Frequency Band Analysis for Stress Detection Using a Teager Energy Operator Based Feature*

Mandar A. Rahurkar, John H.L Hansen,
James Meyerhoff †, George Saviolakis †, Michael Koenig †

Robust Speech Processing Group
Center for Spoken Language Research
University of Colorado, Boulder, C0-80303
{rahurkar,jhlh}@cslr.colorado.edu

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

Abstract

Studies have shown that the performance of speech recognition algorithms severely degrade due to the presence of task and emotional induced stress in adverse conditions. This paper addresses the problem of detecting the presence of stress in speech by analyzing nonlinear feature characteristics in specific frequency bands. The framework of the previously derived Teager Energy Operator(TEO) based feature TEO-CB-AutoEnv is used. A new detection scheme is proposed based on weighted TEO features derived from critical bands frequencies. The new detection framework is evaluated on a military speech corpus collected in a Soldier of the Quarter (SOQ) paradigm. Heart rate and blood pressure measurements confirm subjects were under stress. Using the traditional TEO-CB-AutoEnv feature with an HMM trained stressed speech classifier, we show error rates of 22.5% and 13% for stress and neutral speech detection. With the new weighted sub-band detection scheme, detection error rates are reduced to 4.7% and 4.6% for stress and neutral detection, a relative error reduction of **79.1%** and **64.6%** respectively. Finally we discuss issues related to generation of stress anchor models and speaker dependency.

1. INTRODUCTION

The problem of detecting stress in speech has been the subject of a number of studies[1, 2, 3]. However, depending on the type of emotion or task induced stress condition, reliable detection of stress, even in clean speech, continues to be a challenging task. Reliable stress detection requires that a speaker change their neutral speech production process in a consistent manner so that extracted features can detect and perhaps quantify the change. Unfortunately, speakers are not always consistent in how they convey stress or emotion in speech, and therefore reliable detection typically requires a multi-dimensional solution. In the past, we have considered a variety of approaches to detect stress in speech based on pitch structure, duration, intensity, glottal characteristics, and vocal tract spectral structure using Hidden Markov Models(HMM) or Bayesian Hypothesis testing[1, 2, 3]. Here, we believe that a multidimensional feature obtained across a sub-band structure could be successful. Previously, we have formulated a number of novel nonlinear based features using properties of the Teager Energy Operator. While successful, there remain challenges to ensure consistency in stress detection for a given speaker known to the training set, as well as independent speakers not included within available training data.

The ability to detect stress in speech has many applications in voice communications such as increasing the robustness of speech recognition algorithms, military voice applications and law enforcement. In this paper, we address the problem of stress detection using a nonlinear feature set. We assert that speech production variability caused by stress in speech should be a nonlinear phenomenon which has evidence based on nonlinear studies conducted by Teager. Recently, a new feature based on TEO, TEO-CB-AutoEnv [1] was studied and found to be more responsive to stress. Our focus here, is to explore if specific frequency bands are more sensitive to stress, independent of speaker, and whether these bands can be used to more effectively detect stress in speech. We begin with the introduction of the nonlinear feature followed by description of the experimental setup. We then talk about our new algorithm to ameliorate stress detection. In the end we discuss our future work followed by a conclusion.

2. TEAGER ENERGY OPERATOR

Historically, most approaches to speech modeling have taken a linear plane wave point of view. While features derived from such analysis can be effective for speech coding and recognition, they are clearly removed from physical speech modeling. Teager[5, 6] did extensive research on nonlinear speech modeling and pioneered the importance of analyzing speech signals from an energy point of view. He devised a simple nonlinear, energy tracking operator, for a continuous time signal $x(t)$ as follows:

$$\begin{aligned}\Psi_c[x(t)] &= \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \end{aligned} \quad (1)$$

and for a discrete-time signal $x(n)$ as:

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

where $\Psi[\cdot]$ is the Teager Energy Operator(TEO). These operators were first introduced systematically by Kaiser[7, 8].

*This research was funded by DARPA through SPAWAR under Grant No. N66001-00-2-8906

†Affiliated with Dept. of Neuroendocrinology, Div. Neuroscience, Walter Reed Army Institute of Research(WRAIR), Silver Spring,Maryland

The views of the authors do not purport to reflect the position of Department of the Army or the Department of Defense.

