Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 906789, 10 pages
doi:10.1155/2011/906789

Research Article

Robust Emotional Stressed Speech Detection Using Weighted Frequency Subbands

John H. L. Hansen, Wooil Kim, Mandar Rahurkar, Evan Ruzanski, and James Meyerhoff

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75083-0688, USA

Correspondence should be addressed to John H. L. Hansen, john.hansen@utdallas.edu

Received 25 September 2010; Revised 10 December 2010; Accepted 10 February 2011

Academic Editor: Julien Epps

The problem of detecting psychological stress from speech is challenging due to differences in how speakers convey stress. Changes in speech production due to speaker state are not linearly dependent on changes in stress. Research is further complicated by the existence of different stress types and the lack of metrics capable of discriminating stress levels. This study addresses the problem of automatic detection of speech under stress using a previously developed feature extraction scheme based on the Teager Energy Operator (TEO). To improve detection performance a (i) selected sub-band frequency partitioned weighting scheme and (ii) weighting scheme for all frequency bands are proposed. Using the traditional TEO-based feature vector with a closed-speaker Hidden Markov Model-trained stressed speech classifier, error rates of 22.5/13.0% for stress/neutral speech are obtained. With the new weighted sub-band detection scheme, closed-speaker error rates are reduced to 4.7/4.6% for stress/neutral detection, with a relative error reduction of 79.1/64.6%, respectively. For the open-speaker case, stress/neutral speech detection error rates of 69.7/16.2% using traditional features are used to 13.1/4.0% (a relative 81.3/75.4% reduction) with the proposed automatic frequency sub-band weighting scheme. Finally, issues related to speaker dependent/independent scenarios, vowel duration, and mismatched vowel type on stress detection performance are discussed.

1. Introduction

Speech conveys multiple levels of information which include the speaker identity, the linguistic structure (text, language, etc.), as well as the emotional and stress state of the speaker over time. The topic of automatic speech recognition (ASR) has been concerned with understanding the underlying linguistic message in the utterance and is not focused on speaker traits such as emotion and stress. Studies have shown that stress and emotion play a substantial role on speech recognition performance [1–4], motivating a number of effective approaches to improve ASR performance when speech is under stress [1–3, 5–8]. Recently, extensive research has been conducted in the speech area to improve the performance of stress/emotion classification [9–14]. Many techniques however require some knowledge of the presence and type of stress.

We believe that the speech production process changes significantly, as the affective state, "stress", of the speaker varies. We consider this one such affective state, "stress", and propose a novel scheme to detect physiological stress in speech. This will enable the recognizer to achieve better performance using, either (i) stressed speech models instead of neutral speech models, (ii) robust stress insensitive features, or (iii) feature compensation or model adaptation to reduce the impact of stress. In recent years, researchers have been interested in identifying reliable acoustic correlates of stress. Some studies have considered a variety of approaches to detect stress based on pitch structure, duration, intensity, glottal characteristics, and vocal tract spectral structure [1, 3, 9, 10, 15]. The deviation of fundamental frequency (f_o) from baseline has been found to be a typically strong indicator of stress [16]. While some of these studies allow replication across experimental results, they have failed to produce

reliable indicators of stress due to subjective differences in the method in which the training/test data was collected as well as subjective decisions. For instance, for f_o, some speakers raise or lower their pitch in response to a stressful situation. The formulation of a successful stress classifier will not only help improve the performance of speech recognition systems, it also could serve as an important information resource for medical, military, or telecommunication applications.

This paper focuses on the problem of automatic detection of speech under stress and employs a previously developed TEO-CB-AutoEnv (Teager Energy Operator-based Critical Band Autocorrelation Envelope) area speech feature. The TEO-CB-AutoEnv area feature is extended in number of ways including (i) an improved selected subband frequency partitioned weighting, (ii) a weighting scheme for all frequency bands, and (iii) discussion of anchor bands in the case of open- and closed-speaker situations. In our previous study, a critical-band probe experiment was conducted where a weighted scheme was evaluated that supported our hypothesis [17]. The initial results showed that specific frequency bands are more sensitive to stress while others are more sensitive to neutral. However, the selected frequency bands were determined by testing over the same speakers in the training set, though the test speech tokens were different from those in the training set. Here we propose a new stressed speech classification algorithm that takes advantage of the differences in sensitivity of the TEO-based feature across critical frequency bands. We also discuss issues in stress detection performance for speaker-dependent versus speaker-independent scenarios and the effect of decreased vowel duration and mismatched vowel type on stress detection performance.

