Acoustic correlates for perceived effort levels in male and female acted voices

Mary Pietrowicz, Mark Hasegawa-Johnson, and Karrie G. Karahalios

Citation: The Journal of the Acoustical Society of America **142**, 792 (2017); doi: 10.1121/1.4997189
View online: http://dx.doi.org/10.1121/1.4997189
View Table of Contents: http://asa.scitation.org/toc/jas/142/2
Published by the Acoustical Society of America

Acoustic correlates for perceived effort levels in male and female acted voices

Mary Pietrowicz[a)]
Siebel Center for Computer Science, Department of Computer Science, University of Illinois, 201 North Goodwin Avenue, Urbana, Illinois 61801, USA

Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering, University of Illinois, 2011 Beckman Institute MC 251, 405 North Mathews, Urbana, Illinois 61801, USA

Karrie G. Karahalios
Siebel Center for Computer Science, Department of Computer Science, University of Illinois, 201 North Goodwin Avenue, Urbana, Illinois 61801, USA

(Received 29 April 2016; revised 1 June 2017; accepted 19 July 2017; published online 10 August 2017)

The best actors, particularly classic Shakespearian actors, are experts at vocal expression. With prosodic inflection, change of voice quality, and non-textual utterances, they communicate emotion, emphasize ideas, create drama, and form a complementary language which works with the text to tell the story in the script. To begin to study selected elements of vocal expression in acted speech, corpora were curated from male actors' Hamlet and female actresses' Lady Macbeth soliloquy performances. L1 speakers of American English on Mechanical Turk listened to excerpts from the corpora, and provided descriptions of the speaker's vocal expression. In this exploratory, open-ended, mixed-methods study, approximately 60% of all responses described emotion, and the remainder of responses split evenly between voice quality (including effort levels) and prosody. Also, significant differences were found in the kind and quantity of descriptors applied to male and female speech. Perception-grounded male and female acoustic feature sets which tracked the actors' expressive effort levels through the continuum of whispered, breathy, modal, and resonant speech are presented and validated via multiple models. The best results in applying these features to simple, un-optimized, four-way decision tree classifiers yielded 76% accuracy for male and 73% accuracy for female expressive, acted speech.
[http://dx.doi.org/10.1121/1.4997189]

[JFL] Pages: 792–811

I. INTRODUCTION

"To be, or not to be. That is the question."[1] Each actor who plays the famous Hamlet soliloquy from Act III Scene I of the play speaks the same words; yet, no performance of this soliloquy is like another. Line by line, each performer communicates something different via the paralingual elements of their speech. This is not unique to Hamlet, or to the male actors. Lady Macbeth similarly muses, "The raven himself is hoarse, who croaks the fatal entrance of Duncan, under my battlements."[2] In both cases, the speakers communicate different messages because their vocal expressive qualities are different. An investigation of just five male actors performing Hamlet and five female actresses performing Lady Macbeth soliloquies revealed striking differences in both phonation mode and prosody across the performances. From the prosodic perspective, speakers presented differences in speaking rate, loudness, word and syllable duration, pitch, and prominence. From the vocal quality perspective, speakers ranged across vocal effort levels from whispered, to breathy, to modal, to resonant speech. They were occasionally rough sounding, and sometimes they used vocal fry. Some voices trembled and growled. Other voices used non-word utterances such as extended, audible inhale/exhale, gasps, and squeaks. This expressive difference and variance characterizes acted speech; and professional actors, particularly classically trained Shakespearian actors, are experts at vocal expression. They work many years to acquire control over their voices (similar to music performers), and study their literature of prior performances to gain insight into the impact and effect of various vocal techniques.[3–6] Specifically, they learn how vocal techniques can be used to convey emotion or otherwise help communicate a message. By studying these experts in vocal expression, it is possible to gain insight into the perception and production of vocal expression, particularly the expressive role of vocal quality.

Given the wide range of vocal expressive techniques, this study focuses first on the vocal qualities which listeners without special training in vocal expression would be most likely to hear. To begin exploring what listeners heard in this style of speech, a sampling of utterances from the Lady

a)Electronic mail: mpietro2@illinois.edu

Macbeth and Hamlet soliloquies were presented to 400 Mechanical Turk workers (limited to L1 speakers of American English, and US workers), who provided keywords describing the vocal expression in the voices (not word content). Listeners were free to provide any keywords they thought appropriate to describe vocal expression, but most often provided descriptors of emotion, variance in loudness and speaking rate, and variation across phonation types or effort levels (see Sec. IV for study details and results). Just three effort level keywords (less than 1% of the total number of unique keywords represented) provided a disproportionate 12% of all listener description. Because listeners commented on effort levels so frequently across speakers and gender, because effort levels were low-level voice quality features which could be evaluated independently of the other perceived features, and because the data suggested potential relationships between effort levels and emotion (which could be exploited in future work), the following research questions were selected for analysis:

RQ1: What elements of vocal expression do people consciously perceive in acted voices?
RQ2: How does the perception of vocal expression in male and female acted voices differ?
RQ3: What acoustic features can distinguish each of four levels of vocal effort (whispering, breathiness, modal speech, and resonant speech) in male and female actors' expressive voices?

The primary contributions of this work are an exploratory study in the perception of male and female expressive, acted voices; an analysis of vocal features suitable for distinguishing effort levels (whispering, breathiness, modal speech, and resonance) in male and female voices; the demonstration of the continuum relationship across effort levels, and an analysis of the performance of selected vocal features in the context of four-way effort level classifiers. The investigative process, which is anchored in human perception, is a secondary contribution.

The remainder of the paper is organized as follows: (1) Sec. II presents a summary of the related work; (2) Sec. III provides an overview of the investigative process, (3) Secs. IV and V describe the data set, the curation methods, and the perception study of the data set, which all together address RQ1 and RQ2; (4) Sec. VI presents an analysis of spectral differences across phonation types, or effort levels, which influenced the selection of features (RQ3), (5) Sec. VII discusses the selection of acoustic correlates used to form the model feature vector (RQ3); (6) Sec. VIII discusses the machine learning models, experiments, validation methods, and results (RQ3); and (7) Secs. IX and X raise points of discussion and present conclusions.

II. BACKGROUND AND PRIOR WORK

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[7] 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 do not agree on the definitions of various voice qualities.

Prior work exists in the automated detection of whispered, breathy, and to a much lesser extent, resonant voice. None of it, however, has (1) addressed the range of phonation and effort across whispered, breathy, modal, and resonant speech, (2) examined the transitions across the continuum of effort levels, and (3) compared differences in effort level detection between males and females. A majority of the prior work, even the preliminary work for this project[8] focuses on male voices.

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.[9] 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.[10] 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.[11] Reduced spectral tilt is a frequent observation in unvoiced speech,[11,12] along with shifts in formant frequencies,[13] 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.[14,15]

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.[16,17] 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.[17] 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. 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.[18]

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.[19] 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.[20]

Research which examines differences in phonation types (breathy/modal/pressed) use features characterizing glottal function,[21,22] and find low-frequency spectral density (LFSD) to reflect the differences in open quotient and the corresponding increase in low frequency energy in breathy voices.[22] Amplitude quotient (AQ) and normalized amplitude quotient (NAQ) of the glottal pulse are superior separators, along with harmonic difference H1-H2,[22–24] closing quotient, quasi open quotient, and brightness.[23]

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 relationships 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.[25,26] 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 900–2300 Hz and 1000–2800 Hz.[25,26] 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.[27] Some types of creaky voice have more narrow formant bandwidths than modal speech.[28] Furthermore, singers achieve a more resonant voice by aligning formants slightly above a harmonic.[29] This paper further explores the relationship between phonation types and formant strength and leverages these differences in the selection of classifier features.

Previous studies of voice quality are often motivated by considerations of speech pathology,[30,31] phonology,[22] or speaker identity in speech synthesis;[16] 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.

A. Grounding in perception

Adequate models of paralingual, acted speech, 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 of 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.[32] A goal is discovery of mappings between human-perceived qualities and measurable acoustic features. Thinking this way could ultimately influence what is considered to be 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 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,"[33,34] 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.[35]

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.[36,37]

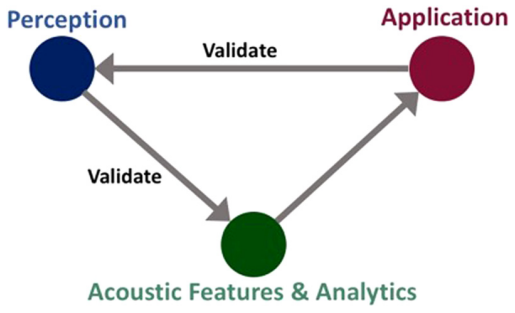

FIG. 1. (Color online) The Desired Relationships among Perception, Analytics, and Applications. Use human perception to guide the acoustic analysis and selection of features so that each measurement has a grounded relationship to human perception. Then, applications can use these grounded features and translate them to meaningful application interactions and presentations, which again, relate back to human perception.

