SPEECH UNDER PHYSICAL STRESS: A PRODUCTION-BASED FRAMEWORK

Sanjay Patil, Abhijeet Sangwan, and John H. L. Hansen

Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A.

ABSTRACT

This paper examines the impact of physical stress on speech. The methodology adopted here identifies inter-utterance breathing (IUB) patterns as a key intermediate variable while studying the relationship between physical stress and speech. Additionally, this work connects high-level prosodic changes in the speech signal (energy, pitch, and duration) to the corresponding breathing patterns. Our results demonstrate the diversity of breathing and articulation patterns that speakers employ in order to compensate for the increased body oxygen demand. Here, we identify the normalized value of breathing energy rate (proportional to minute volume) acquired from a conventional as well as physiological microphone as a reliable and accurate estimator of physical stress. Additionally, we also show that the prosodic patterns (pitch, energy, and duration) of high-level speech structure shows good correlation with the normalized-breathing energy rate. In this manner, the study establishes the interconnection between temporal speech structure and physical stress through breathing.

Index Terms— Physical Stress, Speech under Stress, Breathing Patterns, Physiological Microphone (PMIC)

1. INTRODUCTION

Speech system performance is greatly impacted by speech produced under stress. Speakers encounter various forms of stress, namely, physical (*e.g.* running), cognitive (*e.g.* driving), chemical (*e.g.* medication), fatigue (*e.g.* sleep deprivation), and Lombard-effect (*e.g.* speech in noisy environment) [1]. Here, stress is known to induce changes in the spectral and temporal patterns of speech. As a result, it is important to establish the nature of these changes in order to compensate for the effects and improve speech system performance. This work focuses on studying the impact of physical stress on temporal patterns of speech.

The human body undergoes a number of physiological changes under physical stress. In general, a set of hormones including adrenaline, nor-epinephrine and cortisol get released which boost the heart-rate and change the breathing pattern. Heart-rate is reflective of cardiovascular status and is used to determine exercise intensities as well as VO_2max [2]. In fact, increased heart-rate variability can help in identifying stress, frustration, anger and fear. While both the heart-rate and breathing pattern variability are associated with physical stress, breathing directly impacts speech production. As a result, this study focuses on establishing a relationship between breathing effort and temporal patterns of the speech signal. In particular, our study highlights the changes that occur in the prosodic elements of speech syllable structure.

This project was supported in part by USAF under a subcontract to RADC, Inc. under FA8750-05-C-0029. Approved for public release; distribution unlimited.

The process of breathing causes the tissues close to the trachea to vibrate. These vibrations propagate onto the body surface, and carry the information of the breathing pattern. Using a suitable sensor such as the PMIC (physiological microphone), these vibration signals can be easily recorded [3]. Additionally, the traditional close-talking microphone (CTM) can also be employed for capturing the changes in breathing pattern, albeit with lower reliability as compared to PMIC. This is primarily because CTMs are typically placed at least 10 cm away from the lips, unlike PMIC which are in-contact with the skin. In this work, we compare the capability of PMIC in capturing these breathing patterns to CTM. The results of our study indicate that PMIC offers a more reliable and accurate assessment of physical stress.

Previous studies have focused on examining the impact of breathy articulation on speech. Specifically, the following changes have been observed: (i) increased spectral slope for frequencies above 2kHz, (ii) increased first formant bandwidth, and (iii) increased non-periodic aspiration noise at higher frequencies [4, 5]. This is caused by decreased coordination between subglottal air pressure and medial compression of the vocal folds [6]. However, these studies have predominantly focused on isolated sustained/normal vowels as opposed to continuous speech. Furthermore, we hypothesize that physical stress will also manifest in the temporal patterns of higher-level speech structures (such as syllables) in conjunction with spectral changes. As a result, the current study focuses on examining the temporal variations in speech.

