STRESS AND EMOTION CLASSIFICATION USING JITTER AND SHIMMER FEATURES

Xi Li[1], *Jidong Tao*[1], *Michael T. Johnson*[1], *Joseph Soltis*[2], *Anne Savage*[2], *Kirsten M. Leong*[3],
John D. Newman[4]

[1]Speech and Signal Processing Lab, Marquette University, Milwaukee, WI 53201 USA
{*xi.li*, *jidong.tao*, *mike.johnson*}*@marquette.edu*
[2]Disney's Animal Kingdom, Lake Buena Vista, Florida 32830 USA
{*anne.savage*, *joseph.soltis*}*@disney.com*
[3]Department of Natural Resources, Cornell University, Ithaca, NY 14850 USA
klm47@cornell.edu
[4]Laboratory of Comparative Ethology, National Institute of Child Health and Human Development,
National Institutes of Health, Department of Health and Human Services, MD 20837, USA
newmanj@lce.nichd.nih.gov

ABSTRACT

In this paper, we evaluate the use of appended jitter and shimmer speech features for the classification of human speaking styles and of animal vocalization arousal levels. Jitter and shimmer features are extracted from the fundamental frequency contour and added to baseline spectral features, specifically Mel-frequency Cepstral Coefficients (MFCCs) for human speech and Greenwood Function Cepstral Coefficients (GFCCs) for animal vocalizations. Hidden Markov Models (HMMs) with Gaussian Mixture Models (GMMs) state distributions are used for classification. The appended jitter and shimmer features result in an increase in classification accuracy for several illustrative datasets, including the SUSAS dataset for human speaking styles as well as vocalizations labeled by arousal level for African Elephant and Rhesus Monkey species.

***Index Terms*—** Jitter, Shimmer, MFCC, GFCC, HMM

1. INTRODUCTION

The discrimination of different speaking styles, stresses and emotions has generated considerable research in recent years [1-3]. This task has application to a number of important areas, including security systems, lie detection, video games and psychiatric aid, among others. Similarly, the analysis of arousal levels from animal vocalizations can significantly improve the ability of researchers in bioacoustics to understand animal behavior.

For both these tasks, the performance of emotion recognition largely depends on successful extraction of relevant speaker-independent features. In this work, several acoustic features related to fundamental frequency have been investigated for application to human speech datasets as well as analysis of arousal level from animal vocalizations.

A number of studies have been conducted to investigate acoustic features in order to detect stress and emotion in speech and vocalizations based on HMMs [5, 6], and to examine the correlation between certain statistical measures of speech and the emotional state of the speaker [1-3, 5]. Animal vocalizations have also been analyzed for such correlations, including African Elephants such as those whose vocalizations are examined in the current work [4]. The most often considered features include fundamental frequency, duration, intensity, spectral variation and log energy. However, many of these features are typically discriminatory across a subset of possible speaking styles, so that systems based on a small feature set are unable to accurately distinguish all speaking styles. Improvement in accuracy can be achieved by adding additional features related to measures of variation in pitch and energy contours. Two such recently investigated acoustic features are jitter and shimmer. Fuller *et al.* found increased jitter to be an "indicator of stressor-provoked anxiety of excellent validity and reliability" [7], and both jitter and shimmer can be indicators of underlying stress in human speech and animal vocalizations. The investigation of both human speech and animal vocalizations under the same framework will give the features wider applicability and better support their effectiveness.

The objective of this paper is to investigate the use of jitter and shimmer features in classifying speaking styles in humans and arousal levels in animals. Section 2 provides an overview of the feature extraction method, followed by a discussion of the experimental datasets in Section 3 and experimental results in Section 4. Final discussion and conclusions are then given in Section 5.

1-4244-0728-1/07/$20.00 ©2007 IEEE

2. FEATURE EXTRACTION AND MODELS

2.1. MFCC & Energy features

The most commonly used features for human speech analysis and recognition are Mel-Frequency Cepstral Coefficients (MFCCs) [8], often supplemented by an energy measure. Although there are several possible methods for computation, here the filterbank approach is used, where the spectrum of each Hamming-windowed signal is divided into Mel-spaced triangular frequency bins, then a Discrete Cosine Transform (DCT) is applied to calculate the desired number of cepstral coefficients. Log energy is computed directly from the time-domain signal.