The paper is organized as follows. First, our previously developed TEO-CB-AutoEnv area feature is reviewed as a baseline framework, and a military speech corpus collected in a Soldier of the Quarter (SOQ) paradigm is presented as the evaluation database used in this study. Section 4 develops a selected band weighting scheme for stress classification, and also presents an analysis on frequency band classification sensitivity and its evaluation on a closed-speaker set. In Section 5 an automatic band weighting scheme is proposed, with an evaluation performed on an open-speaker set. Sections 6 and 7 focus on the effects of vowel duration and vowel type on stress detection performance. Finally, Section 8 summarizes the work and draws conclusions.

2. Critical Band-Based TEO Autocorrelation Envelope

Many past research studies on stress have used speech features derived from a linear speech production model which assume that airflow propagates in the vocal tract as a plane wave without dispersing energy in the plane traverse to its direction of propagation. However, studies by Teager [18] and H. Teager and S. Teager [19, 20] have shown that this airflow is actually separated, and that concomitant vortices are distributed through the vocal tract. This observation was supported by theory and experimentation in fluid mechanics as well as by numerical simulation of the Navier-Stokes equation. Previous research has also shown that feature extraction based on nonlinear speech processing (e.g., based on the Teager Energy Operator) is highly successful in the detection of changes in speech production due to the presence of various vocal fold pathologies [28]. In addition, when a speaker is under physiological stress, it is believed that a change occurs in the vocal system physiology during production which further affects the vortex-flow interaction patterns in the vocal tract. We believe that a multidimensional feature obtained across a subband frequency partition would be an effective choice for robust detection of stress. Here our new scheme is based on a previously formulated TEO-CB-AutoEnv feature [9], extending this in a number of ways. This feature has been found to be responsive to speech under stress using audio from the SUSAS corpus (emotional, task stress, or Lombard effect) as well as the emergency-induced stress (NATO SUSC-0 corpus (Speech Under Stress Corpus)) [4].

2.1. Teager Energy Operator. Historically, most approaches to speech modeling have taken a linear plane wave point of view where speech is modeled based on the cross-sectional area along the vocal tract [22, 23]. While features derived from such analysis can be effective for speech coding and recognition, the airflow assumptions are clearly removed from true physical speech modeling. Teager conducted extensive research on nonlinear speech modeling and pioneered the importance of analyzing speech signals from an air-flow energy-based point of view. He devised a simple nonlinear, energy tracking operator, $\Psi[\cdot]$, known as the Teager Energy Operator (TEO), defined for a continuous time signal $x(t)$ as follows:

$$\begin{aligned}\Psi_c[x(t)] &= \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t),\end{aligned} \tag{1}$$

and for a discrete-time signal $x(n)$ as

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

These operators were first introduced systematically by Kaiser [24, 25].

It has been observed [1, 3, 15] that under stressful conditions, a speech signal will display changes in pitch, duration, intensity glottal characteristics, and vocal tract formant structure. It is also known that the fundamental frequency, f_o, will change and hence the distribution pattern of pitch harmonics across a critical frequency band partition will be different than for speech under neutral conditions. Therefore, for a finer resolution of frequencies, the entire audible frequency range can be partitioned into many critical bands. Each critical band possesses a narrow bandwidth (typically 100–400 Hz) thus making this new feature independent of the accuracy of median f_o estimation. This is essential as reliable pitch estimation in emotional speech is difficult, since pitch can increase by more than 200 percent in some high-stress situations [1, 3, 15].

Table 1: Critical band frequency information (bark scale).