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.[38–40] When the text and paralingual channels conflict, research shows that the paralingual elements typically win the conflict in the perception of the listener.[40]

B. Male-female speech differences

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 5 dB prior to age 16 years, and the H1-A3 will drop by about 10 dB between the ages 8–39 years. Male voices' F0 also drops during puberty. In comparison, a female talker's H1-H2 does not change very much. The H1-A3 drops by only about 4 dB from age 8–39 years, 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.[41] Also, adult males have lower formant frequencies than females, because of the differences in vocal tract length;[42] therefore, Hanson recommended correcting for vocal tract resonances. Hanson 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.[43]

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.[44,45] 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 Coefficients (MFCC) 1–2, along with vector quantization, GMM modeling, and SVM techniques. This approach yielded 97%–98% accuracy.[46]

III. RESEARCH PROCESS OVERVIEW

An overview of the end-to-end research process is given here for clarity. The purpose of this process is to allow the perception of the human to drive the investigation. If this is done correctly, the results will better align with human perception, and will therefore be better positioned to fit the needs of application development as discussed in Sec. I. 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), to the domain of quantifiable, computable acoustic features.

A. Step 1

The process began with a user study to understand what everyday listeners heard in expressive speech. In this step, Mechanical Turk workers provided keywords which described the vocal expression in Shakespearian soliloquy. Section V below provides the detailed process, analysis, and results for this step. For the convenience of relating this work to prior work, and to highlight potential relationships among groups of keywords, the analysis includes some clustering of descriptive terms into groups from prior literature (specifically, voice quality, prosody, and emotion). This clustering was the work of researchers, not the Mechanical Turk workers.

B. Step 2

Next, frequently perceived features were selected for in-depth investigation. It is important to note that the items selected for exploration here (breathiness, whispering, and resonance) were selected directly from the descriptors provided by the listeners in the user study, in those terms exactly, and their close synonyms. Section V provides the rationale for selecting these specific features for further analysis. It is also interesting to note that the terms which the users heard (breathiness, whispering, and resonance) and the "other" category of modal speech have been grouped and described in prior literature as "effort levels," "phonation types," or less specifically, "voice qualities." This common group membership suggests a relationship among the selected keywords which is explored in Secs. VI, VII, and VIII.

C. Step 3

Next, expert listeners coded a sample Shakespearian soliloquy corpus with the labels given by the listeners ("breathiness," "whispering," and "resonance") and one "other" category ("modal speech") such that classifiers could be trained to recognize the selected qualities. While it is true that expert listeners are doing the coding to train the classifiers, they are using labels grounded in human perception and discovered in steps 1 and 2. The coding process and the corpus itself are described in Sec. IV.

D. Step 4

Then, an acoustic analysis was conducted to discover acoustic features useful for recognizing and distinguishing breathiness, whispering, modal speech, and resonance. This step mapped the qualitative features which listeners reported hearing with acoustic features which can be measured quantitatively (Secs. VI and VII).

E. Step 5

Finally, classifiers were trained using the selected acoustic features, and cross-validated (Sec. VIII). The entire process is therefore anchored in human perception.

IV. THE HAMLET AND LADY MACBETH CORPORA OF ACTED VOICES

To address the research questions, corpora were created of male actors all speaking the same text, and of female actresses all speaking the same text. The corpora were designed to study experts at their craft, in roles which were typical of their gender, and to explore analytic techniques applicable to recordings which are typically available on the web (not necessarily made in controlled studio conditions). Therefore, .mp4 recordings of predominantly movie and some stage performances were collected because either this format was readily available, or download tools supported efficient conversion to this format. Speakers were selected for their professional acting ability, diversity of expressive speaking style (which occurred naturally), and diversity of origin (which included British, American, and Australian actors). Male and female corpora were curated which had the same speaking style (Shakespearian acting), and similar topic content. The Act III Scene I Hamlet soliloquy[1] ("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[2] served for the women, and featured Lady Macbeth contemplating murder of the king. The Hamlet actors included Mel Gibson,[47] Derek Jacobi,[48] Richard Burton,[49] David Tennant,[50] and Kenneth Branagh;[51] and the Lady Macbeth actresses included Judi Dench,[52] Harriet Walter,[53] Joanne Whalley,[54] Kate Fleetwood,[55] and Allison Jean White.[56]

As an example of the wide variance in expressive style across and within speakers, consider the opening sentence in the Hamlet soliloquy, "To be, or not to be. That is the question." Mel Gibson had the most businesslike interpretation, with the fastest delivery, minimal pauses, no silence, and subtle expressive variance. He took less than 2 s to utter the lines, delivered them all in one breath, and barely paused between the end of the second occurrence of "be" and the word "that." His vocal quality was slightly breathy and soft throughout, and he used prosodic inflections to emphasize the contrastive choices of being and not being. In contrast to Gibson, Branagh took about 8 s to finish the utterance, and broke the utterance into three phrases with distinct silences between them. Both occurrences of the word "be" and the word "not" were elongated, in comparison with the other words. His tone conveyed angst overall. The "to be or not to be" ranged in quality from a soft modal to breathy, and he uttered the remaining words in a chilling whisper. Finally, Jacobi's performance had the most expressive variance within speaker. He also broke the utterance into three distinct phrases, but uttered each phrase in a different expressive style. The first "to be" phrase was breathy and soft, possibly to emphasize his tenuous hold on life. The "or not to be" was resonant, in great contrast to emphasize the dire possibility of not being; and the final phrase had a modal quality with wide variance in pitch and a faster speaking rate overall, to communicate tension, confusion, or indecision.

As with Hamlet, the female actors presented different interpretations of the character, speaking the same text, with varying vocal expression. Consider the phrase from the Lady Macbeth soliloquy, "Unsex me here, and fill me from the crown to the toe top full of direst cruelty," as an example to describe this variance. Dench's performance was breathy, but also trembled, heaved, squeaked, whispered, and growled across the widest range of vocal quality; Dench was also one of the slowest speakers (14 s). Fleetwood's performance tended toward extremes in vocal quality (whispering and resonance), and had a slow speaking rate (14 s), but did not use the range of non-word utterances that Dench used. Walter's, in contrast, was conversational, with localized breathiness and whispering; and she spoke faster (11 s). White's speech was also conversational and punctuated with resonance (as opposed to Walter's breathiness); and it unfolded even faster (9 s). Whalley's performance was the most conversational, and used the smallest range in vocal quality, but had a wide prosodic variance, and unfolded almost twice as fast as the slowest speakers (8 s).

Movie and stage recordings that met the study's quality standards were acquired in .mp4 format because this format was either readily available, or download tools supported efficient conversion to this format. These quality standards excluded from consideration recordings which had high levels of background buzz/hum/crackle, had pervasive echo or reverb, or had a consistently intrusive level of studio or audience noise. Resulting recording quality was similar for stage performances and movies. These recordings were selected in part to force the development of analytic techniques which worked on a widely available recording format across a range of uncontrolled recording environments (see Sec. VIII for the results). Many recorded artifacts which fit these specifications are freely available on the internet, and analysis tools which work across this large class or recordings could have a larger impact than those which only work on recordings made in controlled studio conditions.

Next, recordings were downsampled to 16 kHz and normalized to unit amplitude. Sections with excessive sonic interference, including sections with background music, noise, multiple talkers, significant audio effects processing, and significant reverb were excluded. Then, the text was aligned to the recording, using either the Penn Phonetics Lab Forced Aligner[57] or its online interface, Fave-align.[58] Finally, all vowel sounds of at least 60 msec were extracted with the help of the forced alignment tools.

One expert listener coded the entire Hamlet corpus to the syllable level with effort levels (whispered, breathy, modal, and resonant). Note that the whispered, breathy, and resonant categories were identified as a result of the perception study described in Sec. V. The modal category was added to distinguish these exceptional voice qualities from that of typical, conversational speech (an "other" category for training machine learning models). A second expert listener coded 20 random samples of each condition across all speakers, for an inter-rater agreement of 95%, 85%, 65%, and 90% for the whispered, breathy, modal, and resonant conditions, respectively (kappa = 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, 28 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 are similar, with females having slightly greater use of breathy voice and slightly lesser use of resonant voice than males.