Additionally, delta and delta-delta MFCCs representing the velocity and acceleration profiles of the cepstral and energy features are calculated using linear regression over a neighborhood of five windows.

2.2. GFCC

Greenwood Function Cepstral Coefficients (GFCCs) are a generalization of the MFCC based on using the Greenwood function [9] for frequency warping. These features are appropriate for spectral analysis of vocalizations for a wide variety of species, given basic information about the underlying frequency range [10]. GFCCs are used here as base features for analysis of animal vocalizations, with energy, delta, and delta-delta features computed identically to those for MFCCs described above.

2.3. Fundamental frequency

Fundamental frequency contours are extracted from the vocalizations using the COLEA toolbox [11] cepstrum implementation. Results are post-processed by median filtering. Unvoiced frames are considered to have a frequency (and corresponding jitter and shimmer) of zero.

2.3. Jitter & Shimmer

Jitter is a measure of period-to-period fluctuations in fundamental frequency. Jitter is calculated between consecutive voiced periods via the formula:

$$Jitter = \frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i}$$

where T_i is the pitch period of the i^{th} window and N is the total number of voiced frames in the utterance.

Shimmer is a measure of the period-to-period variability of the amplitude value, expressed as:

$$Shimmer = \frac{|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i}$$

where A_{i} is the peak amplitude value of the i^{th} window and N is the number of voiced frames.

3. DATABASE

3.1. SUSAS Database

This Speech Under Simulated and Actual Stress (SUSAS) dataset was created by the Robust Speech Processing Laboratory at the University of Colorado-Boulder [12]. The database encompasses a wide variety of stresses and emotions. Utterances are divided into two portions, “actual” and “simulated”. In this paper, we use the utterances in the simulated conditions, consisting of utterances from nine male speakers in each of eleven speaking-style classes. The eleven styles include Angry, Clear, Cond50, Cond70, Fast, Lombard, Loud, Neutral, Question, Slow, and Soft. The Cond50 style is recorded with the speaker in a medium workload condition, while in the Cond70 style the speaker is in a high workload condition. The Lombard speaking style contains utterances from subjects listening to pink noise presented binaurally through headphones at a level of 86 dB. The vocabulary includes 35 highly confusable aircraft communication words. Each of the nine speakers (3 speakers from each of 3 dialect regions) in the dataset has two repetitions of each word in each style. All speech tokens were sampled by 16 bits A/D converter at a sample frequency of 8 kHz.

3.2. African Elephant Emotional Arousal Dataset

Elephant vocalizations were collected from six adult nonpregnant, nulliparous female African elephants (*Loxodonta Africana*) by Kirsten M. Leong and Joseph Soltis at Disney’s Animal Kingdom (DAK), Lake Buena Vista, Florida, U.S.A. The data collection occurred from July 2005 to December 2005. Each elephant wore a custom designed collar, containing a microphone and an RF radio that transmitted audio to the elephant barn, where the data was recorded on DAT tapes. The audio was passed through an anti-aliasing filter and stored on computers at a sampling rate of 7518 Hz [13].

There are 131 vocalizations used for these experiments, all low-frequency rumble calls. Each vocalization is labeled by individual ID, social rank, age and arousal levels. Of the six females, three are of high social rank and the remaining three are of low rank. Similarly, three females are of old age and three are of young age. Emotional arousal level was determined from observation of time-synchronized video based on specific social context criteria. The emotional arousal levels are categorized as low (L), medium (M) and high (H), with 51, 46, and 34 calls in each category respectively.

3.3. Rhesus Emotional Arousal Database

Infant cry vocalizations were recorded from rhesus macaques (*Macaca mulatta*) on the island of Cayo Santiago, Puerto Rico by Joseph Soltis and John D. Newman. Ten infants, eight females and two males, between the ages of 4 and 7 months were included in the study. Vocalizations were sampled at a sample rate of 44.1 kHz and 16 bits per sample.

