Acoustic Correlates for Perceived Effort Levels in Expressive Speech

Mary Pietrowicz [1], *Mark Hasegawa-Johnson* [2], *Karrie Karahalios*[1]

[1] University of Illinois, Department of Computer Science
[2] University of Illinois, Department of Electrical and Computer Engineering
mpietro2@illinois.edu, jhasegaw@illinois.edu, kkarahal@illinois.edu

Abstract

Actors and other vocal performers vary their speech across the continuum of vocal effort to express ideas, emphasize thoughts, communicate emotions, and create drama. They are experts at vocal expression. To analyze this range of expression across effort levels, we curated a corpus of professional actors' Hamlet soliloquy performances and present an acoustic feature set and classification model suitable for tracking actors' expressive speech from extreme to extreme – from whispered, to breathy, through modal, to resonant speech.

Index Terms: voice quality, effort levels, acoustic correlates, expressive speech, whispered voice, breathy voice, projected voice, resonant voice, paralingual.

1. Introduction

Each actor who performs the famous Hamlet soliloquy from Act III Scene I of Shakespeare's play speaks the same words, which begin, "To be, or not to be…" Each actor speaks exactly the same words, but communicates something different, because *the expressive qualities of each speaker are different*. A quick survey of just five professional actors performing the soliloquy (David Tennant, Kenneth Branagh, Derek Jacobi, Mel Gibson, and Richard Burton [16-20]) revealed striking differences in speaking rate, pitch variance, use of silence, phrase groupings, accents, and vocal quality (particularly in the use of whispers, breathiness, vocal fry, and projected speech). Expressive difference characterizes this kind of vocal performance, both within speaker and across speakers (Shakespearian actors are experts at vocal expression). To explore these differences, we conducted an exploratory study to find out what features the casual listener would perceive in expressive speech. We presented Mechanical Turk workers with a random sampling of Hamlet soliloquy audio clips from these speakers, and asked workers to provide one or more keywords describing the expression in the voice (not the word content). Listeners could provide any keywords which came to mind, but most often provided keywords describing emotion, changes in loudness, and various aspects of vocal effort (e.g., whispering, breathiness, and "ringing," or resonant/projected speech). Because of the sensitivity of listeners to these features, particularly the elements of vocal effort, we asked the following research question: **What acoustic features can distinguish each of four levels of vocal effort (whispering, breathiness, modal speech, and resonant/projected speech) in male actors' expressive speech?**

To explore this question, we held the spoken text constant by curating a corpus of performances of the Act III Scene I Hamlet soliloquy, and studied the corpus across the continuum of effort, from whispering, through breathiness and modal speech, to resonant speech. The audio content came from movies and stage performances, which were recorded in varying environments. We excluded sections from the corpus which had significant sonic interference, such as music, high levels of background noise, sound effects, competing speech, and excessively reverberant environments (where echoes or "slap-back" was apparent, or effects such as a loud speaker in a small cave would create).

Previous work has found that, normalized autocorrelation in the F0 range produces a strong maximum at the fundamental period, and spikes at regular intervals, which are both lacking in whispered speech [2]. Overall, 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 significant voicing-dependent differences; while the use of MFCC features, standard for speech processing, yields inferior results when compared with spectral entropy and spectral tilt [25]. 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 [5]. Reduced spectral tilt is a frequent observation in unvoiced speech [15,25], along with shifts in formant frequencies [13], differences in the ratios of high-frequency to low-frequency energy (which captures tilt) [6,12,21,24], 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 nonvoiced speech than for voiced speech [6,21].

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 [10,23]. 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 [23]. They also observed that the H1-H2 difference between breathy and modal voice within speaker was measurable, but the H1-H2 value for one speaker's breathiness could be the value for another speaker's modal speech. This finding raises questions about the un-normalized application of these kinds of features across a set of voices with significant variance across speakers, as we have across our Hamlet actors. 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) could distinguish breathy and modal voices [7].

