METHODS FOR STRESS CLASSIFICATION: NONLINEAR TEO AND LINEAR SPEECH BASED FEATURES*

Guojun Zhou, John H.L. Hansen, and James F. Kaiser

Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech gzhou@ee.duke.edu jhlh@ee.duke.edu

ABSTRACT

Speech production variations due to perceptually induced stress contribute significantly to reduced speech processing performance. One approach that can improve the robustness of speech processing (e.g., recognition) algorithms against stress is to formulate an objective classification of speaker stress based upon the acoustic speech signal. In this paper, an overview of recent methods for stress classification is presented. First, we review traditional pitch-based methods for stress detection and classification. Second, neural network based stress classifiers with cepstral-based features, as well as wavelet-based classification algorithms are considered. The effect of stress on linear speech features is discussed, followed by the application of linear features and the Teager Energy Operator (TEO) based nonlinear features for effective stress classification. A new evaluation for stress classification and assessment is presented using a critical band frequency partition based TEO feature and the combination of several linear features. Results using Nato databases of actual speech under stress are presented. Finally, we discuss issues relating to stress classification across known and unknown speakers and suggest areas for further research.

1. INTRODUCTION

The problem of stress classification is to differentiate speech utterances spoken under stressful conditions from those spoken under neutral conditions. Examples of stressful conditions include high workload stress, emergency phone calls, voice communications between aircraft pilots and ground controllers, multi-tasking, fatigue, physical environmental factors (G-force), and emotional moods such as anger, fear, etc. The variations in speech production due to stress can be substantial and therefore have a considerable impact on the performance of speech processing applications like recognition [5]. There have been a number of studies which focus on the variability effects of stress on speech production. For example, the Lombard effect on speech production and on speech recognition performance has been investigated in [9, 4, 10, 16]. Since aircraft flight emergencies often produce high stress conditions, pilot-controller voice communications have therefore been of interest in several studies [18, 23]. The effects of emotion-induced stress on speech production have also been considered [20].

These studies have shown that speech variability due to stress is a challenging research problem, and that traditional techniques generally fall far short of improving the robustness of speech processing performance under stress. For speech recognizers, a typical approach to improve recognition robustness under adverse conditions (e.g., varying communication channels, handset differences) is re-training reference models (i.e., train-test in matched conditions). A similar method, called multi-style training [13], has been used to improve speech recognition under stress, but at the expense of requiring the user to produce speech across a simulated range of stress styles. In a separate study, it was shown that multi-style training only works in speaker-dependent scenarios and that performance actually degrades below neutral training when applied in a speaker independent application [21]. The reason is that stressful conditions are too diverse to be represented by limited training data, and that speakers can at times use a non-uniform set of speech production adjustments to convey their stress state. It is suggested that algorithms which are capable of classifying stress could be used to monitor speaker state, improve the naturalness of speech coding and synthesis algorithms, or increase the robustness of speech recognizers.

While the impact of stress on speech recognition as well as methods of improving recognition performance have been considered [4, 5, 6, 9, 13, 18], there have been limited published studies in the related problem of stress classification. Many traditional methods use pitch-related speech features to detect stress [2, 14, 16]. One previous study employed neural networks to classify stressed speech using Mel-cepstral based parameters [7, 21]. Sarikaya et al. used a wavelet-based method [17] and another study considered a nonlinear feature, where the shape of a pitch normalized TEO profile was used [1]. In our recent studies [23, 24, 26], we proposed several new nonlinear TEO processing based speech features which are effective for stress classification. Also, motivated by the analysis study and results in [3], we investigated five linear classification features [25].

The goal of this paper is to provide a useful overview of different methods of stress classification and to present new results from recent evaluations. Traditional pitch-related methods, neural network based classifiers, a wavelet-based approach, and techniques using nonlinear TEO processing and linear speech based features are discussed in the following sections. Also, new results for stress classification and assessment using a combinational linear feature and TEO-based features are presented. We conclude this paper with a brief discussion of issues related to effective stress classification.

* THIS WORK WAS SUPPORTED BY A GRANT FROM THE U.S. AIR FORCE RESEARCH LABORATORY/IFEC, ROME, N.Y.

