Detection of Speech Under Physical Stress: Model Development, Sensor Selection, and Feature Fusion

Sanjay A. Patil and John H.L. Hansen[1],

Center for Robust Speech Systems (CRSS),
Erik Jonsson School of Electrical Engineering and Computer Science,
University of Texas at Dallas
Richardson, TX 75080

{sanjay.patil, john.hansen}@utdallas.edu

Abstract

Speech system scenarios can require the user to perform tasks which exert limitations on his speech production/physiology thereby causing speaker variability and reduced speech system performance. This is speech under stress, which represents a speech different from speech under neutral conditions. The stress can be physical, cognitive or noise induced (Lombard). In this study, the focus is on physical stress, with specific emphasis on: (i) number of speakers used for modeling, (ii) alternative audio sensors, and (iii) fusion based stress detection using a new audio corpus (UT-Scope). We used a GMM framework with our previously formulated TEO-CB-AutoEnv features for neutral/physical stress detection. Second, stress detection performance is investigated for both acoustic and non-acoustic (P-MIC) sensors. Evaluations show that effective stress models can be obtained with 12 speakers out of a random size of 1-42 subjects, with stress detection performance of 62.96% (for close-talking mic) and 66.36% (for P-MIC) respectively. The TEO-CB-AutoEnv model scores were fused with traditional MFCC based stress model scores using the Adaboost algorithm, resulting in an improvement in overall system performance of 9.43% (absolute, for close-talking mic) and 12.99% (absolute, for PMIC) respectively. These three advances allow for effective stress detection algorithm development with fewer training speakers and/or alternative sensors in combined feature domains.

Index Terms: *Stress Detection, TEO-CB-AutoEnv, P-MIC, Speaker Variability.*

1. INTRODUCTION

Speech production variability is caused by a number of factors which include stress caused by cognitive load, physical task load, or exposure to background noise resulting in Lombard effect. Additional variability factors include accent, dialect, and language differences. These diverse factors degrade automatic speech system performance as well as human speech perception. We note that in adverse noisy, stressful situations where speech technology such as speech recognition, speaker verification, or dialog systems is used, addressing noise is not a sufficient goal to overcome performance loss. In noisy stressful scenarios, even if noise could be completely eliminated, the production variability brought on by stress, including Lombard effect, has a more pronounced impact on speech system performance.

We define the stress component as the external conditions (or environmental conditions) that impact speech production. For example, speech under noisy conditions (Lombard effect), speech while performing time-constraint tasks or decision-sensitive tasks like the decisions such as those taken air traffic controllers or aircraft pilots (Cognitive stress), or speech while under drugs or alcohol use (chemical stress), or speech under task environments such as G-force or roller-coaster rides, reflect the range of scenarios seen in speech under stress. We designate speech under emotion as a separate category since motivation for this speech change is generally under the speaker's control.

Previous research in this field has concentrated on stress classification, stress detection using the SUSAS corpus while more recently on other realistic conversational corpora including CU-Move (in-vehicle route navigation dialog), SOM (Soldier of the Month), FLETC corpus (police/military training scenario), and UT-SCOPE (speech under cognitive and physical stress conditions)[1-4,11-15]. The comprehensive feature domains for research have focused on speech production including: fundamental frequency, intensity, duration, formant locations, spectral slope, including an extensive range of features such as traditional MFCCs features and nonlinear TEO-based features [12]. More recent studies have concentrated on the use of voiced segment (vowels) and the effect of stress on speech/phoneme durations [4,11]. Furthermore, most have concentrated on the use of traditional acoustic microphones. For certain applications, employing an acoustic mic may hinder human task performance, and therefore the use of alternative sensors becomes necessary. For example, if an audio sensor can be fitted to the vocal system of a fire fighter, law enforcement officer, or a aircraft pilot, it will capture not only the speech signal but allow monitoring of their physical status including alertness [6,8]. No study / research has yet been done to investigate speaker population size in formulating effective stress models in conjunction with alternative audio sensors, which represent two goals of this study. We also consider a combination of previously formulated TEO-CB-AutoEnv features with traditional MFCC features for stress detection.

The motivation is to have a reliable stress detection system which can be adopted as front end of speech systems. The paper is organized as follows: first, we describe our corpus development for the study, followed by the stress detection algorithm using TEO-CB-AutoEnv features. We explore experiment analysis related to speaker size in acoustic modeling for stress based on close-talk data and P-MIC data. After analyzing the results for both, we present the fused system using the Adaboost algorithm to leverage the strengths of each method.