Studies comparing resonant with modal voice production suggest that speakers produce a resonant tone via "first formant alignment," which produces a higher harmonic content in the portion of the spectra corresponding to the first formant (4-7dB stronger). Also, resonant voice had stronger harmonics in the 2.0-3.5 kHz band [22]. Actors, especially, train to produce resonant voice. Researchers studying the difference between actors' non-resonant and resonant voices (produced via the Lessac Y-Buzz technique) found a reduction in the difference between the first formant and second harmonic in men [3].

Research which examines differences in phonation types (breathy/modal/pressed) used features characterizing glottal function [4,9], and found low-frequency spectral density (LFSD) to reflect the differences in open quotient and the corresponding increase in low frequency energy in breathy voices [9]. Amplitude quotient (AQ) and normalized amplitude quotient (NAQ) of the glottal pulse were superior separators, along with harmonic difference H1-H2 [1,9,14], closing quotient, quasi open quotient, and brightness [1].

Previous studies of voice quality are often motivated by considerations of speech pathology [8,11], phonology [9], or speaker identifiability in speech synthesis [10]; and therefore, no previous study considers a continuum of expressive speech that includes within-speaker and across-speaker distinctions among whispered, breathy, and resonant voice qualities. There are significant, practical difficulties in the analysis of real-world expressive, acted speech, which we address in this paper. First, acted speech is characterized by greater than usual difference both within speaker and across speakers. In comparison to spontaneous or read speech, it has exaggerated extremes of pitch, volume, speaking rate, phoneme duration, and vocal quality. Second, production of quality acted speech requires expertise. Existing corpora do not contain representative samples of expressive, acted speech; and it is not reasonable to create a suitable corpus from untrained voices. Third, when suitable samples are found, people with expertise to hear the expressive differences must code it. Our primary contribution is a feature set and classifier suitable for parsing the continuum of effort levels, from whispered speech to resonant speech, which will function across widely-varying speaking styles. In addition, the result is robust enough to function across voice recordings from varying environments.

2. The Hamlet Corpus

We selected expert performances of the Hamlet soliloquy (Act III, Scene I) by Mel Gibson, Derek Jacobi, Richard Burton, David Tennant, and Kenneth Branagh [16-20]. These speakers were selected for their collective difference in expressive style across speaker and for their professional acting and speaking ability. This small number of speakers provides a large range of expression for analysis. For example, in just the first sentence of the soliloquy, Jacobi's voice ranges from breathy to resonant, soft to loud, and ranges in pitch over almost an octave. Tennant's voice is breathy, soft, and gently inflected, while Burton's voice is modal and flat in comparison. Branagh's voice is all angst, and ranges from breathy-modal in the first phrase, to a chilling whisper in the second phrase. Gibson's speech is rapid, his pauses, minimal, and tone, almost businesslike. Each speaker's pitch and volume variation, accent points, and phrasing are different, and that is just a high-level observation over just the first sentence. This range is characteristic of actors' speech. One expert hand-coded each performance in our corpus to the syllable level with the 4 conditions (whispered, breathy, modal, and resonant). By our definition, modal speech had an average conversational quality, whispered speech had no voicing, breathy speech had weak voicing with an airy quality, and resonant voice had a ringing, or projected quality in comparison to modal voice. To validate the coding, we randomly selected 20 samples from each condition across all the speakers, and asked a second expert to classify the samples as whispered, breathy, modal, or resonant speech. Before running the experiment, we gave our experts the definition of each type of speech, and demonstrated it with example recordings. We reached 95%, 85%, 65%, and 90% classification agreement over the whispered, breathy, modal, and resonant conditions, respectively, with 85% agreement overall, and a Cohen's kappa of 0.8.