2. PROPOSED MODEL

The proposed model for physical stress is shown in Fig. 1. As shown in the figure, physical exertion impacts the respiration rate which manifests in the speakers breathing. As the demand for oxygen increases, the breathing is expected to be rapid (more breaths per minute) and deep (more tidal volume of air per breath) [7]. As shown in Fig. 1, the changes in breathing, especially exhalation has a direct impact on articulation. Thus, the proposed model identifies breathing as a key intermediate variable to analyze the impact of physical stress on speech. In other words, speech systems dealing with variations in speech due to physical stress can conveniently ignore the actual physical task while focusing on breathing characteristics alone. Therefore, implementing the proposed model requires defining a set of parameters that adequately capture breathing characteristics. Additionally, spectral and temporal characteristics of speech that are most impacted by physical-stress induced breathing also need to be determined. In this work, we focus on identifying key temporal changes at a higher-level of consonant-vowel (CV) clusters. Particularly, we compare and contrast CV pitch, energy, and duration characteristics of speech under neutral and physical exertion. Expanding understanding of the impact of physical stress on higher temporal structures augments to the knowledge of its im-

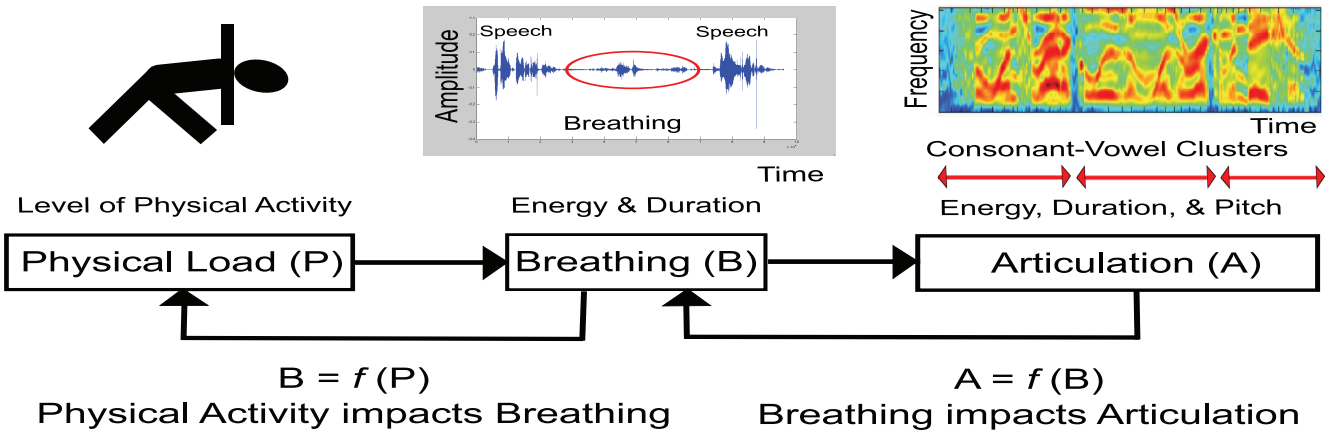

Fig. 1. Proposed Model: The physical load on a speaker impacts the breathing which in turn induces artifacts in the speech signal where (i) the estimates of breathing pattern are drawn from inter-utterance breathing (IUB) segments, and (ii) the articulatory characteristics of speech considered are consonant-vowel pitch (P), energy (E) and duration (D) statistics.

Table 1. Inter-Utterance Breathing (IUB) and Consonant-Vowel (CV) Cluster parameters for speakers under Neutral and Physical Stress.

		For Inter-Utterance Breaths (IUB)		For Consonant-Vowel Segments (CVS)		
		Avg. ($\sum$Energy)	Avg. μ[Energy]	Avg. μ[P] (Hz)	Avg. μ[Energy]	Avg. μ[Duration] (s)
Neutral	PMIC	20818	35	195	46	0.208
	CTM	18022	30	195	46	0.208
Physical	PMIC	14295	34	210	51	0.194
	CTM	11474	28	210	51	0.194

pact on fine spectral structure. When the knowledge of temporal and spectral changes are put together, then a holistic understanding of the phenomenon is complete.

