Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator

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

Studies have shown that distortion introduced by stress or emotion can severely reduce speech recognition accuracy. Techniques for detecting or assessing the presence of stress could help neutralize stressed speech and improve robustness of speech recognition systems. Although some acoustic variables derived from linear speech production theory have been investigated as indicators of stress, they are not consistent. In this paper, three new features derived from the nonlinear Teager Energy Operator (TEO) are investigated for stress assessment and classification. It is believed that TEO based features are better able to reflect the nonlinear airflow structure of speech production under adverse stressful conditions. The proposed features outperform stress classification using traditional pitch by +22.5% for the Normalized TEO Autocorrelation Envelope Area feature (TEO-Auto-Env), and by +28.8% for TEO based Pitch feature (TEO-Pitch). Overall neutral/stress classification rates are more consistent for TEO based features (TEO-Auto-Env: $\sigma = 5.15$, TEO-Pitch: $\sigma = 7.83$) vs. (Pitch: $\sigma = 23.40$). Also, evaluation results using actual emergency aircraft cockpit stressed speech from NATO show that TEO-Auto-Env works best for stress assessment.

1 Introduction

It is well known that the performance of speech recognition algorithms is greatly influenced by the environmental conditions in which speech is produced. Factors which influence recognition performance include background noise, communication channel variations, and environmental task stress or emotion. Examples of adverse environments include aircraft cockpits, 911 emergency telephone calls, factory environments, or speech from controlled experiments such as amusement park roller coaster rides. It is suggested that algorithms, which are capable of estimating and assessing the environmental conditions of speaker, channel, and acoustic environment, would be beneficial in the formulation of more effective speech recognition algorithms in adverse environmental conditions.

Although significant research has been conducted on background noise and channel estimation for speech recognition, there has been limited work performed in the area of stress classification and assessment. The majority of studies in the field of speaker stress analysis have concentrated on pitch, and spectral features derived from a linear model of speech production [4, 1, 2]. The number of studies in stress classification is more limited. One recent study [6] considered stress classification using (1) estimated vocal tract area profiles, (2) acoustic tube area coefficients, and (3) Mel-cepstral based parameters (MFCC) including a new feature based on the autocorrelation of the MFCCs (AC-mel). A later study showed that by using target driven features and context dependent phoneme neural networks, stress classification performance could be measurably improved. Other acoustic features which have also been shown to be useful as indicators of speech under stress include fundamental frequency (F0), phoneme duration and intensity, glottal source structure, and vocal tract formant structure [4].

Our focus here, is to remove word level dependency in the stress classification task, and thereby concentrate on linear or nonlinear excitation characteristics. One previous study [5] considered stress classification using a nonlinear feature, where the shape of a pitch normalized Teager Energy Operator (TEO) profile was used. Good performance was obtained for speech produced under angry, loud, clear, and Lombard effect[1] speaking conditions. That study, however, was limited to binary stress classification of vowels.

In this study, we propose three new features which incorporate TEO based processing, they are the TEO-decomposed FM Variation (TEO-FM-Var), Normalized TEO Autocorrelation Envelope Area (TEO-Auto-Env), and TEO based Pitch (TEO-Pitch). These features explore the prospects of variation in the energy of airflow characteristics within the vocal tract for speech under stress. We compare the performance of TEO based features to traditional pitch information for the task of stress classification. We also perform an open stress assessment evaluation on actual speech under stress from NATO RSG.10 (SUSAS, SUSC-0)[2].

2 TEO Background

Traditional linear acoustic theory assumes that airflow from the vocal folds propagates through the vocal tract as a plane wave, where vocal fold motion or vocal tract constrictions is considered the source of speech production. According to studies by Teager [3], however, this assumption may not hold since vortices are distributed throughout the vocal tract. Teager suggested that the true source of speech production is actually the vortex-flow interactions, which are nonlinear. To measure the energy from speech, which is pro-

[1] Lombard effect occurs when a speaker produces speech in the presence of acoustic background noise. The speaker modifies their speech in order to increase communication quality over the noisy environment.