To prepare the corpus for analysis, we first downsampled it to 16 kHz, normalized the signal within each speaker, and excluded portions with music, excessive noise, sound effects, interfering voices or background, or significant reverb (with noticeable echo, slapback, or delay). Next, we extracted all of the vowel sounds which were at least 60msec long. A forced aligner was helpful in this process, but we overrode it manually when it made errors. We used a window length of 60msec on all features except LFSD which required a smaller 10msec frame [9], applied a Hamming window to each slice, and advanced the window by 15msec over the range of each vowel sample. We experimented with a variety of different window lengths, and found that a window length of 50-65 msec worked best with our feature set (except LFSD). At the end of the process, the Hamlet corpus had a total of 83 whispered, 329 breathy, 353 modal, and 276 resonant speech samples across the range of vowels in the English language. The whispered speech had the fewest samples simply because the actors used it sparingly.

3. Analysis of Features

We selected features for investigation based on the literature and our empirical observations of the characteristics of each condition across the speakers (see Figure 1 for representative spectra). Whispered speech is noiselike, aperiodic, has high-frequency components, and lacks a strong component where F0 would be. Breathy speech has a strong F0, a small number of significant harmonics (often 1-4 spikes on F0 harmonics), and in some cases, some low-energy harmonics at higher frequencies. In general, breathy speech is periodic and lacks significant high-frequency energy. Modal speech is periodic with many multiples of F0, with the presence of formants. Note that its strongest components are below about 500-600 Hz. Resonant speech is still periodic, but it differs from modal speech in that it has proportionally more high-frequency energy, has its strongest components above 500 Hz, has more overall energy, and shows stronger overall formant representation. We observed consistent differences across conditions in the frequency bands 0-300, 300-700, 600-1000, 1000-2000, and 2000-4500 Hz. In general, the 0-300 Hz band showed differences in F0 and glottal formant representation across conditions, the 0-900 Hz bands contained the most significant differences in the amplitude and periodic excitation of F0 and its nearby pitch harmonics, and the bands above

1000 Hz contained difference in formants, high-frequency harmonics, and high-frequency noise. We selected features for exploration because they had the documented ability to provide separation between at least two of the conditions, could leverage the characteristic differences across conditions that we observed empirically, would be robust to uncontrolled recording environments, would work across a wide variance of speaker expressivity, and would introduce the least amount of confusion for classifying the 4 distinct conditions together. We also preferred features that could be computationally efficient enough to use in real time application development.

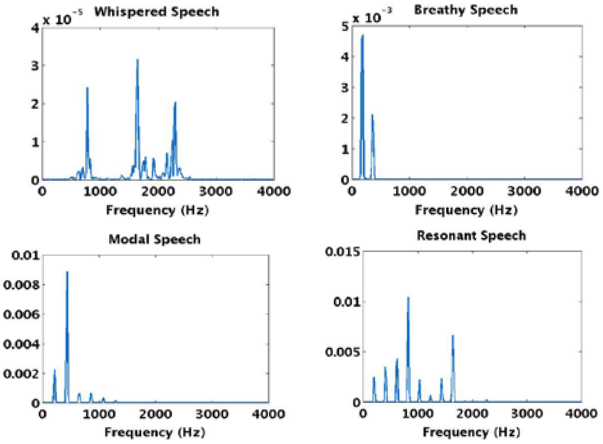

Figure 1: Representative spectra from each effort level type. Top left: Whispered speech is noise-like, flat, and distributed with high-frequency energy. Top right: Breathy speech is focused around F0 with no upper harmonics. Bottom left: Modal speech has F0 with harmonic multiples, and high energy around F1. Bottom right: Resonant speech has relatively weak F0, and high energy around F1 and F2.