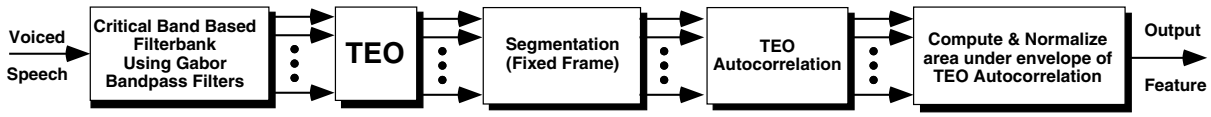

Figure 1: *Feature Extraction Flow Diagram.*

2.1. TEO-CB-Auto-Env:Critical Band Bases Teo Autocorrelation Envelope

Having established the discrete time TEO, a later study [13] produced an AM-FM parameterization(referred to as DESA-1 and DESA-2). These approaches motivate the use of TEO for general speech modelling.

It has been observed[1] that under stressful conditions, a speech signal's fundamental frequency will change and hence the distribution pattern of pitch harmonics across critical bands will be different then for speech under neutral conditions. Therefore, for finer resolution of frequencies, the entire audible frequency range can be partitioned into many critical bands [9, 10]. Each critical band possesses a narrow bandwidth,(i.e., typically 100-400Hz), thus making this new feature independent of the accuracy of median $F0$ estimation.

2.2. Analysis Across Frequency Bands

We can summarize the feature extraction procedure mathematically as follows using bandpass filters(BPF) centered at critical band frequency locations,

$$\begin{aligned} u_j(n) &= s(n) * g_j(n), \\ \Psi_j(n) &= \Psi[u_j(n)] = u_j^2(n) - u_j(n-1)u_j(n) + 1 \ , \\ R_{\Psi_j^{(i)}(n)}(k) &= \sum_{n=1}^{N-k} \Psi_j^{(i)}(n)\Psi_j^{(i)}(n+k), \end{aligned}$$

where,

$g_j(n), j = 1, 2, 3, ...16$, is the BPF impulse response,
$u_j(n), j = 1, 2, 3, ...16$, is the output of each BPF,
"$*$" is the convolution operator
$R_{\Psi_j^{(i)}(n)}(k)$ is the autocorrelation function of the ith frame of the TEO profile from the jth critical band,$\Psi_j^{(i)}(n)$,

and, N is the frame length.

Fig.1 shows a flow diagram of the feature extraction process. The TEO-CB-AutoEnv feature has been shown to reflect variations in excitation characteristics including pitch harmonics, due to its finer frequency resolution. However, we believe that the variation in excitation structure is not uniform across all the bands. Moreover, we also show that specific frequency bands are more sensitive to stress and some to neutral, independent of speaker, and hence we should be able to detect stress in speech without the need of a speaker dependent neutral or stress model as long as there are general reference models.

2.3. EXPERIMENTAL SETUP

2.3.1. Soldier of the Quarter(SOQ) Speech Corpus

A speech under stress corpus was collected at the Walter Reed Army Institute of Research. The speech corpus was constructed using the WRAIR Soldier of the Quarter Board paradigm[11, 12] , by recording the spoken response of 6 individual soldiers to questions in a neutral setting, as well as while seated in front of a seven person military evaluation board (all board members had military rank much above the soldier who faced the panel). The SOQ board is a training exercise and a competition used to prepare soldiers for actual promotion boards. Subjects in this study were candidates in the competition who volunteered to be studied after giving informed consent. Table.1 summarizes average speaker conditions for 6 speakers and 7 speech data collection phases before (A,B,C), during (D), and after (E,F,G) the Board. Changes in mean heart rate(HR), blood pressure(sBP,dSP) and pitch(f0) all confirm a change in speaker state between A,B,C,E,F,G and D. Results confirm a significant shift in biometric measures from the assumed neutral conditions (A,B,C),(E,F,G), versus the assumed stress condition (D). Each soldier was asked to answer all questions by responding "The answer to this question is NO". Each speaker was asked the same set of 6 different militarily-relevant questions on seven occasions. For our evaluations, we focused on the vowel 'o' extracted from the word 'NO'.

