DISCOVERING DIMENSIONS OF PERCEIVED VOCAL EXPRESSION IN SEMI-STRUCTURED, UNSCRIPTED ORAL HISTORY ACCOUNTS

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

University of Illinois, Departments of Computer Science[1] and Electrical and Computer Engineering[2]

ABSTRACT

What do people hear in expressive, unprompted speech? And how can their descriptions be transformed into a representative set of dimensions of vocal expression? This paper presents a methodology for collecting user description of vocal expression, transforms the user descriptions into a set of measurable expressive dimensions, and derives a representative feature set and baseline classifiers across these dimensions. The resulting classifiers recognized the top 13 dimensions over an oral history corpus, with a maximum unweighted recall score of 80.5%

Index Terms— Perception, vocal expression, paralingual speech, acoustic correlates, unscripted speech, oral histories.

1. INTRODUCTION

News interviews, legal proceedings, press conferences, oral histories, and even medical consultations have similar structure and purpose. Typically, an interviewer has a set of information-gathering objectives, with pre-planned lines of questioning; and an interviewee speaks spontaneously in response, doing most of the talking, often in storytelling style. Vocal expression and emotion are natural and spontaneous (in contrast to many of the corpora currently used to study emotion and paralingual expression [28]), as is the context (in contrast to contrived scenarios, such as game play designed to provoke emotion). Listeners are drawn into the stories, and hear and describe a speaker's expression naturally as well. We leverage these qualities in oral history interviews and in listeners to study vocal expression, and ask the following questions:

RQ1: What do people hear, expressively speaking, in semi-structured, unprompted, unscripted speech?
RQ2: How can a set of perceived expressive dimensions be discovered from this kind of speech?
RQ3: What baseline feature sets can represent perceived expressive dimensions in this speech?

Work in emotion detection is often limited to acted speech, which has been shown to differ from unprompted speech [10]. Some prior work focuses on categorical detection of one or more basic emotions identified by a given emotion theory [24]. A typical result is a deep exploration into a single emotion (such as anger or depression) [3,8,16,26], focus on variance of an acoustic parameter across a discrete set of emotions [5,12], or exploration into recognition of the list of basic emotions supporting a given theory [20,22]. We found, however, that human listeners provide nuanced description of the emotions they hear in unscripted speech, which go far beyond the 5-7 emotions which are considered basic. Synonym reduction to basic emotions results in loss of information: it nullifies the expressive perceptual capability of the human listener, and also discards information about the relationships which emotion may have to other expressive elements in the voice, such as voice quality (VQ) or prosody. An alternative approach is an n-dimensional representation of emotion along other axes, such as affect, arousal, and dominance. [11] This approach captures a greater range of emotional expression, but typically does not leverage the average human's description of what they hear. Listeners, for example, will say they hear laughter and embarrassment, or that speech is hesitant, sarcastic, and flat. They do not say that an angry speaker has high arousal, low affect, and high dominance. Our approach instead leverages the nuanced description of the human, and preserves the relationships between emotion, prosody, VQ, and nonverbal vocalization, which are embedded in the description. Furthermore, this approach encourages the development of software analytics which are aligned with human perception and are thus better able to support application development.

Work in VQ and nonverbal quality (NQ) tends to examine qualities such as whispering, breathiness, creakiness, resonance, or laughter [1,14,15,17,29,32,33], or focuses on acoustic measures such as jitter and shimmer. A smaller set of research examines specific relationships among emotion, prosody, and VQ [6,13,27].

Our work extends these approaches by first exploring what people hear with respect to vocal expression in oral history interviews, then uses the natural human description to reveal patterns of expressivity across the corpus of speech. The contributions of this work include 1) an efficient methodology for describing and labeling natural speech in everyday language which preserves perceived relationships among emotion, prosody, and VQ, 2) a set of human-perception-aligned, expressive dimensions for

female oral history speakers, and 3) a baseline feature set suitable for describing vocal expression across these dimensions.

