Stress Level Classification of Speech Using Euclidean Distance Metrics in a Novel Hybrid Multi-Dimensional Feature Space

Evan Ruzanski[1]*, John H.L. Hansen*[1]
James Meyerhoff[2]*, George Saviolakis*[2]*, William Norris*[3]*, Terry Wollert*[3]

[1] Robust Speech Processing Group, Center for Spoken Language Research,
University of Colorado, Boulder, CO, USA
{ruzanski, jhlh}@cslr.colorado.edu

[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

Presently, automatic stress detection methods for speech employ a binary decision approach, deciding whether the speaker is or is not under stress. Since the amount of stress a speaker is under varies and can change gradually, a reliable stress level detection scheme becomes necessary to accurately assess the condition of the speaker. Such a capability is pertinent to a number of applications, such as for those personnel in law enforcement positions. Using speech and biometric data collected from a real-world, variable-stress level law enforcement training scenario, this study illustrates two methods for automatically assessing stress levels in speech using a hybrid multi-dimensional feature space comprised of frequency-based and Teager Energy Operator-based features. The first approach uses a nearest neighbor-type clustering scheme at the vowel token level to classify speech data into one of three levels of stress, yielding an overall error rate of 50.5%. The second approach employs accumulated Euclidean distance metric weighting at the sentence-level to yield a relative improvement of **12.1%** in performance.

1. Introduction

Reliable stress level classification 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 level classification could also be important in stand-alone applications, such as automatic assessment of stress levels of personnel in critical positions (e.g., pilots, air traffic controllers, and security personnel) allowing decisions to be made regarding the suitability of these persons to adequately perform their duties.

Current automatic stress detection methods for speech fail to account for graduated stress levels of the speaker conveyed through speech [1, 2, 3]. This paper presents two approaches for reliably assessing the stress level conveyed through speech. The first method makes decisions on the individual phoneme token level and classifies test tokens into one of 3 stress levels: "low", "mid", or "high". The second method operates on the same basis but makes decisions at the sentence-level. Both show effective performance, with error rates of 50.5% and 44.4%, respectively.

These methods establish centroids for each stress level, and Euclidean distance metrics [4] in an orthogonal three-dimensional feature space "score" the stress level. This "hybrid" space consists of frequency- and Teager Energy Operator (TEO)-based features. Mean pitch and duration features are used as 2 dimensions in this space, as they have been shown to increase when the speaker is under stressful conditions [5]. A new non-linear TEO-based feature, the $\overline{\Delta_{TEO}}$, comprises the third dimension of this space and represents the average value of critical bands 3 and 9 (i.e., 200-300 and 920-1080 Hz, respectively) of the Teager Energy Operator, Critical Band, Autocorrelation Envelope (TEO-CB-AutoEnv). It shows consistent decrease as the level of stress increases.

A simulated hostage scenario used for training at the Federal Law Enforcement Training Center (FLETC) was utilized to generate the speech corpus analyzed in this experiment.

2. The Federal Law Enforcement Training Center simulated hostage scenario

Each of 10 male fully trained students, ranging in age from 24-35 years old, gave written, informed consent to participate as experimental subjects in the two-segment training scenario.

After a baseline period, subjects were paired with a confederate of the experiments (introduced as "another student" who would serve as their duty "partner"). Following a high-speed driving segment, a second segment involved investigation of a domestic complaint. The speech data used in this study was extracted from this second segment, scripted to last 5 minutes to include comparable challenges at similar intervals for all subjects.

In the second segment, the student and his partner entered a house and took a report from a complainant, who claimed his roommate stole a large sum of money from him. The student and his partner, both wearing protective vests and face shields, were given 9mm semi-automatic pistols loaded with paint munitions (similar to paint balls). The third round in the student's weapon was a "dud" requiring the student to perform a clearing maneuver during the gunfight. Two instructors unarmed but wearing protective vests were inside the house and served as role players – one as a cooperative "complainant", the other as a "suspect". The portion of the segment during the complainant interview is defined as the "low" stress condition. The stress level was designed to increase to the "mid" level as the suspect emerged from a backroom and began shouting at the complainant escalating the risk. The "high" stress portion of the segment began as the partner approached the suspect and had his 9mm weapon taken away by the suspect, who promptly shot the partner and complainant and proceeded to fire at the student, who had minimal cover available. The scenario terminated with the suspect falling to the floor immobile after exchanging gunfire with the student.

