Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 327–340
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification

Sanjay A. Patil, John H.L. Hansen *

Dept. of Electrical Engineering, Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, United States

Received 8 January 2009; received in revised form 31 October 2009; accepted 23 November 2009

Abstract

Interactive speech system scenarios exist which require the user to perform tasks which exert limitations on speech production, thereby causing speaker variability and reduced speech performance. 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. Thus, in this study we focus on the use of a silent speech interface (PMIC), with a corresponding experimental assessment to illustrate its utility in the tasks of stress detection and speaker verification. This study focuses on the suitability of PMIC versus close-talk microphone (CTM), and reports that the PMIC achieves as good performance as CTM or better for a number of test conditions. PMIC reflects both stress-related information and speaker-dependent information to a far greater extent than the CTM. For stress detection performance (which is reported in % accuracy), PMIC performs at least on par or about 2% better than the CTM-based system. For a speaker verification application, the PMIC outperforms CTM for all matched stress conditions. The performance reported in terms of %EER is 0.91% (as compared to 1.69%), 0.45% (as compared to 1.49%), and 1.42% (as compared to 1.80%) for PMIC. This indicates that PMIC reflects speaker-dependent information. Also, another advantage of the PMIC is its ability to record the user physiology traits/state. Our experiments illustrate that PMIC can be an attractive alternative for stress detection as well as speaker verification tasks along with an advantage of its ability to record physiological information, in situations where the use of CTM may hinder operations (deep sea divers, fire-fighters in rescue operations, etc.).

Keywords: Physiological sensor; Stress detection; Speaker verification; Non-acoustic sensor; PMIC

1. Introduction

Many interactive systems involving speech require the user to perform tasks which introduce limitations on speech production, thereby causing speaker variability and reduced speech system performance (Baber et al., 1996; Benzeghiba et al., 2007; Bosch, 2003; Corrigan, 1996; Cosmides, 1983; Hansen, 1994; Murray et al., 1996). This results in Speech Under Stress, which represents a challenge for human cognitive coordination because of the alteration to the speech production process. Of the many factors which cause speech production variability such as accent, dialect, language difference, stress (cognitive, physical, or emotional based), represents a significant and typically ill-defined set of production deviations. Stress can result from cognitive load (time-constraint mental workload, mathematical computations, etc.), or from physical work load (cycling, fire-fighters in rescue operations), or from exposure to background noise resulting in Lombard effect (while communicating in noisy conditions such as a party or on a noisy street) (Hansen, 1996). These diverse factors including stress degrade automatic speech system performance as well as cause a loss in intelligibility for human speech perception. In adverse noisy, stressful situations where speech technology such as speech recognition, speaker verification, or where dialog systems are used,

* Corresponding author. Tel.: +1 972 883 2910.
E-mail address: john.hansen@utdallas.edu (J.H.L. Hansen).
URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.11.006

addressing noise is not sufficient to overcome performance loss (Junqua, 1996; Hansen, 1993; Bou-Ghazale, 1996; Viswanathan et al., 1984). 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. Noise in speech can broadly and loosely be categorized into two components, the first being ambient environment noise (AEN), with the second being noise due to alterations in the speech production process. Hence, if AEN is minimized, the study of speech production becomes possible. Any sensor which can record speech before it leaves the speaker's lip/oral cavity would be immune to AEN, allowing researchers to focus on the study of alteration to speech production process due to stress. This is the main focus of this study, which is to evaluate one such silent speech interface entitled the physiological microphone (PMIC).

Most research focused on speech under stress has concentrated on the use of traditional acoustic microphones. Some of these techniques which have been shown to improve performance include:- robust feature methods (developing features which are robust against stress variations); stress equalization methods (compensating feature variations due to stress); model adjustment/training methods (multi-style training, training for specific conditions to be encountered during testing). They can be adopted to mitigate the impact of stress on speech system performance (Bou-Ghazale and Hansen, 1995; Bou-Ghazale and Hansen, 2000; Hansen, 1994; Womack and Hansen, 1996b; Womack and Hansen, 1996a). However, for certain applications, employing an acoustic mic can limit human task performance, and therefore the use of alternative sensors such as the physiological mic (PMIC) should be considered. PMIC is a non-acoustic sensor which captures speech via skin vibrations through contact with the skin near the cricoid and thyroid cartilage. Fig. 1 shows various vocal organs and supporting structure surrounding vocal folds including the cartilages. Thus, the PMIC is a non-acoustic, contact sensor. Fire fighters, law enforcement officers, aircraft pilots, etc. or other subjects who require voice communication or voice engagement could benefit from the use of PMIC in their work environment. In such environments, noise robustness of the PMIC far surpasses that of the acoustic mic (CTM – close-talk microphone). Additionally, PMICs also capture a response related to the heart rhythm, and breathing, allowing for remote monitoring of a subject's physical conditions.

Most silent speech interface studies for speech applications capture articulation-related signal or glottal-fold signals before the sound pressure wave leaves the speaker's oral cavity. However, there are several potential sensors which do not require vocalization of speech, but the artic-

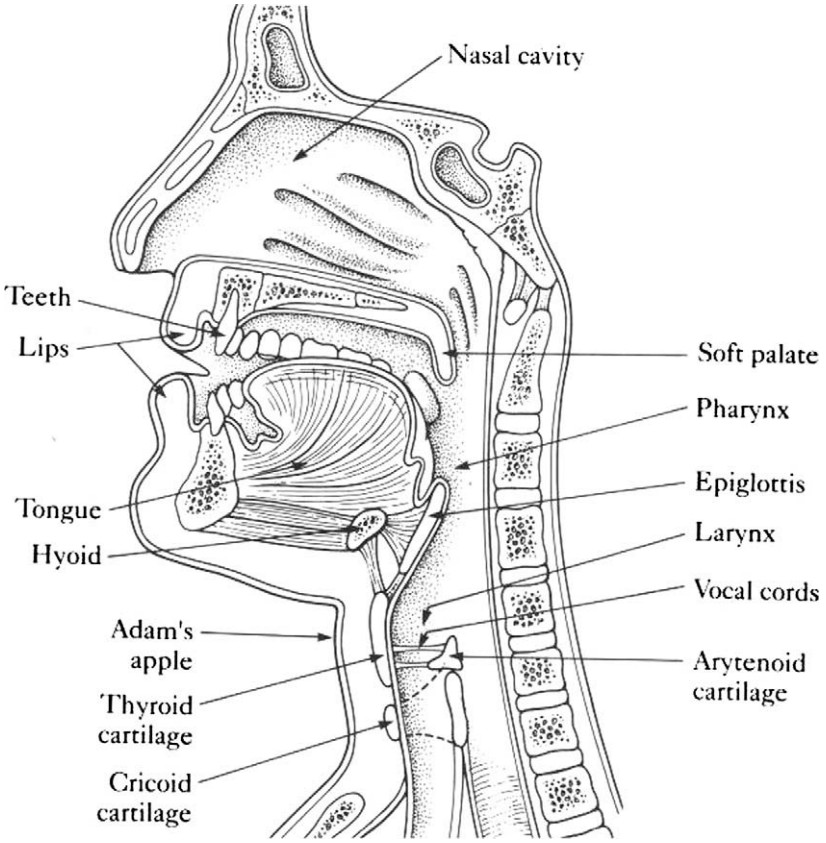

Fig. 1. The human vocal organs surrounding and supporting the vocal folds – specifically Cricoid cartilage and Thyroid cartilage (obtained from Denes and Pinson, 1993).