2. VETERANS' ORAL HISTORY CORPUS

The library of congress Veterans' Oral History Project [30] provides an open collection of oral history interviews which meet the requirement for semi-structured, unscripted speech on the part of the interviewee. Each interview lasts about 0.5-2.0 hours. While the corpus includes both male and female speakers, this paper focuses on analytics for the females (male voices differ). In addition, the structure of the interviews have similar format and questions across the corpus. Almost all interviews, for example, asked subjects to state their names and basic demographic information at the start of the interview; and almost all interviewees responded to these questions with neutral expression (modal voice quality, neutral emotion, and neutral prosody). Many interviewers asked why and how their subjects joined the military, and about their experiences with basic training. Most also asked subjects to relate one or more stories about their individual personal experiences. These characteristics conveniently allow comparison of vocal expression across answers to similar questions. The corpus is unprompted, sparse in non-neutral expression, natural, and realistic.

Quality of the recordings varied, and most were made in public or home environments with non-professional equipment. Our corpus sub-sample included recent interviews collected during the last 10 years on digital recording equipment, with most subjects representing the Iraq and Afghanistan conflicts. We preferred interviews which contained transcripts, and those which included video recordings, for future multimodal analytic work; and we excluded from analysis speech which contained significant background interference (e.g., other voices, street noise, reverb, or high levels of buzz/hum/hiss). We segmented the interviewees' speech starting with turns and sub-segmented the result into successively smaller phrase groups and phrases. Finally, we identified samples from each speaker which covered the range across each speaker's expression.

3. ANALYSIS OF PERCEIVED EXPRESSION

In order to begin to understand what listeners heard in the vocal expression of unscripted, semi-structured speech, we took a non-prototypical approach and presented representative audio samples covering the range of vocal expression for each of 10 speakers (5 male), and asked US English-speaking Mechanical Turk workers to provide three or more keywords describing the vocal expression in the speakers' voices. The survey included 10-15 representative speech segments (4-45 seconds each) for each speaker, and 10 workers evaluated each clip, for a total of over 1000 Turk listeners and over 3000+ keywords describing the range of vocal expression across the speakers. Table 1 summarizes the results. The unprompted listeners provided keywords describing emotion, VQ, prosodic, and conversational dimensions in the voice. The majority of keywords described emotion (about 55% overall), with nearly equivalent proportions of prosodic and voice quality descriptors (about 17% and 16% respectively). The difference in proportion of keyword types between males and females was not statistically significant.

Keyword Class	Male Talkers	Female Talkers	All Talkers
Voice/Nonverbal Quality	μ=15.02% σ=2.65%	μ=16.81% σ=2.58%	μ=15.92 σ=2.64%
Effort Level	μ=1.96% σ=1.00%	μ=1.70% σ=0.83%	μ=1.83% σ=0.88%
Other Quality	μ=13.04% σ=2.35%	μ=15.12% σ=2.51%	μ=14.08% σ=2.54%
Prosody	μ=16.73% σ=1.87%	μ=17.37% σ=1.27%	μ=17.05 σ=1.54%
Pitch	μ=2.32% σ=1.28%	μ=1.87% σ=1.09%	μ=1.87% σ=1.09%
Loudness	μ=5.23% σ=1.57%	μ=5.38% σ=1.26%	μ=5.31% σ=1.34%
Speaking Rate	μ=8.46% σ=2.25%	μ=8.17% σ=2.08%	μ=8.31% σ=2.05%
Articulation	μ=0.674% σ=0.69%	μ=1.39% σ=0.623%	μ=1.03% σ=0.73%
Emotion	μ=57.06% σ=3.07%	μ=53.62% σ=6.17%	μ=55.34% σ=4.94%
Conversation Style	μ=7.76% σ=1.66%	μ=8.31% σ=3.245%	μ=8.04% σ=2.45%
Other	μ=3.42% σ=1.89%	μ=3.88% σ=0.91%	μ=3.65% σ=1.42%

Table 1: Percentage of keywords in each class for male and female speakers. Keyword proportions are nearly equal between males and females.

