected," and "ringing" are close synonyms, and were tagged and counted together under the most frequently-given label. Workers consistently provided a small set of simple, concise, voice quality and prosodic descriptors (such as "slow," "soft," "whispered," or "ringing"), and a wide range of more nuanced emotion-based keywords (e.g., thoughtful, pensive, happy, joyful). Listeners frequently gave the phonation or effort level type when whispering, breathiness, resonant speech, or yelling occurred. The synonym-reduced results in Appendix B show the most frequently-given emotion and non-emotion keywords for each speaker (up to 12 keywords), with unit-frequency words removed.

Listener keywords were clustered into the categories of voice quality, prosody, and emotion. Much of the prior work in paralingual expression has focused on one or more of these areas; and grouping the given keywords in this way allows exploration of 1) the relative frequencies in which listeners perceived qualities in these categories, 2) the variation of keywords given in each category, 3) the specific qualities perceived in Shakespearian acted speech in each category, 4) the relationships among keywords given in each category, and 5) the differences between male and female speakers in each category. Tables 5.1, 5.2, and 5.3 subdivide voice quality and prosody into subcategories, and summarize the results statistically. Tables 5.1 and 5.2 show both raw numbers of keywords provided in each category per speaker and the proportion of keywords provided in each category, per speaker. Since the listeners provided different numbers of keywords for each speaker, Table 5.6 summarizes both the raw numbers and percentages of keywords provided for males and females in each category. Most of the keywords (on the average, about 59%) were emotion keywords, which were nuanced, ranging far beyond emotions considered "basic" by any of the candidate paradigms (Ontony and Turner, 1990; D'Mello and Calvo, 2013). The remaining keywords were nearly evenly divided between voice quality and prosody (on the average, about 21% voice quality and about 20% prosody). The prosodic and voice quality keywords were concise and simple; the same small set of keywords repeated across all the speakers, male and female. Four notable voice quality concepts recurred across speakers, and included whispering, breathiness, yelling, and resonance. These keywords described phonation types, or effort levels, and comprised about 12% of all descriptors in the dataset and more than 57% of all voice quality descriptors provided. The distinct presence of whispering and breathiness show that listeners are sensitive to not only the two qualities, but are aware of the distinction between them and are able to articulate it without any prompting. Similarly, listeners were sensitive to the difference between a resonant quality and yelling. Although these two vocal qualities typically corresponded to louder volumes than the other qualities, listeners heard yelling and resonant quality at multiple levels of loudness, and distinguished resonant speech from non-resonant speech at conversational levels of volume.

The prosodic keywords further divided into pitch, loudness, and speaking rate subclusters (common prosodic categories from the literature). On the average, listeners perceived speaking rate and loudness at similar rates (about 10% vs 9% of keywords provided, respectively). Only about 1% of keywords described pitch, even though many of the speakers had a high degree of pitch variation. Interestingly, listeners did comment on below-average variations in pitch, volume, and speaking rate, but labelled this combined quality as “monotone,” or “flat.” It is probable that listeners hear pitch variation, but use it to infer higher-level qualities at the linguistic layer of language. Results suggest that both prosody and voice quality may help drive the perception of emotion, but understanding and drawing conclusions regarding these potential relationships require further studies which are designed specifically to examine these potential relationships.

Listeners perceived the female talkers’ expressive speech differently from male talkers’ even though the speaking style, topic, and emotional content were similar between the Hamlet and Lady Macbeth soliloquy. These differences are statistically significant at $\alpha=0.05$. A Chi-Square test for Independence between gender (male, female) and descriptor type (prosody, emotion, and voice quality) categories showed that descriptor type depends on gender (χ^2=16.5, df=2, p=0.00026). Additional Chi-Square tests reveal that prosody (χ^2=6.64, df=1, p=0.0099), voice quality (χ^2=5.59, df=1, p=0.018), emotion (χ^2=16.51, df=1, p=0.00005), effort levels (χ^2=7.10, df=1, p=0.0077), and speaking rate (χ^2=6.99, df=1, p=0.0082) all varied significantly with gender. Non-effort-level voice quality (χ^2=0.11, df=1, p=0.742) and loudness (χ^2=0.011, df=1, p=0918) did not vary significantly with gender. The dataset did not have enough data in the pitch category to run a Chi-Square test without violating the Central Limit Theorem. The significant increase in emotion keywords and significant decrease in effort level and speaking rate for females is provocative, and is explored further in Chapter 7.

Listeners perceived the female talkers' expressive speech differently from male talkers' even though the speaking style, topic, and emotional content were similar between the Hamlet and Lady Macbeth soliloquy. These differences are statistically significant at $\alpha=0.05$. A Chi-Square test for Independence between gender (male, female) and descriptor type (prosody, emotion, and voice quality) categories showed that descriptor type depends on gender (χ^2=16.5, df=2, p=0.00026). Additional Chi-Square tests reveal that prosody (χ^2=6.64, df=1, p=0.0099), voice quality (χ^2=5.59, df=1, p=0.018), emotion (χ^2=16.51, df=1, p=0.00005), effort levels (χ^2=7.10, df=1, p=0.0077), and speaking rate (χ^2=6.99, df=1, p=0.0082) all varied significantly with gender. Non-effort-level voice quality (χ^2=0.11, df=1, p=0.742) and loudness (χ^2=0.011, df=1, p=0918) did not vary significantly with gender. The dataset did not have enough data in the pitch category to run a Chi-Square test without violating the Central Limit Theorem. The significant increase in emotion keywords and significant decrease in effort level and speaking rate for females is provocative, but we cannot conclude whether the differences are a result of gender or a result of the inherent differences between the selected male and female Shakespearian parts.

Table 5.1: Allocation of keyword descriptors across keyword classes, for **scripted male speech**. This table shows, for each male speaker, the raw number of keywords given (with percentages in parentheses) for voice quality, prosody and emotion. It further subdivides voice qualities into effort levels and other voice qualities, and further subdivides prosody into pitch, loudness, and speaking rate. In this example, 54% of Kenneth Branagh's keywords described emotion in the voice, 16% described prosodic qualities, and 30% described voice quality. A full 25% of Branagh's keywords described an effort level such as 'breathy'. Listeners provided a total of 77 keywords for Branagh, 69 for Burton, 75 for Gibson, 89 for Jacobi, and 102 for Tennant.

Keyword Class	**Branagh**	**Burton**	**Gibson**	**Jacobi**	**Tennant**
Voice Quality	**23** (29.9%)	**9** (13.0%)	**20** (26.7%)	**18** (20.2%)	**33** (32.3%)
Effort Level	**19** (24.7%)	**2** (2.9%)	**14** (18.7%)	**9** (10.1%)	**20** (19.6%)
Other Quality	**4** (5.2%)	**7** (10.1%)	**6** (8.0%)	**9** (10.1%)	**13** (12.7%)
Prosody	**12** (15.6%)	**21** (30.4%)	**17** (22.7%)	**17** (19.1%)	**32** (31.4%)
Pitch	**0** (0.0%)	**2** (2.9%)	**0** (0.0%)	**3** (3.4%)	**4** (3.9%)
Loudness	**3** (3.9%)	**1** (1.4%)	**5** (6.7%)	**10** (11.2%)	**16** (15.7%)
Speaking Rate	**9** (11.7%)	**18** (26.1%)	**12** (16.0%)	**4** (4.5%)	**12** (11.8%)
Emotion	**42** (54.5%)	**39** (56.6%)	**38** (50.6%)	**54** (60.7%)	**37** (36.3%)

Table 5.2: Allocation of keyword descriptors across keyword classes, for **scripted female speech**. This table shows, for each female speaker, the raw numbers of keywords given for voice quality, prosody, and emotion. It further subdivides voice qualities into effort levels and other voice qualities, and further divides prosody into pitch, loudness, and speaking rate. In this example, 63% of Dench's keywords described emotion in the voice, 15% described prosodic qualities, and 22% described voice quality. 13% of Dench's keywords described an effort level such as 'whispering'. Listeners provided a total of 116 keywords for Dench, 151 for Fleetwood, 83 for Walter, 117 for Whalley, and 109 for White.

Keyword Class	**Dench**	**Fleetwood**	**Walter**	**Whalley**	**White**
Voice Quality	**26** (22.4%)	**27** (17.8%)	**20** (24.1%)	**23** (19.7%)	**12** (11.0%)
Effort Level	**15** (12.9%)	**15** (9.9%)	**11** (13.3%)	**13** (11.1%)	**3** (2.75%)
Other Quality	**11** (9.5%)	**12** (7.9%)	**9** (10.8%)	**10** (8.6%)	**9** (8.25%)
Prosody	**17** (14.7%)	**20** (13.2)	**13** (15.7%)	**22** (18.8%)	**28** (25.7%)
Pitch	**1** (0.9%)	**0** (0.0%)	**0** (0.0%)	**1** (0.8%)	**1** (0.9%)
Loudness	**9** (7.8%)	**12** (7.9%)	**6** (7.2%)	**9** (7.7%)	**14** (12.9%)
Speaking Rate	**7** (6.0%)	**8** (5.3%)	**7** (8.5%)	**12** (10.3%)	**13** (11.9%)
Emotion	**73** (62.9%)	**104** (69.0)	**50** (60.0%)	**72** (61.5%)	**69** (63.3%)

Table 5.3: Comparison of the distribution of keyword descriptor types across **male and female scripted speech**. The results show the means of the numbers of keywords given for voice quality, prosody, and emotion across males, females, and all talkers. Percentages are given shown in parentheses. Listeners provided proportionally 12.8% more keywords describing female talkers' emotions than male talkers' emotions. Conversely, listeners provided 5.6% fewer voice quality descriptors and 7.8% fewer prosodic descriptors for females than males. Almost all of the reduction in voice quality is accounted for in the reduced proportion of effort level descriptors. σ represents standard deviation from the mean here.

Keyword Class	**Male Talkers**	**Female Talkers**	**All Talkers**
Voice Quality	**$\mu=20.6, \sigma=8.7$** (25.0%)	**$\mu=21.6, \sigma=6.0$** (18.8%)	**$\mu=21.1, \sigma=7.0$** (21.3%)
Effort Level	**$\mu=12.8, \sigma=7.5$** (15.5%)	**$\mu=11.4, \sigma=5.0$** (9.9%)	**$\mu=12.1, \sigma=6.0$** (12.2%)
Other Quality	**$\mu=7.8, \sigma=3.4$** (9.5%)	**$\mu=10.2, \sigma=1.3$** (8.9%)	**$\mu=9.0, \sigma=2.7$** (9.1%)
Prosody	**$\mu=19.8, \sigma=7.5$** (24.0%)	**$\mu=20.0, \sigma=5.5$** (17.4%)	**$\mu=19.9, \sigma=6.3$** (20.1%)
Pitch	**$\mu=1.8, \sigma=1.8$** (2.2%)	**$\mu=0.6, \sigma=0.5$** (0.5%)	**$\mu=1.2, \sigma=1.4$** (1.2%)
Loudness	**$\mu=7.0, \sigma=6.0$** (8.5%)	**$\mu=10.0, \sigma=3.1$** (8.7%)	**$\mu=8.5, \sigma=4.8$** (8.6%)
Speaking Rate	**$\mu=11.0, \sigma=5.1$** (13.3%)	**$\mu=9.4, \sigma=2.9$** (8.2%)	**$\mu=10.2, \sigma=4.0$** (10.3%)
Emotion	**$\mu=42.0, \sigma=7.0$** (51.0%)	**$\mu=73.6, \sigma=19.4$** (63.8%)	**$\mu=57.9, \sigma=21.6$** (58.6%)
All Keywords	**$\mu=115.2, \sigma=24.3$**	**$\mu=82.4, \sigma=13.1$**	**$\mu=98.8, \sigma=25.3$**

5.2.2 Perception of Vocal Expression in Semi-structured, Unscripted Speech

We followed a similar data analysis process for unscripted speech. We ran the same synonym-reduction process to collect keywords according to close-synonyms, again as defined by the thesaurus. The listeners still described emotion, prosody, and voice quality; but this time, they also described the conversational interaction between speakers in the interviews which we will call "Interactive Quality" here. Examples of conversational interaction quality keywords include "familiar, "conversational," "testimonial," "narrative," "prompting," and "explaining." Listeners also gave descriptors which did not fit any of these categories, which we will call "Other" here. Many of the keywords in the "Other" category ascribed personal qualities to the speakers, just based on the listener's perception of the voice. Examples of keywords in the "Other" category which attributed personal qualities to speakers included "helpful," "expressive," "honest," "believable," and "uneducated."

Tables 5.4 and 5.5 give the frequency of keywords in each of these categories for males and females, respectively. Listeners perceived emotion slightly over half the time (at about 55%), perceived VQ and Prosody at nearly equal rates (about 16% and 17%, respectively) perceived Interactive Quality at about 7%, and perceived Other qualities at about the remaining 4%. Table 5.6 summarizes the raw numbers and percentages of keywords provided in each category for males, females, and all speakers.

We found no significant differences between males and females in the keyword distributions across prosody, voice quality, emotion, conversation quality, and other vocal qualities in unscripted speech. A Chi-Square test for Independence between the descriptor type (prosody, voice quality, emotion, conversation quality, and other) and gender (male, female) categories failed the dependency test (α=0.05, χ^2=16.5, df=2, p=0.00026). Additional Chi-Square tests of each high-level quality similarly reported no statistical dependency between gender and voice quality (α=0.05, χ^2=1.7, df=1, p=0.19), prosody (α=0.05, χ^2=0.17, df=1, p=0.68), emotion (α=0.05, χ^2=3.27, df=1, p=0.07), interaction quality (α=0.05, χ^2=0.34, df=1, p=0. 56), or other quality (α=0.05, χ^2=3.8, df=1, p=0.053). Lower-level Chi-Square tests for independence between gender and effort levels, loudness, and speaking rate also showed no statistical dependency.

Listeners again reported their perceptions of prosody and voice quality using a small number of focused keywords which repeated with high frequency. Within the Prosody category, listeners reported hearing the same pitch, loudness and duration/speaking rate keywords as were reported in scripted speech. Listeners also began to describe speaker articulation and accent, which were not represented in the description of acted speech. Examples of articulation keywords include "accented," "articulated," and "staccato." Within the VQ category, the set of effort level keywords differed between scripted and unscripted speech. Speakers in oral history interviews rarely whispered. In the rare even that they did whisper, they were usually so overcome by emotion that they could not do otherwise. Instead, listeners reported hearing breathy, resonant, and creaky voice. Speakers liberally used creaky voice (which did not appear at all in the scripted corpus); and this speaking style routinely accompanied negative emotion, sarcasm, or description of negative experiences. Because of the focused repetition of these effort levels, or phonation, keywords, and because of their interesting potential relationships with emotions, we again selected them for analysis. Chapter 6 presents the results of the analysis and detection of breathy, modal, resonant, and creaky voice in unscripted speech. As before, we include the "modal" category as a baseline quality for comparison.

Table 5.4: Allocation of keyword descriptors for semi-structured, **unscripted female speech**. This table shows, for each female speaker, the raw numbers and percentages of keywords given which relate perception of vocal expression to voice quality, prosody, emotion, interactive quality, and other quality.

Keyword Class	**Ferretti**	**Kean**	**Lin**	**Little**	**Sluss**
# Clips	**13**	**11**	**32**	**13**	**11**
# Descriptors	**384**	**319**	**334**	**388**	**338**
Voice Quality	**50** (13.0%)	**64** (20.1%)	**60** (18.0%)	**65** (17.8%)	**55** (16.3%)
Effort Level	**8** (2.1%)	**8** (2.5%)	**6** (1.8%)	**17** (4.4%)	**1** (0.3%)
Other Quality	**42** (10.9%)	**56** (17.6%)	**54** (15.0%)	**48** (12.4%)	**54** (16.0%)
Prosody	**70** (18.2%)	**61** (19.1%)	**56** (16.8%)	**65** (16.8%)	**54** (16.0%)
Pitch	**5** (1.3%)	**7** (2.2%)	**5** (1.5%)	**7** (1.8%)	**1** (0.3%)
Loudness	**28** (7.3%)	**19** (6.0%)	**16** (4.8%)	**16** (4.1%)	**16** (4.7%)
Duration/Rate	**32** (8.3%)	**29** (9.1%)	**28** (8.4%)	**40** (10.3%)	**33** (9.8%)
Articulation	**5** (1.3%)	**6** (1.9%)	**7** (2.1%)	**2** (0.5%)	**4** (1.2%)
Emotion	**240** (62.5%)	**150** (47.0%)	**172** (51.5%)	**194** (50.0%)	**193** (57.1%)
Interactive Quality	**15** (3.9%)	**32** (10.0%)	**31** (9.3%)	**47** (12.1%)	**21** (6.2%)
Other	**9** (2.3%)	**12** (3.8%)	**15** (4.5%)	**17** (4.4%)	**15** (4.4%)

Table 5.5: Allocation of keyword descriptors for semi-structured, **unscripted male speech**.

Keyword Class	**Ancona**	**Carpenter**	**Chier**	**Gamblin**	**Hurtt**
# Clips	**11**	**10**	**11**	**11**	**11**
# Descriptors	**328**	**277**	**309**	**310**	**328**
Voice Quality	**39** (11.9%)	**39** (14.1%)	**44** (14.2%)	**59** (19.0%)	**52** (15.9%)
Effort Level	**3** (0.9%)	**6** (2.2%)	**9** (2.9%)	**9** (2.9%)	**11** (3.4%)
Other Quality	**36** (11.0%)	**33** (11.9%)	**35** (11.3%)	**50** (16.1%)	**41** (12.5%)
Prosody	**56** (17.1%)	**40** (14.4%)	**52** (16.8%)	**49** (15.8%)	**64** (19.5%)
Pitch	**5** (1.5%)	**4** (1.4%)	**11** (3.6%)	**12** (3.9%)	**4** (1.2%)
Loudness	**18** (5.5%)	**7** (2.5%)	**20** (6.5%)	**17** (5.5%)	**21** (6.4%)
Duration/Rate	**31** (9.5%)	**24** (8.7%)	**19** (6.2%)	**20** (6.5%)	**38** (11.6%)
Articulation	**2** (0.6%)	**5** (1.8%)	**2** (0.7%)	**0** (0.0%)	**2** (0.6%)
Emotion	**192** (58.5%)	**171** (61.7%)	**172** (55.7%)	**168** (54.2%)	**181** (55.2%)
Interactive Quality	**19** (5.8%)	**21** (7.6%)	**31** (10.0%)	**27** (8.7%)	**22** (6.7%)
Other	**22** (6.7%)	**6** (2.2%)	**10** (3.2%)	**7** (2.3%)	**9** (2.7%)

Table 5.6: Comparison of the distribution of keyword descriptor types across males and females in semi-structured unscripted speech. The results show the means of the numbers of keywords given for voice quality, prosody, and emotion across males, females, and all talkers. Percentages (in parentheses) are given with respect to the total mean number of keywords.

Keyword Class	**Male Talkers**	**Female Talkers**	**All Talkers**
Voice Quality	**μ=46.6** (15.0%), **σ=8.7**	**μ=58.8** (16.7%), **σ=6.3**	**μ=52.7** (15.9%), **σ=9.6**
Effort Level	**μ= 7.6** (2.4%), **σ=3.1**	**μ= 8.0** (2.3%), **σ=5.8**	**μ= 7.8** (2.4%), **σ=4.4**
Other Quality	**μ=39.0** (12.6%), **σ=3.4**	**μ=50.8** (14.4%), **σ=5.8**	**μ= 44.9** (13.9%), **σ=8.6**
Prosody	**μ=52.2** (16.8%), **σ=8.8**	**μ=61.2** (17.4%), **σ=6.5**	**μ=56.7** (17.1%), **σ=8.7**
Pitch	**μ= 7.2** (2.3%), **σ=4.0**	**μ= 5.0** (1.4%), **σ=2.4**	**μ= 6.1** (1.8%), **σ=3.3**
Loudness	**μ= 16.0** (5.2%), **σ=6.0**	**μ=19.0** (5.4%), **σ=5.2**	**μ= 17.8** (5.4%), **σ=5.2**
Speaking Rate	**μ= 26.4** (8.5%), **σ=8.0**	**μ=32.4** (9.2), **σ=4.7**	**μ= 29.4** (8.9%), **σ=7.0**
Articulation	**μ= 2.2** (0.7%), **σ=1.8**	**μ= 4.8** (1.4%), **σ=1.9**	**μ= 3.5** (1.1%), **σ=2.2**
Emotion	**μ=176.8** (57.0%), **σ= 9.8**	**μ=189.8** (53.8%), **σ=33.3**	**μ=183.3** (55.3%),**σ=24.2**
Interactive Quality	**μ= 24.0** (7.7%), **σ= 4.9**	**μ= 29.2** (8.3%), **σ= 4.9**	**μ= 24.0** (7.2%), **σ= 4.9**
Other	**μ= 10.8** (3.5%), **σ= 6.5**	**μ= 13.6** (3.9%), **σ= 3.1**	**μ= 12.2** (3.7%),**σ= 5.0**
All Keywords	**μ=310.4, σ=20.84**	**μ=352.6, σ=31.1**	**μ=331.5, σ=33.5**

The profiles between scripted and unscripted speech have some similarities. In both cases, the frequency of perception of emotion is around 55%. Also, the relative contribution of VQ and

Prosody is nearly equal within scripted speech and within unscripted speech (about 16% and 21% respectively). But the similarities end there. Listeners perceived the conversational element in the interview, even given only the speech of the interviewee to hear. Furthermore, listeners perceived personal qualities in the speakers of unscripted speech, but did not do this for the actors. The actors used whispering to communicate, but the oral history speakers did not whisper. Conversely, the oral history speakers used creaky voice, and the actors did not employ it. Listeners also heard articulation and accent in the unscripted speech, but did not report hearing it in the acted speech.

Chi-Square tests for independence between scripted and unscripted speech showed no dependency between scripted and unscripted speech and either VQ (α=0.05, χ^2=1.31, df=1, p=0.25) or Prosody (α=0.05, χ^2=0, df=1, p=0.00) in females. The same Chi-Square tests did, however, show statistically significant dependencies for female speakers between scripted/unscripted speech and Emotion (α=0.05, χ^2=17.9, df=1, p=0.000024), Effort Level (α=0.05, χ^2=63.5, df=1, p<0.00001), Other Voice Quality (α=0.05, χ^2=11.8, df=1, p=0.00059), Loudness (α=0.05, χ^2=17.06, df=1, p=0.000036), and Speaking rate/Duration (α=0.05, χ^2=6.44, df=1, p=0.011).

For males, Chi-Square tests for independence between scripted and unscripted speech showed significant dependency between scripted and unscripted speech for VQ (α=0.05, χ^2=22.9, df=1, p<0.00001), Prosody (α=0.05, χ^2=11.3, df=1, p=0.00077), Emotion (α=0.05, χ^2=4.73, df=1, p=0.029), Effort Levels (α=0.05, χ^2=113.2, df=1, p<0.00001), Loudness (α=0.05, χ^2=5.45, df=1, p=0.20), and Speaking Rate (α=0.05, χ^2=8.89, df=1, p=0.003). No statistically significant differences were found between speaking styles and Non-effort Level voice qualities (α=0.05, χ^2=3.0, df=1, p=0.084).

Given the statistically-significant differences between perceived scripted and unscripted speech, particularly for the Emotion category, using acted speech for the analysis of emotion, with the assumption that the analysis generalizes to unscripted speech, may not be recommended. The prosodic and voice quality differences reinforce this caution against generalizing results from analysis of scripted/acted speech to the analysis of unscripted speech.

5.2.3 Perception of Laughter in Semi-structured, Unscripted Speech

We were curious to learn how listeners would perceive laughter, and curious whether listeners would perceive laughter and expressive speech in any similar ways, given the differences in the nature of laughter and speech, and the variation in laughter quality and frequency across speakers. The laughter samples presented to the listeners had no words associated with them, just the expressive gestures within the laughter itself; while speech, by definition, was composed of words. Note that the perception studies of speech used a similar number of samples to represent each speaker. Different speakers laughed at different frequencies and with different purpose, and the study included all instances of laughter from each speaker. Table 5.7 shows the number of laughter events in each female speaker's oral history interview ranging from 7 to 56 laughter events. In general, the happiest speakers did not laugh the most. The speaker who reported the most positive experience in the military (Sluss) actually laughed the least. Speakers laughed not only during the telling of humorous anecdotes, but also during the telling of stories of great difficulty; and the quality of the laughter varied with the emotions of the speaker. Chapters 7 and 8 investigate these variations in detail with the purpose of discovering, exploring, and detecting the different dimensions of laughter present in the corpus and the relationship of laughter to general expressive dimensions. Table 5.7 below summarizes listener perception of laughter per speaker.

Table 5.7: Allocation of keyword descriptors for laughter, for female speakers. The total number of laughter events occurring in each speaker and the total number of laughter keywords collected appear in the top two rows. This table shows, for each female speaker, the raw numbers of keywords given which relate perception of laughter to voice quality, prosody, emotion, interactive quality, and other quality. It further subdivides voice qualities into effort levels, laughter types, gender, age, and other voice quality. Also, it subdivides prosody into pitch, loudness, duration/rate, articulation/rhythm, and other.

Keyword Class	**Ferretti**	**Kean**	**Lin**	**Little**	**Sluss**
# Events	**9**	**11**	**32**	**56**	**7**
# Descriptors	**250**	**401**	**1031**	**1625**	**183**
Voice Quality	**82** (32.8%)	**107 (**26.7%)	**338** (32.8%)	**441** (27.1%)	**36** (19.7%)
Effort Level	**41** (16.4%)	**21** (5.2%)	**110** (10.7%)	**145** (27.1%)	**3** (1.6%)
Laughter Quality	**21** (8.4%)	**58** (14.5%)	**141** (13.7%)	**197** (12.1%)	**21** (11.5%)
Gender	**12** (4.8%)	**10** (2.5%)	**19** (1.8%)	**14** (0.9%)	**3** (1.6%)
Age	**3** (1.2%)	**3** (0.8%)	**8** (0.8%)	**3** (0.2%)	**1** (0.6%)
Other Quality	**5** (2.0%)	**15** (3.8%)	**60** (5.8%)	**82** (5.1%)	**8** (4.4%)

Table 5.7: (cont.)

Keyword Class	Ferretti	Kean	Lin	Little	Sluss
Prosody	**59** (23.6%)	**123** (30.7%)	**256** (24.8%)	**475** (29.2%)	**39** (21.3%)
Pitch	**0**	**10** (2.5%)	**29** (2.8%)	**54** (3.3%)	**5** (2.7%)
Loudness	**28** (11.2%)	**43** (10.7%)	**78** (7.6%)	**131** (8.1%)	**8** (4.4%)
Duration/Rate	**22** (8.8%)	**55** (13.7%)	**125** (12.1%)	**218** (13.4%)	**22** (12.0%)
Articulation	**9** (3.6%)	**15** (3.7%)	**21** (2.0%)	**72** (4.4%)	**4** (2.2%)
Other	**0**	**0**	**3** (0.3%)	**0**	**0**
Emotion	**92** (36.8%)	**113** (28.2%)	**329** (31.9%)	**562** (34.6%)	**87** (47.5%)
Interactive Quality	**3** (1.2%)	**25** (6.2%)	**52** (5.0%)	**35** (2.2%)	**2** (1.1%)
Other	**14** (5.6%)	**33** (8.2%)	**57** (5.5%)	**112** (6.9%)	**19** (10.4%)

Laughter, by nature, has a strong prosodic component. It has a duration, and many kinds of laughter also have a measurable rate. Instead of syllables or pitch accents per unit time, some kinds of laughter can be measured in pulses per unit time. It has rhythmic regularity or variation, and it can have pitch and loudness variation across the duration of it as well. In addition, laughter has variation in articulation. It can have a defined (e.g., “Tuh”) or a soft (e.g., “huh”) attack. It even ranges in voice quality also from unvoiced “whisper” laughs to fully-voiced, melodic, resonant utterances. By examining Table 5.7, we see that listeners report hearing some element of prosody about 21-31% of the time. As with speech, pitch is the most infrequently-perceived component of prosody (0-3.3%), followed by articulation (2.0-4.4%). Loudness and duration/rate qualities were perceived at similar rates (4.4-11.2% and 8.8-13.7% respectively). Listeners used a small set of repeating keywords (and their close synonyms) to describe laughter prosody. Examples of keywords given in this category are very similar to words given to describe voices, and they include words such as “fast,” “slow,” “loud,” “high,” and “getting-lower”. Descriptors which differed from speech usually described articulation (e.g., “heh” vs “tuh”) or described the rhythmic elements in laughter (e.g., “pulsing,” “ripples”) which do not typically appear in speech.

Laughter also varies in voice quality. While some laughter instances are boisterous, projected, pulsing outbursts, other kinds of laughter, in contrast, can be short, single-pulse exhalations of air with no voicing. The quality ranges from unvoiced “whisper” laughs to fully-voiced, melodic, resonant utterances. This qualitative variation is similar to variation in phonation in speech, and listeners reported hearing it using terms similar to the descriptors for effort levels in speech (e.g., “breathy,” “whispery,” “strong,” “air,” “exhale,” “airy”). Effort-level descriptors which appeared for laughter but which did not appear in speech usually described inhalation, exhalation, or “air.” Listeners gave effort-level descriptors between 1-27% of the time, as shown in Table 5.7. Listeners rarely commented on the age or gender of a speaker, but did comment on the age and gender of a person laughing about 1-5% of the time. A marked difference between voice quality perception in speech and laughter is the presence of laughter descriptors, such as “chuckle,” “huff,” or “giggle,” which listeners reported hearing about 8-15% of the time. As with speech, listeners also reported hearing other kinds of vocal quality, such as “gruff” and “rough”. Listeners again used a small number of descriptors (and their close synonyms) to describe voice quality in laughter. This characteristic repetition of a small number of distinct, high-frequency keywords suggests that these qualities would benefit from focused analysis. Listeners perceive these qualities frequently. Chapters 6 and 8 explore a subset of these qualities in detail.

Listeners’ perceptions of emotion in laughter was similar to perception of emotion in the voice. Listeners gave a wide range of nuanced descriptors of emotion that again exceeded the boundaries of emotions considered basic by common theories. Even within the restricted realm of basic emotions, listeners perceived nuance. They rarely just described laughter as “happy.” Instead, they heard “giddy,” “boisterous,” “joyous,” or “bubbly,” each of which is a unique nuance of “happy.” The range and subtlety of emotions perceived in laughter suggest that different analytic techniques will be required to explore them. These techniques will be described further in Chapters 7 and 8.

Since an interview is a conversation, a small percentage of keywords (about 1-6%) described the conversational or interactive elements of the exchange. Some of these keywords described mutual laughter, agreement, or response. A final group of descriptors (about 5-10%) were outside the bounds of the previously mentioned categories. Many of them described a person’s presumed personal qualities, for example, “weird,” “free,” or “honest.” These descriptors were widely-ranging and were not nuanced variations on similar ideas.

The perception of laughter is distinctly different from that of speech. For females, a Chi-Square tests for independence between the perception of unscripted speech and laughter showed significant difference for all the high-level categories, including VQ (α=0.05, χ^2=91.9, df=1, p<0.00001) , Prosody (α=0.05, χ^2=63.2, df=1, p<0.00001), Emotion (α=0.05, χ^2=193.2, df=1, p<0.00001), Interaction Qualities (α=0.05, χ^2=46.0, df=1, p<0.00001), and Other Qualities

(α=0.05, χ^2=17.8, df=1, p<0.000024). The Chi-Square test showed differences in the sub-categories of prosody, including Pitch (α=0.05, χ^2=9.9, df=1, p<0.002), Loudness (α=0.05, χ^2=14.2, df=1, p<0.0002), and Duration (α=0.05, χ^2=13.9, df=1, p<0.0002), and the sub-categories of VQ, including Effort Levels (α=0.05, χ^2=87.3, df=1, p<0.00001) and Other Voice Qualities (α=0.05, χ^2=17.8, df=1, p<0.000024). These differences are extremely statistically significant, and suggest caution against using the same features and analytic techniques for analyzing speech and laughter.

5.3 Perception Study Summary

In this chapter, we presented three Mechanical Turk studies of the perception of vocal expression in scripted and unscripted speech. In these studies, listeners provided open-response keywords describing what they heard, expressively speaking, in audio clips. The data naturally clustered into top-level categories of Emotion, Prosody, and VQ, sub-categories of Prosody (pitch loudness, and duration/rate), and sub-categories of VQ (Effort Levels and other voice qualities). In all cases, listeners described emotion most often. For speech, listeners heard emotion about 55% of the time, and for laughter, about 35% of the time. In all studies, the proportion of Prosody and VQ was nearly equal (about 37.5% for laughter, 16.5% for unscripted speech, and about 20.5% for scripted speech). In the oral history (unscripted speech and laughter) studies, listeners reported human interaction qualities as well, and attributed personal qualities (e.g., “honesty”) to the speakers.

Emotion descriptors were wide-ranging and nuanced, while VQ and Prosodic perception focused on a very narrow range of high-frequency descriptors. This basic difference suggested that fundamentally different analytic techniques were required which allowed focus on just a few qualities for VQ and prosody, but leveraged the wide variances and nuances of Emotion perception. These techniques are described in detail in chapters 6, 7, and 8.

Listeners consistently reported a small number of effort levels in each study, and these qualities will be the focus of detailed analysis in the next chapter (Chapter 6). In scripted speech they are whispering, breathiness, modal speech, and resonance; while in unscripted speech they include breathiness, modal speech, resonance, and creaky voice. Laughter covers the superset of both kinds of speech, and adds inhalation and exhalation to the list of qualities.

Analysis of the perception profiles across scripted speech, unscripted speech, and laughter revealed statistically significant differences, many extremely significant, which suggest caution in extending analytic results from one realm into another. For example, the results of a study of emotion using acted speech may not apply to unscripted speech. Different features and models may be needed to recognize and represent each of these different conditions.

These studies revealed statistically-significant differences between the perception of male and female expressiveness in acted speech. This result requires further study to determine how much of the difference is a result of the difference in the acted parts (even though they had a similar theme and were of the same acting style). The studies also revealed no statistically significant differences in the perception profiles of male and female unscripted expressive speech.

Finally, the studies revealed the most commonly-perceived keywords within each speaker and across the categories of scripted speech, unscripted speech, and laughter. Frequently-appearing keywords may warrant further study in the future. In addition, listeners revealed a list of frequently-perceived categories just for laughter (e.g., “titter,” “chuckle,” and “giggle” to name a few) which may also warrant focused future study.

These study results have implications for future applications work. As a general statement, applications are meant to serve people. In order to serve people, they should operate in the realm of conscious human perception. Consider a sonic search application. Humans are likely to search for things in an audio recording which they remember and perceive (not for things which they cannot describe or articulate), and this work has revealed a number of expressive elements which listeners consistently perceive and describe in both scripted and unscripted speech. Above and beyond all else, these perception studies show that listeners relate to and describe emotion in a nuanced manner. They perceive how someone feels and associate the perceived emotions of other speakers with their own emotional experience. Search should support nuanced emotion, and go beyond the small number of basic emotions typically explored in the literature.

Furthermore, listeners perceive the prosodic elements of duration, speaking rate, and loudness strongly. Search should support these elements directly, with respect to the current speaker and classes of speakers in general. Listeners do not directly articulate hearing pitch variation very often, unless the variation is extreme, or the individual is sensitive to sound. They do perceive pitch changes, but more often interpret these signals within the realm of emotion or emphatic difference (such as with contrastive focus). This suggests the importance of exploring and understanding the relationships and co-occurrences among prosody emotion, voice quality, nonvoiced quality, conversational quality, perceived personal quality, and other expressive elements. Prosodic qualities are well-defined compared to many emotions, and may be useful in the recognition of emotion, given an improved understanding of the relationships among the classes of expressive elements in the voice.

Listeners also clearly hear the vocal qualities of effort levels, or phonation types, in the voice. A voice search application should, therefore, support finding whispering, breathiness, conversational speech, resonance, and creakiness at a bare minimum. Again, discovering the relationships of these voice qualities with emotion, prosody, conversational quality, and perceived personal quality is intriguing. It may be possible to leverage voice quality in the recognition of emotion, for example, if the relationships between voice quality and emotion are understood.

Listeners clearly articulate their perceptions of nonverbal quality, such as laughter, expressive inhalation, sighs, groans, "um" "ah" filler, and silence. These elements, too, have relationships with emotions, voice quality, prosody, etc. which should be understood and exploited for the detection of nuanced emotion and in the context of search applications. Laughter is especially intriguing, since it is not always reflective of happiness and humor.

Future work could also analyze the speaker dialect to discern a dialect's influence on rating of emotion, VQ, NQ, prosody, conversational quality, and personal quality. This could be combined with a study on listener demographic as well, to determine how the dialect of the listener influences perception.

This study has also demonstrated the potential of the voice to telegraph mental well-being. Depression, anxiety, and even post-traumatic stress disorder (PTSD) are potentially reflected in the voice, and applications could be developed which detect these conditions. While these applications would be useful in a doctor's office, they would be even more useful deployed on mobile devices and used in situations where people cannot present themselves physically to a doctor, such as in telemedicine, or in field conditions. An application capable of detecting the potential presence of depression, anxiety, and PTSD could also be used as part of the military discharge process. If an application could catch even a percentage of people with these conditions, and refer them for evaluation and treatment, the impact could be measured in prevented suicides and in thousands fewer cases of homeless veterans. If we can intervene and prevent suffering for people who have served our country, especially in a non-invasive and discreet way, for those who wish it, we have a moral responsibility to do so.

CHAPTER 6: DETECTION OF PERCEIVED VOICE QUALITIES IN SPEECH

This chapter describes the methods used to address research questions RQ4, RQ6, RQ9, and RQ11, along with the experimental results. Specifically, it addresses the process for investigating whispering, breathiness, modal voice, resonance, and creaky voice as appropriate for scripted and unscripted speech. As discussed in the previous chapter, listeners consistently and concisely reported hearing these qualities. A direct, in-depth study of these features, therefore, serves the perception-grounded analytics goals which potentially enable application development by supporting features which listeners hear. We investigate the characteristics of each of these qualities, explore features for their potential ability to model the distinguishing characteristics of each effort level, explore combinations of these features which produce superior results for n-way and binary classifiers, and present a process for investigating and modeling perceived qualities in speech. This process describes 1) ***Interactive Intuitive Analysis*** (evaluation of the spectra and waveforms across conditions for distinguishing characteristics), 2) ***Candidate Feature Selection and Evaluation*** (justification of the selection of candidate features, and analysis of the ability of each candidate feature to separate across conditions), and 3) ***Iterative Feature Group Selection, Model Building and Validation*** (selection of candidate feature groups and performance evaluation within the models). The results suggest a continuum relationship across whispered, breathy, modal, and resonant speech. A summary of the methods and results concludes the chapter. Note that much of this work has been published (Pietrowicz et al., 2015; Pietrowicz et al. 2, 2017).

6.1 Methods

As described in Chapter 4, the curated corpora provided ground-truth-labelled vowels extracted from the Hamlet and Lady Macbeth soliloquy (scripted speech), and ground-truth-labelled vowels extracted from the Veterans Oral History Project at the Library of Congress (unscripted speech). Refer to Chapter 4 for the details of procuring corpora, extracting vowels, labelling each vowel sample for effort level type, and measuring inter-rater agreement.

The effort level analysis process had three steps: 1) ***Interactive Intuitive Analysis***, 2) ***Candidate Feature Selection and Evaluation***, and 3) ***Iterative Feature Group Selection, Model Building and Validation***. The goals of ***interactive intuitive analysis*** were determining the distinguishing characteristics of each effort level condition and understanding the characteristics which best distinguished one condition from another. This meant examining waveforms and spectra samples from each condition and understanding vocal tract production for each phonation type. Each condition had a characteristic spectral profile, with characteristic patterns of energy across the frequency spectrum, and characteristic periodicity patterns. For example, whispering produced seemingly random energy across the spectral region of 50-8000 Hz. In contrast to whispering, the other effort levels exhibited periodicity with distinct patterns of energy expected in specific sub-bands within the 50-8000 Hz range. Some phonation types had multiple spectral profiles to consider, and signature patterns from each sub-type were considered and compared to the average spectral profile for the given condition. After examining each condition in detail, we identified spectral bands of interest which could distinguish one effort level from another. Finally, agreement patterns from ground truth labels were considered to understand how human listeners made distinctions across effort levels. If listeners never confused two effort level types (eg. whispering and resonance), the two types were unlikely to be closely related, unlikely to share common characteristics, and unlikely to reside on a continuum in an adjacent relationship. In contrast, if two effort level types (e.g., whispering and breathiness) were consistently confused, the two effort level types were more likely to share some common characteristics, and might possibly have an adjacent continuum relationship.

Given an understanding of the distinguishing characteristics for each effort level type, the spectral differences among them, the spectral bands of interest, periodicity patterns, and the patterns of perceptual agreement and confusion, we looked for features which would be able to measure and identify the distinguishing characteristics (***Candidate Feature Selection and Evaluation***). These features had potential to be, individually or collectively, the distinguishing features across effort levels. Each candidate feature was defined, and its method of measurement, also defined. To ensure that the features were calculated as specified, the analytic software was either developed from scratch in Matlab, or the source procured and examined for consistency with the desired feature definition. During this candidate feature selection phase, resisting the urge to adopt an available package, with the usual collection of pitch tracking, energy, spectral, and voice quality features (because "most people use these"), was critical. Each feature's ability to distinguish across each effort level was measured via 1) the unequal variance sensitivity d_a from signal detection theory (Pashler and Wixted, eds., 2002), and 2) analysis of each feature's mean and 2-sigma variance across each condition.

Finally, given a list of potential distinguishing features, their clear definitions, and an understanding of their purpose within a model for recognizing effort levels, I proposed feature collections most likely to produce successful 4-way effort level classifiers (***Iterative Feature Group Selection, Model Building and Validation)***. The goals for defining feature collections were maximizing distinguishing features, minimizing potential confusion among features, minimizing duplication of similar features within the model, and maintaining simplicity when possible. Given a set of feature combinations which were most likely to distinguish across conditions, 4-way classifiers were trained and evaluated via 4-way cross-validation. The smallest feature set was evaluated first, and based on the resulting model performance, features were added or removed to fine-tune performance. Each candidate feature set was iteratively evaluated in this way, and the best-performing feature sets and classifier model performance statistics are presented here. The best-performing feature sets were also evaluated for their binary classification ability.

6.2 Scripted Speech

6.2.1 Intuitive Interactive Analysis in Scripted Speech

To begin analysis of effort levels, samples of each of the four effort levels were collected from the Hamlet and Lady Macbeth corpora, and examined to learn what each condition might look like in the context of acted, expressive speech. Representative examples of the spectra for each condition are shown in Figure 6.1; the spectral patterns shown in these examples repeat across the corpus. Overall, the female speech had more variance within each condition than the male speech. Figure 6.1 shows examples of this variation by showing two representative examples of each condition for the female speakers.

Typical male whispered speech lacked a strong spectral component where F0 would be, and appeared noise-like and aperiodic, with many high-frequency components, and formants. It was usually softer than modal or resonant speech, but not always. One of the typical female whispered speech patterns was similar, but it had more high-frequency energy overall; and the female formants (except F1) were higher than the corresponding male formants. If, however, female speech had significant sub-F0 energy, a significant component around F0, low or no periodicity and a relatively low degree of high-frequency energy, listeners also perceived the speech sample as whispered. Noise patterns, with strong sub-F0 components, trumped the presence of F0 in perception.

Typical male breathy speech had a strong component at F0 (around 100 Hz), with usually one or two weaker components at integer multiples of F0, and lacked strong components at formant frequencies. Again, one of the typical female patterns for breathy speech appeared similar, except that F0 (around 180 Hz) and its multiples were higher, as expected for female speech. As Figure 6.1 shows, low levels of aperiodic signal energy did not disrupt the perception of breathy voice, as long as the periodic energy was significantly higher, and as long as significant sub-F0 energy was not present.

Modal speech for males typically had a strong F0 presence, with several components at integer multiples of F0, and frequently a strong F1. Some of the female patterns were similar, with F0 and its multiples at higher frequencies, and sometimes a significant F1 presence. Interestingly, listeners perceived speech to be modal with a small number of F0 multiples as long as a strong F1 was present. In general, with this second female pattern, when the F0 component is the strongest, with very few harmonics, listeners hear breathy voice. In contrast, if F0 is present with very few harmonics and F0 is not the strongest component (instead, a strong F1), listeners hear modal voice.

Male resonant speech has more harmonics than modal speech, stronger formants, and more energy overall at higher multiples of F0 than at F0. Female resonant speech is a bit different, because overall, female speech tends to have fewer harmonics which die out more rapidly than male speech. Figure 6.1 shows that the number of harmonics for female modal and resonant speech can be similar, but in this kind of female resonant speech, a strong F1 is present (but not present in modal speech). The strong presence of this formant is so important to the perception of resonant speech in females, that even speech with extremely weak F0 and minimal harmonics will still be perceived as resonant if it has a strong F1 component.

From these empirical observations, male speech appears to have significant differences in the proportions of energy in the signal across conditions in the following bands: 1) 0-300 Hz: F0, or speaking pitch, 2) 300-700 Hz: Harmonic multiples & F1, 3) 600-900 Hz: Higher harmonic multiples & F1, 4) 1000-2000 Hz: F2, 5) 2000-4500 Hz: High harmonics, higher formants, and noise. To summarize, aperiodicity marks whispered speech, along with the lack of strong F0 in band 1, and formant-aligned energy above 900 Hz in bands 4 and 5. The other conditions were periodic. Strong energy in band 1 marks breathy voice, with very low energy in the higher bands. Strong energy in band 1, moderate energy in band 2, weak energy in band 3, and very weak energy in bands 4 and 5 describe modal voice; and moderate energy in band 1, moderate to strong energy in band 2, strong energy in bands 3 and 4, and weak energy in band 4 characterize resonant voice.

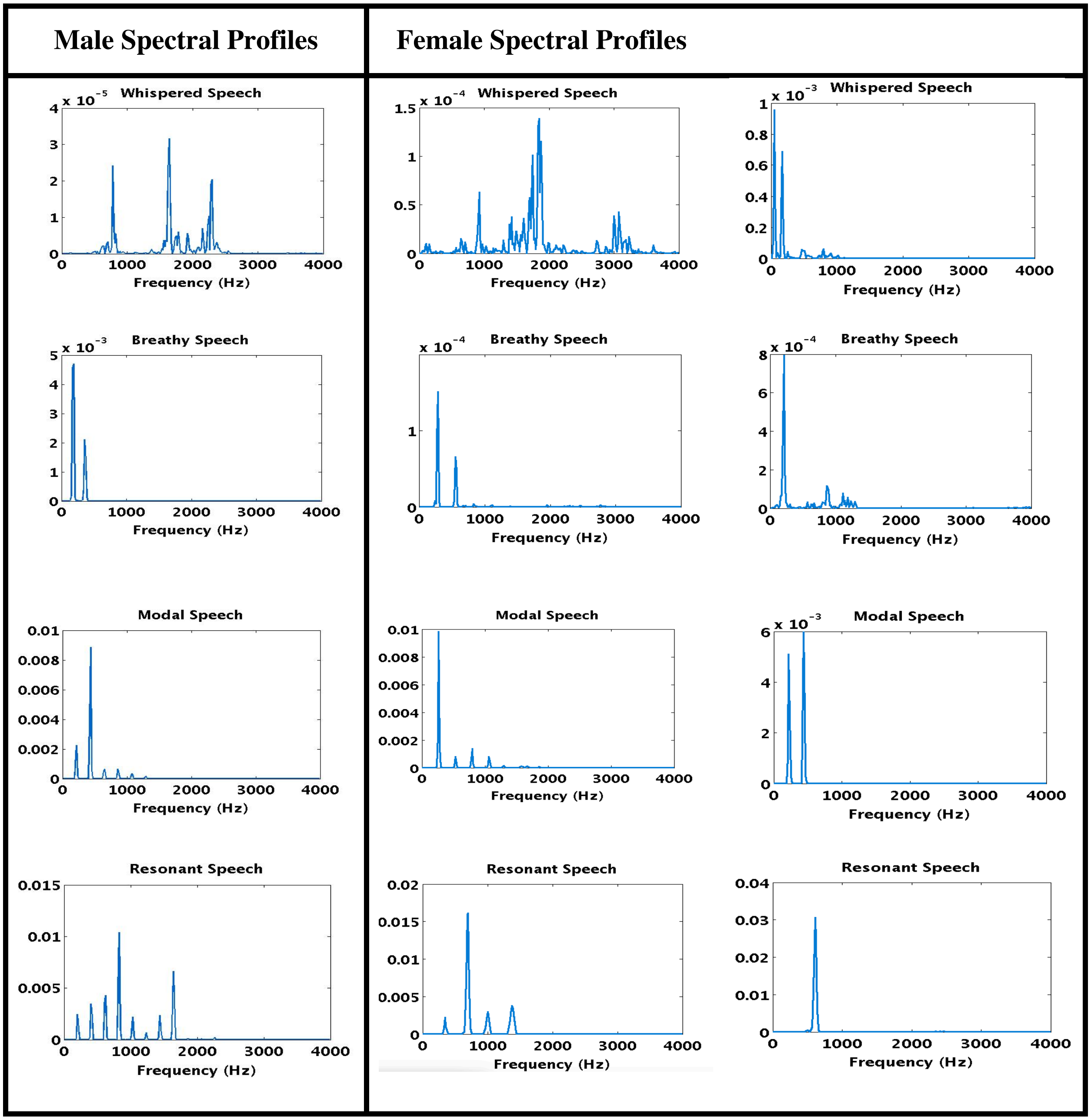


Figure 6.1: Comparison of Male and Female Spectral Profiles Across Effort Levels. These spectra are typical samples of whispered, breathy, modal, and resonant speech. Female speech had more variation, so we show two variants per condition here. The profiles show many similarities between male and female conditions, and the critical differences. They also begin to provide insight in human perception at the condition boundaries. For example, note the similarity between the breathy and modal female profiles, especially the difference between the first female breathy profile and the first modal profile.

