Improved "TEO" Feature-based Automatic Stress Detection Using Physiological and Acoustic Speech Sensors

Evan Ruzanski[1], John Hansen[1], Don Finan[1], James Meyerhoff[2], William Norris[3], Terry Wollert[3]

[1] Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado at Boulder, Boulder, CO USA
[2] Department of Applied Neurobiology, Division of Psychiatry and Neuroscience,
Walter Reed Army Institute of Research (WRAIR), Silver Spring, MD USA
[3] Federal Law Enforcement Training Center, Glynco, GA USA

Abstract

The acoustic pressure microphone has served as the primary instrument for collecting speech data for automatic speech recognition systems. The acoustic microphone suffers from limitations, such as sensitivity to background noise and relatively far proximity to speech production organs. Alternative speech collection sensors may serve to enhance the effectiveness of automatic speech recognition systems. In this study, we first consider an experimental evaluation of the TEO-CB-AutoEnv feature in an actual law enforcement training scenario. We consider feature relation to stress level assessment over time. Next, we explore the use of the physiological microphone, a gel-based device placed next to the vocal folds on the outside of the throat used to measure vibrations of the vocal tract and minimize background noise, as we investigate the effectiveness of a TEO-CB-AutoEnv-based automatic stress recognition system. We employ both acoustic and physiological sensors as stand-alone speech data collection devices as well as consider both sensors concurrently. For the latter, we devise a weighted composite decision scheme using both the acoustic and physiological microphone data that yields relative average error rate reductions of **32%** and **6%** versus sole employment of acoustic and physiological microphone data, respectively, in a realistic stressful environment.

1. Introduction

Reliable stress detection can be used to enhance the performance and robustness of speech recognition systems used in spoken dialog systems, cognitive task assessment, and spoken document retrieval, among other applications. Stress detection is also important in stand-alone applications, such as automatic assessment of stress levels of personnel in critical positions, such as pilots, air traffic controllers, and security personnel, allowing decisions to be made regarding the suitability of such persons to adequately perform their duties and maintain the safety of others.

The acoustic pressure microphone (A-MIC) has long been the standard sensor to collect speech data for use in automatic speech recognition systems. Advances in electrical engineering and acoustics have lead to the development as well as convenient and potentially ubiquitous implementation of alternative sensors for speech data collection, whose design and construction of lead to acoustic characteristics that vary from the traditional A-MIC. We will show in this paper that these unique characteristics can serve to improve the performance of automatic stress recognition systems.

This study considers employing the physiological microphone (P-MIC), a hydrogel/hydrophone-based throat contact sensor placed around the neck at the location of the vocal folds. This sensor, already proposed for use in high-noise military environments [1], is used in this study in a realistic automatic stress detection experiment involving speech as the medium to determine speaker condition (i.e. speaker is or is not under stress) [2]. Speech under the stressful condition is recorded using both A-MIC and P-MIC sensors while the speaker participates in rides at an amusement park chosen for their high level of perceived danger that elicited a stressful response in the speaker [3]. The non-stress condition speech is recorded immediately before and after the rides.

The P-MIC device is believed to be more robust to background noise, which can degrade stress detection performance. This sensor also is believed to more accurate in measuring air turbulence around the vocal folds, an attractive characteristic when using a stress detection system based on the non-linear Teager Energy Operator (TEO) feature. For these reasons, we will show an automatic stress detection system using the P-MIC, which can outperform an A-MIC-based system by 28%.

We will then outline a system using data from both sensors that yields a 32% relative improvement over the system using solely the A-MIC-collected data. This system uses a confidence-score-weighting-based scheme using the Manhattan distance metric between Hidden Markov Model (HMM) scores from each model trained using features collected from each respective sensor (i.e. a greater absolute difference between HMM scores implies greater confidence in the correct decision of speaker condition for a given sensor). These weighted scores are then combined between sensors to arrive at a final decision on the speaker state. This algorithm yields a substantial performance enhancement for automatic stress detection under the actual stress test conditions described above and presents a significant step towards reliable stress detection for spontaneous, unrestricted, conversational speech.