2. STRESS CLASSIFICATION TECHNIQUES

2.1. Traditional Methods

Some early approaches employed voice stress excitation microtremors in an effort to determine a reliable measure for detecting the stress a speaker may exhibit when speaking a lie. It was believed that a physiological microtremor [14, 2] exists, which is associated with the excitation muscles during speech production. A physiological tremor is described as a low amplitude oscillation of the reflex mechanism that controls the length and tension of a stretched muscle, and has a frequency between 8 and 12 Hz. It is assumed that as speaker stress increases, the amplitude of the microtremor decreases; and this microtremor variations can be detected through changes in the fundamental frequency of voice. Based on this assumption, some commercial computerized voice stress analyzers were proposed, although their reliability and claims of success have always been controversial [2]. In spite of this controversy, such excitation related analysis could be beneficial in voice stress detection in monitoring emergency voice communication applications. In another study pitch- and spectral-based analysis was considered for stress detection [16] in laboratory and real stressful conditions. The features used in that study included fundamental frequency, microprosodic variation index, spectral-based indicators from a cumulative histogram of sound level and from statistical analyses of formant frequencies, and distances of formants from the center of the first three formants. Evaluation results showed that a microprosodic variation index is effective in detecting mild stress while the fundamental frequency itself is more efficient for severe stress. Also, spectral-related features were useful although not as effective as pitch-related features.

2.2. Neural Network Stress Classifiers

Methods based on neural networks and an array of features have been employed by Womack and Hansen for stress classification [7, 21]. Features investigated with neural network classifiers included estimated vocal tract area profiles, acoustic tube area coefficients, and Mel-cepstral based parameters which consist of Mel-cepstral (MFCC), delta MFCC, delta-delta-MFCC, and a new feature based on the autocorrelation of the MFCCs (AC-mel). Stress classification performance of these features were determined using separability distance metrics and neural network classifiers. It was shown that stress classification performance varied significantly depending on the vocabulary size and speaker population. However, MFCC and AC-mel performed better than delta-MFCC and delta-delta-MFCC for vocabulary dependent tests. A later study showed that by using target driven features and context dependent phoneme based neural networks, stress classification performance could be measurably improved. Furthermore, the authors extended their work by combining algorithms for stress classification and speech recognition together. An N-dimensional Hidden Markov Model (HMM) framework was used for this purpose. The evaluations proved that stress classification can help in improving speech recognition performance [22].

2.3. Wavelet-Based Stress Classification

One recently reported study [17] used a set of features based on wavelet analysis, or equivalently multirate subband analysis for stress classification. The wavelet analysis can be useful for this task because it can (1) provide a good representation of local spectral variations; (2) be adjusted to account for the human auditory property by using perceptual division of the frequency axis; and (3) achieve better frequency localization than short-time Fourier transform by choosing filters with maximum vanishing moments. Scale energy (SE) of the subsignals for each subband, autocorrelation SE (ACSE), subband based cepstral coefficients (SC), and autocorrelation SC (ACSC) were used as features with a neural network classifier. It was shown that these wavelet based features are better than MFCC-based features, especially the SC feature, for stress classification.

After reviewing stress classification techniques done by other researchers, we present our work on stress classification in the following two sections.

3. LINEAR SPEECH FEATURES

Based on a previous study which considered analysis of speech under stress for recognition [3], five linear speech features were selected for stress classification from five domains including fundamental frequency, glottal source information, duration, intensity and vocal tract spectral structure (formant centers and formant bandwidths) which have been shown to be statistically separable between neutral and stressed speech. We extracted all five features from vowel sections of speech so that they can be compared under the same scenario. The five features used were: frame-based fundamental frequency (pitch), glottal spectral slope, duration of the vowel in msec, mean square value of the vowel as the intensity, and deviations of the first two formant locations (frame-based) from their typical averages representing the vocal track spectral structure [25]. A Bayesian Hypothesis testing approach was employed for classification [25]. Evaluations were based on speech from the SUSAS database [8], which focused on pairwise classification of neutral versus angry, loud, and Lombard speech. For each feature, different vector lengths (1, 5, and 10) were evaluated (for fundamental frequency and formant locations, the value of vector length reflects numbers of consecutive frames, while for the other features, it represents the number of vowel tokens). It is shown [25] that fundamental frequency is the best feature for stress classification among the five features (when using a vector length of 10, classification accuracy was in the following ranges; fundamental frequency: 79–93%, intensity: 64–82%, glottal slope: 64–82%, duration: 54–64%, formant locations: 41–61%). Also, the performance varied across different stress styles, with Lombard typically lower in classification accuracy.