The female voice profile is different, with differences in conditions across the following bands: 1) 0-150 Hz: sub-F0 energy, 2) 0-300 Hz: sub-F0 energy and F0, or speaking pitch, 3) 300-800 Hz: harmonic multiples, and F1, 4) 500-1500 Hz: Higher harmonic multiples, and F1, 5) 1000-2000 Hz: even higher harmonics, and F2, 6) 2000-4500 Hz, high harmonics, higher formants, and noise, and 7) 300-4500: all harmonic multiples, all formants, and noise. Whispering is aperiodic, with the lack of strong F0 again in band 2, possible presence of energy in band 1, weak energy in band 3, strong energy in band 4, strong energy in band 5, moderate energy in band 6, and energy distributed all across band 7. Breathy female voice has a very weak band 1, strong band 2, strong band 3, and weak bands 4, 5, and 6. Modal female voice has a very weak band 1, strong band 2, moderate to strong band 3, and weak to moderate band 4. Resonant voice has greater similarty to modal voice in females than in males, with a very weak band 1, moderate band 2, moderate to strong band 3, and moderate to strong band 4, and weak bands 5 and 6. Band 4 is critical for distinguishing resonant from modal voices in females.

6.2.2 Acoustic Feature Selection & Analysis in Scripted Speech

The features described in this section were selected for detailed analysis based on prior work, empirical observation of the effort level condition spectral properties (discussed in section 6.1.1), and computational efficiency. All features except LFSD (see below) were analyzed using a 60 msec time window with a 15 msec frame advance. LFSD required a smaller 10 msec frame. Feature descriptions, reasons for considering a feature for detailed analysis, and the analytic results of each feature's ability to provide separation across conditions follow. Specific measurements presented here which indicate a feature's ability to provide separation include 1) the unequal variance sensitivity d_a from signal detection theory (Pashler and Wixted, eds., 2002), and 2) analysis of each feature's mean and 2-sigma variance across each condition. A stronger d_a magnitude for a given feature and condition indicates that the feature has a strong ability to separate that condition. A wider separation of a feature's means across conditions with minimal overlap within the 2-sigma variances also indicates strong ability of a feature to separate between conditions. The results of this analysis were used to guide the selection of candidate feature combinations for analysis within classifier models described in section 6.2.3 below.

Zero Crossing Rate (ZCR): This feature gives the rate in which a signal in the time domain changes sign (positive to negative and vice versa). It is included primarily for the detection of whispered voice, because of its prior use in voice activity detection (Campbell and Tremain, 1986). Intuitively, ZCR will be higher for whispered voice because of the greater number of high-frequency components and lack of high-amplitude low-frequency components as compared with voiced speech.

Autocorrelation (AC): Autocorrelation is the cross-correlation of a signal with itself at different delay times, typically examined between about 3.3-16.7 msec, which corresponds to 60-300 Hz, the expected F0 range for adult speech (Atal, 1962). We used 60 msec long signal time windows in our analysis. The maximum value of the magnitude of the autocorrelation in this range usually corresponds to F0, and provides a measure of signal periodicity. Higher values indicate a higher degree of periodicity in the signal. Intuitively, small values are expected for whispered voices (which are noisy and aperiodic), and increasingly larger values are expected to follow the continuum to resonant voice (which is periodic and typically contains many strong harmonics at regular intervals).

Log Low Frequency Spectral Density (LFSD): LFSD is the spectral density at frequencies around the glottal resonance, below the first resonance in the vocal tract (Gowda and Kurimo, 2013). Increases in low frequency energy can occur in voices which have a higher open quotient, as breathy or whispered voices do in comparison with modal and resonant voices. It is included to provide separation between breathy and modal conditions, and as a secondary separator across all conditions.

Number of Spectral Peaks (#peaks): The number of spectral peaks reflects the number of well-defined frequency components in the voice, and is included as a primary separator for the whispered condition (which, by empirical observation, contains many frequencies compared to other conditions), and a secondary separator for the breathy condition, which contains a noisy component, and typically fewer frequencies than whispering, but more frequencies than modal speech. The peak count disregarded frequencies which were below 0.5% of the maximum peak, clustered groups of adjacent frequencies, extracted the maximum value from each cluster, and counted the number of remaining peaks after this pruning and clustering. Peak count calculations were done using the squared magnitude of the signal FFT (linear scale).

H1-H2 (H1-H2): The H1-H2 feature is the difference between the first two harmonics in the voice, and is included as a separator between breathy and modal conditions, as suggested by prior work (Hanson, 1995; Wayland and Jongman, 2003).

Entropy (H-): Spectral entropy is a measure of disorder in a spectrum (Zhang, 2012). It measures how noise-like vs. how tone-like the voice quality is. Intuitively, the whispered condition has a high degree of entropy, because it is noisy. Breathy voice still has a noisy component, but it also has a fundamental frequency and weak harmonics. Modal voice does not have the noisy component, and contains more harmonics, which are stronger. Resonant voice may have even more harmonics, with more energy in the higher frequencies. Intuitively, the overall entropy decreases along the continuum from whispered through resonant, and therefore can be a good separator across conditions.

Prior work (Zhang, 2012) used entropy in two bands to separate whispered vs. non-whispered speech, and also noted its stability across recording conditions. When entropy is examined within selected frequency bands, it reflects qualities such as presence or absence of harmonics and noise, regularity of the spectrum in power and frequency interval, and presence and character of formants. Entropy summarizes the overall spectral character. Frequency bands were selected based on interactive analysis of the spectrum across speaker and across condition, as discussed in Sections VI and VII, yielding, the following bands of interest for both males and females (and for which entropy was measured): 1) 50-300Hz, 2) 300-800Hz, 3) 1000-2000Hz, 4) 300-4500Hz, and 5) 4500-8000Hz. In females, the 50-150Hz, 500-1500Hz, and 2000-4000Hz bands were also distinctive, as were the 600-900Hz, 300-1000Hz, and 2000-4500Hz bands in males. The 0-150 band captured sub-F0 energy in female voices (helpful for separating whisper and breathiness); while the 50-300 range covered F0 and sub-F0 energy for both males and females. The 300-800 band for both genders and the 600-800 band in males reflects harmonic multiples and F1, and is critical for separating modal and resonant voice, especially in males. The 1000-2000 and 500-1500 bands detect higher harmonic multiples and F1, and capture difference in harmonic behavior across conditions. The 2000-4000 band captures high harmonics, higher formants, and noise (especially useful in detecting the whispered and resonant conditions); while the 300-4500 range reflects the character of most significant harmonics and formants. In short, the collection of entropy measurements work together to characterize the spectrum for best separation across conditions. Table 6.1 summarizes the frequency bands for entropy in males and females.

Table 6.1: Summary of entropy features for males and females. Frequency ranges are in Hz.

	H1	H2	H3	H4	H5	H6	H7	H8
Male	50 – 300	300 - 800	600 - 900	1000-2000	2000 – 4500	300 – 1000	300 – 4500	4500 – 8000
Female	50 – 150	50 – 300	300 – 800	500 – 1500	1000 – 2000	2000 – 4000	300 – 4500	4500 – 8000

Zhang also noted the stability of entropy to varying recording conditions and amplitude levels, which is noted here, along with robustness to a wide range of expressive speech and speakers. Variance, difference, aperiodicity, and frequent extremes characterize expressive speech; and entropy, by definition, captures these qualities.

Entropy Ratio (HR-): Entropy ratios enhance separation between conditions by comparing the character of two frequency bands, as Zhang (Zhang, 2012) noted in his work on whisper detection. In our work, breathy voice in males has an organized spectrum around F0, but very weak harmonics in the range 400-600 Hz. Modal voice, in contrast, has organized spectra in both bands. Furthermore, modal harmonics in the range 400-600 Hz are stronger than resonant voice harmonics are in that range. Whispering has aperiodic spectra in both bands. Therefore, the entropy ratio (50-300Hz) / (400-600Hz) provides potential separation across all conditions in males, particularly between the breathy and modal conditions. The ratio (50-600Hz) / (400-600Hz) was examined for similar reasons in males; and the ratio (50-300Hz) / (2000-8000) provided enhanced separation in females. Table 6.2 summarizes the entropy ratio relationships explored in males and females.

Table 6.2: Summary of entropy ratio features for males and females. Frequency ranges are in Hz.

	HR1	HR2	HR3	HR4	HR5	HR6
Male	(50-600) / (400-600)	(50-300) / (400-600)	(50-300) / (2000-8000)	(450-650) / (2800-3000)	(50-900) / (300-900)	(50-300) / (50-900)
Female	(50-300) / (50-150)	(50-500) / (500-1000)	(300-800) / (50-300)	(50-500) / (500-1500)	(50-300) / (2000-8000)	(450-650) / (2800-3000)

Power Ratio (PR-): Power measurements alone depend on recorded amplitudes, but power ratios look instead at the relationships among bands, which differ across conditions and provide secondary separation. For males, the ratio (50-900Hz) / (300-900Hz) provides additional separation across all conditions (but particularly the modal case) by examining the relative strength of the combined F0, F1 (in some cases), low harmonics, and noise with the combined F1, low harmonics, and noise. In females the ratios (50-300Hz)/(50-150Hz) and (50-500Hz)/(50-1000Hz) provide similar secondary separation (particularly for the breathy case). Table 6.3 summarizes the power ratios explored in males and females.

Table 6.3: Summary of power ratio features for males and females. Frequency ranges are in Hz.

	PR1	PR2	PR3
Males	(50-900)/(300-900)	(50-300)/(300-900)	(50-300)/(50-900)
Females	(50-300)/(50-100)	(50-500)/(50-1000)	(300-800)/(50-300)

Vowel Duration: Vowel duration can indicate speech rate, and unvoiced speech is often slower than voiced speech. If speech rate is related to voicing, then this feature could potentially provide separation across conditions.

Statistical Measures: In signal detection theory (Pashler and Wixted, eds., 2002), the detection process is continuous and assumes that signal is present along with noise. If the noise mean and standard deviation are μ_n and σ_n, and the signal (present with noise) mean and standard deviation are μ_s and σ_s respectively, the unequal-variance sensitivity d_a can be calculated by:

$$d_a = \frac{(\mu_s - \mu_n)}{\sqrt{\frac{(\sigma_s^2 + \sigma_n^2)}{2}}} \tag{6.1}$$

In this case, μ_s represents the mean value of a given feature within a given condition, such as the ZCR mean for whispered data. The μ_n value is the mean average across other conditions, for this example, the mean of ZCR across the breathy, modal, and resonant data. σ_s^2 is the variance of ZCR for the given condition (whisper in this example), and σ_n^2 is the combined variance of the other conditions (breathy, modal, and resonant). The combined variance can be approximated by

$$\sigma_n^2 = E(n^2) - \mu_n^2 \tag{6.2}$$

$$\sigma_n^2 = p_1\sigma_1^2 + p_2\sigma_2^2 + p_3\sigma_3^2 + p_1\mu_1^2 + p_2\mu_2^2 + p_3\mu_3^2 - \left(\frac{\mu_1 + \mu_2 + \mu_3}{3}\right)^2 \tag{6.3}$$

Where p_1, p_2 , and p_3 are the probabilities of the three other conditions (in this example, the probabilities of the breathy, modal, and resonant conditions, respectively). For the test dataset, p_1, p_2, and p_3 are equally probable. Intuitively, the magnitude of d_a for a given feature and condition indicates how easily that feature can distinguish the condition in question from the other conditions. A larger magnitude indicates a higher sensitivity, or ease of detection.

Figures 6.2 and 6.3 show the error bar and sensitivity plots, respectively, for male and female acted voices. The error bars show the means and 2-sigma variances within a feature, across conditions. The sensitivity plots do not show means and variances directly, but instead provide a quantifiable measure of the ability of a feature to distinguish each condition. A feature does a good job distinguishing a condition if the sensitivity magnitude for the condition is large, and the feature's mean and 2-sigma variance range for that condition has minimal overlap with other conditions. The plots show that for females whispering is the most easily separated case. ZCR, AC, H3, H4, H7, and PR1 provide good separation for whispering in females; and LFSD, #peaks, and the remaining entropy features provide secondary separation.

The plots also show that breathiness is the most difficult condition to separate in females, with ZCR providing the most single-feature breathiness separation. The entropy features work together to provide separation across all conditions, including breathiness. The strongest separators for female modal speech were ZCR, AC, and H7; while the strongest separators for resonant speech were AC and PR1. Note that the entropy features outperformed the TILT and H2-H1 features proposed in prior work. The TILT feature did reflect some changes in condition within speaker, but performed poorly as a feature across speakers. This finding raises questions about the un-normalized application of this kind of feature across a set of voices with significant variance across speakers (typical of expressive speech), and it may also be reflective of the variance in recording conditions across speakers.

As with females, ZCR and AC are the strongest separators for whispering in male voices, with many entropy features also providing good separation for whispering. In contrast to females, modal voice was the most difficult condition to separate in male voices, and ZCR again provided the strongest separator for the weakest condition. Resonant voice was easier to detect in male voices than in female voices, as shown by the stronger overall sensitivity of features to male resonance, perhaps because female voices have about half as many harmonics as male voices. AC provided the strongest resonant separation, with many entropy features also providing strong separation. Again, the combined entropy, entropy ratio, and power ratio features, with band boundaries selected to fit female voices, provided potentially easier separation compared to the spectral tilt and H2-H1 features. Vowel duration was not a reliable separator for any condition.

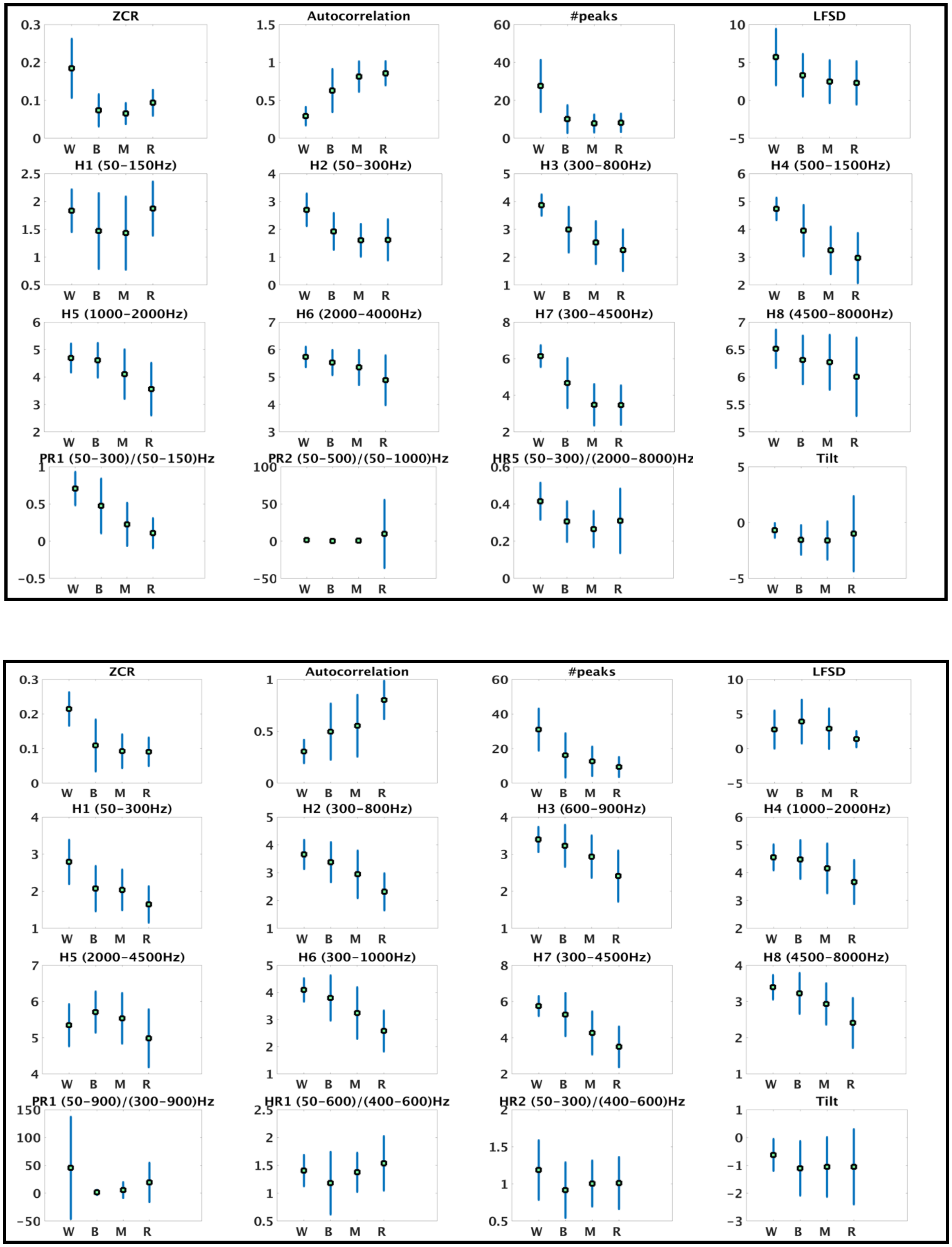


Figure 6.2: Error Bar Plots for Female (top) and Male (bottom) Acted Voices. These diagrams show the mean and 2 sigma ranges around the means for selected features across the continuum of whispered, breathy, modal, and resonant speech. ***For female voices (top):*** ZCR, autocorrelation,

Figure 6.2 (cont.): and #peaks show clean separation between whispering and the other conditions. The entropy features (H1-H8) work together to provide separation, and show the nature of effort levels as a continuum. Power and entropy ratios reinforce general separation and boost detection of specific conditions (e.g., PR2's separation of resonance by its wide variance compared to the other conditions' near zero variance). ***For male voices (bottom):*** These plots show many of the same trends present in female voices, with bands adjusted for differences in male voices (note the different frequency ranges on the Entropy features H1-H8, and on the Entropy Ratios). LFSD differences between male and female speech are apparent, and match empirical observations of greater sub-F0 energy in female voices. Observed differences in resonance show up as subtle differences in the Autocorrelation, ZCR, and Entropy features, and suggest that resonance is easier to detect in male than female voices. Most of the features are monotonically increasing or decreasing across the conditions for both males and females, which reveals the nature of the conditions as a continuum instead of discrete states.

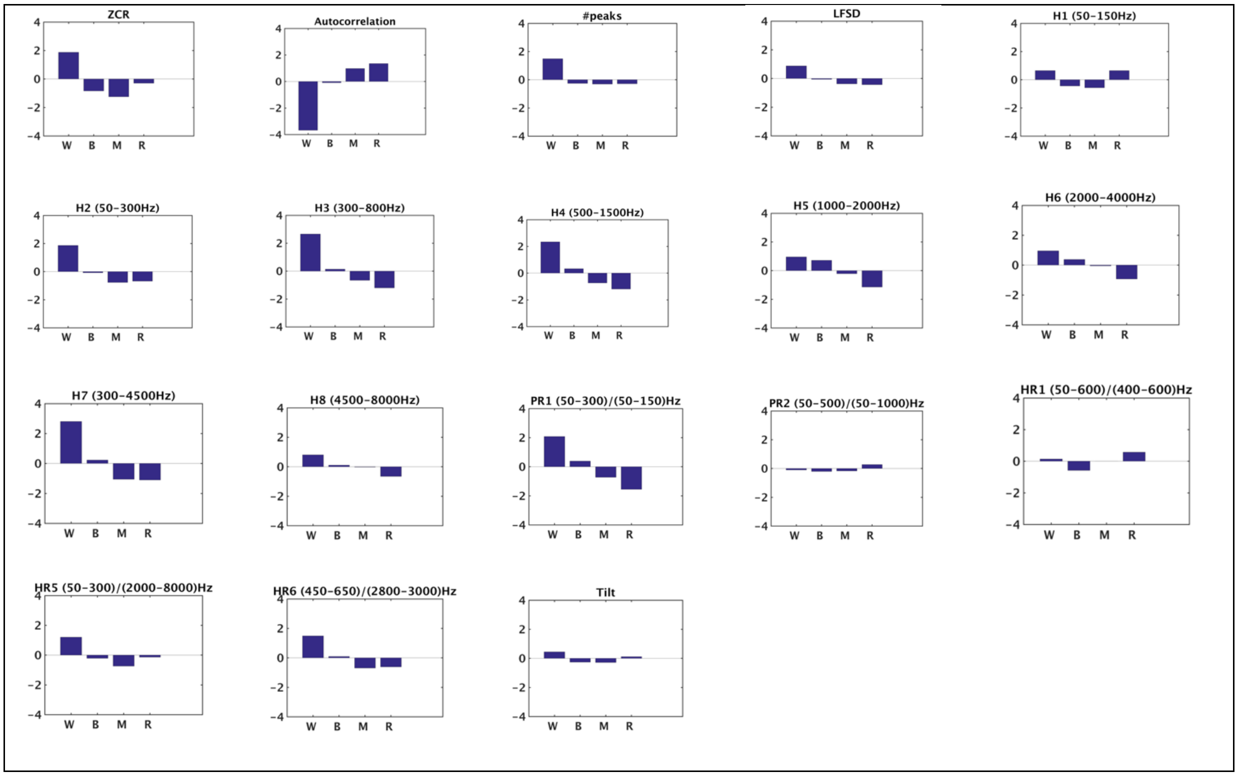


Figure 6.3: Sensitivity Plots for Female Acted Voices. These plots show d_a, the unequal-variance sensitivity, across conditions for a selection of features. A larger magnitude of a bar shows a greater ability of a feature to provide separation for that condition. Many of the features provide strong separation for the whisper case, and smaller separation power for the other features. The sensitivity measurements suggest that breathiness is the most difficult condition to separate from the other three in females, and that the features (particularly entropy) work together to provide separation for the other conditions.

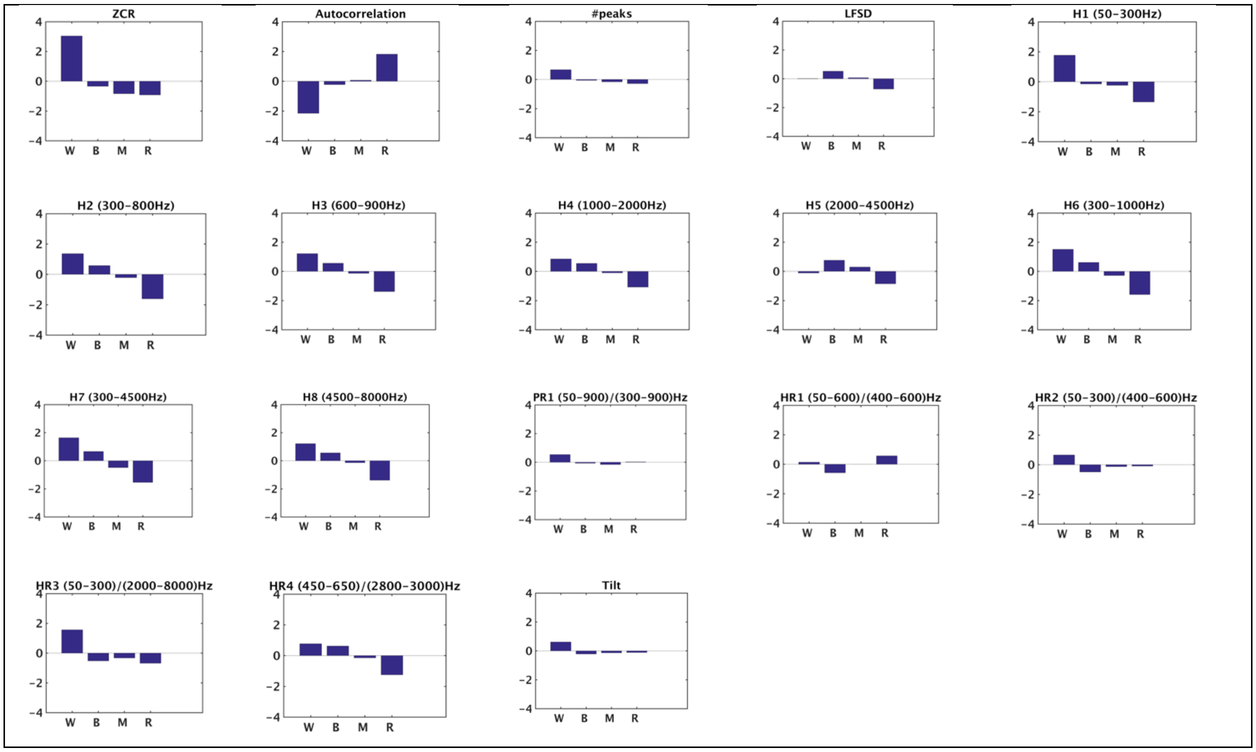


Figure 6.3 (cont.): Sensitivity Plots for Male Acted Voices. These plots show d_a, the unequal-variance sensitivity, across conditions for a selection of features. A larger magnitude of a bar shows a greater ability of a feature to provide separation for that condition. The ZCR, Autocorrelation, #peaks, LFSD, ant Tilt features are identical to the female features, and the rest differ by band boundaries. The results suggest that the modal condition is the most difficult to distinguish in male speech, and that resonance is easier to detect in male than female speech, and that autocorrelation does a poor job distinguishing modal speech compared to the female case. Again, the features work together to provide separation, particularly for the modal and breathy conditions.

6.2.3 Iterative Feature Group Selection, Model Building and Validation in Scripted Speech

To address the research questions, feature combinations were selected (based on the sensitivity measurements and mean/variance separations) which would best work together on the acted speech corpora to provide maximum separation across conditions. 4-way decision tree classifiers were trained using a series of the most promising feature combinations, pruned to guard against overfitting and tune performance, and cross-validated to measure performance for *sample independence*, *text independence*, and *speaker independence*. These measurements provide insight into the requirements for training a model on a corpus using these features (e.g., speaker coverage vs. text coverage), and provide insight into how well the model will work when trained under different conditions. To validate sample level independence, 5-way cross validation was used across all speakers and phrases in the corpora, such that each trained model saw none of the test samples. To validate text independence, the male and female corpora were divided into segments of similar size, (5 segments for females and 6 segments for males), segmented where speaking styles were likely to change, and text segments held out. Speaker independence, was validated by holding one speaker out. Finally, binary classifiers were trained for the sample independent case, using the same feature sets, and tested to distinguish each of the four conditions against the rest. Results are presented in terms of *precision* (the fraction of retrieved, or recognized, instances which were relevant, or correctly recognized), *recall* (the fraction of relevant, or available, cases which were retrieved, or recognized), and overall *accuracy*. The process of feature set selection, classifier algorithms, and results are described below.

Classification Feature Sets: The interactive analysis was the primary driver of feature selection, and was based on observed complementarity among the discriminant characteristics of individual features. Analysis of these features revealed the best feature separators for each condition, showed features which provided general separation across multiple features, revealed the most difficult conditions to detect in males and females, revealed male and female differences, and suggested that the entropy features worked best as a collective. Given the interactive analytic results, and as a secondary technique, over 20 feature combinations were evaluated via the described cross-validation techniques for both males and females. The first feature collection evaluated for use with a 4-way classifier included the strongest separators for each condition, the best general features, and the full collection of entropy features. Next, features were selectively removed, starting with those most likely to be redundant, or those which would cause the most confusion. At this stage, we found that while AC was a superior separator for the whispered condition, including it with ZCR degraded performance in most cases. We also found that frequencies above 4500 did not contribute significantly to the results, that vowel length was not a reliable separator for any of the conditions, and that H2-H1 and TILT either did not contribute significantly or degraded results.

In the next phase of evaluation, power ratio and entropy ratio features were successively added, and tested for their ability to boost performance, particularly for the most difficult-to-classify conditions. The error rates were calculated using the same cross-validation methods

The two best-performing feature collections (measured by global accuracy and average recall) for males (N of 25) and females (N of 20) are reported in Table 6.4, and discussed in detail in the "Results" section.

Table 6.4: Male and Female Feature Sets. The two best-performing feature set combinations for males and females are listed here. The features listed are those described in Figures 5 and 6 and discussed in Section VII. Note that the frequency bands are different between males and females, to account for gender differences in the spectrum across conditions. Frequency ranges are given in Hz.

Name	Gender	Features
SET1	Male	ZCR, H1(50-300), H2(300-800), H3(600-900), H4(1000-2000), H5(2000-4500), H6(300-1000), H7(300-4500), PR1(50-900)/(300-900), HR1(50-600)/(400-600), HR2(50-300)/(400-600), #peaks
SET2	Male	SET1 features, plus LFSD
SET3	Female	ZCR, AC, H3(300-800), H4(500-1500), H5(1000-2000), H6(2000-4000), H7(300-4500), PR1(50-300)/(50-150), PR2(50-500)/(50-1000)
SET4	Female	ZCR, H1(50-150), H3(300-800), H4(500-1500), H5(1000-2000), H6(2000-4000), H7(300-4500), PR1(50-300)/(50-150), PR2(50-500)/(50-1000), HR1(50-300)/(50-150), HR5(50-300)/(2000-8000), LFSD

Classifier Algorithms: The emphasis of the work here is on feature selection, finding the acoustic correlates for effort levels, and understanding their function, both individually and collectively. For these reasons, the classifier algorithm remains simple, so that the results reflect more the power of the features to characterize effort levels, and less the power of the classification algorithm (or optimizations to classifier algorithms). Future work can isolate classifier selection and optimization as separate goals. For these reasons, simple decision trees were used.

Overfitting, however, was a risk, and to address this issue, each classifier was pruned by factors of 5, 10, 15, and 20, and individually evaluated. Pruning turns selected branch nodes of a tree into leaf nodes, and removes the leaf nodes under that branch. Pruning to level n turns the nodes at tree height n into leaf nodes and removes the leaf nodes. Best results consistently corresponded with pruning factors of 10 or 15, and the pruned classifiers were used for evaluation here.

Results: The research questions asked what acoustic features could distinguish across the four levels of vocal effort, from whispering, to breathiness, to modal speech, and to resonant speech. Table 6 summarizes the precision and recall for each condition, along with the global accuracy of the top two male and female 4-way decision tree classifiers for the sample, text, and speaker-independent cases.

The feature sets were similar between males and females, except for spectral ranges on frequency bands of interest; and recognition results (when feature sets were applied with cross validation to 4-way decision tree classifiers) were correspondingly similar. The difference between male and female accuracy at sample independence for the best-performing feature sets (as measured by global accuracy) was not statistically significant per chi-square test (χ^2 =0.42, df=1, p=0.52); and differences in recall rates for each condition were also not statistically significant between males and females. The male/female accuracy difference was also not significant at text independence per chi-square test (χ^2 =0.99, df=1, p=0.32); however, females had a significantly better whisper recall rate according to t-test for independent means (p=0.003). Finally, female accuracy was significantly better than male accuracy for the speaker independent tests per chi square test (χ^2 =4.5, df=1, p=0.036).

Table 6.5 summarizes the average precision and recall for each condition (whispered, breathy, modal, and resonant) for male acted voices, using feature sets Set1 and Set2, applied to 4-way decision tree classifiers and cross validation techniques. The data show that, on the average, both of these sets perform at three times chance for sample independence and about twice chance for both text and speaker independence. The overall accuracy for Set1 was greatest at sample independence (μ=0.7, σ=0.2), and had similar values for both text independence (μ=0.51, σ=0.01) and speaker independence (μ=0.5, σ=0.04). Set 2 followed the same trend, with greatest accuracy at sample independence (μ=0.76, σ=0.08) and again similar levels for text-independence accuracy (μ=0.52, σ=0.01) and speaker independent accuracy (μ=0.5, σ=0.01). The differences in accuracy between the two sample models were not statistically significant per chi-square test (p=0.65 for text independence, p=0.67 for text independence, and p=0.20 for sample independence). The difference in overall accuracy at sample independence, however, was nearly 6%; and it is intriguing that the only difference between the two models was the inclusion of LFSD. Future work may re-evaluate this result with a larger acted voice dataset and a wider range of speakers. The best-recognized condition was resonance. Note that recall values between the two models are comparable in the sample and text independent cases, just less accurate for text independence; a larger training set of male acted speech could possibly overcome the loss in recall rates in practice, especially when combined with an optimized machine learning model.

When the entire set of data from a speaker was reserved for testing, and not included in the training data set (i.e., withheld for testing), the training data did not have sufficient samples per speaker to validate speaker-independent whispering and resonance in males. The remaining conditions (breathy, modal, and female resonance), however, evaluated at similar accuracy to text independence. See Table 6.5 for details. Losses in recall rates for speaker independence might be overcome by increasing the size of the dataset and number of speakers.

Table 6.5 also provides the recall results for the female speakers and corresponding feature sets (Set3 and Set4), again applied to 4-way decision tree classifiers. Average accuracy for both models is similar, and is about three times chance for sample independence, 2.5 times chance for text independent tests (better than the rates for males), and again about twice chance for speaker independent tests. Set3 accuracy (μ=0.7/σ=0.024; μ=0.6/σ=0.08, and μ=0.4/σ=0.12 for sample, text, and speaker tests, respectively) was not significantly different per t-tests of 2 independent means (p=0.49, p=0.74, and p=0.31 for sample, text, and speaker tests, respectively) from that of Set4 accuracy (μ=0.72/σ=0.01; μ=0.6/σ=0.02; and μ=0.5/σ=0.085 for sample, text, and speaker tests). Accuracy is defined as is typical for a confusion matrix: *Accuracy* = (Σ *TP* + Σ *TN*)/(Σ *total population*), where TP and TN are the number of true positives and negatives for each condition.

Table 6.5: Classifier Precision, Recall, and Accuracy. This table summarizes the mean classification results across all of the folds of the two best-performing feature sets for males and females (defined in Table 5), and compares the Sample (SMP), Text (TXT), and Speaker (SPK) Independent Cases. It shows the precision and recall (p/r) for the whispered, breathy, modal, and resonant (W, B, M, R) cases, and the global accuracy (A).

	SMP					TXT					SPK				
	W p/r	B p/r	M p/r	R p/r	A	W p/r	B p/r	M p/r	R p/r	A	W p/r	B p/r	M p/r	R p/r	A
SET1	.63/.63	.70/.68	.67/.68	.71/.74	.69	.43/.58	.49/.45	.48/.46	.59/.59	.51	*	.44/.50	.69/.48	*	.41
SET2	.71/.68	.76/.74	.73/.74	.77/.80	.76	.39/.59	.49/.47	.48/.47	.60/.60	.52	*	.56/.59	.49/.47	*	.41
SET3	.70/.80	.80/.74	.59/.62	.65/.72	.70	.60/.68	.63/.13	.52/.51	.56/.67	.57	*	.52/.59	.22/.36	.76/.53	.51
SET4	.76/.75	.78/.76	.64/.66	.69/.73	.72	.72/.71	.59/.57	.54/.51	.55/.65	.59	*	.52/.59	.22/.36	.77/.53	.51

The classifier performance was less evenly distributed across conditions for females than for males. The condition with the highest recall rates varied, but the modal condition consistently had the lowest performance across all conditions. The modal condition dropped disproportionally in performance for the speaker-independent case, but the other conditions were comparable to the text-independent case (like the male data). As with males (see Figure 6.4), whispering is not evaluated in the speaker-independent test, because the speakers did not all have sufficient representation of whispered speech to support a valid test.

As a final test, the best performing male and female feature sets (Set 2 and Set 4, respectively), were used to train binary one-vs-all classifiers for the sample independent case. Accuracies for the male binary one-vs-all classifiers were 0.95, 0.84, 0.64, and 0.67; while accuracies for the female binary one-vs-all classifiers were 0.88, 0.63, 0.66, and 0.66 for whispering, breathiness, modal speech, and resonance, respectively. The male whisper and breathy classifiers performed better per t-tests of independent means (p=0.0007, and p<0.00001, respectively); while the female modal classifier performed better (p=0.013), also per t-test of independent means. The performance difference between resonant classifiers was not statistically significant (p=0.17).

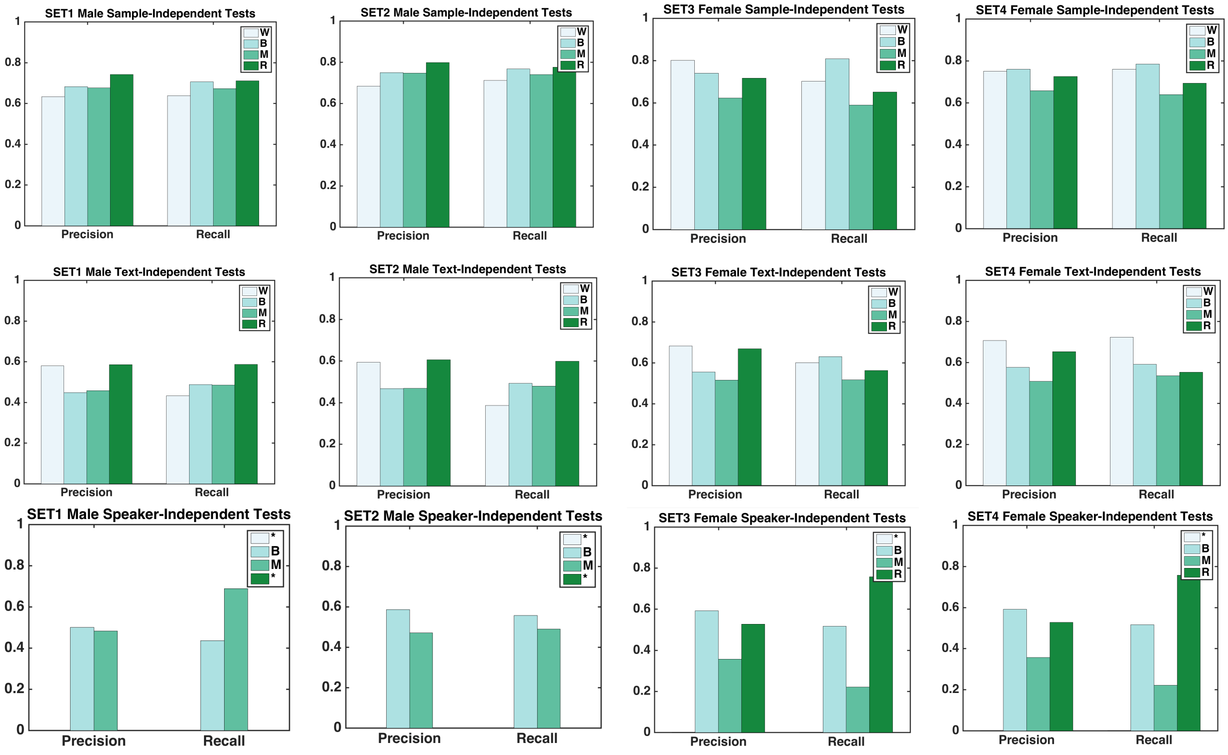


Figure 6.4: Cross Validation Results for Male and Female Acted Voices. This figure shows the feature set performance for sample-independent (top), text-independent (middle), and speaker-independent (bottom) cross validation. ***For males:*** Sample-independent performance is the strongest, followed by text independence, followed by speaker independence. Set2 slightly

Figure 6.4 (cont.): outperformed Set1 in each case. Note that for the speaker-independent case, the dataset does not have enough whispered or resonant samples within every speaker to ensure a valid 95% confidence interval for accuracy. Therefore, accuracies are not reported for entries with n < 25 (marked * in the key). ***For females:*** Note that for the speaker-independent case, the dataset also did not have enough whispered data to validate these cases. As for male acted voices, sample-independent performance is strongest, followed by text and speaker independence.

6.3 Unscripted Speech

6.3.1 Intuitive Interactive Analysis in Unscripted Speech

The interactive analysis process was similar to that of scripted speech. Samples of each of the four effort levels, this time, breathy, modal, resonant, and creaky voice, were collected from oral history interviews, and examined to learn the characteristics of each condition in unscripted speech. This larger collection of samples, unrestricted from constraints of acting a fixed part, had more variance than the scripted speech, and male and female unscripted speech was equally variant. Because of the larger sample and wider variance, mean spectra (magnitude squared) across the samples in each category were examined and are shown in Figure 6.5. This was a good first step for observing the major differences across effort levels and between gender. For males, breathy voice shows the strongest spectral component at F0, around 100 Hz, with a quick dropoff and much smaller energy content between about 300-700 Hz, where F1 and small harmonic multiples lie. An even smaller region of energy appears around 1000 Hz, where F2 for some vowels appears, along with higher harmonic multiples. Energy about 1500 Hz does not register on this graph. For females, the strongest spectral component still appears at F0, around 200 Hz, again with a quick dropoff. The energy content between about 300-700 Hz was proportionally larger than that of the male speakers, and energy above about 1100 Hz does not register on this graph.

Male modal speech had clearly-defined F0 and H1 (around 100 and 200 Hz, respectively) with strong energy in the 300-900 Hz region. H1 was clearly smaller than F0. The energy in the 300-900 Hz region had two clearly defined pulses, which were comparable to or greater than F0. Energy dropped at around 1000 Hz and peaked again between 1000-1700 Hz (the region of F2 and higher harmonics). The female modal speech profile differed as expected from the male profile. F0 and H1 were clearly defined again (around 200 and 400 Hz respectively). F0 and H1 had comparable energy. The region of F1 and some F2 (around 400-1100) was a high-variance energy mass, which rapidly dropped off rapidly around 700Hz. Energy above 1100 Hz does not register on the graph.

Male resonant speech featured a relatively modest F0 with an increasing energy contour to about 700 Hz, and a quick dropoff between 700 and 1000 Hz. Higher formants have strong energy associated with them which distinguishes male resonant speech from the other effort level types examined here. Female resonant speech has a similar profile up to 1000 Hz. F0 is modest, and the contour increases to about 800 Hz, where it quickly drops off around 1000 Hz. The contour does not show the formant outlines so clearly above 1000 Hz. The energy content between 1000-2000 Hz also distinguishes resonant speech from the other types for females, but it is proportionally smaller than that of male speech. H1 appears to be larger than F0 in both genders.

The male creaky voice profile had three regions of spectral peaks, one around F0, another between 500-750 Hz, and the final peak around 1400 Hz. The spectra showed variable energy in the 1000-2000 Hz region, lower-energy than the resonant speech, and more variant than the modal speech. Energy above 2000-3000 Hz appeared to be variant and slightly greater than that of modal speech. Female creaky voice featured a peak at F0 and the greatest energy between about 400-700 Hz, peaking around 600 Hz, and rapidly dropping off. The region between 1000-2000 Hz had low-level variable energy less than that of resonant speech.

While this approach shows the major differences across condition and between gender, it prevents identification of sub-types within each class, prevents evaluation of periodicity patterns, and prevents accurate examination of the relationships between spectral energy among regions or harmonics of interest. Seeing periodicity patterns is not only informative, it is essential for establishing spectral bands for entropy measures. Also, longer vowels evolve. In the initial frames, spectra may begin looking less like the profile of the perceived type, but after a frame or two, they often evolve into one of the prototypical patterns. Over very short periods of time, a single vowel sound can “blossom” into a strong example of the type. When this occurred, listeners usually perceived the pattern heard at the end of the sound. For these reasons, individual spectra were examined, and frequently-occurring types identified. For the cases of breathy and creaky speech, multiple subtypes were identified. Finally, longer vowels were examined over time for the “blossoming” effect.

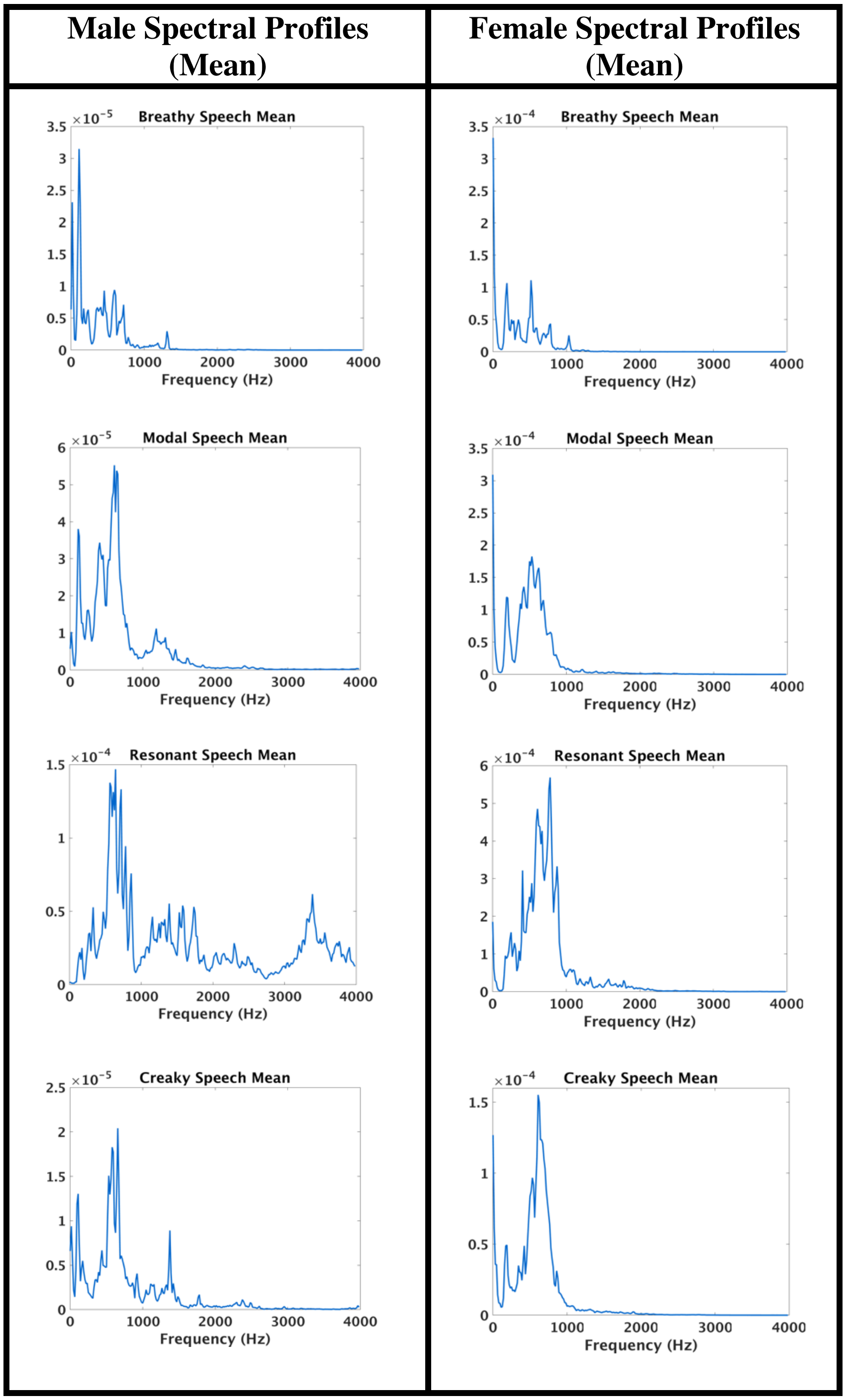


Figure 6.5: Mean spectral profiles for males and females.

Representative Breathy Speech Frames: Breathy speech, for both males and females, tends to fall into one of two sub-classes. The first breathy profile, which we call “classic” breathy, has a strong F0, a weaker H1 (first harmonic), and zero to a small number of additional harmonics, which are much weaker than H1. Figure 6.6 shows prototypical “classic” breathy speech for males and females.

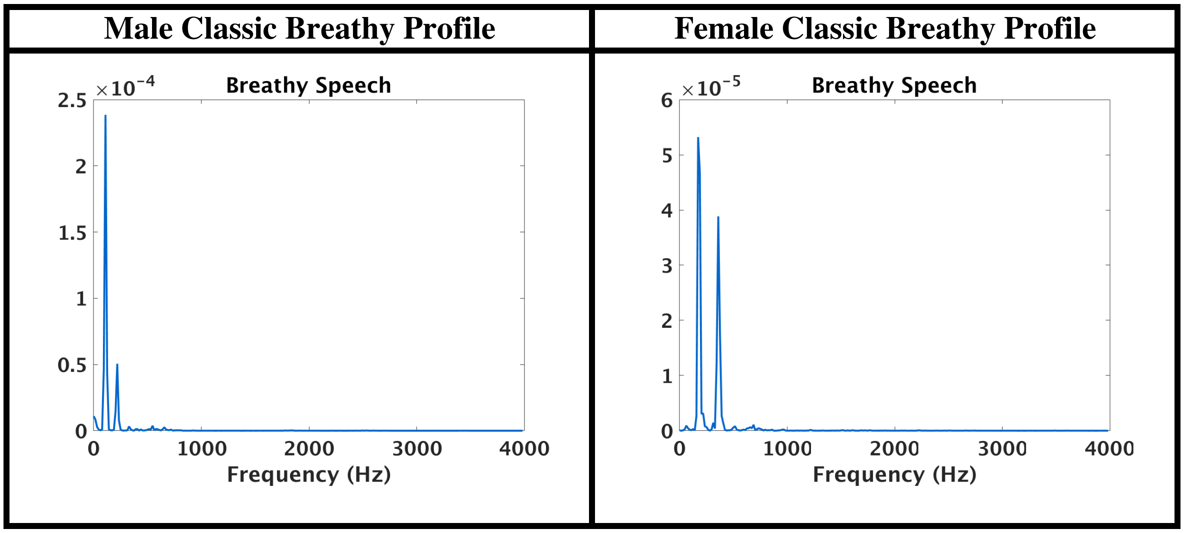


Figure 6.6: Classic breathy spectra. F0 and H1 are closer in magnitude in the female sample.