Prosodic and voice quality descriptors included a small, repeating set of keywords and their close synonyms. Some of the most common voice quality descriptors included laughter, stuttering, trembling, monotone, and effort levels, such as breathy, creaky, and resonant. The most common prosodic descriptors included speaking rate and loudness, but direct mention of pitch was uncommon. Emotion description, however, was much more nuanced; and listeners used a wide vocabulary to describe what they heard. To discover clusters of speaking styles, potentially *expressive dimensions*, from the given perceptual data, and discover relationships among perceived keyword qualities, we ran latent semantic analysis (LSA) [21] to analyze the distribution of descriptive keywords versus audio clips, and derived 61 concept factors. Typically, the meaning of the factors from an LSA process is not known, but the most positively and negatively-associated keywords can be

interpreted as indicating the meaning of each hidden dimension. Qualitative analysis suggested that it was useful to define strong positive keyword-concept or clip-concept associations as those with projection weights >= 0.085, and negative associations with weights <= -0.085. By projecting the acoustic clips (documents) onto the hidden LSA factors, we also see which clips provided the strongest examples of each LSA concept. Further, the perceived valence and arousal of each keyword [31] were used to derive weighted mean and standard deviation valence and arousal scores for each LSA dimension (see Figure 1). Table 2 describes each of the top thirteen expressive dimensions.

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

Table 2: Description of the top-13 LSA Concept Factors. A short description of the factor is given, followed by the strongest positively and negatively-associated keywords. The top 13 dimensions had multiple keyword concepts with strong weights.

In the next sections, we derive an acoustic feature set capable of describing the characteristics of the LSA factors, and validate the ability of the feature set to capture the expressive information in the oral history speech clips by training models to recognize representative speech in each factor, and running 4-way cross validation to validate whether the resulting models can recognize the most significant 12 concepts.

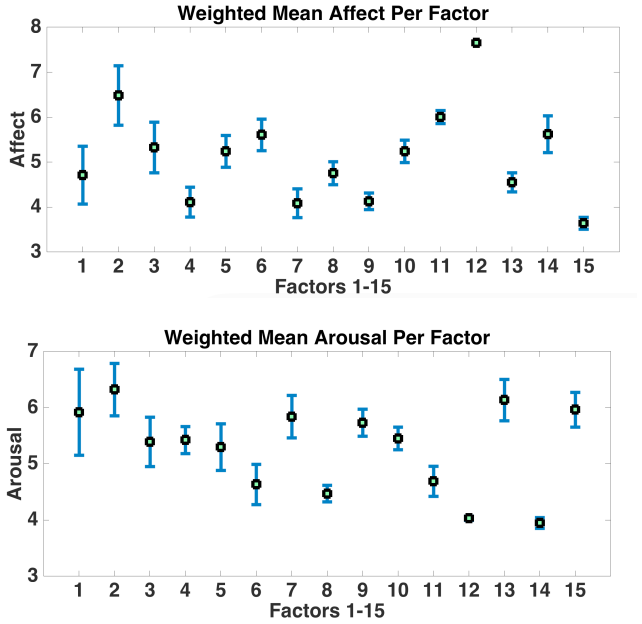

Figure1: Error bar graphs show mean and variance of perceived valence and arousal within LSA concept factors. Both arousal and affect vary from low to high on a scale of 1-9 [31]. Factors overlap within arousal and affect individually, but differentiate when the combination of affect and arousal is considered. Affect and arousal values were linked to keywords, then weighted according to the projection of each keyword onto LSA factor space.