Band number	Critical band frequency information (Hz)			
	Lower	Center	Upper	Bandwidth
1	0	50	100	100
2	100	150	200	100
3	200	250	300	100
4	300	350	400	100
5	400	450	510	110
6	510	570	630	120
7	630	700	770	140
8	770	840	920	150
9	920	1000	1080	160
10	1080	1170	1270	190
11	1270	1370	1480	210
12	1480	1600	1720	240
13	1720	1850	2000	280
14	2000	2150	2320	320
15	2320	2500	2700	380
16	2700	2900	3150	450
17	3150	3400	3700	550

2.2. TEO-CB-Auto-Env Feature. The feature extraction procedure can be mathematically summarized as follows using Gabor bandpass filters (BPF) centered at critical-band frequency locations (as shown in Table 1)

$$\begin{aligned} u_j(n) &= s(n) * g_j(n), \\ \Psi_j(n) &= \Psi\Big[u_j(n)\Big] = u_j^2(n) - u_j(n-1)u_j(n+1), \\ R_{\Psi_j^{(i)}(n)}(k) &= \sum_{n=1}^{N-k} \Psi_j^{(i)}(n)\Psi_j^{(i)}(n+k), \end{aligned} \tag{3}$$

where, $g_j(n)$, $j = 1, 2, 3, \ldots, 17$, is the BPF impulse response as shown in Figure 1, $u_j(n)$, $j = 1, 2, 3, \ldots, 17$, is the output of each BPF, "$*$" is the convolution operator, $R_{\Psi_j^{(i)}(n)}(k)$ is the autocorrelation function of the ith frame of the TEO profile from the jth critical band, $\Psi_j^{(i)}(n)$, $j = 1, 2, \ldots, N$, and N is the frame length.

Figure 2 shows a flow diagram of the previously proposed TEO-CB-AutoEnv feature extraction process [9]. As seen in the figure, the TEO profile is segmented on a short-term basis, followed by an autocorrelation operation. The operation is intended to determine the level of "regularity" in the resulting segmented TEO response (for a detailed discussion on capturing the "regularity" of the harmonic excitations for modeling phenomena like voice pathology, see the studies by Michaelis et al. [26, 27].) Once the autocorrelation response is found, the area under the autocorrelation envelope is calculated and normalized over a lag range of 25 msec. A single area coefficient is obtained for each frequency band. The obtained area coefficients have been shown to be large for neutral speech (i.e., speech has high "regularity") and low for speech that is produced with irregular excitation structure (i.e., for speech under stress and/or speech under vocal fold pathology [28]). The TEO-CB-AutoEnv area feature has been shown to reflect variations in excitation characteristics [9] including pitch harmonics. However, we believe that the variation in the excitation structure is not uniformly consistent across all frequency bands.

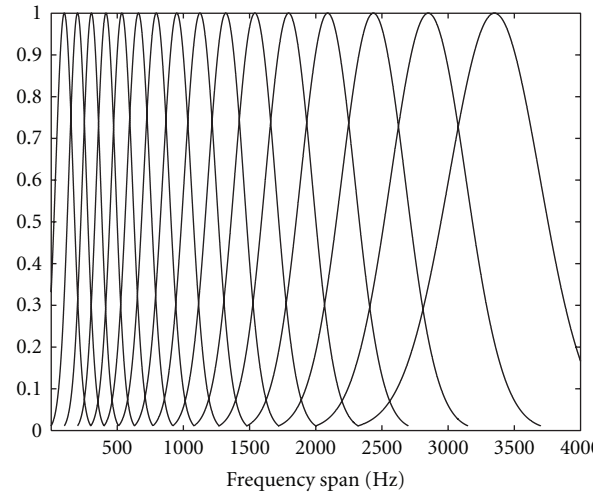

Figure 1: Critical band frequency partition (Hz).