Breathy voice can also have a “noisy” profile. In noisy breathy spectra, random frequencies are present, along with the expected F0 and H1 in the classic breathy profile. This occurs because, as discussed earlier in the chapter, breathy speech and whispered speech are adjacent to one another in a continuum relationship. Recall that in the whispered profile, F0 is absent, and a large number of randomly-distributed frequencies are present across the vocal range. The noisy breathy voice has an F0, and usually one or more harmonics, but the voice is in a transitional state between whispering and classic breathy mode. The presence of noise in the spectra is perceptually important. A spectral profile which would otherwise resemble a soft modal voice will sound breathy to a listener if it also has a high, whisper-like noise content. Figure 6.7 shows noisy breathy profiles.

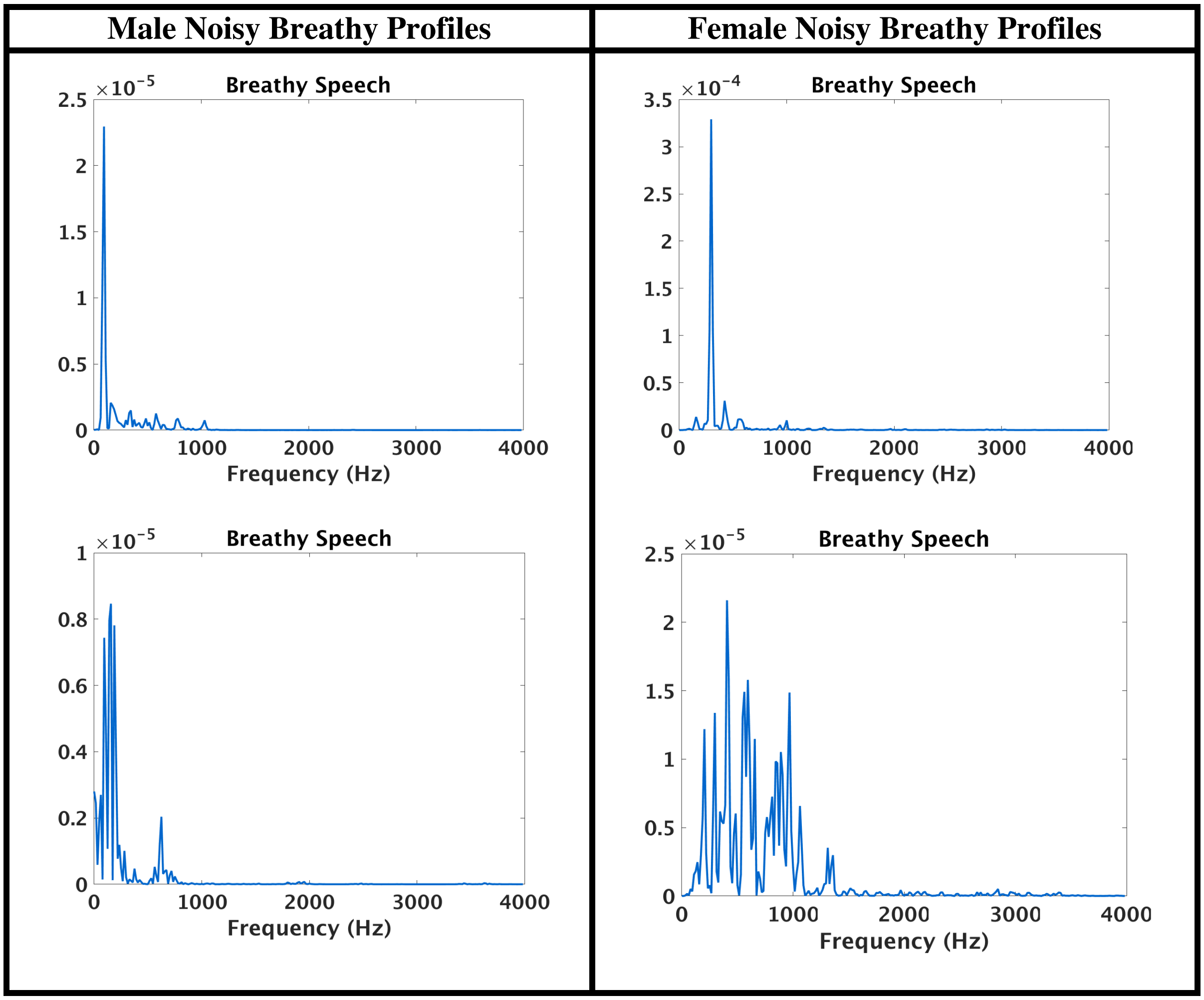


Figure 6.7: Noisy breathy spectra. The samples on the top and bottom show smaller and larger amounts of noise, respectively. Even with small amounts of noise, a breathy voice can have more harmonics than the classic breathy profile, strong presence in the low formant regions, and a first harmonic which is larger than F0. Some of these profiles also show energy in the sub-F0 range, which prior work has noted for breathy voice (Gowda and Kurimo 2014).

Representative Modal Speech Frames: Modal voice, for both genders, tends to have decreasing harmonics in general, with more energy in the formant regions, as expected. The specific formant regions depend on the utterance, specifically, the vowel. Most of the time, harmonic energy levels are higher below 1200 Hz than they are above 1200 Hz (both genders). Modal voice has strong periodicity across the spectrum, and usually does not contain noise. As discussed earlier in this chapter, modal voice and resonant voice are adjacent in a continuum relationship. As modal voice approaches resonant voice, more energy will appear in higher harmonics and formant regions. Figure 6.8 shows three typical male and female modal samples.

Note from Figure 6.8 that, in general, the slope of the frequency contour (defined here as a line fitted through the spectral frequencies visible on the graph) can be either positive or negative. If the sample has few harmonics (as in the bottom male sample), the relationship between F0 and H1 is important. The stronger H1 is then the differentiating factor between modal and breathy voice for modal samples with few strong harmonics. On the other hand, if the sample has more harmonics (as in the bottom female sample) or higher formant energy, the relationship between F0 and H1 may be helpful in differentiating between modal and resonant voice. Also note that in Figure 6.8, the energy in formants above 1200 Hz is less than or equal to the energy in F0 and in formant frequencies below 1200 Hz.

Since the spectral shape follows the varying contour of the formants, and because modal voice lies on a continuum between breathy and resonant voice (with edge cases adopting characteristics of breathy voice or resonant voice), defining sub-types for modal voice is not practical. Furthermore, we did not find any evidence of different kinds of modal voice production.

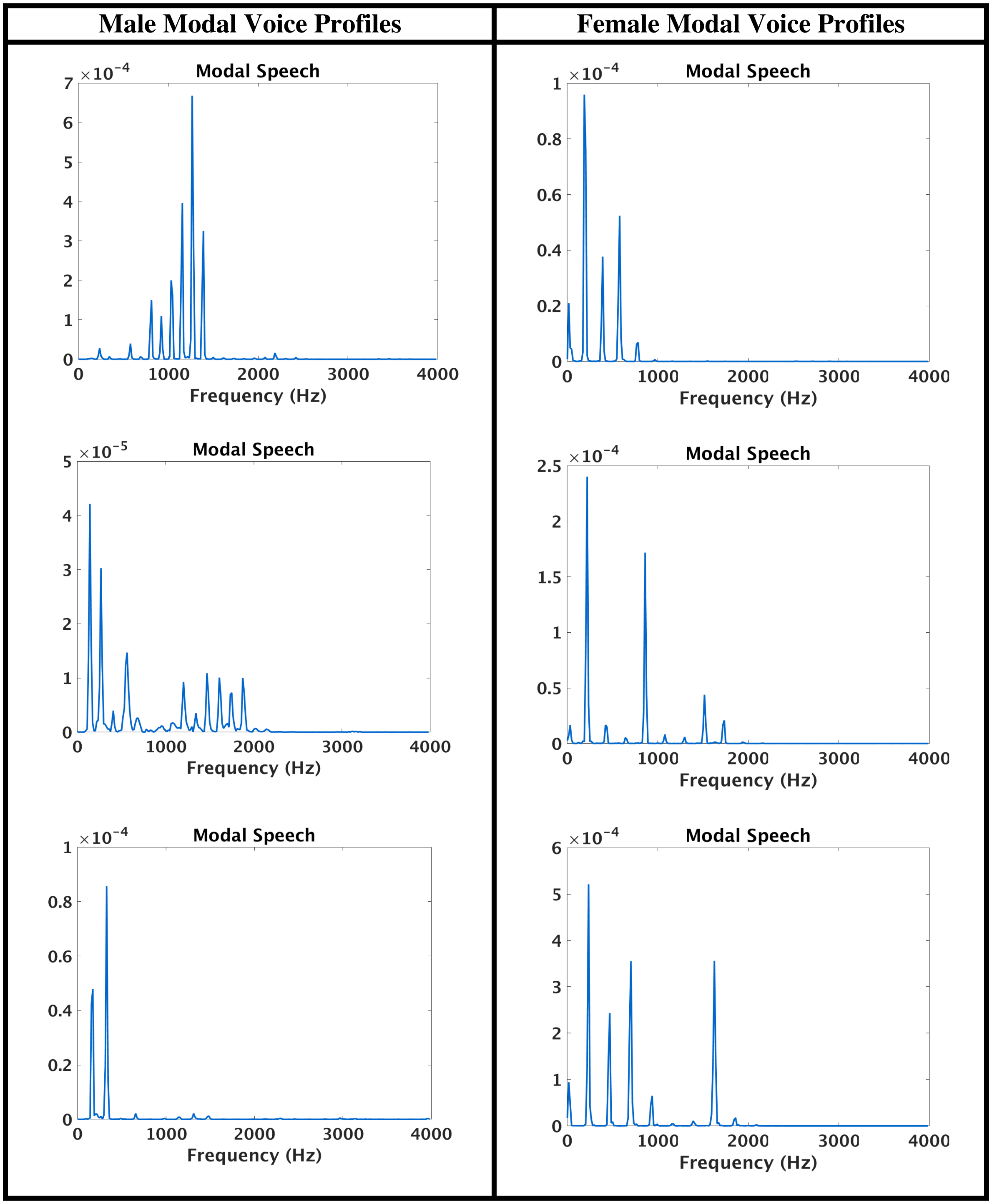


Figure 6.8: Modal voice spectra. The samples show variability in spectral slope and relative energy in F0, H1, and various formant regions. In all cases, the spectra are periodic, with little noise, and more energy (in general) in the lower frequencies than the higher frequencies.

Representative Resonant Speech Frames: The corpus included two kinds of resonant voice. The first kind of resonant voice, which we call "classic resonant" here, had strong a F0, with strong low multiple harmonics (stronger than the modal harmonics, with respect to F0), and strong resonances at formant frequencies (stronger than the modal resonances, with respect to F0). Figure 6.9 shows examples of classic male and female resonant voice. Note that modal and resonant voice are adjacent to one another in a continuum relationship, so some natural perceptual confusion occurs at the boundaries. Listeners will perceive a voice to be resonant when the low harmonic multiples are strong (greater than or closer to F0 levels) with respect to F0, and resonance at the formant frequencies are strong (also greater than or closer to F0 levels).

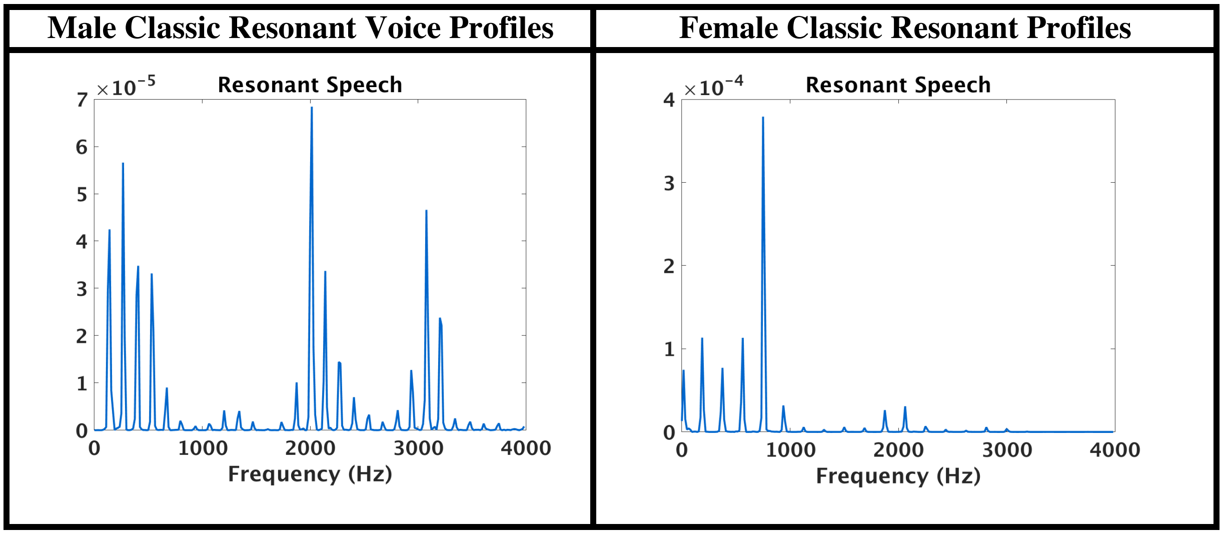
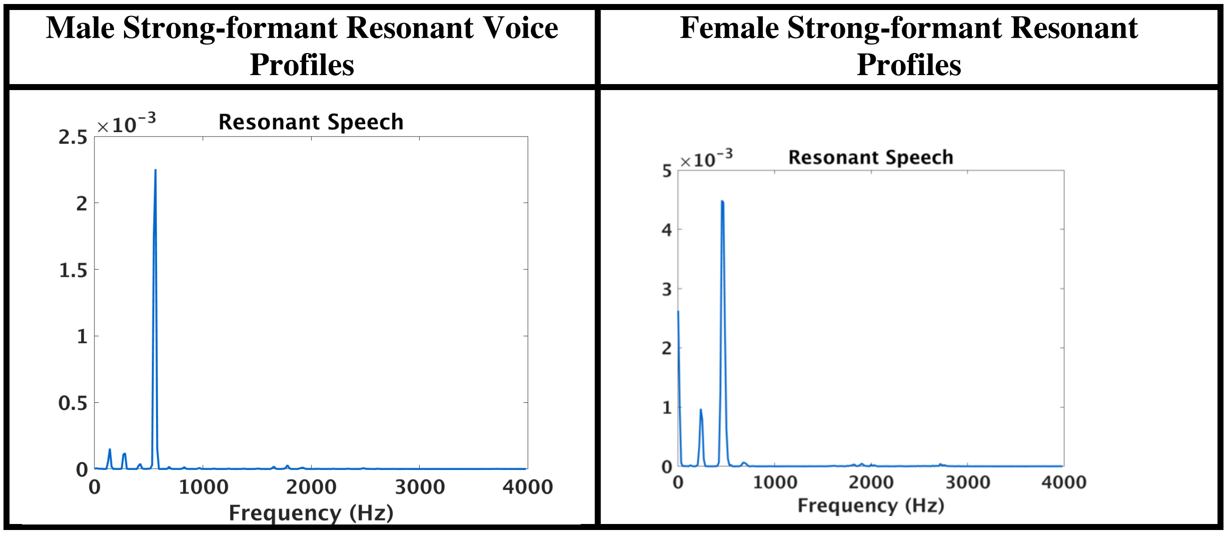


Figure 6.9: Classic resonant voice spectra. Note the strong low harmonics (with respect to F0) and the stronger formant frequencies (again with respect to F0). The formants appear to align with the harmonics as well.

The second kind of resonance occurs when the strength of the first formant is extremely strong in comparison to the power of F0. In the examples in Figure 6.10, the female formant aligns with the first harmonic, and is more than 4 times greater than F0. In the male example, the harmonic multiple also aligns with the first formant, and as a result, the first formant is more than 9 times greater than F0. It is not enough for the first formant or first harmonic to be greater than that of F0 for listeners to hear resonant voice; it must be several times greater than that of F0. Figure 6.10 shows an example of a male modal voice with a similar profile. The difference is that the modal first harmonic/first formant is only greater than F0 by a factor of 2. The ratio between the power around F0 and power around the first formant is critical for differentiating modal and resonant voices here.



Male Strong-formant Resonant Voice Profiles

Female Strong-formant Resonant Profiles

Figure 6.10: Strong-formant resonant voice spectra. Note the ratio between the first formant and F0. Also note the alignment of the formant with harmonic multiple frequencies.

Representative Creaky Speech Frames: Several confirmed creaky voice subtypes appeared in the unscripted speech corpus. These types include 1) prototypical creaky voice, 2) vocal fry, 3) multiply-pulsed creaky voice, and 4) aperiodic creaky voice. These types (and others) have been recorded in the literature, although much of the creaky voice literature focuses on one or two of these subtypes without considering the others (Cullen et al., 2013; Drugman et al., 2014; Keating et al., 2015; Narendra and Rao, 2015). Prototypical creaky voice has a low F0 compared to modal phonation and a variable (unstable) F0. The glottis is constricted, with a slow airflow, and is closed for longer periods of time (low open quotient) when compared to modal voice. Figure 6.11 shows samples of prototypical creaky voice.

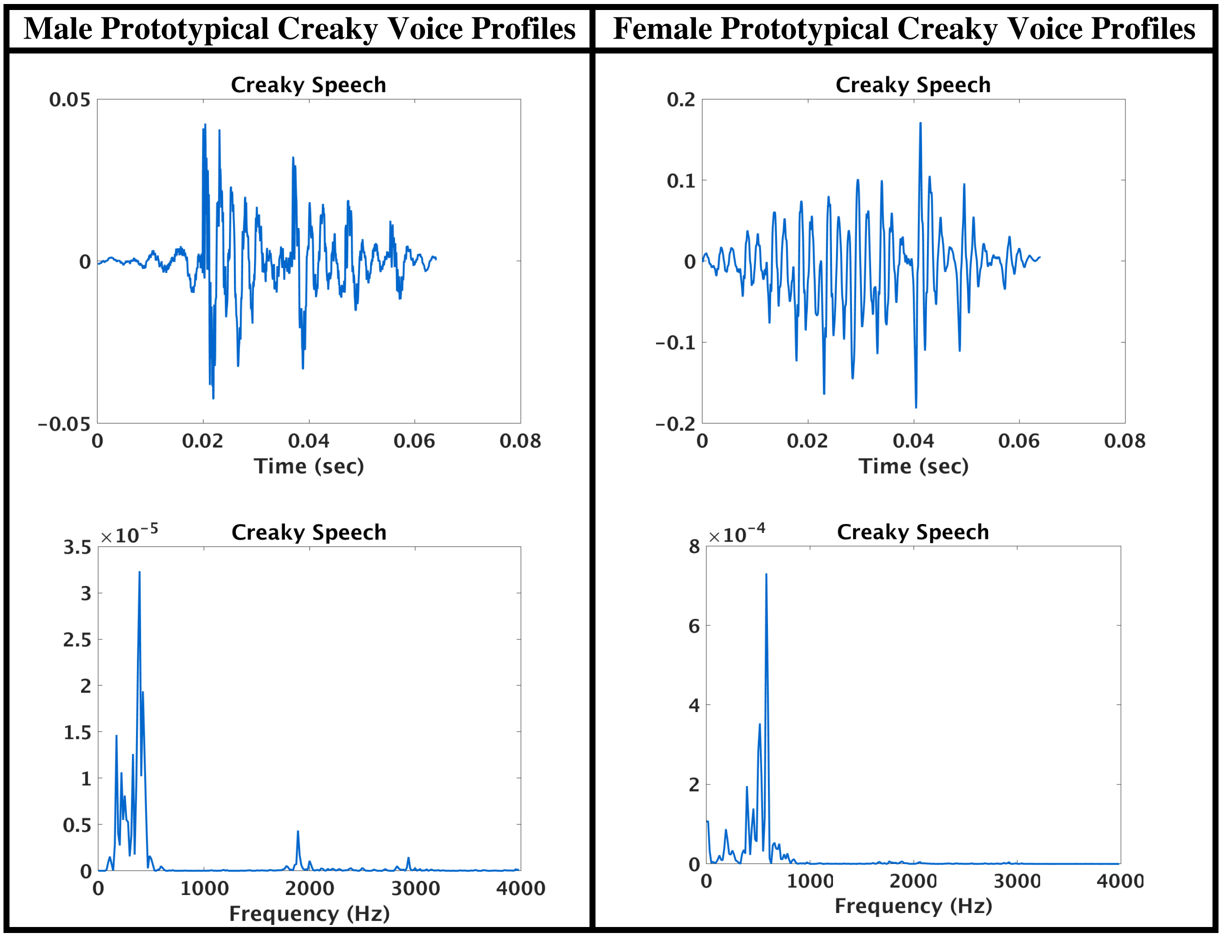


Figure 6.11: Prototypical Creaky Voice. The waveforms appear irregular in this kind of creaky voice because F0 is unstable. This instability is even more obvious in the spectra. Most of the spectral energy for this type of creaky voice is below 1000 Hz for both males and females, below about 600 Hz for males and 800 Hz for females. Smaller amounts of energy are present at higher frequencies.

Vocal fry is similar to prototypical creaky voice, except that the waveform pulses are dampened, meaning that the initial part of each period has a much higher amplitude than the rest of it. The waveform is often highly periodic, and this is also apparent in the spectra. Figure 6.12 shows samples of vocal fry (waveforms and spectra) from the corpus.

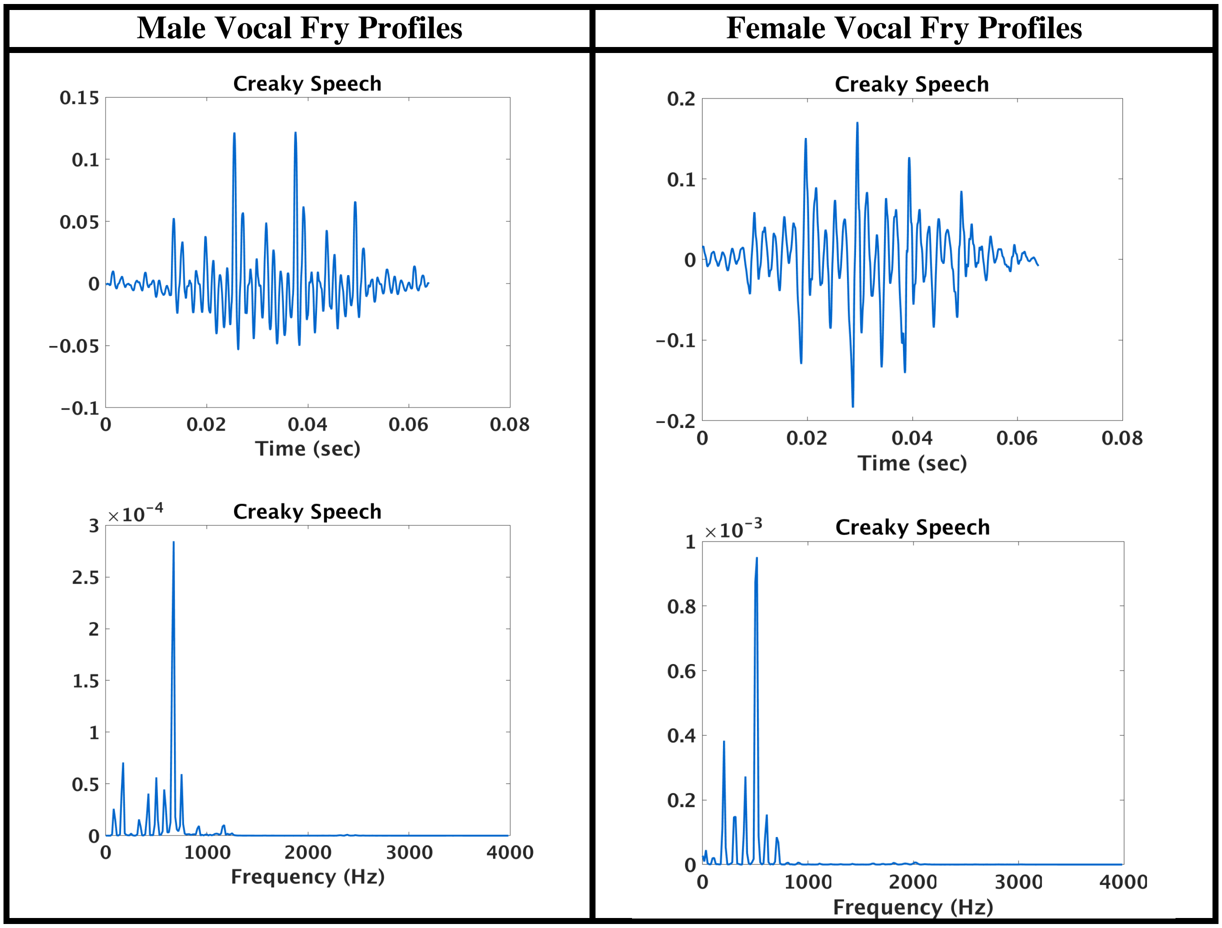


Figure 6.12: Vocal Fry. Dampening is apparent in the waveform samples, and both the waveforms and spectra show strong periodicity. F0 is quite low. The male F0 averages around 110 Hz for modal voices, but is around 80 Hz here. The female F0 is around 100 Hz here, almost half of the typical frequency for this speaker; the female waveform also appears to contain multiple F0 frequencies.

Multiply-pulsed creaky voice has multiple periodicities, and effectively, multiple F0 frequencies. Detecting a stable F0 for this kind of voice can be difficult for both humans and analytic software. The lowest F0 can be extremely low (half the modal F0 or below), which adds further difficulty for pitch tracking software when the frequency is below the range of a given software algorithm. Usually this type of creaky voice has a double frequency, but higher periodic multiples sometimes occur. Figure 6.13 shows multiply-pulsed creaky voice samples.

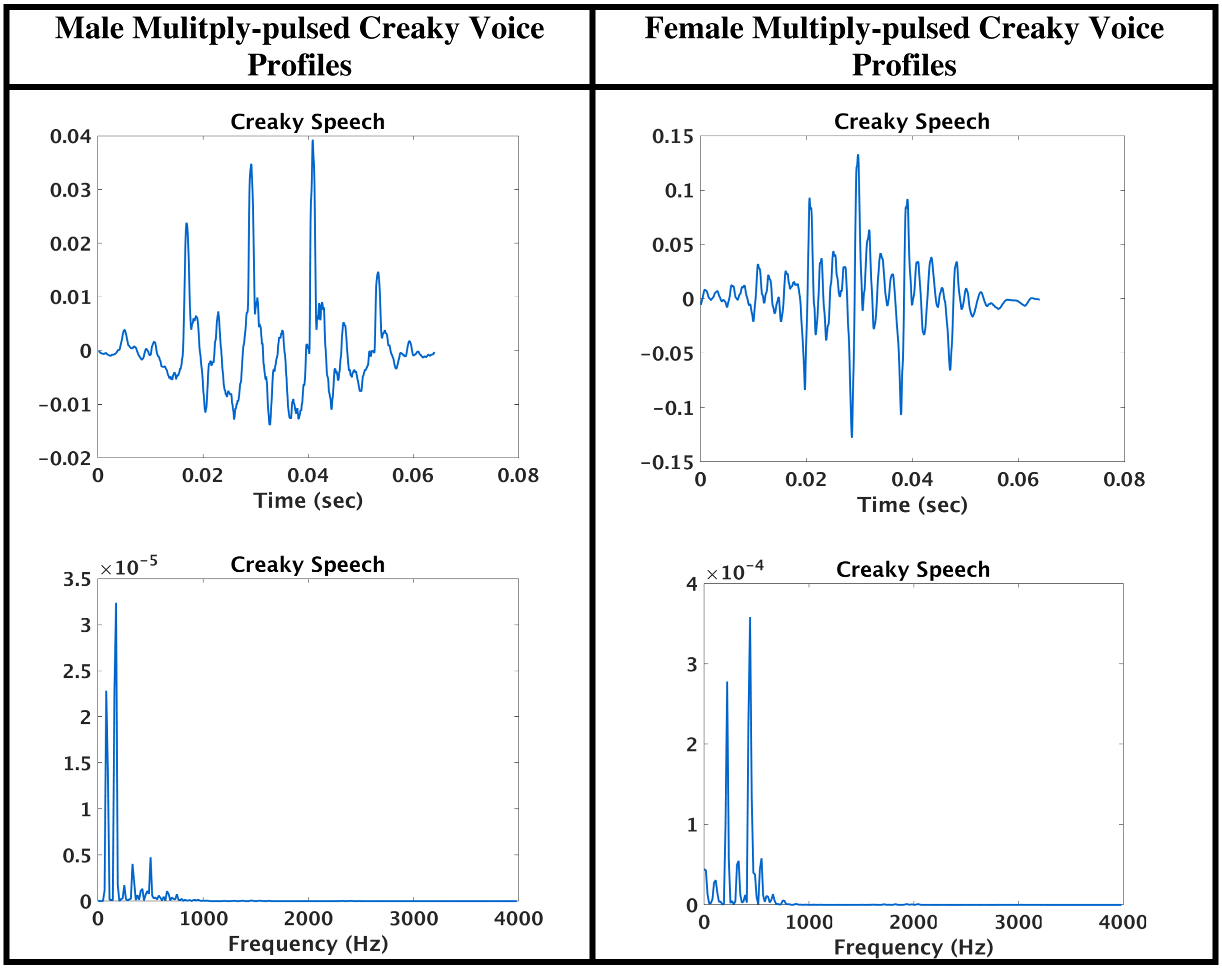


Figure 6.13: Multiply-pulsed Creaky Voice. Multiple periodicities are apparent in the waveforms (note the alternating high and low pulses). Note the first two strong frequencies in the male spectra, the second approximately double the first; these are effectively two F0 frequencies. Note also the alternating harmonic power levels in both spectra (weak, strong, weak, strong). Also note the lowered frequency of the first spectral element. The male sample is approximately 80 Hz (compared to about 120 for this speaker's modal voice), and the female sample is approximately 110 Hz (compared to about 220 for this speaker's modal voice).

The hallmark of aperiodic creaky voice is extreme irregularity. This kind of creaky voice can be interpreted as an extreme form of prototypical creaky voice, where irregularity has progressed to the extreme condition with no detectable periodicity. Pitch tracking algorithms and voice detection algorithms usually fail when presented with this kind of voice. This type was extremely common in the male voices sampled. Figure 6.14 shows aperiodic creaky voice samples.

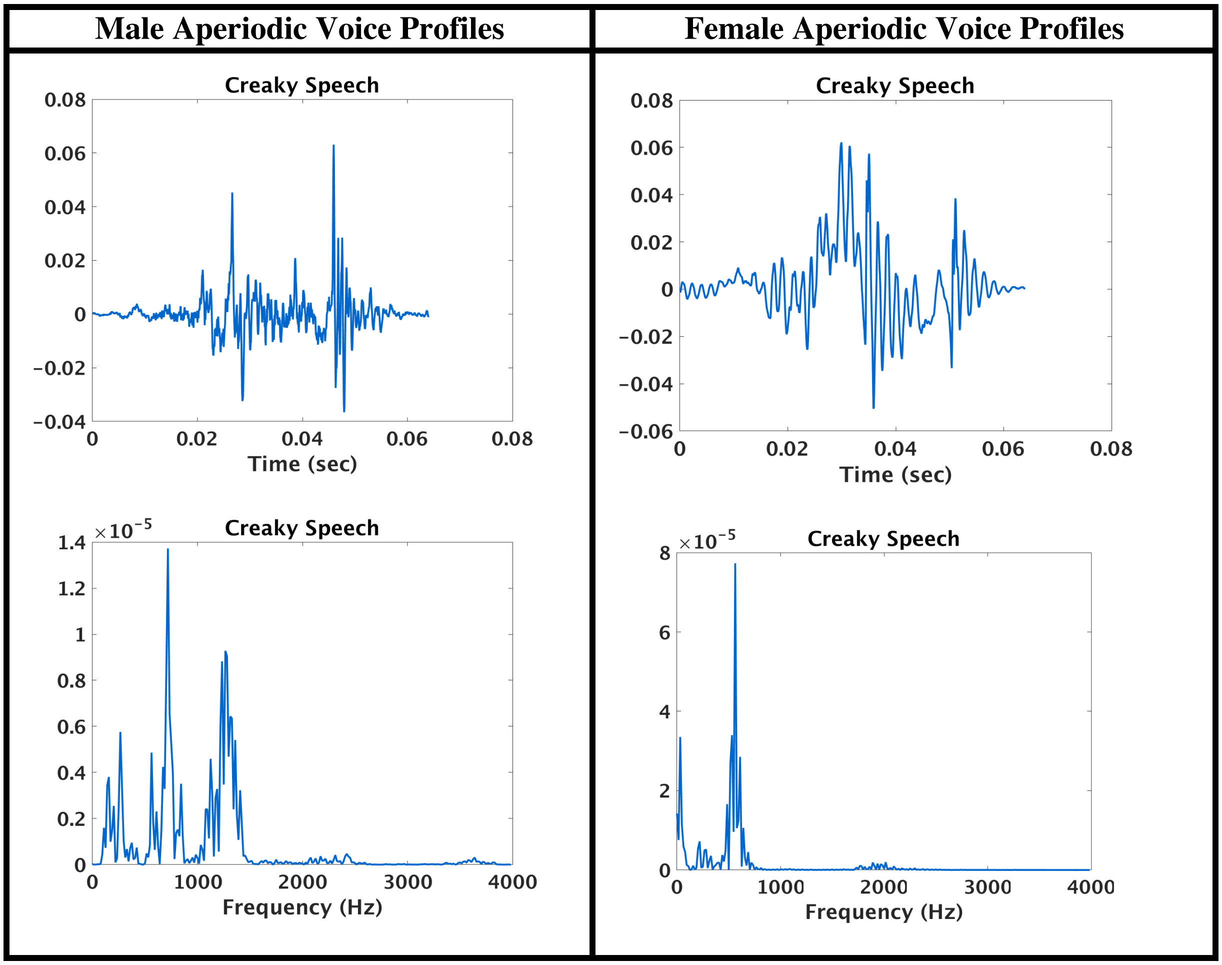
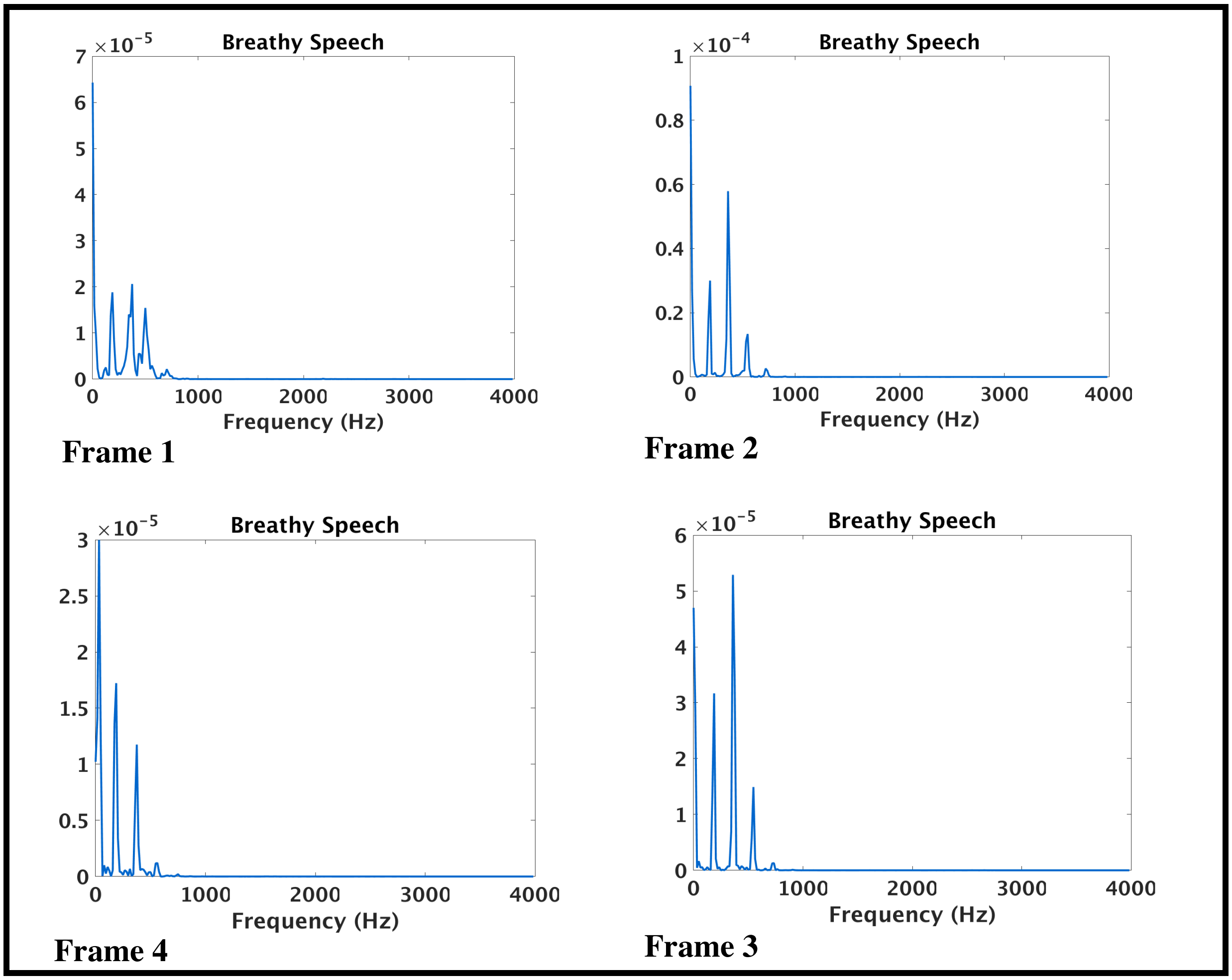


Figure 6.14: Aperiodic Creaky Voice. Extreme aperiodicity is apparent in both the waveforms and spectra (all regions). Most of the spectral energy is below about 1500 Hz for the male sample and below about 800 Hz for the female sample.

This analysis shows that a creaky voice model should support at least four different spectral profiles; creaky voice isn't a single entity. At least, creaky voice should model the irregular kinds of creaky voice (prototypical and aperiodic) and the highly periodic (usually vocal fry and multiply-pulsed).

Spectral Blossoming: As mentioned earlier in this chapter, spectra often evolve over time in longer vowel samples; and listeners tend to perceive the voice quality which they hear at the end of the sample. This could cause difficulty in stream processing if the analytics did not account for this phenomenon, particularly if earlier frames had characteristics of another voice quality.

Figures 6.15 shows an example of spectral blossoming. Frame sequences start at the top left and move clockwise.



Frame 1

Frame 2

Frame 4

Frame 3

Figure 6.15: This vowel begins with a spectral profile which could possibly be interpreted as soft modal speech. The sound later "blossoms" into a classic breathy voice profile in Frame 4, and listeners classified this sound as breathy, not modal speech.

6.3.2 Acoustic Feature Selection & Analysis in Unscripted Speech

The interactive analysis in section 6.3.1 showed that the scripted and unscripted conditions were similar for breathy, modal, and resonant speech. The candidate features from the scripted speech section, therefore, were used to help distinguish these conditions. The "bands of interest" defined for scripted speech in Table 6.1 also applied. An additional feature which examined entropy patterns in the 50-85 Hz band for males (H9) was explored because of the lowered frequency common in creaky voice. Expanded power and entropy ratios were also proposed to examine low-frequency patterns and relationships with upper frequency bands, for purposes of distinguishing creaky voice from the other voice quality conditions. Tables 6.6 and 6.7 summarize the expanded entropy and power ratios used for unscripted speech, respectively.

The features described in this section were selected for detailed analysis based on prior work, empirical observation of the effort level condition spectral properties (discussed in section 6.1.1), and computational efficiency. All features except LFSD (see below) were analyzed using a 60 msec time window with a 15 msec frame advance. LFSD required a smaller 10 msec frame. Feature descriptions, reasons for considering a feature for detailed analysis, and the analytic results of each feature's ability to provide separation across conditions follow. Specific measurements presented here which indicate a feature's ability to provide separation include 1) the unequal variance sensitivity d_a from signal detection theory (Pashler and Wixted, eds., 2002), and 2) analysis of each feature's mean and 2-sigma variance across each condition. A stronger d_a magnitude for a given feature and condition indicates that the feature has a strong ability to separate that condition. Also, a wider separation of feature means across conditions with minimal overlap within the 2-sigma variance range indicates strong ability of a feature to separate between conditions. The results of this analysis were used to guide the selection of candidate feature combinations for analysis within machine learning models in section 6.2.3 below.

Table 6.6: Summary of entropy ratio features for unscripted male and female speech. Frequency ranges are in Hz. The unshaded portion of the table shows features carried over from scripted speech analysis, and the shaded portion of the table shows new features added for distinguishing creaky voice from other voice qualities.

	Male	Female
HR1	(50-600) /(400-600)	(50-300) /(50-150)
HR2	(50-300) /(400-600)	(50-500) /(500-1000)
HR3	(50-300) /(2000-8000)	(300-800) /(50-300)
HR4	(450-650) / (2800-3000)	(50-500) /(500-1500)
HR5	(50-900) /(300-900)	(50-300) /(2000-8000)
HR6	(50-300) /(50-900)	(450-650) /(2800-3000)
HR7	(50-300)/(1000-2000)	(50-150)/(300-800)
HR8	(50-85)/(50-900)	(50-300)/(1000-2000)
HR9	(50-300)/(600-900)	---

Table 6.7: Summary of power ratio features for males and females. Frequency ranges are in Hz. The unshaded portion of the table shows the features carried over from scripted speech analysis, and the shaded portion shows the new features added for distinguishing creaky voice from other voice qualities.

	PR1	PR2	PR3	PR4
M	(50-900)/(300-900)	(50-300)/(300-900)	(50-300)/(50-900)	---
F	(50-300)/(50-100)	(50-500)/(50-1000)	(300-800)/(50-300)	(50-150)/(300-800)

Figures 6.16, 6.17, 6.18, and 6.19 show the error bar and sensitivity plots, respectively, for male and female unscripted voices. As before, the error bars show the means and 2-sigma variances within a feature, across conditions. The sensitivity plots do not show means and variances directly, but instead provide a quantifiable measure of the ability of a feature to distinguish each condition. A feature does a good job distinguishing a condition if the sensitivity magnitude for the condition is large, and the feature's mean and 2-sigma variance range for that condition has minimal overlap with other conditions.

The error bar plots suggest that creaky voice has qualities which overlap with the other effort levels to a greater degree than whispered voice did. They also suggest that the range of expressivity was larger within the unscripted corpus than in the scripted corpus. The actors collectively had a wide range of expression, but this range was bounded by the Shakespearian speaking style and by the characters they played. The oral history interview speakers, in contrast, were all individuals acting as unique persona, describing their widely-varying individual experiences. The error bar graphs also do not make a strong case for a continuum relationship between creaky voice and the other qualities, particularly for male speakers. The perception studies reinforce this in that creaky voice is rarely, if ever, confused with other vocal types.

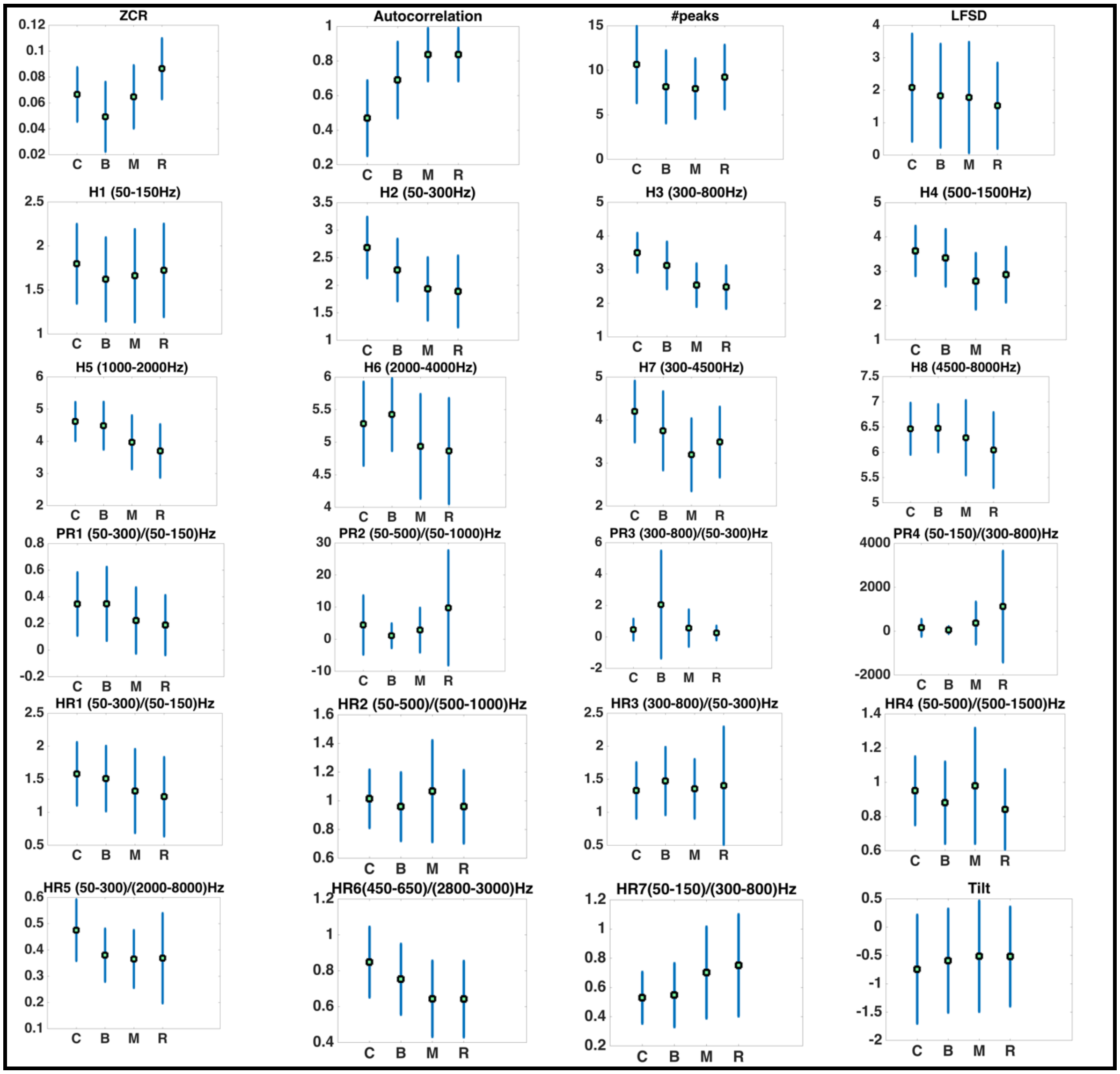


Figure 6.16: Female Feature Error Bar Plots.

As before, the plots show breathiness to be a difficult condition to separate in females, with ZCR again providing the most single-feature breathiness separation. PR3 provides secondary separation, with the entropy features again working together to provide separation across all conditions, including breathiness. The strongest separators for female modal speech were Autocorrelation, PR3, and H5; while the strongest separators for resonant speech were ZCR, AC, and H3. Many features separated creaky voice, including AC, H2, H3, H7, HR5, and HR6.

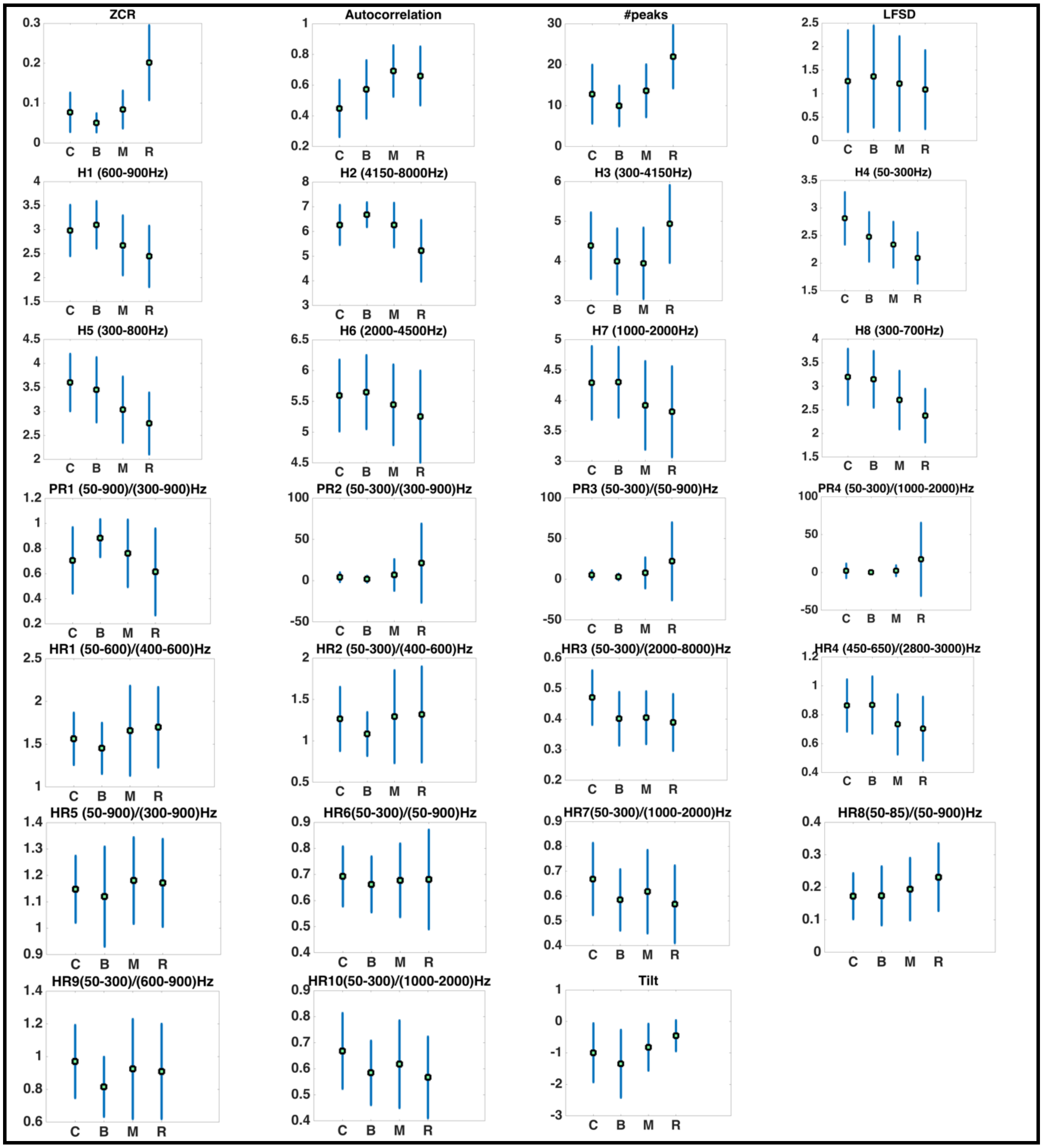


Figure 6.17: Male Feature Error Bar Plots.