There are 150 calls used for these experiments. Vocalizations are labeled by individual ID, social rank, age, gender and arousal level. As with the African elephant vocalizations, social rank and age were equally distributed across the individuals; five each of low rank and high rank, and five each of old age and young age. The emotional arousal levels are labeled as non-arousal (NA) versus arousal (A) based on observed social contexts, with 61 and 89 calls in the two categories, respectively.

4. EXPERIMENT RESULTS

4.1. Using SUSAS in Speaking Style Classification

For the SUSAS experiments, 24 MFCCs with log energy, plus delta and delta-delta coefficients, are used as the baseline feature vector. Data is Hamming windowed with a 30 ms window size and 10ms step size.

Of the eleven speaking styles labeled in the SUSAS dataset, six emotion-related speaking styles are selected for the classification study (the remainders are considered to be related to noise conditions rather than speaking style). These include Angry, Fast, Lombard, Question, Slow and Soft. Training and testing is implemented through a three-fold cross-validation scheme within each dialect type, so that the results are speaker independent but dialect dependent. This results in training sets of 140 utterances per speaking style and corresponding testing sets of 70 utterances. The average accuracy value across the three folds in each dialect type was used to determine the overall accuracy. Three state left-to-right HMMs with four-mixture GMMs in each state are used for classification models. The programming toolkit HTK 3.2.1 from Cambridge University [14] is used for all training and testing.

	Accuracy (%)
MFCC (baseline)	65.5
MFCC + Jitter	68.1
MFCC + Shimmer	68.5
MFCC + Jitter + Shimmer	69.1

Table 1 SUSAS classification results across six speaking styles for different feature combinations

Jitter and shimmer are added to the baseline feature set both individually and in combination. Table 1 shows the overall results. The absolute accuracy increase is 2.6% and 3.0% after appending jitter and shimmer individually, while there is 3.6% increase when used together.

4.2 Emotional Arousal Recognition using Elephant Data

For the African elephant dataset, 12 GFCCs and the normalized log energy are extracted from 300ms Hamming windows, with a step size of 100 ms. Frequency warping for the GFCCs is done 26 filterbanks spaced across the range of 10 – 150 Hz to emphasize the infrasonic vocal range of the vocalizations. Classification models are 3 state left-to-right HMMs with a single Gaussian per state, in addition to a silence model.

Four acoustic model systems are built: caller-independent (CI), rank-dependent (RD), age-dependent (AD) and caller-dependent (CD). Again, the HTK toolkit is used for training and testing. Leave-one-out cross-validation is used for evaluation of overall accuracy.

Results are shown in Table 2. In all four cases, there is improvement in accuracy as a result of adding either jitter or shimmer, and the use of both together results in the highest overall accuracy in all cases. Comparing the results of the caller-dependent (CD) system to the others, it is clear that individual variability is the largest confounder in determining arousal level.

	CI	RD	AD	CD
GFCC (baseline)	38.9	58.0	58.0	76.5
GFCC + Jitter	42.0	60.3	61.1	81.6
GFCC + Shimmer	42.8	63.4	58.8	80.6
GFCC + Jitter + Shimmer	44.3	64.9	62.6	82.7

Table 2 Elephant arousal classification accuracy for different feature combinations in 4 acoustic model systems

4.3 Emotional Arousal Recognition using Rhesus Data

For the rhesus macaques data, 12 GFCCs plus the normalized log energy, with delta and delta-delta coefficients, are again used as the baseline feature set. Frequency warping for this species is done with 26 filterbanks spaced across the range of 20 – 3500 Hz. The vocalizations are Hamming windowed with frame and step sizes of 25 ms and 10 ms. Classification models for the experiments are 6-state left-to-right HMMs with each state containing 4-mixture Gaussian Mixture Models (GMMs).