3. SOQ Corpus

In a number of previous studies, we have considered the detection of speech under emotional and task-induced stress using the SUSAS corpus, and NATO SUSC-0 military voice communications corpus [4]. In these corpora, the presence of stress in speech is apparent to listeners (i.e., in most cases stress is quite extreme). We note that SUSAS and SUSC-0 are both effective and useful corpora for analysis and algorithm development in speech spoken under stressful conditions. However, neither corpus contains associated biometrics which can confirm that the subjects were under stress. In such situations, collecting biometric data is typically not necessary as stress levels are extreme. In our study presented here, we are interested in stressed speech detection when the stress is not as apparent as in the SUSAS and SUSC-0 corpora. Therefore, associated biometric data is needed to establish ground truth for building models.

A "speech under stress" corpus was collected by researchers at the Walter Reed Army Institute of Research (WRAIR). The speech corpus was constructed using the WRAIR Soldier of the Quarter (SOQ) paradigm [29, 30], by collecting the spoken response from 6 questions in neutral settings as well as while seated in front of a formal seven person military evaluation board where all board members had a military rank much above the soldier who faced the panel [17]. Table 2 summarizes average speaker conditions for 6 speakers and 7 data collection times before the board (i.e., sets A, B, and C), during the board (i.e., set D), and after the board (i.e., sets E, F, and G). Changes in mean heart rate (HR), blood pressure, both systolic (sBP) and dystolic (dBP), and

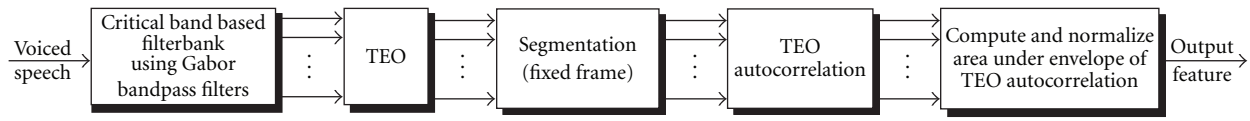

Figure 2: TEO-CB-autoEnv feature extraction flow diagram (as in [9]).

Table 2: Summary of mean biometrics for SOQ subjects: HR: heart-rate in beats per minute, sBP: Systolic blood pressure in mm, dBP: Dystolic blood pressure in mm, f_o: Fundamental frequency in Hz.

Measure	Data set				
	A and B −7 day	C −20 min	D Board	E +20 min	F and G +7 day
HR	70.3	70.8	**93.2**	69.5	67.2
sBP	118	146	**178**	154	117
dBP	77.5	74.8	**89.7**	71.2	69.5
f_o	103.4	102.7	**136.9**	104.3	103.1

pitch (f_o) all confirm a measurable change in speaker state between neutral (sets A, B, C, E, F, and G) and assumed stress condition (set D). For our evaluations, we focused our analysis on the word "no" extracted from the carrier phrase.

4. Algorithm Development Using Closed-Speaker Set

A series of experiments were first performed to help motivate the algorithm development [17]. For these experiments, a Hidden Markov Model (HMM) was used with 3 states and 2 Gaussian mixtures per state. The evaluations were performed on tokens of the word "no" extracted from sentences in the SOQ audio corpus. Here a closed-speaker set was used consisting of six speakers.

4.1. TEO Autocorrelation Envelope Analysis. In this experiment, the goal was to study the area under the TEO autocorrelation envelope across 16 frequency bands for neutral and stress speech conditions. As previously noted, a change in the area under the autocorrelation envelope is expected to reflect a change in the regularity or consistency in the excitation structure during speech production. Figure 1 shows the shape of the frequency band partition and Table 1 summarizes the center frequencies, bandwidths, and cutoff frequencies for all critical bands. The TEO-CB-AutoEnv area feature was calculated across all speakers and averaged for all sixteen bands. Using the critical band frequency structure from Figure 1, the TEO-CB-AutoEnv area features were extracted on a frame-by-frame basis for all neutral and stress tokens of the word "no". Figure 3 shows the average feature profile before the board and after the board averaged across all speakers. The solid line (with solid-circle) in the plot represents the stressed speech scenario. We observe that the total area under the autocorrelation envelope, referred to as frequency-band area feature, is measurably distinct for a number of frequency bands in neutral and stress tokens.