2. Teager Energy Operator-based Stress Classification

2.1 Introduction to the TEO-CB-AutoEnv Feature

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, they are clearly removed from physical speech modeling. Teager [4, 5] did extensive research on non-linear speech modeling

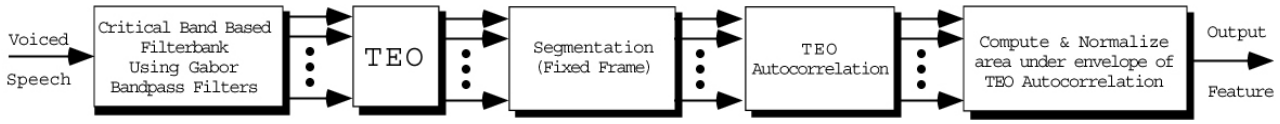

Figure 1. TEO Feature Extraction Flow Diagram

and pioneered the importance of analyzing speech signals from an "energy" point-of-view. His studies showed that airflow is separated, with concomitant vortices distributed throughout the vocal tract. The differences in linear vs. non-linear vocal tract airflow modeling is illustrated in Figure 2.

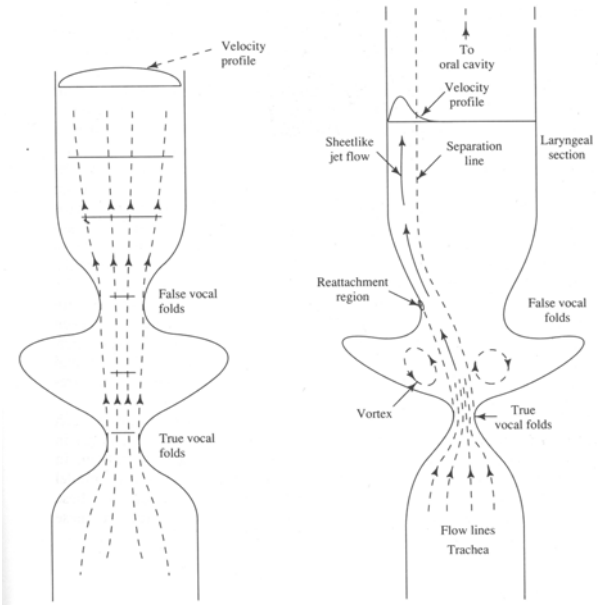

Figure 2. Linear (Left) vs. Non-linear (Right) Vocal Tract Airflow Model

It is believed that when a speaker is under stress, a change occurs in the vocal system physiology that further affects vortex-flow interaction patterns.

Teager devised a simple nonlinear energy-tracking operator that models the airflow through the vocal tract, shown mathematically for discrete-time signals as follows:

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

where $\Psi[\cdot]$ is the Teager Energy Operator (TEO). Kaiser first systematically introduced the TEO in [6, 7].

The Teager Energy Operator, Critical Band, Autocorrelation Envelope (TEO-CB-AutoEnv) feature employed here has been shown to reflect variations in excitation under stressful conditions [8]. A speech signal's fundamental frequency will change and hence the distribution pattern of pitch harmonics across critical bands will be different for speech under non-stressful conditions [8]. This finer frequency resolution comes from partitioning of entire audible frequency range into critical bands [9, 10].

The TEO-CB-AutoEnv is extracted through a process shown in the flow diagram of Figure 1 and illustrated mathematically using critical bandpass filters (BPF) as,

$$u_j(n) = s(n) * g_j(n) \qquad (2)$$

$$\Psi_j(n) = \Psi[u_j(n)] = u_j^2 - u_j(n-1)u_j(n+1) \qquad (3)$$

$$R_{\Psi_j^{(i)}(n)}(k) = \sum_{n=1}^{N-1} \Psi_{j(n)}^{(i)} \Psi_{j(n)}^{(i)}(n+k) \qquad (4)$$

where,

$g_j(n), j = 1,2,...,17$ is the bandpass filter impulse response,

$u_j(n), j = 1,2,...,17$ is the output of each bandpass filter,

"$*$" denotes the convolution operator,

$R_{\Psi_j^{(i)}(n)}(k)$ is the autocorrelation function of the i[th] frame of the TEO profile from the j[th] critical band, $\Psi_j^{(i)}(n), j = 1,2,...,M$ and N is the frame length.