1-4244-0469-X/06/$20.00 ©2006 IEEE

The resulting speech corpus used for this experiment shows a definite build-up of speaker stress over time as verified by trainee heart rate and blood pressure readings taken at regular intervals [6]. The behavior of the biometric data is shown in Fig. 1.

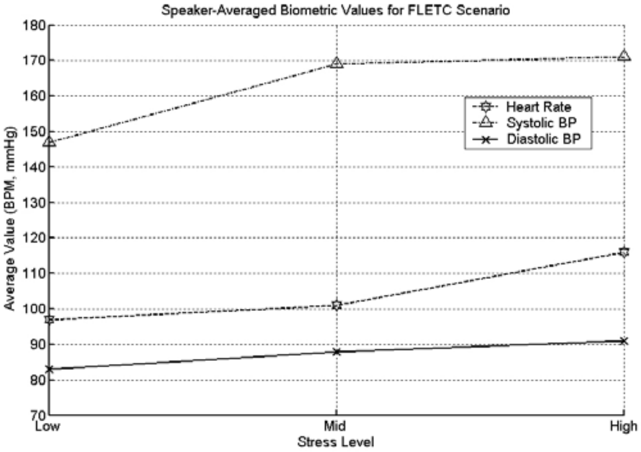

Fig. 1. Speaker biometric profile for FLETC scenario

Vowels are an attractive class of phonemes to use as tokens in automatic speech recognition systems due to their definite quasi-periodic nature [3]. There is little variance between vowel types used as tokens in an automatic TEO-based stress detection system [7]. A total of 756 vowel tokens of suitable duration [7] were extracted from 10 male speakers completing the FLETC scenario as subjects. The vowel set contains the phonemes /AA/, /AE/, /AO/, /AX/, /EH/, /EY/, /IH/, /IY/, /OW, and /UW/. The vowels were manually extracted. Further details are shown in Table 1.

Table 1. Speech data analysis for FLETC scenario

Parameter	Stress Condition		
	Low	Mid	High
# of Tokens	163	384	209
# of Sentences	41	84	53
μ #. of Tokens per Sentence	3.98	4.52	3.90
σ^2 of #. of Tokens per Sentence	2.82	2.63	2.00

Five tokens were extracted from each speaker during a "relative neutral" condition (i.e., conditions occurring approximately 30 minutes before the hostage scenario as subjects were speaking to a radio dispatcher prior to the beginning of the driving scenario). These tokens were used for normalization of the feature dimensions described in the next section.

3. Hybrid multi-dimensional feature selection and analysis

The features chosen for this experiment are hybrid (i.e., frequency-domain-based and non-linear) and multi-dimensional. The dimensions were normalized using respective values for true neutral speech pooled for all speakers. The three dimensions chosen for stress level assessment in this experiment, mean pitch, token duration, and $\overline{\Delta_{TEO}}$, are described below. The following analyses consider a set of the 10 speakers in the FLETC corpus.

3.1 Mean pitch

It has been shown that mean pitch increases significantly relative to a non-stress condition [5]. We chose to explore the validity of using mean pitch as a feature in the assessment of stress level in speech in the FLETC scenario. Frame-by-frame pitch values are extracted from each of the tokens using a sub-harmonic-to-harmonic ratio algorithm [8]. The mean value was then determined and the results tabulated for each of the 3 stress levels.

3.2 Vowel duration

It has been shown that vowel duration increases significantly for a speaker under stress [5]. We chose vowel duration as one feature to measure in the assessment of stress level in speech in the FLETC scenario. Duration is independent of mean pitch, assuming quasi-periodicity of the vowel tokens and the removal of the temporal component during the averaging operation. Pitch and duration dimensions are thus orthogonal.

3.3 TEO-CB-AutoEnv-based $\overline{\Delta_{TEO}}$ feature

Historically, most approaches to speech modeling have taken a linear plane wave point-of-view. Teager [9, 10] did extensive research on non-linear speech modeling 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. It is believed that when a speaker is under stress, a change occurs in the vocal system physiology that affects vortex-flow interaction.

Teager devised a nonlinear energy-tracking operator that models vocal tract airflow, shown mathematically as:

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{1}$$

where $\Psi[\cdot]$ is the TEO, first introduced by Kaiser in [11, 12].