2.1. Inter-Utterance Breathing (IUB)

In order to estimate the changes in breathing patterns, we rely on inter-utterance breathing (IUB). As shown in Fig. 1, IUB occurs when speakers takes a pause between two utterances to catch their breath. Under physical stress, the characteristics of IUB changes with speakers tending to gulp larger quantities of air to meet increased body oxygen demand. We have observed that speakers tend to employ two general strategies for inter-utterance breathing : (i) a series of uniform gulps, or (ii) a large gulp of air succeeded by smaller gulps. Hence, the tidal volume of air flowing in and out of the speakers body is a more reliable and accurate measure of physical stress than breathing rate alone [7]. Here, the exhalation cycle of breathing tends to be more forceful under physical stress, which is adequately captured by the CTM and PMIC. The energy patterns of breathing (exhalation) are easily measured in either modality and form the basis of the proposed features in this work. Particularly, we examine the role of breathing-energy and breathing-energy-rate (energy divided by duration) of IUB in differentiating between neutral and physical exertion.

Let $\mathbf{b}_i$, $i = 1, \ldots, N$ be the i^{th} frame of inter-utterance breathing, and $\mathbf{E}_{\mathbf{b}_i}$ be the energy of frame $\mathbf{b}_i$. Now, let the total and average energy of inter-utterance breathing be given by $\sum(\mathbf{E}_{\mathbf{b}_i})$ and $\mu(\mathbf{E}_{\mathbf{b}_i})$, respectively. The total energy ($\sum(\mathbf{E}_{\mathbf{b}_i})$) is assumed to be proportional to the *tidal volume* of air [7]. Similarly, the mean energy is assumed to be proportional to *minute volume* [7]. We expect the estimation of tidal and minute volumes to be better on PMIC than CTM. This is because while CTM captures exhalation more efficiently than inhalation, PMIC captures inhalation as well as exhalation with equal effectiveness.

2.2. Consonant-Vowel Cluster Features

This study focuses on higher-level prosodic changes that occur in the speech signal due to physical stress. In particular, we are interested in the relationship between breathing features (mentioned above) and temporal properties of speech. As shown in Fig. 1, the speech signal is first decomposed into a series of consonant-vowel (CV) clusters and the corresponding energy (E), pitch (P), and duration (D) contours are extracted. The CV segmentation is achieved with the help of phonological features where the manner of articulation cues are used to identify the vowel and consonant sections of speech [8]. Thereafter, the mean values of energy (E), pitch (P), and duration (D) are computed for the CV clusters.

In this work, we test the hypothesis that a functional relationship exists between IUB (inter-utterance breathing) parameters and the values of E, P, and D in the adjacent spoken utterance. As speech is a form of controlled exhalation, a relationship between these parameters can be expected. Here, we assume a linear relationship between the parameters, *i.e.*,

$$\mu(\mathbf{E}_{\mathbf{b}_i}) = w_p P + w_e E + w_d D, \tag{1}$$

and

$$\sum(\mathbf{E}_{\mathbf{b}_i}) = w'_p P + w'_e E + w'_d D, \tag{2}$$

where w_p, w'_p, w_e, w'_e, w'_d, and w_d are the weight parameters. The value of the weights are determined using linear regression.