V. PERCEPTION OF SHAKESPEARIAN ACTED VOICES

This initial analysis sought insight into the expressive qualities which naive listeners would hear in acted voices. Of all possible expressive characteristics listeners could describe, which ones would win out in their perception? Would they notice whispering, or breathiness; and would they notice the difference between the two? Which prosodic features would they perceive and articulate? How would listeners handle emotive expression? This exploratory perception study (approved by institutional board review procedures at the University of Illinois) was used to gain insight into the perception of vocal expression, that is, to address RQ1, not to draw detailed conclusions yet about the relationships among the elements which listeners perceived, and not to provide ground truth coding for acoustic modeling. The study results were also used to guide our initial selection of perceived features for acoustic analysis.

A. Methods

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 via open response describing what they heard in the vocal expression (not in the speech text). Many workers provided multiple keywords. Multiple keywords were requested, but not required, because simplifying the task, and reducing cognitive load in Mechanical Turk tasks has been shown to improve the quality of results on this crowdsourcing platform, especially when dealing with sound.[59] Workers were not given a list of keywords to choose from, or led in any way, other than to ask them 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 five professional actors (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 in its entirety, and separately, divided into the three most likely phrases: "To be," "Or not to be," and "That is the question." This exploratory study included a total of 40 Mechanical Turk tasks for each speaker, 10 listeners for each clip. This way, listeners (as a collective) had the opportunity to hear and comment on an excerpt in its larger context, and then comment on isolated, specific, expressive styles featured in 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 the 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). 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 toe 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 including 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, 40 Mechanical Turk tasks were presented per speaker, 10 distinct listeners per each of the 4 phrases. As with Hamlet, the female actors presented different interpretations of the character, speaking the same text, with varying vocal expression.

This study design had several advantages over the traditional approach of bringing subjects into the laboratory, 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 problem of within-subject bias in small numbers of listeners. By using larger numbers of listeners, the collective response approached a population normal. Next, the Turk platform allowed limiting listeners to native U. S. speakers of English, which the study accomplished by using only U. S. workers and then collecting demographic data designed to identify qualified listeners. Also, because 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 any phrase more than once, and did not experience any 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 three 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 phase 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, a qualification task was designed to verify the ability and willingness of a worker to provide valid input, and was used to screen 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 RQ1 asks what people consciously perceive in acted voices and addresses this question in the manner listeners that 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.

B. Results

The keywords were collected for each speaker, consolidated by close synonym (as defined by thesaurus[60]), 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," "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 Figs. 2 and 3 show the most frequently given 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) differences between male and female speakers in each category. Figures 2 and 3 show the frequency of the top emotion and non-emotion keywords for all speakers in the dataset. Note that these figures do not show all of the keywords, just the most frequently given words. Tables I, II, and III subdivide voice quality and prosody into subcategories, and summarize the results statistically. Tables I and II 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 III 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.[61] 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 the speakers. Four notable voice quality concepts recurred across the speakers, and included whispering, breathiness, yelling, and resonance. Note "ringing" and "projection" are close synonyms of "resonance," and these synonymous terms appear in Tables I and II. These keywords describe phonation types and 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 as descriptors show that listeners are sensitive to not only the two qualities, but are aware of distinction between them and 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 listener 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-

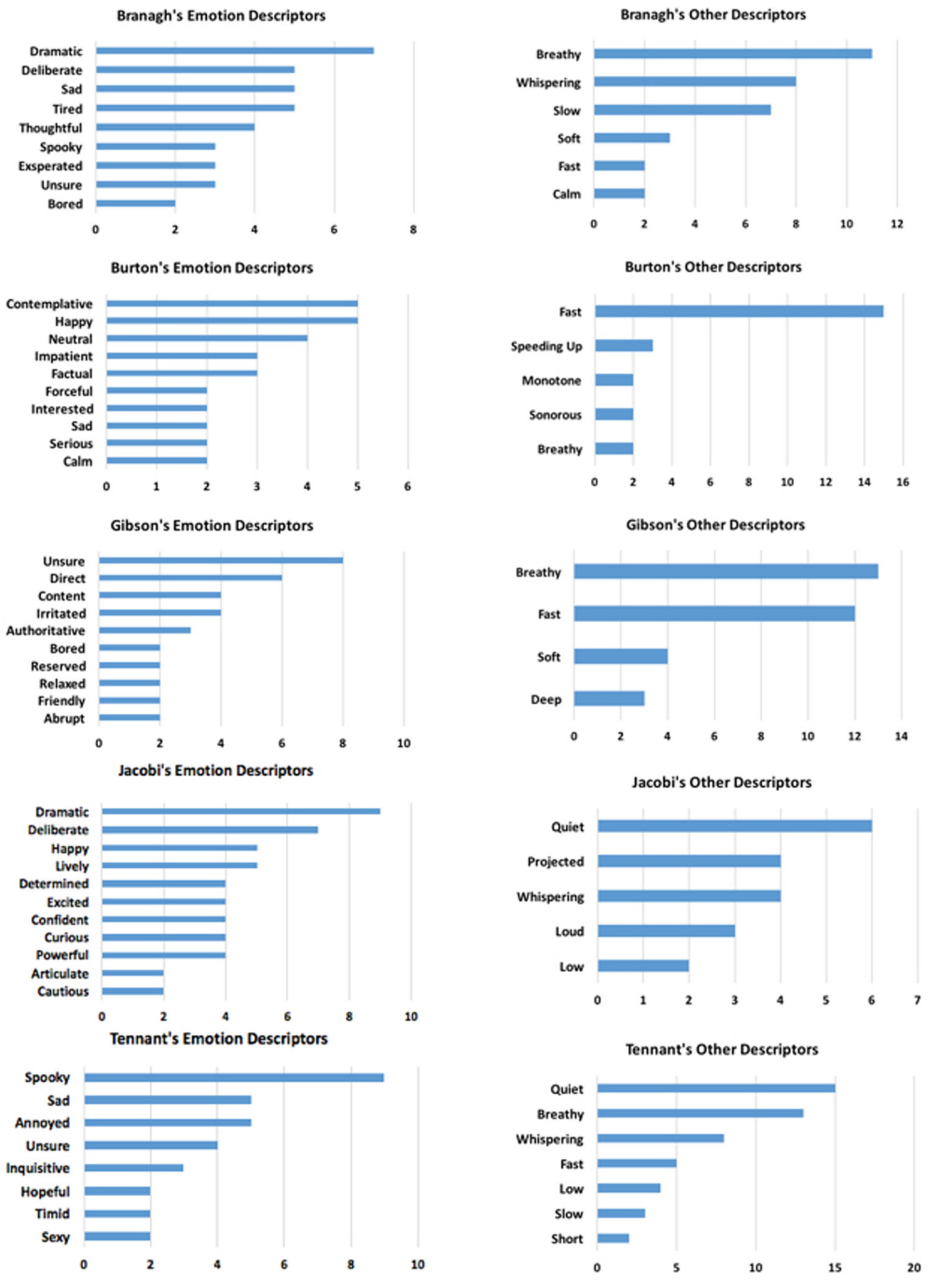

FIG. 2. (Color online) Top Keyword Descriptors for Male Acted Voices. This figure shows the most frequent emotion and non-emotion descriptors (synonym-reduced) given for each male speaker performing the Hamlet soliloquy. The emotion descriptors are nuanced, even though they have been synonym reduced. Most are not basic emotions, and few repeat across speakers. The other descriptors are concise, frequently repeating prosodic and voice quality descriptors. Whispering, breathiness, and resonance (synonomous with terms such as "sonorous," "projected," or "ringing") are all observed here. Note that not all of the descriptors given are listed here, just the most frequently given ones.

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 ($\chi^2 = 16.5$, $\mathrm{df} = 2$, $p = 0.00026$). Additional chi-square tests reveal that prosody ($\chi^2 = 6.64$, $\mathrm{df} = 1$, $p = 0.0099$), voice quality ($\chi^2 = 5.59$, $\mathrm{df} = 1$, $p = 0.018$), emotion ($\chi^2 = 16.51$, $\mathrm{df} = 1$, $p = 0.00005$), effort levels ($\chi^2 = 7.10$, $\mathrm{df} = 1$, $p = 0.0077$), and speaking rate ($\chi^2 = 6.99$, $\mathrm{df} = 1$, $p = 0.0082$) all varied significantly with gender. Non-effort-level voice quality ($\chi^2 = 0.11$, $\mathrm{df} = 1$, $p = 0.742$) and loudness ($\chi^2 = 0.011$, $\mathrm{df} = 1$, $p = 0.918$) did not vary significantly with gender. The dataset did not have a sufficient amount of 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 keywords for females is provocative. Future studies will be needed to explore the potential relationships between emotion and voice quality, especially, and in further exploring the reasons for the significantly larger proportion of emotion descriptors given for female speech.