Since the TEO models airflow through the vocal tract, as shown in the above equations and as illustrated in Figure 2, it is assumed that a sensor placed closer to the vocal folds (e.g. P-MIC) will yield data creating a more accurate vocal excitation model versus data collected from a sensor placed further from the vocal folds (e.g. A-MIC). It is also assumed that a more accurate speech production model will lead to better system performance. Finally, since the P-MIC and A-MIC are expected to sample different aspects of speech production, a question might be, "Can we use data from both sources to further improve stress detection performance?" Before exploring an answer to this question, we consider A-MIC-collected TEO-CB-AutoEnv feature behavior for an actual stressful scenario with variable stress levels.

2.2 A-MIC Probe Study: Relation of the TEO-CB-AutoEnv Feature to the Level of Stress in Speech

Speech data used for this study was collected from an actual training scenario at the Federal Law Enforcement Training Center (FLETC) in Glynco, GA. An A-MIC was used to record a single male trainee as he completed a simulated hostage negotiation scenario. As this scenario became progressively more hostile, a gunfight began with "simunitions" (9mm paint munitions). This resulted in elevated stress levels as quantified via psychometric, physiological, and stress hormone indices [11]. The scenario included simulated casualties.

The TEO-CB-AutoEnv was analyzed for an /AE/ vowel token under a clear neutral and stress condition. This comparison is shown in Figure 3. Figure 3 shows the TEO-CB-AutoEnv values vary for each of the 17 frequency bands (ranging from 0.035 to 0.527) and are generally higher for the token under the neutral condition.

To investigate the behavior of the TEO-CB-AutoEnv feature over time in the FLETC variable-stress level scenario, we choose the 4 frequency bands that show the greatest difference between neutral and stress conditions from the analysis of Figure 3. Here, these bands are 3, 7, 9, and 10.

We manually extracted 9 sentences at various times in the FLETC scenario, and extracted 3 vowels from each of these sentences. We found average values for the TEO-CB-AutoEnv for each of the 4 bands above across each of the 3

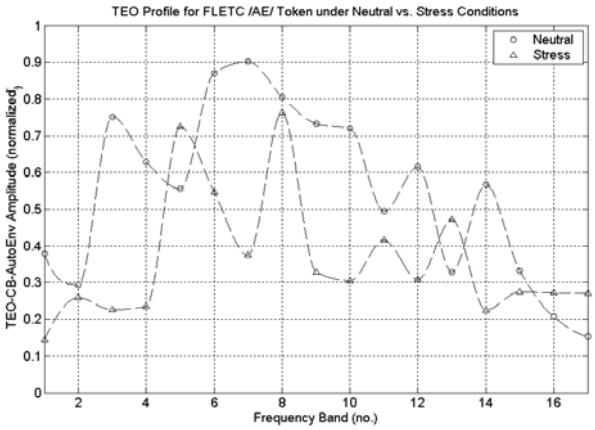

Figure 3. TEO-CB-AutoEnv Profile for /AE/ Token, Neutral vs. Stress Condition

vowels in each sentence. Finally, we averaged the average values for each of the 3 vowels to yield one value representative of the behavior of the TEO-CB-AutoEnv at each time. The results are shown in Figure 4.

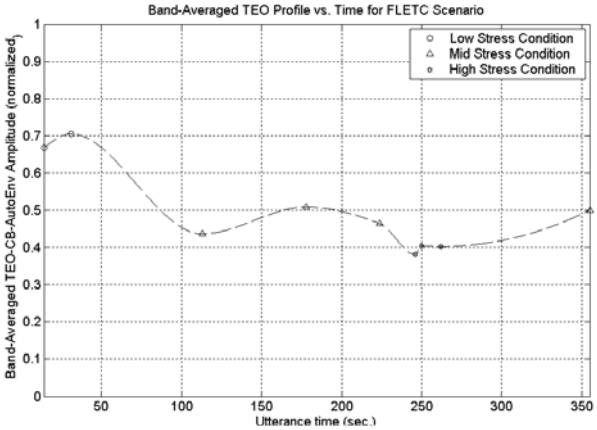

Figure 4. Band-Averaged TEO-CB-AutoEnv Profile for FLETC Scenario vs. Time

Figure 4 shows that the band-averaged TEO-CB-AutoEnv values across various phonemes in sentences spoken at different times in a stressful scenario are inversely proportional to the level of stress the speaker is under at that particular time.