The TEO-CB-AutoEnv reflects variations in excitation under stressful conditions [3]. A speech signal's fundamental frequency will change and the distribution of pitch harmonics across critical bands will differ for speech under non-stressful conditions [3].
This finer frequency resolution comes from partitioning of the entire audible frequency range into critical bands [13, 14].
The TEO-CB-AutoEnv is extracted through a process shown in the flow diagram of Fig. 2 and illustrated mathematically using critical bandpass filters as,

$$u_j(n) = s(n) * g_j(n) \tag{2}$$

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

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

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 ith frame of the TEO profile from the jth critical band, $\Psi_j^{(i)}(n),\ j = 1, 2, ..., M$ and N is the frame length.

The following experimental procedure is employed for stress level assessment. Two different vowel tokens are selected for each of two speakers from the corpus. For the first speaker, vowels

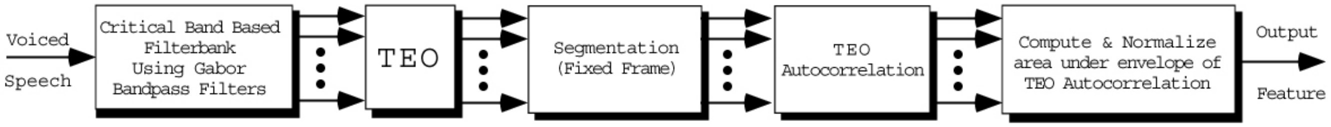

Fig. 2. TEO-CB-AutoEnv feature extraction flow diagram

/OW/ and /AY/ were selected. For the second speaker, vowels /IH/ and /AO/ were selected. The vowels were selected such that the same vowel was spoken during low, mid, and high stress conditions. The normalized TEO-CB-AutoEnv values were extracted for each band from each of the tokens in each of the 3 stress conditions and the difference taken from low-to-high and low-to-mid stress conditions. The analysis showed the bands that exhibit ubiquitous decreasing behavior across the 4 tokens across both speakers from the low to high stress conditions are bands 3 and 9 (200-300 Hz and 920-1080 Hz, respectively). We believe increased airflow turbulence energy levels are present within these 2 frequency bands, a claim that should be substantiated by further physiological research. By averaging the TEO-CB-AutoEnv values across bands 3 and 9, we create a new $\overline{\Delta_{TEO}}$ feature, shown mathematically as,

$$\overline{\Delta_{TEO}} \triangleq \frac{TCA_3 + TCA_9}{2} \qquad (5)$$

where

TCA_i, $i = \{3, 9\}$ is the TEO-CB-AutoEnv value for the ith band.

By differencing this value from unity, there is consistent quasi-linear increasing behavior of each feature with respect to increasing stress level. This is shown in Fig. 3.

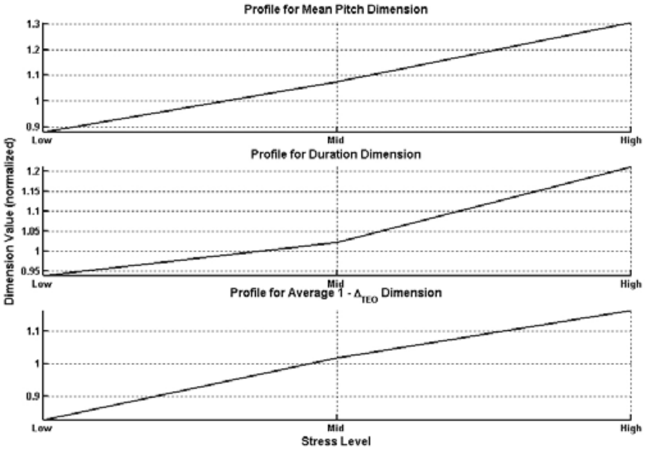

Fig. 3. Centroid plot for hybrid 3-D feature space (f_o, duration, $\overline{\Delta_{TEO}}$, shown respectively)

4. Stress level classification schemes

We began with the establishment of centroids encompassing the 3 dimensions explained in the previous section. It was determined experimentally that use of the 3-dimension feature space yielded better results than any of the other 6 combinations of lower-dimensional class combinations. The available speech data was collected and the norms of the 3 dimensions are computed for each token. Subsets of the lowest and highest 1/3 of this data are created and the resulting centroids become the low-stress and high-stress level centroids, respectively. Based on the consistent quasi-linear behavior of each of the 3 dimensions, the mid-stress level centroid is taken to be the mean of the low- and high-stress level centroids. The selection of 1/3 of the data was determined to give the best experimental results for the partitioned stress-level space.