Summary Of Mean Biometrics for SOQ subjects					
Measure	A B -7Day	C -20min	D Board	E +20min	F G +7day
HR	70.3	70.8	**93.2**	69.5	67.2
sBP	118	146	**178**	154	117
dBP	77.5	74.8	**89.7**	71.2	69.5
f_o	103.4	102.7	**136.9**	104.3	103.1

Table 1: *HR - heartrate (in beats per minute), sBP - Systolic blood pressure in mm, dBP - Dystolic blood pressure in mm, f_o - Fundamental frequency in Hz.*

2.3.2. HMM Baseline Classification System

A baseline Hidden Markov Model(HMM) system was formed using SOQ corpora. Acoustic models consisted of three state HMM's each with two Gaussian mixtures. A total of 191 tokens were used for training the neutral model, while 30 tokens were used for training the stress model in a traditional round robin manner. The front-end feature consists of a sixteen dimensional TEO-CB-AutoEnv vector. The speech data obtained during the SOQ Board scenario was assumed to be "Stress" and the remaining speech data was grouped together as "Neutral" based upon biometric results. Thus, two HMM models termed "Neutral" and "Stress" result after the training phase. Using the entire critical band TEO-CB-AutoEnv feature, a round-robin open error classification rate was found to be 22.5% for stress and 13% for neutral.

2.3.3. HMM for Frequency Band Analysis

For frequency band analysis, a second HMM classification system was trained with a front-end feature made up of the TEO-CB-AutoEnv of each individual band, forming an independent system. A seperate Neutral and Stress model, was therefore constructed for every band. In addition to single band neutral and stress models, we also trained models using the first four bands(1-4), bands 5-8, and the last four bands(12-16) grouped together, which we believe will play an important role in distinguishing between neutral and stress speech. Thus, we have

thirty-two single band models, sixteen of which are neutral and sixteen under stress. We also have six four-band models again classified in a similar manner.

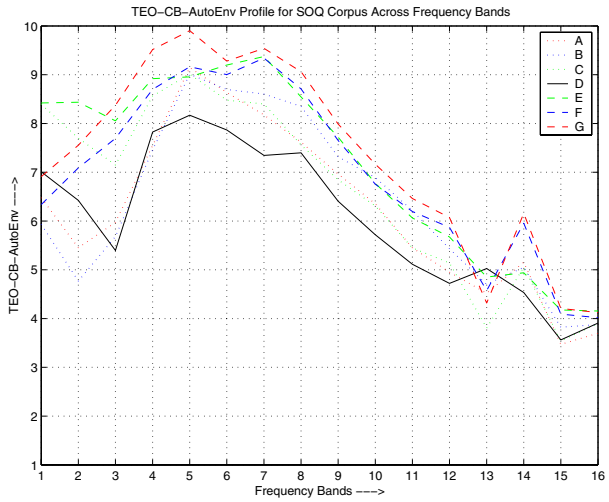

Figure 2: *Area Under Auto-Correlation Envelope across all 16 Bands.*

2.4. EXPERIMENTS

2.4.1. TEO Autocorrelation Envelope Analysis

In this initial experiment, we studied the area under the TEO autocorrelation envelope across sixteen frequency bands. The area under the auto-correlation envelope was calculated across all speakers and averaged for all sixteen bands. Fig.2, shows average feature profile before the board(A,B,C) and after the board(E,F,G). The continuous line represents the stress scenario(D).