ulation movement. A second group of sensors concentrate on face contour mapping by use of image processing of facial videos (Knudsen et al., 1994; Otani and Hasegawa, 1995). To capture articulation-related signal and glottal-fold signals, most sensors are placed near or around the neck region. For example, the glottal electromagnetic sensor (GEMS) (Brady et al., 2004; Burnett, 1999) is a sensor which is attached near the neck region. The GEMS sensor works on the principle of electromagnetic wave scattering from body organs to reflect movement as speech information. GEMS sensor studies have considered detection of voiced speech, analysis of excitation characteristics (Holzrichter, 2002), improvement of ASR performance as well as speaker verification task (Gable, 2000). GEMS has also served as the primary signal capture sensor for the US DARPA Advanced Speech Encoding (ASE) program for ultra-low-bit (< 600 bits/sec) speech coding. GEMS require precise setup for effective speech capture and are costly sensors, requiring EM transmission and antenna setup. A variant of GEMS called the tuned electromagnetic resonator collar (TERC) (Brown et al., 2005) has been considered in a study on glottal activity. As compared with GEMS, TERC does not require precise alignment of the sensor and receiver, but TERC still requires the speaker to be positioned in one place. Some research groups are focused on the use of electrical signal from various muscles associated with speech production. These sensors can include for example electromyograms (EMG) (Chan, 2003; Titze et al., 2000), and electroglottalgrams (EGG) (Courteville et al., 1998; Titze et al., 2000). EMG exists in two variations; one EMG uses surface electrodes to capture muscle movement, while the second is more invasive with needle-like electrodes to capture specific muscles responses from speech. EMG (Jou et al., 2007; Wand et al., 2007) and EGG have been applied for automatic speech recognition, and speech synthesis. There also exists several silent speech interfaces (SSIs) which work on vibration pickup for various bone structures around the neck and face regions. These bone-conduction sensors (Quatieri et al., 2006) are mostly placed near the ear bone, or inside the ear, or forehead region. The purpose of these bone-conduction sensors as a SSI is to improve speech intelligibility in high ambient noise conditions, to allow differentiation between speech and non-speech, and also for ASR applications. Several other SSIs work on the principle of acquiring speech information via skin vibrations. These sensors and interfaces include the throat microphone (Ingalls, 1987; Mainardi and Davalli, 2007; Shahina and Yegnanarayana, 2005; Roucos et al., 1986), skin-conduction accelerometers (Akargun and Erzin, 2007; Graciarena et al., 2003), symmetrical differential capacitors (SDC) (Peters, 1995; Mohamad et al., 2007), and also include the physiological microphone (PMIC) (Scanlon et al., 2002) which is considered in detail in our study. These skin vibration SSIs convert surface vibrations into what is believed to reflect articulation movement, as well as vocal fold movements. The main motivation again with such a vibration-based sensor is to mitigate the harsh ambient noisy environment. Sometimes, these SSIs are used along with close-talk microphones in a multi-sensor configuration to boost the ASR performance by employing a data fusion scheme, where sensor signal usage is conditioned on the level (or type) of background noise (represented as SNR). Quite a few SSIs use an ultrasonic strategy (Heuber et al., 2007) to acquire speech information by mapping the articulation movement via a Doppler imaging concept, while some polymer-embedded microphone devices, like non-audible murmur (NAM) sensors, have been used for converting whisper into speech (Noma et al., 2005; Tran et al., 2008). Some SSIs use optical imaging and photoelectric sensors for speech interfaces (Knudsen et al., 1994; Otani and Hasegawa, 1995; de-Paula et al., 1992). As a SSI, some sensors such as the GEMS, and TERC have a critical procedure, while some sensors like SDC, PMIC, NAM or throat microphones are simple to use without any necessary prior setup time.

In this paper, the role of the physiological microphone (PMIC) is evaluated for two tasks – stress detection and speaker verification. Stress specific information is largely present in the glottal waveform. Useful speaker specific information can also be found in the glottal waveform. Therefore, it is expected that the PMIC would be useful in these two tasks due to its close proximity to the larynx. In other words, the PMIC will faithfully capture the glottal waveform under all conditions in a robust manner.

This study considers experiments conducted using the PMIC. The outline of the paper is as follows – the next section describes the collected database for experiments. Section 3 describes the PMIC and illustrates the difference between PMIC and CTM based on frequency structure. Subsequent sections describe the investigation of PMIC for stress detection and speaker verification. Finally, the last section illustrates the use of PMIC for heart rate monitoring for speech under stress.

2. UT-Scope database

This study is focused on speech data obtained from the UT-Scope database (Ikeno et al., 2007). The database currently consists of 85 speakers, with 63 females and 22 males. More than half of the speakers in the database participated in two to four data collection sessions in order to explore session variability. All data collection sessions were performed in an ASHA (American Speech-Language-Hearing Association) certified single wall sound booth. Two synchronous recordings are obtained, with one using a close-talk Shure Beta-54 microphone (CTM), and the second with a PMIC sensor positioned at the cricoid and thyroid cartilage. The data is sampled at 44.1 kHz and stored with 16 bits per sample. It is noted that all acoustic channels were recorded on a Fostex 8-channel digital recorder, so all data sets are recorded synchronously. A Polar heart rate monitor (Goodie et al., 2000) is used to record heart

rate information every 5 s, with blood-pressure measurements taken for some subjects before and after the tasks.

Speakers produced speech under two stress tasks (cognitive stress, and physical stress) in addition to producing speech under a neutral state while seated and relaxed (neutral relaxed state). So, the database consists of three segments of recordings per speakers, (i) neutral relaxed state, (ii) physical stress, and (iii) cognitive stress. The two stress tasks consist of a cognitive stress task and a physical stress task. The cognitive stress includes recordings while the subject operates a car in a racing car simulation with a Playstation™ console, while the physical task consists of the speaker using a stair-stepper at a constant speed of 10 miles/h. The physiological metrics such as blood-pressure and heart rate are recorded for a majority of the speakers. The ambient noise floor level is checked during all tasks, specifically for the physical task, since noise from the mechanical movement while stair-stepping was thought to impact (and thus skew) stress task identification. An environment noise assessment showed that the subject produced utterances at 50 dBA higher than the ambient noise floor. Thus, ambient noise does not drastically impair the task performances. The corpus consists of 35 TIMIT sentences and spontaneous conversation recordings for each speaker. Each TIMIT sentence[1] is between 2 and 5 s in duration. These sentences are chosen so as to be phonetically balanced. The spontaneous conversation includes discussion and/or views on various daily activities on the university campus. Thus, no personal information is recorded during the spontaneous conversation (eg, ensure IRB compliance).

For this study, speech files of 42 females (native speakers of American English) are used for algorithm development and evaluations. The age distribution is between 18 and 46 years, with more than 50% within the age group of 22–27 years. Analysis on blood-pressure and heart rate is not directly used for either reporting or enhancing the task performances, but the heart rate extracting algorithm is compared with the Polar heart rate monitor ground truth readings. Here, we focus on the comparative performance analysis between CTM and PMIC. The contiguous 35 TIMIT sentence recordings for each sensor is divided into individual sentences for our evaluations. The two sensor evaluation domains of stress detection and speaker verification are used to compare the two sensors. It is noted that most of the evaluations are carried on neutral relaxed state and physical stress segments. Thus, experiments are carried out on a pool of 42 (number of speakers) $\times$ 35 (TIMIT sentences) $\times$ 2 (mic types) $\times$ 2 (stress types) utterances. The spontaneous speech segment is not used for task assessments in this study. The cognitive stress task is used in Section 4.3, for performance comparison across two stress styles and Section 5, for speaker verification evaluations.

[1] The text for the 35 sentences were obtained from a selection used in the TIMIT corpus [http://www.ldc.upenn.edu/].

3. Physiological microphone (PMIC)

The PMIC is a physiological microphone which contains a piezo-electric crystal sensor enclosed in a gel-filled pad [refer Fig. 2]. Gel-filled pads are designed to have the impedance close to that of the skin, thus allowing for maximum transfer of vibration from skin to the sensor. Secondly, they conserve the bandwidth of the speech spectra for the best possible transfer of signal structure. As shown in Fig. 2, the PMIC is attached around the neck to the side of the cricoid and thyroid cartilage of the speaker using a Velcro strap. In this manner, it faithfully captures the skin vibrations during speech. The piezo-crystal in the PMIC converts skin vibrations into electrical signals which are subsequently recorded. The insulating material covering the piezo-crystal minimizes the acoustic (airborne) coupling between the piezo-crystal and the ambient environment. Thus, the PMIC acquires skin vibrations near the neck region during speech but minimizes transfer of ambient noise, making the sensor robust under high ambient noisy conditions (Scanlon et al., 2002). Also, the PMIC captures speech before it leaves the speaker's mouth, converting the articulation-related signal and vocal-fold signal through the skin vibrations into an electrical signal. Thus, the PMIC is one of the many silent speech interfaces, since it does not directly capture acoustic sound pressure from the speech airflow.

Even though the PMIC has a robust design against ambient noise, being a contact sensor makes PMIC vulnerable to other signal corruption. Artifacts tend to be introduced because of: (a) body movement near the contact surface, (b) improper sensor placement, and (c) nature of contact between skin and gel-pad (caused by perspiration). Thus, schemes to mitigate or overcome these short-comings need to be devised. While these issues have been reported in some studies, to our knowledge no prior research has been conducted to evaluate their impact on speech technology.