For males, breathiness was the most difficult condition to separate, followed by modal voice. AC was the strongest modal separator. ZCR provided primary breathy voice separation for males, and the combined entropy features provided secondary separation (particularly H2). Surprisingly, TILT provided strong separation for the breathy and resonant conditions in males, even given the variability of TILT we observed within condition. ZCR, H8, and H5 provided strong separation for resonant voice.

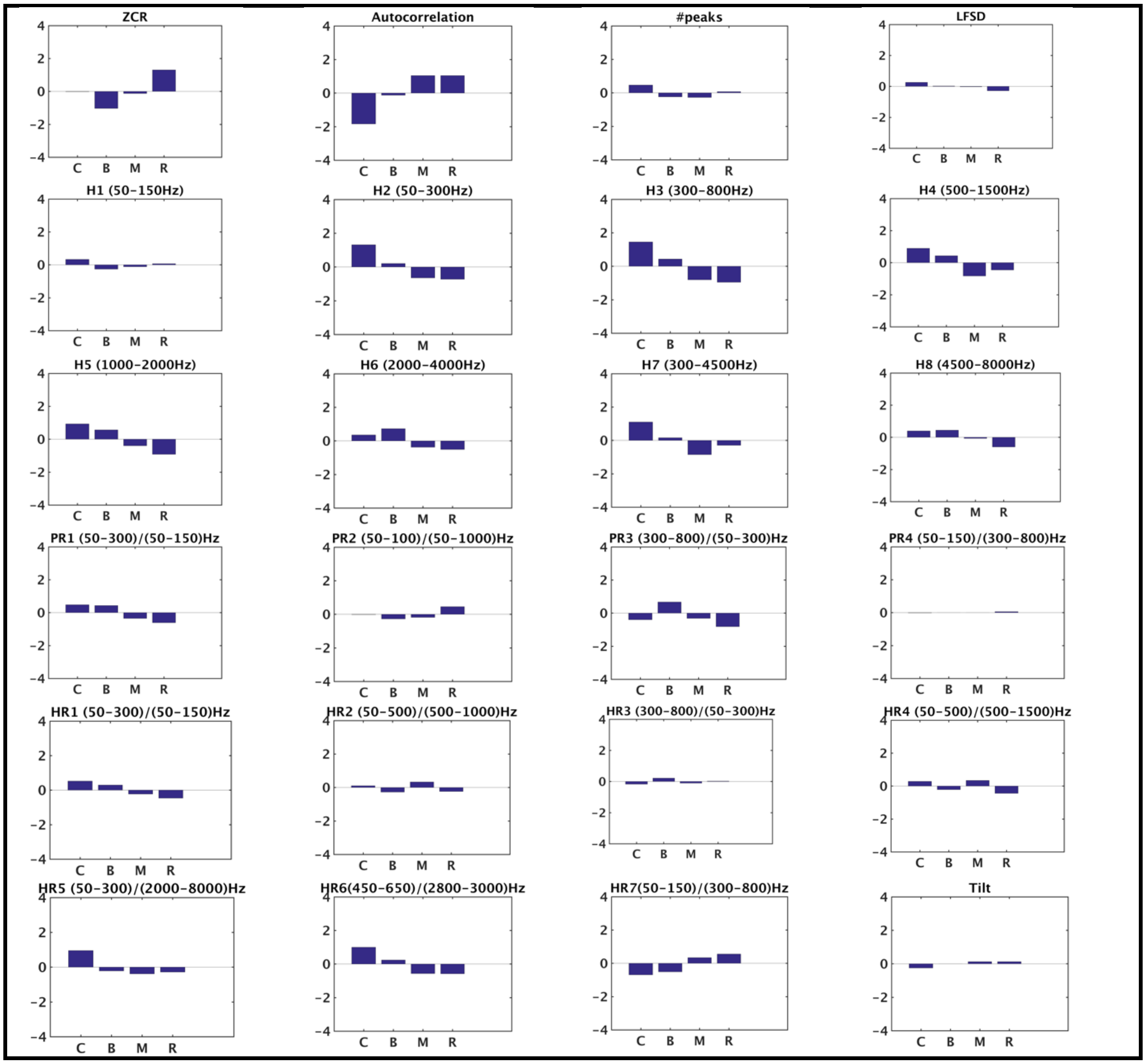


Figure 6.18: Female Sensitivity Plots.

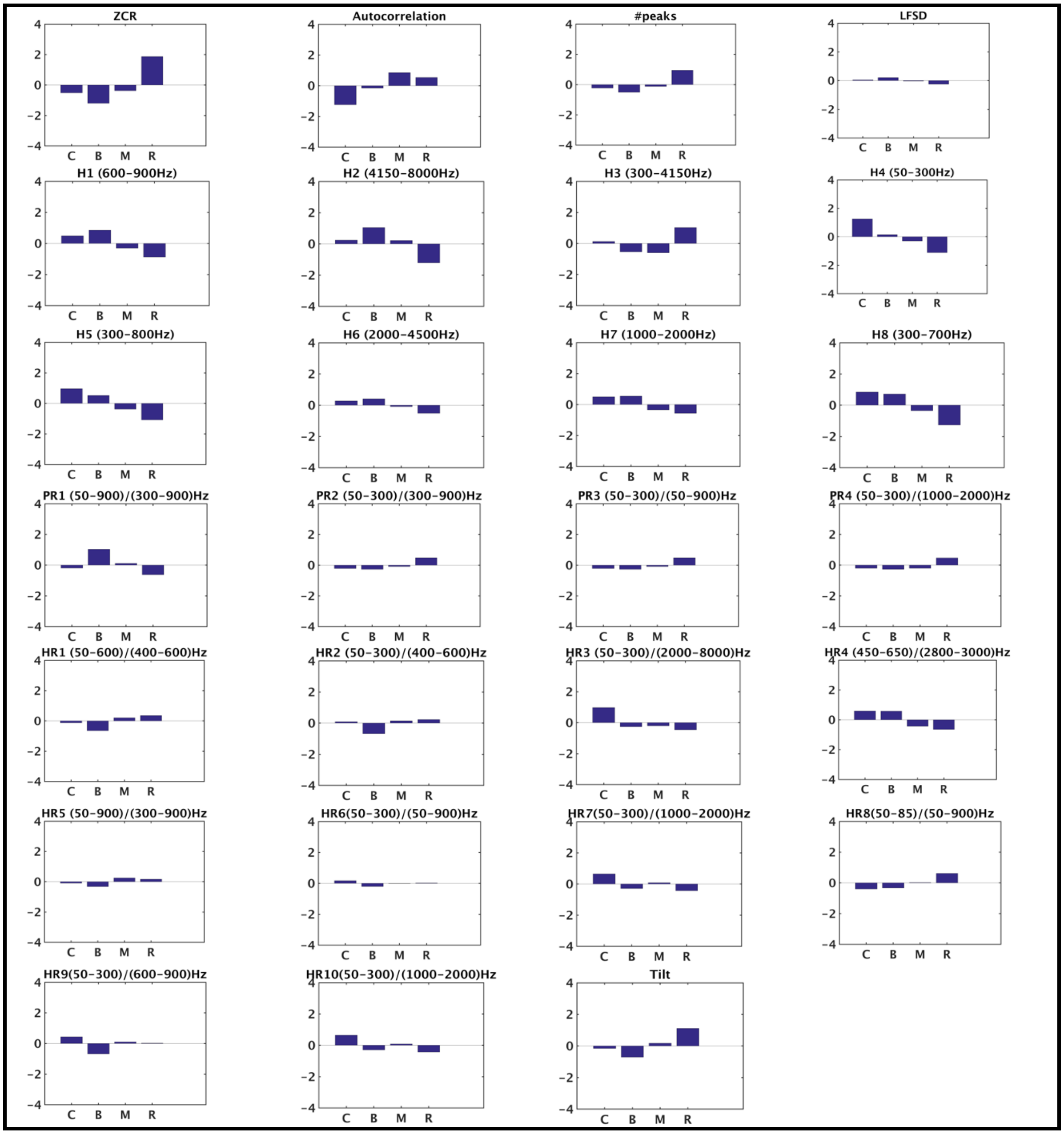


Figure 6.19: Male Sensitivity plots.

6.3.3 Models for Perceived Effort Levels in Unscripted Speech

As before, feature combinations were selected which would best work together to provide maximum separation across conditions. 4-way decision tree classifiers were trained using a series of the most promising feature combination, again pruned to guard against overfitting and to tune performance, and cross-validated to measure performance *sample independence*, *text independence*, and *speaker independence*. The best-performing feature sets from the scripted speech corpora were also tested on the unscripted speech for comparison. To validate sample level independence, 5-way cross validation was used across all speakers and phrases in the corpora, such that each trained model saw none of the test samples. To validate text independence, about 5% of the data from each speaker was reserved for testing and excluded from the models. Speaker independence was validated by holding one speaker out. Results are again presented in terms of precision, recall, and overall accuracy.

Classification Feature Sets: The interactive analysis again revealed which features provided general separation across multiple conditions and revealed the primary separators for each single condition. It also revealed potential difficulties in 4-way classification, which included 1) wider variation in expressivity across the unscripted speech corpora than across scripted speech, 2) multiple classes of breathy and creaky voice to consider, with differing spectral patterns within each perceptual class, 3) a non-continuum relationship between creaky voice and the other voice qualities considered, and because of this non-continuum relationship, 4) greater spectral overlap between creaky voice and the other voice qualities (compared to that of the overlap with whispered voice for unscripted speech). The same secondary technique was applied such that 20 feature combinations each for males and females were evaluated via the cross-validation methods described above. The analysis began with the best-performing feature sets from scripted speech to gain a baseline separation across breathiness, modal speech, and resonance. Other features were selectively added and removed to reinforce separation across all conditions and to boost creaky voice separation, since this condition was new.

Table 6.8: Male and Female Feature Sets. SET1-SET2 (male) and SET6-SET7 (female) are the best-performing feature sets from the unscripted speech analysis, considered here for comparison and to provide baseline separation across breathiness, modal speech, and resonance. SET4-SET6 and SET8-SET10 are the best-performing male and female feature combinations, respectively. Note that the frequency bands are different between males and females, to account for gender differences in the spectrum across conditions. Frequency ranges are given in Hz.

Name	Gender	Features
SET1	Male	ZCR, H1(50-300), H2(300-800), H3(600-900), H4(1000-2000), H5(2000-4500), H6(300-1000), H7(300-4500), PR1(50-900)/(300-900), HR1(50-600)/(400-600), HR2(50-300)/(400-600), #peaks
SET2	Male	SET1 features, plus LFSD
SET3	Male	SET2 features, plus AC, H8(4500-8000), HR4(450-650) / (2800-3000)
SET4	Male	SET3 features, plus TILT

Table 6.8: Male and Female Feature Sets (continued).

Name	Gender	Features
SET5	Male	SET3 features, plus HR7(50-300)/(1000-2000)
SET6	Female	ZCR, AC, H3(300-800), H4(500-1500), H5(1000-2000), H6(2000-4000), H7(300-4500), PR1(50-300)/(50-150), PR2(50-500)/(50-1000)
SET7	Female	ZCR, H1(50-150), H3(300-800), H4(500-1500), H5(1000-2000), H6(2000-4000), H7(300-4500), PR1(50-300)/(50-150), PR2(50-500)/(50-1000), HR1(50-300)/(50-150), HR5(50-300)/(2000-8000), LFSD
SET8	Female	SET6 features, plus HR7(50-150)/(300-800)
SET9	Female	SET6 features, plus H1(50-150), HR6(450-650)/(2800-3000), and HR7(50-150)/(300-800)
SET10	Female	SET6 features, plus PR3(300-800)/(50-300)

Classifier Algorithms: The classifier algorithm is constrained for simplicity again, to emphasize the power of the features to characterize effort levels, and de-emphasize the effects of the classification algorithm. Simple decision trees were used for those reasons, and to remain consistent with the approach used for scripted speech. Again, the classifiers were pruned to factors of 5, 10, 15, and 20, and individually evaluated. Best results for the unscripted classifiers corresponded with pruning factors of 15 or 20 (larger than the factor of 10-15 used in scripted classifiers).

Results: The research questions asked what acoustic features could distinguish across the four levels of effort, from creakiness, to breathiness, to modal speech, and to resonant speech. Table 6.9 summarizes the precision and recall for each condition, along with the global accuracy of the top male and female 4-way decision tree classifiers for the sample, text, and speaker-independent cases. Note that since this is a 4-way classifier, chance is 25%, or 0.25 in the table. Note that the feature sets were again similar between males and females, except for the frequency band ranges.

Overall, the results were consistent between males and females, and across the sample, text, and speaker independent tests. All of these cases performed at slightly above twice chance, and classifier performance degradation was minimal when portions of the text or entire speakers were removed from the training set. While the sample-independent unscripted classifiers were less accurate than the sample-independent scripted classifiers, many of the speaker and text-independent unscripted classifiers performed better than their scripted counterparts. The improved unscripted classifier stability is probably a result of a greater natural expressive range within and across the oral history speech, which guarded against overfitting. The Shakespearian actors, in contrast, tended to exaggerate and enunciate more, but the overall expressive range was bound by the acting style and given parts. In addition, the actors each had a distinct expressive style. Some actors were almost exclusively whispery and breathy; while others were frequently resonant. When one speaker was removed from the training set, therefore, a disproportionate number of training cases for one or two of the classes could be removed from the training set.

The lower performance of the sample-independent case for unscripted speech is probably, in part, a result of the overlap of creaky voice spectral characteristics with some of the other effort level types. For example, aperiodic creaky voice and noisy breathy voice can share high entropy values in the lower bands. Also, periodic creaky voice and resonant voice can share periodic (low-entropy) activity in the upper bands. Creaky voice may often have a much lower F0, but not always, and creaky voice which is periodic and does not have a low F0 shares some characteristics with modal voice. The overall performance for the unscripted classifiers might, then, be improved by adding hierarchy to the creaky and breathy classifiers so that noisy breathy voice is recognized as a separate entity from classic breathy voice. Breathy voice is the worst-performing condition, so improving breathy voice recognition will improve the overall results. Also, each of the four kinds of creaky voice present in the corpus could also be detected separately, or at least, the periodic and aperiodic creaky voice types. Detecting a lower-than-average F0 *for a given speaker* could also be helpful in distinguishing some kinds of creaky voice, as could peak counts in the lower bands (for multiply-pulsed, “doubled” creaky voice). Since breathy voice, modal voice, and resonance are on a continuum together, and creaky voice is not, teasing creaky voice apart from the other voice qualities could help performance approach that of scripted speech.

The differences between male and female accuracies at sample independence ($\chi^2 =0$, df=1, p=0.99) and text independence ($\chi^2 =1.27$, df=1, p=0.26) for the best-performing feature sets (SET5 for males and SET10 for females) were not statistically significant per chi-square test at a significance level of 0.05. The same speaker-independent classifier feature sets were significantly more accurate for males than for females, however ($\chi^2 =44.8$, df=1, p<0.00001).

The difference between sample, text, and speaker independence accuracy within the male SET5 classifier was not statistically significant; nor was the difference between sample and text independence accuracy within the female SET10 classifier. SET10 female sample independence, however, performed significantly better than SET10 female speaker independence (χ^2=8.77, df=1, p=0.003).

Table 6.9: Classifier Precision, Recall, and Accuracy. This table summarizes the mean classification results across all the folds of the two best-performing feature sets for males and females (defined in Table 5), and compares the Sample (SMP), Text (TXT), and Speaker (SPK) Independent Cases. It shows the precision and recall (p/r) for the whispered, breathy, modal, and resonant (W, B, M, R) cases, and the global accuracy (A).

	SMP					TXT					SPK				
	C p/r	B p/r	M p/r	R p/r	A	C p/r	B p/r	M p/r	R p/r	A	C p/r	B p/r	M p/r	R p/r	A
SET1	.39/ .56	.18/ .67	.80/ .42	.54/ .74	.51	.76/ .40	.31/ .73	.58/ .42	.52/ .60	.50	.24/ .37	.11/ .34	.83/ .36	.15/ .39	.37
SET2	.38/ .56	.17/ .60	.81/ .41	.54/ .74	.51	.64/ .34	.36/ .73	.59/ .33	.40/ .63	.47	.52/ .58	.20/ .52	.88/ .48	.28/ .58	.54
SET3	.42/ .59	.20/ .61	.82/ .48	.56/ .76	.56	.76/ .46	.27/ .35	.41/ .60	.53/ .58	.50	.44/ .59	.16/ .59	.82/ .45	.54/ .75	.54
SET4	.47/ .61	.13/ .49	.81/ .50	.57/ .72	.55	.80/ .50	.21/ .35	.30/ .38	.62/ .40	.47	.47/ .64	.15/ .41	.82/ .50	.56/ .72	.55
SET5	.43/ .62	.20/ .60	.84/ .46	.53/ .76	.55	.79/ .60	.23/ .30	.36/ .53	.78/ .65	.54	.43/ .62	.20/ .61	.84/ .76	.53/ .76	.55
SET6	.65/ .70	.22/ .66	.84/ .49	.34/ .60	.56	.55/ .55	.24/ .59	.45/ .40	.75/ .45	.49	.53/ .61	.19/ .38	.85/ .52	.29/ .57	.55
SET7	.67/ .71	.20/ .58	.85/ .49	.35/ .59	.56	.51/ .58	.30/ .67	.37/ .38	.86/ .42	.48	.52/ .58	.20/ .52	.88/ .48	.29/ .57	.54
SET8	.66/ .69	.21/ .64	.86/ .49	.34/ .62	.56	.51/ .59	.22/ .44	.45/ .40	.79/ .51	.50	.49/ .50	.13/ .36	.81/ .49	.28/ .58	.51
SET9	.66/ .67	.21/ .69	.84/ .50	.35/ .62	.56	.50/ .52	.27/ .63	.54/ .50	.78/ .49	.51	.48/ .44	.15/ .46	.79/ .50	.25/ .51	.50
SET10	.66/ .70	.23/ .65	.85/ .47	.32/ .62	.55	.51/ .58	.34/ .59	.52/ .53	.78/ .52	.55	.48/ .53	.15/ .43	.82/ .48	.23/ .52	.49

6.4 Summary

This chapter described the methods used to address research questions RQ4, RQ6, RQ9, and RQ11, along with the experimental results. An interactive analysis process was used to explore the waveform patterns and spectral profiles of whispered, breathy, creaky, modal, and resonant voice, as appropriate, in scripted and unscripted speech. This exploration revealed a continuum relationship from whispering, to breathiness, to modal speech, to resonance; and this pattern was confirmed in both perception and in acoustic feature measurements. Regression analysis reinforces this continuum relationship. Appendix C shows the results of training a neural network model with the male scripted speech data, and performing regression analysis on the resulting predictions (overall R = 0.813). Creakiness did not appear to lie on this continuum, and instead was found to share spectral qualities with other voice quality types in certain bands. A deeper exploration into creakiness revealed four distinct classes of spectra in the corpora which listeners perceived as creaky voice; and these could be grouped into periodic and aperiodic forms of creakiness. Frequently, F0 was lower for creaky voice than for modal speech, but not always. Sometimes multiple F0 were present, or F0 was not discernable at all. This spectral diversity in perceived creaky voice causes confusion in machine recognition, and future work could improve performance by addressing each of these spectral patterns individually. As a specific example, prior research has identified, with respect to modal voice, a reduced F0 and H2>H1 (first and second harmonic relationship) (Yoon et al.). Our research has shown that these relationships hold some of the time, for some kinds of creaky voice, but not for all types of spectral patterns which listeners identify as creaky voice. Furthermore, many of the relationships identified in prior creaky voice literature alone will not do the job of distinguishing among non-modal phonation types. They are meant to distinguish between creaky and modal phonation; and some of the literature acknowledges this (Yoon et al.). In our experiments, F0 and H2-H1 did not improve results when creaky voice was considered as a single type and when used to distinguish creaky voice from non-modal quality. The next steps should take each of these prior findings into account.

Breathiness also had two distinct spectral patterns, a classic profile, and a noisy profile; the noisy profile is likely a result of adjacency to whispered voice on the continuum. Resonant voice, too, had multiple profiles. Again, dealing with each of the spectral types of breathiness and resonance separately may help in the modeling and recognition of breathy and resonant voice and their distinctions from other kinds of modal and nonmodal voices.

The analysis revealed spectral "bands of interest," or frequency ranges, different for males and females, which exhibited different behavior across voice qualities. The proposed feature sets measured periodic and aperiodic behavior across these different bands and used entropy, entropy ratios, and power ratios within and across these bands of interest to distinguish the different voice qualities. Similar feature sets were used for both scripted and unscripted speech, with differences to accommodate recognition of whispered voice in the acted speech corpus and creaky voice in the unscripted corpus.

The proposed feature sets were validated by training 4-way decision tree classifiers and testing them via 4-way cross validation. They were examined for sample independence, text independence, and speaker independence. The acted speech classifiers had a higher recognition rate for sample independence than the unscripted speech, but degraded when text or entire speakers were removed from the training set. The unscripted speech classifier performance remained stable when text and entire speakers were removed from the training set. This was probably a result of wider expressive variation in the unscripted speech than in the scripted speech, and a tendency for the individual actors to focus on just one or two voice qualities in their soliloquys.

While this chapter explored frequently-perceived individual voice qualities and the relationships among them, the next chapter explores the discovery and analysis of entire dimensions of expression, or repeating, co-occurring sets of expressive elements. It extends the detailed analysis of individual voice qualities described here, and places these qualities into a larger context. The next chapter explores relationships between of the voice qualities explored here (whispering, breathiness, resonance, and creaky voice), and other elements in the voice, particularly emotion.

CHAPTER 7: ANALYSIS AND DISCOVERY OF EXPRESSIVE DIMENSIONS IN SPEECH

This chapter describes the methods, experiments, and results associated with the organic discovery and analysis of expressive dimensions in scripted and unscripted corpora. It begins with and uses the listener descriptions of vocal expression collected in Chapter 4; therefore, the process continues to be grounded in the answer to the question, “What do people hear?” The results reveal a collection of expressive modalities, or dimensions, representing frequently co-occurring combinations of emotion, voice quality, prosody, and conversational style. Analysis of these discovered dimensions reveal relationships among these components, particularly between emotion and voice quality, and therefore, address RQ10 and RQ11. Each discovered dimension is modeled and validated. Note that some of this work has been published (Pietrowicz et al., 2017).

7.1 Methods Overview

Discovery of expressive dimensions in a corpus begins with human perception. In the perception studies described in Chapter 5, listeners were presented with sound clips from scripted Shakespearian speech and unscripted oral history interviews and asked to provide keywords describing what they heard, expressively speaking. Their description included a rich range of nuanced emotion descriptors and small, concise sets of prosodic, voice quality, and conversation quality descriptors. These results are summarized in Tables 5.1-5.6. The next steps include 1) *Dimensional Discovery* (using this listener description to discover expressive dimensions present in the corpora), 2) *Dimensional Analysis* (discovering relationships among emotion, voice quality, nonverbal quality, prosody, and conversation quality within these dimensions), 3) *Dimensional Modeling* (training and validation of classifiers to recognize the discovered dimensions), and 4) limited *Dimensional Mapping* (exploring the relationship between these organically-discovered dimensions and the predefined dimensions of affect, arousal, and dominance from prior work).

7.1.1 Dimensional Discovery

To discover repeating patterns of expressive speech, or dimensions, from the given perceptual data, and to discover relationships among perceived keyword qualities, latent semantic analysis (LSA) (Landauer et al., 1998) was used to analyze the distribution of descriptive keywords versus audio clips. The steps to the LSA technique are the following:

1. **Create a matrix**. Put the keyword descriptors from the perception study along the rows, and list the contexts (audio clips) along the columns. Each cell in the matrix contains the number of times which the descriptive keyword occurred in the given context. For best results, remove keywords from the analysis (and matrix) which occur only once. Next, weight each cell with a value expressing the word's importance in the passage and the degree to which the word carries information. This involves taking the log of each cell value, computing the entropy of each word, and dividing each cell value by the row entropy value. This emphasizes the words which bear the most meaning across the corpus.
2. **Apply SVD to the matrix**. This step decomposes the matrix into three separate matrices. One matrix describes the rows (keyword descriptors) as orthogonal factors, and another matrix describes the columns (audio clips) as orthogonal factors. The remaining matrix is a diagonal matrix of weights such that the product of the three matrices gives back the original matrix. Figure 7.1 shows this process.

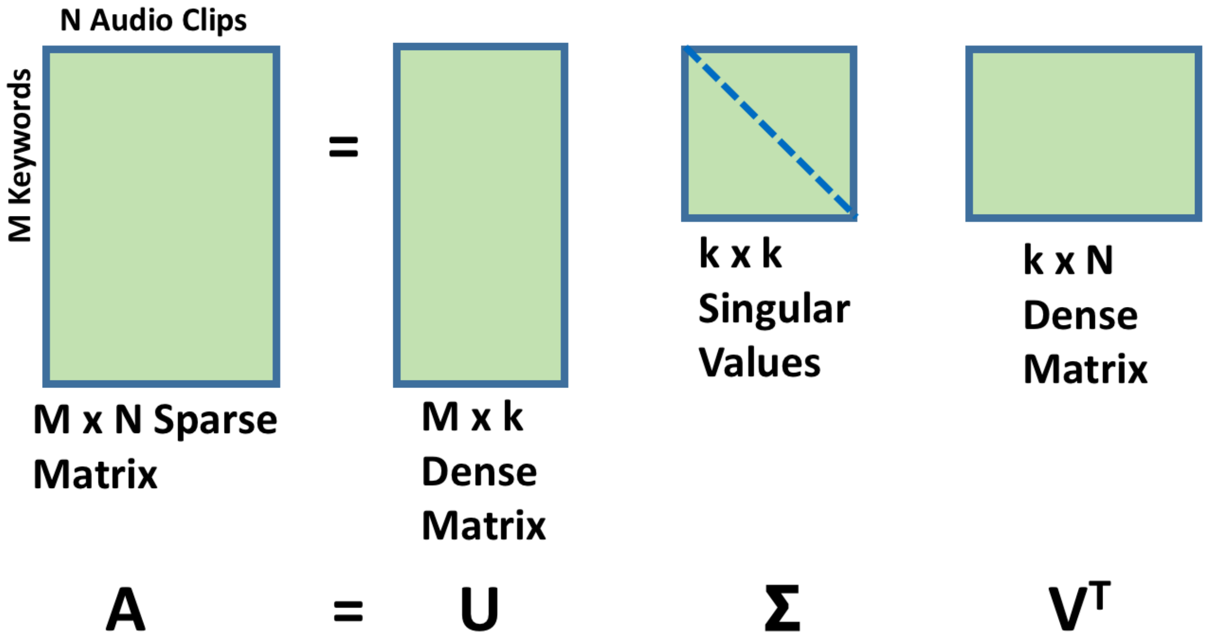


Figure 7.1: The application of SVD to the keyword-audio clip matrix in LSA. This shows the decomposition of the M x N matrix into a singular values matrix of k expressive dimensions, a matrix of the column entities as orthogonal factors and a matrix of the row entities as orthogonal factors.

3. **Reduce the dimensionality**. This step involves zeroing out the lesser dimensions (similar to dimensionality reduction via principal component analysis) (Abdi et al., 2010) and retaining the strongest dimensions present in the data set. Usually this is done by deleting the smallest coefficients in the diagonal matrix. For the scripted and unscripted male and female corpora, examining 12-25 dimensions produced good results; best results were obtained by using 9-13 dimensions. The cutoff is determined empirically. An initial technique for determining the number of dimensions to consider is calculating the correlations among the emotion keywords using successively fewer dimensions. If the correlations make sense at a given level, and that particular level maximizes the number of sensible correlations and minimizes the number of correlations which do not make sense, then it is reasonable to use that many dimensions. A sensible positive correlation, for example, would be one between "joyful" and "happy". The negative correlations should also be sensible (for example, a result that says that "happy" and "sad" are negatively correlated). The full matrix may give poor results because not all the information in the full matrix is equally important. The lower dimensions may have the effect of introducing "noise," or confusion, into the analysis which can make it difficult to discover the major patterns and strong relationships present in the data. Dropping these lower dimensions usually results in improvement in both the number of statistically significant correlations found in the data, and the number of correlations which are sensible. Results may continue to improve as dimensions are dropped until, at some point, the matrix has lost too much important information. When this happens, fewer statistically significant correlations will be found; and an increasing number of nonsensical correlations (such as a strong positive correlation between "happy" and "sad") will be reported. Note that not only can dimensionality reduction improve the analysis, but it can also simplify and speed up the computation.

   This step also requires empirically determining what numeric values correspond with sensible "strong," significant correlations. For emotion-emotion correlations on the unscripted corpora, for example, setting the boundary for a strong positive correlation to 0.7 and the boundary for a strong negative correlation to -0.7 produced reasonable results. Significant results have $p <= 0.05$. Keyword correlations are calculated by taking the Spearman correlation between the rows corresponding to the keywords of interest in the dimensionally-reduced matrix.

A secondary guideline for determining a reasonable number of keywords to consider is requiring at least one strong positive or strong negative keyword correlation in a dimension. The weaker dimensions usually have fewer and fewer strong keyword associations. When a dimension does not have at least one strong keyword association, the dataset usually does not have enough information in it about a particular concept to train a model to recognize that dimension reliably.

4. **Re-compute the weighted keyword-audio clip matrix.** Compute an estimate of the original matrix using only the retained dimensions. The resulting estimate will be a linear combination of values from the dimensions which were retained. Both the audio clips and the keyword descriptors are both related to the retained dimensions, or "abstract expressive dimensions" present in the corpus; therefore, the audio clips and descriptors are related to each other through these abstract expressive dimensions. Figure 7.1 illustrates these relationships. In this way, the descriptors which the listeners gave are described by having a given amount of each retained expressive dimension. The strongest dimension-descriptor relationships describe each dimension. In the typical application of LSA (mapping text to documents), the meanings of the dimensions are not known. Since the keywords are the listener perceptions of vocal expression, the cluster of descriptors most strongly associated with a given dimension can be interpreted to describe that dimension.

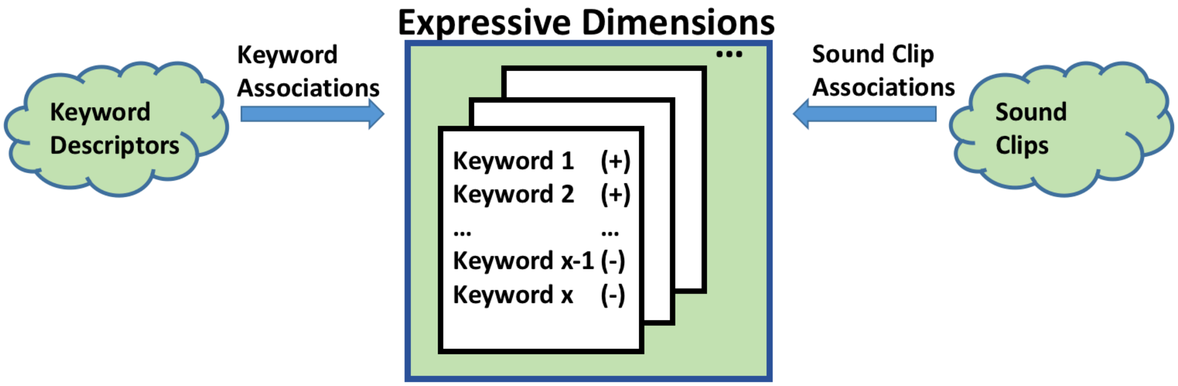


Figure 7.2: The relationship among keyword descriptors, sound clips, and expressive dimensions. The sound clips and descriptors are associated with the expressive dimensions. The strongest associations between keywords and a dimension describe the dimension. Similarly, the strongest associations between sound clips and dimensions show which sound clips most strongly represent a given dimension.

7.1.2 Dimensional Analysis

After the dimensionality has been reduced and the descriptive keyword – audio clip matrix recomputed, the next steps are 1) analyzing the associations between keywords and dimensions, 2) analyzing the associations between audio clips and dimensions, and 3) exploring associations between keywords both within dimensions and globally. Step 1 results in a human-perception-grounded description of each expressive dimension discovered by the LSA process. Step 2 results in identification of audio clips which can be used to train classifiers to recognize each of the expressive dimensions. The final step provides a way to analyze the strength of the relationships among descriptive keywords (for example, the relationship between sadness and creaky voice). Step 3 provides a technique for discovering relationships among the different categories of keywords, such as emotion, prosody, voice quality, conversational quality, and personal human quality. The step-by-step procedure is described below.

1. **Analyze keyword-dimension relationships.** In this step, the descriptors are projected onto each of the top concept dimensions. If A is the weighted keyword descriptor-audio clip matrix with singly-occurring keywords removed, and $A_k = U_k \Sigma_k V_k^T$ is the singular value decomposition of A (considering k dimensions), then $U_k \Sigma_k$ is the projection of the keyword descriptors onto k dimensions. The rows in this $U_k \Sigma_k$ matrix represent the descriptors. The columns are the discovered dimensions, and the matrix

values are the weights. The weights, therefore, describe the association of the descriptors with the dimensions. A strong positive weight indicates a strong positive association of a given descriptor with a dimension, and a strong negative weight indicates a strong negative association of a given descriptor with a dimension. Because the keyword descriptors are human-perception-grounded descriptions of vocal expression, the strong positive and negative keyword associations with a dimension can be interpreted to describe that expressive dimension.

2. **Analyze audio clip-dimension relationships.** In this step the audio clips are projected onto each of the top concept dimensions. If A is the weighted keyword descriptor-audio clip matrix with singly-occurring keywords removed, and $A_k = U_k \Sigma_k V_k^T$ is the singular value decomposition of A (considering k dimensions), then $\Sigma_k V_k^T$ is the projection of the audio clips onto k dimensions. The rows in this $\Sigma_k V_k^T$ matrix represent the discovered dimensions. The columns are the audio clips, and the matrix values are the weights. As with the keyword-dimension relationships, the weights describe the association of the audio clips with the dimensions. The collection of audio clips which have strong positive weights associated with a given dimension can be used to train classifiers to recognize vocal expression belonging to that expressive dimension.
3. **Analyze the associations between keywords.** In this step, the descriptors are clustered into categories of emotion, prosody, voice/nonverbal quality, conversation quality, personal quality, and other quality. Please refer to the co-occurrences of strong positive and negative keyword associations within each discovered dimension (from step 1). This gives a dimension-by-dimension view of relationships among voice quality, emotion, prosody, and conversational quality. This is a good method for finding strong associations between keywords across the entire corpus.

Next, a global view across k dimensions is considered, where k is the number of dimensions retained after dimensionality reduction. Keyword correlations are calculated again by taking the Spearman correlations between the rows corresponding to the keywords of interest in the matrix. To explore the relationships between emotion and voice quality, for example, each voice quality keyword is correlated with each emotion keyword. As in the dimensionality reduction process, strong, statistically significant positive and negative correlations are observed and noted. The correlation value cutoffs obtained during dimensionality reduction are used as a measure for "strong" correlations here as well. This is a good method for discovering correlations which are present in the data, but which do not necessarily occur in the context of discovered dimensions, and do not necessarily occur for all speakers.

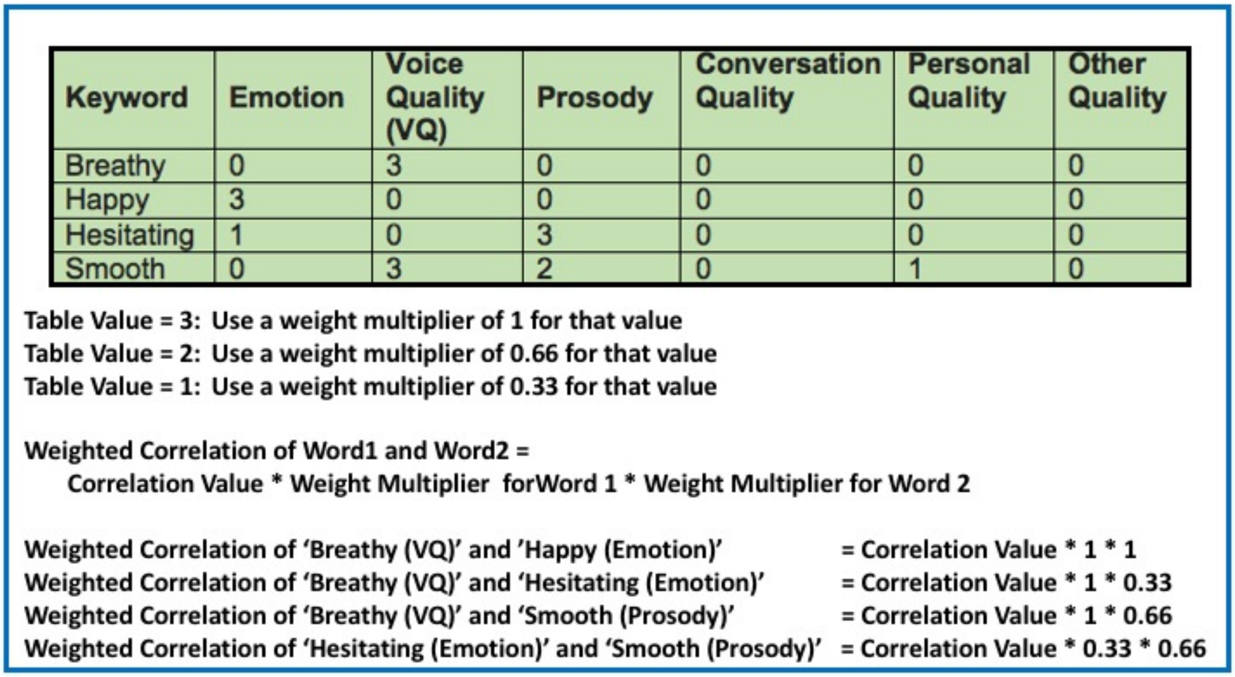
To handle the case in which a word could be associated with more than one category, each word is given a rating of 0-3 in each category. The ratings indicate the following relationships between the keyword and category:

0: Keyword and category are unrelated
1: Keyword and category are weakly related
2: Keyword and category are moderately related
3: Keyword and category are strongly related

The keywords, categories, and ratings form a keyword-category matrix with the keyword labels on the rows and category labels on the columns, as shown in Figure 7.3. In this example, "Breathy" is clearly a voice quality word, and is not considered to be a member of other categories at all. "Hesitating," however, has been designated as a keyword strongly related to prosody and weakly related to emotion. Since these are subjective ratings, two individual raters (L1 English speakers with training in Speech Processing, Linguistics, or Music) categorized each keyword, and a Cohen's kappa rating was calculated to indicate the level of agreement between individuals. This rating system allows flexibility in methods of handling multiple-category membership. The simplest method restricts correlations to words with ratings of 3 each category only.

A method which more accurately considers the strength of each keyword-category relationship applies weight multipliers to the correlation results which are proportional to the strength of the keyword relationship to each category. In the example in Figure 3, Keywords with a '0' relationship generate a '0' multiplier for the correlation calculation, and result in correlation values of 0. This is sensible because a zero implies that a given word is unrelated to a category. On the other end of the spectrum, words with a '3' rating results in a multiplier of '1'. This makes sense because the word is strongly correlated with a given category. Keywords with a '1' relationship generate a '0.33' multiplier for the correlation result, and keywords with a '2' relationship generate a '0.66' multiplier for the correlation result. Figure 7.3 shows examples of this method.

Voice and prosodic qualities are often more directly related to acoustic-level parameters than emotion. For example, perceived pitch and loudness are closely related to F0 and energy. In contrast, the relationship of emotional happiness to these acoustic correlates varies. By finding relationships between emotion and voice quality or prosody, we may be able to leverage the acoustic correlates of voice quality and prosody in the recognition of emotion. To quantify these relationships, we examine the Spearman correlations of emotion-voice quality descriptors and emotion-prosody descriptors across the top k dimensions and within each single dimension, and note the statistically-significant strong positive and negative correlations.



Keyword	Emotion	Voice Quality (VQ)	Prosody	Conversation Quality	Personal Quality	Other Quality
Breathy	0	3	0	0	0	0
Happy	3	0	0	0	0	0
Hesitating	1	0	3	0	0	0
Smooth	0	3	2	0	1	0

Table Value = 3: Use a weight multiplier of 1 for that value
Table Value = 2: Use a weight multiplier of 0.66 for that value
Table Value = 1: Use a weight multiplier of 0.33 for that value

Weighted Correlation of Word1 and Word2 =
Correlation Value * Weight Multiplier forWord 1 * Weight Multiplier for Word 2

Weighted Correlation of 'Breathy (VQ)' and 'Happy (Emotion)' = Correlation Value * 1 * 1
Weighted Correlation of 'Breathy (VQ)' and 'Hesitating (Emotion)' = Correlation Value * 1 * 0.33
Weighted Correlation of 'Breathy (VQ)' and 'Smooth (Prosody)' = Correlation Value * 1 * 0.66
Weighted Correlation of 'Hesitating (Emotion)' and 'Smooth (Prosody)' = Correlation Value * 0.33 * 0.66

Figure 7.3: Correlations can be weighted based on strength of association of a keyword with a category. This sample data shows sample shows the strength of association between four sample keywords and six keyword categories. It defines keyword multiplier values with respect to the weights in the table with the strongest associations having table value of 3 and weight multipliers of 1. In contrast, the weakest associations have tables values of 1 and corresponding weight multipliers of 0.33. To calculate a weighted correlation between two words, multiply the

Figure 7.3 (cont.): correlation value times the weight multipliers for each word/category combination as shown in the sample calculations.

Taking correlations globally does find associations which may not be associated with specific dimensions, and which may not be associated with every speaker. While this is useful, the technique could possibly uncover strong correlations which occur in relatively few speakers, even just one, particularly if the number of speakers in a study is small. This possibility can be explored by examining descriptor correlations within individual speaker data versus descriptor correlations across the remaining speaker set.

7.1.3 Dimensional Modeling

In this step, classifiers are trained to recognize each of the discovered dimensions which were retained after dimensionality reduction. Each audio file which is a strong example of a given dimension is a member of the positive examples for that class, and the other audio files are negative examples. Training samples are balanced via random undersampling, and the resulting models are tested via 5-fold cross validation. Features are considered for inclusion in the model if they have been used in the past in the recognition of the emotions, voice qualities, or prosodic dimensions described by the keywords in the retained dimensions. Features are also considered for inclusion if they have been shown to distinguish affect, arousal, or dominance levels in speech.

7.1.4 Dimensional Mapping

We found it informative to compare the dimensions discovered organically via LSA with the commonly-used predefined dimensions of emotional arousal, valence, and dominance. This step is insight-giving, and it provides a bridge to (and possibly leverage of) prior work along those predefined dimensions. The perceived valence, arousal, and dominance scores (Warriner et al., 2013 and Pietrowicz et al. 2014) of each strong positively and negatively-associated keyword were used to derive the weighted mean and standard deviation scores for arousal, valence, and dominance for each LSA dimension. The boundary for what is considered a “strong” association is corpus-dependent, and therefore determined empirically. For unscripted speech, a word was considered to have a strong positive association for LSA weights >= 0.85 and a strong negative association for LSA weights <= -0.85.

The emotional valence, arousal, and dominance scales ranged from 1-9, with with 1 being the most strongly-negative, unaroused, and non-dominant rating, and 9 being the most positive, aroused, and dominant rating. The weighted score of a dimension was calculated by first inverting the negatively-associated valence, arousal, and dominance scores (around 5.0), and then calculating the weighted average across all the ***strongly-related*** descriptor keywords. In this calculation *n* is the number of strong descriptor associations, w_i is the LSA weight for the *i*th keyword descriptor, and D_i is the dimensional value (valence, arousal, or dominance) for the *i*th keyword descriptor.

$$weighted\ averge = \frac{\sum_{i=1}^{i=n}(w_i * D_i)}{\sum_{i=1}^{i=n} w_i} \qquad (7.1)$$

7.2 Dimensional Analysis of Unscripted Speech

This section describes the explorations, experiments, and results in 1) organic dimensional discovery, 2) dimensional analysis, 3) dimensional modeling, and 4) dimensional mapping (organically-discovered dimensions to the predefined dimensions of affect, arousal, and dominance).

7.2.1 Explorations and Experiments

Organic Dimensional Discovery: In this experiment, we ran LSA over the dataset (male and female data separately) using the keywords collected in the perception studies described in Chapter 5. Male and female descriptor-audio clip matrices were created and weighted, and decomposed via SVD as described in Section 7.1. For females, LSA resulted in the discovery of 61 dimensions. An analysis of the amount of variance covered by accumulating dimensions showed that the first 10, 20, 30, 40, and 50 dimensional factors covered 31%, 51%, 67%, 81%, and 92% of the variance present in the data, respectively. Examining the number of strongly-related keywords associated with each dimension suggested that up to 20 dimensions could be recognized reliably. A dimensionality-reduction to 12 dimensions (following the process described in section 7.1) gave the best emotion correlation conclusions at this level of reduction. Table 7.1 shows the number of statistically-significant correlations which were found at various levels of dimensionality reduction, the number of sensible correlations found at each level, and the percentage of correlations which were determined to be sensible. Two individuals examined the candidate keyword correlations, and deemed them "sensible" or "not sensible." An example of a correlation which is "not sensible" is a positive correlation between "happy" and "sad". Measuring the agreement between these two individuals yielded a Cohen's kappa = 0.76.

Table 7.1: Dimensionality Reduction “Sanity” Check for Unscripted Female Speakers. This table shows the number of dimensions retained in the left column, the number of statistically-significant (p<0.05) strong correlations (rho>=0.7) found between emotion keywords in the center column, and the number of these correlations which were deemed “reasonable” in the third column. These results show poor performance before dimensionality reduction, with increasing performance to a peak of 70 strong, sensible, statistically-significant correlations found at 12 dimensions. Note that 80% of the statistically-significant correlations found at this level are sensible. Performance tapers off after more dimensions are removed, as shown by the decreasing number of sensible correlations and/or a reduction in the percent of correlations found to be sensible.

# Dimensions	**# Statistically-Significant Strong Correlations Found**	**# Sensible Correlations Found**
All (61)	61	16 (26%)
1-40	11	5 (45%)
1-30	17	11 (65%)
1-20	41	25 (61%)
1-15	67	37 (55%)
1-14	70	41 (57%)
1-13	76	53 (70%)
1-12	88	70 (80%)
1-11	113	66 (58%)
1-10	140	88 (65%)
1 - 9	185	107 (57%)

Given the success of finding statistically-significant, reasonable emotion keyword correlations at 12 dimensions, and the strength of the keyword associations projected over the top 12 dimensions, we retained the first 12 dimensions for the dimensional analysis step for females.

The same LSA process resulted in 54 candidate dimensions for the male oral history data. An analysis of the amount of variance covered by accumulating dimensions showed that the first 10, 20, 30, 40, and 50 dimensional factors covered 34%, 56%, 73%, 87%, and 98% of the variance present in the data, respectively. A dimensionality-reduction to 11 dimensions performed the best, providing the largest number of statistically-significant strong correlations with the highest rate of sensible correlation across emotion keywords.

Table 7.2: Dimensionality Reduction “Sanity” Check for Unscripted Male Speakers. This table shows the number of dimensions retained in the left column, the number of statistically-significant (p<0.05) strong correlations (rho>=0.7) found between emotion keywords in the center column, and the number of these correlations which are reasonable in the third column. Percentage of sensible correlations with respect to the number of statistically-significant, strong correlations, is in parentheses.

# Dimensions	# Statistically-Significant Strong Correlations Found	# Sensible Correlations Found
All (54)	19	12 (63%)
1-40	10	2 (20%)
1-30	8	5 (63%)
1-20	16	10 (62%)
1-15	38	18 (47%)
11-14	43	28 (65%)
1-13	48	33 (69%)
1-12	54	34 (63%)
1-11	72	51 (71%)
1-10	110	67 (61%)
1 - 9	128	78 (61%)

Dimensional Analysis: In this experiment, we explored relationships among co-occurring classes of keyword descriptors, particularly between emotion and voice quality keywords. The methods are described in section 7.1, and the results in section 7.2.2 below. The first 12 dimensions for females and the first 11 dimensions for males were retained for this analysis.