Effort levels, specifically whispered, breathy, and resonant, were selected for in-depth investigation because

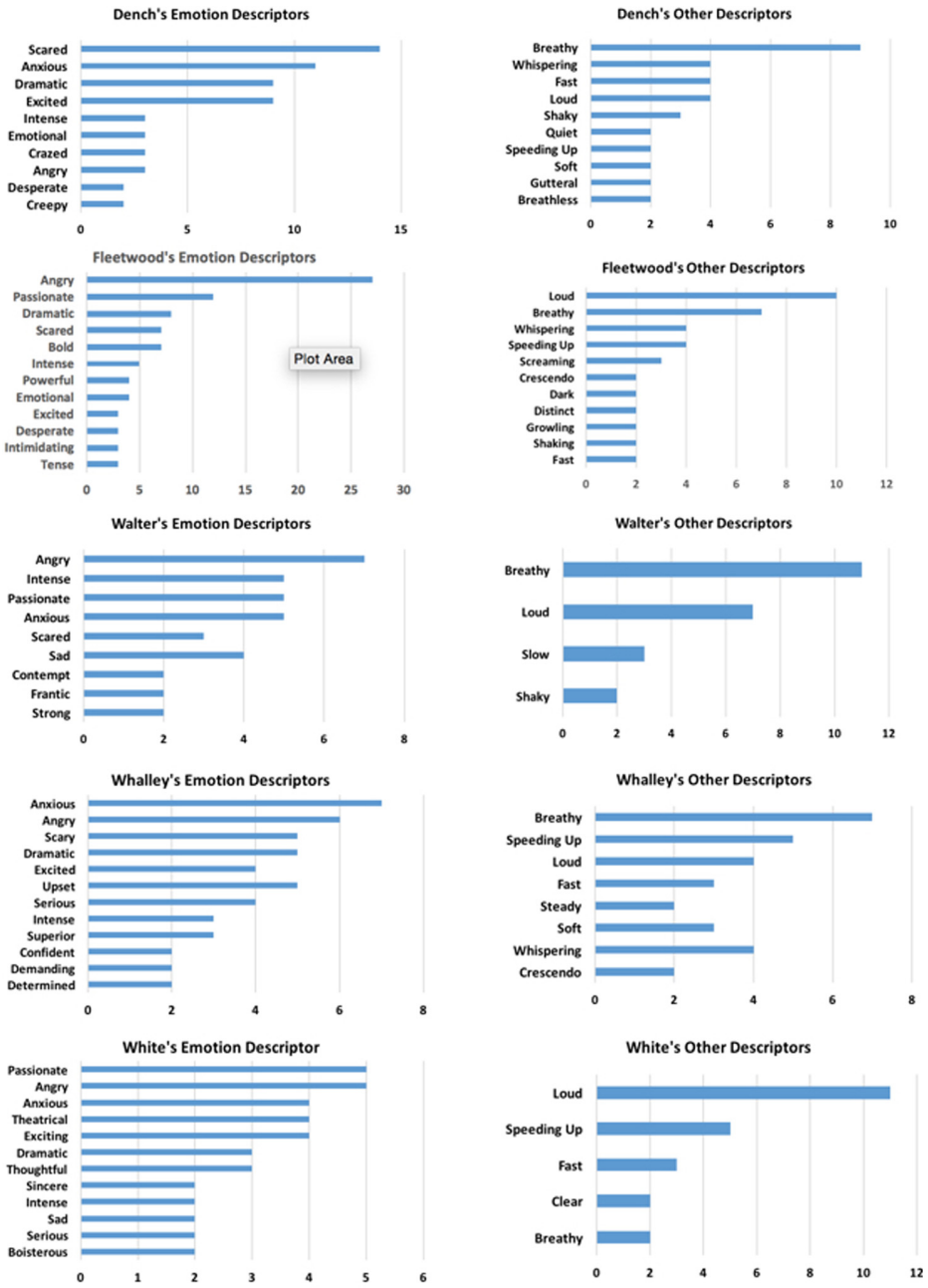

FIG. 3. (Color online) Top Keyword Descriptors for Female Acted Voices. This figure shows the most frequent emotion and non-emotion descriptors given for each female speaker performing the Lady Macbeth soliloquy. The patterns of nuanced, infrequently repeating emotion keywords and concise, frequently repeating prosodic and voice quality descriptors are similar to the descriptor patterns in male acted voices. Whispering, breathiness and resonance were all perceived, but resonance less frequently than the other keywords displayed here. Note again that not all of the descriptors given are listed here, just the most frequently given ones.

listeners consistently perceived them for both males and females, and because these three descriptors (and their close synonyms) made up a disproportionately large proportion of listener feedback. Because these qualities contrast both with each other, and with conversational, or "modal" speech, the modal quality was added to the list of effort levels for analysis and to provide a valid "other" category for training classifiers. Furthermore, high expert listener agreement regarding these qualities reinforced grounding in perception, and our decision to investigate them.

VI. SPECTRAL ANALYSIS OF EFFORT LEVELS

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 Fig. 4; 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 4 attempts to show 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 to this, but it had more high-frequency energy overall; and the female formants (except F1) were higher than the corresponding male formants. If, however,

TABLE I. Allocation of keyword descriptors across keyword classes, for male speakers, in percent. 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 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 II. Allocation of keyword descriptors across keyword classes, for female speakers, in percent. This table shows, for each female speaker, the raw number of keywords given 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, 63% of Judi 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 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%)

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 Fig. 4 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 4 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)

TABLE III. Comparison of the distribution of keyword descriptor types across males and females. 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 again 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	$\mu = 115.2$, $\sigma = 24.3$	$\mu = 82.4$, $\sigma = 13.1$	$\mu = 98.8$, $\sigma = 25.3$

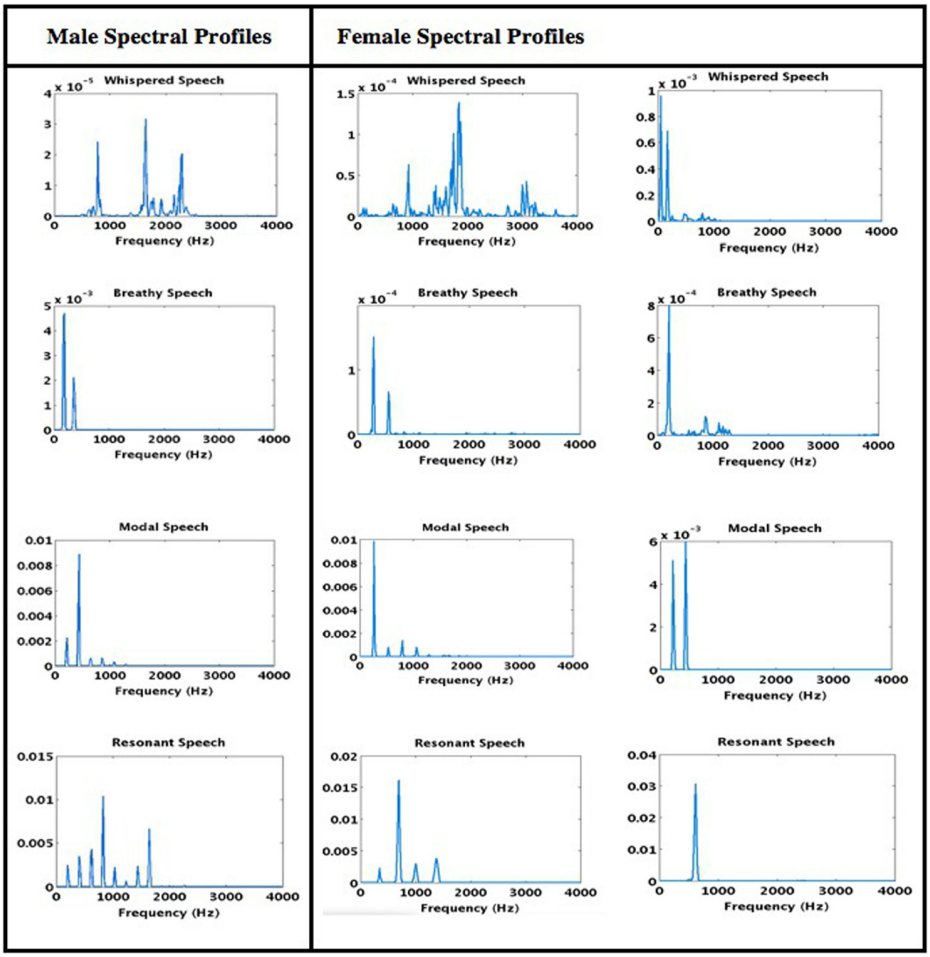

FIG. 4. (Color online) Comparison of Male and Female Spectral Profiles Across Effort Levels. These spectra are typical samples of whispered, breathy, modal, and resonant speech. Female speech had more variation, so we show two variants per condition here. The profiles show many similarities between male 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 second female modal profile (the relationship between the two spectral components is reversed).