Since the PMIC captures articulation-related signals as well as vocal-folds signals via skin vibrations, and the CTM records air pressure variations near the lips, the two will have different frequency content. Spectrograms shown in Figs. 3 and 4 illustrate the frequency content of CTM and PMIC under two different stress conditions – neutral and physical stress. As a side note, on careful observation on the magnified spectrum on the right hand side in Figs. 3 and 4 both for the CTM and PMIC sensor, there exists a 60 Hz hum, a recording artifact which does not impact our analysis and other important conclusions drawn in this study. PMIC shows prominent lower frequency-band profiles for all phones, and clear structured high-frequency content from sustained voiced segments. Alternatively, as compared to CTM, the PMIC predominantly records the frequency structure in the 0–2 kHz band. Though a restricted frequency structure does suggest that potential crucial speech information might be lost, the PMIC might actually compress the complete frequency band information [0–4 kHz] within this smaller band [0–

Fig. 2. The physiological sensor (PMIC) (on right) and its placement (left) around the neck near cycloid thyroid cartilage.

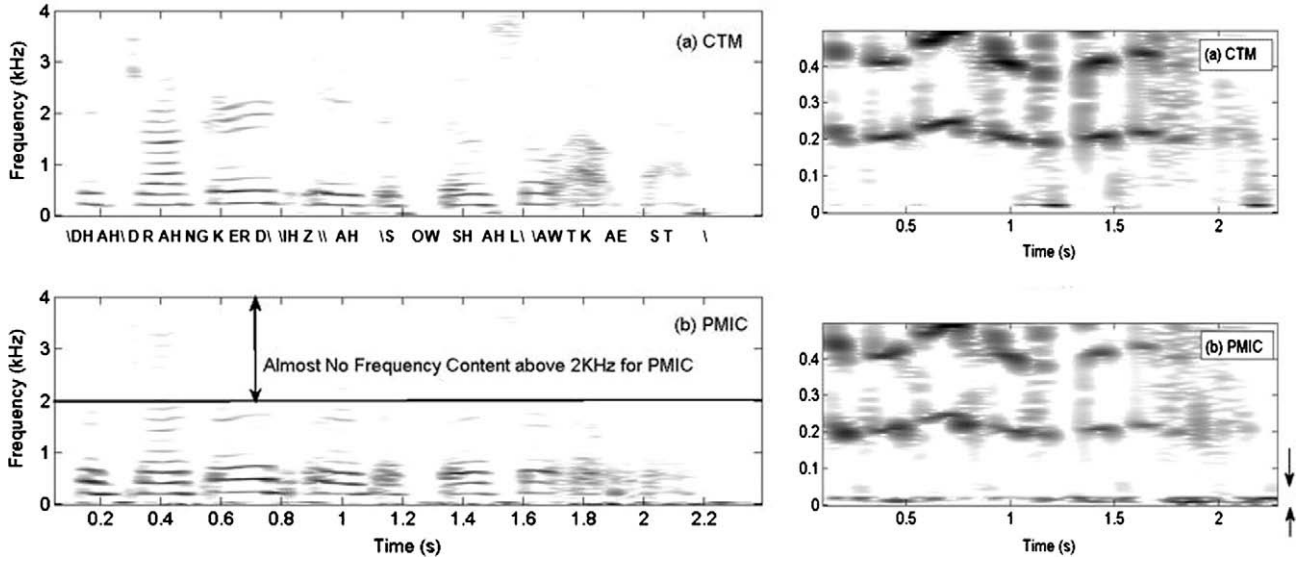

Fig. 3. Speech spectrogram for the utterance *the drunkard is a social outcast* (i) CTM [top] and (ii) PMIC [bottom] under neutral state – PMIC has almost no frequency contents above 2 kHz, (right) shows activity for frequencies less than 100 Hz (indicated by side arrows) for PMIC recordings.

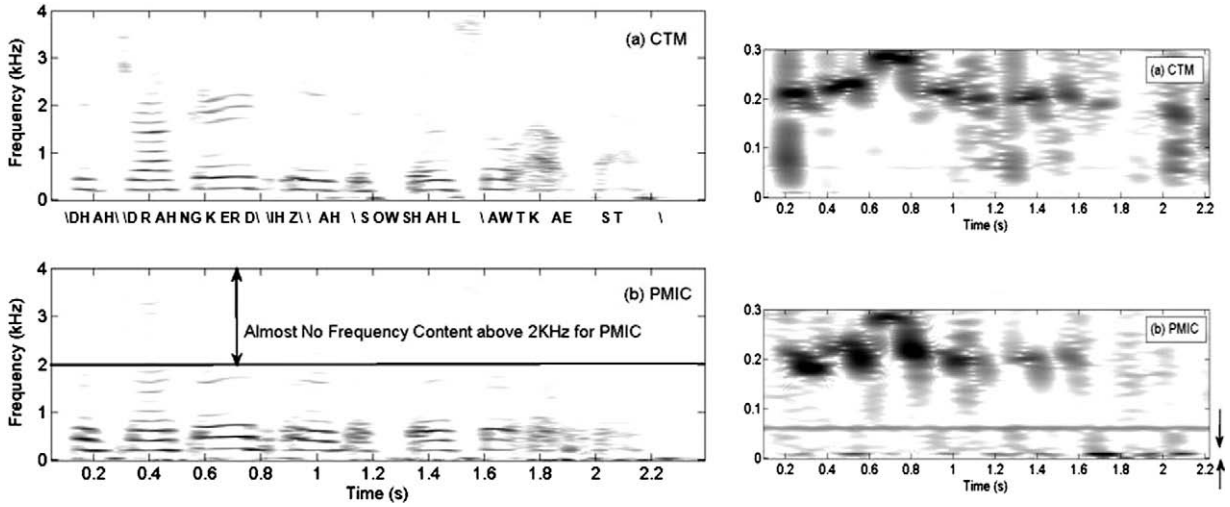

Fig. 4. Speech Spectrogram for the utterance *the drunkard is a social outcast* (i) CTM [top] and (ii) PMIC [bottom] under physical stress – PMIC has almost no frequency contents above 2 kHz, (right) shows higher activity for frequencies less than 100 Hz (indicated by side arrows), indicates probable recording of heart-beats and breathing pattern.

2 kHz]. This is because PMIC signal is significantly intelligible. Fig. 4 shows speech spectrograms under physical stress. Here, it is observed that the lowest frequency [0–100 Hz] for PMIC shows distinct activity which could be related to physiological activity. By comparing Fig. 4a and b, it can be seen that some potential physiological

activity is represented in the PMIC but not seen in CTM. The physiology of the user is reflected while the speaker is under physical exertion (Fig. 4b). This ability to reflect physiology (heart-beat information) is reported in Section 6 wherein the algorithm results are compared with baseline ground truth for neutral relaxed state and physical stress. An informal listener test with 3 subjects was conducted in which listeners evaluated the CTM and PMIC recordings for two recording segments: (a) speech segment, and (b) silence segment. The evaluation was used to find the differences between the two recordings. The listeners reported that (i) while the PMIC recordings were not as intelligible as CTM recordings, and listening to the PMIC speech over a long duration produced annoyance/fatigue, the listener was able to understand the sentence prompts, (ii) for silent segments, listeners did not hear any speech or other artifacts, yet upon increasing the audio gain, each subject would hear a regular beat. When the signal was inspected visually, the signal bore similarity with QRS complex pattern (representing the heart depolarization and repolarization electrical activity) as observed for heart-beat (Klabunde, 2005); (iii) the breathing pattern became recognizable at the very end of the physical stress session, when the subject becomes tired.

In the next section, we perform a comparative analysis between CTM and PMIC for two speech applications – stress detection (SD) and speaker verification (SV).

4. PMIC-based stress detection

The focus here is to compare the performance of PMIC and CTM, with the goals to:

(1) determine the impact of using different number of speakers in modeling stress,
(2) determine the impact of fusion scheme on PMIC and CTM-based systems, and
(3) determine stress detection performance for two different stress types (cognitive–neutral and physical–neutral).

The stress detection (SD) task is a binary (two-way) decision task in which the test signal is classified either as belonging to neutral or physical stress. Analysis involving cognitive stress is not reported in Sections 4.1 and 4.2, since larger deviations in speech production are expected under the physical stress condition. To evaluate the performance under the above mentioned scenario, 50% of the UTScope utterances from a speaker are used for training models, and the remaining 50% are employed as test data. The task is evaluated using the prompted speech segments from the UTScope database in which 35 sentences were recorded per speaker. Hence, 17 sentences were used for training and the remaining 18 sentences to evaluate the algorithm.