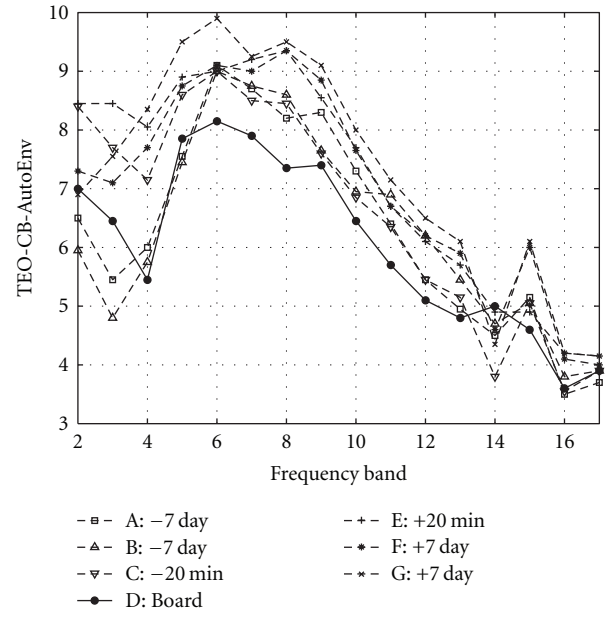

Figure 3: Area under autocorrelation envelope across all 16 bands.

The TEO-CB-AutoEnv area feature is smaller in magnitude for low- and mid-band frequency bands (i.e., bands 4 and 6–8) for stress versus neutral conditions. For high-band frequency bands, the stress condition generally produced the smallest value, while for band 14 it was the largest. These results strongly suggest a frequency-dependent nature of TEO-CB-AutoEnv area feature for speech under stress.

4.2. Band Classification Sensitivity for Neutral versus Stressed Speech. The results so far suggest a frequency sensitive nature for the TEO-CB-AutoEnv area feature. Next, we determine if some bands are more reliable in their ability to detect neutral or stressed speech. Therefore, a series of stress classification experiments was performed where stress detection was based on single individual frequency bands. To accomplish this, a neutral/stress classification experiment was performed where an HMM classifier was trained for each critical frequency band. Neutral versus stress detection was performed individually on each band. Figure 4 shows results for both stressed and neutral speech classification using the frequency band partition summarized in Table 3. We observe that bands 6, 7, 10, and 14 are very sensitive to neutral speech (i.e., above 85% correct neutral classification), while bands 8, 13, 15 and 17 are sensitive to speech under stress (i.e., above 70% correct stress classification). Moreover we also observe that bands which are sensitive to stress are complementary to those sensitive to neutral. Note that all stress classification

Table 3: Percentage error rate in stress/neutral recognition for individual frequency bands 2–17 and 3 sets of 4 band groups.

Band	Stress	Neutral
2	38.09	28.04
3	57.14	22.35
4	57.14	38.99
5	38.09	32.91
6	76.19	13.97
7	66.67	13.97
8	28.57	58.94
9	80.95	14.92
10	80.95	13.07
11	66.67	22.46
12	33.33	62.46
13	23.81	77.54
14	71.43	11.21
15	23.81	67.22
16	42.86	73.84
17	14.29	67.25
2, 3, 4, 5	42.86	15.87
6, 7, 8, 9	47.62	24.15
14, 15, 16, 17	57.14	8.37

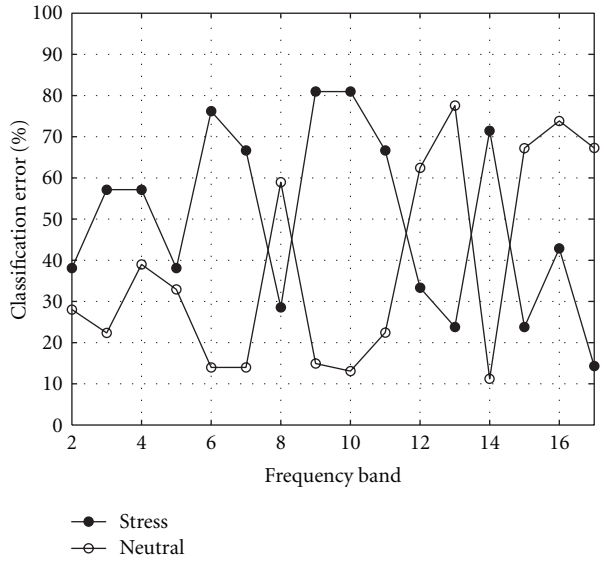

Figure 4: Stress and neutral speech classification results for individual frequency bands.

rates are based on "single" phonemes tests using the /OW/ phoneme extracted from the word "no".