0–300 Hz—F0, or speaking pitch; (2) 300–700 Hz—harmonic multiples and F1; (3) 600–900 Hz—higher harmonic multiples and F1; (4) 1000–2000 Hz—F2; and (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 mark resonant voice.

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 Hz—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 is more similar 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.

VII. ANALYSIS OF ACOUSTIC CORRELATES OF EFFORT LEVELS

A. Acoustic features analyzed

The features described in this section were selected for analysis based on prior work, empirical observation of the effort level condition spectral properties, 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, and motivation for inclusion follow.

1. 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.[11]

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.

2. Autocorrelation (AC)

Autocorrelation is the cross-correlation of a signal with itself at different delay times, typically examined between about 3.3 and 16.7 msec, which corresponds to 60–300 Hz, the expected F0 range for adult speech.[9] 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).

3. 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.[22] 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.

4. 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).

5. Spectral tilt (TILT)

The spectral tilt as used here is the slope of a line fitted to the spectrum of a voice source. The best results from this feature are obtained by (1) taking the square of the magnitude of the spectrum to magnify difference, (2) extracting the spectral peaks (as described in the Number of Spectral Peaks feature, which emphasizes the components with the most energy and filters out frequencies which were below 0.5% of the maximum peak), (3) taking the log of the magnitude-squared, peak-filtered signal, and (4) plotting it against the log of the frequency. This method of calculating tilt generally limits the calculations to frequencies below 4500 Hz. Because prior work observed reduced TILT in unvoiced speech,[11,12] it is included as a separator for whispered and breathy voices, compared to modal and resonant conditions, which are clearly voiced.

6. 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.[16,17]

7. Entropy (H-)

Spectral entropy is a measure of disorder in a spectrum.[10] 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[10] 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 Secs. VI and VII, yielding, the following bands of interest for both males and females (and for which entropy was measured): (1) 50–300 Hz, (2) 300–800 Hz, (3) 1000–2000 Hz, (4) 300–4500 Hz, and (5) 4500–8000 Hz. In females the 50–150 Hz, 500–1500 Hz, and 2000–4000 Hz bands were also distinctive, as were the 600–900 Hz, 300–1000 Hz, and 2000–4500 Hz 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 IV summarizes the frequency bands for entropy in males and females.

TABLE IV. Summary of entropy features for males and females. Frequency ranges are given 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.

8. Entropy Ratio (HR-)

Entropy ratios enhance separation between conditions by comparing the character of two frequency bands, as Zhang[10] 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 (50–300 Hz)/(400–600 Hz) provides potential separation across all conditions in males, particularly between the breathy and modal conditions. The ratio (50–600 Hz)/(400–600 Hz) was examined for similar reasons in males; and the ratio (50–300 Hz)/(2000–8000 Hz) provided enhanced separation in females. Table V summarizes the entropy ratio relationships explored in males and females.

9. 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 (50–900 Hz)/(300–900 Hz) 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 (50–300 Hz)/(50–150 Hz) and (50–500 Hz)/(50–1000 Hz) provide similar secondary separation (particularly for the breathy case). Table VI summarizes the power ratios explored in males and females.

10. 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.

B. Statistical measures

In signal detection theory,[62] 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

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

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

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

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.

C. Analytic results

Figures 5 and 6 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

TABLE V. Summary of entropy ratio features for males and females. Frequency ranges are given 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)

TABLE VI. Summary of power ratio features for males and females. Frequency ranges are given 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)

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

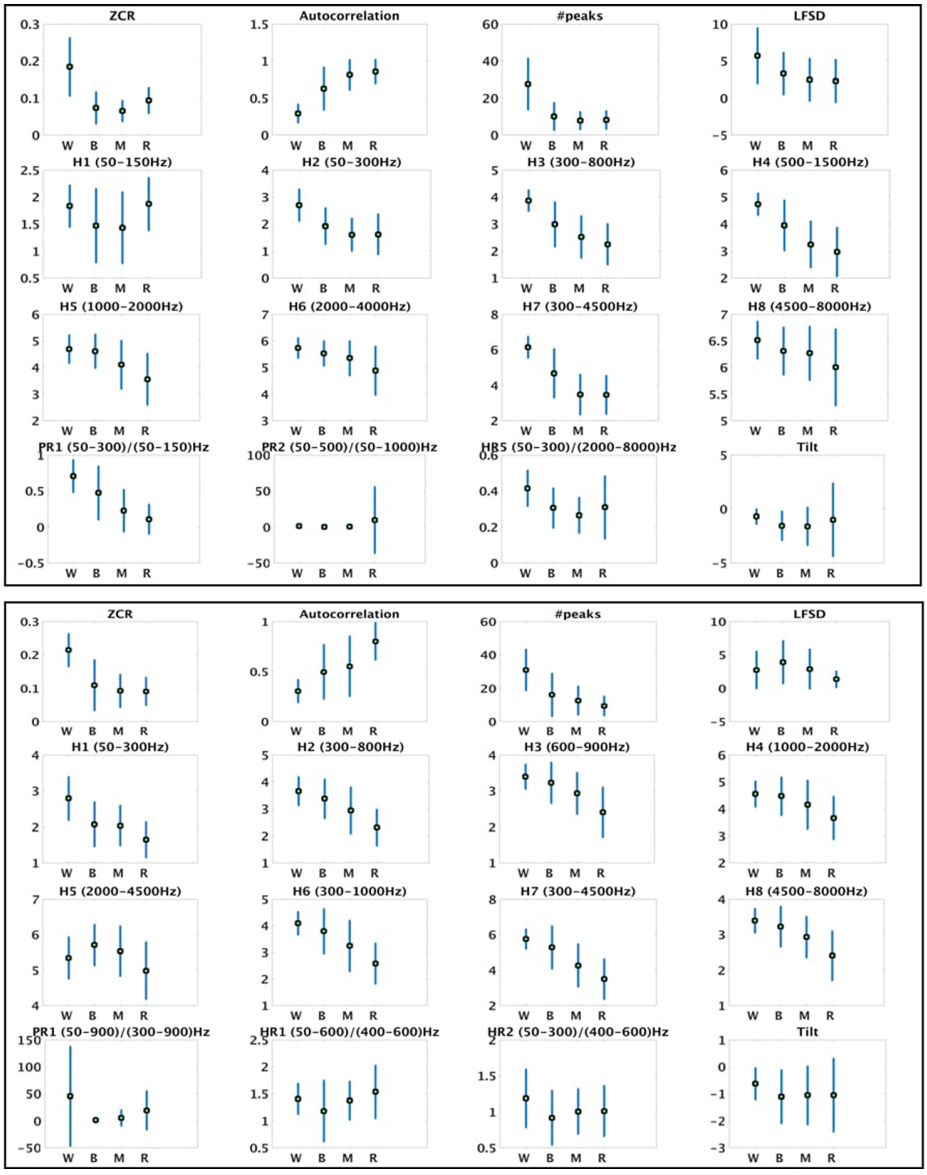

FIG. 5. (Color online) Error Bar Plots for Female (top) and Male (bottom) Acted Voices. These diagrams show the mean and 2 sigma ranges around the means for selected features across the continuum of whispered, breathy, modal, and resonant speech. ***For female voices (top):*** ZCR, autocorrelation, and #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.