[1] This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029, and Univ. of Texas at Dallas under project EMMITT.

2. CORPUS

This study employs UT-Scope corpus for algorithm development and experimental evaluations. For corpus details please refer [16]. For each stress condition, the speaker was instructed to repeat the sentences prompts, while performing the task. The physical task consisted of using a stair-stepper for 10 minutes at a constant speed of between 9-11 miles per hour (digital readout display). Data is collected with three distinct audio sensors – close-talking mic (Shure beta-54), far-field mic (Shure) and P-MIC. It is noted that range of the study would be expanded to address gender, audio structure (voiced segments/consonants/silence), spontaneous or prompted speech, domains (close-talk, pmic or far field), native / non-native speakers for stress detection. For the purposes here we focus exclusively on prompted speech for native female subjects.

For this study, we employ 42 native female speakers with neutral and prompted physical stress conditions. Each speaker task segment is divided into 35 speech utterances (2-4 sec), resulting in a total of 5880 speech segments (42 spkrs * 35 utterances per task * 2 tasks * 2 mics). The short duration (2-4 sec) of the test utterances as well pose a challenge for the evaluation.

Ten-fold cross validation scheme is employed. The results reported represent the average over all test utterances. Furthermore, the test data is speaker exclusive (speaker used for building the models are not used for testing the model) as well as utterance exclusive (the TIMIT sentences used to build the model are not used for testing the model). The amount of data per speaker used to build the model is consistent and same for all the speakers in the model. Each stress model was build with the same amount of data per speaker. The distribution of test labels is 1:1 (i.e, number of utterances belonging to stressed speech is equal to number of utterances belonging to neutral category).

3. STRESS DETECTION

3.1. Algorithm Development

Physical airflow for speech production can be considered as a nonlinear system [10]. By modeling speech production as a spring-mass mechanical assembly, it is possible to explore fundamental issues of linear and nonlinear oscillations as a part of the excitation process [9]. Kaiser showed [9] that energy contained in speech could be modeled more accurately using a nonlinear operator, which is termed as the Teager Energy Operator (TEO),

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

Where $\psi[.]$ is the discrete-time TEO.

When a speaker is under stress, a number of factors will influence speech production including, air dynamics, muscle tension in true and false vocal folds, viscosity of oral tissue for airflow, ability and muscle control, movement of vocal tract articulators. Hence, TEO and/or TEO derived features will represent the change in nonlinear speech by modeling the airflow characteristics [1].

The illustration in Fig. 1 shows a bandpass filtered speech (from critical band #9, frequency range 1080-1270Hz) with the resulting TEO profile, autocorrelation response, and estimated envelope (in the lower plots) for neutral and stressed (angry) speech. The reduction in area of the autocorrelation response (bottom, right as compared to the bottom, left plot) reflects the change in regularity of the excitation under stress. Thus TEO derived features will help represent the variations in speech due to stress.

TEO-CB-AutoEnv (Teager Energy Operator – Critical Band – Auto Correlation Envelope) feature employs a critical band based filterbank to filter the speech signal followed by TEO processing. Each filter in the filterbank is a Gabor bandpass filter, with effective RMS bandwidth being the corresponding critical band.

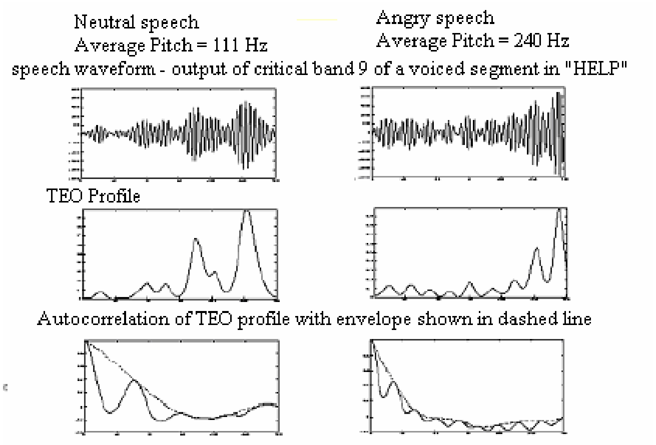

Fig. 1: TEO-CB-AutoEnv profile with change in stress

For our study, the TEO-CB-AutoEnv features on a per frame basis are extracted by segmenting the TEO profile output into 20ms frames with 10ms overlap between adjacent frames. M normalized TEO autocorrelation envelope area parameters are extracted for each time frame (i.e, one for each critical band), where M is the total number of critical bands (for this study M = 18). This is the TEO-CB-AutoEnv feature vector per frame [1].