[2] For further information on NATO RSG.10 efforts on stress, see their speech under stress web page: *http://www.ee.duke.edu/Research/Speech/stress.html*.

duced by a nonlinear process, Teager developed an energy operator, which was described by Kaiser [7, 8] as follows,

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1), \quad (1)$$

where $\Psi[\cdot]$ is the Teager Energy Operator (TEO), and $x(n)$ is the sampled speech signal.

The TEO is typically applied to a bandpass filtered speech signal, since its intent is to reflect the energy of the nonlinear energy flow within the vocal tract for a single resonant frequency. Under this condition, the resulting TEO profile can be used to decompose a speech signal into its AM-FM components within a certain frequency band via,

$$f(n) \approx \frac{1}{2\pi T} \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right), \quad (2)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{\left[1 - \left(\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2\right]}}, \quad (3)$$

where $y(n) = x(n) - x(n-1)$, $\Psi[\cdot]$ is the TEO operator as shown in Eq. 1, $f(n)$ is the FM component at sample n, and $a(n)$ is the AM component at sample n [9].

3 Stress Classification Features

3.1 TEO-FM-Var: FM Variation

Our previous studies have shown that vowels spoken under stress generally have more instantaneous pitch variations than vowels spoken under neutral conditions. This suggests that features which represent fine excitation variations, would be useful for stress classification. To some extent, it is believed that these variations are due to the effects of modulations. According to the work of Maragos, Kaiser, and Quatieri [9], the TEO is a nonlinear differential operator that can detect modulations in the speech signal and further decompose the signal into its AM and FM components. It is not difficult to understand that the AM/FM decomposition of a speech signal over a wide bandwidth will not provide correct estimation of the real modulations. AM-FM signal analysis requires a carrier frequency which must be higher than the modulating frequencies within the signal. Since we are interested in fine excitation variations, we filter the raw input speech through a Gabor bandpass filter (BPF) centered at the median fundamental frequency, $F0$, with a bandwidth of $F0/2$. The average magnitude difference function (AMDF) is employed to estimate the median fundamental frequency, $F0$, based on the TEO profile of the entire input. After the Gabor BPF, the TEO is performed and the resulting profile is used to separate the input speech signal into its AM and FM components using Eq. 2 and Eq. 3. A flow diagram for extracting the TEO-FM-Var feature is shown in Fig. 1.

We believe that the FM component of stressed speech should have greater fluctuations than that of neutral speech due to the more instantaneous pitch variation present in stressed speech. While it might seem straightforward to apply a standard pitch estimation algorithm to estimate these variations, the large and erratic pitch changes under stress cause traditional estimation algorithms to fail, thus requiring human pitch labeling. An alternative is to use the FM variation of each frame as the feature for stress classification as illustrated in Fig. 1.