3. A-MIC/P-MIC Experimental Preparation

The speech corpus used in this experiment consists of repetition of a set of eleven words taken from the Speech Under Actual and Simulated Stress (SUSAS) [3] corpus spoken by one male American speaker under non-stress (i.e. neutral) and stress conditions. The SUSAS corpus also contains speech data from roller coaster rides, but used only an A-MIC signal. In the present study, the speaker was fitted with an acoustic microphone approximately 1-2 inches from the opening of the mouth, but also employed a physiological microphone fitted around the neck at the location of the true vocal folds. The set-up is shown in Figure 5.

The stress speech data was collected while the speaker participated in several amusement park rides, chosen for their high speed and rapid change of direction, inducing the perception of danger. The neutral speech data was collected while the speaker was stationary immediately before and after the rides.

The speech data was collected using a digital audio tape (DAT) recorder, segmented, manually extracted, digitized at a 44.1kHz-sampling rate and downsampled to 8kHz for analysis. The segmentation portion of the process consisted of separating neutral and stress data, whose classification was readily apparent through informal listener tests.

Figure 5. Speaker Set-up, A-MIC and P-MIC

It has been determined that vowels are an attractive class of phonemes to use as tokens in such stress recognition systems due to their definite quasi-periodic nature [8]. As such, the following vowels were extracted from utterances in neutral and stressful conditions: /EY/, /IY/, /IH/, and /OW/. Each set of words contained 5, 3, 1, and 2 occurrences of the /EY/, /IY/, /IH/, and /OW/ phonemes, respectively. A total of 374 phonemes were extracted, 187 under neutral conditions and 187 under stressful conditions, and verified manually to be of suitable duration [12]. Three-mixture, three-state, multi-phoneme Hidden Markov Models (HMMs) were trained using the combination of these vowel tokens and tests were conducted in "round-robin" format as in [12].

4. Performance Evaluation

The performance evaluation for this study consists of 2 parts: evaluation of automatic stress detection performance using the A-MIC and P-MIC sensors separately in Section 4.1 and the employment of a weighted composite decision scheme using both sensors concurrently in Section 4.2.

4.1 Comparison of A-MIC Versus P-MIC Performance

The average combined error rates (i.e. (neutral + stress error rate) / 2) for the matched train-to-test tokens collected from the A-MIC and P-MIC are shown in Table 1. We see that the P-MIC far outperforms the A-MIC for combined neutral/stress classification.

Table 1. Automatic Stress Detection Performance Using Separate Acoustic and Physiological Microphone-Collected Speech Data

Sensor	Percent Error	Relative Improvement
A-MIC	**20.06**	**28%**
P-MIC	**14.44**	

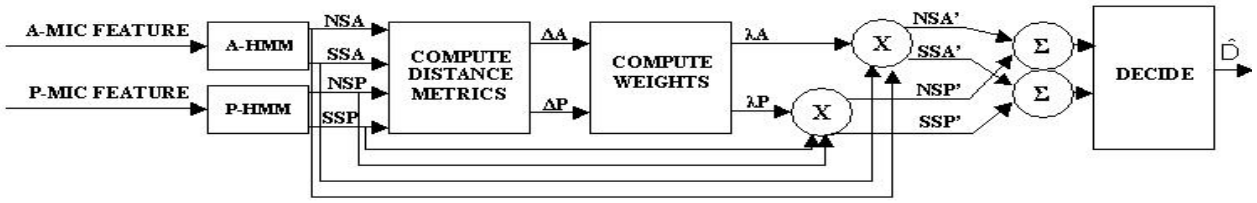

Figure 6. Weighted Composite Decision Detection Scheme Flowgraph

4.2 Novel Weighted Composite Decision Scheme for Improved Automatic Stress Detection

Since the speaker considered in this experiment was fitted with both sensors while speaking, we employ a stress detection decision scheme that uses data from both sensors. The diagram for a weighted composite decision process is shown in Figure 6.