Our candidate features included 1) Zero Crossing Rate (ZCR), 2) Number of Significant Spectral Peaks (PK), 3) Normalized Autocorrelation Maximum in the pitch period range 5-60msec (AC), 4) Log Low-frequency Spectral Density (LFSD), 5) Entropy 50-300 Hz (H1), 6) Entropy 300-700 Hz (H2), 7) Entropy 600-950 Hz (H3), 8) Entropy 1000-2000 Hz (H4), 9) Entropy 2000-4500 Hz (H5), 10) Entropy 300-1000 Hz (H6), 11) Entropy 300-4500 Hz (H7), 12) Entropy 4500-8000 Hz (H8), 13) Normalized Power Ratio 50-900/50-600 Hz (PR1), 14) Entropy Ratio 50-300/400-600 Hz (HR1), 15) Entropy Ratio 450-650/2800-3000 (HR2), 16) Spectral Tilt (TILT), 17) Difference between the First Two Harmonics (H1-H2). We selected the first three features to detect voicing, the frequency bands to align with observed spectral differences across conditions, entropy for its robustness across a widely-varying set of speaking styles and ability to measure the degree to which a sound is noise-like or tone-like, LFSD for its potential to reflect glottal and open quotient differences across conditions, and power and entropy ratios to magnify spectral differences across conditions.

We calculated entropy (H) [25], entropy ratio (HR) [25], LFSD [9] as described in the literature, and additionally took the log of LFSD to enhance separation. Normalized power ratio (PR) is similar to spectral density, except that we use the magnitude squared of the spectral components, and normalize each power value by the sum of the power over the spectral range. Then, we take the ratio of the normalized power in the high band of interest to the normalized power in the low band of interest. For peak detection, we first zeroed out frequencies which were less than 0.5% of the maximum peak, clustered groups of adjacent frequency spikes together, and extracted the maximum frequency spike from each cluster.

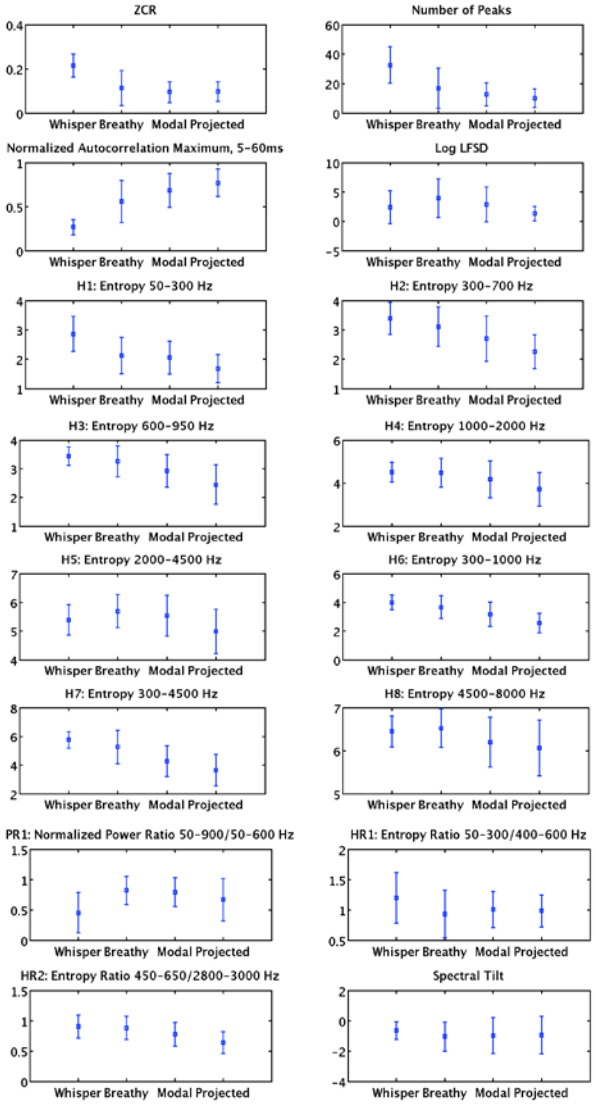

Figure 2: Analysis of features across the continuum of whisper, breathy, modal, and resonant speech. The square markers show the mean at each condition, and the error bars show the 2-sigma distribution around the mean.