Dimensional Modeling: We selected acoustic features for investigation based on the literature, the results of our human perception analysis, and the representation of VQ, prosody, and emotion components in the LSA expressive dimensions. Clips were downsampled to 16Khz, and features were computed based on 60ms frames with a 15msec advance (except LFSD, which required 10msec frames). We mapped the range of emotion keywords onto affect and arousal dimensions, and selected features (a mix of energy, VQ, F0, and spectral features) which have been shown to represent the affect and arousal dimensions (Busso et al., 2009). Typical VQ feature sets include jitter and shimmer, but these are disconnected from human description. To address this, we augmented jitter and shimmer with features which are known acoustic correlates for perceived vocal effort levels (breathy, whispered, and projected voice); these features are entropy, entropy ratios, and power ratios across selected frequency bands which differentiate among vocal qualities in female voices (Pietrowicz et al., 2014). These are also useful for laughter detection. Autocorrelation, low frequency spectral density, and peak count have also been used in the detection of vocal quality, particularly breathiness (Gowda et al., 2013). Tables 7.3 and 7.4 list the acoustic features in each category for females and males, respectively.

Table 7.3: Acoustic features for perceived vocal expression of *female* speech by category. Each feature and its derivative were tested for correlation with LSA dimensions.

Class	**Name**	**Description**
Energy	RMS	RMS Energy
	ZCR	Zero Crossing Rate
	RMS_u	RMS Energy / Mean RMS for clip
	PKRate	Energy peak rate
	PKDUR	Energy Peak Duration
F0	F0	Fundamental Frequency
	F0_u	F0 / Mean F0 for Clip
VQ Support	Jitter	Jitter
	Shimmer	Shimmer
	AC	Normalized Autocorrelation Maximum
	LFSD	Log low frequency spectral density
	PkCount	Number of spectral peaks
	H1	Entropy 50-150 Hz
	H2	Entropy 50-300 Hz
	H3	Entropy 300-800 Hz
	H4	Entropy 500-1500 Hz
	H5	Entropy 1000-2000 Hz
	H6	Entropy 2000-4000 Hz
	H7	Entropy 300-4500 Hz
	H8	Entropy 4500-8000 Hz
	PR1	Spectral Power Ratio(50-300)/(50-150)
	PR2	Spectral Power Ratio(50-500)/(500-1000)
	PR3	Spectral Power Ratio(300-800)/(50-300)
	HR1	Entropy Ratio (50-300)/(50-150)
	HR2	Entropy Ratio (50-500)/(500-1000)
	HR3	Entropy Ratio (300-800)/(50-300)
	HR4	Entropy Ratio (50-500)/(50-1500)

Table 7.3: (cont.)

Class	**Name**	**Description**
VQ Support	HR5	Entropy Ratio (50-300)/(2000-8000)
	HR6	Entropy Ratio (450-650)/(2800-3000)
Spectral	MFCC	Mel-frequency cepstrum coefficients

Table 7.4: Acoustic features for perceived vocal expression of *male* speech by category. Each feature and its derivative were tested for correlation with LSA dimensions.

Class	**Name**	**Description**
Energy	RMS	RMS Energy
	ZCR	Zero Crossing Rate
	RMS_u	RMS Energy / Mean RMS for clip
	PKRate	Energy peak rate
	PKDUR	Energy Peak Duration
F0	F0	Fundamental Frequency
	F0_u	F0 / Mean F0 for Clip
VQ Support	Jitter	Jitter
	Shimmer	Shimmer
	AC	Normalized Autocorrelation Maximum
	LFSD	Log low frequency spectral density
	PkCount	Number of spectral peaks
	H1	Entropy 50-300 Hz
	H2	Entropy 300-800 Hz
	H3	Entropy 600-900 Hz
	H4	Entropy 1000-2000 Hz
	H5	Entropy 2000-4500 Hz
	H6	Entropy 300-1000 Hz
	H7	Entropy 300-4500 Hz
	H8	Entropy 300-700 Hz
	PR1	Spectral Power Ratio(50-900)/(300-900)
	PR2	Spectral Power Ratio(50-300)/(300-900)
	PR3	Spectral Power Ratio(50-300)/(50-900)
	PR4	Spectral Power Ratio(50-300)/(1000-2000)
	HR1	Entropy Ratio (50-600)/(400-600)
	HR2	Entropy Ratio (50-300)/(400-600)
	HR3	Entropy Ratio (50-300)/(2000-8000)
	HR4	Entropy Ratio (450-650)/(2800-3000)
	HR5	Entropy Ratio (50-900)/(300-900)
	HR6	Entropy Ratio (50-300)/(50-900)

Table 7.4: (cont.)

Class	**Name**	**Description**
VQ Support	HR7	Entropy Ratio (50-300)/(1000-2000)
	HR8	Entropy Ratio (50-85)/(50-900)
	HR9	Entropy Ratio (50-300)/(600-900)
Spectral	MFCC	Mel-frequency cepstrum coefficients

Forty binary decision tree classifier sets each, for males and females, were trained to classify each clip sample for membership within LSA dimensions. Features and delta-features were included in the classifiers. The majority class in each fold of the training data was randomly undersampled to achieve a balanced training set. As in the Paralingual Challenges for INTERSPEECH 2009-2013, Average Unweighted Recall (AUR) was used as a validation measure. The best performing classifier sets and classifier performance are reported in the results section.

Dimensional Mapping: In this experiment, the listener descriptors were associated with arousal, affect, and dominance scores as described in section 7.1. Then, the weighted mean and variance of arousal, affect, and dominance were calculated, also as described in section 7.1, for each dimension organically discovered via LSA. Male and female data were analyzed separately. The resulting mappings for the top 15 dimensions for males and females are presented and discussed in the results section below.

7.2.2 Results

Organic Dimensional Discovery: The discovered dimensions for female unscripted speech are described in Table 7.5 below. Note that each of the discovered dimensions is distinctly different from the others. Although dimensions 2 and 3, for example, both include laughter, dimension 2 describes sincere, high-affect, high-energy expression; and dimension 3 describes nervousness and sarcasm (lower affect, and insincerity). Creaky voice is associated with sarcasm and nervousness as well (dimension 3), but is absent in the higher-affect, high-sincerity expression (dimension 2). Also note the difference between dimensions 3 and 4. While both involve nervousness, dimension 3 is high affect; but dimension 4 is low affect. Also note the absence of laughter and creakiness in dimension 4.

Table 7.5: Description of the top-13 LSA Concept Factors in ***Female*** Unscripted Speech. A short description of the factor is given, followed by the strongest positively and negatively-associated keywords. The top 13 dimensions had multiple keyword concepts with strong weights.

#	**Expressive Dimensions (ie, LSA Concept Factors) in Female Speech**
1	**High-variance, opposing qualities.**
Neg:	*Clear, happy, loud, slow, calm, fast, confused, sad, and others.*
2	**Sincere, high energy/affect, with laughter.**
Pos:	Happy, excited, proud, enthusiastic, engaged, joyful, loud, laughing, fast, clear.
Neg:	*Sad, unsure, confused, calm, upset, bored, breathy, mumbling, monotone, quiet.*
3	**Joking, sarcastic, laughing, nervous.**
Pos:	Happy, nervous, joyful, amused, sarcastic, cheerful, embarrassed, hesitant, joking, laughing, creaky.
Neg:	*Clear, excited, confident, proud, loud, sincere.*
4	**Low affect, with nervous energy.**
Pos:	Excited, unsure, nervous, upset, hesitant, confused.
Neg:	*Friendly, calm, upbeat, monotone, unclear, creaky, soft, fast.*
5	**Positive reflection and calm.**
Pos:	Calm, unsure, confused, confident, clear, pauses.
Neg:	*Sad, upset, excited, mumbling, monotone, soft, quiet, fast.*
6	**Lower-energy, medium-affect, quiet, and slow.**
Pos:	Slow, low, quiet, mumbling.
Neg:	*Confused, creaky, thoughtful, annoyed, upset, hesitant.*
7	**High-energy anger/frustration.**
Pos:	Mad, frustrated, angry, anxious, defensive, upset, sad, loud, fast.
Neg:	*Slow, creaky.*
8	**Slow, low-energy sadness and annoyance.**
Pos:	Sad, breathy, annoyed, angry, slow, nasal.
Neg:	*Nervous, bored, unsure, speeding-up, slow, mumbling.*
9	**Loud, anxious, fearful.**
Pos:	Scared, emotional.
Neg:	*Relaxed, soft, angry, unsure, enthusiastic.*
10	**Happy, emotional, serious, and proud.**
Pos:	Happy, serious, proud, emotional, confident.
Neg:	*Calm, excited, interested.*
11	**Even-ness interspersed with laughter.**
Pos:	Calm, serious, thoughtful, monotone, laughing.
Neg:	*Slow, quiet, annoyed.*
12	**Friendly, happy, and relaxed.**
Pos:	Friendly.
Neg:	*Angry, embarrassed.*
13	**High-energy embarrassment, without pauses.**
Pos:	Unsure, embarrassed, passionate.
Neg:	*Pauses.*

Table 7.6: Description of the top-13 LSA Concept Factors in ***Male*** Unscripted Speech. As with the female speech, a short description of the factor is given, followed by the strongest positively and negatively-associated keywords. The top 13 dimensions for male speech also had multiple keyword concepts with strong weights.

#	**Expressive Dimensions (ie, LSA Concept Factors) in Male Speech**
1	**High-variance, opposing qualities.**
Pos:	Clear, happy, slow, fast, sad, loud, soft, low, amused, upset, and others.
2	**Sincere, high-energy, high-affect happiness and enthusiasm.**
Pos:	Excited, fast, loud, happy, lively, enthusiastic, passionate, energetic, thrilled.
Neg:	*Plain, frustrated, breathy, serious, creaky, heistant, calm, slow, low, soft, sad.*
3	**Quiet, calm, steady, confidence and contentment.**
Pos:	Confident, calm, content, plain, and clear.
Neg:	*Sad, amused, breathy, loud, excited, slow, hesitant, upset, frustrated, confused.*
4	**Humorous joking, a mix of expressive deadpan, breathiness, and resonance.**
Pos:	Happy, calm, humorous, amused, deep, breathy, monotone, confident.
Neg:	*Sincere, speeding-up, soft, plain, clear.*
5	**Positive, introspective, thoughtfulness and contentment with creakiness.**
Pos:	Soft, happy, plain, content, thoughtful, creaky.
Neg:	*Anxious, concerned, nervous, tired, calm, serious, fast.*
6	**Friendly, loud, plodding, indifferent conversational speech.**
Pos:	Slow, hesitant, friendly, loud, normal, bored, upbeat, indifferent.
Neg:	*Fast, concerned, frustrated, confused, breathy, sad.*
7	**Thoughtful, calm, hesitant, reflection.**
Pos:	Hesitant, content, slow, thoughtful, confused, clear, calm.
Neg:	*Deep, depressed, low, sad, creaky.*
8	**Soft, fast, caution with contentment.**
Pos:	Content, unsure, fast, quiet, speeding-up, low.
Neg:	*Amused, bored.*
9	**Nervous and unsure, but expressively flat and steady.**
Pos:	Nervous, matter-of-fact, unsure, monotone.
Neg:	*Sad, enthusiastic, slow, fast, creaky.*
10	**Relaxed, with some nervous energy.**
Pos:	Nervous, relaxed.
Neg:	*Speeding-up, upset, amused, monotone, clear.*
11	**Friendly concern or confusion.**
Pos:	Friendly, concerned, confused.
Neg:	*Steady, upset, strong, speeding-up, sad.*
12	**Calm, easygoing, and loud.**
Pos:	Calm, loud, relaxed.
Neg:	*Confident, sad, proud, slow, matter-of-fact.*
13	**Loud, clear, upbeat, and expressively flat.**
Pos:	Clear, monotone, loud, upbeat.
Neg:	*Funny, amused, low.*

The discovered male dimensions are described in Table 7.6. As with the female speech, each dimension is distinct. Dimension 2 is sincere, high-energy, animated happiness (similar to the female Dimension 2, but without the laughter). Note the differences between the male Dimensions 2 and 4. Both are high-affect and happy. Dimension 4, however, is lower-energy and humorous; it seems to have a bit of a "teasing" quality to it. It is closest to the female Dimension 3 (joking and sarcasm). Dimensions 3 and 5 are similar as well; they both describe low-energy contentment. Dimension 5 is punctuated with creaky voice, however, and seems to be more introspective than Dimension 3. Dimensions 5 and 7 are both thoughtful; but Dimension 7 is not creaky, and appears to be lower-affect than Dimension 5.

The wide range of dimensions discovered here reflects the range of expressivity in both the male and female oral history corpora. They also reflect the differences between male and female oral history accounts. Females laugh more, and are more frequently breathy and creaky. In this corpus, they express sarcasm more directly, express hot anger more clearly, and express sadness, depression, fear, and joy more openly. They seem to have a collectively wider range of expressivity for both positive and negative emotions and a wider range of application of prosodic and voice quality devices. Females have a wide range of dimensions for both positive and negative emotion, and they are equally represented in the higher dimensions. The males, however, do not have a single strongly-negative emotion represented in the top 13 dimensions. They have positive and neutral-affect dimensions instead, with varying levels of energy. Given these differences, males and females still share some similar dimensions. They both have a high-energy, high-affect dimension (male and female D2). They both have a "joking" dimension, still high-affect, but not as high as D2 (male D4, female D3). They both have a calm, reflective dimension (male D5, D7, and D12, and female D5); and they both have tension and nervousness (male D9 and D10, female D4). Finally, both male and females share a high-variance dimension with strong, oppositional qualities (male and female D1).

Dimensional Analysis via Joint Association with Expressive Dimensions: To understand the potential relationships among emotion, prosody, and VQ, we first considered the keywords which were strongly and jointly associated with the same dimensions, or expressive modalities. Associations are considered "strongly positive" if the LSA weight is >= 0.85 and "strongly negative" if the LSA weight <= -0.85. For example, "creaky" and "sarcastic" both have strong positive associations with the third dimension in females; and we interpret this result as a strong probability that vocal creakiness accompanies emotional sarcasm, much of the time (for females). Conversely, in females, “angry” has a strong positive association with the 7th dimension, and “creaky” has a strong negative association with the 7th dimension. We interpret this oppositional relationship as a strong probability that creakiness does not accompany anger, most of the time (at least within the scope of that expressive dimension). The results of this analysis are summarized in Table 7.7.

In females, ***breathiness*** is positively-associated with low-energy negative emotion (namely, sadness and anger in dimension 8) and negatively associated with high-energy positive emotions (e.g., enthusiasm, happiness, engagement, pride, joy, and excitement in dimension 2). Males share this relationship between breathiness and negative emotion; but the relationship is less direct than it is for females, because it is implied by the negative association with the descriptors from dimensions 2, 3, and 6. Males did not have any strongly-negative emotions represented in the top 13 dimensions. In males only, breathiness co-occurs with amusement, humor, and sarcasm through dimension 4. They “lighten” their voices when they joke around.

Males and females share some characteristics in their uses of ***clarity*** as a vocal quality as well. It co-occurs in both genders with an emotional state of positive reflection, calm, and contentment in male dimensions 3 and 7 and in female dimension 5. It also occurs in both genders with positive, high-energy, sincere emotion (female dimension 2 and male dimension 13). Clarity does *not* co-occur with sarcasm, insincerity, or “joking around” (male dimension 4 and female dimension 3). In males only, vocal clarity accompanied “relaxed nervousness,” as when someone is working out a problem in their head while talking.

Creakiness had multiple functions which differed between males and females. In females, it accompanied high-affect, high-energy speech with sarcasm and joking (dimension 3). Further examination of these speech segments revealed that the speakers were making negative comments about an experience or were describing a negative experience while outwardly affecting positive emotions in their voices. The conversational topic affect was in direct conflict with the affect reflected in their voices. This opposition between two or more communication channels is a central quality of irony and sarcasm. This combination of creakiness, positive expressive affect in the voice, and negative affect within the word content was present in males as well; but it was not reflected in the top 13 dimensions. It probably did not appear in the top 13 dimensions for males because it occurred less frequently than other kinds of expression; and it appeared to be limited to the younger men who had positive, growth experiences in the military, who were not exposed to combat or other traumatic experiences. Traumatic experiences had the effect of dampening, or “flattening,” all vocal expression.

Note that creakiness did *not* co-occur with sincere high-affect, high-energy speech in either males or females. Creakiness also did not occur with hot anger (dimension 7) in females (which is sensible, because someone in that state would be more likely to become resonant) and did not occur with nervous, high-energy, negative-affect emotions (also sensible because the nervous tension can work against the production of creaky voice). In males, creakiness punctuated introspection, which was otherwise plain, happy, thoughtful, calm, and content speech (dimension 5). Creakiness, did not co-occur with calm, thoughtful speech in males or females, however, if the speaker was also nervous, lacking in confidence, or was confused (dimensions 7 and 9 for males and dimension 4 for females).

Monotone quality was associated with calmness for both males (dimension 4) and females (dimension 11). In females, the calm-monotonality was more serious in affect; in males, it was clearly happy, as in friendly conversation or in just giving information. In both males and females, monotonality did not occur with strong, high-arousal emotions, either positive or negative (dimensions 2, 4, and 5 for females and 2 and 9 for males). In both genders, monotone quality occurs when the speakers are talking about a traumatic or depressing experience, particularly if the speakers have not fully dealt with the trauma in their lives.

The remaining VQs were unique to males or females in the analysis. The male VQs included “***plain,***” “***strong,***” and “***deep.***” Male plain voice aligned with low-energy, positive emotion such as calmness, contentment, thoughtfulness, and confidence (dimensions 3 and 5). Plain voice did not co-occur with high-arousal positive emotions or expressive humor (dimensions 2 and 4). “***Strong***,” or resonant, voice was negatively related to calmness, confusion, and reflection (dimension 11). “***Deep***” is synonymous with low, resonant voice, and occurs in combination with joking, sarcasm, and positive affect (dimension 4). Males use resonance for emphasis in this expressive dimension. “***Deep***” voice was associated with strong, positive, high-energy emotion (dimension 4), but negatively related to thoughtful reflection (dimension 7).

The remaining female VQs (and non-vocal qualities) include ***laughing***, ***mumbling***, ***nasality***, and ***lack of clarity***. ***Laughter*** was most strongly associated with strong, positive emotional affect, both sincere and sarcastic. (dimensions 2 and 3) It was also associated with quiet, thoughtful, positive, even emotion and laughter (dimension 11). In dimension 11, however, laughter co-occurs with monontonality and a positive vocal affect; but the laughter does not reflect humor here. When segments in this dimension are examined more closely, the topics of discussion are often quite serious. ***Mumbling*** was associated with slowness, lowness, and low-energy quiet. It was negatively associated with 1) sincere, high-affect, high-arousal emotions such as joy, excitement, pride, and enthusiasm (dimension 2), 2) low-energy sadness or annoyance (dimension 8), and 3) positive calm and reflection (dimension 5). ***Nasality*** correlated with low-energy sadness and annoyance (dimension 8). Finally, ***lack of clarity*** was negatively associated with low-affect nervousness.

Table 7.7: Joint associations between emotion and voice quality keywords in males and females. This table shows the emotion and voice quality keyword descriptors which were jointly strongly associated with the same dimensions (females on the left, males on the right). The keyword associations were determined by projection of the descriptors across the dimensions discovered via LSA, as described in section 7.2. An association was considered a strong positive association if the keyword-dimension projection matrix weight >= 0.85, and considered to be a strong negative association if the projection matrix weight <= -0.85. The top 13 dimensions were considered here. As an example, the table shows a strong positive association between "breathy" and "sad" for females, and that this association occurred in the 8^{th} dimension. We can also see a strong negative association between "breathy" and "happy" for females, and that this association occurred in the 2^{nd} dimension. When an emotion was correlated with a voice quality across multiple dimensions, the emotion is highlighted in blue.

Females				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Breathy	Angry	(8)	Cheerful	(2)
	Annoyed	(8)	Engaged	(2)
	Sad	(8)	Enthusiastic	(2)
			Excited	(2)
			Happy	(2)
			Joyful	(2)
			Proud	(2)
			Bored	(8)
			Nervous	(8)
			Unsure	(8)
Clear	Cheerful	(2)	Bored	(2)
	Engaged	(2)	Calm	(2)
	Enthusiastic	(2)	Confused	(2)
	Excited	(2)	Hesitant	(2)
	Happy	(2)	Sad	(2)
	Joyful	(2)	Thoughtful	(2)

Males				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Breathy	Amused	(4)	Amused	(2)
	Calm	(4)	Energetic	(2)
	Confident	(4)	Enthusiastic	(2)
	Happy	(4)	Excited	(2)
	Humorous	(4)	Happy	(2)
			Interested	(2)
			Lively	(2)
			Passionate	(2)
			Thrilled	(2)
			Calm	(3)
			Confident	(3)
			Content	(3)
			Sincere	(4)
			Bored	(6)

Table 7.7: (continued)

Females				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Clear	Proud	(2)	Upset	(2)
			Unsure	(2)
	Calm	(5)		
	Confident	(5)	Amused	(3)
	Confused	(5)	Cheerful	(3)
	Unsure	(5)	Embarrassed	(3)
			Happy	(3)
			Hesitant	(3)
			Joking	(3)
			Joyful	(3)
			Nervous	(3)
			Sarcastic	(3)
Creaky	Amused	(3)	Confident	(3)
	Cheerful	(3)	**Excited**	(3)
	Embarrassed	(3)	Proud	(3)
	Happy	(3)	Sincere	(3)
	Hesitant	(3)		
	Joking	(3)	Confused	(4)
	Joyful	(3)	**Excited**	(4)
	Nervous	(3)	Hesitant	(4)
	Sarcastic	(3)	Nervous	(4)
			Unsure	(4)
			Upset	(4)
			Angry	(7)
			Anxious	(7)
			Defensive	(7)
			Frustrated	(7)
			Mad	(7)
			Sad	(7)
			Upset	(7)
Laughing	**Cheerful**	(2)	Bored	(2)
	Enthusiastic	(2)	Calm	(2)
	Excited	(2)	Confused	(2)
	Happy	(2)	Hesitant	(2)
	Hesitant	(2)	Sad	(2)
	Nervous	(2)	Thoughtful	(2)
	Proud	(2)	Unsure	(2)
	Sarcastic	(2)	Upset	(2)
	Amused	(3)	Confident	(3)
	Cheerful	(3)	Excited	(3)
	Embarrassed	(3)	Proud	(3)
	Engaged	(3)	Sincere	(3)
	Hesitating	(3)		
	Joking	(3)	Annoyed	(11)
	Joyful	(3)		
	Nervous	(3)		
	Serious	(3)		
	Calm	(11)		
	Thoughtful	(11)		

Males				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Breathy	Amused	(4)	Friendly	(6)
	Calm	(4)	Hesitant	(6)
	Confident	(4)	Indifferent	(6)
	Happy	(4)	Normal	(6)
	Humorous	(4)	Upbeat	(6)
Clear	Calm	(3)	**Amused**	(3)
	Confident	(3)	Confused	(3)
	Content	(3)	Excited	(3)
			Frustrated	(3)
	Confused	(7)	Hesitant	(3)
	Content	(7)	Nervous	(3)
	Hesitant	(7)	Unsure	(3)
	Thoughtful	(7)	Upset	(3)
	Nervous	(10)	**Amused**	(4)
	Relaxed	(10)	Calm	(4)
			Confident	(4)
	Upbeat	(13)	Humorous	(4)
			Depressed	(7)
			Sad	(7)
			Amused	(13)
			Funny	(13)
Creaky	Content	(5)	Amused	(2)
	Happy	(5)	Energetic	(2)
	Thoughtful	(5)	Enthusiastic	(2)
			Excited	(2)
			Happy	(2)
			Interested	(2)
			Lively	(2)
			Passionate	(2)
			Thrilled	(2)
			Anxious	(5)
			Bored	(5)
			Calm	(5)
			Concerned	(5)
			Nervous	(5)
			Serious	(5)
			Tired	(5)
			Calm	(7)
			Confused	(7)
			Content	(7)
			Hesitant	(7)
			Thoughtful	(7)
			Matter-of-fact	
			Nervous	(9)
			Unsure	(9)

Table 7.7: (continued)

Females				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Monotone	Calm	(11)	Cheerful	(2)
	Serious	(11)	Engaged	(2)
	Thoughtful	(11)	Enthusiastic	(2)
			Excited	(2)
			Happy	(2)
			Joyful	(2)
			Proud	(2)
			Confused	(4)
			Excited	(4)
			Hesitant	(4)
			Nervous	(4)
			Unsure	(4)
			Upset	(4)
			Calm	(5)
			Confident	(5)
			Confused	(5)
			Unsure	(5)
			Annoyed	(11)
Mumbling			Cheerful	(2)
			Engaged	(2)
			Enthusiastic	(2)
			Excited	(2)
			Happy	(2)
			Joyful	(2)
			Proud	(2)
			Calm	(5)
			Confident	(5)
			Confused	(5)
			Unsure	(5)
			Annoyed	(6)
			Confused	(6)
			Hesitant	(6)
			Thoughtful	(6)
			Upset	(6)
			Angry	(8)
			Annoyed	(8)
			Sad	(8)
Nasal	Angry	(8)	Bored	(8)
	Annoyed	(8)	Nervous	(8)
	Sad	(8)	Unsure	(8)
Unclear			Confused	(4)
			Excited	(4)
			Hesitant	(4)
			Nervous	(4)
			Unsure	(4)
			Upset	(4)

Males				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Monotone	Amused	(4)	Amused	(2)
	Calm	(4)	Energetic	(2)
	Humorous	(4)	**Enthusiastic**	(2)
	Happy	(4)	Excited	(2)
			Happy	(2)
	Nervous	(9)	Interested	(2)
	Matter-of-	(9)	Lively	(2)
	fact	(9)	Passionate	(2)
	Unsure		Thrilled	(2)
	Upbeat	(10)	**Enthusiastic**	(9)
			Sad	(9)
			Relaxed	(10)
			Nervous	(10)
			Amused	(13)
			Funny	(13)
Plain	Calm	(3)	**Amused**	(2)
	Confident	(3)	Energetic	(2)
	Content	(3)	Enthusiastic	(2)
			Excited	(2)
	Content	(5)	**Happy**	(2)
	Happy	(5)	Interested	(2)
	Thoughtful	(5)	Lively	(2)
			Passionate	(2)
			Thrilled	(2)
			Amused	(3)
			Confused	(3)
			Excited	(3)
			Frustrated	(3)
			Hesitant	(3)
			Nervous	(3)
			Sad	(3)
			Unsure	(3)
			Upset	(3)
			Amused	(4)
			Calm	(4)
			Confident	(4)
			Happy	(4)
			Humorous	(4)
			Anxious	(5)
			Bored	(5)
			Calm	(5)
			Concerned	(5)
			Nervous	(5)
			Serious	(5)
			Tired	(5)

Table 7.7: (continued)

Females		
VQ Keyword	**Positively Correlated Emotions**	**Negatively Correlated Emotions**

Males				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Strong (resonant, deep)	Amused	(4)	Concerned	(11)
	Calm	(4)	**Confused**	(11)
	Confident	(4)	Friendly	(11)
	Happy	(4)		
	Humorous	(4)	Calm	(7)
			Confused	(7)
			Content	(7)
			Hesitant	(7)
			Thoughtful	(7)

Dimensional Analysis via Keyword Correlations: Emotion-VQ relationships were also evaluated by taking the Spearman correlation between keyword rows in the weighted, dimension-reduced keyword-audio clip matrix as described in section 7.2. Using the process described in section 7.1 and the dimensionality reduction test results in section 7.3.1 as a guide, we systematically reduced the dimensionality of the keyword-audio clip matrix and examined correlations between the emotion and VQ keywords (rows in this matrix). The best results were obtained by retaining relatively few dimensions (between 10-13, out of a total of 61 dimensions for females and 54 dimensions for males). Tables 7.8 and 7.9 show the results of a short series of dimensionality reductions and the resulting emotion-VQ correlations for males and females, respectively. The other variable parameter in this investigation was rho, the correlation value threshold for defining strong positive and negative correlations. The best results were obtained by setting the rho threshold to 0.6 for positive correlations and -0.6 for negative correlations.

Examining correlations in this manner gave a corpus-wide understanding of global correlations among emotion and voice quality (a *global* view), while examining associations among keywords within each dimension provided the details of emotion-VQ relationships in the scope of frequently-repeating expressive modalities, or dimensions. The global view reinforced many of the relationships revealed by close examination of co-occuring keywords within the top 13 dimensions, but not all of them. The global approach also revealed relationships which the dimensional approach did not. The two approaches taken together, therefore, provided the best understanding of the relationships among emotion, voice quality, prosody, conversational quality, and personal qualities, and other qualities in the corpora. As an example, in females, breathiness still correlated with sadness, and creakiness with pensive thoughtfulness. The relationship between creakiness and irony, however, was not revealed by examining descriptor correlations globally.

Table 7.8 summarizes the VQ-Emotion descriptor correlations for ***females***. Some of the new discoveries presented by the global view over the dimensional view include 1) ***growling*** quality associated with aroused unhappiness, 2) ***musicality*** associated with friendliness, 3) ***smoothness*** associated with sincere concern, 4) ***flatness*** (similar to monotone with tension) associated with insincere happiness, 5) ***lightness*** associated not only with flirtation, joking, and playfulness, but also with sarcasm, irony, and disinterest, 6) ***raspy*** quality was not only associated with passion, determination, and seriousness, but also with misery and boredom, 7) ***throatiness*** associated with shock, regret, depression, contemplation, and disbelief, 8) ***wavering*** associated with sadness and discomfort, 9) ***youth*** associated with earnestness, excitement, and high-energy negative emotion, 10) ***old age*** associated with surprise and high-energy positive emotion, 11) ***stuttering*** associated with confidence, cooperation, and driven-ness, and 12) ***southern*** accent associated with thinking, insecurity, caution, and comfort. Perhaps some of the most intriguing relationships are the differences between ***laughter*** and ***giggling*** (showing that laughter itself has multiple dimensions and is multi-functioned), and the flattening of expression with insincerity.

Table 7.9 summarizes the VQ-Emotion descriptor correlations for ***males***. Some of the new discoveries beyond the dimensional view for males include 1) ***growling*** quality associated with focus, indifference, depression, and annoyance, 2) ***flat*** quality associated with boredom, annoyance, calm, and seriousness, 3) ***throatiness*** with nostalgia, happiness, comfort, and surety, 4) ***shakiness*** associated with lack of emotion, confusion, and being sad and upset, 5) ***youth*** associated with earnestness, confusion, and high-energy negative emotion, 6) ***stuttering*** with thinking, confusion, uncertainty, and upset, 7) ***southern accent*** with kindness, pride, and authority, 8) ***dull quality*** with concern, neutrality, and calm, 9) ***gravelly voice*** with seriousness, pleasantness, and joy, 10) ***laughing*** with confidence, 11) ***masculinity*** with content, thoughtfulness, and sternness, 12) ***male*** quality with nervousness, irritation, engagement, and excitement, 13) ***taut*** quality (tightness) with embarrassment, disappointment, and niceness, and 14) ***strong voice*** (resonance) with steadiness, directness, neutrality, indifference, and joviality. Note that growliness, shakiness/wavering, and youthful qualities had strikingly similar associations with emotion between males and females. Flat, throaty, stuttering, southern, resonant, and creaky qualities all had different emotional associations between males and females.

Table 7.8: Emotion/VQ *positive* correlations across dimensions in **female unscripted speech**. This table shows the emotion and voice quality descriptors which correlated with rho >= 0.6 and p <= 0.5. The best quality associations at this rho appear to be across 8-10 dimensions.

VQ	**1-8 Dim**	**1-9 Dim**	**1-10 Dim**	**1-11 Dim**	**1-12 Dim**	**1-13 Dim**
Breathy	Sad	Sad	Sad	Sad	Sad	Sad
Bright						Present
Clear	Sincere	Sincere	Sincere	Sincere	Sincere	
Creaky	Pensive	Pensive	Pensive	Pensive	Pensive	Pensive
Female			Thinking			
Flat	Careful Cheerful Happy Insincere Lighthearted	Careful Cheerful Happy Insincere	Careful Cheerful Insecure Insincere Lighthearted	Cheerful Happy Insincere	Cheerful Happy Lighthearted Insincere	Cheerful Happy Insincere Lighthearted
Giggling	Amused Cheerful Embarrassed Flirty Fun Funny Happy Ironic Joyful	Amused Cheerful Embarrassed Flirty Fun Funny Happy Ironic Joyful	Amused Cheerful Embarrassed Flirty Fun Funny Happy Ironic Joyful	Amused Cheerful Embarrassed Flirty Fun Happy Ironic Joyful	Amused Cheerful Flirty Fun Happy Ironic Joyful	Cheerful Flirty Fun Happy Ironic Joyful
Grandma (old)	Engaged Enthusiastic Excited Proud Shocked Surprised	Engaged Enthusiastic Proud Shocked Surprised	Engaged Enthusiastic Shocked Surprised	Engaged Enthusiastic Shocked Surprised	Engaged Enthusiastic Shocked Surprised	Engaged Enthusiastic Shocked Surprised
Growling	Agitated Miserable	Agitated Miserable Unhappy	Unhappy		Miserable	Miserable
Laughing	Embarrassed Happy	Embarrassed Happy	Embarrassed			
Light	Disinterested Flirty Fun Ironic Joking Playful Sarcastic	Disinterested Flirty Fun Ironic Joking Playful Sarcastic	Disinterested Flirty Fun Ironic Joking Playful Sarcastic	Disinterested Ironic Pensive Sarcastic	Disinterested Ironic Pensive Sarcastic	Disinterested Fun Ironic Sarcastic
Musical	Friendly	Friendly				
Nasal	Agitated Sincere	Agitated	Agitated	Agitated	Agitated	Agitated
Pleasant	Shocked	Shocked				
Raspy	Attentive Bored Calm Certain Determined Firm Matter-of-fact Miserable Passionate Serious Thinking	Calm Determined Firm Matter-of-fact Passionate Matter-of-fact Passionate Serious	Calm Determined Firm	Calm Determined Firm Matter-of-fact	Calm Content Determined Miserable Serious	Content Determined Indifferent Miserable Serious Uneasy

Table 7.8: (cont.)

VQ	1-8 Dim	1-9 Dim	1-10 Dim	1-11 Dim	1-12 Dim	1-13 Dim
Smooth	Concerned Sincere	Concerned Sincere	Concerned Sincere			
Southern	Careful Comfortable Insecure Thinking	Careful Comfortable Insecure Thinking	Careful Insecure Thinking	Careful Insecure Thinking	Careful Thinking	Careful
Strong (resonant, bold)	Calm Confident Confrontational Cooperative Defensive Driven Frustrated Mad Matter-of-fact Passionate Questioning Relieved Stern Worry	Angry Calm Confrontational Defensive Frustrated Mad Passionate Questioning Stern Worry	Angry Calm Confrontational Defensive Frustrated Mad Passionate Stern	Defensive Frustrated Mad Passionate Stern	Angry Defensive Frustrated Stern	Angry Content Frustrated
Stuttering	Comfortable Cooperative Confident Friendly Stern	Confident Cooperative Driven Friendly Matter-of-fact Relieved	Confident Driven Friendly Matter-of-fact	Confident Driven	Confident Driven	Confident
Throaty	Agitated Contemplative Depressed Disbelief Present Regretful Shocked	Contemplative Disbelief Indifferent Present Regretful	Indifferent Regretful Worried	Indifferent Regretful	Disbelief Indifferent Worried Regretful	Indifferent Regretful
Unclear		Defensive		Defensive	Defensive	Defensive
Wavering	Sad Uncomfortable	Sad Uncomfortable	Sad Uncomfortable	Uncomfortable	Sad Uncomfortable	Uncomfortable
Whisper	Apathetic Sad	Apathetic Bored				
Young	Cautious Confused Distracted Hesitant Sadness Relaxed	Cautious Confused Distracted Hesitant Sadness	Cautious Distracted Hesitant Sadness	Cautious Distracted Hesitant Sadness	Cautious Distracted Sadness	Cautious Distracted Sadness

Table 7.9: Emotion/VQ *positive* correlations across dimensions in **male unscripted speech**. This table shows the emotion and voice quality descriptors which correlated with rho >= 0.6 and p <= 0.5. The best quality associations at this rho appear to be across 8-10 dimensions.

VQ	1-8 Dim	1-9 Dim	1-10 Dim	1-11 Dim	1-12 Dim	1-13 Dim
Creaky	Matter-of-fact Uncertain	Friendly				
Dull	Concerned Mellow Neutral	Concerned Mellow Neutral		Concerned Mellow Worried	Concerned Mellow Worried	Concerned Mellow Worried
Flat	Annoyed Apathetic Bored Calm Pleasant Serious	Annoyed Apathetic Bored Calm Serious Tired	Apathetic Bored	Annoyed Bored Neutral Persuasive Steady	Annoyed Bored Persuasive	Annoyed Bored
Gravelly	Joy Nice Pleasant Serious	Joy Matter-of-fact Nice Pleasant Serious	Joy Nice Pleasant Serious	Nice Pleasant Serious	Matter-of-fact	Matter-of-fact
Growly (Growling)	Bothered Depressed Focused Friendly Indifferent Matter-of-fact Perturbed Sincere	Bothered Depressed Focused Friendly Indifferent Perturbed Sincere	Focused Indifferent	Focused Indifferent Matter-of-fact	Focused Indifferent	Focused Indifferent
Laughing	Sure	Sure	Sure	Sure	Sure	
Male	Engaged Excited	Irritated Nervous	Irritated Nervous	Irritated Nervous		
Masculine	Content Thoughtful Stern	Content Thoughtful	Content Thoughtful	Content	Content Stern	Content Stern
Monotone	Matter-of-fact	Matter-of-fact	Matter-of-fact	Matter-of-fact		
Mumbling (Mumbly)	Afraid Assertive Certain Depressed Distraught Emotional Irritated Mellow Unemotional Unhappy Unsure Worried	Afraid Assertive Certain Depressed Distraught Emotional Mellow Unemotional Unsure Worried	Afraid Assertive Certain Distraught Emotional Mellow Normal Unemotional Unsure Worried	Afraid Certain Distraught Emotional Mellow Uncertain Unemotional Unsure	Distraught Unemotional	Afraid Distraught Unemotional
Pleasant	Nostalgic Serious	Nostalgic Serious	Serious	Serious	Serious	
Shaky	Confusion Emotionless Sad Upset	Confusion Emotionless	Emotionless Upset	Emotionless	Emotionless	Emotionless
Southern	Authoritative Kind Pride	Authoritative Kind Pride	Authoritative Kind	Kind Proud	Proud Stern	Proud

Table 7.9: (cont.)

VQ	1-8 Dim	1-9 Dim	1-10 Dim	1-11 Dim	1-12 Dim	1-13 Dim
Strong (Resonant, Deep)	Focused Indifferent Matter-of-fact Neutral Steady Tired	Cheerful Indifferent Jovial Matter-of-fact Neutral Steady Tired	Indifferent Neutral Normal Steady Tired	Neutral Steady Matter-of-fact	Neutral Steady	Neutral Steady
Stuttering	Confused Thinking Uncertain Unsure Upset	Confused Thinking Uncertain Upset	Thinking	Thinking	Thinking	Thinking
Taut	Disappointed Embarrassed Nice	Disappointed Embarrassed Nice	Nice			
Throaty	Confident Happy Interested Nostalgic Sure	Confident Happy Interested Nostaltic Sure	Comfortable Confident Content Happy Sure	Comfortable Content Happy Sure	Happy	
Young (Youthful)	Angry Anxious Confused Dragging Earnest Excited Irritated Nervous Upset	Angry Anxious Dragging Ernest Excited Irritated Nervous Upset	Angry Dragging Ernest Irritated Nervous Upset	Angry Dragging Earnest Excited Irritated Nervous	Angry Dragging Earnest Excited Irritated Nervous	Angry Dragging Earnest Excited Irritated Nervous

Dimensional Modeling: Table 7.10 shows the ability of two representative feature sets to discern dimensional membership for female audio clips in LSA dimensions 2-13. SET2 is minimal but representative, and includes RMS, ZCR, RMS_u, F0, F0_u, Jitter, Shimmer, and MFCCs. Inclusion of deltas did not significantly change the result. SET1 extended this base to include additional features in support of VQ and NQ (LFSD, H1, H3-H7, PR1-PR2, HR1, and HR5), which improved results in 7 of the 12 perceptual dimensions.

Table 7.11 shows the ability of two representative feature sets to discern dimensional membership for male audio clips in LSA dimensions 2-15. SET1 contains RMS, ZCR, RMS_u, F0, F0_u, # peaks, normalized autocorrelation maximum, jitter, shimmer, LFSD, H1-H8, PR1, HR1, HR2, HR4, and MFCC 1-12. SET2 contains all of the SET1 features, and HR7.

Table 7.10: Dimensional Classifier Performance for Females. This table shows the Average Unweighted Recall (AUR) in % for each dimension's binary classifier. ***SET1 Content***: RMS, RMS_u, ZCR, F0, F0_u, Jitter, Shimmer, LFSD, H1, H3-7, PR1-2, HR1, HR5, and MFCC1-12. ***SET2 content***: RMS, RMS_u, ZCR, F0, F0_u, Jitter, Shimmer, and MFCC1-12.

LSA #	SET1 AUR	SET2 AUR	LSA #	SET1 AUR	SET2 AUR
2	78.5	75.5	8	65.5	59.6
3	59.5	57.5	9	78.0	75.5
4	80.5	80.5	10	67.5	63.0
5	61.5	66.0	11	61.5	55.0
6	69.0	70.5	12	64.0	68.5
7	65.0	62.0	13	57.0	57.0

Table 7.11: Dimensional Classifier Performance for Males. This table shows the Average Unweighted Recall (AUR) in % for each dimension's binary classifier. ***SET1 Content***: RMS, RMS_u, ZCR, F0, F0_u, #peaks, Autocorrelation, Jitter, Shimmer, LFSD, H1-H8, PR1, HR1, HR2, HR4, and MFCC1-12. ***SET2 content***: Set1 content plus HR7.

LSA #	SET1 AUR	SET2 AUR	LSA #	SET1 AUR	SET2 AUR	LSA #	SET1 AUR	SET2 AUR
2	80.0	74.0	8	61.0	65.3	14	74.2	79.2
3	63.1	65.1	9	62.0	54.0	15	79.2	86.1
4	65.0	63.4	10	78.8	72.1			
5	59.0	57.0	11	71.5	81.9			
6	65.0	64.5	12	65.3	71.5			
7	79.2	83.3	13	62.0	67.1			

Dimensional Mapping: The mean and variance of affect, arousal, and dominance are calculated for each expressive dimension discovered via LSA using the calculated norms from the Warriner/Kuperman/Brysbaert (Warriner et al., 2013) database. This database contained weights, on a scale of -7 to 7, of almost 13,000 words and their perceived associations with affect, arousal, and dominance. On this scale a "-7" represented the strongest possible negative association of a word with a dimension; and a "+7" represented the strongest possible association of a word with a dimension. Had a more complex analysis been desired, or had listeners given their perceptions in the form of natural language instead of simple keywords, a LIWC analysis (Tausczik and Pennebaker, 2010) could have been a reasonable alternative to gauging affect, arousal, dominance, or any other category which LIWC directly or indirectly supports (it supports over 80 categories). LIWC considers the full language model with all parts of speech (not just keywords), and during analysis, counts the occurrences of words associated with each LIWC category, and provides a summary of occurrence rates across each category. The mechanical turk keywords were, however, just keywords, and not natural language. The LIWC model also counts keywords associated with categories, and does not distinguish the degree of association between a specific keyword and a category the same way the Warriner database does.

Affect, arousal, and dominance alone do singly differentiate the LSA dimensions, but the combination of affect and arousal does differentiate the 15 dimensions examined in Figure 7.4. The affect and dominance dimensions are highly correlated, particularly when compared to the correlations between either affect and arousal or dominance and arousal. Because of these relationships, using either the dominance or affect dimension together with arousal is sufficient for mapping between the organically-discovered dimensions from LSA and affect, dominance, and arousal.

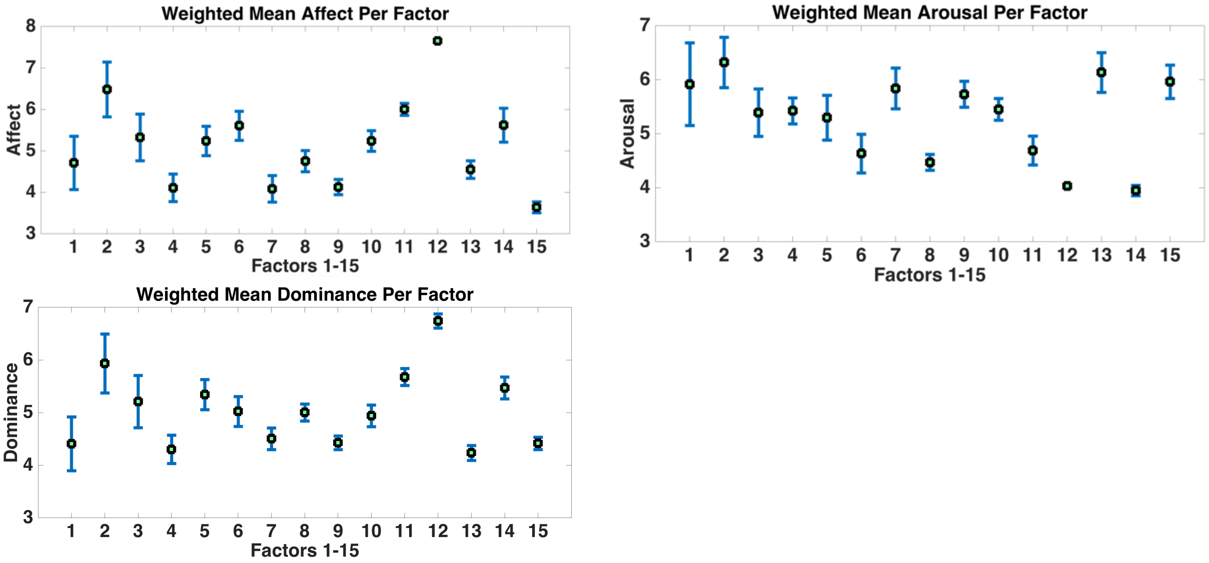


Figure 7.4: Error bar graphs show mean and variance of perceived valence, arousal, and domenance within LSA concept factors for females. Each of these factors vary from low to high on a scale of 1-9 (Warriner et al., 2013). Factors overlap within arousal and affect individually, but differentiate when a combination of affect and arousal is considered. Note the similarity of the shape of the affect and dominance results; only dimensions 5 and 6 break the shape similarity between the two graphs. Affect, arousal, and dominance values were linked to keywords, then weighted according to the projection of each keyword onto LSA factor space.

7.3 Dimensional Analysis of Scripted Speech

This section describes the discovery, analysis, and validation of dimensions in the scripted, Shakespearian corpora. We followed the perception-grounded methods given in Section 7.1, and executed them on the Shakespearian, scripted male and female corpora. Even given the similarities in the speaking styles between males and females, and given the topic similarity, the methods revealed different expressive dimensions present in the male and female speech. The experiments and results for dimensional discovery are discussed below.

7.3.1 Explorations and Experiments

Organic Dimensional Discovery: Latent Sematic Analysis (LSA) was applied to the keyword descriptors, and a matrix derived following the process discussed in section 7.1. In order to determine a level of dimensionality reduction which yielded sensible results, we correlated emotion keywords, and evaluated the strong positive and negative relationships for plausibility as we successively eliminated weaker dimensions. These results showed poor performance before dimensionality reduction, with increasing performance to a peak of 53 strong, sensible, statistically-significant correlations found at 11 dimensions in females. Note that 90% of the statistically-significant correlations found at this level are sensible. Performance tapered off after more dimensions were removed, as shown by the reduction in the percent of correlations found to be sensible. For the males, sensible emotion keyword correlation peaked at about 19 sensible correlations found at 13 dimensions. These results were used to help determine the dimensionality reduction suitable for evaluating relationships between emotion and voice quality.

Table 7.12: Dimensionality Reduction "Sanity" Check for ***Scripted Female*** Speakers. This table shows the number of dimensions retained in the left column, the number of statistically-significant ($p<0.05$) strong correlations ($rho>=0.7$) found between emotion keywords in the center column, and the number of these correlations which appear reasonable in the third column. Examples of unreasonable correlations would be "happy" and "sad" or "furious" and "content".

# Dimensions	# Statistically-Significant Strong Correlations Found	# Sensible Correlations Found
All (34)	29	27 (93%)
1-30	55	43 (78%)
1-20	18	15 (83%)
1-15	31	24 (77%)
1-14	34	27 (79%)
1-13	39	34 (87%)
1-12	52	45 (87%)

Table 7.12: (cont.)

# Dimensions	# Statistically-Significant Strong Correlations Found	# Sensible Correlations Found
1-11	59	53 (90%)
1-10	76	64 (84%)
1 - 9	108	89 (82%)
1 - 8	140	85 (61%)

Table 7.13: Dimensionality Reduction "Sanity" Check for ***Scripted Male*** Speakers. This table shows the number of dimensions retained in the left column, the number of statistically-significant (p<0.05) strong correlations (rho>=0.7) found between emotion keywords in the center column, and the number of these correlations which are reasonable in the third column.

# Dimensions	# Statistically-Significant Strong Correlations Found	# Sensible Correlations Found
All (24)	95	57 (60%)
1-20	14	12 (86%)
1-15	19	14 (74%)
1-14	22	15 (68%)
1-13	23	19 (83%)
1-12	35	25 (71%)
1-11	45	31 (69%)
1-10	45	31 (69%)
1 - 9	63	39 (62%)
1 - 8	100	62 (62%)

Dimensional Analysis: In this experiment, we explored relationships among co-occurring classes of keyword descriptors, particularly between emotion and voice quality keywords. The methods are described in section 7.1, and the results in section 7.3.2 below. The first 11 dimensions for females and the first 13 dimensions for males were retained for this analysis.