4. ANALYSIS OF FEATURES & EXPERIMENTS

We selected acoustic features for investigation based on the literature, the results of our human perception analysis, and the representation of VQ, prosody, and emotion components in the LSA expressive dimensions. Clips were downsampled to 16Khz, and features were computed based on 60ms frames with a 15msec advance (except LFSD, which required 10msec frames). We mapped the range of emotion keywords onto affect and arousal dimensions, and selected features (a mix of energy, VQ, F0, and spectral features) which have been shown to represent the affect and arousal dimensions [5]. Typical VQ feature sets include jitter and shimmer, but these are disconnected from human description. To address this, we augmented jitter and shimmer with features which are known acoustic correlates for perceived vocal effort levels (breathy, whispered, and projected voice); these features are entropy, entropy ratios, and power ratios across selected frequency bands which differentiate among vocal qualities in female voices [25]. These are also useful for laughter detection. Autocorrelation, low frequency spectral density, and peak count have also been used in the detection of vocal quality,

particularly breathiness [14]. Table 3 lists the acoustic features in each feature category.

Class	Name	Description
Energy	RMS	RMS Energy
	ZCR	Zero Crossing Rate
	RMS_u	RMS Energy / Mean RMS for clip
	PKRate	Energy peak rate
	PKDUR	Energy Peak Duration
F0	F0	Fundamental Frequency
	F0_u	F0 / Mean F0 for Clip
VQ Support	Jitter	Jitter
	Shimmer	Shimmer
	AC	Normalized Autocorrelation Maximum
	LFSD	Log low frequency spectral density
	PkCount	Number of spectral peaks
	H1	Entropy 50-150 Hz
	H2	Entropy 50-300 Hz
	H3	Entropy 300-800 Hz
	H4	Entropy 500-1500 Hz
	H5	Entropy 1000-2000 Hz
	H6	Entropy 2000-4000 Hz
	H7	Entropy 300-4500 Hz
	H8	Entropy 4500-8000 Hz
	PR1	Spectral Power Ratio(50-300)/(50-150)
	PR2	Spectral Power Ratio(50-500)/(500-1000)
	PR3	Spectral Power Ratio(300-800)/(50-300)
	HR1	Entropy Ratio (50-300)/(50-150)
	HR2	Entropy Ratio (50-500)/(500-1000)
	HR3	Entropy Ratio (300-800)/(50-300)
	HR4	Entropy Ratio (50-500)/(50-1500)
	HR5	Entropy Ratio (50-300)/(2000-8000)
	HR6	Entropy Ratio (450-650)/(2800-3000)
Spectral	MFCC	Mel-frequency cepstrum coefficients

Table 3: Acoustic features for perceived vocal expression by category. Each feature and its derivative were tested for correlation with LSA dimension.

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

Table 4: Average unweighted recall (AUR)_in % for each dimension's binary classifier. ***SET1 Content***: RMS, RMS_u, ZCR, F0, F0_u, Jitter, Shimmer, LFSD, H1, H3-7, PR1-2, HR1, HR5, and MFCC1-12. ***SET2 content***: RMS, RMS_u, ZCR, F0, F0_u, Jitter, Shimmer, and MFCC1-12.

Forty binary decision tree classifier sets were trained to classify each clip sample for membership within LSA dimensions. Features and delta-features were included in the classifiers. The majority class in each fold of the training data was randomly undersampled to achieve a balanced training set. As in the Paralingual Challenges for INTERSPEECH 2009-2013, Average Unweighted Recall (AUR) was used as a validation measure. Table 4 shows the ability of two representative feature sets to discern audio clip membership in LSA dimensions 2-13. SET2 is minimal but representative, and includes RMS, ZCR, RMS_u, F0, F0_u, Jitter, Shimmer, and MFCCs. Inclusion of deltas did not significantly change the result. SET1 extended this base to include additional features in support of VQ and NQ (LFSD, H1, H3-H7, PR1-PR2, HR1, and HR5), which improved results in 7 of the 12 perceptual dimensions.

