Acoustic Correlates for Perceived Effort Levels in Expressive Speech

...And Beyond

M. Pietrowicz
10/12/2015
Hamlet Act III Scene I

- A maze of 278 twisty little words of text, *all the same*
  
  “To be or not to be. That is the question…”

- A maze of 278 twisty little words, *all expressively different*

- They speak the same text, but each speaker communicates something different!
Putting the puzzle together...

• Where, exactly, is the expression?
• Derek Jacobi again: 
• Prosody
  – Louder and softer
  – Higher and lower
  – Faster and slower
  – Longer and shorter
• Vocal Quality
  – Resonant, whispering, breathy
  – Many others possible
What cues do people perceive in vocal expression?

• To find out, we asked Mechanical Turk workers to provide *keywords* describing *vocal expression* they heard in audio clips from actors playing *Hamlet*.

• The most popular kinds of keywords?
  – Perceived loudness, e.g., “soft, quiet, loud”
  – Perceived emotion, e.g., “angry, sad”
  – Perceived vocal effort, e.g., “whisper, breathy, ringing”
  – Perceived emphasis, e.g., “emphatic, strong, loud”
What to Explore First, and Why

• The continuum of vocal effort
  Silent ➔ Whisper ➔ Breathy ➔ Modal ➔ Resonant ➔ Yelling

• Why vocal effort?
  – Listeners are sensitive to it!
  – It’s a *primitive* feature of vocal quality for expressive speech

• **RQ:** *What acoustic features can distinguish each of four levels of vocal effort (whispering, breathiness, modal speech, and resonant speech) in male actor’s expressive speech?*
Effort Level Distinctions for Perception

• Whispering (NO voicing)
• Breathy (Small amount of voicing, lots of air)
• Modal (Quality of average conversation)
• Resonant (Fully voiced, with a powerful, “ringing” quality. It sounds “rich.”)
Challenges for Analysis of Expressive, Acted Speech

• Acted speech, compared to spontaneous or read speech, has exaggerated extremes.
  – Pitch, volume, speaking rate, phoneme duration, and vocal quality.

• Production of quality acted speech requires expertise.

• Expert listeners also must code it.
The Hamlet Corpus

• Curated expert performances of the Hamlet Soliloquy (Act III Scene I)
• Speakers selected for their professional speaking ability and wide range of expressive style.
• Recordings taken from movies and the stage
• Recording environments uncontrolled
• Excluded sections containing sonic interference
The Hamlet Corpus – Coding and Validation

• 1 expert coded all of the soliloquys, to the syllable level.
• 20 random samples of each condition across all speakers in the corpus coded by another expert listener.
• Inter-rater reliability
  – Whisper: 95%, Breathy: 85%, Modal: 65%, Resonant: 90%
  – Kappa: 0.8
Hamlet Corpus Pre-Processing

- Normalize within each sample
- Downsample from 44 kHz -> 16 kHz
- Exclude sections with excessive sonic interference
- Extract all “long enough” vowel sounds with help of forced alignment tool
- Window size = 60 msec, or 10 msec
- Hop size = 15msec
- Applied Hamming window to each frame
Hamlet Corpus Result

- Utterance count
  - 83 whispered
  - 329 breathy
  - 353 modal
  - 276 resonant

- The actors used whispered speech sparingly
- Some actors used more of one speech style than others
Previous Work

• Motivation
  – Speech pathology, phonology, criminology, speaker ID
  – Not effort levels
  – Very little for acted or expressive speech
• Prior work in analysis of acoustic correlates
  – Whispered/non-whispered
  – Breathy/non-breathy
  – Resonant/non-resonant
  – Phonation type (breathy/modal/pressed)
  – Primarily binary conditions, or related to airflow through glottis
Empirical Observations

- Whispered Speech
- Breathy Speech
- Modal Speech
- Resonant Speech
Empirical Observations – Bands of Interest

- **0-300 Hz**: F0, or speaking pitch.
- **300-700**: Harmonic multiples & F1
- **600-900**: Higher harmonic multiples & F1
- **1000-2000**: Even higher harmonics & F2
- **2000-4500**: High harmonics, higher formants, and noise

Note that you can measure features in these bands, supersets of these bands, and ratios of these bands to differentiate across the 4 conditions.
How to select candidate features to explore?

- Consider the most promising from the literature for each condition.
- Create features which leverage our empirical observations of the spectra.
- Prefer features which are more efficient to compute.
- Prefer a combined feature set that gives best performance as a 4-way classifier.
- Robust to varying recording environments.
- Robust to large ranges of acoustic difference.
Candidate Features - 1

- **Zero crossing rate (ZCR):** rate at which a signal changes from positive to negative.

- **Normalized Autocorrelation (AC) in the F0 range:** the cross-correlation of a signal with itself, that is, the similarity between observations as a function of the time lag. It picks up on periodicity.

\[ y(k) = \frac{\sum_{n=1}^{N-K} x(n)x(n+k)}{\sum_{n=1}^{N-K} x(n)^2} \]

\[ \text{max} [\frac{Fs}{200} \leq k \leq \frac{Fs}{80}] \]
Candidate Features - 2

- **Number of spectral peaks (PK):** number of spikes in the spectrum, above a critical power threshold (another empirical observation).

- **Log Low-Frequency Spectral Density (LFSD):** measure of how much power is in the signal at frequencies below F0, i.e., how much influence the glottal formant has.