Features are extracted from both A-MIC and P-MIC speech data and submitted to HMMs trained on A-MIC and P-MIC features, respectively, yielding a matched train-to-test scenario and the decision for each speaker condition is made according to the following equation:

$$\hat{D}=\begin{cases}\text{"}Neutral\text{"} \ if \ (NSA'+NSP') \geq (SSA'+SSP') \\ \text{"}Stress\text{"} \ if \ (SSA'+SSP') \geq (NSA'+NSP')\end{cases} \quad (5)$$

where,

$$NSA' = (\lambda A)(NSA) \quad (6)$$

$$NSP' = (\lambda P)(NSP) \quad (7)$$

$$SSA' = (\lambda A)(SSA) \quad (8)$$

$$SSP' = (\lambda P)(SSP) \quad (9)$$

$$\lambda A = \frac{\Delta A}{\Delta A + \Delta P} \quad (10)$$

$$\lambda P = \frac{\Delta P}{\Delta A + \Delta P} \quad (11)$$

$$\Delta A = |NSA - SSA| \quad (12)$$

$$\Delta P = |NSP - SSP| \quad (13)$$

where we define the terms,

$NSA \triangleq$ Neutral score from HMM trained on A-MIC tokens,

$SSA \triangleq$ Stress score from HMM trained on A-MIC tokens,

$NSP \triangleq$ Neutral score from HMM trained on P-MIC tokens,

$SSP \triangleq$ Stress score from HMM trained on P-MIC tokens.

The result for this scheme with comparison to previous results is shown in Table 2. We see the combination of P-MIC and A-MIC-based stress classification further improves overall system performance.

Table 2. Automatic Stress Detection Performance Using Weighted Composite Decision (WCD) Scheme

Sensor	Percent Error	Rel. Change vs. A-MIC	Rel. Change vs. P-MIC
WCD	**13.63**	**-32%**	**-6%**

5. Results and Conclusions

In this study, we considered the use of a P-MIC as an alternative speech sensor for automatic stress detection. It was shown that using the P-MIC as the alternative sensor resulted in a relative performance gain of 28% using speech data from amusement park roller coaster rides. Furthermore, it was shown that using speech data from both the A-MIC and P-MIC sensors in a composite weighted decision scheme yielded an additional 6% performance improvement.

These results suggest the effectiveness of using speech data collected from a P-MIC sensor when used for automatic stress detection in a TEO feature-based system. The close proximity to the excitation vocal tract airflow that accurately reflects change in vortex-flow interaction of air in the vocal system from neutral to stressful conditions, and robustness to background noise are possible reasons for this gain.

Our probe study suggests that the TEO-CB-AutoEnv-based stress classification system provides a potential measure for stress level over time. Future research could consider stress detection in a similar scenario as presented here across a range of speakers, employing additional sensors (e.g. GEMS-MIC) in conjunction with the sensors used here, and considering a range of stressful speaker environments.

6. References

[1] Scanlon, M.V., "Acoustic monitoring pad for combat casualty care," Army Science Conf. Proc., June 1996.

[2] Finan, D.S., Hansen, J.H.L., "Toward a Meaningful Model of Speech Under Stress", Conf. on Motor Speech Disorders, Motor Track Control, 2004.

[3] Hansen, J.H.L., "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, Vol. 20, pp. 151-173, Nov. 1996.

[4] Teager, H., "Some Observations on Oral Air Flow During Phonation", *IEEE Trans. Acoustics, Speech & Signal Proc.,* 28(5): 599-601, 1990.

[5] Teager, H., Teager, S. "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", *Speech Production and Speech Modeling,* NATO Advanced Study Inst., vol. 55, Kluwer Acad. Pub., Boston, 1990.

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

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

[8] Zhou, G., Hansen, J.H.L., Kaiser, J.F., "Nonlinear feature-based classification of speech under stress", *IEEE Trans. Speech & Audio Process.*, March 2001.

[9] Scharf, B., "Critical Bands", *Foundations of Modern Auditory Theory,* Tobias (Ed), Acad. Press, 1970.

[10] Yost, W., "Fundamentals of Hearing", 3rd Ed., Acad. Press, 1994.

[11] Meyerhoff, J., Norris, W., Saviolakis, G., et. al., "Evaluating Performance of Federal Law Enforcement Personnel During a Stressful Training Scenario", *Annals of the New York Acad. of Sci,* vol. 1032, pp. 250-253, 2004.

[12 Ruzanski, E., Hansen, J.H.L., Meyerhoff, J., et. al., "Effects of Phoneme Characteristics on TEO-based Automatic Stress Detection in Speech", ICASSP-05, pp. 357-360, 2005.