Dimensional Modeling: We used the same feature sets for male and female Shakespearian acted voices which were used in the modeling of unscripted expressive speech in oral history interviews. Tables 7.3 and 7.4 list the acoustic features in each category for males and females.

7.3.2 Results

Organic Dimensional Discovery for Females: The discovered dimensions for ***female scripted speech*** are described in Table 7.14 below. Note that each of the discovered dimensions is distinctly different from the others. Also note the differences between the dimensions discovered in dramatic, Shakespearian scripted speech and unscripted oral history interviews.

Table 7.14: Dimensions discovered in female, scripted Shakespearian speech. The top 13 dimensions are described here via a summary statement (in bold), and the strongest positive and negative keyword associations. Positive keyword associations are given in plain type, and negative associations are given in italics.

#	Expressive Dimensions (ie, LSA Concept Factors) in Female Speech
1	**High-variance, weakly-weighted, opposing qualities.**
Neg:	*Angry, breathy.*
2	**Low-energy nervous worry and fear, with a breathy, whispered, dark, shaky, haunting, spooky, aged quality. Speech is slow and soft overall, with some variation in speed.**
Pos:	Scared, afraid, anxious, worried, tense, haunted, prayer, alert, desperate, soft, slow, hushed, low, slowing-down, rushed, whispering, breathy, quiet, spooky, shaky, old, deep, breathless, eerie, haunted, wizened, intense, dark.
Neg:	*Passionate, angry, forceful, powerful, excited, confident, emotional, dramatic, emotional, aggressive, spitting, loud, empathic, bold, strong, expressive, screaming, shouting, overacting, clear.*
3	**Higher-energy, dramatic, anxious, nervous fear with a rapid pace and a whispered, shaky, quality.**
Pos:	Excited, anxious, dramatic, scared, nervous, tense, desperate, excitement, fearful, pleading, suspenseful, demanding, fast, speeding-up, soft, whispering, shaky.
Neg:	*Sad, determined, serious, upset, furious, concerned, possessed, spiteful, indignant, fearless, defiance, chilled, urgent, worried, slow, quiet, deathly, intense, clear.*
4	**Dramatic, passionate, worry, fear, and sadness, with a breathy, shaky, quiet quality.**
Pos:	Scared, anxious, worried, sad, emotional, desperate, determined, afraid, frightened, excited, passionate, urgent, fearful, dramatic, slow, breathy, shaky, intense, quiet.
Neg:	*Mad, angry, prayer, haunted, powerful, alert, crazy, serious, vengeful, anger, soft, distinct, rushed, speeding-up, slowing-down, low, hushed, dark, old, deep, raspy, wizened, haunted, eerie, spooky, clear, whispering.*
5	**Forceful, passionate, fearful, desperate aggression, with a raspy, whispery, spitting, dark quality and slow pace**.
Pos:	Passionate, emotional, desperate, aggressive, pleading, forceful, demanding, serious, eager, horny, passion, afraid, scared, spitting, slow, whispering, overacting, old, raspy, dark.
Neg	*Excited, upset, worried, frightened, anxious, dramatic, cheerful, speeding-up, fast, crescendo, empathic, steady, rushed, whisper, steady, terse, screaming.*

Table 7.14 (cont.)

#	Expressive Dimensions (ie, LSA Concept Factors) in Female Speech
6	**High-energy, nervous, serious, passionate, sadness and concern. Quality includes theatrical, expression, breathlessness, resonance, shouting, varying loudness, rapid pace, and an eerie, raspy, wizened quality.**
Pos:	Serious, nervous, alert, excited, boisterous, sad, emotional, concerned, determined, passionate, haunted, prayer, anxious, zealous, crazy, loud, fast, soft, expressive, theatrical, deep, breathless, shouting, eerie, haunted, wizened, raspy.
Neg	*Anger, mad, forceful, urgent, vengeful, scared, cold, deliberate, angered, sharp, punctuated, crescendo, whispering, intense, spitting, terse, whisper, growling, strong.*
7	**Aggressive, spiteful, high-energy upset and anger, with varying speed, and whispered, breathless, spitting quality.**
Pos:	Excitement, fear, emotional, urgent, intensity, excited, furious, desperate, chilled, defiance, fearless, indignant, spiteful, upset, nervous, concerned, aggressive, angry, fast, slow, low, whispering, intensity, overacting, deathly, breathless, spitting.
Neg:	*Passionate, stern, frightened, tense, contemptuous, confident, scared, exciting, assertive, interested, calm emphatic, thoughtful, deliberate, soft, intense, strong, clear, shaky, rehearsed.*
8	**Sharp, cold, sad, anger with a slow, whispered, intense quality.**
Pos:	Biting, content, deliberate, melancholy, interested, mad, emotional, cold, curious, sharp, calm, sad, passionate, fast, hushed, low, slowing-down, steady, slow, whisper, strong, whispering, steady, intense, expressive.
Neg:	*Tense, forceful, dramatic, urgent, happy, anxious, contempt, demanding, furious, crazy, cheerful, energetic, determined, distinct, acting, growling, spooky, shaky.*
9	**Dramatic, high-energy anger and fear with variable speed and variable quality, ranging from whispering to screaming.**
Pos:	Angry, dramatic, powerful, afraid, pissed, exciting, determined, biting, stressed, stern, contemptuous, fast, empathic, slow, screaming, whispering, quiet.
Neg:	*Excited, upset, desperation, vengeful, serious, forceful, anger, cold, content, happy, calm, melancholy, emotional, steady, soft, mysterious, whisper, breathy, deep, intense.*
10	**High-energy, strong, forceful, bold happiness.**
Pos:	Happy, determined, confident, urgent, tense, demanding, excited, suspenseful, aggression, anger, panic, anxious, forceful, strong, whispering, clear, bold.
Neg:	*Frightened, afraid, emotional, desperation, vengeful, passionate, dramatic, upset, mysterious, hoarse, deep.*
11	**High-energy, dramatic expression with mixed-emotions. Speech is soft, terse, and forceful.**
Pos:	Dramatic, sad, emotional, afraid, fear, powerful, melancholy, excited, concerned, forceful, happy, content, worried, interested, soft, crescendo, quiet, terse.
Neg:	*Serious, panicked, scared, desperate, determined, urgent, anxious, cold, speeding-up, loud, low, halting, mysterious, intense.*

Table 7.14 (cont.)

#	**Expressive Dimensions (ie, LSA Concept Factors) in Female Speech**
12	**Upset, vengeful, powerful anger. Resonant, dark, strong, breathless quality.**
Pos:	Urgent, upset, dramatic, afraid, angered, pleading, vengeful, nervous, powerful, low, clear, strong, breathless, dark.
Neg:	*Passionate, tense, stern, deliberate, angry, sharp, anxious, fear, exciting, pissed, rushed, soft, slowing-down, whisper.*
13	**Theatrical, forceful, aggressive anger, with varying speed, hushed volume, and varied quality (hoarseness, shakiness, resonance).**
Pos:	Excited, frantic, forceful, aggression, boisterous, zealous, anger, speeding-up, slowing-down, slow, hushed, quiet, bold, shaky, hoarse, intense, theatrical.
Neg:	*Anxious, mad, sad, crazy, desperate, worried, exciting, dramatic, fear, thoughtful, passionate, excitement, calm, distinct, dark, rehearsed.*

Note the nuanced dimensions of fear (which include low-energy, high-energy/high-speed, high-energy/slow speed, fear and sadness, and aggressive fear) and anger (similar range of variation). The remaining two dimension include mixed-emotion/high-drama and a single positive-affect dimension of happiness and boldness. Affect is overwhelmingly negative here.

Organic Dimensional Discovery for Males: The discovered dimensions for ***male scripted speech*** are described in Table 7.15 below. Note that each of the discovered dimensions is again distinctly different from the others, and that they represent a wider range of affect and arousal levels that was present in the female scripted speech.

Table 7.15: Dimensions discovered in male, scripted Shakespearian speech. The top 13 dimensions are described here via a summary statement (in bold), and the strongest positive and negative keyword associations. Positive keyword associations are given in plain type, and negative associations are given in italics.

#	**Expressive Dimensions (ie, LSA Concept Factors) in Male Speech**
1	**High-variance, strongly-weighted, opposing qualities.**
Pos:	Happy, confident, calm, excited, nervous, hesitant, content, confused, bored, serious, sad, amused, unsure, friendly, upbeat, tired, relaxed, upset, anxious, matter-of-fact, thrilled, funny, enthusiastic, normal, frustrated, neutral, energetic, indifferent, steady, thoughtful, lively, humorous, interested, uncertain, sincere, mellow, scared, proud, authoritative, slow, loud, fast, low, soft, speeding-up, quiet, clear, plain, creaky, breathy, monotone, strong, deep, gravelly, stuttering.
2	**Genuine, high-energy, loud, fast, confident, happiness.**
Pos:	Excited, happy, lively, enthusiastic, thrilled, passionate, amused, energetic, interested, engaged, confident, fast, loud.
Neg:	*Sad, calm, hesitant, confused, bored, tired, serious, unsure, frustrated, indifferent, slow, low, soft, quiet, creaky, breathy, plain, monotone.*

Table 7.15: (cont.)

#	**Expressive Dimensions (ie, LSA Concept Factors) in Male Speech**
3	**Confident, calm, soft, positive, contentment in plain voice.**
Pos:	Confident, calm, content, steady, soft, plain, clear.
Neg:	*Confused, nervous, unsure, frustrated, upset, hesitant, excited, amused, sad, Anxious, slow, loud, breathy, stuttering.*
4	**Confident, calm, insincere happiness and amusement with varying quality, including monotone, resonance, and breathiness.**
Pos:	Happy, calm, humorous, amused, confident, upbeat, proud, cheerful, monotone, deep, breathy, reminiscent.
Neg:	*Sincere, nervous, excited, sad, soft, speeding-up, fast, low, clear, plain.*
5	**Content, happy, thoughtfulness in soft, slow, clear voice, punctuated by creakiness.**
Pos:	Happy, content, thoughtful, uncertain, soft, plain, creaky, clear.
Neg:	*Bored, serious, calm, nervous, tired, concerned, anxious, fast, low.*
6	**Indifferent, bored, upbeat friendliness, with varying speed, creakiness, and growling.**
Pos:	Hesitant, friendly, normal, bored, upbeat, indifferent, proud, matter-of-fact, slow, loud, speeding-up, creaky, growling.
Neg:	*Sad, confused, frustrated, concerned, serious, content, unemotional, fast, breathy.*
7	**Calm, hesitant, confused, calm thoughtfulness, spoken slowly and clearly.**
Pos:	Hesitant, content, thoughtful, confused, calm, slow, clear.
Neg:	*Sad, depressed, friendly, scared, low, soft, creaky, deep, growling.*
8	**Mumbling, content, and unsure, low and soft, with varying speed.**
Pos:	Content, unsure, fast, quiet, speeding-up, low, slow, mumbling.
Neg:	*Bored, amused, funny, humorous, clear.*
9	**Nervous, unsure, and scared, in soft monotone.**
Pos:	Nervous, matter-of-fact, unsure, scared, quiet, soft, monotone.
Neg:	*Enthusiastic, sad, bored, frustrated, fast, slow, creaky.*
10	**Loud, fast, upbeat, stuttering nervousness.**
Pos:	Relaxed, nervous, upbeat, content, loud, fast, stuttering.
Neg:	*Amused, upset, irritated, steady, normal, annoyed, speeding-up, quiet, monotone, clear, strong.*
11	**Loud, friendly, relaxed concern, worry, and confusion.**
Pos:	Concerned, confused, friendly, tired, worried, mellow, uncertain, loud.
Neg:	*Sad, upset, steady, emotional, speeding-up, slow, strong, breathy.*
12	**Calm, humble, and hesitant.**
Pos:	Calm, relaxed, hesitant.
Neg:	*Matter-of-fact, proud, pride, sad, confident, surprised, nervous, gravelly.*
13	**Fear with upbeat affect, and monotone and stuttering quality.**
Pos:	Upbeat, scared, loud, clear, monotone, stuttering.
Neg:	*Amused, funny, pleasant, low, slow, pleasant.*

The male scripted dimensions show much more variability across the emotional spectrum than the female scripted speech. While the topic is still dramatic (suicide vs. murder), 5 of the dimensions could be considered to have positive affect. Overall, affect is significantly higher for the males than for the females. While the female dimensions had higher negative affect, they did not include a sarcasm or irony dimension. The male dimension 4 reflects upbeat qualities, but is not sincere, and therefore could describe sarcasm and irony. The male dimensions reflect 5 nuanced contentment and calm variants, 2 dimensions of nervousness, and only 1 dimension of fear. Clearly, the dimensions discovered are dependent on the expression in the corpus and the listeners' interpretation of that expression.

Dimensional Analysis via Joint Association within Expressive Dimensions in Female Scripted Speech: Examination of strongly-related, co-occurrent descriptors within each dimension revealed expressive relationships among voice quality, emotion, prosody, personal quality, and conversation interaction type. This discussion, as with the unscripted speech, will focus on the relationships between voice quality and emotion, but this time, within expressive Shakespearian, scripted speech. Tables 7.14 and 7.15 present dimensional summaries along with all of the descriptors most strongly associated with each of the top 13 dimensions. Appendix D provides the details of the strong emotion-VQ relationships. Highlights of VQ-emotion relationships within dimensions are summarized in this section. For ***females***, ***breathlessness or breathiness*** was positively associated with fear, anxiety, nervousness, and desperation across multiple dimensions. Depending on the specific dimension in which breathiness occurred, the intensity and affect varied. On one end of the spectrum (Dimension 2, or D2), the affect was low, but arousal was too, with the behavior described as "prayer-like" and "haunted." In D6, the arousal was high, with the presence of higher-energy descriptors such as "boisterous," "zealous," and "crazy." High-energy breathlessness was associated with increasingly hostile emotions, such as aggression, anger, vengefulness, and drama.

"***Eerie-ness***" and "***spooky***" quality shared the lower-arousal emotion associations with breathiness, that is, fear, anxiety, nervousness, and desperation. Eerie and spooky were negatively associated with hot anger, urgency, aggression, power, and vengeance.

"***Deepness***" in female voices represented a lowered pitch and a higher degree of resonance on the continuum. It was positively associated with genuine happiness, humor, and confidence. It was negatively associated with forcefulness, anger, and high emotional variance. Deepness had a secondary function in signaling lower-affect passion and determination.

"Shakiness" was associated with fear, drama, desperation, anxiety, excitement, and tension. It was negatively associated with anger, prayer, aggression.

"***Forceful-ness***" in women was synonymous with either resonance or tension in the voice. In this corpus it was associated with high-energy aggression, passion, power, and excitement. It was usually associated with negative emotions, such as sadness, fear, melancholy, and desperation; but it could also be associated with happiness, and with high-energy neutral emotions such as urgency, zealousness, and boisterousness.

Growliness had a negative association in females with sadness, nervousness, zealousness, anger, passion, and intense, raging emotion. The opposite implies positive association with positive, low-arousal emotion, such as relaxed calm; but this is not explicitly stated.

"***Hoarse-ness***" was positively associated with frantic aggression and excitement, and hot anger. "***Raspiness***" was associated with seriousness, passion, and high-ranging emotions (imagine the roughness in a voice when a person becomes emotional and tears up).

Female ***"Screaming"*** was positively associated with anger, drama, contempt, determination, and stress and negatively associated with fear, passion, calm, and happiness. ***"Shouting"***, surprisingly, was not a synonym for screaming, and did not have similar associations with emotion. Shouting was positively associated with determination, passion, and haunted seriousness/sadness/concern; it had negative associations with anger.

"***Whispering,***" like many of the VQs listed here, had multiple classes of expressive emotional association in females. The first kind of association in female acted speech was with fear, anxiety, and nervousness, and the second, with anger and comtempt. The other associations were variations of these two basic groups of emotions, across 8 different dimensions.

"***Old***" or "***wizened***" sounding female voices were associated with fear, anxiety, aggression, passion, and desperation. These specific relationships were probably a reflection of the corpus content and do not necessarily hold across other corpora.

Dimensional Analysis via Joint Association within Expressive Dimensions in Male Scripted Speech: The male associations between VQ and emotion share some similarities, but also have distinct differences, many which are likely to be the result of corpora content, particularly across acted speech. For example, male "***breathiness*** corresponded with an upbeat, happy, calm, and humorous affect, in contrast to the female breathiness. Breathiness was negatively associated with worry, fear, and anxiety in the male corpus.

"Clarity" in quality was, overall, associated with calm, contentment, and thoughtfulness. It varied in its association with confidence and confusion, according to dimension, and could be either happy or neutral in affect.

"Creakiness" appeared in the male scripted speech (but not the female scripted speech), probably because of corpus content. In the speech samples we examined, creakiness corresponded with positive affect and thoughtfulness, which occurred when a speaker tried to recall prior events. Creakiness frequently occurred with boredom when the speaker was just relating information. As an incidental observation, creakiness frequently occurred when the topic of conversation was a negative, but not traumatic, memory. To document this relationship between conversation topic and creakiness formally, this study would need to be augmented with topic content data.

"***Deep***" quality in males (lowered pitch and increased resonance) was associated with happy, confident, upbeat, and humorous affect, similar to the function in female scripted speech. Male "***growliness"*** was associated with friendliness, upbeat affect, and indifference; it was negatively correlated with seriousness, confusion sadness, and frustration. Depending on the expressive dimension examined, it was also negatively correlated with calm, hesitancy, thoughtfulness, and confusion. Both genders had a negative association between qualitative growliness and emotional sadness and calm.

"***Monotone***" quality in males (not present in the female sample) had multiple functions. Many speakers would flatten their expressivity and take a matter-of-fact expressive stance when relating or recalling fear and nervousness. Other speakers would flatten expressivity when calm and upbeat; this could be a natural stance for people who were just relating facts.

"**Strong**" speech (resonance) had negative association with nervousness and worry (tense emotions), and a negative association with low-energy positive affect. This is sensible; nervous tension and low energy work against production of strong, resonant voice.

"***Plain***" and "***Steady***" speech shared an association with calm, confident, contentment. Both of these VQs, as Appendix D shows, had variable expressive modalities. Plain speech, especially, had a range of positive and negative affect; while steady speech was clearly not nervous, upset, or sad.

Dimensional Analysis via Keyword Correlations: As with unscripted speech, the Emotion-VQ relationships were evaluated by taking the Spearman correlation between keyword rows in the weighted, dimension-reduced keyword-audio clip matrix as described in section 7.3. Tables 7.16 and 7.17 show the results of systematically reducing the dimensionality of the keyword-audio clip matrix and examining the correlations between emotion and VQ keywords. The dimensionality reduction test results in section 7.3.1 guided the dimensionality reduction here. The best results were obtained by taking a similar number of dimensions to the unscripted speech: between 10-13 out of a total number of 24 dimensions for males and 34 dimensions for females. Note that although the number of dimensions retained was similar in scripted and unscripted speech, the number of retained dimensions in scripted speech represented a greater percentage of the expressive variance because the scripted speech had fewer dimensions. The parameter rho (used for defining strong correlations) was again set at -0.6 for negative correlations and 0.6 for positive correlations.

Examining the correlations between emotion and VQ keywords again provided a corpus-wide, high-level, global understanding of relationships between emotion and VQ, while examining associations within specific dimensions provided a low-level, more detailed view within specific expressive modalities. The global view reinforced much of what was found by examining within-dimension associations. Additional relationships and differences found by examining the global view are summarized here.

Notable differences in females included the strong association of ***hoarseness*** with frantic fear, not anger. ***Whispering*** had strongest associations with curiosity and sadness from the global point of view, and ***growling*** was associated with both anger and happiness. ***Old*** or ***wizened*** voices also had a haunted, prayerful quality, as did ***raspy*** voices.

Additional relationships between female VQ and emotion included ***rehearsed or acted*** voices being associated with contempt. “Rehearsed” voices tended to be low-affect (sad and stern), while “acted” voices were high-affect (happy and energetic). ***Theatrical*** voices were high-energy and were associated with boisterousness and alertness. By describing the Shakespearian voices in these terms, listeners were probably acknowledging the expressive exaggerations in Shakespearian acted speech, when compared to everyday expression. ***Deathly*** (deathlike, with resonance) voices were associated with fury, and ***boldness*** (high in resonance) with passion and anger. ***Spitting*** voices were full of anger, aggression and depression. ***Steadiness,*** in contrast, was associated with lower-arousal curiosity, contentment, and melancholy affect.

Notable differences from what was found by the within-dimension analysis in males included the association of ***growling*** with friendliness, indifference, and upbeat affect. ***Monotone*** voices were perceived as simply rushed from a global point of view, and ***strong*** voices were authoritative, confident, and emotional.

Additional associations found in male voices via global correlations included ***boldness and booming*** quality with confidence, seriousness, and authority. ***Deep*** and ***sonorous*** voices were confident; ***crisp*** voices were happy, and ***British*** voices were bored. ***Forceful*** voices were authoritative, confident, determined, and emotional (in strength of emotion and range).

Flat (low variance with high tension) voices were exasperated, and ***bland*** voices (low variance with low tension) were deliberate, depressed, and indifferent. Both ***sinister*** and ***spooky*** voices were anxious. The sinister voices were also annoyed and scared, while the spooky voices were soulful. ***Raspy*** voices were hopeful and pensive in males, while ***whispered*** voices were unsure. The "***male***" quality was interestingly, simply happy.

Table 7.16: Emotion/VQ *positive* correlations across dimensions in **female scripted speech**. This table shows the emotion and voice quality descriptors which correlated with rho >= 0.6 and p <= 0.5. The best quality associations at this rho appear to be across 8-10 dimensions.

VQ	1-8 Dim	1-9 Dim	1-10 Dim	1-11 Dim	1-12 Dim	1-13 Dim
Acting	Contempt Dramatic Energetic Frightened Happy Tense	Contempt Energetic Happy Tense	Contempt Energetic Happy Tense	Contempt Energetic Happy Tense	Contempt Energetic Tense	Contempt Energetic
Bold	Angered Excited Forceful Passion Powerful	Angered Forceful Passionate Powerful	Angered Forceful Passion Powerful	Forceful Passion Powerful	Passion Powerful	Forceful Passion Powerful
Breathless	Afraid Nervous	Nervous	Nervous	Nervous Pleading	Nervous	Nervous
Deathly	Chilled Furious Pissed	Chilled Concerned Furious Pissed	Chilled Concerned Pissed	Chilled Pissed	Chilled Furious	Chilled Furious
Deep	Alert					
Eerie	Alert	Alert	Alert	Alert	Alert	Alert
Forceful	Aggressive Anger Angry	Aggressive Anger	Aggressive Anger	Anger	Aggressive Anger	Anger

Table 7.16: (continued)

VQ	1-8 Dim	1-9 Dim	1-10 Dim	1-11 Dim	1-12 Dim	1-13 Dim
Growling	Anger Angered Dramatic Forceful Happy	Anger Happy Forceful				Anger
Hoarse	Frantic Frightened Scared	Frantic Frightened Scared	Frantic Frightened	Frantic Frightened	Frantic Frightened	Frantic
Intensity	Desperate Emotional Excitement Fear Overacting	Emotional Excitement Fear Overacting	Excitement Fear Overacting	Excitement Fear Overacting	Excitement Fear	Excitement Fear
Mysterious	Cold Upset	Cold Vengeful	Cold Desperation	Desperation		
Old	Haunted	Haunted	Haunted			
Overacting	Aggressive Desperation Emotional	Emotional Excitement	Emotional	Emotional	Intensity	Emotional
Raspy	Haunted Prayer	Haunted Prayer	Haunted Raspy	Haunted Prayer	Haunted Prayer	Alert Haunted Prayer
Rehearsed	Content Contemptuous Exciting Interested Sad Stern	Contemptuous Interested Sad Stern	Contemptuous Stern	Contemptuous Sad		
Screaming	Biting Powerful	Angered Powerful Biting	Angered Biting Powerful	Angered Biting Powerful	Angered Biting Powerful	Anger Biting Powerful
Shaky	Afraid Frantic Frightened Scared	Afraid Frantic Frightened Scared	Frantic Scared	Afraid Frantic Scared	Frantic Scared	Frantic Scared
Shouting	Biting, Excited		Powerful	Powerful		Powerful
Spitting	Aggressive Anger Desperation	Aggressive Anger Forceful	Aggressive Anger	Aggressive		
Spooky	Haunted Nervous Prayer	Nervous				
Steady	Content Curious Interested Melancholy Sad	Content Curious Interested Melancholy	Content Interested Melancholy	Content	Melancholy	
Terse	Angered, Contempt, Happy	Happy				
Theatrical	Alert Boisterous Stressed	Alert Boisterous	Alert Boisterous	Alert Boisterous	Boisterous	Boisterous
Whisper	Curious Melancholy	Curious Melancholy	Curious Melancholy	Curious		
Whispering	Pleading	Pleading				
Wizened	Alert Prayer	Alert Haunted Prayer	Alert Haunted Prayer	Alert Haunted Prayer	Alert Haunted Prayer	Alert Haunted Prayer

Table 7.17: Emotion/VQ *positive* correlations across dimensions in **male scripted speech**. This table shows the emotion and voice quality descriptors which correlated with rho >= 0.6 and p <= 0.5. The best quality associations at this rho appear to be across 8-10 dimensions.

VQ	**1-8 Dim**	**1-9 Dim**	**1-10 Dim**	**1-11 Dim**	**1-12 Dim**	**1-13 Dim**
Acting	Angry					
Bland	Deliberate Depressed Indifferent	Contemplative Deliberate Depressed Indifferent	Content Relaxed	Content Relaxed		
Bold	Authoritative Confident Serious	Authoritative Confident	Authoritative Confident	Authoritative Confident	Authoritative Confident	Authoritative
Booming	Authoritative Confident	Authoritative Confident	Authoritative Confident	Authoritative Confident	Confident	Authoritative Confident
British	Bored	Bored	Bored	Pensive		
Crisp	Happy Resigned	Happy		Happy		Happy
Deep	Confident					
Direct	Angry	Angry				
Flat	Exasperated	Exasperated	Exasperated			
Forceful	Authoritative Confident Determined Emotional	Authoritative Confident Determined Emotional	Authoritative Confident Emotional	Authoritative Emotional	Authoritative Confident Emotional	Confident Emotional
Growling	Hurried Neutral Reflective Yearning	Hurried Reflective Yearning	Hurried			
Male	Firm Happy	Happy				
Manly	Exasperated					
Monotone	Hurried	Hurried				
Mumbling	Hurried Reflective Yearning	Hurried Yearning	Hurried Yearning	Yearning	Hopeful Yearning	Hopeful
Pleasant	Confident Happy	Confident Happy	Confident Neutral	Confident	Confident	Confident
Raspy	Hopeful Tired	Hopeful Pensive	Hopeful Pensive			
Seductive	Purposeful	Purposeful	Passive			
Sinister	Annoyed Anxious Disinterested Intense Scared Unemotional Wistful	Annoyed Anxious Scared Intense	Annoyed Anxious Scared Intense	Annoyed Anxious Scared Unemotional Intense	Annoyed Scared Intense	Commanding
Sonorous	Confident	Confident	Confident	Confident	Confident	Confident
Spooky	Anxious Dramatic Questioning	Anxious	Anxious	Anxious Soulful	Anxious	Anxious Soulful
Strong	Authoritative Confident Emotional Passive Sure	Authoritative Confident Emotional	Authoritative Confident	Authoritative Confident	Confident	Authoritative

Table 7.17: (continued)

VQ	1-8 Dim	1-9 Dim	1-10 Dim	1-11 Dim	1-12 Dim	1-13 Dim
Unintelligible	Commanding Irritated Resigned Unsure	Commanding Irritated Unsure Resigned	Commanding Resigned Irritated	Commanding Firm Irritated		Commanding Irritated
Weak	Indifferent			Neutral		
Whisper	Unsure					
Whispering	Unsure	Unsure	Secretive Unsure	Unsure	Unsure	
Whispery	Secretive	Secretive	Unsure	Secretive Unsure	Secretive Unsure	Secretive Unsure

Dimensional Modeling: Table 7.18 shows the ability of two representative feature sets to discern dimensional membership for female scripted speech audio clips in LSA dimensions 2-13, and Table 7.19 shows the ability of two representative feature sets to discern dimensional membership for male scripted speech audio clips. The feature set content is identical to that of the unscripted speech and is given again in the table annotations. For females, the simpler feature set (SET 2) produces superior results overall. For SET 1, only 1 dimension had recall rates over 70%, but for SET 2 9 dimensions did. The results for male speech were better overall than the female results; and again, the simpler feature set performed better. For males, 8 dimensions had recall rates over 80% for the best-performing feature set. Some possible reasons for the differences in performance include 1) expressive range in the male voices was smaller, 2) the male clip sizes were smaller, and the resulting models were trained using more representative clips with less variance within each of them, and 3) the retained number of dimensions for males contained a larger percentage of the total original variance than the retained dimensions for females.

Table 7.18: Dimensional Classifier Performance for **female scripted speech**. This table shows the Average Unweighted Recall (AUR) in % for each dimension's binary classifier. ***SET1 Content***: RMS, RMS_u, ZCR, F0, F0_u, Jitter, Shimmer, LFSD, H1, H3-7, PR1-2, HR1, HR5, and MFCC1-12. ***SET2 content***: RMS, RMS_u, ZCR, F0, F0_u, Jitter, Shimmer, and MFCC1-12.

LSA #	SET1 AUR	SET2 AUR	LSA #	SET1 AUR	SET2 AUR	LSA #	SET1 AUR	SET2 AUR
2	73.1	74.6	8	57.3	57.3	14	62.1	72.8
3	65.2	74.7	9	67.7	66.7	15	69.4	71.9
4	56.0	76.0	10	57.6	77.4			
5	66.7	65.4	11	55.6	74.0			
6	68.0	64.1	12	56.7	73.2			
7	68.0	66.7	13	57.6	73.8			

Table 7.19: Dimensional Classifier Performance for **male scripted speech**. This table shows the Average Unweighted Recall (AUR) in % for each dimension's binary classifier. ***SET1 Content***: RMS, RMS_u, ZCR, F0, F0_u, #peaks, Autocorrelation, Jitter, Shimmer, LFSD, H1-H8, PR1, HR1, HR2, HR4, and MFCC1-12. ***SET2 content***: Set1 content plus HR7.

LSA #	**SET1 AUR**	**SET2 AUR**	**LSA #**	**SET1 AUR**	**SET2 AUR**	**LSA #**	**SET1 AUR**	**SET2 AUR**
2	89.3	86.1	8	77.1	78.6	14	83.3	75.0
3	86.4	83.7	9	72.9	80.6	15	74.3	70.1
4	78.5	75.7	10	70.8	74.5			
5	90.5	86.9	11	80.8	79.2			
6	68.8	68.7	12	90.8	90.8			
7	89.1	85.4	13	88.2	80.6			

7.4 Summary

This chapter presented a methodology for 1) grounding analysis of expressive voice in human perception, 2) discovering the expressive dimensions present in scripted and unscripted corpora organically, 3) analyzing the relationships among categories of perceived qualities in the voice (e.g., emotion, voice quality, conversation quality, prosody, and personal quality), 4) relating the organically-discovered dimensions to predetermined dimensions (e.g., affect, arousal, and dominance), and 5) modeling expressive dimensions found in the voice. The potential impact of this work is the production of analytics which better align with human perception and therefore better support application development. Two immediate target application domains include search, and human health and well-being. Future work could extend the current ability of search to find expressive speech sections in terms which humans perceive and describe. Also, non-invasive applications could be developed which analyze human expressivity in the diagnosis and monitoring of human physical and mental health. These applications could be made available on inexpensive, ubiquitous mobile platforms, and used in telemedicine. Furthermore, creativity platforms could leverage human expressivity in therapeutic treatment of depression, anxiety, and PTSD and could also be used in physical therapy in some cases.

More immediately, future work could extend the analytic technique to incorporate relationships among all the categories identified here, not just between emotion and voice quality. Prosody especially could be useful in emotion and voice quality recognition, and as features supporting the recognition of organically-recognized dimensions. Personal qualities such as honesty or creativity could be explored this way as well. How does an unusually creative person express themselves, for example, and could this be modeled? Then, some of the voice qualities examined here had multiple dimensions themselves (such as laughter and creaky voice), and should be explored further, broken into their own expressive dimensions, and used in the modeling of higher-level expressive dimensions. Next, exploring the scope of the correlations found by running correlations among words at the global scale, could measure the number of speakers and contexts in which the associations among words occur. Within-speaker correlations vs other-speaker and all-speaker correlations would be particularly interesting. Finally, we have shown that dimensional mapping between pre-defined fixed expressive dimensions and organically-discovered dimensions is possible. Further exploration of this relationship could enable leverage of prior work. Finally, the machine learning techniques themselves could be optimized. The larger-scale oral history corpus might benefit from deep learning techniques, for example. Furthermore, given the superior results of the lower-variance male scripted corpus models with the corresponding smaller clip sizes (lower variance within the clips themselves), exploring techniques for 1) partitioning a corpus itself into lower variance sections, and 2) varying clip sizes used in training and evaluation could improve results.

Because of 1) the repeated presence of laughter in the organically-discovered dimensions explored in this chapter, 2) the frequency of the perception of laughter (particularly in women's voices), and 3) the range of nuanced description of laughter discovered in Chapter 5, the next chapter explores laughter in depth. It extends the dimensional analysis techniques applied to general expression in this chapter to an in-depth exploration of laughter in the next chapter.

CHAPTER 8: ANALYSIS OF LAUGHTER IN UNSCRIPTED SPEECH

This chapter presents the discovery of the dimensions of laughter present in the female unscripted speech corpus, and answers portions of RQ8, RQ9, and RQ10. The methods used here were similar to those described in Chapter 7 for the general discovery of descriptive dimensions. A semi-automated Mechanical Turk study collected descriptors for each instance of laughter by presenting each laughter clip in its own Mechanical Turk task, and asking the listeners to provide 3 or more words describing what they heard, expressively speaking, in the laughter. Most of the laughter clips contained no speech. The exceptions were the relatively smaller number of instances which contained 1) simultaneous talking and laughter within a single speaker, and 2) simultaneous talking and laughter of two interacting people. In the second case, either the interviewer or the interviewee laughed, while the other talked. Chapter 5 describes the laughter study in detail. This chapter describes the resulting top ten dimensions of laughter present in the corpus, and the relationships among perceived emotion and VQ discovered via Latent Semantic Analysis with dimensional reduction.

8.1 Dimensional Analysis of Laughter: Explorations and Experiments

Organic Dimensional Discovery: In this experiment, we ran LSA over the female unscripted laughter dataset using the keywords collected in the perception studies described in Chapter 5. A descriptor-audio clip matrix was created and weighted, and decomposed via SVD as described in Section 7.1. This resulted in the discovery of 116 dimensions. Figure 8.1 shows an analysis of the amount of variance covered by accumulating dimensions.

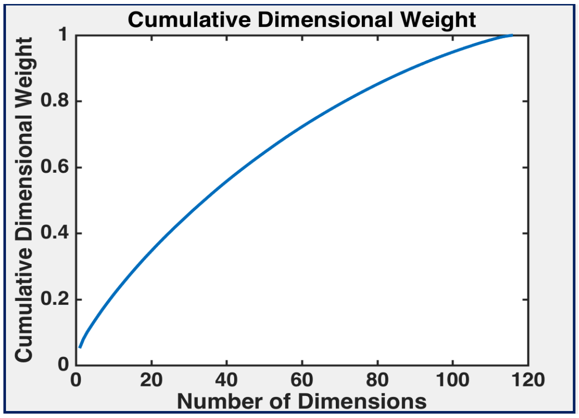


Figure 8.1: Cumulative Dimensional Weight. About half of the variance is in the first 38 dimensions. Retaining about half of the dimensions retains about 75% of the variance.

The analysis of general expressive dimensions in this corpus showed that on a data set this size, about 20-25 dimensions were easily differentiable by humans. Table 8.1 shows the normalized dimensional weights of the first 25 dimensions resulting from the LSA analysis.

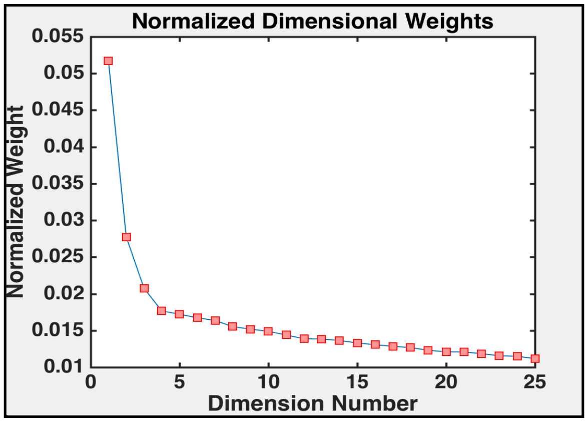


Figure 8.2: Normalized dimensional weights for the top 25 dimensions. Of the top 25 dimensions, the top 5 dimensions contain most of the information. At this point, the weight curve becomes linear between dimensions 5 and 25, and then asymptotic.

A heuristic similar to that used in Chapter 7 was used to determine a suggested number of dimensions to retain. As shown in Chapter 5, listeners gave laughter descriptors which aligned with emotion, prosody, voice quality, personal quality, and other (similar to the classes of descriptors for expressive voice, but in slightly different proportions across categories). The dimensionality was reduced successively, and a sanity check applied at each level of reduction. The Spearman correlation between emotion keyword vectors was taken, and again the threshold is set at 0.75 for detecting a "strong" correlation. As in Chapter 7, many of the emotion keyword correlations appear "sensible," but some will not appear to be sensible (such as a strong correlation between "happy" and "sad). Table 8.1 shows the number of statistically-significant correlations found at each level, and the percentage of correlations which were determined to be sensible.

Table 8.1: Dimensionality Reduction “Sanity” Check for Laughter in Unscripted Female Speech. This table shows the number of dimensions retained in the left column, the number of statistically-significant (p<0.05) strong correlations (rho>=0.7) found between emotion keywords in the center column, and the number of these correlations which were deemed “reasonable” in the third column. Above 15 dimensions, the number of strongly-correlated sensible correlations dropped to less than 50%. The best performance, in terms of number of strong, sensible correlations retrieved, is between 10-14 dimensions. This is consistent with the dimensionality reduction analysis for expressive speech.

# Dimesions	# Statistically-Significant Strong Correlations Found	# Sensible Correlations Found
1-15	195	135 (69%)
1-14	142	112 (79%)
1-13	113	81 (72%)
1-12	71	58 (82%)
1-11	62	51 (82%)
1-10	48	41 (85%)
1 - 9	15	13 (87%)

Dimensional Analysis: In this experiment, we again explored relationships among co-occurring classes of keywords and keyword descriptors, particularly between emotion and voice quality keywords, according to the methods given in Chapter 7.

8.2 Dimensional Analysis of Laughter: Results

Organic Dimensional Discovery: Table 8.2 describes the discovered dimensions of laughter in female unscripted speech. Note that because the information is distributed more evenly across more dimensions than in the expressive speech experiments, fewer keywords have strong correlations with each dimension. Lowering the threshold of strong correlation would add more descriptors to the list, but it would also introduce error into the resulting models if the model design did not compensate for this.

Table 8.2: Description of the top-12 LSA Concept Factors for Laughter in Female Unscripted Speech. A short description of the factor is given, followed by the strongest positively and negatively-associated keywords. The top 12 dimensions had multiple keyword concepts with strong weights.

#	Expressive Dimensions (ie, LSA Concept Factors) in Female Laughter
1 ***Neg:***	**High-variance, opposing qualities.** *Low, fast, slow, happy, scared, shy, forced, funny, and others.*

Table 8.2: (cont.)

#	**Expressive Dimensions (ie, LSA Concept Factors) in Female Laughter**
2	**Genuinely happy, sustained, voiced giggle.**
Pos:	Happy, amused, genuine, chuckle, funny, long, giggle.
Neg:	*Scared, quiet, short, soft, breathy, exhale, air, gasp.*
3	**A sad, short, low-pitched, voiced chuckle.**
Pos:	Short, low, chuckle.
Neg:	*Happy, long, gasp, inhale, exhale, giggle.*
4	**Fast, confident, feminine laughter with simultaneous talking.**
Pos:	Fast, talking, female.
Neg:	*Surprised, nervous.*
5	**Resonant and slow.**
Pos:	Sincere, slow, deep.
Neg:	*Surprised, nervous.*
6	**Soft, fast, and masculine.**
Pos:	Quiet.
Neg:	*Feminine, slow.*
7	**Gentle, soft, sustained, and nervous.**
Pos:	Nervous, soft, quiet.
Neg:	*Surprised, loud, short.*
8	**Surprise and alarm.**
Pos:	Surprised, alarmed.
Neg:	*Happy, sad.*
9	**Amused, nervous, soft, and unsure.**
Pos:	Amused, nervous, unsure, quiet.
Neg:	*Soft.*
10	**Sustained, voiced, fast, and nervous.**
Pos:	Nervous, long, fast.
Neg:	*Airy.*
11	**A loud, masculine, syllable.**
Pos:	huh
Neg:	*Quiet, female.*
12	**Sarcastic and confident.**
Pos:	Sarcastic.
Neg:	*Surprised.*

Dimensional Analysis via Joint Associate with Expressive Dimensions: Examining the associations within-dimension provides a detailed view of the relationships between emotion and voice quality.

Breathiness, ***airy*** quality, ***exhalation***, and ***gasping*** (all of these involve air) were negatively-associated with happiness and humor (dimension 2, similar to the relationships found in expressive speech). ***Airy*** quality was also negatively-associated with nervousness (dimension 10).

Chuckling and ***giggling***, both laughter-specific voice qualities, were positively-associated with happiness and humor (dimension 2). Interestingly, ***chuckling*** was also associated with sadness (dimension 3); but ***giggling*** retained a negative associated with sadness (dimension 3).

Talking while laughing and a ***feminine*** voice quality were associated with confidence (via negative association with nervousness and surprise in dimension 4). ***Feminine*** voice quality was negatively associated with short, syllabic utterances (dimension 11) and with surprise and nervousness (dimension 4).

Deep, resonant voice quality in laughter was positively associated with sincerity and negatively associated with surprise and nervousness (dimension 5).

Dimensional Analysis via Keyword Correlation: Emotion-VQ relationships were also evaluated by taking the Spearman correlation between keyword rows using the weighted, dimension-reduced keyword-audio clip matrix as described in section 7.2. Fewer strong correlations were found globally between voice quality and emotion. Successive reduction in dimensions from 15 dimensions to 8 dimensions uncovered the following ***strong negative correlations*** between ***VQ and emotion***:

- Negative correlation between breathiness and amusement.
- Negative correlation between musicality and surprise.
- Negative correlation between musicality and nervousness.
- Negative correlation between musicality and sadness.
- Negative correlation between sighing and nervousness.
- Negative correlation between giggling and amusement.
- Negative correlation between giggling and nervousness.

An interesting generalization here is that musicality is associated with positive affect and steady levels of arousal.

A similar successive reduction in dimensions from 15 to 5 yielded the following ***strong positive correlations*** between ***VQ and emotion***:

- Positive correlation between chuckling and amusement.

- Positive correlation between shakiness and feeling obligated.

8.3 Summary

This chapter presented the results of dimensional discovery for laughter. The descriptors were surprisingly similar to those given for expressive speech, so they were clustered and analyzed (emotion versus voice quality) according to the dimensional analysis methods which were used for expressive speech overall. The prior pattern of having a single high-variance, opposing quality dimension held for laughter as well. It is interesting to note that a larger number of correlations between prosody and emotion are found in laughter. This is sensible, since many kinds of laughter are pseudo-syllabic and rhythmic; and speed and duration are elements of prosody. Furthermore, the variance in pitch across laughter varies with emotion. Ongoing research is exploring feature sets for laughter, developing models for laughter, relating prosodic quality to emotion in laughter, and using the dimensions of laughter to improve dimensional modeling of expressive speech. As discussed in Chapter 7, relationships discovered via direct correlations could be explored further to reveal the number of speakers in which they occur (many vs. just one, for example), and the number of contexts in which these correlations occur. This additional step, along with scaling up the number of speakers, will guard against correlation biases introduced by single speakers.

The next chapter considers the results found in this exploration of laughter, and the results from the prior chapters, and presents a discussion of the results, implications, and next steps for research.

CHAPTER 9: DISCUSSION

The results uncovered in this dissertation enable and invite further investigation. Some thoughts for extension of the work and future directions are described below. The first set of extensions involves continued exploration of vocal expression qualities and models. The second set of extensions described below involves expansion of analytic technique. The final set of thoughts covers potential applications for the work.

9.1 Exploration of Vocal Expression

The exploration of vocal expression described in this thesis do have some limitations which invite further exploration. First, the range of voice qualities and nonverbal qualities could be greatly expanded. This thesis emphasized some of the most frequently heard qualities in the scripted and unscripted corpora, including whispering, breathiness, resonance, creakiness, and laughter. Many others could be explored. Second, relationships among categories other than emotion and VQ could be explored using similar dimensional analysis techniques. Next, each of the voice qualities could be explored and modeled at a deeper level. This thesis began to explore this in its examination of multiple spectral types of voice qualities which were all perceived as the same quality to listeners (for example, four different spectral patterns of creaky voice, multiple patterns of creaky voice, etc.).

With respect to the limitations of the resulting models themselves, the models were trained with relatively small data sets and a small number of speakers. The models should be expanded to incorporate larger data sets, the results of which could be used as background models in a wide range of projects. Labeling ground truth data was the bottleneck to the methods described in this thesis. Given the results in this thesis, however, active learning techniques could be integrated with crowdsourcing and model training to bootstrap models which incorporate a larger number of speakers across dialects to support an ever-increasing range of expressivity. With a larger scale of data available for analysis, the simple model architecture described in this thesis could be revised, and improved learning techniques, including deep learning, incorporated. Finally, a higher level of modeling could provide optimizations for streaming data, and include voice activation and error correction. Details of these limitations and potential future research are described below.

9.1.1 Exploration of Additional Voice Qualities

This dissertation explores the most frequently-perceived voice qualities present in the corpora. Many other less frequently-perceived VQs, however, such as growling, hoarseness, raspiness, spitting, youth, age, yelling, tenseness, terseness, melodic quality, and others could be explored using the same processes outlined in the prior chapters. Listeners heard these qualities, and they clearly have relationships to emotion, prosody, and personal quality. The curated corpora should be expanded to gain additional samples of these less-frequently observed qualities across more speakers; and the Library of Congress Veterans' History Project contains many more suitable samples which could be curated. Other styles of speech could also be examined and compared with the oral history samples.

9.1.2 Expanded Exploration of Relationships Within Corpora

The exploration of relationships between voice quality and emotion discussed in this dissertation could be expanded to include relationships among voice quality, emotion, prosody, personal quality, conversational quality, and other descriptor types. Extending these explorations, and examining them from multiple points of view, will provide further insight into both perception and modeling. For example, the research contained in this dissertation revealed a relationship between creaky voice and sarcasm (or lack of sincerity). Informal observation suggests that prosodic patterns might also align with sarcasm and creakiness, and prior literature has suggested relationships between prosody and sarcasm. From the point of view of emotion recognition, the presence of both creaky voice and relevant prosodic patterns could improve modeling and recognition of this dimension of sarcasm.

Relationships between perceived personal qualities and prosody, voice quality, and emotion, for example, could be explored. Personal qualities such as trustworthiness, intelligence, strength, persistence, etc. are, like many emotions, difficult to quantify or even define. Users, however, used personal quality descriptors when they could have selected other descriptors. What are they perceiving? By using the techniques presented in this dissertation, it would be possible to gain a better understanding of, for example, perceived intelligence and trustworthiness by learning the prosodic and VQ markers associated with this quality. Then, models could be constructed to detect or produce speech corresponding to a given personal quality. A potential application of this is an avatar in telemedicine; it would be desirable for humans to listen to and trust an avatar which gives information critical to a person's health.

9.1.3 Deepened Exploration of Voice Qualities Explored in this Dissertation

This research has demonstrated that a perceived quality such as creakiness may have multiple, sometimes opposing, spectral profiles. Creakiness, for example, can be either aperiodic or highly periodic. It can have a single or multiple F0 frequencies. Also, the H2-H1 relationship is not the same across all types of creaky voice and all speakers. In such cases, the models could be enriched, or deepened, such that each subtype is optimally modeled. The feature sets could be tuned for the subtypes, and the model architecture should reflect the perceived feature's multiple profiles. Dimensional analysis techniques applied to the features themselves, not just the descriptors, might be useful in improving the models here.

9.1.4 Stream Processing

Creating an overall model architecture which distinguishes across modal and non-modal phonation types is necessary for true stream processing. Much of the prior work distinguishes a single, target non-modal phonation type from modal phonation. The results from these single-phonation type studies may degrade when multiple non-modal phonation types are in the mix, particularly when a phonation type is not in a continuum relationship with the others. The results from the experiments in this dissertation were consistent when conducted across whispered, breathy, modal, and resonant effort levels; these are all in continuum relationship. When creaky voice was introduced into the mix (not in continuum relationship with the other qualities), performance degraded, compared to performance of the models when it was not in the mix. Creaky voice had multiple spectral profiles, some of which had overlapping qualities with other phonation types, particularly breathy voice. Future work should address and improve this finding.

Incorporating error correction strategies into the models is also important for robust stream processing. When this is not done, the output is "jittery." That is, regions of continuous single-type phonation will have sporadic errors where individual frames are incorrectly judged. The literature has many techniques for error correction which could be incorporated here for the improvement of recognition rates across modal and non-modal phonation in unscripted speech.