- **Entropy 50-300 Hz (H1):** measure of how noisely or tonal-like the sound is.
Candidate Features (Entropyfest) - 3

- Entropy 300-700 Hz (H2)
- Entropy 600-900 Hz (H3)
- Entropy 1000-2000 Hz (H4)
- Entropy 2000-4500 Hz (H5)
- Entropy 300-1000 Hz (H6)
- Entropy 300-4500 Hz (H7)
- Entropy 4500-8000 Hz (H8)
- Entropy measured across bands which differentiate the vocal qualities
Candidate Features - 4

- **Normalized Power Ratio (PR1)**: 50-900/50-600 Hz
- **Entropy Ratio (HR1)**: 50-300 / 400-600 Hz
- **Entropy Ratio (HR2)**: 450-650/2800-3000 Hz
- **Spectral Tilt (TILT)**: Slope of regression line fitted to spectrum.
- **Difference Between First Two Harmonics (H1-H2)**:
Analysis of Selected Features - 1

- ZCR
- Normalized Autocorrelation Maximum, 5-60ms
- H1: Entropy 50-300 Hz
- H2: Entropy 300-700 Hz

Number of Peaks

Log LFSD
Analysis of Selected Features - 2

H3: Entropy 600–950 Hz

H4: Entropy 1000–2000 Hz

H5: Entropy 2000–4500 Hz

H6: Entropy 300–1000 Hz

H7: Entropy 300–4500 Hz

H8: Entropy 4500–8000 Hz

Whisper Breathy  Modal Projected

Whisper Breathy  Modal Projected

Whisper Breathy  Modal Projected

Whisper Breathy  Modal Projected
Analysis of Selected Features - 3

PR1: Normalized Power Ratio 50–900/50–600 Hz

HR1: Entropy Ratio 50–300/400–600 Hz

HR2: Entropy Ratio 450–650/2800–3000 Hz

Spectral Tilt
Feature Summary-1

- **ZCR, PK, AC, and H1** provided best separation between **whispered** (unvoiced) speech and the rest.

- **Breathy-modal boundaries** were difficult to draw with single features
  - AC and H7 provide some separation
  - Cluster of features together separate well

- **Log LFSD** did not provide strong **breathy-modal** separation as expected
Feature Summary-2

- The best modal-resonant feature separators were AC, H2, H3, H4, H5, and H6.

- **Spectral tilt** and **H1-H2** did not provide significant separation in the Hamlet corpus.
  - These features *degraded* classifier performance

  - Possible differences in source recording
  - Extreme expressive difference within and across speakers
Feature Summary-3

• Normalized autocorrelation performed surprisingly well across the board
• Entropy performance thrived over the differences in expressive speech
Experiments: 4-way Classifier

• Experiment 1: 4-way Decision Tree classifier
  – Pruned tree, with 4-fold cross validation
  – 76% Accuracy
Experiments: Single-feature 4-way Classifiers

ZCR

AC

H1

H3

H6

H7

Precision Recall

Precision Recall

Precision Recall

Precision Recall

Precision Recall

Precision Recall
# Experiments: Binary Classifiers

Performance of Binary Decision Tree Classifiers for Feature set ZCR, PK, PR1, HR1, H1, H3, H6, H7. (Best performing 4-way classifier set)

<table>
<thead>
<tr>
<th>Condition</th>
<th>p/r</th>
<th>Other</th>
<th>p/r</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whispered</td>
<td>99/73</td>
<td>Non-whispered</td>
<td>98/99</td>
<td>98</td>
</tr>
<tr>
<td>Breathy</td>
<td>99/67</td>
<td>Non-breathy</td>
<td>87/99</td>
<td>90</td>
</tr>
<tr>
<td>Modal</td>
<td>74/74</td>
<td>Non-modal</td>
<td>87/87</td>
<td>83</td>
</tr>
<tr>
<td>Resonant</td>
<td>71/76</td>
<td>Non-resonant</td>
<td>91/89</td>
<td>86</td>
</tr>
</tbody>
</table>
Conclusions & Contributions

• **RQ**: *What acoustic features can distinguish each of four levels of vocal effort (whispering, breathiness, modal speech, and resonant speech) in male actor’s expressive speech?*
• **Feature set** suitable for the continuum of effort levels (whispered -> resonant)
• **Annotated Hamlet corpus** of expressive acted speech (male speakers)
• **Validation of performance** via classification experiments, which out-performed previous research on whisper and breathy-speech detection
Discussion Points

- Used only male speakers (analysis required for male and female speakers is different!)
- Acted, expressive speech is characterized by extremes and difference, and requires analytic techniques that tolerate and leverage this.
- Curating corpora to study this is time intensive
- We believe that the perceptual labels such as “breathy” have subtypes
- Spoken context affects perception
- Realtime stream processing
Exploring Vocal Expression: The Birds-Eye View

How can we discover how people perceive, interpret, and produce vocal expression?

- Acoustic Correlates
- Semantic Context Correlation
- Linguistic Context Correlation

How can we measure these qualities?

How can we reflect and use The results of vocal expressive Analysis?

Validate!
What’s next?

• What about *acted female speech*?
• *Lady MacBeth* to the rescue...
  – Same speaking style, and similar topic
What about spontaneous speech?

• Desired characteristics of a corpus
  – Conversational, but with long utterances
  – Similar prompts for all speakers enables comparison of expression in the answers
  – Topics elicit vocal expression range typical for spontaneous speech and conversation
  – Speaker metadata available for correlating vocal expression trends with demographics
  – Sectioned (easier to manage)
  – Transcribed (no time to do this myself)
  – Aligned (bonus)
  – Annotations of interest (bonus)
To the Rescue...

- The **HistoryMakers**!
  - It has everything on the list, except alignment and extra annotations.
  - It also has a community of real users who want to use it for purposes other than linguistic study.