4.1 Token-level classification

The establishment of stress level centroids and test set is performed in a "round robin" fashion establishing a speaker-independent test scenario. A test token's class membership is determined by the minimum Euclidean distance [4] to the centroid of that class. The decision scheme for the token-level classification scheme is shown in Eq. 6:

$$\hat{D}_T = \begin{cases} \text{"Low", if } \|x - C_{LOW}\| = \min\left(\|x - C_{LOW}\|, \|x - C_{MID}\|, \|x - C_{HIGH}\|\right) \\ \text{"Mid", if } \|x - C_{MID}\| = \min\left(\|x - C_{LOW}\|, \|x - C_{MID}\|, \|x - C_{HIGH}\|\right) \\ \text{"High", if } \|x - C_{HIGH}\| = \min\left(\|x - C_{LOW}\|, \|x - C_{MID}\|, \|x - C_{HIGH}\|\right) \end{cases}$$

where (6)

$\hat{D}_T$ is the token-level stress level classification decision,

x is the 3-dimensional feature vector for the test token,

$C_{\{LOW,MID,HIGH\}}$ is the 3-dimensional {"Low Stress", "Mid Stress", "High Stress"} centroid.

4.2 Sentence-level classification

It has been shown that utilizing a weighted majority rule decision algorithm yields considerable improvement for automatic stress detection over token-level algorithms [7]. We next employ a sentence-level classification scheme using accumulated Euclidean distances for each vowel token within a given sentence.

The stress level centroids are established and speaker-independent testing is conducted in a "round robin" fashion as was used with the token-level scheme. The stress classification decision is made according to Eq. 7.

$$\hat{D}_S = \begin{cases} \text{"Low", if } \sum_{i=1}^{N} W_{LOW,i} = \min\left(\sum_{i=1}^{N} W_{LOW,i}, \sum_{i=1}^{N} W_{MID,i}, \sum_{i=1}^{N} W_{HIGH,i}\right) \\ \text{"Mid", if } \sum_{i=1}^{N} W_{MID,i} = \min\left(\sum_{i=1}^{N} W_{LOW,i}, \sum_{i=1}^{N} W_{MID,i}, \sum_{i=1}^{N} W_{HIGH,i}\right) \\ \text{"High", if } \sum_{i=1}^{N} W_{HIGH,i} = \min\left(\sum_{i=1}^{N} W_{LOW,i}, \sum_{i=1}^{N} W_{MID,i}, \sum_{i=1}^{N} W_{HIGH,i}\right) \end{cases}$$

where (7)

$\hat{D}_S$ is the sentence-level stress level classification decision,

$W_{\{LOW,MID,HIGH\},i} = \left\| x - C_{\{LOW,MID,HIGH\}} \right\|$ is the normalized distance from the centroids for the i^{th} vowel token in the test sentence,

x is the 3-dimensional feature vector for the test token,

N is the number of vowel tokens per test sentence, and

$C_{\{LOW,MID,HIGH\}}$ is the 3-dimensional {“Low Stress”, “Mid Stress”, “High Stress”} centroid.

4.3 Comparison of classification schemes

The performance of the token- and sentence-level stress level classification schemes is shown in Table 2.

Table 2. Comparison of stress-level classification schemes

Comparison of Automatic Stress-Level Detection Schemes (Percent-Error, with random-decision error rate = 66.7%)				
Classification Scheme	Stress Level			
	Low	Mid	High	**Overall**
Token-Level	17.2	62.7	50.8	**50.5**
Sentence-Level	14.6	53.6	52.8	**44.4**

The results presented in Table 2 show effectiveness of both schemes in classifying a vowel token at each stress level, as a “chance” decision yields an asymptotic error rate of 66.7%. As shown in [7], a weighted majority decision scheme yields improved performance, here a relative improvement of 12.1%.

The results presented in Table 2 raise other issues concerning the choice of paradigm for stress level assessment in speech. It is seen that the mid-stress level condition consistently yields worse performance than that for the low- and high-stress level conditions. It is believed more stress level graduations within the mid-stress region may improve performance.

5. Summary and conclusions

The ability to assess the stress level in speech is important for a number of applications involving personnel in critical positions. This paper presents a significant step towards optimally assessing the stress level conveyed through the speech of a speaker, with independent biometrics to confirm speaker state.