Mathematically, TEO-CB-AutoEnv [1] using critical bandpass Gabor filters can be illustrated as,

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

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

$$R_{\psi_j^i(n)}(k) = \sum_{n=1}^{N-1} \psi_j^i(n)\, \psi_j^i(n+k) \tag{4}$$

where,

$g_j(n)$, j = 1,2,,18 is the bandpass filter impulse response,

$u_j(n)$, j = 1,2,,18 is the output of each bandpass filter,

* denotes the convolution operator,

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

3.2. Model Development

To investigate the optimal speaker size to reliably represent the stress model, stress models are constructed using different numbers of speakers. The term speaker size in modeling represents the number of speakers used to train the model, and *not* the actual physical size of the speaker. To date, no effort has been made to quantify the number of speakers needed to build effective stress models. The different speaker combinations for this study are: 3, 6, 12, 20, and 35; to construct stress models from a pool of 42 native females.

A GMM framework is used with our previously formulated TEO-CB-AutoEnv features, with a 20 ms window size and 10 ms skip rate, with 18-dimen. TEO-CB-AutoEnv features per frame.

The TEOmix1 is a GMM stress model using a reduced number of mixtures, while TEOmix2 uses twice as many mixtures than TEOmix1. For example, while training a stress model (one model per stress style, hence two stress models, one for Neutral speech, and another one for Physical stress speech), using 12 speakers, TEOmix1 has 32 mixtures, while TEOmix2 has 64 mixtures. Table 1 summarizes the number of mixtures used to build the model.

# spkrs →	**3**	**6**	**12**	**20**	**35**
TEOmix1	16	16	32	32	64
TEOmix2	32	32	64	64	128

Table 1: Number of mixtures in Stress GMM

Table 2 indicates the stress detection performance of the TEOmix1 and TEOmix2 for different speaker population size for the stress model with data from the close-talk mic. TEO features perform above 60% for all speaker conditions, with performance slightly improving as number of speakers increases in the stress model.

# spkrs →	**3**	**6**	**12**	**20**	**35**
TEOmix1	60.93	61.37	63.71	63.46	63.58
TEOmix2	60.87	61.47	62.96	63.16	62.70

Table 2: Stress Detection Performance (%) for Close-talk Mic

Table 2 indicates that the stress detection performance for 12 speakers is 63.71% for 32 mixtures and 62.98% at 64 mixtures, while 20 speaker, 64 mixtures GMM is 63.46% (better by 0.18% over 12 speaker, 64 GMM model), and 35 speaker, 128 mixture GMM is 62.67% (better by -0.28% over 12 speaker, 64 GMM model). Thus, performance with 12 speakers, 64 GMM model is similar to those achieved by using more speakers. Also, the difference between the TEOmix1 and TEOmix2 for the same speaker size model is negligible, indicating that a reduced number of mixtures is acceptable to build the stress models. As seen, the performance remains consistent beyond 12 speakers. *So, no major gain is achieved if we increase the training size significantly. As such, it is recommended that a training speaker set of 12 with 32 GMM mixtures is acceptable for stress model construction.*

3.3. Alternative Sensor

While a close-talk acoustic mic is effective, it would be useful to explore performance with alternative input sensors for stress detection. The physiological microphone (P-MIC) was developed by Scanlon at US Army Research Lab. with the goal to assist for soldier monitoring and research on Sudden Infant Death Syndrome[6, 7]. This sensor is about one inch square in size with a gel-filled chamber, conical focusing aperture and a piezoelectric sensor behind the chamber [6,7] (see Fig. 3).

As the gel-pad has an impedance similar to that of skin and a bandwidth permitting intelligible voice, the vibrations sensed on the throat can easily be transferred to the piezoelectric sensor and hence to the recording unit. The insulating material covering the sensor (not the gel-pad) helps to keep the acoustic (airborne) coupling with the ambient background to a minimum. Fig. 3 shows the placement of P-MIC at the vocal system of a user, held closely by the use of a Velcro strap.