3.1. Combinational Linear Feature

Motivated by classification results of individual linear features, we considered an approach which combines different individual features together as one classification vector. Due to the difficulty in obtaining the glottal spectral slope for limited data and unsatisfactory performance of formant location, we chose to combine fundamental frequency, phone duration, and intensity. Since fundamental frequency is frame-based while the other two are phone-based, we used phone-based mean fundamental frequency as one of the three components in the combinational feature. The Bayesian Hypothesis testing approach was also used for classification between neutral and stressed speech. Although duration and intensity do not have high classification rates individually, the combinational feature can achieve very high accuracy. Table 1 lists classification error rates with different length input test vectors. For a vector length of 10, the feature can achieve 83.33–100% accuracy, which is measurably higher than fundamental frequency alone. Furthermore, the combinational feature outperforms the individual fundamental frequency for all stress styles (angry, loud, and Lombard).

4. NONLINEAR TEO FEATURES

While linear speech production features have been considered for stress classification, several recent studies have revealed the promise of nonlinear features. One study which considered stress classification using a nonlinear feature focused on the shape of a pitch normalized TEO profile [1]. Good classification performance was obtained for speech produced under angry, loud, clear, and the Lombard effect speaking conditions. The approach used a Mellin transform to remove the impact of pitch period variations before stress classification. Although that study was limited to binary stress classification of front and mid vowels, it did suggest that nonlinear speech analysis might be helpful for stress classification. Based on studies by Teager [19], Kaiser [11, 12], and Maragos et al. [15], evidence suggests that speech production is not actually a linear process, which is typically assumed by linear acoustic theory, but a nonlinear process where speech can be decomposed into amplitude (AM) and frequency (FM) modulated components via the TEO. In our previous studies [23, 24, 26], we proposed three nonlinear TEO-based features for stress classification. These included the TEO-decomposed FM variation (TEO-FM-Var), Normalized TEO Autocorrelation Envelope Area (TEO-Auto-Env), TEO profile-based pitch (TEO-Pitch), and TEO-Auto-Env with critical band based frequency partition (TEO-CB-Auto-Env). To evaluate the performance of these features for stress classification, the SUSAS database and a baseline 5-state HMM-based stress classifier with continuous Gaussian mixture distributions were employed. Two HMM models (neutral and stressed) were trained for each pairwise classification. For comparison purposes, we also evaluated MFCC and pitch along with our four TEO-based features. It was shown that TEO-based features are effective for stress classification, especially the two TEO autocorrelation-based features, TEO-Auto-Env and TEO-CB-Auto-Env. These features outperform traditional MFCC and pitch in terms of accuracy and consistency across three stress styles (angry, loud, and Lombard).

4.1. Assessment of Actual Speech Under Stress Using TEO-Based Features

In addition to stress classification using TEO-based features, we also evaluated their ability of assessing stress. In [23, 26], we reported some preliminary results for stress assessment using the NATO SUSC-0 database[1]. In that evaluation, four sentences with a single voiced portion each were extracted for assessment. To achieve more reliable results, we extended the sentence number to 12 and used 4 different voiced portions across each sentence. Table 2 summarizes the sentences with all extracted voiced portions highlighted. Sentence 1 was extracted from the initial ground aircraft system check. Sentences 2 to 7 represent phrases from preliminary discovery of engine problems to problem allocation. Sentences 8 to 11 were spoken while emergency actions were taken. The last sentence indicates the safe resolution of the emergency. To assess stressed speech, two HMM models were trained, representing neutral and stressed conditions, respectively. The assessment score was computed as the difference of likelihood scores obtained from the two HMM models from the input token. In this evaluation, the assessment score of each sentence was obtained from 4 voiced portions. We trained the neutral HMM model from SUSAS neutral training data. For the stressed HMM model, we trained two models, one from SUSAS actual domain and the other from SUSAS simulated angry, loud and Lombard domains. Evaluation results are shown in Fig. 1. We can see that all scores from MFCC and TEO-Auto-Env are close to zero, indicating that these two features are not as effective in assessing the degree of stress. The other two features, TEO-CB-Auto-Env and pitch, show variations that correspond to the degree of stress as perceived by a listener. Both simulated and actual stressed HMM models show similar score variations although the actual anchor model results in larger fluctuations due to higher degree of stress from the training data. The peaks in pitch for sentence 9 are attributed to incorrect pitch estimation.