This paper illustrates the effectiveness of 2 approaches to effectively assess the levels of stress conveyed in speech using the FLETC simulated hostage scenario corpus, a real-world development constructed to present increasing stress stimulus to the participant over time as verified by biometric data. Both algorithms utilize vectors located in an orthogonal 3-dimensional space extracted from vowel tokens from the FLETC simulated hostage scenario. The dimensions comprising this space include mean pitch, token duration, and a new TEO-based non-linear feature, the $\overline{\Delta_{TEO}}$ feature. It was shown that these features exhibit ubiquitous and monotonic quasi-linear behavior over increasing stress levels.

The first classification scheme, based on the ability to categorize test tokens into 1 of 3 stress levels, “low”, “mid”, or “high”, showed effective results relative to “chance” decisions. Both token- and sentence-level classification schemes based on minimum Euclidean distance metrics and accumulated Euclidean distance-weighted metrics, respectively, showed this. The results shown reflect those results presented by speaker biometric analysis and present a significant step toward optimal stress level assessment in spontaneous, unrestricted speech.

The results present the issue of “quantization error” in the establishment of stress levels, with higher error rates in the “mid” stress section, suggesting decisions could “miss” towards both the “low” stress and “high” stress regions. The ability to accurately determine the locations and sizes of stress levels motivate an “assessment”-type paradigm where stress levels can be “tracked” over time. Additionally, the experimental observations leading to the selection of the $\overline{\Delta_{TEO}}$ feature for the feature space should be substantiated by further experimental and theoretical analysis of the TEO-profile irregularity in frequency bands 3 and 9.

6. Acknowledgements

Volunteers participated in this research only after giving their free and informed consent. Investigators adhered to AR 70-25 and USARMDC Reg. 70-50 on the use of volunteers in research. This work was supported by the FLETC and the U.S. Army MRMC. Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of U.S. Army. The authors wish to thank Val Atkins, Dr. Charles Speilberger, Bob Burge, and Sgt. Nadia Kendall-Diaz.

REFERENCES

[1] D.A. Cairns and J.H.L. Hansen, “Nonlinear analysis and detection of speech under stressed conditions,” *JASA.,* v. 6, pp. 3392-3400, 1994.

[2] M.A. Rahurkar, J.H.L. Hansen, M.A. Oleshansky, J.L. Meyerhoff, and M. Koenig, “Frequency Band Analysis for Stress Detection Using a Teager Energy Operator-based Feature,” in *Proc. of ICSLP-02,* Denver, CO, 2002, pp. 2021-2024.

[3] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, “Nonlinear feature-based classification of speech under stress,” *IEEE Trans. Speech & Audio Process.*, vol. 9, pp. 201-216, Mar. 2001.

[4] S. Theodoridis and K. Koutroumbas, *Pattern Recognition,* 2nd ed. San Diego: Elsevier Academic Press, 2003.

[5] J.H.L. Hansen, “Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition,” PhD dissertation, Dept. of Elect. Eng., Georgia Tech., Atlanta, GA, 1988.

[6] J.L. Meyerhoff, W. Norris, G. Saviolakis, T. Wollert, B. Burge, V. Atkins, and C. Spielberger, “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.

[7] E. Ruzanski, J.H.L. Hansen, J.L. Meyerhoff, M. Koenig, and G. Saviolakis, “Effects of Phoneme Characteristics on TEO-based Automatic Stress Detection in Speech”, in *Proc. of ICASSP-05*, Philadelphia, PA, 2005, pp. 357-360.

[8] X. Sun, "Pitch Determination and Voice Quality Analysis Using Subharmonic-to-Harmonic Ratio," in *Proc. of ICASSP-02*, Orlando, FL, 2002, pp. 333-336.

[9] H. Teager, “Some observations on oral air flow during phonation,” *IEEE Trans. Acous., Speech, Sig. Proc.,* v. 5, pp. 599-601, 1990.

[10] H. Teager and S. Teager, “Evidence for Nonlinear Production Mechanisms in the Vocal Tract,” *Speech Production and Speech Modeling,* NATO Adv. Study Inst., v. 55, Boston: Kluwer Acad. Pub., 1990.

[11] J. F. Kaiser, “On a Simple Algorithm to Calculate ‘Energy’ of a Signal”, ICASSP-90, pp. 381-384, 1990.

[12] J. F. Kaiser, “On Teager’s Energy Algorithm, its Generalization to Continuous Signal”, in *Proc. 4th IEEE DSP Workshop,* Sept. 1990.

[13] B. Scharf, “Critical Bands”, in *Foundations of Modern Auditory Theory,* edited by J. V. Tobias, Acad. Press, v. 1, pp. 157-202, 1970.

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