	CI	RD	AD	GD	CD
GFCC (baseline)	66.7	74.0	72.7	67.3	93.3
GFCC + Jitter	70.7	74.7	74.0	69.3	94.0
GFCC + Shimmer	69.3	75.3	74.7	69.3	94.7
GFCC +Jitter+Shimmer	72.7	76.0	76.0	70.0	96.0

Table 3 Rhesus arousal classification accuracy for different feature combinations in 5 acoustic model systems

As with the African elephant experiments, multiple experimental setups are implemented, including caller-independent (CI), rank-dependent (RD), age-dependent (AD), gender-dependent (GD) and caller-dependent (CD). Evaluation is done using leave-one-out cross-validation across the data set.

Results in Table 3 above show a similar pattern to the elephant arousal experiments. In all cases, adding jitter or shimmer individually increases the accuracy, with shimmer having slightly better performance, while using the two together gives substantially better results than using them individually. Individual variation again seems to be the strongest confounding factor in accurate classification of arousal, as indicated by the 96% peak accuracy for the caller-dependent experiments.

5. CONCLUSIONS

Jitter and shimmer features have been evaluated as important features for analysis and classification of speaking style and arousal level in both human speech and animal vocalizations. Adding jitter and shimmer to baseline spectral and energy features in an HMM-based classification model resulted in increased classification accuracy across all experimental conditions. In evaluation of animal arousal levels, the largest obstacle to accurate classification is shown to be individual variability, rather than rank, gender, or age factors.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge Marek B. Trawicki for provision of initial experiment code. This contribution is based on work supported by the National Science Foundation under Grant No. IIS-0326395.

7. REFERENCES

[1] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communications, Special Issue on Speech Under Stress*, vol. 20(2), pp. 151-170, November 1996.

[2] S. Bou-Ghazale and J. H. L. Hansen, "A novel training approach for improving speech recognition under adverse stressful environments," *EUROSPEECH-97*, vol. 5, pp. 2387-2390, Sept. 1997.

[3] S. Bou-Ghazale and J. H. L. Hansen, "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8(4), pp. 429-442, July 2000.

[4] J. Soltis, K. M. Leong, and A. Savage, "African elephant vocal communication II: rumble variation reflects the individual identity and emotional state of callers," *Animal Behaviour*, vol. 70, pp. 589-599, 2005.

[5] A. Nogueiras, A. Moreno, A. Bonafante, and J. Maririo, "Speech Emotion Recognition Using Hidden Markov Models," *Eurospeech 2001, Poster Proceedings*, pp. 2679-2682, 2001.

[6] K. Oh-Wook, C. Kwokleung, H. Jiucang, and L. Te-Won, "Emotion Recognition by Speech Signals," presented at Eurospeech, Geneva, 2003.

[7] B. F. Fuller, Y. Horii, and D. A. Conner, "Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety," *Research in Nurse & Health*, vol. 15(5), pp. 379-389, Oct. 1992.

[8] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Upper Saddle River, New Jersey: Prentice Hall, 2001.

[9] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," *The Journal of the Acoustical Society of America*, vol. 33, pp. 1344-1356, 1961.

[10] P. J. Clemins, M. B. Trawicki, K. Adi, J. Tao, and M. T. Johnson, "Generalized perceptual features for vocalization analysis across multiple species," *Proceedings of the IEEE ICASSP*, vol. 1, pp. I253 - I256, May 2006.

[11] P. Loizou, "COLEA: A MATLAB software tool for speech analysis." Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, 1999.

[12] J. H. L. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting Started with the SUSAS: Speech Under Simulated and Actual Stress Database," Robust Speech Processing Laboratory April 15, 1998.

[13] K. M. Leong, A. Ortolani, K. D. Burks, J. D. Mellen, and A. Savage, "Quantifying acoustic and temporal characteristics of vocalizations of a group of captive African elephants (*Loxodonta africana*)," *Bioacoustics*, vol. 13, pp. 213-231, 2003.

[14] S. Young, et al., *the HTK Book (for HTK Version 3.2.1)*, 2002.