4.1. Impact of using different number of speakers in training stress model

To study the impact of variable number of speakers in the stress model, we form groups of 3, 6, 12, 20, 35 speakers derived from a larger pool of 42 female speakers. This process of forming groups of 3, 6, 12, 20, and 35 speakers was repeated five times to obtain five different sets for each speaker group [five-fold cross validation]. The different set of speakers is chosen to build the models across these five runs. Using multiple runs (five in our case) eliminates any speaker bias in the performance assessment. An overall average of the five runs is then reported as the final system performance. Stress detection is carried out using a GMM framework (128 mixtures), where the difference in log likelihood scores of the neutral model and physical stress model is used to classify a test token. The GMM models are trained using TEO-CB-AutoEnv features. The features are extracted from the signal using the process as described in (Zhou et al., 2001), and have been shown to be effective features for stress classification. For purposes of completeness, we include a flow diagram of the TEO-CB-AutoEnv feature in Fig. 5 (Zhou et al., 2001). We ensure that test and train utterances do not have overlap [no speaker overlap as well as data overlap]. Table 1 shows the average performance of stress detection for all five groups and both microphone sensors (PMIC, CTM) for TEO-CB-AutoEnv features. It is observed that the PMIC based system shows increasing performance with increasing group size. It is possible that the PMIC contains more diverse information, thus requiring more speakers to model both stress and related speech/speaker structure. Both systems (CTM-based and PMIC-based) show a jump in performance when moving from 3 speakers to 6 speakers in the stress model [SD accuracy of 87.07% and 88.69% for CTM, and PMIC-based system has accuracy of 83.49% and 86.27% for 3 and 6 speakers respectively]. The stress detection performance with CTM levels out for more than 6 speakers (less than 0.5% across 12 speakers, 20 speakers and 35 speakers). This indicates that the CTM may not contain as diverse information as compared to PMIC. Also, having 12 speakers with CTM may suffice for SD, while having an increased number of speakers with PMIC is better.

Table 2a indicates binary stress detection (neutral–physical stress) performance for CTM using both MFCC (first

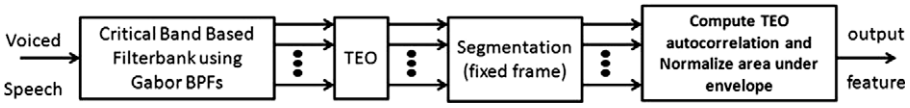

Fig. 5. TEO-CB-AutoEnv feature extraction algorithm [adapted from Zhou et al., 2001].

Table 1
Stress detection performance [in terms of accuracy (%)] for stress models build using TEO-CB-AutoEnv features for neutral/physical stress two-way detection – PMIC performs close to CTM-based system for stress models built using at least 12 speakers.

	Number of speakers in models				
	3	6	12	20	35
Close-talk mic (CTM)	87.07	88.69	**88.49**	88.21	88.07
PMIC	83.49	86.27	87.51	87.98	**89.42**
diff. w.r.t. CTM	−3.58	−2.49	−0.98	−0.23	1.35

Table 2
Stress detection performance for MFCC, TEO-CB-AutoEnv and fused system [in terms of accuracy (%)] combining MFCC and TEO-CB-AutoEnv based scores using Adaboost algorithm, (a) CTM and (b) PMIC. PMIC performs close to CTM-based system for stress models built using different speakers.

	Number of speakers in models				
	3	6	12	20	35
(a) CTM-based Adaboost system					
MFCC	90.83	91.93	**92.80**	93.00	92.96
TEO	87.07	88.69	**88.49**	88.21	88.07
Fusion	90.82	94.03	**95.16**	93.96	95.16
(b) PMIC-based Adaboost system					
MFCC	87.65	91.70	**93.46**	94.40	96.22
TEO	83.49	86.20	**87.51**	87.98	89.42
Fusion	92.56	93.55	**97.18**	95.38	96.28

row) and TEO-CB-AutoEnv (second row) features. Similar conclusions can be drawn, as seen with TEO-CB-AutoEnv features, with stress detection accuracy leveling off at about 92.80% (+/−1% variation) for 12 or more speakers in the stress models. Table 2b indicates stress detection performance for PMIC. MFCC-based system shows a jump from the 3 speaker based stress model to the 6 speaker based stress model (accuracy 90.83% jumped 91.93% for CTM, and 87.65% to 91.70% for PMIC). The PMIC SD system shows a gradual increase in stress detection accuracy with an increase in the number of speakers used to train the stress model. Secondly, the MFCC-based PMIC system shows improved accuracy as compared with the CTM system for 12 or more speakers (92.80% with CTM, and 93.46% with PMIC at 12 training speakers). The performance variations (jumps and lower values) for less than 12 speakers in the stress model can be clearly attributed to a reduced data diversity for training models, but still a SD accuracy of greater than or equal to 87.65% is still significant. Thus, the experiments with both TEO-CB-AutoEnv (18-dimension), MFCC (19-dimension C1–C19) present similar conclusions. The results indicate that PMIC performs as good as CTM or slightly better when the number of speakers is greater than 12. Thus, this is an indication that PMIC can be a viable alternative for CTM while operating under unfavorable conditions, such as emergency personnel in rescue operations (e.g., fire-fighters). During such operations, the subject will have significant gear and needs to be free of hand-held communication devices. Extreme levels of noise render CTM less successful, while the PMIC will be able to sustain performance as acoustic noise conditions evolve/change.

4.2. Fusion scheme

In this section, we explore if the two features (MFCC, TEO-CB-AutoEnv) carry complementary information. If the features were to be complementary, then it is expected that a fusion system will outperform the individual classifier systems. The fusion system combines the individual systems based on MFCC and TEO-CB-AutoEnv features for CTM and PMIC. The Adaboost algorithm is selected to fuse the systems and evaluate this hypothesis. The details on Adaboost algorithm implementation can be found in (Freund and Schapire, 1999; Huang et al., 2007). The LLR scores from the MFCC-based approach are combined with the TEO-CB-AutoEnv-based system. The fusion scheme performances are compared for binary stress detection tasks which detects whether the test signal belongs to either neutral relaxed state or physical stress. Table 2a and b illustrates the performance for the two sensors. Table 2a illustrates the performance of CTM with Adaboost fusion scheme, while Table 2b shows the performance of PMIC with Adaboost fusion scheme. The results clearly show that the individual system using MFCC or TEO-CB-AutoEnv feature for PMIC may not outperform CTM for all the speaker conditions. The combined PMIC Adaboost fusion outperforms CTM Adaboost fusion system in nearly all speaker group sizes. This suggests that the MFCC feature-set and TEO-CB-AutoEnv feature-set for PMIC sensor may store a subtle complementary information that is not possible with only the CTM sensor. Thus, stress differentiating information is distinctly stored by the PMIC sensor. PMIC-based fused system clearly outperforms CTM-based system by at least 1% over different speaker groups in stress models, except for the 6 speaker group where CTM based stress detection accuracy is 94.03% and PMIC based is 93.55% (slightly better by 0.08% more, which is not statistical significant). At 12 speakers, CTM fused system has a performance of 95.16% while PMIC has 97.18% (2.02% higher). This also indicates that for systems with 12 speakers or more in the stress model, PMIC becomes a viable alternative for CTM-based system. It is noted that while the error rate reduction between the CTM fusion system and the PMIC fusion system is 2.02%, this reflects a 41.7% relative reduction in stress detection error rate.

4.3. Performance for different stress types

The UTScope database consists of two stress types, cognitive and physical stress. Though evaluations in previous subsections focused on SD between neutral and physical

stress (NX–PP), another set of experiments were carried out to compare the two sensors performance also for neutral/cognitive (NX–CP) stress. As mentioned is Section 2, cognitive stress data recordings are done while the speaker operates a car in a racing car simulation with a Playstation™ console. The separate set of experiments were carried out and results are represented in Fig. 6.

The cognitive stress model and neutral state models were constructed using speech data from 12 speakers, since it is reported in Section 4.2, that both PMIC and CTM-based SD have comparable performances which level off with 12 or more training subjects. However, due to data convergence problems using the cognitive stress data, the accuracies are reported with 64 mixtures used in the Gaussian mixture models instead of 128 mixtures as reported in the earlier section for physical stress. MFCC features were used to build the stress models. The performance is lower for the NX–PP as compared to that reported in Table 2, due to the reduced number of mixtures in the GMM models. In spite of this, we feel that the conclusions drawn on the sensor performance for the two stress styles are still effective.

Fig. 6 shows the comparative performance of CTM and PMIC across two SD systems. For binary SD for cognitive/neutral, CTM gives 65.62% accuracy, whereas PMIC has 76.86% accuracy, a performance difference of +11.24% absolute, indicating that PMIC can effectively represent the subtle differences between cognitive and neutral conditions to a far greater degree than the CTM sensor. The SD accuracy for physical/neutral is 73.29% with CTM, while PMIC represents better performance by +5.73% absolute at 79.02% accuracy. By comparing the relative difference between CTM and PMIC performances for the CP–NX system, it is clear that the PMIC stores the cognitive stress information distinctly as compared with CTM. This would suggest that the glottal folds undergo drastic changes during their vibratory pattern for speech production which can be caught by skin vibrations, thus by PMIC.