Sentences from Mayday2 Domain of SUSC-0	
No.	Sentence
1	avionics l**IGH**t hydr**AU**lic oil pressure l**IGH**t engine indications **ARE** ...
2	**AND** you'er g**ONNA** declare an em**ER**gency or am **I**
3	... checklist **OI**l pressure malfunction **G** one-hundred ... cruise altitude st**OR**e jett ... throttle minimize m**O**vement ...
4	roger that **OI**l indic**A**tion is n**OW** z**ERO**
5	... **ALRIGH**t newt ... engine fault l**IGH**t still lit ... hydr**AU**lics are ... total p**OUN**ds six ...
6	and I'm going there and I'm there I'm des**EN**ding down to ten gr**AN**d right I'm n**O**t picking up a t**A**can lock
7	no I'**M** doing **ALRIGH**t now and the r**A**dial is wh**A**t
8	ok**AY** give me imm**E**diate vectors this is an em**ER**gency I'm engine **OU**t
9	g**I**ve me h**EA**dings I n**EE**d headings n**OW**
10	put the c**A**ble d**OWN** p**U**t the c**A**ble down
11	I'm h**O**t I n**EE**d the c**A**b**L**e ...
12	m**A**n I th**OUGH**t I w**A**s g**O**ne

Table 2: Sentences from SUSC-0 for Stress Assessment Evaluation

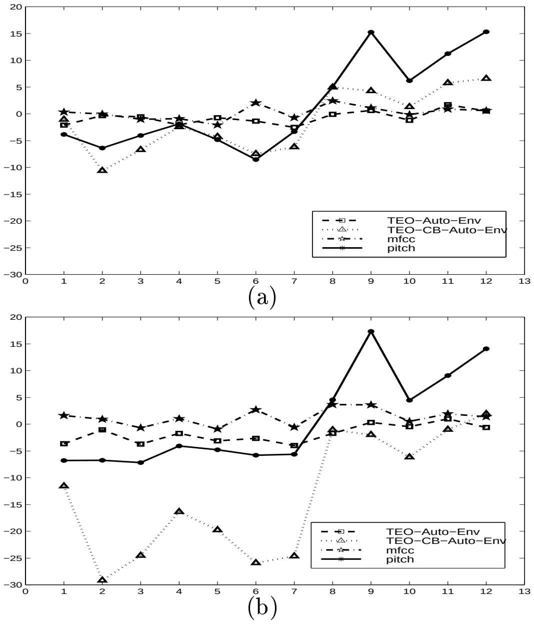
5. CONCLUSIONS AND DISCUSSIONS

In this paper, we presented an overview of different methods for stress classification. After reviewing studies performed by other researchers, we discussed our recent methods using nonlinear TEO and linear speech based features for stress classification. Furthermore, two new evaluations were presented. The evaluation of combinational linear feature with three components: mean fundamental frequency, duration, and intensity showed a significant performance improvement over individual linear features (fundamental frequency, duration, intensity, glottal source, and formant lo-

[1] SUSC-0 consists of fighter cockpit speech communication under emergency conditions. Audio examples of this and SUSAS can be found at
http://www.ee.duke.edu/Research/Speech/stress.html

Vector Length	Speaking Style of Submitted Test Speech						OVERALL ERROR RATES	
	Neutral	Angry	Neutral	Loud	Neutral	Lombard	Mean $\bar{m}_{\text{ALL}}$	stand. dev. σ_{ALL}
1	17.58	17.03	11.67	11.97	19.85	21.21	16.55%	3.97
5	6.15	5.00	4.62	4.62	13.08	13.08	7.76%	4.16
10	1.67	0.00	3.03	3.03	13.64	16.67	6.34%	6.98

Table 1: Error Rates (percentage) of Open-set Pairwise Stress Classification Using the combination of mean pitch, duration and intensity as the feature.



(a)

(b)

Figure 1: Assessment results for pilot's speech from Mayday2 domain of SUSC-0 database (Log likelihood ratio is shown along Y-axis while sentence number is shown along X-axis): (a) HMMs of Neutral vs SIMULATED stress (Loud, Angry and Lombard); (b) HMMs of Neutral vs ACTUAL stress

cations) for stress classification. A second evaluation using NATO SUSC-0 actual stress database showed that the TEO autocorrelation envelope feature with critical band based frequency partition is promising for stress assessment. Since both linear and nonlinear features are effective, a combination of linear and nonlinear features may be needed for universal speaker stress classification. Issues which are important to consider for stress classification include: (1) how to establish neutral and stress anchor models, (2) consistency for a given speaker and across unseen speakers, (3) types of speaker stress, and (4) relationship between automatic methods and human stress assessment. Since no universal objective standard exists to quantify stress classification feature performance, it is suggested that further research is needed.