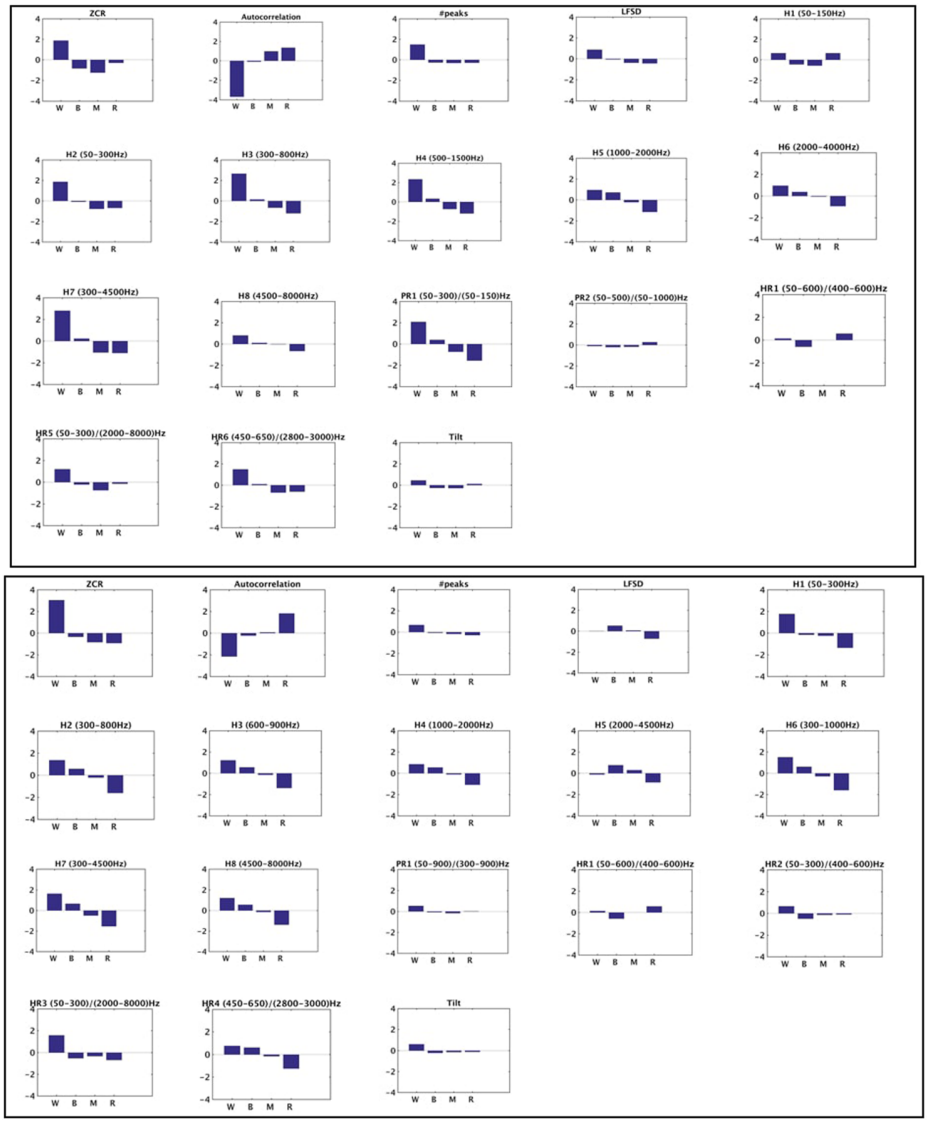

FIG. 6. (Color online) Sensitivity Plots for Female (top) and Male (bottom) Acted Voices. These plots show d_a, the unequal-variance sensitivity, across conditions for a selection of features. A larger magnitude of a bar shows a greater ability of a feature to provide separation for that condition. Many of the features provide strong separation for the whisper case, and smaller separation power for the other features. ***For female voices (top):*** 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. ***For male voices (bottom):*** 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.

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 to measure 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 across any of the conditions.

VIII. CLASSIFICATION EXPERIMENTS

In order to validate the research questions, feature combinations were selected which would best work together on the acted speech corpora to provide maximum separation across conditions. Four-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, five-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, (five segments for females and six segments for males), and segmented where speaking styles were likely to change. 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.

TABLE VII. 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 Figs. 5 and 6 and discussed in Sec. 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
Set 1	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
Set 2	Male	SET1 features, plus LFSD
Set 3	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)
Set 4	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

A. Classification feature sets

The interactive analysis described in detail in Secs. VI and VII 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 four-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 VII, and discussed in detail in Sec. VIII C.

B. 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.

C. 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 VIII summarizes the precision and recall for each condition, along with the global accuracy of the top two male and female four-way decision tree classifiers for the sample, text, and speaker-independent cases.

TABLE VIII. 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 VII), 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
Set 1	0.63/0.63	0.70/0.68	0.67/0.68	0.71/0.74	0.69	0.43/0.58	0.49/0.45	0.48/0.46	0.59/0.59	0.51	*	0.44/0.50	0.69/0.48	*	0.41
Set 2	0.71/0.68	0.76/0.74	0.73/0.74	0.77/0.80	0.76	0.39/0.59	0.49/0.47	0.48/0.47	0.60/0.60	0.52	*	0.56/0.59	0.49/0.47	*	0.41
Set 3	0.70/0.80	0.80/0.74	0.59/0.62	0.65/0.72	0.70	0.60/0.68	0.63/0.13	0.52/0.51	0.56/0.67	0.57	*	0.52/0.59	0.22/0.36	0.76/0.53	0.51
Set 4	0.76/0.75	0.78/0.76	0.64/0.66	0.69/0.73	0.72	0.72/0.71	0.59/0.57	0.54/0.51	0.55/0.65	0.59	*	0.52/0.59	0.22/0.36	0.77/0.53	0.51

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 four-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 ($\chi^2 = 0.42$, 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 ($\chi^2 = 0.99$, 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 ($\chi^2 = 4.5$, df $= 1$, $p = 0.036$).

Figure 7 summarizes the average precision and recall for each condition (whispered, breathy, modal, and resonant) for male acted voices, using feature sets Set 1 and Set 2, applied to four-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 Set 1 was greatest at sample independence ($\mu = 0.7$, $\sigma = 0.2$), and had similar values for both text independence ($\mu = 0.51$, $\sigma = 0.01$) and speaker independence ($\mu = 0.5$, $\sigma = 0.04$). Set 2 followed the same trend, with greatest accuracy at sample independence ($\mu = 0.76$, $\sigma = 0.08$) and again similar levels for text-independence accuracy ($\mu = 0.52$, $\sigma = 0.01$) and speaker independent accuracy ($\mu = 0.5$, $\sigma = 0.01$). 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

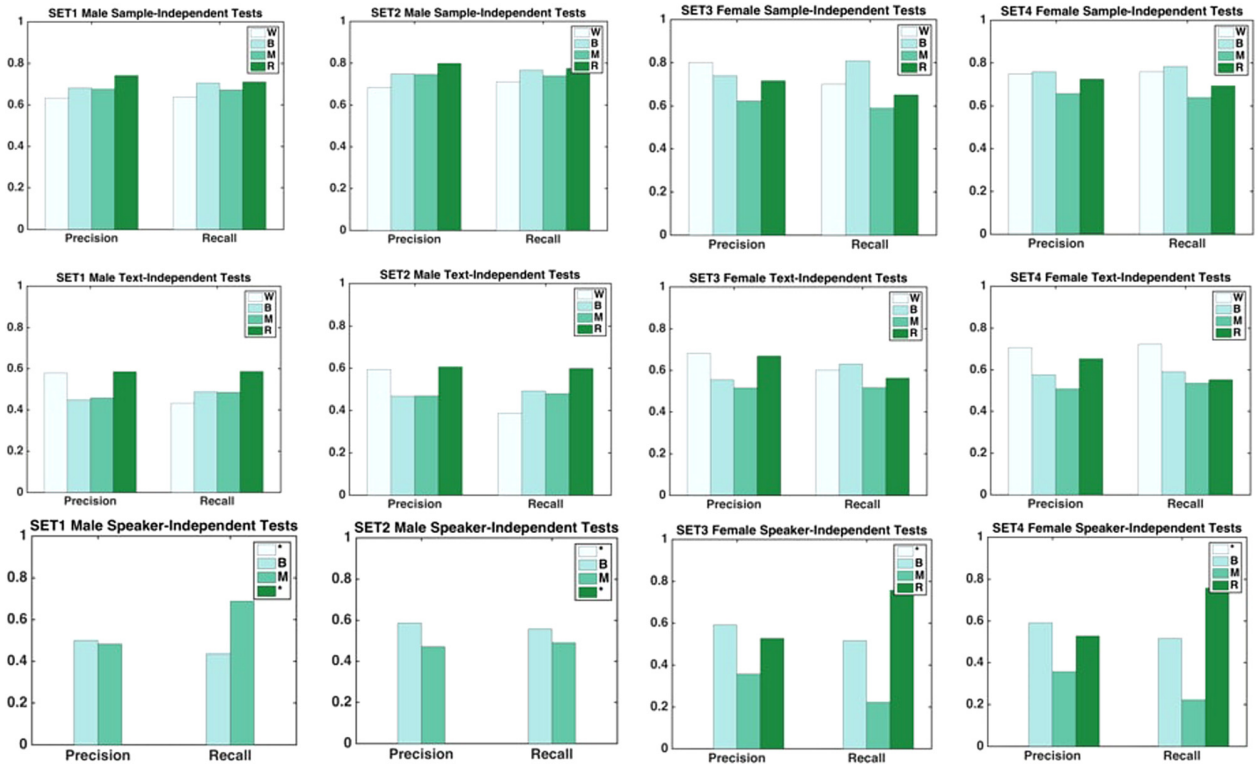

FIG. 7. (Color online) Cross Validation Results for Male and Female Acted Voices. This figure shows the feature set performance for sample-independent (top), text-independent (middle), and speaker-independent (bottom) cross validation. ***For males:*** Sample-independent performance is the strongest, followed by text independence, followed by speaker independence. Set 2 slightly outperformed Set 1 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 independence, followed by speaker independence.