Figure 2 summarizes the characteristics of each feature across conditions. ZCR, PK, AC, and H1 provided the best separation between whispered speech and the rest. The difference between breathy and modal speech was the most difficult distinction to draw with our feature set, but was provided to a certain extent by normalized autocorrelation and by H7. We expected LFSD to provide strong breathy-modal separation, but it did not perform as well as the entropy features, even when we took the log of LFSD to amplify the differences across conditions. It is interesting to observe that the best voiced-unvoiced features introduced confusion for the breathy-modal distinction. Finally, the best modal-resonant separators were AC, H2, H3, H4, H5, and H6; and LFSD provided secondary separation.

The spectral tilt and H1-H2 features did not provide the separation we expected for the Hamlet corpus. Tilt showed only weak separation between voiced and unvoiced speech, and did not distinguish across the other conditions. The

harmonic difference (H1-H2) provided limited differentiation between modal and resonant conditions, and did not distinguish well across the remaining conditions. Tilt features, in general, degraded classifier performance.

4. Methods & Experiments

To address our research question, we trained a 4-way decision tree classifier on the Hamlet corpus, pruned the result to guard against overfitting and tune performance, and used 4-fold cross-validation to validate our approach. Figure 3 shows precision and recall for an assortment of 4-way classifiers over the best performing feature subsets and single features. By precision, we mean the fraction of retrieved (recognized) instances that were relevant (correctly recognized); and by recall, we mean the fraction of relevant (available) cases that were retrieved (recognized). The best 4-way classifier had a 76% overall accuracy; while binary classifiers which used this best-performing feature set had 83-98% accuracy (Table 1).

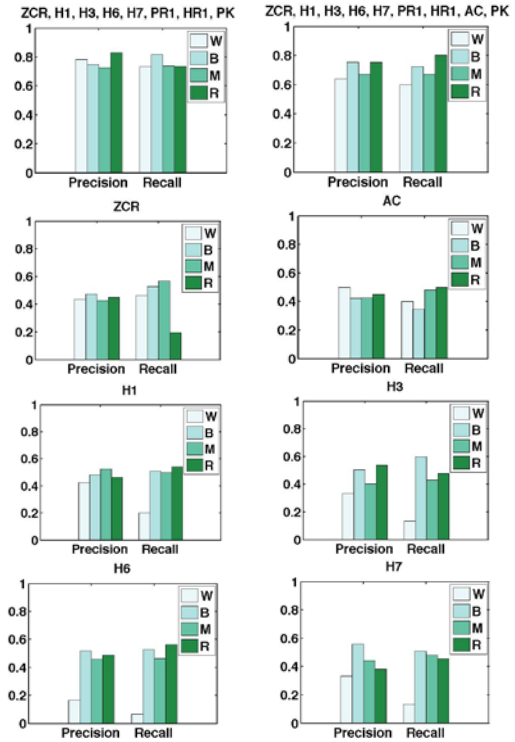

Figure 3: Performance of 4-Way Decision Tree Classifiers, for single features and two feature combinations. (W=whisper, B=Breathy, M=Modal, R=Resonant).

Condition	p/r	Other	p/r	Accuracy
Whispered	99/73	***Non-whispered***	98/99	98
Breathy	99/67	***Non-breathy***	87/99	90
Modal	74/74	***Non-modal***	87/87	83
Resonant	71/76	***Non-resonant***	91/89	86

Table 2: Performance of Binary Decision Tree Classifiers for feature set ZCR, PK, PR1, HR1, H1, H3, H6, H7. Precision (p), recall (r), and accuracy are given in percent.

5. Discussion and Conclusions

Our research question asked what acoustic features could distinguish across four levels of vocal effort, from whispering, to breathiness, to modal speech, to resonant speech. We explored a range of features which the literature suggested could provide separation, and added to this list our empirical observations of representative spectra across each condition. We observed differences across conditions in the bands 0-300, 300-700, 600- 950, 1000-2000, and 2000-4500 Hz within the Hamlet corpus, and found that each condition had one or more characteristic spectral "fingerprints". Entropy measurements within each of these bands captured the important spectral relationships, were robust to varied recording conditions, and functioned well across speakers and the range of speaker expression. Some of the entropy features, such as H3, provided good general separation, but we found that using a collection of entropy features together provided the best results. Entropy ratios, too, could be used to highlight differences between two bands. Normalized power ratios were also useful, but in general did not separate the conditions as well as entropy or entropy ratio features.