6. DISCUSSION & CONCLUSIONS

This work addressed RQ1 and RQ2 by curating an oral history corpus from the Library of Congress, asked Mechanical Turk workers for descriptive keywords with respect to their perception of vocal expression, and used LSA to derive a set of vocal expression dimensions from listener perception of the clips. The result was 61 expressive dimensions, and we analyzed the strongest 12 to address RQ3. We developed a purposely simple baseline feature set, cross-validated it, and demonstrated improved performance by including VQ support for vocal effort levels.

Future work can improve performance by augmenting the baseline results with additional elements which specifically address strongly-perceived VQ, NQ, and emotion, such as laughter, sarcasm, creakiness, mumbling, roughness, filler, and distribution of silence. A closer look at the strongest and weakest dimensions supports this. Factors 2 and 3 were both marked by laughter, but factor 3 was sarcastic and nervous, while factor 2 was sincere. The more complex factor 3 was recognized about 18% less often than factor 2. Strongly-recognized factors also tended to have clips which were representative of the factor for the duration of the clip (as in Factors 2 and 9). Factor 7, however had bursts of angry speech embedded in a more modal background; but strong anger has a high impact on the human listener, and listeners will perceive and report it. Reflecting the presence of a quality embedded in a clip, or adjusting weights to reflect impact on human perception will also help improve results. Next, the Turk survey methods were simple and exploratory in this study. Improving them to allow marking of localized perception within a clip (such as the angry bursts) will improve understanding about perception, provide better information to LSA processes, and improve classification performance in the future. Finally, LSA relationship among keyword classes can be explored

7. ACKNOWLEDGEMENTS

Parts of this research were supported by AHRQ grant 1-483711-392030-191100. All findings and opinions are those of the authors, and are not endorsed by sponsors of the research.

12. REFERENCES

[1] Gouzhen An, David Guy Brizan, and Andrew Rosenberg, "Detecting Laughter and Filled Pauses Using Syllable-based Features," INTERSPEECH 2014.

[2] Matti Arias and Paavo Alku, "Comparison of Multiple Voice Source Parameters in Different Phonation Types," INTERSPEECH 2007.

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

[4] Carlos Busso and Tauhidur Rahman, "Unveiling the Acoustic Properties that Describe the Valence Dimension," INTERSPEECH, 1179-1182, 2012.

[5] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection," IEEE Transactions on Audio, Speech, and Language processing, 17(4):582-596, 2009.

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

[7] Sidney D'Mello and Rafael A. Calvo, "Beyond the basic emotions: what should affective computing compute?" CHI EA 2013.

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

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

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

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

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

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

[14] D. Gowda and M. Kurimo, "Analysis of breathy, modal, and pressed phonation based on low frequency spectral density," INTERSPEECH 2013.

[15] 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.

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

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

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

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

[20] Shashidhar G. Koolagudi, Sourav Nandy, and K. Sreenivasa Rao, "Spectral Features for emotion Classification," IACC 2009.

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

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

[23] N.P. Narendra and K. Sreenivasa Rao, "Automatic detection of creaky voice using epoch parameters," INTERSPEECH 2015.

[24] Andrew Ortony and Terence J. Turner, "What's Basic About Basic Emotions?" Psychological Review, 97(3): 315-331, 1990.

[25] Mary Pietrowicz, Mark Hasegawa-Johnson, and Karrie Karahalios, "Acoustic Correlates for Perceived Effort Levels in Expressive Speech," INTERSPEECH 2014.

[26] Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner, "Anger recognition in speech using acoustic and linguistic cues," Speech Communication, 53:1198-1209, 2011.

[27] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency, "Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD," INTERSPEECH 2013.

[28] Bjorn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions in affect in speech: State of the art and lessons learnt from the first challenge," Speech Communication, 53:1062-1087, 2012.

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

[30] Library of Congress Veterans History Project, Available at https://www.loc.gov/vets/, accessed 9/10/16.

[31] A.B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 12,915 English lemmas," Behavior Research Methods, 45:1191-1207, 2013.

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

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