4.3. HMM Baseline Classification System. Having identified frequency bands which are more sensitive to either neutral or stressed speech conditions, we now turn to formulating a frequency band sensitive stress classifier. Before embarking on this task, we consider a baseline classifier which uses all frequency bands. A baseline HMM system was formed using audio material from the SOQ corpora. Acoustic models consist of three-state HMMs with two Gaussian mixtures per each state. A total of 191 tokens was used for training the neutral model, while 30 tokens were used for training the stress model in a traditional round-robin manner (i.e., train with 80% of the data and test with 20% of the data). The front-end feature consists of a 16-dimensional TEO-CB-AutoEnv area feature vector. The speech data obtained during the SOQ Board scenario (e.g., token set D) was assumed to be under "stress", and the remaining speech data was grouped together as a "neutral" set based upon reported biometric results. Thus, we obtained 2 HMM models termed "Neutral" and "Stress" after the training phase. Using the entire critical band TEO-CB-AutoEnv feature, a round-robin open error classification rate was found to be 22.5% for stress and 13.0% for neutral tokens.

4.4. HMM Modeling for Frequency Band Analysis. For frequency band analysis, a second HMM classification system was developed with a feature made up of the TEO-CB-AutoEnv area feature for each individual band, formulating an independent system. A separate "Neutral" and "Stress" model was therefore constructed for every band in a manner similar to that discussed in Section 4.2. In addition to single band neutral and stress models, we also trained models using the first 4 bands (i.e., bands 2–5), bands 6–9, and the last 4 bands (i.e., 14–17) grouped together, which we believe will play an important role in establishing classifiers that can distinguish between neutral and stressed speech. Thus, we have 32 single band models, 16 of which are neutral and 16 under stress. We also have 6 four-band models, again organized in a similar manner.

4.5. Stress Classification Employing Weighted Band Scoring Scheme. In this section, we develop a novel scheme for speech under stress detection based on the findings from the preceding section. The approach is to construct a weighted band scoring scheme in which each band is assigned a weight, depending upon its sensitivity to a stress or neutral condition, with the constraint that all weights sum to unity. The weights used in the formulation are determined experimentally, and the same sets of weights were used for all evaluations in their respective categories (i.e., "Stress" or "Neutral" classification). The equation below shows how individual band HMM scores are weighted using stress and neutral sensitive bands to obtain an overall stress classifier decision:

$$\text{Score} = \sum_{n=1}^{4} W_{(n)} \text{SNB}_{(n)}^{(j)} - \sum_{n=1}^{4} W_{(n)} \text{SSB}_{(n)}^{(j)}, \tag{4}$$

where, $\text{SNB}_{(n)}^{(j)}$ = Score of Sensitive Neutral Bands: $j = 6, 7, 10, 14$ corresponds to $n = 1, 2, 3, 4$. $\text{SSB}_{(n)}^{(j)}$ = Score of Sensitive Stress Bands: $j = 8, 13, 15, 17$ corresponds to $n = 1, 2, 3, 4$. (Note: $n = 1$ corresponds to $j = 6$ in the neutral case and $j = 8$ in stress case, etc.) $W_{(n)}$ = band "n" weight, $n = 1, 2, 3, 4$.

Table 4: Evaluation using new 4-band weighted stress detection scheme for closed-speaker scheme.