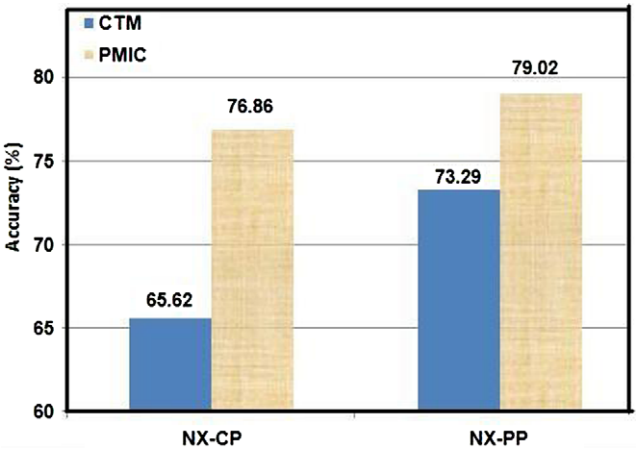

Fig. 6. Stress detection (Accuracy %) for PMIC and CTM sensors.

5. Speaker verification experiments

The prior section considered the PMIC and CTM for stress detection. In this section, the focus shifts to a speaker verification application carried out to evaluate the suitability of PMIC versus CTM. As PMIC is snuggly attached around the neck, close to the larynx (vocal folds), it is expected that PMIC will record speaker-dependent information. To test our hypothesis, the best option is to check the degree of variability across different speakers. Hence, a speaker verification task will be a good indicator. Speaker verification (SV) is a binary decision task to declare whether a test utterance belongs to a speaker (target model) or not (hence, an outside imposter). Evaluations were carried out on the pool of 42 female speakers, with the individual speaker GMM model constructed using 50% of the data, and the anti-speaker GMM model obtained from 50% of the utterances belonging to 12 speakers. The choice of using 12 speakers for the anti-speaker model is motivated by the stress detection results reported in the earlier subsection, wherein the results indicated that CTM-based system peaked for models formed with 12 training speakers. This allows for a comparison of the best case for CTM with PMIC while having sufficient trials for evaluations.

A single run of SV task consists of scoring test files against both the speaker model and anti-speaker model. If the difference between the scores (target model scores – imposter model scores) is greater than a threshold, the test file is classified as the target speaker, otherwise when the difference is less than or equal to the threshold, the test file is categorized as an outside imposter. To avoid a bias that may exist within a single run, the speaker model is tested against five different anti-speaker models, each time building the anti-speaker model with a different set of 12 speakers. This represents a five-fold cross validation assessment. The results are reported by averaging the results from all five-fold cross validations. The percent Equal Error Rate (%EER) is the measure used to compare SV performance. Table 3 reports the results for the speaker verification tasks. The boldface figures are under matched stress conditions, wherein both the test and train speech belongs to the same stress condition. Fig. 7 shows a DET curve for a single run of speaker verification task for the matched neutral condition. The DET curve for mismatched will intersect the diagonal (which represents the EER line) higher (for e.g.,

Table 3
Speaker verification performance [in terms of EER (%)] for matched and mismatched conditions. For matched conditions, PMIC system outperforms CTM-based system (boldface are figures for matched conditions).

Train	Neutral		Cognitive stress		Physical stress	
Test	CTM	PMIC	CTM	PMIC	CTM	PMIC
Neutral	**1.69**	**0.91**	4.32	11.33	6.39	13.15
Cognitive	6.08	7.10	**1.49**	**0.45**	6.31	12.36
Physical	15.08	16.15	12.93	19.75	**1.80**	**1.42**

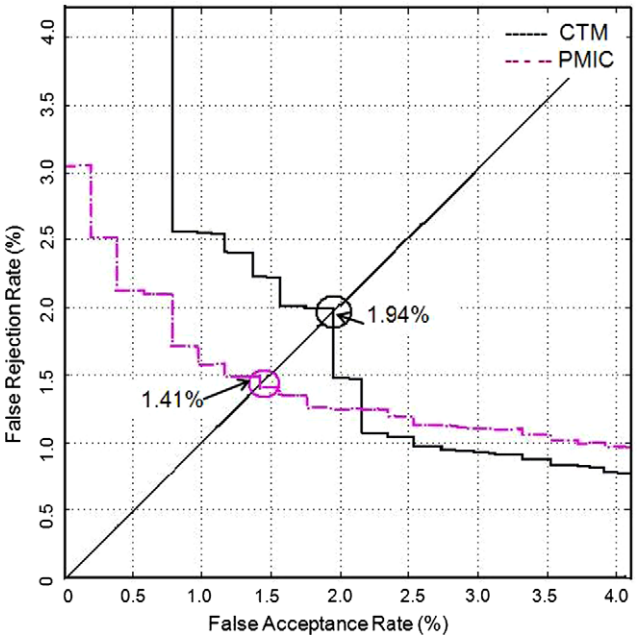

Fig. 7. DET curve for speaker verification task for Neutral conditions, EER with CTM is 1.94%, with PMIC 1.41%, (EER is point as which DET plot intersects the diagonal, lower EER indicates a better system).

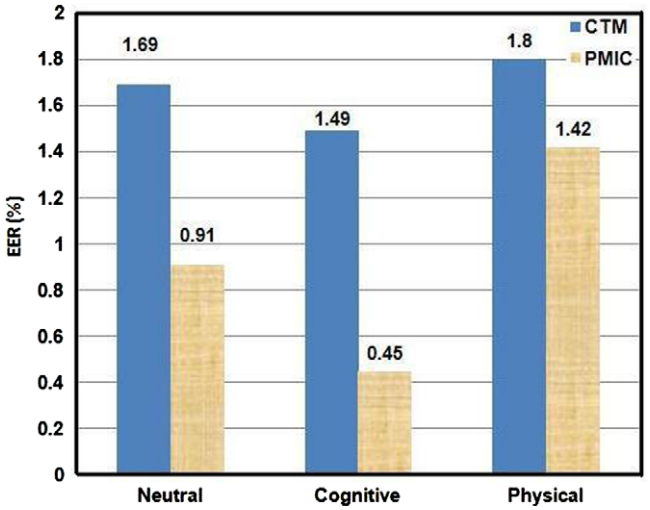

Fig. 8. Speaker verification (ERR %) under matched train-test stress conditions for PMIC and CTM sensor. Lower the value, better the system.

for PMIC sensor data, SV system trained with cognitive stress style files and tested for neutral files, it will interact at 11.33%), but the curve will be similar. DET (Detection Error Trade-off) curve plots variation in two types of errors, reflecting the false acceptance rate (FAR) and false rejection rate (FRR) as a change in threshold. The point on the DET curve where FAR equals FRR is termed the Equal Error Rate (ERR), which lies along the diagonal.

The FRR is a percentage ratio of the true tests being assigned as false versus the total true tests [test signal belonging to the speaker model but decided as NOT belonging to the speaker]. FAR is the percentage ratio of the false tests being decided as true versus the total false tests [test signal NOT belonging to the speaker but decided as belonging to it]. Hence, it becomes obvious that the lower the value of EER, better the resulting system. For details related to speaker verification, please refer to (Bimbot et al., 2004; Reynolds, 1995).

Fig. 8 indicates that under all matched conditions, the PMIC-based SV system outperforms the CTM-based system. For example, under matched cognitive stress conditions (wherein both the models and test files belong to cognitive stress), the CTM-based system has a speaker verification EER of 1.49%, while the PMIC-based yields 0.45% ERR. Thus, PMIC-based Speaker verification system makes fewer errors of assigning target test files as imposters and vice-versa as compared with CTM-based SV in the matched stress condition. Thus, subtle speaker differentiating information is captured by the PMIC to a greater extent as compared to CTM. On a different note, the SV system overall performs better under the cognitive versus physical stress condition, indicating that speakers tend to preserve more consistent speaker dependent structure under the cognitive stress as compared to that under physical stress, or even neutral relaxed state. This highlights speaker traits under cognitive stress in comparison to other states. The SV performance deteriorates for physical stress, indicating that there is a greater increase in stress-related information as compared to speaker-related information under physical stress speech conditions. As the PMIC sensor is close to the neck region, it will capture source excitation information faithfully. Thus, the changes in speech production apparatus due to the physical stress will also be captured. The most predominant changes due to physical stress will be change in the breathing patterns, and also changes to the control on the vocal folds. The changes due to physical stress are likely to overshadow the differences in speech features across different speakers, an attribute essential for speaker verification task. The key findings is that, under matched stress conditions, the PMIC performs better as compared to CTM.