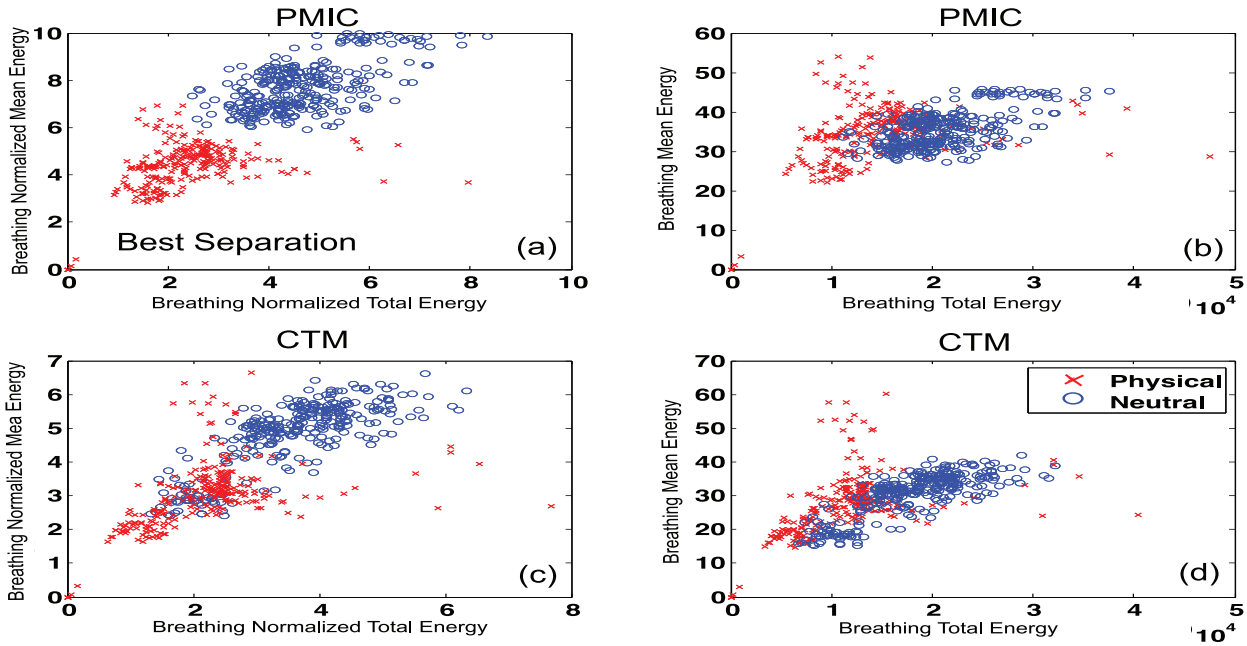

Fig. 2. Clear Separation between Physical Stress and Neutral Condition in space of Inter-Utterance Breathing (B) Features: (a) Normalized-Total Energy *Vs.* Normalized-Mean Energy in PMIC, (b) Total Energy *Vs.* Mean Energy in PMIC, (c) Normalized-Total Energy *Vs.* Normalized-Mean Energy in CTM, and (d) Total Energy *Vs.* Mean Energy in CTM.

Table 2. Different Articulation Strategies Employed By Speakers under Physical Stress identified in the space of Consonant-Vowel (CV) parameters for 9 different Speakers

Strategy	Spkr	Avg. Pitch		Avg. Energy		Avg. Duration	
		NX	PP	NX	PP	NX	PP
EDS	1	194	207	51	57	0.211	0.203
EDS	2	201	212	45	48	0.201	0.201
EDS	3	189	198	43	52	0.177	0.186
DDS	4	170	187	50	50	0.215	0.185
CS	5	177	179	40	48	0.239	0.200
CS	6	208	219	46	56	0.205	0.196
PDS	7	202	223	46	48	0.216	0.205
CS	8	200	204	48	55	0.203	0.207
CS	9	215	219	43	43	0.201	0.181

NX: Neutral & PP: Physical Stress

3. ANALYSIS AND RESULTS

The proposed model is analyzed on the UTScope corpus which consists of data under neutral condition, cognitive and physical stress from 85 speakers (63 females and 22 males). Each utterance in the corpus was collected synchronously on a close-talk Shure Beta-54 microphone (CTM), and a physiological microphone (PMIC). The data for speech under physical stress was obtained while speakers were exercising on a stair-stepper at a constant speed of 10 miles/hour. For the current study, we employ data from 9 female speakers of native American English. The heart-rate plot of these speakers exhibited an incremental increase with duration in the physical task and an almost flat trend for the neutral task.