System	Error in stress (%)	Error in neutral (%)
Baseline	22.5	13.0
Weighted CB	**4.7**	**4.6**

Experimental evaluations were performed as outlined in the previous section. The frequency-band analysis results using the new detection scheme are shown in Table 5. Using the entire TEO-CB-AutoEnv area feature from the entire frequency range, baseline stress and neutral error rates are 22.5% and 13.0%. Using the results from the experimental procedure discussed in Section 4.2 to establish stress and neutral sensitive bands, our new weighted subband stress classification algorithm using (4) was able to achieve error rates of 4.7% and 4.6% for stress and neutral speech detection, respectively. This corresponds to relative 79.1% reduction in the stress speech detection error rate, and a 64.6% relative reduction in the neutral speech detection error rate.

5. Open-Speaker Stress Classification Employing Automatic Band Weighting Scheme

In the previous section, the critical-band probe experiment was conducted, which showed that specific frequency bands are more sensitive to stress while others are more sensitive to neutral. However, the selected frequency bands were determined by testing over the same speakers in the training set, though the test speech tokens were different from those in the training set. In this section we propose a new stressed speech classification algorithm that takes advantage of the differences in sensitivity of the TEO-based feature across critical frequency bands. In a same manner as the closed-speaker system in Section 4, an HMM classifier with 3-state with 2-Gaussian mixture is used for an evaluation. The evaluation was performed also on the same SOQ corpus using the extracted word "no" for 6 speakers.

5.1. Baseline System Development. For our baseline evaluation, the available SOQ corpus was divided into 3 sets: training set, development test set, and an open test set. Four speakers were used for training, one for development, and one speaker was set aside for testing, thus allowing us to carry out open-speaker evaluation. Since the corpus is not large enough to allow for a large independent test set, we performed a round-robin procedure using the 6 speakers, where each speaker was tested against the HMM trained using the combination of the remaining 5 speakers. Thus, each of the 5 training speakers acted as a development speaker. Baseline results were averaged over all 30 of these evaluations.

5.2. HMM System Development for Frequency Band Analysis. In a same manner as the closed-speaker set, for frequency band analysis, a second HMM classification system was trained with a feature made up of the TEO-CB-AutoEnv area feature from each individual band, resulting in an independent HMM system for each band, where a separate "Neutral" and "Stress" model is available. Therefore, 34 single band models for 17 neutral and 17 stress, respectively, are built. Evaluations were carried in the same manner as in the baseline evaluation using three different sets for training, development, and testing. Development sets were used to determine the band-weights using the new band-weighting scheme discussed in the following section.

5.3. Automatic Band Weighting Scheme. In Section 4.5, the effectiveness of a subband-based stress classification scheme for a closed group of individual speakers was demonstrated. However, we realize that the selected bands for classifying emotional speech may not be consistent across different speakers. Hence, in this section all sub-bands in a progressive weighting scheme are employed instead of selecting sub-bands with simply an overall low error percentage. One issue is to determine the balance of subband performance between stress and neutral speech, even if it is very subtle. In order to address this issue, we developed a novel automatic weighting scheme for bands where weights are assigned based on performance from training and development test sets. The scheme takes full advantage of prior stress/neutral speech analysis and is also computationally simple. In the proposed scheme, first the N most successful bands for each speaker are identified at every evaluation of the development set, generating a frequency distribution for each band. Next, by summing up the frequency distributions across all speakers and normalizing them, a speaker-independent weight for each band can be obtained. The speaker-independent weights are obtained separately for neutral and stressed speech.

Here, for neutral speech, the frequency distribution for each band that occurs in the top 5 positions in each evaluation was computed (i.e., we test each speaker for each of the 17 critical bands and identify the 5 most successful bands for each speaker). Tables 5 and 6 summarize these results under neutral and stress conditions, respectively. The frequency distributions results were summed for all speakers so as to produce generic weights. Next, each speaker frequency distribution was normalized by dividing each subband by the sum, for a speaker-independent framework. As can be seen in Tables 5 and 6, all the subband weights sum to 1. For stressed speech, the frequency distribution was computed by selecting the bands in the top 4 positions. The selection of the number of bands to use was determined empirically. The last column in Tables 5 and 6 shows the computed weights which are obtained by using the frequency distribution term (FD) divided by the total number of band hits (i.e., 150 for neutral and 120 for stress). After computing the speaker-independent weights, these weights are incorporated into the subband stress classification scheme. Therefore, the classification algorithm employs a combination of weighted subband scores instead of simply selecting from the 4 highest

Table 5: Weights for neutral speech (where FD is the combined frequency distribution across all speakers).