Table 3 indicates performance under matched (diagonal) as well as mismatched (off-diagonal) stress conditions. The performance for speaker verification clearly degrades from neutral to either cognitive or physical stress conditions. The PMIC is measurably better than CTM for matched neutral, cognitive or physical stress conditions (0.45–1.42% EER for PMIC versus 1.49–1.80% for CTM). However, for mismatched conditions, the CTM performs slightly better for neutral models (6.08% versus 7.10% and 15.08% versus 16.15%). Mismatched conditions using cognitive or physical stress models, showed much better EER performance for CTM versus PMIC. Since these error rates are so large, the reason for the difference in CTM versus PMIC performance is left for future work.

From the DET curves and EER table, experiments indicate that the PMIC performs better as compared to CTM under matched stress conditions, thus indicating that PMIC sensor captures speaker traits/information more

effectively as compared to CTM. This is by virtue of its proximity at the cricoid and thyroid cartilage making it possible to capture glottal movement/structure, which might represent some speaker variability.

6. Extracting physiological information

Studies have shown that non-acoustic sensors such as the NAM microphone, SDC, or a combination of multiple sensors can be useful in the study of heart rate and respiratory function, or even assessing lung sounds (Noma et al., 2005; Peters, 1995; Wang and Wang, 2003). Also, studies have been conducted that suggest the existence of a relationship between the amount of physical exertion and heart rate/heart rate variability (Brouha et al., 1961; Brouha et al., 1963; Courteville et al., 1998; Mulder and Mulder, 1981; Maxfield and Brouha, 1963). The ability of the current non-acoustic sensor – PMIC – to provide additional information would in general be impossible with a standard CTM sensor. By virtue of the PMIC being a contact sensor and its proximity to the cartoid arteries (which physicians use to measure heart pulse), it is in theory possible to obtain heart rate information. Thus, if placed correctly, the PMIC can reflect the physiology traits of the speaker. The silent segment of PMIC recordings shows heart-beat pattern. The distance between the two successive peaks which is the time between successive heart-beats is called the Inter-Beat Interval (IBI). For a silent segment of the PMIC sensor recording for speaker performing physical task as shown in Fig. 9, the three successive IBI values are 0.61 s, 0.62 s, and 0.62 s, so for this particular speaker the heart rate does not vary much for the time duration considered in the figure. This pattern is computed using an algorithm described here. The idea in the current set of experiments is to evaluate and study the feasibility of the use of the PMIC sensor for heart rate extraction. Thus, our current focus is not to evaluate and compare performance of different heart rate extraction algorithms, but to illustrate the PMICs ability to capture this knowledge, and therefore a basic autocorrelation approach is implemented with certain constraints to suit our goal. The autocorrelation approach may occasionally produce spurious peaks consisting of period halving or doubling, thus making this approach impractical for a direct heart rate estimator. However, steps are included in the algorithm to suppress these anomalies and limit their impact. Thus, it is possible to draw desired conclusions. Fig. 10 compares algorithm results with an external Polar heart monitor which is used as a baseline reference. The Polar monitor was set to record the subjects heart rate every 5 s. The top part of Fig. 10 shows plots for the neutral relaxed session, while the bottom part shows plots for physical stress session for a specific female speaker (subject: fac1_1). The reported speaker used 254 s to complete the neutral relaxed speech capture session, while 312 s were used for the same speech capture during the physical stress session. Thus, the time taken for a session will vary from session-to-session as well as speaker-to-speaker. The time taken by a speaker defines the duration of the recordings. The evaluation is done for all the speakers, but evaluation statistics are reported for a single speaker (fac1_1) in the key of Fig. 10.

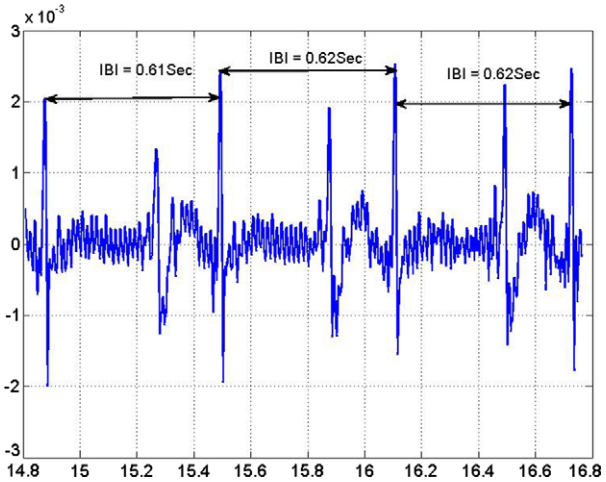

Fig. 9. Speech waveform (magnified) clearly shows heart-beat pattern during the silence segment as recorded with PMIC under physical stress task (time between two successive heart-beats is called Inter-Beat Interval IBI).

Here, heart rate information is extracted from the PMIC signal as follows. An autocorrelation is performed using a window (2 s duration) of PMIC signal data, with a skip rate of 500 ms between subsequent analysis windows. The peak of the autocorrelation signal represents the time elapsed between two heart-beats. This time duration represents the Inter-Beat Interval (IBI) that is, the instantaneous heart rate. The window duration of 2 s is chosen since this duration will include at least two heart-beats for the neutral relaxed state. The minimum heart rate recorded in the UTScope database was 72 beats per minute (which is more than 1 beat per second). Though the heart rate can change between subsequent beats, the constraint that the heart rate will not change by more than 30% as compared to the previous value helps mitigate drawbacks of period halving or doubling associated with traditional autocorrelation based methods. A five-point median filter helps average random spikes which can persist unnoticed. To match the baseline readings from the Polar heart rate monitor, the IBI calculations over a period of 5 s were averaged. The performance of the heart rate algorithm is reported in terms of root mean square error (RMSE) for neutral and physical exertion. RMSE is computed over a complete session duration for the speaker. As reported in Tables 4 and 5, RMSE for the neutral condition is 7.2 beats/min, and for physical stress, it is 5.9 beats/min for a single speaker (speaker tag – fac1_1). The results from Tables 4 and 5 clearly confirm the ability of the PMIC to accurately represent heart rate actively and subject heart rate statistics in both neutral as well as physical stress conditions. This is important in physical stress task such as in current scenario stair-stepper, as neck movement may introduce artifacts which will distort the heart rate measurements.

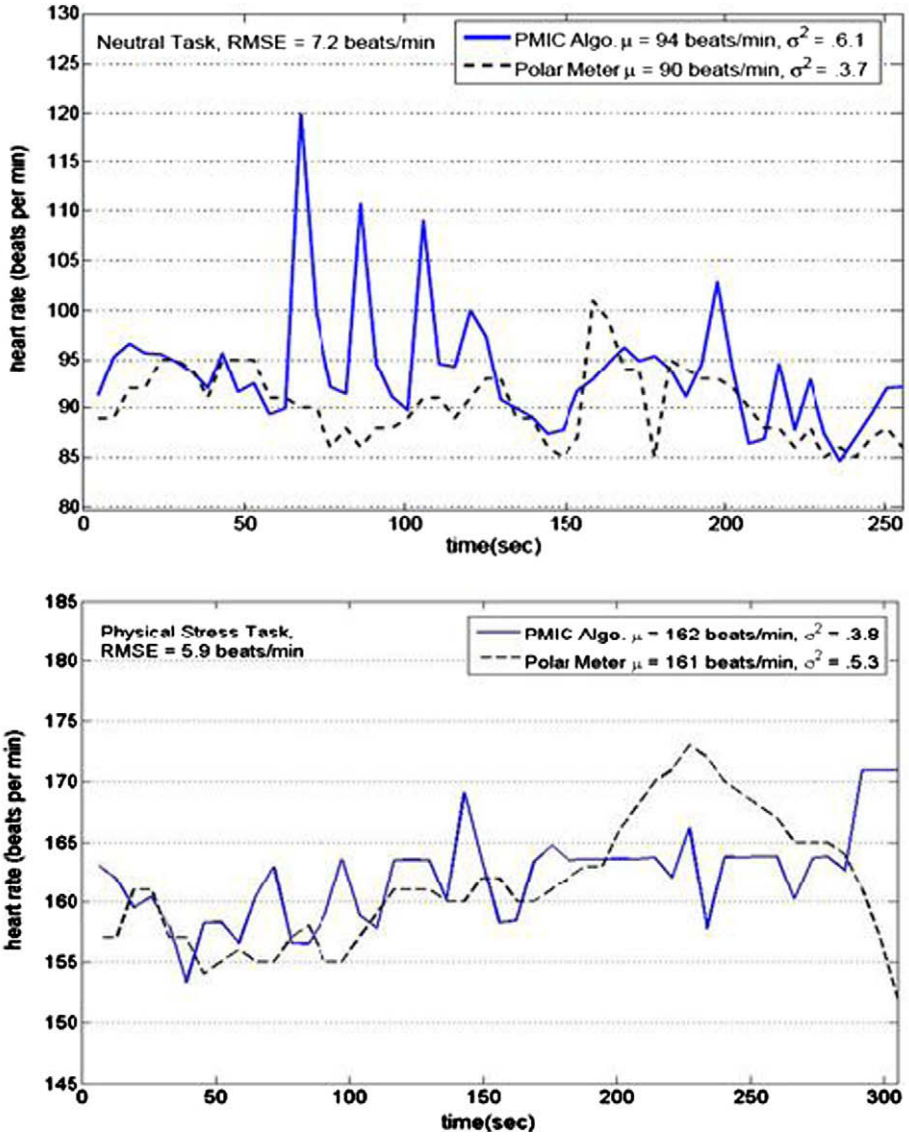

Fig. 10. Heart Rate extraction for the neutral (top) and physical stress (bottom) recordings of a speaker.