Table 1 shows the average values of Inter-Utterance Breathing (IUB) as well consonant-vowel cluster parameters. Specifically, the total and mean energy of IUB (corresponding to tidal and minute volumes) are shown as an average for the entire data under consideration (all speakers). The average values are shown separately for PMIC and CTM under physical and neutral stress conditions. Similarly, the average values of pitch (P), energy (E), and duration (D) corresponding to the CV clusters are also shown under physical and neutral stress conditions for PMIC and CTM. As seen in the table, there is a general tendency of speakers to employ a higher pitch under physical stress when compared to neutral condition. Additionally, speakers shorten the duration of CV segments while pumping more energy into them. Furthermore, it is also observed that the total energy of inter-utterance breathing falls from neutral to physical. This observation is perhaps explained as an attempt by speakers to re-direct constrained exhalation resources to speaking under physical stress. However, as expected the combination of speaking (E) and breathing (B) energies are greater for physical stress than neutral condition.

The separability of breathing patterns under neutral and physical stress conditions are shown in Fig. 2. The separability is analyzed in the 2-dimensional space of total breathing energy ($\sum(\mathbf{E}_{\mathbf{b}_i})$, tidal volume) *vs.* mean breathing energy ($\mu(\mathbf{E}_{\mathbf{b}_i})$, minute volume) for PMIC and CTM. Additionally, the separation is also shown in space of normalized values of total and mean breathing energy (obtained by dividing by their standard-deviation values). Here, the normalized values of total and mean breathing energy are good indicators of physical stress as they clearly separate the physical and neutral data clusters (see Fig. 2 (a)).

Table 2 shows the average values of pitch (P), energy (E) and duration (D) for all 9 speaker used in the analysis data-set. The variations in the pitch, energy and duration values reflect upon the various strategies that can be employed by speakers to compensate for increased and sustained breathing requirement (in physical stress). Here, speakers 1-3 adopt an energy dominant strategy (EDS) where