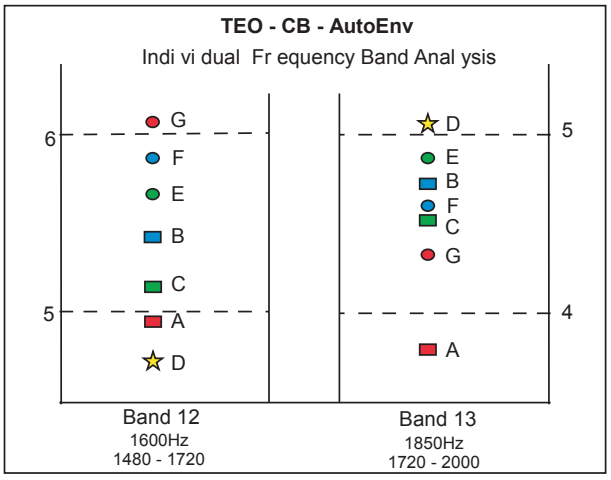

Figure 3: *Band 12 and 13 Zoomed.*

We observed that the area under the autocorrelation envelope for some bands is more distinct for some neutral and stress models. The TEO-CB-AutoEnv is lower in magnitude for low and mid-band frequencies (i.e., bands 3,5-7) for stress versus neutral. To better illustrate the result, band 12 and 13 have been shown in detail in Fig.3. For band 12,the stress condition D produced the lowest score, while for band 13 it was largest. These results strongly suggest a frequency dependent nature of TEO-CB-AutoEnv feature.

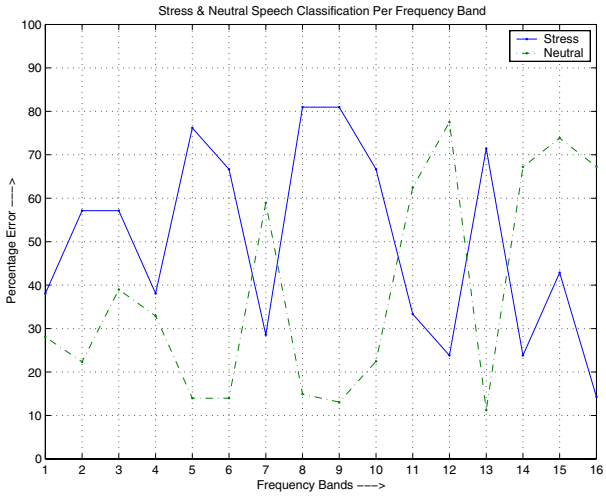

Figure 4: *Stress and Neutral speech Classification Results.*

Band	Stress	Neutral
1	38.09	28.04
2	57.14	22.35
3	57.14	38.99
4	38.09	32.91
5	76.19	13.97
6	66.67	13.97
7	28.57	58.94
8	80.95	14.92
9	80.95	13.07
10	66.67	22.46
11	33.33	62.46
12	23.81	77.54
13	71.43	11.21
14	23.81	67.22
15	42.86	73.84
16	14.29	67.25
1,2,3,4	42.86	15.87
5,6,7,8	47.62	24.15
13,14,15,16	57.14	8.37

Table 2: *Percentage Error rate in Stress/Neutral Recognition.*

2.4.2. Band Classification Sensitivity towards Neutral & Stress Speech

The results so far suggest a frequency sensitive nature for the TEO-CB-AutoEnv feature. Next, we want to determine if some bands are consistently more reliable in detecting neutral or stressed speech, and therefore we perform stress classification within each band. Fig.4 shows results for both stressed and neutral speech classification. We observe that bands 5, 8, 9 and 13 are sensitive to neutral speech (i.e., above 85% correct neutral classification), while bands 7, 12, 14 and 16 are sensitive to speech under stress (i.e., above 70% correct stress classification). Individual error rates for stress and neutral speech recognition are summarized in Table 2. Results are also shown for combined 1-4, 5-8, and 13-16 band-sets. Clearly, some com-

binations significantly outperform individual frequency bands for stress speech classification. Moreover we also observe that bands which are sensitive to stress are complementary to those sensitive to neutral. Note that all stress classification rates are based on single phonemes tests using /o/ in word 'no'.

3. NEW FEATURE FOR STRESS DETECTION

Here, we develop a novel scheme for stress detection based on the findings from the preceding section. We construct a weighted band scoring scheme in which each band is weighted, depending upon its sensitivity to stress or neutral, with the condition that all weights sum to unity. The weights used in the formulation were determined experimentally and the same set of weights were used for all evaluations in their respective categories (stress or neutral classification). In the future we will attempt to optimize the weight calculations using optimization algorithms on a large corpus. The equation below shows how an overall model score is obtained using stress and neutral sensitive bands :

$$Score = \sum_{n=1}^{4} W_{(n)} SNB_{(j)} - \sum_{n=1}^{4} W_{(n)} SSB_{(j)} \qquad (3)$$

where,
$\text{SNB}_{(j)}$ =Sensitive Neutral Bands: $j = 5, 8, 9, 13$
$SSB_{(j)}$ =Sensitive Stress Bands: $j = 7, 12, 14, 16$
$W_{(n)}$ =band 'n' Weight.