These results indicate that the PMIC pcan be employed to extract heart rate information reliably for most speech segments, with spikes reported for neutral recordings in the 50–200 s region (refer to Fig. 10). A listener assessment of the PMIC recordings indicate a lateral movement of the speakers neck, hence introducing unwanted artifacts in the heart rate information. Since we have limited change in heart rate variation across subsequent computation not to exceed $\pm 30\%$ and added a five-point median filtering stage, our algorithm will not show sudden spikes as seen in Polar heart monitor. Furthermore, experiments not only indicate the ability to extract heart rate information, but also that the two different stress conditions have a distinctly different heart rate response, a vital biometric aspect for stress detection and an ideal metric to indicate the presence of stress for the speaker. The plot also indicates an increase in heart rate activity for physical stress as compared to the relaxed neutral task, which is expected, but the increase and the rate of increase in heart rate depends on the phys-

Table 4
Statistics of polar heart rate monitor and physiology information from PMIC (neutral condition) – total task recording 254 s.

Statistical measure	Polar heart rate monitor	From PMIC
Mean (μ beats/min)	90	94
Variance (σ^2)	3.7	6.1
RMSE	7.2	beats/min

Table 5
Statistics of polar heart rate monitor and physiology information from PMIC (physical stress) – total task recording 312 s.

Statistical measure	Polar heart rate monitor	From PMIC
Mean (μ beats/min)	161.4	162.1
Variance (σ^2)	5.3	3.8
RMSE	5.9	beats/min

ical fitness of an individual and the familiarity with performing the required task. Thus, heart rate measurement on its own may not be as useful as a physiological stress indicator, as formulating a mathematical relationship would be a difficult task. In spite of these points, the current experiments do indicate that the PMIC can be a useful tool to extract physiological indicators of stress, which when appended with the stress features previously considered in this study should help stress detection performance/accuracy for practical applications in speech technology. Also, heart rate knowledge can be used to enhance our understanding on speech parameterization and heart rate change. This analysis, along with the ability to record speech signals makes the PMIC a good alternative to CTM in next generation speech technologies.

7. Conclusion

In this study, the prospect of silent speech interface and an alternative sensor (PMIC) for stress detection and speaker recognition was considered. The PMIC performance as compared to a close-talk mic (CTM) was studied under various conditions. The PMIC was shown to perform similar to CTM or better under all conditions. For the stress detection task, a wide range of experiments were carried out to assess the PMIC sensor – to study the impact of different number of speakers in the stress model, different feature-sets, score fusion schemes, performance across different stress types, and also for a speaker verification application.

First, Tables 4 and 5 illustrate a difference in the mean heart rate for neutral and physical stress, for a particular speaker. Secondly, studies have indicated that heart rate variability and breathing patterns can be an indicator of stress. Hence, the additional advantage of the PMIC to represent the physiology state (heart rate and breathing pattern) of the speaker will add a vital dimension/feature for stress detection as well as stress level assessment tasks.

The PMIC sensor is connected near the cricoid cartilage, hence acquiring the articulation-related signal and vocal-fold signal before it leaves the speaker's oral cavity via skin vibration. For a study on the impact of an increasing number of training speakers in the stress model, the SD performance improved with an increase in the number of speakers for PMIC recordings, while performance levels off for CTM at 12 speakers or more in the stress model. This implies that the PMIC contains diverse information requiring more speakers to model stress (PMIC SD performance is 86.27% (93.16%) for 12 training speaker level which improved to 89.42% (96.22%) at 35 speakers level with TEO-CB-AutoEnv (MFCC) features respectively). The PMIC sensor not only helps capture complementary information with two feature-set (TEO-CB-AutoEnv and MFCC), it also contributes for a fusion system combining the two features with the Adaboost-based fusion scheme, indicating that the PMIC sensor provides complementary knowledge to the CTM by about 2% absolute. For two different stress/neutral scenarios, the PMIC outperformed the CTM by +11.24% (neutral/cognitive stress detection), and +5.73% (neutral/physical stress detection) respectively with 12 speakers and a reduced number of mixtures. These experiments indicate that the PMIC may reflect stress-related information to a greater extent as compared to the CTM.

To evaluate the utility of the PMIC at capturing speaker-dependent information, a speaker verification (SV) application was considered. The study on speaker verification (with results compared using DETs and %EER) shows that the PMIC sensor consistently outperforms the CTM for neutral (0.91% EER as compared to 1.69% for CTM), cognitive (0.45% as compared to 1.49% EER), and physical task stress (1.42% as compared to 1.80%) under matched stress conditions. Thus, the PMIC stored speaker specific information to a greater extent as compared to the CTM. Additionally, with the stress mismatched condition (SV task when the train and test stress conditions did not match), the PMIC performed worse as compared to CTM, indicating that stress-related information takes precedence over speaker-related information. Although further experiments are required, we postulate that as the PMIC is connected near the neck, it should capture the speaker dependent glottal excitation signal faithfully and thus impact of stress on the speech production will be more pronounced. This may cause the two speakers to occupy the same feature space degrading the speaker verification performance under stress mismatched condition. Hence, the PMIC indicates worse speaker verification performance as compared to CTM for stress mismatched condition.

Since the PMIC is a skin-contact silent speech interface (SSI), it has an ability to reflect heart rate information. A signal processing scheme based on correlation analysis was used to extract heart rate information from the PMIC, where the heart rate value shows good agreement with an external baseline Polar heart rate monitor, with an RMS error (RMSE) of 7.2 beats per min for neutral condition, and 5.9 beats per min. for physical stress task for a particular speaker. As demonstrated in this study, the heart rate extraction from the PMIC signal is in good agreement with the baseline ground truth, and the PMIC can be exploited to extract useful heart rate information for future speech technologies. Though heart rate cannot be used as a single indicator of stress, it is suggested that when coupled with the stress detection (SD) system presented in this study, the combined performance will certainly improve, even for a reduced number of training speakers in the stress model.

Future work could include robust heart rate extraction strategies and investigate the impact of (i) placement of the sensor at different locations around the throat, (ii) quality of contact between skin and sensor, (iii) impact of body movement, and (iv) impact of breathing patterns on heart rate measurements. It is also of interest to assess the suitability of PMIC under high ambient noise conditions, a

major advantage of the PMIC over CTM sensors. Thus, based on the results from this study, the PMIC is a viable option (as a SSI) for tasks which require humans to operate under different stress conditions, as well as under conditions wherein CTM may limit/impact an individual's mobility for workplace task scenarios.

References

Akargun, U.C., Erzin, E., 2007. Estimation of acoustic microphone vocal tract parameters from throat microphone recordings. IEEE Signal Process. Comm. Appl., 1–4.

Baber, C., Mellor, B., Graham, R., Noyes, J.M., Tunley, C., 1996. Workload and the use of automatic speech recognition: The effects of time and resource demands. Speech Comm. 20 (1–2), 37–53.

Benzeghiba, M., Mori, R.D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., 2007. Automatic speech recognition and speech variability: A review. Speech Comm. 49 (10–11), 763–786.

Bimbot, F., Bonastre, J., Fredouille, C., et al., 2004. A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. 2004 (4), 430–451.

Bosch, L., 2003. Emotions, speech and the ASR framework. Speech Comm. 40 (1–2), 213–225.

Bou-Ghazale, S., 1996. Analysis, modeling and perturbation of speech under stress with applications to speech synthesis and recognition. Ph.D. thesis, Duke University, NC.

Bou-Ghazale, S.E., Hansen, J.H.L., 1995. A source generator based modeling framework for synthesis of speech under stress. In: IEEE Conf. on Acoust., Speech, Signal Process. (ICASSP '95) 1, pp. 664–667.

Bou-Ghazale, S.E., Hansen, J.H.L., 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. Speech Audio Process. 8 (4), 429–442.