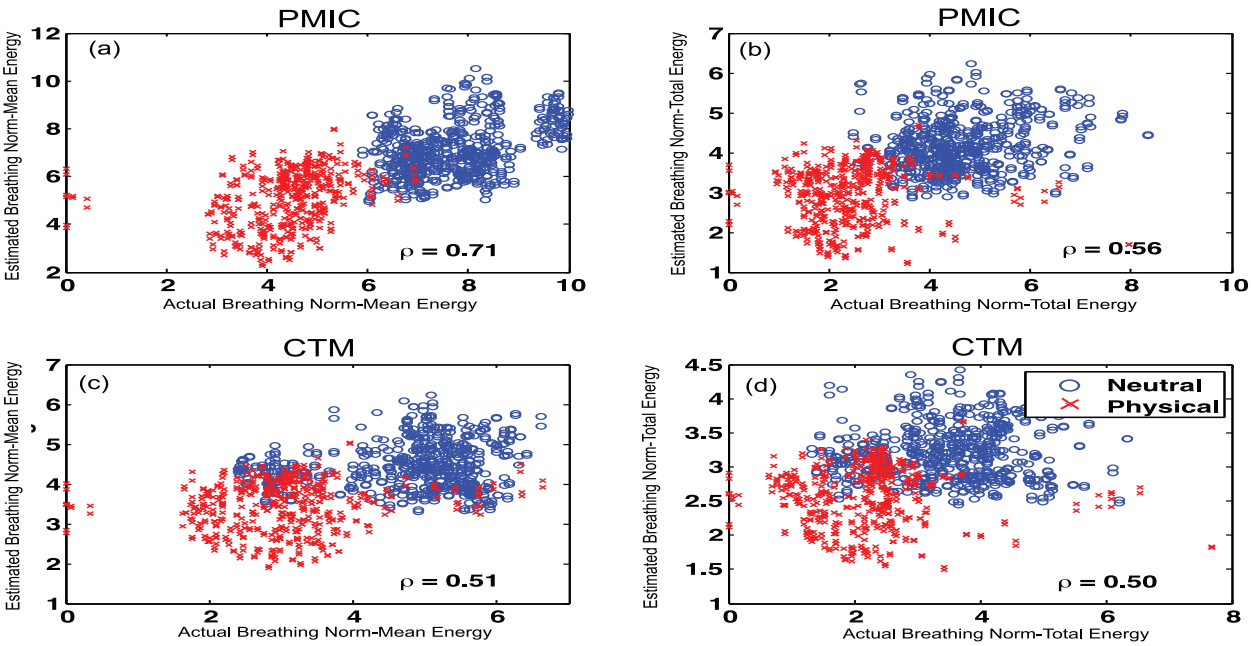

Fig. 3. Correlation between Estimated and Actual Breathing Normalized-Mean Energy (a) in PMIC, (c) and CTM; Correlation between Estimated and Actual Breathing Normalized-Total Energy (b) in PMIC, and (d) in CTM.

the amount of energy used in CV clusters is increased under physical stress. This is also accompanied by a small increase in pitch values but no marked decrease in duration is observed. Under this strategy, speakers divide the articulatory load into smaller more manageable chunks (such as speaking words of an utterance one at a time). Thus, speech may no longer be continuous but a string of isolated words. Speaker 4 adopts a duration dominant strategy (DDS) where the speaking rate increases dramatically under physical stress. It is also accompanied by a small increase in pitch but no increase in energy values. Furthermore, it is also seen that speaker 7 applies a pitch dominant strategy (PDS). Other speakers are observed to apply a mixture of these pure strategies (CS: combined strategy). The above observations reflect the diversity of options that are available to speakers while encountering physical stress. It also hints at the difficulty that speech systems can potentially face when encountering speech under stress.

Finally, we employ the pitch (P), duration (D), and energy (E) values of CV clusters to predict the normalized value of mean and total breathing energy (minute and tidal volumes, respectively) in a linear regression setup (using Eqns 1 and 2). Fig. 3 (a) and (c) shows the correlation between the actual and estimated values of minute volume (normalized mean breathing energy) for PMIC and CTM. Similarly, 3 (b) and (d) shows the correlation between the actual and estimated values of tidal volume (normalized total breathing energy). It can be observed that the highest correlation is obtained while estimating minute volume (normalized mean breathing energy) on the PMIC channel (see Fig. 3 (a)).

4. CONCLUSION

In this paper, it has been shown that the normalized value of breathing-energy rate can be a reliable and accurate indicator of physical stress. This breathing measure is easily extracted from the conventional (CTM) or physiological (PMIC) microphones. Additionally, we have also shown the diversity of prosodic patterns that speakers employ to compensate for increased breathing demand during physical stress. However, the variations in pitch, duration, and energy patterns show a reasonably strong linear relationship with normalized-breathing energy rate. These changes in the articulation patterns have implications for design of speech system compensation strategies (speech, speaker, emotion, dialect/accent identification systems).

5. REFERENCES

[1] "Special issue on speech under stress," *Speech Communication archive*, vol. 20, no. 1-2, 1996.

[2] R. Robergs and R. Landwehr, "The surprising history of the hrmax=220-age equation," *Journal of Exercise Physiology*, vol. 2, no. 2, May 2002.

[3] J.D. Bass, M. V. Scanlon, T. K. Mills, and John J. Morgan, "Getting two birds with one phone: An acoustic sensor for both speech recognition and medical monitoring," in *Acoustic Society of America*, Nov. 1999.

[4] B. Blankenship, "The timing of nonmodla phonation in vowels," *Journal of Phonetics*, vol. 30, pp. 163–191, 2002.

[5] C. Ishi, "A new acoustic measure for aspiration noise detection," in *ICSLP*, 2004.

[6] S. England and D. Bartlett, "Changes in respiratory movements of the human chords during hyperpnea," *American Physiological Society*, vol. 52, no. 3, pp. 780–785, 1982.

[7] Willard R. Zemlin, *Speech and hearing science; anatomy and physiology*, Prentice-Hall Englewood Cliffs, N.J.,, 1968.

[8] A. Sangwan and John H. L. Hansen, "On the use of phonological features for automatic accent analysis," in *Interspeech-09*, Sept. 2009.