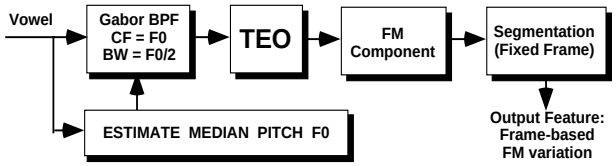

Figure 1: TEO-FM-Var Feature Extraction

3.2 TEO-Auto-Env: Normalized TEO Autocorrelation Envelope Area

The second feature, named TEO-Auto-Env, also reflects the instantaneous excitation variations of speech. A flow diagram is shown in Fig. 2. This feature is based on the idea that the presence of stress may affect modulation patterns within the frequency bands of speech differently. It is obtained by passing the raw input speech through a filterbank consisting of 4 bandpass filters (BPF). Each BPF output stream is processed by a TEO to estimate each profile. Our experiments show that the TEO profile of an AM-FM signal has the same periodicity as the modulating signals. Furthermore, the TEO profile periodicity is generally dominated by an amplitude modulating signal frequency. This explains why the TEO profile reflects the same periodicity as the pitch profile since both are affected by amplitude modulations. Therefore, we obtain a feature representing the fine pitch variation by analyzing the TEO autocorrelation envelope. If we consider the fact that pitch is a slow-changing variable, we can bandpass-filter each TEO output stream through a Gabor BPF centered at the median $F0$, with the 3dB bandwidth being roughly $F0/2$. $F0$ is obtained using the AMDF based pitch detection method on the TEO profile instead of the raw speech. Subsequently, each Gabor-filtered TEO stream is segmented into frames. In order to have equivalent averaging effects, the frame length is set to 4 times the median pitch period. Furthermore, the normalized autocorrelation function is computed for each frame. If there is no pitch variation within a frame, its normalized autocorrelation function should be a damped sinusoidal response with a straight line envelope. The area under the ideal envelope (without pitch variation) should be the same for each frame for a specified vowel, that is, $N/2$, where N is the frame length. In the case when pitch variation is present in a frame, its normalized autocorrelation envelope will not be an ideal straight line, and hence the area under the envelope will not be $N/2$. By computing the area under the normalized autocorrelation envelope and normalizing it by $N/2$, we can obtain 4 normalized TEO autocorrelation envelope area parameters for each time frame (i.e., one for each frequency band). This represents the TEO-Auto-Env feature per frame.

3.3 TEO-Pitch: TEO based Pitch

Unlike the previous two features, the TEO-Pitch feature is a direct estimate of the pitch itself as opposed to a secondary parameter for representing frame-to-frame excitation variations. Since it is very difficult for current available techniques to correctly detect pitch of speech under stress, especially under extreme stress, we first apply the

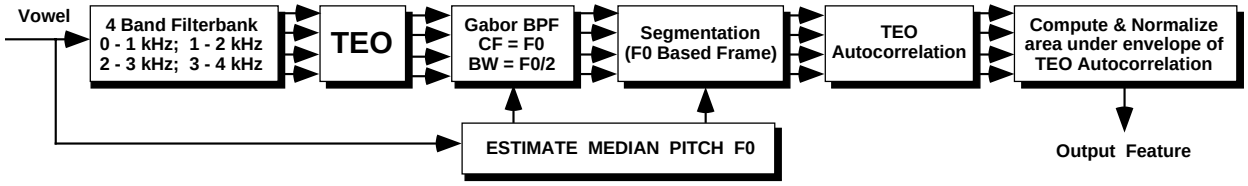

Figure 2: TEO-Auto-Env Feature Extraction

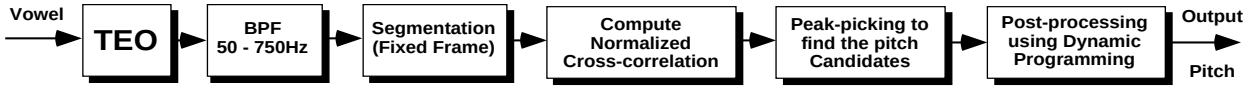

Figure 3: TEO-Pitch Feature Extraction

TEO to the raw vowel speech. As explained in Sec. 3.2, the TEO profile has the same periodicity as pitch. Furthermore, experiments determined that it generally showed better periodicity than raw stressed speech. Since we found that pitch usually falls within the extreme range of 50Hz to 750Hz (female speech from actual high stress can have pitch as high as 700Hz), the TEO profile is bandpass filtered over (50:750Hz). As shown in Fig. 3, after the BPF and segmentation, a normalized cross-correlation function (NCCF) and dynamic programming [10] is applied to detect the pitch structure. Here the waveform is first down-sampled, and candidate peaks in the NCCFs are determined. Subsequently, the peaks are fine-tuned by using the NCCF of the original waveform (before down-sampling). The candidate frame-based pitch periods are determined by the average distance of two neighboring peaks within that frame. Finally, dynamic programming is employed to decide the pitch of each frame.