6. REFERENCES

[1] D.A. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions", *J. Acoust. Soc. Am.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[2] V.L. Cestaro, "A Comparison between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy", Tech. Report, DoD Polygraph Inst., Aug. 1995.

[3] J.H.L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition", Ph.D. Thesis, Georgia Inst. of Tech., Atlanta, GA, 1988.

[4] J.H.L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect", *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 598–614, Oct. 1994.

[5] J.H.L. Hansen, S. Bou-Ghazale, "Robust speech recognition training via duration and spectral-based stress token generation", *IEEE Trans. Speech Audio Proc.*, **(3)**:415–421, 1995.

[6] J.H.L. Hansen, "Analysis and Comparison of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Comm.*, vol. 20, pp. 151-173, Nov. 1996.

[7] J.H.L. Hansen, B.D. Womack, "Feature Analysis and Neural Network Based Classification of Speech Under Stress" *IEEE Trans. Speech Audio Proc.*, vol. 4, no. 4, pp. 307–313, 1996.

[8] J.H.L. Hansen, S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database", *EUROSPEECH–97*, pp. 1743–1746.

[9] B.A. Hanson, T. Applebaum, "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: Experiments with Lombard and noisy speech", *ICASSP–90*, pp. 857–860.

[10] J.C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers", *J. Acous. Soc. Am.*, vol. 93, pp. 510–524, Jan. 1993.

[11] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *ICASSP–90*, pp. 381–384.

[12] J.F. Kaiser, "Some Useful Properties of Teager's Energy Operator," *ICASSP–93,* vol. 3, pp. 149–152.

[13] R. Lippmann et al., "Multi-style training for Robust Isolated-word Speech Recognition", *ICASSP–87*, pp. 705–708.

[14] O. Lippold, "Physiological Tremor", *Scientific American*, vol. 224, no. 3, pp. 65–73, Mar. 1971.

[15] P. Maragos et al., "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Proc.*, vol. 41, no. 10, pp. 3025–3051, Oct. 1993.

[16] R. Ruiz et al., "Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions", *Speech Comm.*, vol. 20, pp. 111–129, 1996.

[17] R. Sarikaya, J.N. Gowdy, "Subband Based Classification of Speech under Stress", *ICASSP–98*, pp. 569–573.

[18] B. Stanton et al., "Robust recognition of loud and Lombard speech in the fighter cockpit environment", *ICASSP–89*, pp. 675–678.

[19] H. Teager, S. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modeling*, NATO Advanced Study Institute, vol. 55, Bonas, France, (Boston: Kluwer Academic Pub.), pp. 241–261, 1990.

[20] C. Williams, K. Stevens, "Emotions and Speech: Some Acoustic Correlates", *J. Acoust. Soc. Am.*, **(52)**4:1238–1250, 1972.

[21] B.D. Womack, J.H.L. Hansen, "Classification of Speech under Stress Using Target Driven Features", *Speech Comm.*, vol. 20, pp. 131–150, 1996.

[22] B. D. Womack, J.H.L. Hansen, "N-D HMMs for Combined Stress Speech Classification and Recognition," submitted to *IEEE Trans. Speech Audio Proc.*, Sept. 1996. Revised March 1998.

[23] G. Zhou, J.H.L. Hansen, J.F. Kaiser, "Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator", *ICASSP–98*, pp. 549–552.

[24] G. Zhou, J.H.L. Hansen, J.F. Kaiser, "A New Nonlinear Features for Stress Classification", *NORSIG–98*, pp. 89–93.

[25] G. Zhou, J.H.L. Hansen, J.F. Kaiser, "Linear and Nonlinear Speech Feature Analysis for Stress Classification", to appear in *ICSLP–98*, Sydney, Australia, Nov. 30 to Dec. 4, 1998.

[26] G. Zhou, J.H.L. Hansen, J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress", submitted to *IEEE Trans. Speech Audio Proc.*, Dec. 1997.