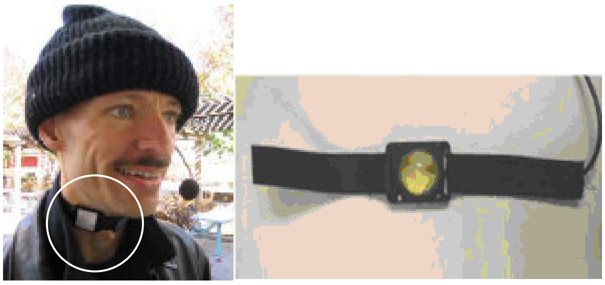

Fig. 3: The Physiological Microphone (PMIC) and it's placement around neck

There are other non-acoustic sensors such as TERC (tuned electromagnetic resonator collar), EGG (electroglottalgram), GEMS (radar-based glottal electromagnetic sensors), ultrasonic or photoelectric sensors, bone- or skin-conduction accelerometers which can be used for the purpose, but we choose to use the P-MIC for the reasons mentioned above including lightweight and ease of use.

# spkrs →	**3**	**6**	**12**	**20**	**35**
TEOmix1	65.11	64.67	66.36	66.15	63.03
TEOmix2	64.85	64.54	66.91	66.11	62.87
wrt close	**+4.18**	**+3.30**	**+2.65**	**+2.69**	**-0.54**
wrt close	**+3.98**	**+3.08**	**+3.94**	**+2.95**	**+0.17**

Table 3: Stress Detection Performance across different speaker number in models

Table 3 summarizes stress detection performance for the P-MIC sensor. The last two rows compare P-MIC performance with that obtained from close-talk mic. *For a reduced number of speakers in the model, P-MIC is about 3% better than close.* Even for 35 speakers in the stress model, the performance is almost the same, or better with P-MIC sensor. The results indicate that P-MIC can be **good** alternative to a close-talk mic for applications in which the close-talk mic may hinder man-machine interaction, (e.g., emergency response person when ambient conditions are noisy). Along with the performance shown above, P-MIC has (i) intelligible speech bandwidth, (ii) less acoustic coupling with ambient noise, and (iii) an option to help analyze the physical status of the user by acquiring heart-rate and breathing pattern.

3.4. Combined/Fused System

The performance of the TEO-CB-AutoEnv features based model is close to 63%. TEO-CB-AutoEnv features are believed to represent excitation, and regularity / correlation of the signal with some spectral dependency, while MFCC represents spectral structure, vocal tract information along with spectral tilt. Hence, a fusion system based on a machine learning scheme such as the Adaboost algorithm could combine scores from TEO-based models and MFCC-based models and produce better overall performance. 19-dimension MFCCs extracted from 20ms window with a skip rate of 100 using HTK toolkit. Adaboost (adaptive boosting) algorithm is an adaptive structure in which a weak classifier is repeatedly called, with weights adjusted in a way to bias in favor of misclassified instances during the previous iterations [5].

Fusion scheme is implemented for both the mic types on the optimal model size (model with 12 speakers) obtained above from our experiment.

Table 4 present the results for the fusion system for close-talk and P-MIC respectively. The fusion scheme with MFCC stress model scores and TEOmix2 based model scores gave a performance of 72.39%, for close-talking mic data while the accuracy of 79.35% with P-MIC data.

Hence, we see that combining excitation-based TEO-CB-AutoEnv features with vocal-tract based MFCC features improves system performance by +9.43% in close-talk mic in the best case, while an improvement of +12.99% absolute occurs for P-MIC over the TEO-based system.