4 Evaluations

In this study, evaluations were conducted using the SUSAS, *Speech Under Simulated and Actual Stress* database. SUSAS consists of five domains spoken under a wide range of stresses and emotions. In our experiments, the following subset of SUSAS words were used: "freeze", "help", "mark", "nav", "oh", and "zero". Angry, loud and Lombard styles were used for simulated stress (speakers were requested to speak in that style; $85dB$ SPL pink noise played through headphones was used to simulate Lombard effect). Data for actual stress were selected from the subject motion-fear domain. In the actual domain, a series of controlled speech data collection experiments were performed with speakers riding amusement park roller coaster rides. Background noise levels and stress levels were monitored during the completion of each ride. Since the TEO is more applicable for vowels than for consonants, only voiced vowel sections of all word utterances were used for evaluation. All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8 kHz. A baseline 5-state HMM-based stress classifier with continuous Gaussian mixture distributions was employed for the evaluations. For the purposes of comparison, the traditional pitch feature was used based on the algorithm proposed in [10].

4.1 Simulated Stressed Speech

For each stress style, we trained an HMM model for each word vowel using 18 tokens from 9 speakers. One neutral HMM model per word was trained using 18 neutral tokens; and 90 neutral tokens per word were used for pairwise testing between neutral and stress style trained HMMs. Since only 18 stressed tokens per word for each style are available, a round-robin method was employed for training and scoring. A total of 648 tokens were used for open test evaluation. The results are shown in Fig. 4. The results show that while traditional pitch characteristics are important for speech under stress, they are not consistent for stress classification, especially under milder forms of simulated stress. TEO based features were always more effective in neutral and stress speech classification, with improved performance for the TEO-Auto-Env and TEO-Pitch features.

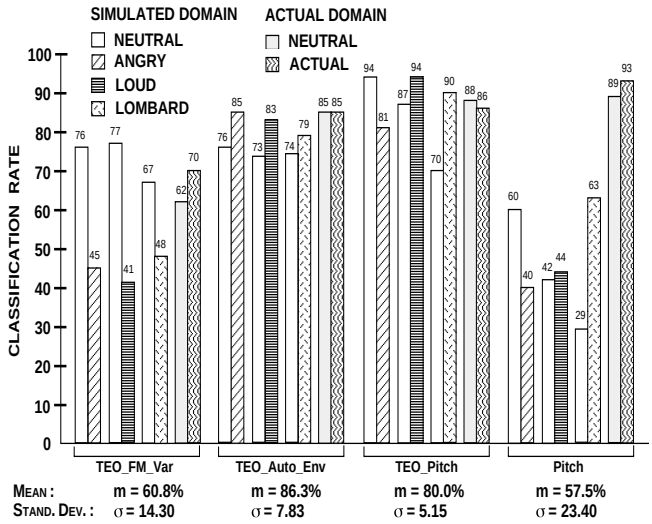

Figure 4: Pairwise Stress Classification Results (Mean and standard deviation of overall neutral/stress classification rates are shown)

4.2 Actual Speech Under Stress

Next, an evaluation was performed using the actual domain of SUSAS. For this evaluation, we used 7 speakers producing 20 tokens of "freeze", 9 tokens of "help", 16 tokens of "mark", 16 tokens of "nav", 15 tokens of "oh", and 18 tokens of "zero" for neutral and actual stressed conditions, respectively. A total of 188 tokens were used for open test evaluations. Since the speech data from the actual stress domain contained background noise, a previously formulated speech enhancement method was first applied as a pre-processing phase [11]. Round-robin training and scoring was employed for both neutral and actual data (pairwise

Sentence (from *Mayday2* of SUSC-0) (Extracted Phonemes Listed as "**I**")	Subjective Stress Level (1-10) (Listener Test)	Vowel Extracted for Use	HMM Likelihood Evaluation Score of Stress Level (neutral) "—" $\longleftrightarrow$ "+" (stress)			
			TEO-FM-Var	TEO-Auto-Env	TEO-Pitch	Pitch
Avionics l**I**ght hydraulic oil pressure light engine indicators are ...	2.32	/AY/	−0.1530	−1.8530	−1.4670	−2.9740
Okay give me imm**E**diate vectors this is an emergency I'm engine out	6.15	/IY/	0.0290	0.0330	−0.1100	0.0390
I'm h**O**t I need the cable ...	7.95	/AA/	−0.0320	0.2450	0.0070	−0.1400
M**A**n I thought I was gone	5.55	/AE/	−0.0120	−0.9080	1.5140	0.5300

Table 1: Evaluation Results of Stress Assessment Using SUSC-0 Database

test results shown in Fig. 4). Since the stress level of speech for the actual domain is far more severe, stress classification rates were generally higher. Performance for direct pitch feature was better than that under simulated stress, however, some human interaction was needed to ensure proper pitch estimates. The results for the three nonlinear TEO based features were comparable to that under simulated stress, with TEO-Auto-Env and TEO-Pitch features performing the best. These results suggest the consistency of the TEO features from simulated to actual speech under stress domains.