Brady, K., Quatieri, T., Campbell, J., Campbell, W., Brandstein, M., Weinstein, C.J., 2004. Multisensor MELPe using parameter substitution. In: IEEE Internat. Conf. on Acoust., Speech, Signal Process. (ICASSP '04) 1, pp. I-477–480.

Brouha, L., Smith, P., Lanne, R., Maxfield, M., 1961. Physiological reactions of men and women during muscular activity and recovery in various environments. J. Appl. Physiol. 16 (1), 133–140.

Brouha, L., Maxfield, M., Smith, P., Stopps, G., 1963. Discrepancy between heart rate and oxygen consumption during work in the warmth. J. Appl. Physiol. 18 (6), 1095–1098.

Brown, D., Keenaghan, K., Desimini, S., 2005. Measuring glottal activity during voiced speech using a tuned electromagnetic resonating collar sensor. Measure. Sci. Technol. 16, 2381–2390.

Burnett, G., 1999. The physiological basis of glottal electromagnetic micropower sensors (GEMS) and their use in defining an excitation function for the human vocal tract. Ph.D. thesis, University of California, Davis.

Chan, C., 2003. Multi-expert automatic speech recognition system using myoelectric signals. Ph.D. thesis, University of New Brunswick, Canada.

Corrigan, G., 1996. Speaker understandability as a function of prosodic parameters. Ph.D. thesis, Northwest University, WA.

Cosmides, L., 1983. Invariances in the acoustic expression of emotion during speech. J. Exp. Psychol.: Human Percept. Perform. 9 (6), 864–881.

Courteville, A., Gharbi, T., Cornu, J.Y., 1998. MMG measurement: A high-sensitivity microphone-based sensor for clinical use. IEEE Trans. Biomed. Eng. 45 (2), 145–150.

Denes, P., Pinson, E., 1993. The Speech Chain: The Physics and Biology of Spoken Language. W.H. Freeman and Company, New York, USA.

de-Paula, M.H., Vinha, C.A., Badini, R.G., 1992. High-sensitivity optical microphone for photoacoustics. Rev. Sci. Instrum. 63, 3487–3491.

Freund, Y., Schapire, R., 1999. A short introduction to boosting. J. Jpn. Soc. Artif. Intell. 14 (5), 771–780.

Gable, T., 2000. Speaker verification using acoustic and glottal electromagnetic micropower sensor (GEMS) data. Ph.D. thesis, University of California, Davis.

Goodie, J., Larkin, K., Schuass, S., 2000. Validation of the polar heart rate monitor for assessing heart rate during physical and mental stress. J. Psychophysiol. 14 (3), 159–164.

Graciarena, M., Franco, H., Sonmez, K., Bratt, H., 2003. Combining standard and throat microphones for robust speech recognition. IEEE Signal Process. Lett. 10 (3), 72–74.

Hansen, J.H.L., 1993. Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments. In: IEEE Internat. Conf. on Acoust., Speech, Signal Process. (ICASSP-93) 2, pp. 95–98.

Hansen, J.H.L., 1994. Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect. IEEE Trans. Speech Audio Process. 2 (4), 598–614.

Hansen, J.H.L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Comm. 20 (1–2), 151–173.

Heuber, T., Chollet, G., Denby, B., Stone, M., 2007. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. INTERSPEECH 2007, pp. 658–661.

Holzrichter, J.F., Ng, L.C., 2002. Speech coding using EM sensor and acoustic signals. In: IEEE 10th Digital Signal Process. Workshop, pp. 35–36.

Huang, R., Hansen, J., Angkititrakul, P., 2007. Dialect/accent classification using unrestricted audio. IEEE Trans. Audio, Speech, Lang. Process. 15 (2), 453–464.

Ikeno, A., Varadarajan, V., Patil, S., Hansen, J.H.L., 2007. UT-Scope: Speech under Lombard effect and cognitive stress. In: IEEE Aerospace Conference, pp. 1–7.

Ingalls, R., 1987. Throat microphone. J. Acoust. Soc. Am. 81, 809.

Jou, S., Schultz, T., Waibel, A., 2007. Multi-stream articulatory feature classifiers for surface electromyographic continuous speech recognition. In: IEEE Internat. Conf. on Acoust., Speech, Signal Process. (ICASSP-2007).

Junqua, J., 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. Speech Comm. 20 (1–2), 13–22.

Klabunde, R., 2005. Cardiovascular Physiology Concepts. Lippincott Williams and Wilkins.

Knudsen, S., Yurek, A., Tveten, A., Dandridge, A., 1994. High-sensitivity fiber optic planar ultrasonic microphone. In: Tenth SPIE Internat. Conf. on Optical Fibre Sensors 2360, pp. 396–399.

Mainardi, E., Davalli, A., 2007. Controlling a prosthetic arm with a throat microphone. In: Annual Internat. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBS 2007), pp. 3035–3039.

Maxfield, M., Brouha, L., 1963. Validity of heart rate as an indicator of cardiac strain. J. Appl. Physiol. 18 (6), 1099–1104.

Mohamad, N., Iovenitti, P., Vinay, T., 2007. High sensitivity capacitive MEMS microphone with spring supported diaphragm. In: Proc. of the SPIE 6800, p. 40.

Mulder, G., Mulder, L., 1981. Information processing and cardiovascular control. Psychophysiology 18 (4), 392–402.

Murray, I.R., Baber, C., South, A., 1996. Towards a definition and working model of stress and its effects on speech. Speech Comm. 20 (1–2), 3–12.

Noma, H., Kogure, K., Nakajima, Y., Shimonomura, H., Ohsuga, M., 2005. Wearable data acquisition for heartbeat and respiratory information using NAM (non-audible murmur) microphone. In: Ninth IEEE Internat. Symp. on Wearable Computers, pp. 210–211.

Otani, K., Hasegawa, T., 1995. The image input microphone – a new nonacoustic speech communication system by media conversion from oral motion images to speech. IEEE J. Select. Areas Comm. 13 (1), 42–48.

Peters, R.D., 1995. Remote respiratory monitor. In: The Eighth IEEE Symp. on Computer-Based Medical Systems, pp. 204–211.

Quatieri, T.F., Brady, K., Messing, D., Campbell, J.P., Campbell, W.M., Brandstein, M.S., Weinstein, C.J., Tardelli, J.D., Gatewood, P.D., 2006. Exploiting nonacoustic sensors for speech encoding. IEEE Trans. Audio, Speech, Lang. Process. 14 (2), 533–544.

Reynolds, D., 1995. Speaker identification and verification using gaussian mixture speaker models. Speech Comm. 17 (1–2), 91–108.

Roucos, S., Viswanathan, V., Henry, C., Schwartz, R., 1986. Word recognition using multisensor speech input in high ambient noise. In: IEEE Internat. Conf. on Acoust., Speech, Signal Process. (ICASSP '86) 11, pp. 737–740.

Scanlon, M., Fisher, F., Chen, S., 2002. Physiological sensors for speech recognition. In: Multimodal Speech Recognition Workshop, pp. 1–5.

Shahina, A., Yegnanarayana, B., 2005. Language identification in noisy environments using throat microphone signals. In: Internat. Conf. on Intelligent Sensing and Information Process, pp. 400–403.

Titze, I.R., Story, B.H., Burnett, G., Holzrichter, J., Ng, L., Lea, W., 2000. Comparison between electroglottography and electromagnetic glottography. J. Acoust. Soc. Am. 107 (1), 581.

Tran, V., Bailey, G., Loevenbruck, H., Jutten, C., 2008. Improvement to a NAM captured whisper-to-speech system. In: INTERSPEECH 2008, pp. 1465–1568.

Viswanathan, V., Karnofsky, K., Stevens, K., Alakel, M., 1984. Multisensor speech input for enhanced immunity to acoustic background noise. In: IEEE Internat. Conf. Acoust., Speech, Signal Process. (ICASSP '84) 9, pp. 57–60.

Wand, M., Jou, S., Schultz, T., 2007. Wavelet- based front-end for electromyographic speech recognition. In: INTERSPEECH 2007.

Wang, H., Wang, L.Y., 2003. Multi-sensor adaptive heart and lung sound extraction. In: Proc. of IEEE Sensors 2, pp. 1096–1099.

Womack, B.D., Hansen, J.H.L., 1996a. Classification of speech under stress using target driven features. Speech Comm. 20 (1–2), 131–150.

Womack, B.D., Hansen, J.H.L., 1996b. Improved speech recognition via speaker stress directed classification. In: IEEE Internat. Conf. Acoust., Speech, Signal Process. (ICASSP-96) 1, pp. 53–56.

Zhou, G., Hansen, J.H.L., Kaiser, J.F., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. 9 (3), 201–216.