Features	**Close-mic**	**Pmic**
MFCC	73.61	77.77
TEOmix1	63.71	66.91
Fusion1	72.63	79.18
Improvement (%)	**+8.92**	**+12.27**
MFCC	73.61	77.77
TEOmix2	62.96	66.36
Fusion2	72.39	79.35
Improvement (%)	**+9.43**	**+12.99**

Table 4: Performance (% accuracy) for the fused system for 12-speaker size model

5. CONCLUSIONS

Speech systems generally show performance drop because of speaker-stress variability. This study has focused primarily on speaker size in modeling along with choice of acoustic sensors to improve stressed speech detection. The study focused on having a reliable framework for stress detection for physical stress, where speakers used a stair-stepper task at a constant speed during exercise and repeated sentence prompts. We found the stress detection performance (62.96%, close talk mic, 66.36% for P-MIC) was consistent with twelve or more training speakers in the stress model, for either audio sensor. The study also shows the P-MIC to be a very good alternative to the close-talk acoustic sensor and can be deployed in situations wherein close-talk mics may hinder man-machine interactions, (e.g., fire-fighters, or personnel working in hazardous conditions). By fusing TEO-CB-AutoEnv feature model scores with traditional MFCC features using the Adaboost algorithm, the fused system shows an improvement of +9.43% (absolute, for close-talking mic) and +12.99% (absolute, for P-MIC), indicating that both features model different aspects of speech under stress.

Future work will focus on effect of scale change (when the database involves more number of enrolled speakers, say 100, or even more). Even still, for a smaller database or enrollment unit or application arena, we feel our results will hold true. Also, in future we will extend out study to include other variations like impact of non-nativity, different enrollment language, impact of age variations, and gender.

The three advances discussed in this paper will allow for effective stress detection algorithm development, with fewer training speakers and/or alternative audio sensors. The combined feature domains will contribute to improving voice-interaction systems. The contributions here establish general yet flexible guidelines to build more effective stress detection algorithms for new man-machine/voice-interactive domains.

6. REFERENCES

[1] Zhou G., Hansen, JHL, and Kaiser, J.F., “Nonlinear Feature Based Classification for Speech Under Stress,” IEEE trans. of SAP, vol. 9, no. 3, March 2001, pp. 201-216.

[2] Hansen JHL, Swail C., South A.J., et al, “The impact of Speech Under Stress on Military Speech Technology,” NATO Research and Technology Organization RTO-TR-10, March 2000, vol AC/323(IST)TP/5 IST/TG-01.

[3] Speech Communications, Special Issue on Speech under Stress, vol. 20, Nov 1996.

[4] Bou-Ghazale S.E., Hansen JHL, “HMM-based Stressed Speech Modeling with Application to Improved Synthesis and Recognition of Isolated Speech under Stress,” IEEE trans. of SAP, vol. 6, no. 3, May 1998, pp. 201-216.

[5] Schapire R.E., Singer Y., “BoosTexter: A boosting-based system for Text Categorization,” Machine Learning, vol 39 (2/3), pp. 135-168, 2000.

[6] Bass J.D, Scanlon M.V., Mills, T.K., “Getting Two Birds with one Phone: An Acoustic sensor for both speech recognition and medical monitoring,” The JASA, vol 106, no. 4, October 1999, pp. 2180.

[7] Campbell, W.M., Quatieri, T.F., et al, “Multimodel Speaker Authentication using Nonacoustic Sensors,” Workshop on Multimodal User Authentication, pp. 215-222, Dec 2003

[8] Stanton, B.J., “Robust recognition of loud and Lombard speech in the fighter cockpit environment,” PhD thesis, Air Force Inst. Of Tech., 1988.

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

[10] Teager H., Teager S., “Evidence for Nonlinear Production Mechanisms in the Vocal Tract,”, NATO Advanced Study Institute, vol 55, Bonas, France, pp. 241-261, 1990.

[11] Ruzanski, E., Hansen JHL, et al, “Effect of Phoneme Characteristics on TEO Feature-based Automatic Stress Detection in Speech,” ICASSP’05, vol 1, pp. 357-360, 2005.

[12] Bou-Ghazale, S.E., and Hansen JHL, “A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress,” IEEE trans. SAP, vol 8(4), pp. 429-442, 2000.

[13] Rahurkar M., Hansen JHL, et al, “Frequency Band Analysis for Stress Detection Using A Teager Energy Operator Based Feature,” ICSLP’02, vol 3, pp. 2021-2024, 2002.

[14] Varadarajan, V., Hansen, JHL, “Analysis and Normalization of Lombard Speech Under Different types and levels of Noise with Application to In-set Speaker ID systems,” IEEE trans. [in press, 2007]

[15] Ikeno, A, Varadarajan, V, Patil, S, Hansen, J.H.L., “UT-Scope: Speech under Lombard Effect and Cognitive Stress” AeroSpace Conference, March 2007 Big Sky, Montana.