Distinguishing voice from silence or other noises in a recording is also important. A voice-activation detector from prior literature would be a useful addition for stream processing. Given that a stream section is voice, syllable and/or vowel detection would be useful, because many of the effort level detectors described in the scope of this dissertation use the vowel sounds for modeling and validation. Techniques and tools for vowel extraction are also described in the prior literature.

9.1.5 Improvements for Scale and Model Architecture

The scale of data used to train the models for this work was small because of the effort and cost necessary to annotate the recordings with ground truth. This could be remedied by developing semi-automated, human-in-the-loop corpora component screening and annotation techniques on crowdsourcing platforms. The resulting models could also be validated and tuned in a semi-automated and incremental way using crowdsourcing techniques, by comparing model output with human perception. This will require some lower-level process definition and software to automated it, as described in section 9.2 below. Models trained with many more speakers across a wider range of expression (for example, using more speakers from the Veterans History Project, and/or using speech from TED talks) could be used as background models for a variety of applications.

9.1.6 Exploration of Other Classes of Vocalization

This dissertation explored acted Shakespearian and oral history interviews. What about other categories of speech, such as TED talks, university lectures, phone support conversations, or sermons? Specifically, what about vocal music? Song has many similarities to spoken expression, but it is its own entity. Song has a text channel like speech, but also a very different overlay of musical language. A composition teacher changed the way I think about music by asking the question, "Yes, but what does it mean? Why is that a B and not a Bb? What would it mean then?" It is an elusive question, "What does it mean," related to both human perception and the unfolding of acoustic construction, at multiple scales, over time. Again, part of the answer lies in the question, "What do people hear," but this time, in musical expression. The investigative process will be similar, but music will impose special requirements on it.

9.2 Expansion of Technique

The exploration of vocal expression described in this thesis do have some limitations which could be addressed by expanding the analytic technique. First, the current findings only examine vocal expression, which is only a part of the picture of human expression. Next steps should investigate vocal expression in conjunction with other modalities, such as the text in the speech, physical gesture, and bio signals such as electro-dermal conductivity, temperature, or heart rate.

Next, the techniques presented here measure external behaviors as humans observe them, and cannot reveal that which humans do not describe. A speaker could be very good at hiding his or her emotions and fool the listeners. This limitation invites collaboration with neuroscientists who measure activity within the brain, and invites investigations which may discover correlations between brain activity and observed behavior. Also, because the technique depends on human description to measure perception, biases are introduced based on the cultural backgrounds dialect differences, and gender of both speaker and listener. The current work collected demographic data on the listeners, and this could be scaled up and used to reveal 1) differences in reported perception introduced by specific demographic, dialect, and gender combinations of the speaker and listener, and 2) differences in expressive nuance related to speaker dialect.

Another issue related to human differences across listeners depends on the listeners' differing vocabularies which they use to describe what they hear. LSA provides built-in methods to identify, manage, and use these differences; and synonym reduction techniques approach the problem from the other direction by reducing the number of descriptors to the most commonly articulated synonym. These are helpful techniques, but the vocabulary of the listener could be assessed and measured as well, along with the range and number of descriptors which a given listener is capable of providing for test cases.

Finally, dimensional analysis techniques could be expanded to 1) incorporate dimensional differences present in the acoustic data, 2) further explore mapping of organically-discovered perceived dimensions onto predefined axes (such as arousal, dominance, and valence), 3) explore hierarchical modeling of low-level discovered dimensions (e.g., laughter) within higher-level expressive dimensions, and 4) explore expressive dimensions across a range of phrase sizes. More detailed discussion follows.

9.2.1 Multimodal Investigations

This dissertation focused on the acoustic channel, however, examining just the acoustic channel alone is as limiting as only examining the text. Some questions require incorporating multiple channels. For example, sarcasm is more easily detected when the text is available, along with the vocal expression because a mismatch exists between what is being said via the text versus via vocal expression. Furthermore, trauma, particularly PTSD, may also be more easily detected by having the text available for analysis. Traumatized people tend to flatten out their voices and compress expressive range when relating a traumatic experience. This was a personal observation of mine in the Veterans' History Corpus, and an effect documented by other researchers as well. Examining text with expression can reveal what text or expression alone cannot. Physical expression functions hand-in-hand with the voice. The hands, face, eyes, body posture, and body orientation all work with voice in communication. Sarcasm could also be detected via a mismatch among text, vocal expression, and physical expression. Much prior work which analyzes physical expressive gesture could be leveraged in this task.

9.2.2 Active Learning for Scalability

This dissertation describes the use of crowdsourcing for learning what people hear. Studies conducted here could not be done easily at scale without this technique. The limiting factor in scaling the process out over larger data sets is still the determination of ground truth. We could use crowdsourcing to label ground truth, but this is still labor intensive for a person to build a system which can present audio clips, take in crowdsourced labels at selected points in the clips, validate the results, and save the coded data into an appropriate format. This approach scales poorly.

What if, instead, we built a system which could label unlabeled data, present sound clips and generated labels back to listeners, and request validation or correction from the listeners? The corrections and confirmations would be used to tune the models, and the iteratively-improved models would be used to label the next set of unlabeled data. This iterative model training process would continue until a performance target had been achieved. The techniques described in prior literature in crowdsourcing and active learning would help address the data labeling scalability problems.

9.2.3 Dimensional Mapping Bridge

Prior work has mapped emotions, and emotional reactions to words, onto predefined dimensions (most commonly affect, arousal, and dominance). Understanding the relationships among these predefined dimensions and the organically-discovered ones could enable leveraging prior work. A bi-directional mapping and model would extend techniques and findings in both directions. The initial experiments in Chapter 7 suggest that a mapping between dimensional worlds is possible, and that one organically-discovered dimension can still be distinguished from another when mapped onto the predefined axes of affect, arousal, and dominance.

9.2.4 Hierarchical Dimensional Mapping

Mapping among multiple sets of organically-discovered dimensions is also possible, and discovering these mappings may improve the vocal expression models. For example, laughter was a strongly-perceived nonverbal quality in this dissertation. Listeners described multiple dimensions of laughter, just as they described higher-level expressive dimensions. Many of these dimensions included laughter. Could the higher-level models of general expression leverage the lower-level models of laugher for improved results? As an example, sarcasm is often accompanied by laughter, but the laughter has a different quality from that of sincere humor and happiness. The kind (or dimenson) of laughter present in a dimension is important, not just the presence of laughter itself.

9.2.5 Optimizing the Analysis Window and Dimensional Scope

The results in this dissertation have shown that the quality of the analytic results (precision, recall, accuracy) depend on the size of the labeled analysis window used in modeling and the number of retained dimensions. The male scripted speech had the best results for dimensional analysis. This speech had the shortest labeled analysis windows. Shortening the labeled window size could potentially improve results in the other corpora (scripted female and unscripted speech for both genders), but what is a reasonable and optimal size? Furthermore, shorter windows mean more ground truth labeling, an important factor for scalability.

The experiments across all corpora in this dissertation used approximately the same number of retained dimensions for all corpora; however, the percentage of information retained in that number of dimensions varied. The retained dimensions in the male scripted speech contained more of the available information (about half of the available dimensions), versus about 20% for the unscripted speech.

9.3 Summary and Potential Applications

Many applications could be enabled by this work, including sonic search, health and wellness applications, voice training applications, and resource curation tools, to name a few. Sonic search would enable finding resources which contain the desired expressive qualities, and finding locations within the relevant resources where the desired qualities are exhibited. These tools ultimately will become multimodal, and capable of handling search requests for text content, vocal expressive quality, physical gesture, or other expressive quality. Resource curation tools are closely linked to search, because resources are meant to be found and used.

Health and wellness applications would include ubiquitous tools for detecting mental and physical health status, which leverage patterns in the voice in some way. Depression, anxiety, and PTSD may share similar vocal markers, and may be useful in the non-intrusive detection of these conditions, especially in populations who do not visit doctors. As an example, what if the military had a tool to record and analyze interviews of service men and women at discharge, and screen for PTSD? People with potential problems could be referred for diagnosis and treatment, instead of being released to the street, with a high probability of becoming homeless. For a person with PTSD especially, homelessness adds more stress, makes their existing problem worse, and reduces the chance of recovery.

What if the medical community had similar tools for ongoing monitoring of mental and physical health status? These applications could record and analyze expressive speech, gesture, or other interaction quality, and analyze a patient's periodic health status, instantaneously, and over time. Sudden changes in a person's usual patterns would be visible, as would similarities with known pattens common in disease. They could also be made available on mobile devices, for those unable to visit a clinic.

And what if interactive applications for therapy using expressive creativity could be developed and deployed online? Art therapy is in common use today; therefore, why not explore ways to make it more accessible, where human interactions could be analyzed, along with the resulting electronic creative works?

Platforms for creativity could be built and used to analyze human expressivity, both in the voice and in other forms of human expression. By enabling multimodality, such platforms could extend the current capability for multimodal expression, and new expressive modalities studied.

Finally, teaching applications could help people learn to improve the way they express themselves. Professional speakers practice for many years to become experts at their craft, and so do singers. What if we could develop interactive applications which teach people to become effective speakers, and teach specific expressive techniques for desired impact on the listener? These training platforms could help actors, teachers, ministers, musicians, newscasters, and other public performers. The training content and feedback mechanisms could be tuned by experts in the expressive modality in question, by audience reaction to live performance, or by a crowdsourced response. Teaching applications could potentially train service robots and avatars as well, so machines become target students, different from, but not unlike, human students.

Many other applications are possible. These are just a few. Given the results of the research in prior chapters, and the discussion points given here, the next and final chapter summarizes and concludes this thesis.

CHAPTER 10: CONCLUSIONS

This chapter summarizes my work. It reviews the research questions and investigations, and summarizes the contributions.

10.1 Summary of Investigation

This dissertation explores human vocal expressivity, and at the highest level, asks how people hear vocal expression and how vocal expression should be modeled, based on how people hear it (RQ1, RQ2). Then, it addressed scripted and unscripted speech separately, and asked, for each kind of speech, what people heard in the expressivity of males and females. Given the answers to the perception question, it asked what features supported the modeling of several voice qualities which were frequently-perceived in both scripted and unscripted speech, for both males and females. These qualities were whispering, breathiness, modal voice, resonance, and (for unscripted speech) creakiness. The in-depth investigation produced a proposed feature set, with simple, cross-validated models. It also 1) revealed the continuum relationship from whispered, to breathy, to modal, to resonant speech, 2) showed creakiness not to lie on this continuum, and 3) revealed the multiplicity of spectral profiles which mapped to single, perceived voice qualities.

Not all perceived features, however, were described via small sets of clearly-articulated, frequently-repeating keywords. People reported hearing emotion more often than prosody, voice quality, and conversation quality, and described emotion in nuanced terms, many of which were related, but which did not frequently repeat. “Joy,” “delight,” “happiness,” and “jolliness” are all happy emotions, for example, but being just “happy” is not the same as being “jolly” or “joyful.” Humans heard the difference, and the nuances are important. A different analysis technique was needed which aligned with and leveraged the subtlety of human perception, in the terms which humans reported hearing it. Latent Semantic Analysis applied to the keyword descriptors preserved human expressive subtlety while enabling the analysis and generalization of relationships among the individual descriptors (such as joy and happiness) and among entire categories of descriptors (such as voice quality, prosody, and emotion). LSA also enabled expressive dimensional discovery for both male and female scripted and unscripted speech; and the top discovered expressive dimensions were modeled and validated.

My research also explored laughter, and discovered that listeners perceived it in ways similar to general high-level expressivity in speech. They still heard and articulated prosody, voice quality, personal quality, conversational quality, and emotion. Dimensional analysis of the descriptors again via LSA revealed the diversity of laughter, and the multiple functions of it in vocal expression. Not all laughter is happy, and not all laughter is sustained and syllabic. Some laughter occurred simultaneously with speech. Given the presence of laughter in the organically-discovered expressive dimensions of speech, it is probable that the dimensions of laughter could be used in the detection of higher-level, expressive dimensions in the voice. Extensions to this work can explore relationships between expressive dimensions at higher and lower levels in a hierarchy, or between non-hierarchical sets of dimensions.

My research has addressed the questions of what people hear in vocal expression for males and females, and in scripted an unscripted speech (RQ3, RQ5, RQ8). It has explored feature sets and models for frequently-perceived voice quality features, and for entire dimensions of expressivity, again for males and females, in scripted and unscripted speech (RQ4, RQ6, RQ7, RQ9, RQ10, and RQ11). Finally, it addressed the relationships between voice quality and emotion (RQ11), and presented a process for investigating relationships among individual qualities and entire categories of qualities, particularly voice quality, emotion, prosody, conversational quality, and personal quality.

10.2 Summary of Contributions & Final Statement

My work has produced the following contributions, which were discussed in detail in Chapter 1:

1) **An end-to-end cross-disciplinary process for grounding human expression analytics in human perception *(RQ1-11).*** This process bridges disciplines and produces software artifacts which are better suited for application development, because the resulting services are rooted in human perception and human needs.
2) **The confirmation of the continuum relationship across whispered, breathy, modal, and resonant voice, both from human perception and acoustic analysis *(RQ1-6, RQ8-9).*** This relationship had not been recognized across the entire continuum before, probably because prior work had been driven by a range of specific questions across multiple disciplines such as speech pathology, security, speech processing, and vocal performance in theater.
3) **Baseline models for recognizing effort levels within male and female scripted and unscripted speech *(RQ1-6, RQ8-9).*** These models used features that went beyond the

standard feature set used in speech analysis. They resulted from a detailed analysis in spectral behavior across conditions, which resulted in the definition of spectral bands of interest for both males and females, and features designed to distinguish across all conditions in the continuum. Much of the prior work distinguished modal speech from a single non-modal condition; the approach presented in this thesis considered the entire continuum. Final feature selection depended on unequal variance sensitivity measurements (d_a from signal detection theory), analysis of means and 2-sigma variances across conditions, and interactive empirical evaluation of best feature set collections.

4) **Comparison between effort levels in male, female, scripted, and unscripted speech *(RQ6-7).*** Gender comparison across a continuum of effort levels, from whispering through resonance, are new.

5) **A perception-grounded technique for discovering dimensions of expressive speech present in corpora *(RQ10).*** This technique can be applied to the entire body of vocal expression, or just one quality, such as laughter. The technique was validated by modeling each of the dimensions detected, and validating the resulting models. This blend of crowdsourcing and latent semantic analysis applied to the human descriptors is applicable to any set of qualities which humans can describe and perceive, not just expressive speech; therefore, its potential impact crosses many disciplines. ***For this research, the technique enabled an entirely new way of exploring the perception and recognition of human emotion, which potentially impacts, at a minimum, the fields of HCI, speech and language processing, computational linguistics, and psychology. It also enabled the organic discovery of dimensions present in a corpus, and did not rely on reduction to pre-defined dimensions such as arousal and affect***. Discovering mappings between organically-discovered dimensions and other predefined dimensions, however, is possible for the leverage of and contribution to prior analytic techniques.

6) **A technique for discovering relationships among perceived emotion, voice quality, prosody, personal quality, conversational quality, and other elements of expression in the voice *(RQ10).*** The techniques presented here enable systematic exploration among multiple categories of human perception and exploration, not only

to expressive speech, but to any domain which humans can perceive and describe. It is a generalized technique which applies to many disciplines.

7) **Organic discovery of the expressive dimension of laughter present in unscripted speech *(RQ1-2, RQ8-10).*** Dimensional exploration of laughter is new, and this exploration shows that dimensional discovery applies to lower-level expressive features as well as high-level expressive dimensions.
8) **Curated corpora for exploring male and female scripted and semi-structured, unscripted speech *(RQ1-11).*** The curation methods support the end-to-end investigative process, and have potential impact on digital curation and archival, as well as speech processing and HCI.

These discoveries give us tools for understanding how human expression works, not just in the domain of speech, but in other modalities, too. The process presented here helps us ground analytic models in human perception, so that the results can be better aligned with how humans think and with what they do. This alignment better supports application development, because it is in sync with natural human abilities. Then the resulting models leverage what humans already do instead of forcing humans to adapt to something which is not natural for them.

These results encourage development in several key application areas, especially health and human wellness, search and archival, and explorations in human creativity. What if analysis of human expressivity could help diagnose illness, and provide therapy for people who are suffering? What if we could archive, discover, and browse sonic artifacts at least as easily as we can browse text artifacts? And what if we could understand and augment possibilities for human creativity, across multiple modalities? Exploring human expression opens the doors for these possibilities.

APPENDIX A: VOICE AND LAUGHTER SAMPLE NAMING CONVENTIONS

This appendix describes the naming convention which encodes, for voice samples, the interview and speaker identification and the specific section within the interview text.

For laughter samples, the naming convention encodes the interview and speaker identification (speaker with the floor), the specific section within the interview text in which the laughter occurred, the laughter type, the laughter instance number within the named section, and optional laughter interaction indicator.

All **voice sample clips** are named as follows:

InterviewID_SpeakerName_QAExchange#_Section#_Subsection#_Sub-subsection#

> The **InterviewID** uniquely identifies the Library of Congress Oral History interview. It is a numeric ID.
>
> The **SpeakerName** is the name of the speaker, the person who currently has the floor. It is a human-readable text ID name.
>
> The **QAExchange#** is the nth question-answer exchange in the interview. It is a numeric ID.
>
> The **Section#** is the nth section within a question-answer exchange. It is a numeric ID. It is usually a speaker floor change. Usually, when this value is a 1, the speaker is the interviewer. Usually when this value is a 2 or greater, the speaker is the interviewee (or one of the interviewees).
>
> The **Subsection#** is the first hierarchical layer under the Section#. It is a numeric ID. Often, an interviewee has multiple sections, particularly when telling a story, and not just answering direct yes-no or single-phrase questions.
>
> The **Sub-subsection#** is the next hierarchical layer beneath the Subsection. This layering continues to an arbitrarily large number of hierarchical levels, separated by underscores. All are numeric IDs.

Voice Sample Clip Example: **2_Ancona_15_4_1.wav**

> The **InterviewID** is 2, which represents the interview of Joseph Ancona.
>
> The name of the speaker is "**Ancona**" (Joseph Ancona).

It represents the **15th** question-answer exchange.

It is the **4th** section within the 15th question-answer exchange.

It is the **1st** sub-section beneath the 4th section of the 15th question-answer exchange.

An arbitrary number of subsection levels could have been named here.

All **laughter sample clips** are named as follows:

InterviewID_SpeakerName_LaugherName_QAExchange#_SectionHierarchy_LaughterInstance_LaughterInteraction

The **InterviewID** uniquely identifies the Library of Congress Oral History interview. It is a numeric ID.

The **SpeakerName** is the name of the speaker, the person who currently has the floor. It is a human-readable text ID name.

The **LaugherName** is the name of the person who is laughing. It is a human-readable text ID name.

The **QAExchange#** is the nth question-answer exchange in the interview. It is a numeric ID.

The **SectionHierarchy** is the identification of the clip's position within the hierarchical layers of sections, sub-sections, sub-sub-sections, etc. beneath the QAExchange#**.** These are an arbitrary number of numeric values sepearated by underscores, and they are in the same format as the voice sample clips.

The **LaughterInstance** identifies the basic laughter type, and the nth instance of that laughter type within the section. The LaughterInstance has the format **L#** or **LS#,** where the '**L**' indicates laughter alone, and the '**LS**' indicates simultaneous laughter and speech.

The **LaughterInteraction** is an optional tag, and identifies the nature of interactive laughter. A "*****" indicates mutual, or joint laughter, with 1 or more other speakers in the interview. A "**&**" indicates laughter over another speaker.

Laughter Sample Clip Example: **32_Lim_Lim_12_2_16_L1_*.wav**

The **InterviewID** is 32, which represents the interview of Ingrid Lim.

The name of the person laughing is **Ingrid Lim**.

It represents the **12th** question-answer exchange.

It represents the 2nd section beneath the 12th question-answer exchange, and the 16th sub-section beneath the 2nd section (**entire hierarchical section identification of 12_2_16**).

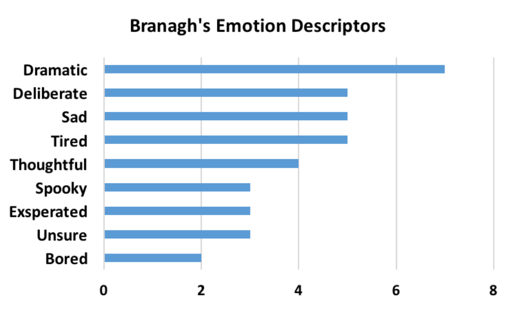
It is the **first instance of laughter** in this section.

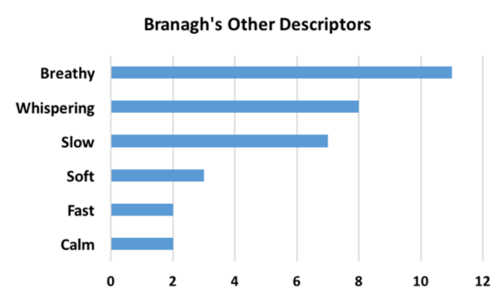
It is **mutual laughter** with one or more other speakers in the interview (in this case, the interviewee).

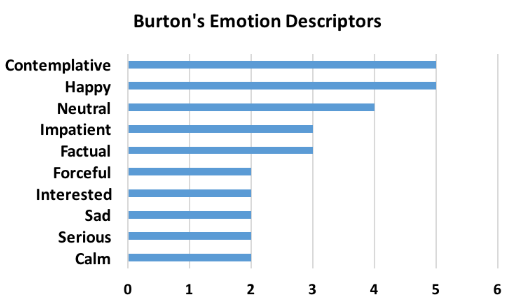
APPENDIX B: KEYWORD DESCRIPTORS

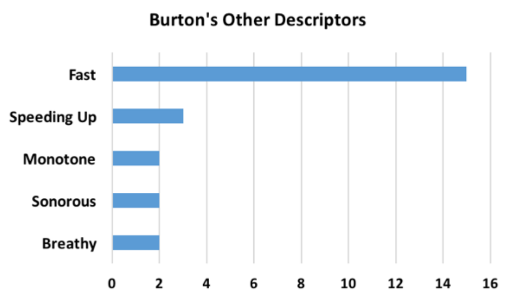
Top Descriptors for Male Acted Voices

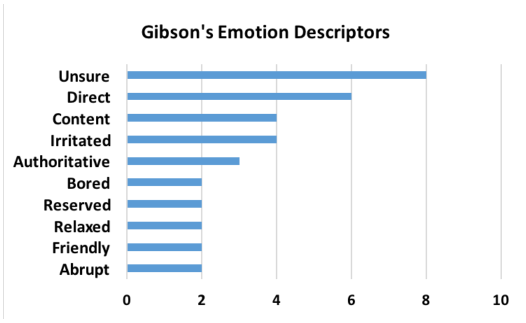
These graphs show the most frequently-given emotion and non-emotion descriptors given for each male speaker performing the Hamlet soliloquy. Note that not all keywords given are listed here, just up to 12 of the most frequently-occurring descriptors in each category.

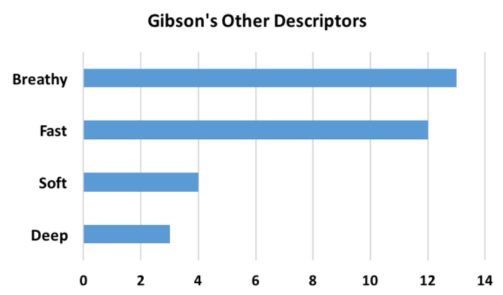


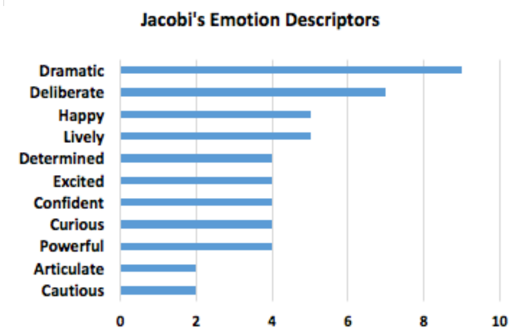


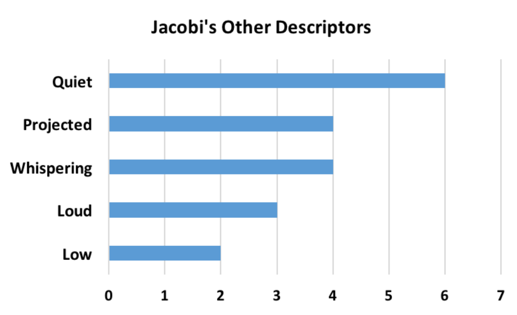


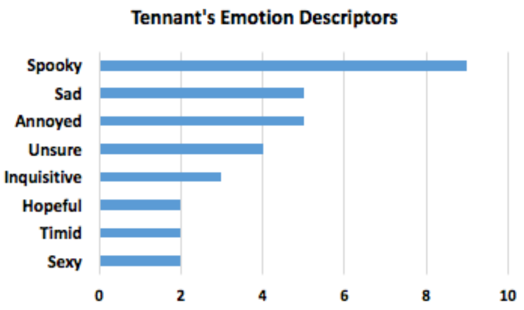


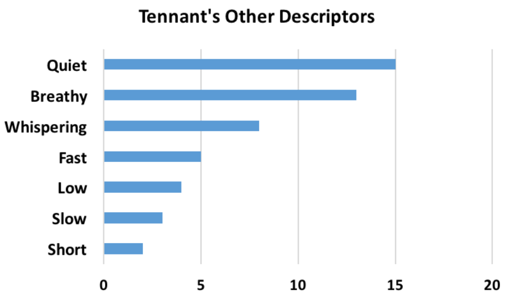






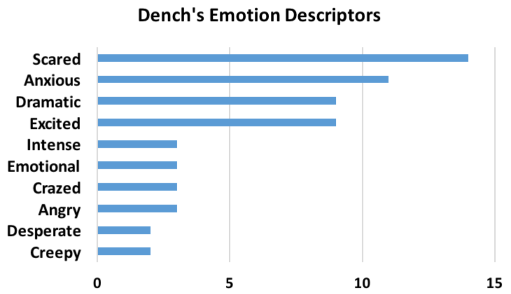


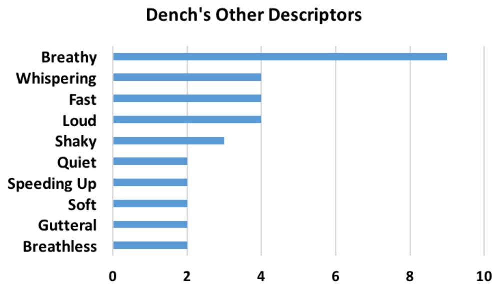


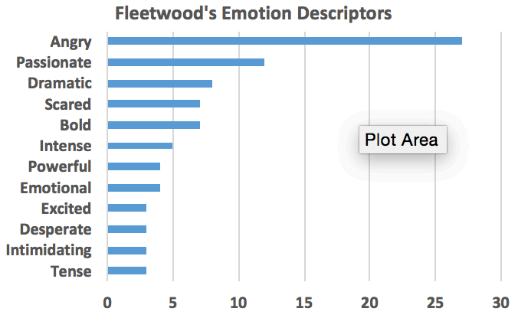


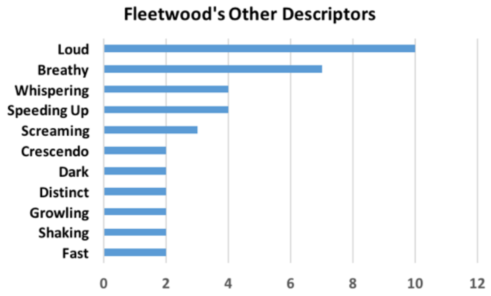
Keyword Descriptors: Top Descriptors for Female Acted Voices

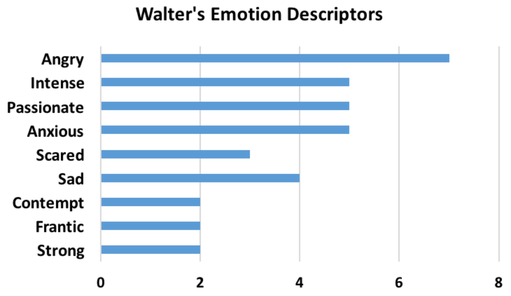
These graphs show the most frequently-given emotion and non-emotion descriptors given for each female speaker performing Lady Macbeth's soliloquy. Note that not all keywords given are listed here, just up to 12 of the most frequently-occurring descritprs in each category.

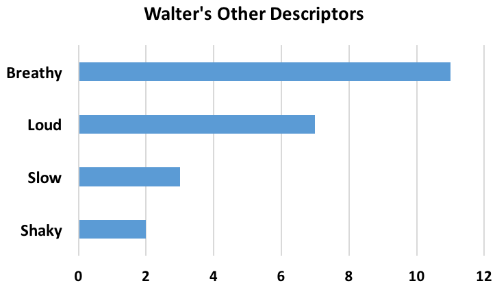






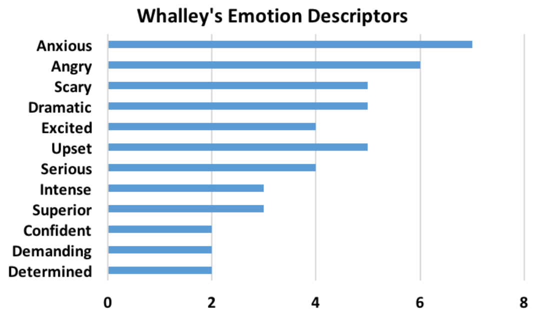


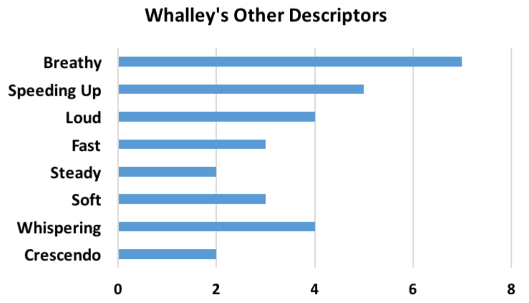


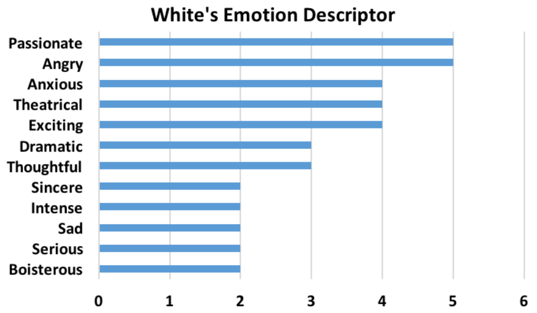


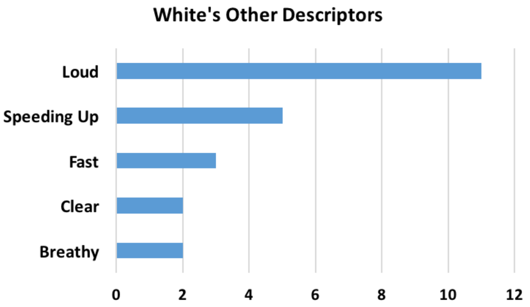
APPENDIX B: (cont.)

Keyword Descriptors: Top Descriptors for Female Acted Voices (cont.)


Whalley's Emotion Descriptors

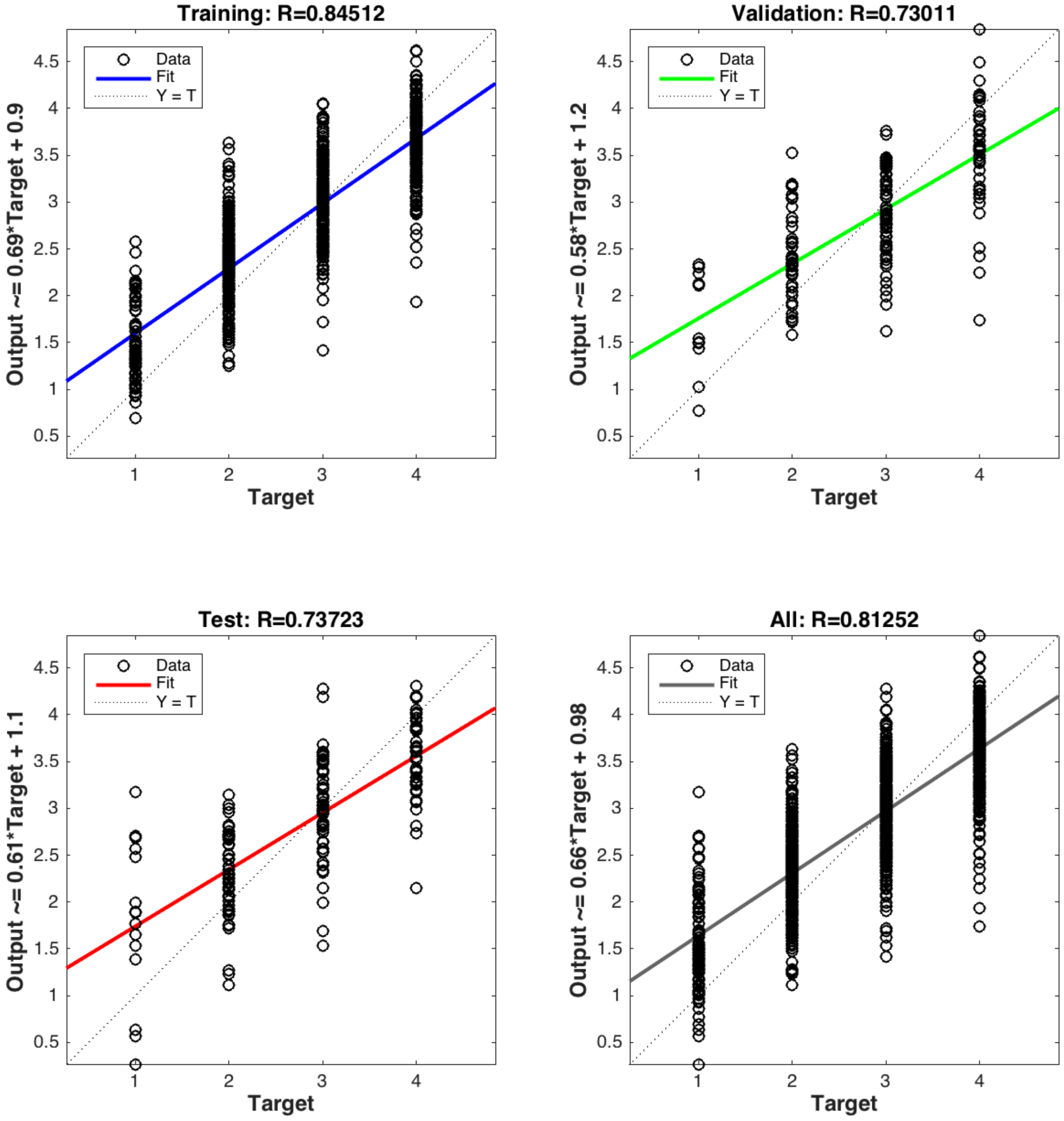

Whalley's Other Descriptors


White's Emotion Descriptor


White's Other Descriptors

APPENDIX C: REGRESSION ACROSS CONDITIONS IN MALE SCRIPTED SPEECH

These graphs show the results of training a neural network model with the male, scripted whispered (Target=1), breathy(Target = 2), modal (Target = 3), and resonant (Target = 4) speech data, and plotting a regression line through the predictions. This monotonically-increasing linear relationship across conditions shown here aligns with the linear relationships observed across the means and variances observed for individual features in Figure 6.2. This result reinforces the continuum relationship revealed across whispering, breathiness, modal speech, and resonance.



APPENDIX D: JOINT ASSOCIATIONS BETWEEN EMOTION AND VQ KEYWORDS

Table Appendix D.1: This data shows the emotion and voice quality keyword descriptors which were jointly strongly associated with the same dimensions (females on the left, males on the right). The keyword associations were determined by projection of the descriptors across the dimensions discovered via LSA, as described in section 7.2. An association was considered a strong positive association if the keyword-dimension projection matrix weight >= 0.85, and considered to be a strong negative association if the projection matrix weight <= -0.85. The top 13 dimensions were considered here. When an emotion was correlated with a voice quality across multiple dimensions, the emotion is highlighted in blue.

Females

VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Acting			Content	8
			Deliberate	8
			Melancholy	8
			Interested	8
			Mad	8
			Emotional	8
			Cold	8
			Curious	8
			Calm	8
			Sad	8
			Passionate	8
Bold	Happy	10	Scared	2
	Determined	10	Afraid	2
	Confident	10	Anxious	2
	Urgent	10	Worried	2
	Demanding	10	Tense	2
	Excited	10	Haunted	2
	Suspenseful	10	Prayer	2
	Aggression	10	Alert	2
	Anger	10	Desperate	2
	Panic	10		
	Anxious	10		
	Excited	13		
	Frantic	13		
	Aggression	13		
	Boisterous	13		
	Zealous	13		
	Anger	13		

Males

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Breathy	Happy	4	Excited	2
	Calm	4	Happy	2
	Humorous	4	Lively	2
	Amused	4	Enthusiastic	2
	Confident	4	Thrilled	2
	Upbeat	4	Passionate	2
	Proud	4	Amused	2
	Cheerful	4	Energetic	2
			Interested	2
			Engaged	2
			Confident	2
			Confident	3
			Calm	3
			Content	3
			Sincere	4
			Nervous	4
			Excited	4
			Sad	4
			Hesitant	6
			Friendly	6
			Normal	6
			Bored	6
			Upbeat	6
			Indifferent	6
			Proud	6
			Matter-of-fact	6
			Concerned	11
			Confused	11
			Friendly	11
			Tired	11
			Worried	11
			Mellow	11
			Uncertain	11

Table Appendix D.1: (cont.)

Females

VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Breathless	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Powerful	2
	Worried	2	Excited	2
	Haunted	2	Confident	2
	Prayer	2	Emotional	2
	Alert	2	Dramatic	2
	Desperate	2	Aggressive	2
			Empathic	2
	Intense	2		
			Anger	6
	Serious	6	Mad	6
	Nervous	6	Urgent	6
	Alert	6	Vengeful	6
	Excited	6	Scared	6
	Boisterous	6	Cold	6
	Sad	6	Deliberate	6
	Emotional	6	Angered	6
	Concerned	6	Sharp	6
	Determined	6		
	Passionate	6	Passionate	7
	Haunted	6	Stern	7
	Prayer	6	Frightened	7
	Anxious	6	Tense	7
	Zealous	6	Contemptuous	7
	Crazy	6	Confident	7
			Scared	7
	Excitement	7	Exciting	7
	Fear	7	Assertive	7
	Emotional	7	Interested	7
	Urgent	7	Calm	7
	Excited	7	Emphatic	7
	Furious	7	Thoughtful	7
	Desperate	7	Deliberate	7
	Chilled	7		
	Defiance	7	Passionate	3
	Fearless	7	Tense	3
	Indignant	7	Stern	3
	Spiteful	7	Deliberate	3
	Upset	7	Angry	3
	Nervous	7	Sharp	3
	Concerned	7	Anxious	3
	Aggressive	7	Fear	3
	Angry	7	Exciting	3
			Pissed	3
	Urgent	12		
	Upset	12		
	Dramatic	12		
	Afraid	12		
	Angered	12		
	Pleading	12		
	Vengeful	12		
	Nervous	12		

Males

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Clear	Confident	3	Confused	3
	Calm	3	Nervous	3
	Content	3	Unsure	3
			Frustrated	3
	Happy	5	Upset	3
	Content	5	Hesitant	3
	Thoughtful	5	Excited	3
	Uncertain	5	Amused	3
			Sad	3
	Hesitant	7	Anxious	3
	Content	7		
	Thoughtful	7	Happy	4
	Confused	7	Calm	4
	Calm	7	Humorous	4
			Amused	4
			Confident	4
			Upbeat	4
			Proud	4
			Cheerful	4
			Bored	5
			Serious	5
			Calm	5
			Nervous	5
			Tired	5
			Concerned	5
			Anxious	5
			Sad	7
			Depressed	7
			Friendly	7
			Scared	7
			Content	8
			Unsure	8
			Relaxed	10
			Nervous	10
			Upbeat	10
			Content	10
			Upbeat	13
			Scared	13

Table Appendix D.1: (cont.)

Females

VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Deathly	Excitement	7	Passionate	7
	Fear	7	Stern	7
	Emotional	7	Frightened	7
	Urgent	7	Contemptuous	7
	Excited	7	Confident	7
	Furious	7	Scared	7
	Desperate	7	Exciting	7
	Chilled	7	Assertive	7
	Defiance	7	Interested	7
	Fearless	7	Calm	7
	Indignant	7	Emphatic	7
	Spiteful	7	Thoughtful	7
	Upset	7	Deliberate	7
	Nervous	7		
	Concerned	7		
	Aggressive	7		
	Angry	7	Excited	3
			Anxious	3
			Dramatic	3
			Scared	3
			Nervous	3
			Tense	3
			Desperate	3
			Excitement	3
			Fearful	3
			Pleading	3
			Suspenseful	3
			Demanding	3

Males

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Creaky	Happy	5	Excited	2
	Content	5	Happy	2
	Thoughtful	5	Lively	2
	Uncertain	5	Enthusiastic	2
			Thrilled	2
	Hesitant	6	Passionate	2
	Friendly	6	Amused	2
	Normal	6	Energetic	2
	Bored	6	Interested	2
	Upbeat	6	Engaged	2
	Indifferent	6	Confident	2
	Proud	6		
	Matter-of-fact	6	Bored	5
			Serious	5
			Calm	5
			Nervous	5
			Tired	5
			Concerned	5
			Anxious	5
			Sad	6
			Confused	6
			Frustrated	6
			Concerned	6
			Serious	6
			Content	6
			Unemotional	6
			Hesitant	7
			Content	7
			Thoughtful	7
			Confused	7
			Calm	7
			Nervous	9
			Matter-of-fact	9
			Unsure	9
			Scared	9

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Deep	Happy	2	Passionate	2
	Calm	2	Angry	2
	Humorous	2	Forceful	2
	Amused	2	Powerful	2
	Confident	2	Excited	2
	Upbeat	2	Confident	2
	Proud	2	Dramatic	2
	Cheerful	2	Emotional	2
			Aggressive	2
	Serious	6		
	Nervous	6	Scared	4
	Alert	6	Anxious	4
	Excited	6	Worried	4
	Boisterous	6	Sad	4
	Sad	6	Emotional	4
	Emotional	6	Desperate	4
	Concerned	6	Determined	4
	Determined	6	Afraid	4
	Passionate	6	Frightened	4
	Haunted	6	Excited	4
	Prayer	6	Passionate	4
	Anxious	6	Urgent	4
	Zealous	6	Fearful	4
	Crazy	6	Dramatic	4
			Anger	6
			Mad	6
			Forceful	6
			Urgent	6
			Spitting	6
			Vengeful	6
			Scared	6
			Cold	6
			Deliberate	6
			Angered	6
			Sharp	6
			Excited	9
			Upset	9
			Desperation	9
			Vengeful	9
			Serious	9
			Forceful	9
			Anger	9
			Cold	9
			Content	9
			Happy	9
			Calm	9
			Melancholy	9
			Emotional	9

Males				
VQ Keyword	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Deep	Happy	4	Sincere	4
	Calm	4	Nervous	4
	Humorous	4	Excited	4
	Amused	4	Sad	4
	Confident	4		
	Upbeat	4	Hesitant	7
	Proud	4	Content	7
	Cheerful	4	Thoughtful	7
			Confused	7
			Calm	7
Gravelly			Calm	12
			Relaxed	12
			Hesitant	12
Growling	Hesitant	6	Sad	6
	Friendly	6	Confused	6
	Normal	6	Frustrated	6
	Bored	6	Concerned	6
	Upbeat	6	Serious	6
	Indifferent	6	Content	6
	Proud	6	Unemotional	6
	Matter-of-fact	6	Hesitant	7
			Content	7
			Thoughtful	7
			Confused	7
			Calm	7
Monotone	Happy	4	Excited	2
	Calm	4	Happy	2
	Humorous	4	Lively	2
	Amused	4	Enthusiastic	2
	Confident	4	Thrilled	2
	Upbeat	4	Passionate	2
	Proud	4	Amused	2
	Cheerful	4	Energetic	2
			Interested	2
	Nervous	9	Engaged	2
	Matter-of-fact	9	Confident	2
	Unsure	9	Sincere	4
	Scared	9	Nervous	4
			Excited	4
	Upbeat	13	Sad	4
	Scared	13		
	Loud	13	Enthusiastic	9
			Sad	9
			Bored	9
			Frustrated	9
			Relaxed	10
			Nervous	10
			Upbeat	10
			Content	10

Table Appendix D.1: (cont.)

Females

VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Deep			Angry	10
			Dramatic	10
			Powerful	10
			Afraid	10
			Pissed	10
			Exciting	10
			Determined	10
			Biting	10
			Stressed	10
			Stern	10
			Contemptuous	10
Eerie	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Powerful	2
	Worried	2	Excited	2
	Tense	2	Confident	2
	Haunted	2	Emotional	2
	Prayer	2	Dramatic	2
	Alert	2	Aggressive	2
	Desperate	2	Empathic	2
			Expressive	2
	Serious	6		
	Nervous	6	Scared	4
	Alert	6	Anxious	4
	Excited	6	Worried	4
	Boisterous	6	Sad	4
	Sad	6	Emotional	4
	Emotional	6	Desperate	4
	Concerned	6	Determined	4
	Determined	6	Afraid	4
	Passionate	6	Frightened	4
	Haunted	6	Excited	4
	Prayer	6	Passionate	4
	Anxious	6	Urgent	4
	Zealous	6	Fearful	4
	Crazy	6		
			Anger	6
			Mad	6
			Forceful	6
			Urgent	6
			Spitting	6
			Vengeful	6
			Scared	6
			Cold	6
			Deliberate	6
			Angered	6
			Sharp	6

Males

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Mumbling	Content	8	Bored	8
	Unsure	8	Amused	8
			Funny	8
			Humorous	8
Normal	Hesitant	6	Sad	6
	Friendly	6	Confused	6
	Normal	6	Frustrated	6
	Bored	6	Concerned	6
	Upbeat	6	Serious	6
	Indifferent	6	Content	6
	Proud	6	Unemotional	6
	Matter-of-fact	6	Relaxed	10
			Nervous	10
			Upbeat	10
			Content	10
Plain	Confident	3	Excited	2
	Calm	3	Happy	2
	Content	3	Lively	2
			Enthusiastic	2
	Happy	5	Thrilled	2
	Content	5	Passionate	2
	Thoughtful	5	Amused	2
	Uncertain	5	Energetic	2
			Interested	2
			Engaged	2
			Confident	2
			Confused	3
			Nervous	3
			Unsure	3
			Frustrated	3
			Upset	3
			Hesitant	3
			Excited	3
			Amused	3
			Sad	3
			Anxious	3
			Happy	4
			Calm	4
			Humorous	4
			Amused	4
			Confident	4
			Upbeat	4
			Proud	4
			Cheerful	4

Table Appendix D.1: (cont.)

Females

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Forceful	Passionate	5	Scared	2
	Emotional	5	Afraid	2
	Desperate	5	Anxious	2
	Aggressive	5	Worried	2
	Pleading	5	Tense	2
	Demanding	5	Haunted	2
	Serious	5	Prayer	2
	Eager	5	Alert	2
	Horny	5	Desperate	2
	Passion	5		
	Afraid	5	Excited	5
	Scared	5	Upset	5
			Worried	5
	Happy	10	Frightened	5
	Determined	10	Anxious	5
	Confident	10	Dramatic	5
	Urgent	10	Cheerful	5
	Tense	10		
	Demanding	10	Serious	6
	Excited	10	Nervous	6
	Suspenseful	10	Alert	6
	Aggression	10	Excited	6
	Anger	10	Boisterous	6
	Panic	10	Sad	6
	Anxious	10	Emotional	6
			Concerned	6
	Dramatic	11	Determined	6
	Sad	11	Passionate	6
	Emotional	11	Haunted	6
	Afraid	11	Prayer	6
	Fear	11	Anxious	6
	Powerful	11	Zealous	6
	Melancholy	11	Crazy	6
	Excited	11		
	Concerned	11	Biting	8
	Happy	11	Content	8
	Content	11	Deliberate	8
	Worried	11	Melancholy	8
	Interested	11	Interested	8
			Mad	8
	Excited	13	Emotional	8
	Frantic	13	Cold	8
	Aggression	13	Curious	8
	Boisterous	13	Sharp	8
	Zealous	13	Calm	8
	Anger	13	Sad	8
			Passionate	8

Males

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Plain			Bored	5
			Serious	5
			Calm	5
			Nervous	5
			Tired	5
			Concerned	5
			Anxious	5
Pleasant			Upbeat	13
			scared	13
Steady	Confident	3	Confused	3
	Calm	3	Nervous	3
	Content	3	Unsure	3
			Frustrated	3
			Upset	3
			Hesitant	3
			Excited	3
			Amused	3
			Sad	3
			Anxious	3
			Relaxed	10
			Nervous	10
			Upbeat	10
			Content	10
			Concerned	11
			Confused	11
			Friendly	11
			Tired	11
			Worried	11
			Mellow	11
Strong			Relaxed	10
			Nervous	10
			Upbeat	10
			Content	10
			Concerned	11
			Confused	11
			Friendly	11
			Tired	11
			Worried	11
			Mellow	11

Table Appendix D.1: (cont.)