(breathy, modal, and female resonance), however, evaluated at similar accuracy to text independence. See Table VIII for details. Losses in recall rates for speaker independence might be overcome by increasing the size of the dataset and number of speakers.

Figure 7 also provides the recall results for the female speakers and corresponding feature sets (Set 3 and Set 4), again applied to four-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. Set 3 accuracy ($\mu = 0.7/\sigma = 0.024$; $\mu = 0.6/\sigma = 0.08$, and $\mu = 0.4/\sigma = 0.12$ 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 Set 4 accuracy ($\mu = 0.72/\sigma = 0.01$; $\mu = 0.6/\sigma = 0.02$; and $\mu = 0.5/\sigma = 0.085$ for sample, text, and speaker tests). Accuracy is defined as it typically is in a confusion matrix: $Accuracy = (\Sigma\ TP + \Sigma\ TN)/(\Sigma\ total\ population)$, where TP and TN are the number of true positives and true negatives for each condition.

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 Fig. 7), 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 ($p = 0.17$).

IX. DISCUSSION

A. Effort level perception

The perception study yielded the following general findings: (1) On the average, about 60% of descriptors were related to emotion, (2) the remaining descriptors were split almost evenly between prosody and voice quality, (3) emotion descriptors were nuanced, and varied widely, (4) both voice quality and prosodic descriptors focused on a small number of repeating keywords, (5) most of the voice quality descriptors were effort levels, (6) listeners rarely commented on prosodic pitch, and (7) despite similar topic content and speaking style, listeners provided different quantities and distributions of descriptors between males and females. People responded to female speech with significantly more descriptors overall, and significantly more emotion descriptors than male speech. Findings 1–6 directly address RQ1, which asked what listeners hear in expressive speech. Finding 7 addresses RQ2, which asked about differences in perception between male and female speakers. Further studies are required to understand exactly how much of the difference can be attributed to the speaker gender, difference in text, and the demographic characteristics of the listeners themselves.

By beginning with RQ1 and RQ2, asking what people hear, and following up with feature analysis that matches what people hear, we present both a process for grounding voice quality analytics in human perception, and evidence of the elements of vocal expression which people hear clearly in acted voices. Furthermore, the process leverages proven human-computer interaction methodology. The results of following this process, then, can better support future application development, because the findings and resulting analytics are provided in human terms for the development of applications which humans will use (a potentially large impact).

Finally, the study results suggest relationships among prosody, voice quality (especially effort levels), and emotion. What classes of emotion correlate with statistical measures of effort levels and other voice qualities? For example, what classes of emotions correlate with whispering, in different contexts? Examining the quality of the descriptors themselves could yield further insight into these questions. Listeners responded strongly to perceived emotion in the voice, and an analysis[63] of the affect ratings of the descriptors provided in the perception study revealed that the mean affect of the female speech descriptors ($\mu = 4.61$) was significantly lower than that of the male speech descriptors ($\mu = 5.2$) per t-test with independent means and large sample size ($p = 0.0011$). As a grounding reference, the affect rating for the word "sad" is 2.1, and the affect rating for the word "happy" is 8.47.[63] Future work can delve more deeply into these questions and relationships.

Beyond acted speech, we ask the question what do people hear in semi-structured, unscripted speech? What are the similarities and differences? Can analytics developed for acted speech translate to unscripted speech? Future work can also address these questions.

B. Effort level analytics

Analysis of effort levels yielded the following findings: (1) bands of interest for males and females exist which show different characteristics across effort levels, (2) similar feature collections are useful for detecting effort levels in male and female speech, but the band boundaries differ, (3) both perception study and analytic results suggest that effort levels distribute across a continuum from whispered, to breathy, to modal, through resonant speech, (4) both perception study and analytic results suggest that resonance should be

included in analysis as an effort level, (5) breathiness is the most difficult quality to distinguish from among the four effort levels in females, (6) modal speech is the most difficult quality to distinguish from among the four effort levels in males, and (7) the feature collections are applicable at different levels of accuracy in situations where the trained model has not seen the sample, the text, or the speaker. These findings all address RQ3, which asks which features can distinguish across the continuum of effort levels for males and females.

Viewing effort levels as a continuum from whispered through resonant is not a traditional way of viewing these qualities, but both the perception study and the analytic results support this view. First, listeners repeatedly described both male and female acted speech in terms of effort levels. Second, the expert-coded inter-rater agreement data showed that when the expert listeners did not agree, they classified the speech segment in question along adjacent values on the continuum. For example, the raters *never* simultaneously classified a speech segment as whispered and resonant; but sometimes they disagreed by labeling speech segments as "whispered" and "breathy," or "breathy" and "modal." The spectra in Fig. 4 also suggests a continuum of effort levels. The female speech shows, for all the effort levels, both a profile that resembles that of male speech at the same effort level, and a profile with some of the character of the adjacent level. For example, the second female breathy profile has a strong F0, but also a small amount of noise at higher frequencies (like whispering). Also note the breathy and modal female profiles which have 2 components in the same positions on the spectrum, but the power relationship between the components is reversed between the conditions. Finally, the error bar graphs in Fig. 4 show that many of the features (for both males and females), particularly entropy, follow a monotonically increasing or decreasing trend along the continuum of breathy through resonant conditions. This collection of feature trends models a continuum.

The data suggest that the transitions between adjacent conditions have unique characteristics. The transition between breathy to modal, for example, does not function the same way as the transition from whispered to breathy. The transitions also seem to flow at different rates, and appear to be different between males and females. Future work could study these transitions and better quantify what distinguishes adjacent qualities, both in perception and in the acoustic correlates.

The combined results present a strong case for considering effort levels as part of a standard feature set for describing expressive speech. People hear these qualities consistently, reliable correlates grounded to perception exist, and models can be trained to detect the conditions across a range of speakers and text.

Finally, the sensitivity measurements in Fig. 6 suggest that it is reasonable to develop a standardized metric and methodology which can identify the best feature collections for detecting specific expressive conditions from within a pool of possible features, and quantify the efficacy of each collection. This methodology would be extensible to all feature sets which are grounded in human perception.

X. SUMMARY AND CONCLUSIONS

RQ1 and RQ2 asked what untrained listeners would hear and consciously describe with respect to an actor's vocal expression. The perception study addressed these questions, and pointed to a second tier of perception-oriented questions; options for next steps were discussed. Investigating the relationships among the three categories of descriptors, including emotion, prosody, and voice quality, is particularly interesting. The analytic investigation addressed RQ3 and resulted in proposed, validated feature sets for both male and female acted speech. Investigations of all three research questions grounded the results in human perception, and presented evidence for working across a continuum of effort levels, as opposed to treating effort levels as unrelated, isolated conditions. Next steps for investigating this continuum, focusing on the transitions, were also discussed.

Beyond these fundamental results, the work suggests a general methodology for grounding analytics in human perception, which could accelerate future speech-related application development. It is difficult for an application developer to use analytics to build tools for humans unless the analytics produce results which are in human terms. It also presents effort levels as a potential standard feature for expressive voice analysis.

This work has not addressed the question of optimized classification, however. Future work should explore methods which are more efficient and accurate, and investigate a range of acted speech styles. This will require expanded corpora, which will also help address the question of speaker-independent whispered speech. Furthermore, the work has only addressed acted, or scripted voices. What are the characteristics of perceived vocal expression in unscripted voices? Are voice qualities perceived in similar ways between scripted and unscripted speech? Finally, can similar analytic methods be used for scripted and unscripted speech? In the case of effort levels, can the same analytic methods be used? The findings from this study suggest several paths for future investigation.

ACKNOWLEDGMENTS

This work was funded in part by grant R21HS022948 from AHRQ. All findings and opinions are those of the authors, and are not endorsed by AHRQ.

[1]W. Shakespeare, *The tragedy of Hamlet, Prince of Denmark* (Penguin, New York, 1998).
[2]W. Shakespeare, *The tragedy of Macbeth*, edited by B. A. Mowat and P. Werstine (Simon and Schuster, New York, 2013).
[3]L. Gates, *Voice for Performance – Training the Actor's Voice*, 2nd ed. (Limelight Editions, Milwaukee, WI, 2011).
[4]C. Jones, *Make Your Voice Heard, An Actor's Guide to Increased Dramatic Range Through Vocal Training* (Backstage Books, New York, 2005).
[5]R. Smallwood, ed., *Players of Shakespeare 4: Further Essays in Shakespearian Performance by Players with the Royal Shakespeare Company* (Cambridge University Press, New York, 2003).
[6]R. Smallwood, ed., *Players of Shakespeare 5* (Cambridge University Press, New York, 2006).
[7]National Center for Voice and Speech, "Tutorial on voice qualities," http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html (Last viewed April 22, 2016).