The result of evaluations using the new detection scheme are shown in Table.3. Using the entire TEO-CB-AutoEnv frequency range for the feature, baseline stress and neutral error rates are 22.5% and 13%. With results from experiments in section 2.4 to establish stress and neutral sensitive bands, the new weighted algorithm is able to achieve error rates of **4.7%** and **4.6%** for stress and neutral detection respectively. This corresponds to relative **79.1%** reduction in the stress speech detection error rate, and a **64.6%** relative reduction in the neutral speech detection error rate.

System	% Error in Stress	% Error in Neutral
Baseline	22.5	13
Weighted CB	**4.7**	**4.6**

Table 3: *Evaluation using New Detection Scheme*

4. DISCUSSION

The resulting weighted TEO-CB-AutoEnv detection scheme has resulted in a substantial improvement in stress and neutral speech detection. However, there are a number of important issues that remain to be addressed. For example, in our evaluations, we trained and tested with a closed speaker population. In many situations, it is not possible to have prior training data in both neutral and stressed conditions. Therefore, how well will models of speech under neutral and stress for a given speaker group, be applicable to new speakers ? We performed a test using the present corpus and found neutral detection rates to vary between 77.7-97.2%. However, because of limited number of speakers, there was insufficient test data tokens to obtain results for stressed speech detection for individual open speakers. We believe that it will be more important to determine performance for stress detection versus neutral detection, since from our earlier work [4], we believe there will be more speaker variability across speakers under stress. Another issue to raise is that all detection experiments were performed on a single vowel. Our earlier results suggest more consistent performance using a set of vowels from complete sentences. We will consider this in the future with a larger speaker corpus. The proposed solution here is well positioned to address the increased speaker question, or unseen speaker question, since we have independent weights for frequency sensitive bands in stress and neutral conditions. This issue will also help address the problem of anchor model construction for new speakers.

5. CONCLUSION

In this paper we have proposed a novel algorithm for stressed speech detection.This approach was based on nonlinear analysis using features derived from the Teager Energy Operator(TEO). Speech data under a SOQ paradigm obtained from WRAIR, independently showed a statistically significant change in blood pressure, heart rate and blood chemical composition between neutral and stress conditions. Individual stress detection experiments across critical sub-band frequencies showed some bands to be more sensitive for stress detection, while others were sensitive to neutral. Objective evaluations showed that this novel scheme leads to a substantial improvement in stress detection performance. We also discussed issues related to generation of stress anchor models and speaker dependency, however further evaluations will be needed on a larger population to determine if stress and neutral speech detection performance will hold for unseen speakers.

6. References

[1] G.Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress", *IEEE Trans. Speech & Audio Process.*, **9** (3):201-216, Mar. 2001.

[2] 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.

[3] 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.

[4] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," Speech Communications, vol. **20**(2), pp. 151-170, November 1996.

[5] H. Teager, "Some Observations on Oral Air Flow During Phonation", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.ASSP-28,No.5, pp. 599-601, Oct. 1990.

[6] H. Teager, S. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", *Speech Production and Speech Modeling*, NATO Advanced Study Institute, vol. 55, Kluwer Academic Pub., pp. 241-261, 1990.

[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, "On Teager's Energy Algorithm, its Generalization to Continuous Signals," in *Proc. 4th IEEE Digital Signal Processing Workshop,*,Sept 1990.

[9] B. Scharf, "Critical Bands", in *Foundations of Modern Auditory Theory,* Edited by J.V. Tobias, Academic Press, Vol. 1, pp. 157-202, 1970.

[10] W.A. Yost, "Fundamentals of Hearing", 3rd Edition, Academic Press, pp. 153-167,1994.

[11] Meyerhoff, J.L., Oleshansky, M.A. and Mougey, E.H. Psychological stress increases plasma levels of prolactin, cortisol and POMC-derived peptides in man. Psychosomatic Medicine 50(3): 295-303, 1988.

[12] Oleshansky, M.A. and Meyerhoff, J.L. Acute catecholaminergic responses to mental and physical stressors in man. Stress Medicine 8:175-179, 1992.

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