Females

VQ Keyword	Positively Correlated Emotions	Negatively Correlated Emotions	
Forceful		Angry	9
		Dramatic	9
		Powerful	9
		Afraid	9
		Pissed	9
		Exciting	9
		Determined	9
		Biting	9
		Stressed	9
		Stern	9
		Contemptuous	9
		Frightened	10
		Afraid	10
		Emotional	10
		Desperation	10
		Vengeful	10
		Passionate	10
		Dramatic	10
		Upset	10
		Mysterious	10
		Serious	11
		Panicked	11
		Scared	11
		Desperate	11
		Determined	11
		Urgent	
		Anxious	11
		Cold	11
			11
		Anxious	
		Mad	13
		Sad	13
		Crazy	13
		Desperate	13
		Worried	13
		Exciting	13
		Dramatic	13
		Fear	13
		Thoughtful	13
		Passionate	13
		Excitement	13
		Calm	13
			13

Males

VQ Keyword	Positively Correlated Emotions		Negatively Correlated Emotions	
Stuttering	Relaxed	10	Confident	3
	Nervous	10	Calm	3
	Upbeat	10	Content	3
	Content	10		
			Amused	10
	Upbeat	13	Upset	10
	Scared	13	Irritated	10
			Steady	10
			Normal	10
			Annoyed	10
			Amused	13
			funny	13

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Growling			Serious	6
			Nervous	6
			Alert	6
			Excited	6
			Boisterous	6
			Sad	6
			Emotional	6
			Concerned	6
			Determined	6
			Passionate	6
			Haunted	6
			Prayer	6
			Anxious	6
			Zealous	6
			Crazy	6
			Biting	8
			Content	8
			Deliberate	8
			Melancholy	8
			Interested	8
			Mad	8
			Emotional	8
			Cold	8
			Curious	8
			Sharp	8
			Calm	8
			Sad	8
			Passionate	8
Hoarse	Excited	13	Happy	10
	Frantic	13	Determined	10
	Aggression	13	Confident	10
	Boisterous	13	Urgent	10
	Zealous	13	Tense	10
	Anger	13	Demanding	10
			Excited	10
			Suspenseful	10
			Aggression	10
			Anger	10
			Panic	10
			Anxious	10

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Hoarse			Anxious	13
			Mad	13
			Sad	13
			Crazy	13
			Desperate	13
			Worried	13
			Exciting	13
			Dramatic	13
			Fear	13
			Thoughtful	13
			Passionate	13
			Excitement	13
			Calm	13
			Rehearsed	13
Intensity	Excitement	7	Passionate	7
	Fear	7	Stern	7
	Emotional,	7	Frightened	7
	Urgent,	7	Tense	7
	Excited,	7	Contemptuous	7
	Furious,	7	Confident	7
	Desperate,	7	Scared	7
	Chilled,	7	Exciting	7
	Defiance	7	Assertive	7
	Fearless	7	Interested	7
	Indignant	7	Calm	7
	Spiteful Upset	7	Emphatic	7
	Nervous	7	Thoughtful	7
	Concerned	7	Deliberate	7
	Aggressive	7		
	Angry	7		
		7		
Mysterious			Angry	9
			Dramatic	9
			Afraid	9
			Pissed	9
			Exciting	9
			Determined	9
			Biting	9
			Stressed	9
			Stern	9
			Contemptuous	9
			Happy	10
			Determined	10
			Confident	10
			Urgent	10
			Tense	10
			Demanding	10
			Excited	10
			Suspenseful	10
			Aggression	10
			Anger	10

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Mysterious			Panic	10
			Anxious	10
			Dramatic	11
			Sad	11
			Emotional	11
			Afraid	11
			Fear	11
			Melancholy	11
			Excited	11
			Concerned	11
			Happy	11
			Content	11
			Worried	11
			Interested	11
Old	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Forceful	2
	Worried	2	Powerful	2
	Tense	2	Excited	2
	Haunted	2	Confident	2
	Prayer	2	Emotional	2
	Alert	2	Dramatic	2
	Desperate	2	Aggressive	2
	Passionate	5	Scared	4
	Emotional	5	Anxious	4
	Desperate	5	Worried	4
	Aggressive	5	Sad	4
	Pleading	5	Emotional	4
	Forceful	5	Desperate	4
	Demanding	5	Determined	4
	Serious	5	Afraid	4
	Eager	5	Frightened	4
	Horny	5	Excited	4
	Passion	5	Passionate	4
	Afraid	5	Urgent	4
	Scared	5	Fearful	4
			Dramatic	4
			Excited	5
			Upset	5
			Worried	5
			Frightened	5
			Anxious	5
			Dramatic	5
			Cheerful	5

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Overacting	Passionate	5	Scared	2
	Emotional	5	Afraid	2
	Desperate	5	Anxious	2
	Aggressive	5	Worried	2
	Pleading	5	Tense	2
	Demanding	5	Haunted	2
	Serious	5	Prayer	2
	Eager	5	Alert	2
	Horny	5	Desperate	2
	Passion	5		
	Afraid	5	Excited	5
	Scared	5	Upset	5
			Worried	5
	Excitement	7	Frightened	5
	Fear	7	Anxious	5
	Emotional	7	Dramatic	5
	Urgent	7	Cheerful	5
	Excited	7		
	Furious	7	Passionate	7
	Desperate	7	Stern	7
	Chilled	7	Frightened	7
	Defiance	7	Tense	7
	Fearless	7	Contemptuous	7
	Indignant	7	Confident	7
	Spiteful	7	Scared	7
	Upset		Exciting	7
	Nervous	7	Assertive	7
	Concerned	7	Interested	7
	Aggressive	7	Calm	7
	Angry	7	Emphatic	7
		7	Thoughtful	7
			Deliberate	7
Raspy	Passionate	5	Scared	4
	Emotional	5	Anxious	4
	Desperate	5	Worried	4
	Aggressive	5	Sad	4
	Pleading	5	Emotional	4
	Forceful	5	Desperate	4
	Demanding	5	Determined	4
	Serious	5	Afraid	4
	Eager	5	Frightened	4
	Horny	5	Excited	4
	Passion	5	Passionate	4
	Afraid	5	Urgent	4
	Scared	5	Fearful	4
			Dramatic	4
	Serious	6		
	Nervous	6	Bored	5
	Alert	6	Serious	5
	Excited	6	Calm	5
	Boisterous	6	Nervous	5
	Sad	6	Tired	5

Table Appendix D.1: (cont.)

Females				
VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Raspy	Emotional	6	Concerned	5
	Concerned	6	Anxious	5
	Determined	6		
	Passionate	6	Anger	6
	Haunted	6	Mad	6
	Prayer	6	Forceful	6
	Anxious	6	Urgent	6
	Zealous	6	Spitting	6
	Crazy	6	Vengeful	6
			Scared	6
			Cold	6
			Deliberate	6
			Angered	6
			Sharp	6
Screaming	Angry	9	Scared	2
	Dramatic	9	Afraid	2
	Powerful	9	Anxious	2
	Afraid	9	Worried	2
	Pissed	9	Tense	2
	Exciting	9	Haunted	2
	Determined	9	Prayer	2
	Biting	9	Alert	2
	Stressed	9	Desperate	2
	Stern	9		
	Contemptuous	9	Passionate	5
			Emotional	5
			Desperate	5
			Aggressive	5
			Pleading	5
			Forceful	5
			Demanding	5
			Serious	5
			Eager	5
			Horny	5
			Passion	5
			Afraid	5
			Scared	5
			Excited	9
			Upset	9
			Desperation	9
			Vengeful	9
			Serious	9
			Forceful	9
			Anger	9
			Cold	9
			Content	9
			Happy	9
			Calm	9
			Melancholy	9
			Emotional	9

Table Appendix D.1: (cont.)

Females				
VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Shaky	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Forceful	2
	Worried	2	Powerful	2
	Tense	2	Excited	2
	Haunted	2	Confident	2
	Prayer	2	Emotional	2
	Alert	2	Dramatic	2
	Desperate	2	Aggressive	2
	Excited	3	Sad	3
	Anxious	3	Determined	3
	Dramatic	3	Serious	3
	Scared	3	Upset	3
	Nervous	3	Furious	3
	Tense	3	Concerned	3
	Desperate	3	Possessed	3
	Excitement	3	Spiteful	3
	Fearful	3	Indignant	3
	Pleading	3	Fearless	3
	Suspenseful	3	Defiance	3
	Demanding	3	Chilled	3
			Urgent	3
	Scared	4	Worried	3
	Anxious	4		
	Worried	4	Mad	4
	Sad	4	Angry	4
	Emotional	4	Prayer	4
	Desperate	4	Haunted	4
	Determined	4	Powerful	4
	Afraid	4	Alert	4
	Frightened	4	Crazy	4
	Excited	4	Serious	4
	Passionate	4	Vengeful	4
	Urgent	4	Anger	4
	Fearful	4		
	Dramatic	4	Excitement	7
			Fear	7
	Excited	13	Emotional	7
	Frantic	13	Urgent	7
	Forceful	13	Intensity	7
	Aggression	13	Excited	7
	Boisterous	13	Furious	7
	Zealous	13	Desperate	7
	Anger	13	Chilled	7
			Defiance	7
			Fearless	7
			Indignant	7
			Spiteful	7
			Upset	7
			Nervous	7
			Concerned	7

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Shaky			Aggressive	7
			Angry	7
			Biting	8
			Content	8
			Deliberate	8
			Melancholy	8
			Interested	8
			Mad	8
			Emotional	8
			Cold	8
			Curious	8
			Sharp	8
			Calm	8
			Sad	8
			Passionate	8
			Anxious	13
			Mad	13
			Sad	13
			Crazy	13
			Desperate	13
			Worried	13
			Exciting	13
			Dramatic	13
			Fear	13
			Thoughtful	13
			Passionate	13
			Excitement	13
			Calm	13
Shouting	Serious	6	Scared	2
	Nervous	6	Afraid	2
	Alert	6	Anxious	2
	Excited	6	Worried	2
	Boisterous	6	Tense	2
	Sad	6	Haunted	2
	Emotional	6	Prayer	2
	Concerned	6	Alert	2
	Determined	6	Desperate	2
	Passionate	6		
	Haunted	6	Anger	6
	Prayer	6	Mad	6
	Anxious	6	Forceful	6
	Zealous	6	Urgent	6
	Crazy	6	Spitting	6
			Vengeful	6
			Scared	6
			Cold	6
			Deliberate	6
			Angered	6
			Sharp	6

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Spitting	Passionate	5	Scared	2
	Emotional	5	Afraid	2
	Desperate	5	Anxious	2
	Aggressive	5	Worried	2
	Pleading	5	Tense	2
	Forceful	5	Haunted	
	Demanding	5	Prayer	2
	Serious	5	Alert	2
	Eager	5	Desperate	2
	Horny	5		2
	Passion	5	Bored	
	Afraid	5	Serious	5
	Scared	5	Calm	5
			Nervous	5
	Excitement	7	Tired	5
	Fear	7	Concerned	5
	Emotional	7	Anxious	5
	Urgent	7		5
	Excited	7		5
	Furious	7	Serious	
	Desperate	7	Nervous	6
	Chilled	7	Alert	6
	Defiance	7	Excited	6
	Fearless	7	Boisterous	6
	Indignant	7	Sad	6
	Spiteful	7	Emotional	6
	Upset	7	Concerned	6
	Nervous	7	Determined	6
	Concerned	7	Passionate	6
	Aggressive	7	Haunted	6
	Angry	7	Prayer	6
			Anxious	6
			Zealous	6
			Crazy	6
				6
			Passionate	
			Stern	7
			Frightened	7
			Tense	7
			Contemptuous	7
			Confident	7
			Scared	7
			Exciting	7
			Assertive	7
			Interested	7
			Calm	7
			Emphatic	7
				7
Spooky	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Forceful	2
	Worried	2	Powerful	2

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Spooky	Tense	2	Excited	2
	Haunted	2	Confident	2
	Prayer	2	Emotional	2
	Alert	2	Dramatic	2
	Desperate	2	Aggressive	2
			Scared	4
			Anxious	4
			Worried	4
			Sad	4
			Emotional	4
			Desperate	4
			Determined	4
			Afraid	
			Frightened	4
			Excited	4
			Passionate	4
			Urgent	4
			Fearful	4
			Dramatic	4
				4
			Biting	
			Content	8
			Deliberate	8
			Melancholy	8
			Interested	8
			Mad	8
			Emotional	8
			Cold	8
			Curious	8
			Sharp	8
			Calm	8
			Sad	8
			Passionate	8
				8
Steady	Biting	8	Passionate	5
	Content	8	Emotional	5
	Deliberate	8	Desperate	5
	Melancholy	8	Aggressive	5
	Interested	8	Pleading	5
	Mad	8	Demanding	5
	Emotional	8	Serious	5
	Cold	8	Eager	5
	Curious	8	Horny	5
	Calm	8	Passion	5
	Sad	8	Afraid	5
	Passionate	8	Scared	5
			Tense	8
			Dramatic	8
			Urgent	8
			Happy	8

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Steady			Anxious	8
			Contempt	8
			Demanding	8
			Furious	8
			Crazy	8
			Cheerful	8
			Energetic	8
			Determined	8
			Angry	9
			Dramatic	9
			Powerful	9
			Afraid	9
			Pissed	9
			Exciting	9
			Determined	9
			Biting	9
			Stressed	9
			Stern	9
			Contemptuous	9
Terse	Dramatic	11	Passionate	5
	Sad	11	Emotional	5
	Emotional	11	Desperate	5
	Afraid	11	Aggressive	5
	Fear	11	Pleading	5
	Powerful	11	Demanding	5
	Melancholy	11	Serious	5
	Excited	11	Eager	5
	Concerned	11	Horny	5
	Happy	11	Passion	5
	Content	11	Afraid	5
	Worried	11	Scared	5
	Interested	11		
			Serious	6
			Nervous	6
			Alert	6
			Excited	6
			Boisterous	6
			Sad	6
			Emotional	6
			Concerned	6
			Determined	6
			Passionate	6
			Haunted	6
			Prayer	6
			Anxious	6
			Zealous	6
			Crazy	6

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Terse			Serious	11
			Panicked	11
			Scared	11
			Desperate	11
			Determined	11
			Urgent	11
			Anxious	11
			Cold	11
Theatrical	Serious	6	Anger	6
	Nervous	6	Mad	6
	Alert	6	Forceful	6
	Excited	6	Urgent	6
	Boisterous	6	Spitting	6
	Sad	6	Vengeful	6
	Emotional	6	Scared	6
	Concerned	6	Cold	6
	Determined	6	Deliberate	6
	Passionate	6	Angered	6
	Haunted	6	Sharp	6
	Prayer	6		
	Anxious	6	Anxious	13
	Zealous	6	Mad	13
	Crazy	6	Sad	13
			Crazy	13
	Excited	13	Desperate	13
	Frantic	13	Worried	13
	Forceful	13	Exciting	13
	Aggression	13	Dramatic	13
	Boisterous	13	Fear	13
	Zealous	13	Thoughtful	13
	Anger	13	Passionate	13
			Excitement	13
			Calm	13
Whispering	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Forceful	2
	Worried	2	Powerful	2
	Tense	2	Excited	2
	Haunted	2	Confident	2
	Prayer	2	Emotional	2
	Alert	2	Dramatic	2
	Desperate	2	Aggressive	2
	Excited	3	Sad	3
	Anxious	3	Determined	3
	Dramatic	3	Serious	3
	Scared	3	Upset	3
	Nervous	3	Furious	3
	Tense	3	Concerned	3
	Desperate	3	Possessed	3

Table Appendix D.1: (cont.)

Females

VQ Key-word	Positively Correlated Emotions		Negatively Correlated Emotions	
Whispering	Excitement	3	Spiteful	3
	Fearful	3	Indignant	3
	Pleading	3	Fearless	3
	Suspenseful	3	Defiance	3
	Demanding	3	Chilled	3
			Urgent	3
	Passionate	5	Worried	3
	Emotional	5		
	Desperate	5	Scared	4
	Aggressive	5	Anxious	4
	Pleading	5	Worried	4
	Demanding	5	Sad	4
	Serious	5	Emotional	4
	Eager	5	Desperate	4
	Horny	5	Determined	4
	Passion	5	Afraid	4
	Afraid	5	Frightened	4
	Scared	5	Excited	4
			Passionate	4
	Excitement	7	Urgent	4
	Fear	7	Fearful	4
	Emotional	7		
	Urgent	7	Excited	5
	Excited	7	Upset	5
	Furious	7	Worried	5
	Desperate	7	Frightened	5
	Chilled	7	Anxious	5
	Defiance	7	Dramatic	5
	Fearless	7	Cheerful	5
	Indignant	7		
	Spiteful	7	Serious	6
	Upset	7	Nervous	6
	Nervous	7	Alert	6
	Concerned	7	Excited	6
	Aggressive	7	Boisterous	6
	Angry	7	Sad	6
			Emotional	6
	Biting	8	Concerned	6
	Content	8	Determined	6
	Deliberate	8	Passionate	6
	Melancholy	8	Haunted	6
	Interested	8	Prayer	6
	Mad	8	Anxious	6
	Emotional	8	Zealous	6
	Cold	8	Crazy	6
	Curious	8		
	Calm	8	Passionate	7
	Sad	8	Stern	7
	Passionate	8	Frightened	7
			Tense	7
	Angry	9	Contemptuous	7
	Dramatic	9	Confident	7

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Whispering	Powerful	9	Scared	7
	Afraid	9	Exciting	7
	Pissed	9	Assertive	7
	Exciting	9	Interested	7
	Determined	9	Calm	7
	Biting	9	Emphatic	7
	Stressed	9		
	Stern	9	Tense	8
	Contemptuous	9	Dramatic	8
			Urgent	8
			Happy	8
			Anxious	8
			Contempt	8
			Demanding	8
			Furious	8
			Crazy	8
			Cheerful	8
			Energetic	8
			Determined	8
			Excited	9
			Upset	9
			Desperation	9
			Vengeful	9
			Serious	9
			Forceful	9
			Anger	9
			Cold	9
			Content	9
			Happy	9
			Calm	9
			Melancholy	9
			Emotional	9
			Urgent	12
			Upset	12
			Dramatic	12
			Afraid	12
			Angered	12
			Pleading	12
			Vengeful	12
			Nervous	12
			Powerful	12

Table Appendix D.1: (cont.)

Females				
VQ Key-word	**Positively Correlated Emotions**		**Negatively Correlated Emotions**	
Wizened	Scared	2	Passionate	2
	Afraid	2	Angry	2
	Anxious	2	Forceful	2
	Worried	2	Powerful	2
	Tense	2	Excited	2
	Haunted	2	Confident	2
	Prayer	2	Emotional	2
	Alert	2	Dramatic	2
	Desperate	2	Aggressive	2
	Serious		Scared	4
	Nervous	6	Anxious	4
	Alert	6	Worried	4
	Excited	6	Sad	4
	Boisterous	6	Emotional	4
	Sad	6	Desperate	4
	Emotional	6	Determined	4
	Concerned	6	Afraid	4
	Determined	6	Frightened	4
	Passionate	6	Excited	4
	Haunted	6	Passionate	4
	Prayer	6	Urgent	4
	Anxious	6	Fearful	4
	Zealous	6		
	Crazy	6	Anger	6
		6	Mad	6
			Forceful	6
			Urgent	6
			Spitting	6
			Vengeful	6
			Scared	6
			Cold	6
			Deliberate	6
			Angered	6
			Sharp	6

REFERENCES

H. Abdi and L.J. Williams, “Principal Components Analysis,” Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002.

“AES57-2011: AES standard for audio metadata – Audio object structures for preservation and restoration,” Audio Engineering Society standard available at http://www.aes.org/publications/standards/

“AES60-2011: AES standard for audio metadata – core audio metadata,” Audio Engineering Society standard available at http://www.aes.org/publications/standards/

H. Abdi and L.J. Williams, “Principal component analysis,” Wiley Interdisciplinary Reviews: Computational Statistics, 2:433-459, 2010.

Matti Airas and Paavo Alku, “Comparison of Multiple Voice Source Parameters in Different Phonation Types,” Proc. INTERSPEECH, 2007.

The American Psychological Association Monitor on Psychology, available at http://www.apa.org/monitor/2013/03/ptsd-vets.aspx , last accessed 8/20/2017.

AMI Corpus, a multi-modal data set consisting of 100 hours of meeting recordings, available at http://groups.inf.ed.ac.uk/ami/corpus/, last accessed 8/20/2017.

Mansour Alsulaiman, Zulfiquar Ali, and Ghulam Muhammad, “Gender Classification with Voice Intensity,” 5th European Symposium on Computer Modeling and Simulation, 2011.

Mansour Alsulaiman, Zulfiqar Ali, and Ghulam Muhammad, “Voice Intensity Based Gender Classification by Using Simpson’s Rule with SVM,” IWSSIP 2012.

Alzheimer’s Association, available at http://www.alz.org, last accessed 8/20/2017.

Amazon Mechanical Turk, available at https://www.mturk.com/mturk/welcome , last accessed 8/20/2017.

Gouzhen An, David Guy Brizan, and Andrew Rosenberg, “Detecting Laughter and Filled Pauses Using Syllable-based Features,” INTERSPEECH 2014.

S. Ananthakirshnan and S. Narayanan, “Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence,” IEEE transactions on audio speech, and language processing, 2008.

M. B. Arnold, “Emotion and personality,” New York: Columbia University Press, 1960.

Bishnu S. Atal, “Generalized short-time power spectra and autocorrelation function,” J. Acoust. Soc. Am., 34, 1679-1683, 1962.

Tanja Banziger and Klaus R. Scherer, “The role of intonation in emotional expressions,” Speech Communications, 46(3-4):252-267, 2005.

Viviane Barrichelo-Lindstrom and Mara Behlau, “Resonant Voice in Acting Students: Perceptual and Acoustic Correlates of the Trained Y-Buzz by Lessac,” poster from the 2nd IALP International Composium, 2007.

D.A. Boyd, “OHMS: enhancing access to oral history for free,” In D. Boyd, S. Cohen, B. Rakerd, and D. Rehberger (Eds.), Oral history in the digital age, Institute of Library and Museum Services. Retrieved from http://ohda.matrix.msu.edu/2012/06/ohms-2/

B. Bozkurt, B. Doval, C. D’Alessandro, T. Dutoit, “A Method for Glottal Formant Frequency Estimation,” Proc. INTERSPEECH - ICSLP, 2004.

Elif Bozkurt, Orith Toledo-Ronen, Alexander Sorin, and Ron Hoory, “Exploring Modulation Spectrum Features for Speech-Based Depression Level Classification,” INTERSPEECH 2014.

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, “Analysis of Emotionally Salient Aspects of fundamental Frequency for Emotion Detection,” IEEE Transactions on Audio, Speech, and Language Processing,” 17(4):582-596, 2009.

Carlos Busso and Tauhidur Rahman, “Unveiling the Acoustic Properties that Describe the Valence Dimension,” INTERSPEECH, 2012.

John P. Campbell and Thomas E. Tremain, “Voice/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm,” ICASSP 1986.

W.N. Campbell, “Analog i/o nets for syllable timing,” Speech Communication, 9:57-61, Feb. 1990.

W.N. Campbell, “Syllable based segment duration,” in Talking Machines: Theories, Models, and Designs (G. Bailey, C. Benoit, and T.R. Sawallis, eds.), 211-224, Elsevier, 1992.

Jean Carletta, “Unleasing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus,” 41(2):181-190, 2007.

M.A. Carlin, B.Y. Smolenski, and S.J.Wendt, “Unsupervised detection of whispered speech in the presence of normal phonation,” Proc. INTERSPEECH, 2006.

D.G. Childers and C.K. Lee, “Vocal quality factors: Analysis, synthesis, and perception,” J. Acoust. Soc. Am.,90(5): 2394- 2410, 1991.

Michael Christel and Michael Frisch, “Evaluating the contributions of video representation for a life oral history collection,” JCDL ’08.

Michael Christel, Julieanna Richardson, and Howard Wactlar, “Facilitating access to large digital oral history archives through informedia technologies,” JCDL ’06.

Michael Christel, Scott Stevens, Bryan Maher, and Julianna Richardson, “Enhanced exploration of oral history archives through processed video and synchronized text transcripts,” MM ’10.

Michael Christel, “Establishing the utility of non-text search for news video retrieval with real world users,” MULTIMEDIA ’07.

Michael Christel, “Examining User Interactions with Video Retrieval Systems,” Proc SPIE Multimedia Content Access: Algorithms and Systems, 6506.

Michael Christel and Rong Yan, “Merging storyboard strategies and automatic retrieval for improving interactive video search,” CIVR ’07.

Steve Cohen, Brad Rakerd, Doug Boyd, Dean Rehberger, “Oral History in the Digital Age: The Imperative for Rethinking Best Practices based on a Survey of the Field(s),” available at http://ohda.matrix.msu.edu/2012/07/ohda-survey/

Jennifer Cole, Yoonsook Mo, and Mark Hasegawa-Johnson, “Signal-based and expectation-based factors in the perception of prosodic prominence,” Laboratory phonology 1.1:425-452, 2010.

Pierre Comon, “Independent Component Analysis: a new concept? Signal Processing, 36(3):287-314, 1994.

CrowdFlower, available at https://www.crowdflower.com/, last accessed 8/20/17.

Ailbhe Cullen, John Kane, Thomas Drugman, and Naomi Harte, “Creaky Voice and the Classification of Affect,” Workshop in Affective and Social Speech Signals, WASSS 2013.

G. Corrigan, N. Massey, O. Karaali, “Generating Segment Durations in a Text-to-speech system: A Hybrid Rule-based/Neural Network Approach,” Eurospeech, 1997.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, and Jarek Krajewski,” Probabilistic Acoustic Volume Analysis for Speech Affected by Depression,” INTERSPEECH 2014.

George Edward Dahl, “Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing,” PhD Dissertation, University of Toronto, 2015.

T. Edward Damer, “Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments,” Wadsworth Publishing/Cengage Learning, 2009.

Li Deng, “Deep Learning for Speech/Language Processing – machine learning & signal processing perspectives,” Tutorial given at INTERSPEECH, Sept 6, 2015.

Li Deng and Dong Yu, “Deep Learning Methods and Applications,” Foundations and Trends in Signal Processing, 7(3-4):198-387, 2014.

Duo Ding, Florian Metze Shourabh Rawat, Peter Franz Schulam, Susanne Buerger, Ehsan Younessian, Lei Bao, Michael Christel, and Alexander Hauptmann, “Beyond audio and video retrieval: towards multimedia summarization,” ICMR ’12.

Alan Dix, Janet Finlay, Gregory D. Abowd, and Russell Beale, “Human-Computer Interaction,” Pearson/Prentice Hall, 2004.

Sidney D’Mello and Rafael A. Calvo, “Beyond the basic emotions: what should affective computing compute?” CHI EA 2013.

Thomas Drugman, John Kane, and Christer Gobl (1), “Modeling the Creaky Excitation for Parametric Speech Synthesis,” INTERSPEECH 2012.

Thomas Drugman, John Kane, and Christer Gobl (2), “Resonator-based Creaky Voice Detection,” INTERSPEECH 2012.

Thomas Drugman, John Kane, and Christer Gobl, “Data-driven Detection and Analysis of the Patterns of Creaky Voice,” Computer Speech & Language, 28(5): 1233-1253, 2014.

Dublin Core Metadata Initiative, “User Guide,” 2011. Available at http://wiki.dublincore.org/index.php/UserGuide

ELAN, a professional tool for the creation of complex annotations on video and audio resources, available at https://tla.mpi.nl/tools/tla-tools/elan/, last accessed 8/20/17.

P. Elkman, “Emotion in the human face (2nd ed.),” New York, Cambridge University Press, 1982.

Julien Epps, Roddy Cowie, Shrikanth Narayanan, Bjorn Schuller, and Jianhua Tao, “Emotion and mental state recognition from speech,” EURASIP Journal on Advances in Signal Processing 15, 2012.

D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, “Exploratory study of some acoustic and articulatory characteristics of sad speech,” Phonetica, 63:1-25, 2004.

S. Erk, M. Kiefer, J. Grothe, AP Wunderlich, M. Spitzer, and H. Walter, “Emotional context modulates subsequent memory effect,” Neuroimage, 18, pp 439-447, 2003.

Florian Eyben, Martin Wollmer, and Bjorn Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," ACM Transactions on Interactive Intelligent Systems (TiiS) – Special Issue on Affective Interaction in Natural Environments, 2(1):2012.

G. Fant and A. Kruckenberg, "On the quantal nature of speech timing, Spoken Language, 1996," Proceedings of the ICSLP, pp:2044-2047, 1996.

FAVE-align, an online interface for the Penn Forced Aligner, available at http://fave.ling.upenn.edu/FAAValign.html , last viewed 4/22/16.

Peter W. Foltz, "Latent Semantic Analysis for Text-Based Research," Behavior Research Methods, Instruments, and Computers, 28(2):197-202, 1996.

M. Forsell, "Acoustic Correlates of Perceived Emotions in Speech," Master of Science Thesis at the School of Media Technology, Royal Institute of Technology, 2007.

Mark Gales and Steve Young, "The Application of Hidden Markov Models in Speech Recognition," Foundations and Trends in Signal Processing, 1(3):195-304, 2008.

P. Gangamohan, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty, and B. Yegnanarayana, "Excitation Source Features for Discrimination of Anger and Happy Emotions," INTERSPEECH 2014.

James Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, "An Introduction to Statistical Learning," Springer, New York, 2015.

Linda Gates, "Voice for Performance – Training the Actor's Voice Second Edition," Applause Books, 2000.

D. Gerhard, "Pitch Extraction and Fundamental Frequency: History and Current Techniques," Technical Report, Dept. of Computer Science, University of Regina, 2003. Available at http://www.cs.uregina.ca/Research/Techreports/2003-06.pdf.

Bruce R. Gerratt and Jody Kreiman, "Toward a taxonomy of nonmodal phonation," Journal of Phonetics, 29:365-381, 2001.

Christer Gobl, and Ailbhe Ni Chasaide, "The role of voice quality in communicating emotion, mood, and attitude," Speech Communication 40:189-212, 2003.

Matthew Gordon, "Disentangling stress and pitch accent: Toward a typology of prominence at different prosodic levels," in Harry van der Hulst (ed.), Word Stress: Theoretical and Typological Issues, pp. 83-118, Oxford University Press, 2014.

D. Gowda and M. Kurimo. “Analysis of breathy, modal and pressed phonation based on low frequency spectral density,” Proc. INTERSPEECH, 2013.

David Graff, “An Overview of Broadcast News corpora,” Speech Communications, 37: 15-26, 2005.

J.A. Gray, “The neurophysiology of anxiety,” Oxford: Oxford University Press, 1982.

Hamlet Soliloquy Performance, David Tennant, Available at https://www.youtube.com/watch?v=xYZHb2xo0OI .

Hamlet Soliloquy Performance, Derek Jacobi, Available at https://www.youtube.com/watch?v=-elDeJaPWGg .

Hamlet Soliloquy Performance, Kenneth Branagh, Available at https://www.youtube.com/watch?v=SjuZq-8PUw0 .

Hamlet Soliloquy Performance, Mel Gibson, Available at http:// https://www.youtube.com/watch?v=Vf2TpWsPvgI .

Hamlet Soliloquy Performance, Richard Burton, Available at https://www.youtube.com/watch?v=lsrOXAY1arg .

Helen Hanson, “Glottal characteristics of female speakers: acoustic correlates,” JASA, 101(1):466-81.

Helen Hanson, “Glottal characteristics of female speakers,” PhD Dissertation, Harvard University, 1995.

Helen Hanson and Erika Chuang, “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data,” JASA106(2):1064-1077, 1999.

James Hillenbrand and Michael Clark, “The role of F0 and formant frequencies in distinguishing the voices of men and women,” Attention, Perception, & Psychophysics, 71(5):1150-1166, 2009.

James Hillenbrand and Michael Clark, “The role of F0 and formant frequencies in distinguishing the voices of men and women,” Attention, Perception, and Psychophysics, 71(5):1150-1166, 2009.

James Hillenbrand and Robert Houde, “Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech,” Journal of Speech and Hearing Research 39:31-321, 1996.

The History Makers Video Oral History Collection, available at http://www.thehistorymakers.com

Florian Honig, Anton Batliner, Elmar Noth, Sebastian Schnieder, and Jarek Krajewski, "Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender," INTERSPEECH 2014.

Panagiotis G. Ipeirotis, "Analyzing the Amazon Mechanical Turk Marketplace," ACM XRDS (Crossroads), 17(2), Winter 2010.

T. Itoh, K. Takida, and F. Itakura, "Analysis and recognition of whispered speech," Speech Communications, 45: 139-152, 2005.

Carlos Toshinori Ishi, "Analysis of Autocorrelation-based Parameters for Creaky Voice Detection," Proceedings of Speech Prosody, 2004.

Carlos Toshinori Ishi, Ken-Ichi Sakakibara, Hiroshi Ishiguro, and Norihiro Hagita, "A Method for Automatic Detection of Vocal Fry," IEEE Transactions on Audio, Speech, and Language Processing, 16(1):47-56, 2008.

C.E. Izard, "The face of emotion," New York: Appleton-Century-Crofts, 1971.

Tom Johnstone and Klaus R. Scherer, "The Effects of Emotions on Voice Quality," Proc. 14th International Conference on Phonetic Sciences, 2029-2032, 1999.

Chuck Jones, "Make Your Voice Heard, An Actor's Guide to Increased Dramatic Range Through Vocal Training," Back Stage Books, 2005.

K.J. Kallail and F.W. Emanuel, "Formant-frequency differences between isolated whisper and phonated vowel samples produced by adult female subjects," Journal of Speech and Hearing Research, 27: 245-251, 1984.

J. Kane and C. Gobi, "Identifying regions of non-modal phonation using features of the wavelet transform," in Proc. INTERSPEECH, Florence, Italy, Aug. 2011, pp. 177-180.

Lakshmish Kaushik, Abhijeet Sangwan, and John H.L. Hansen, "Laughter and Filler Detection in Naturalistic Audio," INTERSPEECH 2015.

Patricia Keating, Marc Garellek, and Jody Kreiman, "Acoustic properties of different kinds of creaky voice," 18th International Congress of Phonetic Sciences, Glasgow, Scotland, 2015.

Lyndon S. Kennedy and Daniel P.W. Ellis, "Laughter Detection in Meetings," In Proceedings of the NIST Meeting Recognition Workshop, 2004.

D.H. Klatt, "Synthesis by rule of segmental durations in English sentences, " in Frontiers of Speech Communication Research, pp. 287-300, New York: Academic Press, 1979.

D. Klatt and L. Klatt, “Analysis, synthesis, and perception of voice quality variations among male and female talkers,” J. Acoust. Soc. Am. 87:820-857, 1990.

Mary Tai Knox and Nikki Mirghafori, “Automatic Laughter Detection Using Neural Networks,” INTERSPEECH 2007.

G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, “Loudness predicts prominence: fundamental frequency lends little,” J Acoust Soc Am. 118(2):1038-1054.

Shashidhar G. Koolagudi, Sourav Nandy, and K. Sreenivasa Rao, “Spectral Features for Emotion Classification,” IACC 2009.

Margarita Kotti and Constantine Kotropoulos, “Gender Classification in Two Emotional Speech Databases,” International Conference on Pattern Recognition (ICPR) 2008.

Sreenivasa Rao Krothapalli and Shashidhar G. Koolagudi, “Emotion Recognition using Speech Features,” Springer New York, 2013.

S.R.R. Kumar, “Significance of durational knowledge for speech synthesis in an Indian language, “ Master’s thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology, Madras, Mar. 1990.

Lady Macbeth Soliloquy Performance, Judi Dench, Available at https://www.youtube.com/watch?v=2xHlngY6Bgk&list=PLJ7XsNHDDVSkFLkEohiEtqO3g184DXft7&index=5 , last viewed 4/22/16.

Lady Macbeth Soliloquy Performance, Harriet Walter, Available at https://www.youtube.com/watch?v=hREqqNr9AyI , last viewed 4/22/16.

Lady Macbeth Soliloquy Performance, Joanne Whalley, Available at https://www.youtube.com/watch?v=TiBBLlb_pZ0 , last viewed 4/22/16.

Lady Macbeth Soliloquy Performance, Kate Fleetwood, Available at https://www.youtube.com/watch?v=RM8QQuz5BP4 , last viewed 4/22/16.

Lady Macbeth Soliloquy Performance, Allison Jean White, Available at https://www.youtube.com/watch?v=ft2Lthl9q5Y , last viewed 4/22/16.

Sachin Lakra, Juhi Singh, and Arun Kumar Singh, “Automated Pitch-Based Gender Recognition using an Adaptive Neuro-Fuzzy Interface System,” ISSP 2013.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham, “An Introduction to Latent Semantic Analysis,” Discourse Processes, 25(2&3):259-284: 1998.

Patricia Leavy, “Oral History, Understanding Qualitative Research,” Oxford University Press, 2011.

Chi-Chun Lee, Emily Mower, Carlos Busso, Sunbok Lee, and Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Communication 53:1162-1171, 2011.

Daniel D. Lee and H. Sebastian Seung, "Algorithms for Non-negative Matrix Factorization," NIPS 2000.

Library of Congress Civil Rights History Project, available at https://www.loc.gov/collections/civil-rights-history-project/

Library of Congress Veterans History Project, available at http://memory.loc.gov/diglib/vhp/html/search/search.html or http://www.loc.gov/vets/, accessed 9/10/16.

Library of Congress (LOC) Veterans History Project Interview: Joseph Daniel Ancona Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Dax Ashlee Carpenter Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Andrew James Chier Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Nicole Cabral Ferretti Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Christopher M. Gamblin Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Jeremy Brandon Hurtt Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Amanda R. Fichera Kean Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Ingrid C. Lim Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC) Veterans History Project Interview: Teresa Michelle Little Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of Congress (LOC)Veterans History Project Interview: Elida Trinidad Fernandez Sluss Collection, available at https://memory.loc.gov/diglib/vhp/html/search/search.html , last accessed 8/20/2017.

Library of congress (LOC) Veterans History Project, Use of Collection Materials, available at https://www.loc.gov/vets/vets-questions.html#research7, last accessed 10/31/2017.

Boon Pang Lim. "Computational Differences in Whispered and Non-Whispered Speech," Dissertation, University of Ilinois at Urbana-Champaign, 2010.

P.A. Lewis and H.D. Critchley, "Mood-dependent memory," Trends in Cognitive Sciences, 7(9) 2003.

Linguistic Data Consortium (LDC), Memberships Page, available at https://catalog.ldc.upenn.edu/memberships, last accessed 8/14/2017.

Macmillan Dictionary, "Words used to describe someone's voice," available at http://www.macmillandictionary.com/us/thesaurus-category/american/words-used-to-describe-someone-s-voice , last viewed 4/22/16.

Elinor A. Maze, "Metadata: Best Practices for Oral History Access and Preservation," Baylor University Institute for Oral History, Available at http://ohda.matrix.msu.edu/2012/06/metadata/, accessed 11/11/2015.

A. Mehrabian, "Silent Messages," Wadsworth, 1971.

A. Minard, M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Xchoen, and C. van Son, "MEANTIME, the NewsReader Multilingual Event and Time Corpus," Proc LREC, 2016.

B. Mobius and J.P.H. vanSanten, "Modeling segmental duration in German text-to-speech synthesis," Proc ICSLP, 4:2395-2398, 1996.

Tsuyoshi Moriyama, Hideo Saito, and Shinji Ozawa, "Evaluation of the Relation Between Emotional Concepts and Emotional Parameters in Speech," ICASSP 1997.

R.W. Morris, "Enhancement and recognition of whispered speech," PhD Dissertation, Georgia Institute of Technology 2003.

O.H. Mowrer, "Learning theory and behavior," New York: Wiley, 1960.

N.P. Narendra and K. Sreenivasa Rao, “Automatic detection of creaky voice using epoch parameters,” INTERSPEECH 2015.

Md Nasir, Wei Xia, Bo Xiao, Brian Baucom, Shrikanth S. Narayanan, and Panayiotis Georgiou, “Still Together: The Role of Acoustic Features in Predicting Marital Outcome,” INTERSPEECH, 2015.

National Center for Voice and Speech, Tutorial on Voice Qualities. Available at http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html.

Tilda Neuberger and Andras Beke, “Automatic Laughter Detection in Spontaneous Speech Using GMM-SVM Method,” TSD 2013.

J. Panksepp, “Toward a general psychobiological theory of emotions,” The Behavioral and Brain Sciences, 5:407-409, 1982.

Phouc Nguyen, Dat Tran, Xu Huang, and Dharmendra Sharma, “Automatic Classification of Speaker Characteristics,” ICCE 2010.

R. Plutchik, “A general psychoevolutionary theory of emotion,” In R. Plutchik and H. Kellerman (Eds.), “Emotion: Theory, research, and experience: Vol. 1. Theories of emotion,” New York: Academic Press, 1980.

Jieun Oh, Eunjoon Cho, and Malcolm Slaney, “Characteristic Contours of Syllabic-level Units in Laughter,” INTERSPEECH 2013.

Andrew Ortony and Terence J. Turner, “What’s Basic About Basic Emotions?” Psychological Review, 97(3):315-331, 1990.

The Penn Phonetics Lab Forced Aligner, available at https://www.ling.upenn.edu/phonetics/old_website_2015/p2fa/ , last viewed 4/22/16.

Praat, phonetics investigation software, available at http://www.fon.hum.uva.nl/praat/, last accessed 8/20/2017.

Online thesaurus of English, available at http://www.thesaurus.com , last viewed 4/22/16.

Oral History Association OHDA Essay Collection (best practices), available at http://www.oralhistory.org/ohda-essays/

Oral History Online, Available at http://historymatters.gmu.edu/mse/oral/online.html

Andrew Ortony and Terence J. Turner, “What’s Basic About Basic Emotions?” Psychological Review, 97(3):315-331.

Hal Pashler and John Wixted, eds., “Stevens’ Handbook of Experimental Psychology, Third Edition, Volume 4: Methodology in Experimental Psychology,” John Wiley& Sons, 2002.

Marc Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja Kotz, “Factors in the recognition of vocally expressed emotions: A comparison of four languages,” J Phonetics, 37:417-435, 2009.

Elizabeth A. Peterson, Nelson Roy, Shaheen Awan, Ray M. Merrill, Russell Banks, Kristine Tanner, “Toward Validation of the Cepstral Spectral Index of Dysphonia (CSID) as an Objective Treatment Outcomes Measure,” J. Voice, 27(4):401-10, 2013.

G.E. Peterson and H.L. Barney, “Control methods used in the study of the vowels,” JASA 24(2):175-184, 1952.

Hartmut R. Pfitzinger, “Intrinsic phone durations are speaker-specific,” ICSLP – INTERSPEECH, 2002a.

Harmut R. Pfitzinger, “Local Speech Rate as a Combination of Syllable and Phone Rate,” Proc. ICSLP, 3:1087-1090, 1998.

Hartmut R. Pfitzinger, “Reducing Segmental Duration Variation by Local Speech Rate Normalization of Large Spoken Language Resources,” LREC, 2002b.

J. Pierrehumbert and J. Hirschberg, “The meaning of intonational contours in the interpretation of discourse,” In: P.R. Cohen, J. Morgan, M.E. Pollack, (Eds.), Intentions in Communication, MIT Press, Cambridge, MA, pp. 271-311, 1990.

Mary Pietrowicz, Danish Chopra, Amin Sadeghi, Puneet Chandra, Brian P. Bailey, and Karrie Karahalios, “CrowdBand: An Automated Crowdsourcing Sound Composition System,” HCOMP 2013.

M. Pietrowicz, M. Hasegawa-Johnson, K. Karahalios, “Acoustic Correlates for Perceived Effort Levels in Expressive Speech,” INTERSPEECH 2015.

M. Pietrowicz, M. Hasegawa-Johnson, K. Karahalios (1), “Discovering Dimensions of Perceived Vocal Expression in Semi-structured, Unscripted Oral History Accounts,” International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2017.

M. Pietrowicz, M. Hasegawa-Johnson, K. Karahalios (2), “Acoustic correlates for perceived effort levels in male and female acted voices,” Journal of the Acoustical Society of America (JASA), accepted for publication July 2017.

Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner, “Anger recognition in speech using acoustic and linguistic cues,” Speech Communication, 53:1198-1209, 2011.

Lawrence Rabiner and Ronald Schafer, “Theory and Applications of Digital Speech Processing,” Pearson, 2011.

K. Sreenivasa Rao, “Predicting Prosody from Text for Text-to-Speech Synthesis,” Springer Science+Business Media New York, 2012.

M.D. Riley, “Tree-based modeling of segmental durations,” in G. Bailey, C. Benoit, and T.R. Sawallis, eds., “Talking Machines: Theories, Models and Designs,” pp265-273, 1992.

Amy J. Schafer, Shari R. Speer, Paul Warren, and S. David White, “Intonational Disambiguation in Sentence Production and Comprehension, “ Journal of Psycholinguistic Research, 29(2), 2000.

Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency, “Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD,” INTERSPEECH 2013.

M. R. Schroeder and Bishnu S. Atal, “Generalized short-time power spectra and autocorrelation functions,” J. Acoust. Soc. Am., 34(11):1679-1683, 1962.

Bjorn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, “Recognizing realistic emotions in affect in speech: State of the art and lessons learnt from the first challenge,” Speech Communication, 53:1062-1087, 2012.

G. Schwarz. Estimating the dimension of a model. Ann. Statist., 6(2):461–464, 1978.

William Shakespeare, “Hamlet (Signet Classic Shakespeare),” Signet Classic, 1998.

William Shakespeare, and eds. Barbara A. Mowat and Paul Werstine, “Macbeth,” Simon & Schuster Paperbacks, 2013.

D.S. Shete, and S.B. Patil, “Zero crossing rate and Energy of the Speech Signal of Devanagari Script,” IOSR-JVSP 4(1): 1-5, 2014.

Yen-Liang Shue and Marngus Iseli, “The Role of Voice Source Measures on Automatic Gender Classification, “ ICASSP 2008.

Robert Smallwood (ed), “Players of Shakespeare 4: Further Essays in Shakespearian Performance by Players with the Royal Shakespeare company,” Cambridge University Press, 2003.

Robert Smallwood (ed), “Players of Shakespeare 5,” Cambridge University Press,

Cara G. Smith, Eileen M. Finnegan, and Michael P. Karnell, “Resonant Voice: Spectral and Nasendoscopic Analysis,” Journal of Voice, 19(4): 607-622, 2005.

Peng Song, Jinglei Liu, Yun Jin, and Yanwei Yu, “Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization,” Speech Communication, 2015

Peng Song, Yun Jin, Cheng Zha, and Li Zhao, “Speech emotion recognition method based on hidden factor analysis,” IEEE Electronics Letters, 51(1): 112-114, 2015.

SoX, Sound eXchange sound processing software, available at http://sox.sourceforge.net, last accessed 8/20/2017.

The StoryCorps collection, available at https://storycorps.org, and via the Library of Congress.

Yla R. Tausczik and James W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, Journal of Language and Social Psychology 29(1) 24-54, 2010.

TED, a collection of short talks, 18 minutes or less, on Technology, Entertainment, and Design, available at https://www.ted.com, last accessed 8/20/2017.

Paul Thompson, “The Voice of the Past Oral History,” Third Edition, Oxford University Press, 2000.

Ingo Titze, “Acoustic Interpretation of Resonant Voice,” Journal of Voice, 14(4):519-528, 2001.

S. S. Tomkins, “Affect theory,” In K.R. Scherer & P. Elkman (Eds.), “Approaches to emotion,” Hillsdale, NJ: Erlbaum, 1984.

Tumblr Writing Helpers Blog, “55 Words to Describe Someone’s Voice,” available at http://writinghelpers.tumblr.com/post/41621570418/55-words-to-describe-someones-voice , last viewed 4/22/16.

J.P.H. van Santen, “Assignment of segment duration in text-to-speech synthesis,” Computer Speech and Language, 8:95-128, Apr. 1994

J.P.H. van Santen, “Segmental duration and speech timing,” in Computing Prosody, Computational models for processing spontaneous speech, ch. 15, pp. 225-249, Springer-Verlag, 1996.

J.P.H. van Santen, “Timing in text-to-speech systems,” Proc. EUROSPEECH, 2:1397-1404, 1993.

Switch Audio File Converter, available at https://switch.en.softonic.com, last accessed 8/20/2017.

Dimitrios Ververidis and Constantine Kotropoulos, “A Review of Emotional Speech Databases,” Proc. Panhellenic Conference on Informatics (PCI), 560-574, 2003.

Wavepad Audio Editing Software, available at http://www.nch.com.au/wavepad/index.html, last accessed 8/20/2017.

X. Wang, “Incorporating knowledge on segmental duration in HMM-based continuous speech recognition,” PhD thesis, University of Amsterdam, 1997.

A.B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 12,915 English lemmas,” Behavior Research Methods, 45:1191-1207, 2013.

J.B. Watson, “Behaviorism,” Chicago: University of Chicago Press, 1930.

Ratree Wayland and Allard Jongman, “Acoustic correlates of breathy and clear vowels: the case of Khmer,” Journal of Phonetics, 31:181-201, 2003.

Colin Wightman, Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti J. Price, “Segmental durations in the vicinity of prosodic phrase boundaries,” J. Acoust. Soc. Am. (JASA), 91(3):1707-1717, 1992.

Tian Wu, Yingchun Yang, Zhaohui Wu, and Dongdong Li, “MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition,” Speaker and Language Recognition Workshop, 2006.

Tae-Jin Yoon, Xiaodan Zhuang, Jennifer Cole, and Mark Hasegawa-Johnson, “Voice Quality Dependent Speech Recognition,” Language and Linguistics, 2007.

YouTube Fair Use Policy, available at https://www.youtube.com/yt/about/copyright/fair-use/#yt-copyright-protection, last accessed 10/31/2017.

YouTube Statistics, available at https://www.youtube.com/yt/press/statistics.html

Chi Zhang, “Whisper Speech Processing: Analysis, Modeling, and Detection with Applications to Keyword Spotting,” PhD Dissertation, University of Texas at Dallas, 2012.