[8]M. Pietrowicz, M. Hasegawa-Johnson, and K. Karahalios, "Acoustic correlates for perceived effort levels in expressive speech," in *INTERSPEECH* (2015), pp. 3720–3724.
[9]B. S. Atal, "Generalized short-time power spectra and autocorrelation functions," J. Acoust. Soc. Am. **34**, 1679–1683 (1962).
[10]C. Zhang, "Whisper speech processing: Analysis, modeling, and detection with applications to keyword spotting," Ph.D. dissertation, University of Texas at Dallas, Dallas, TX, 2012.
[11]J. P. Campbell and T. E. Tremain, "Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm," ICASSP **11**, 473–476 (1986).
[12]B. P. Lim, "Computational differences in whispered and non-whispered speech," dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 2010.
[13]K. J. Kallail and F. W. Emanuel, "Formant-frequency differences between isolated whisper and phonated vowel samples produced by adult female subjects," J. Speech Hearing Res. **27**, 245–251 (1984).
[14]M. A. Carlin, B. Y. Smolenski, and S. J. Wendt, "Unsupervised detection of whispered speech in the presence of normal phonation," in *INTERSPEECH* (2006), pp. 685–688.
[15]R. W. Morris, "Enhancement and recognition of whispered speech," Ph.D. dissertation, Georgia Institute of Technology, Atlanta, GA, 2003.
[16]H. Hanson, "Glottal characteristics of female speakers," Ph.D. dissertation, Harvard University, Cambridge, MA, 1995.
[17]R. Wayland and A. Jongman, "Acoustic correlates of breathy and clear vowels: The case of Khmer," J. Phonetics **31**, 181–201 (2003).
[18]D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**(5), 2394–2410 (1991).
[19]C. G. Smith, E. M. Finnegan, and M. P. Karnell, "Resonant voice: Spectral and nasendoscopic analysis," J. Voice **19**(4), 607–622 (2005).
[20]V. Barrichelo-Lindstrom and M. Behlau, "Resonant voice in acting students: Perceptual and acoustic correlates of the trained Y-buzz by Lessac," in *2nd IALP International Symposium*, 2007.
[21]B. Bozkurt, T. Dutoit, B. Doval, and C. D'Alessandro, "A method for glottal formant frequency estimation," in *INTERSPEECH-ICSLP* (2004), pp. 2421–2424.
[22]D. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density," in *INTERSPEECH* (2013), pp. 49–53.
[23]M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *INTERSPEECH* (2007), pp. 1410–1413.
[24]J. Kane and C. Gobi, "Identifying regions of non-modal phonation using features of the wavelet transform," in *INTERSPEECH* (2011), pp. 177–180.
[25]G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184 (1952).
[26]J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995).
[27]P. Ladefoged, *Vowels and Consonants*, 2nd ed. (Blackwell Publishing, Malden, MA, 2005).
[28]P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *International Congress of Phonetic Sciences (ICPhS)*, 2015.
[29]I. Titze, "More about resonant voice: Chasing the formants but staying behind them," J. Singing **59**(5), 413–414 (2003).
[30]B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," J. Phonetics **29**, 365–381 (2001).
[31]J. Hillenbrand and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Hear. Res. **39**, 31–321 (1996).
[32]T. E. Damer, *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments* (Wadsworth, Boston, MA, 2009).
[33]Macmillan Dictionary, "Words used to describe someone's voice," http://www.macmillandictionary.com/us/thesaurus-category/american/words-used-to-describe-someone-s-voice (Last viewed April 22, 2016).
[34]Tumblr Writing Helpers Blog, "55 words to describe someone's voice," http://writinghelpers.tumblr.com/post/41621570418/55-words-to-describe-someones-voice (Last viewed April 22, 2016).
[35]A. Dix, J. Finlay, G. D. Abowd, and R. Beale, *Human-Computer Interaction* (Pearson/Prentice Hall, Upper Saddle River, NJ, 2004).
[36]P. A. Lewis and H. D. Critchley, "Mood-dependent memory," Trends Cognit. Sci. **7**(9), 431–433 (2003).
[37]S. Erk, M. Kiefer, J. O. Grothe, A. P. Wunderlich, M. Spitzer, and H. Walter, "Emotional context modulates subsequent memory effect," Neuroimage **18**, 439–447 (2003).
[38]A. J. Schafer, S. R. Speer, P. Warren, and S. D. White, "Intonational disambiguation in sentence production and comprehension," J. Psycholinguist. Res. **29**(2), 169–182 (2000).
[39]M. Forsell, "Acoustic correlates of perceived emotions in speech," M.S. thesis, School of Media Technology, Royal Institute of Technology, 2007.
[40]A. Mehrabian, *Silent Messages* (Wadsworth, Belmont, CA, 1971).
[41]Y.-L. Shue and M. Iseli, "The role of voice source measures on automatic gender classification," in *ICASSP* (2008), pp. 4493–4496.
[42]E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," ICASSP **1**, 346–348 (1996).
[43]H. Hanson and E. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," J. Acoust. Soc. Am. **106**(2), 1064–1077 (1999).
[44]M. Alsulaiman, Z. Ali, and G. Muhammad, "Gender classification with voice intensity," in *5th European Symposium on Computer Modeling and Simulation*, 2011.
[45]M. Alsulaiman, Z. Ali, and G. Muhammad, "Voice intensity based gender classification by using Simpson's Rule with SVM," in *IWSSIP*, 2012.
[46]P. Nguyen, D. Tran, X. Huang, and D. Sharma, "Automatic classification of speaker characteristics," in *ICCE*, 2010.
[47]M. Gibson, "Hamlet soliloquy performance," https://www.youtube.com/watch?v=Vf2TpWsPvgI (Last viewed April 22, 2016).
[48]D. Jacobi, "Hamlet soliloquy performance," https://www.youtube.com/watch?v=-elDeJaPWGg (Last viewed April 22, 2016).
[49]R. Burton, "Hamlet soliloquy performance," https://www.youtube.com/watch?v=lsrOXAY1arg (Last viewed April 22, 2016).
[50]D. Tennant, "Hamlet soliloquy performance," https://www.youtube.com/watch?v=xYZHb2xo0OI (Last viewed April 22, 2016).
[51]K. Branagh, "Hamlet soliloquy performance," https://www.youtube.com/watch?v=SjuZq-8PUw0 (Last viewed April 22, 2016).
[52]J. Dench, "Lady Macbeth soliloquy performance," https://www.youtube.com/watch?v=2xHIngY6Bgk&list=PLJ7XsNHDDVSkFLkEohiEtqO3g184DXft7&index=5 (Last viewed April 22, 2016).
[53]H. Walter, "Lady Macbeth soliloquy performance," https://www.youtube.com/watch?v=hREqqNr9AyI (Last viewed April 22, 2016).
[54]J. Whalley, "Lady Macbeth soliloquy performance," https://www.youtube.com/watch?v=TiBBLlb_pZ0 (Last viewed April 22, 2016).
[55]K. Fleetwood, "Lady Macbeth soliloquy performance," https://www.youtube.com/watch?v=RM8QQuz5BP4 (Last viewed April 22, 2016).
[56]A. J. White, "Lady Macbeth soliloquy performance," https://www.youtube.com/watch?v=ft2Lthl9q5Y (Last viewed April 22, 2016).
[57]The Penn Phonetics Lab, "Forced aligner," https://www.ling.upenn.edu/phonetics/old_website_2015/p2fa/ (Last viewed April 22, 2016).
[58]FAVE-align, "an online interface for the Penn Forced Aligner," http://fave.ling.upenn.edu/FAAValign.html (Last viewed April 22, 2016).
[59]M. Pietrowicz, D. Chopra, A. Sadeghi, P. Chandra, B. P. Bailey, and K. Karahalios, "CrowdBand: An Automated Crowdsourcing Sound Composition System," in *HCOMP*, 2013.
[60]Online thesaurus of English, http://www.thesaurus.com (Last viewed April 22, 2016).
[61]A. Ortony and T. J. Turner, "What's basic about basic emotions?," Psychol. Rev. **97**(3), 315–331 (1990).
[62]H. Pashler and J. Wixted, eds., *Stevens' Handbook of Experimental Psychology, Methodology in Experimental Psychology*, 3rd ed. (John Wiley and Sons, New York, 2002), Vol. 4.
[63]A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 12,915 English lemmas," Behav. Res. Meth. **45**, 1191–1207 (2013).