4.3 Evaluation with SUSC-0

In addition to the above evaluations with the SUSAS database, we evaluated stress assessment performance of the proposed features using SUSC-0 database. SUSC-0 is a stressed speech database from NATO, which consists of actual aircraft pilot communications under emergency situations. In this evaluation, we extracted 4 sentences from the *Mayday2* test portion. *Mayday2* contains speech data between a pilot and controller collected from the initial ground aircraft system check, through preliminary discovery of engine problems during flight, followed by emergency pilot actions to correct engine emergency, until safe resolution of the emergency. Table 1 summarizes the text and degree of stress obtained from a formal listener test of the selected 4 sentences. For the listener evaluation, 10 subjects rated the SUSC-0 speech utterances on a scale $(1:10)$, using neutral (1) and actual (10) stress tokens from SUSAS as listener judgment anchors. From each sentence, we extracted one vowel for the evaluation. Each feature discussed in Sec. 3 was extracted and submitted to the corresponding neutral and actual vowel HMM models of the word "help", trained for the previous evaluation in Sec. 4.2 (note vowel independence for this test). Here the actual stress HMM represents the speech under the extreme stress conditions. If a feature is able to assess the degree of stress regardless of the text, the difference between the likelihood scores from the actual stress and neutral HMMs should indicate the degree of stress. It is suggested that the level of stress will be more severe as the difference in HMM likelihood scores increases. The resulting score differences are shown in Table 1. The TEO-Auto-Env feature reflects the same stress level variation trend as observed in the listener test, While the others were less successful.

5 Conclusions

In this study, we proposed three new Teager Energy based nonlinear features, TEO-FM-Var, TEO-Auto-Env, and TEO-Pitch, for stress classification and assessment. TEO based features strive to reflect the variation in nonlinear airflow excitation of speech under stress. Evaluation results using the SUSAS database showed that the TEO-Pitch feature is the best for stress classification of loud, angry, and Lombard effect. Although traditional pitch can work very well for extreme stress, the TEO-Pitch feature performs more consistently, with overall stress classification rates of (Pitch: $m = 57.5\%, \sigma = 23.40$) vs. (TEO-Pitch: $m = 86.3\%$, $\sigma = 7.83$). In addition, evaluation results with NATO SUSC-0 database showed that the TEO-Auto-Env feature is the best for actual stress assessment, while TEO-FM-Var feature was less reliable. We conclude that the TEO-Auto-Env feature is a promising feature for both stress classification and assessment, and we believe its performance could be further improved by increasing the number of filterbank partitions to better reflect subtle energy changes across frequency for excitation.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] D. A. Cairns, J. H. L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions", *J. Acoust. Soc. Am.*, **96**(6):3392-3400, 1994.

[6] J. H. L. Hansen, B. D. Womack, "Feature Analysis and Neural Network Based Classification of Speech Under Stress" *IEEE Trans. Speech Audio Process.*, **4**(4):307-313, 1996.

[7] J. F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *ICASSP-90*, pp. 381-384, 1990.

[8] J. F. Kaiser, "Some Useful Properties of Teager's Energy Operator," *ICASSP-93*, vol. 3, pp. 149-152, 1993.

[9] P. Maragos, J. F. Kaiser and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Proc.*, **41**(10):3025-3051, Oct. 1993.

[10] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*, Edited by W. B. Kleijn and K. K. Paliwal, Elsevier Science, pp. 497-518, 1995.

[11] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Process.*, **39**(4):795-805, Apr. 1991.