The spectral tilt features (TILT and H1-H2) and LFSD did not provide the expected separation across conditions over the Hamlet corpus; and we suspect that this result reflects the character of expressive voices, which vary greatly both within and across speaker. Expressive speech has more variance in it than a corpus in which a similar population (e.g., students in a lab) reads text or speaks phonemes. Variance, difference, aperiodicity, and extremes characterize expressive speech, and our human perceptual system is wired to perceive it. For these reasons, we seek to study and emphasize these differences in our future work, not normalize them away. Exploring models of expressive speech that include the excitation and highlight differences may be helpful.

While our results were generally positive, we believe we could improve them by further exploring and understanding perception of the different conditions. We suspect that the perceptual label "breathy," for example, is a large umbrella over a collection of breathy subtypes which have different acoustic fingerprints. It is also possible that the human perception of "breathy" depends on the context, so that the same sound may be perceived as "breathy" when surrounded by one kind of speech and "modal" when surrounded by another kind of context, or when taken out of context. Furthermore, the distinctions do lie on a continuum, and the distance between a "strong breathy" and a "soft modal" may be acoustically and perceptually small and variant.

We also think that studying female actors' voices and comparing male and female voices would be useful, as would extending our scope into other phonation types (e.g., pressed, yelling) and other frequently-perceived features of vocal expression such as emotion and vocal emphasis.

Finally, we believe that the study of acted expressive speech has the potential to inform a range of applications, particularly those in the areas of speech therapy, vocal performance coaching, language learning, medical diagnostics, multimodal art, and automated voice agents.

6. Acknowledgements

Parts of this research were supported by QNRF grant NPRP 7-776-1-140. All findings and opinions are those of the authors, and are not endorsed by sponsors of the research.

7. References

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

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

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

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

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

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

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

[8] Bruce R. Gerratt and Jody Kreiman, “Toward a taxonomy of nonmodal phonation,” Journal of Phonetics, 29:365-381, 2001.

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

[10] H. Hanson, “Glottal characteristics of female speakers,” PhD Dissertation, Harvard University, 1995.

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

[12] T. Itoh, K. Takida, and F. Itakura, “Analysis and recognition of whispered speech,” Speech Communications, 45: 139-152, 2005.

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

[14] J. Kane and C. Gobi, “Identifying regions of non-modal phonation using features of the wavelet transform,” in Proc. Interspeech, Florence, Italy, Aug. 2011, pp. 177-180.

[15] Boon Pang Lim. “Computational Differences in Whispered and Non-Whispered Speech,” Dissertation, University of Ilinois at Urbana-Champaign, 2010.

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

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

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

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

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

[21] R.W. Morris, “Enhancement and recognition of whispered speech,” PhD Dissertation, Georgia Institute of Technology 2003.

[22] Cara G. Smith, Eileen M. Finnegan, and Michael P. Karnell, “Resonant Voice: Spectral and Nasendoscopic Analysis,” Journal of Voice, 19(4): 607-622, 2005.

[23] Ratree Wayland and Allard Jongman, “Acoustic correlates of breathy and clear vowels: the case of Khmer,” Journal of Phonetics, 31:181-201, 2003.

[24] D.S. Shete, and S.B. Patil, “Zero crossing rate and Energy of the Speech Signal of Devanagari Script,” IOSR-JVSP 4(1): 1-5, 2014.

[25] Chi Zhang. “Whisper Speech Processing: Analysis, Modeling, and Detection with Applications to Keyword Spotting,” PhD Dissertation, University of Texas at Dallas, 2012.