Band	Weights for neutral speech							
	Spkr1	Spkr2	Spkr3	Spkr4	Spkr5	Spkr6	F.D	Weight (i)
1	4	3	4	1	2	4	18	0.12
2	4	5	4	5	5	5	28	0.19
3	1	2	0	0	1	0	4	0.03
4	0	2	1	1	0	1	5	0.03
5	1	2	0	1	0	1	5	0.03
6	1	0	1	0	1	1	4	0.03
7	5	4	5	5	5	3	27	0.18
8	1	0	0	0	0	1	2	0.01
9	0	0	2	2	5	1	10	0.07
10	1	1	1	0	1	1	5	0.03
11	3	2	1	3	3	4	16	0.106
12	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0
14	3	2	4	5	2	1	17	0.11
15	0	0	0	0	0	1	1	0.01
16	0	0	0	0	0	0	0	0
17	1	2	2	2	0	1	8	0.05
Sum	25	25	25	25	25	25	150	1

Table 6: Weights for stressed speech (where FD is the combined frequency distribution across all speakers).

Band	Weights for stressed speech							
	Spkr1	Spkr2	Spkr3	Spkr4	Spkr5	Spkr6	F.D	Weight (i)
1	4	3	0	1	2	4	14	0.1166
2	3	4	1	3	4	3	18	0.1500
3	0	0	0	0	1	0	1	0.0083
4	0	2	2	1	0	1	6	0.0500
5	1	2	3	0	0	1	7	0.0583
6	0	0	2	0	1	1	4	0.0333
7	5	4	0	5	5	3	22	0.1833
8	1	0	2	0	0	1	4	0.0333
9	0	0	0	2	5	1	8	0.0666
10	0	1	2	0	0	0	3	0.0250
11	3	1	0	3	1	4	12	0.1000
12	0	0	1	0	0	0	1	0.0083
13	0	0	4	0	0	0	4	0.0333
14	3	2	0	3	1	1	10	0.0833
15	0	0	2	0	0	0	2	0.0166
16	0	0	1	0	0	0	1	0.0083
17	0	1	0	2	0	0	3	0.0250
Sum	20	20	20	20	20	20	120	1

scores as discussed in Section 4.5. The weighted score is calculated below based on our development in(4)

$$\text{Score} = \sum_{n=1}^{17} W_{(n)}\text{NCS}_{(n)} - \sum_{n=1}^{17} W_{(n)}\text{SCS}_{(n)}, \tag{5}$$

Table 7: Evaluation using new detection scheme for open-speaker scheme.

System	Error in stress (%)	Error in neutral (%)
Baseline	69.7%	16.2%
Autoweighted CB	**13.1**%	**4.0**%

where, $\text{SCS}_{(n)}$ = Stress Classification Score from Subband n, $\text{NCS}_{(n)}$ = Neutral Classification Score from Subband n, $W_{(n)}$ = band "n" Weight, $n = 1, \ldots, 17$.

The results from these evaluations using this proposed stress classification algorithm are shown in Table 7. Using the TEO-CB-AutoEnv area feature vector directly produces an open-speaker classification baseline stress and neutral error rate of 69.7% and 16.2%, respectively. We note here that these open-speaker set error rates are quite different from those reported in Table 4 for the set from the closed-speaker case. Using the proposed automatic subband weighting scheme, the percentage stress and neutral error rates drop to 13.1% and 4.0%, respectively. This corresponds to a relative 81.3% reduction in stress speech detection error rate, and a 75.4% percent reduction in neutral speech detection rate. While the error reduction is significant, the more important result is that the autoweighted CB stress classification results for open-speaker set evaluations have moved closer to the results seen for the closed-speaker set evaluations. This suggests some degree of reduction in the level of speaker-dependent structure, and focusing the classification on traits which are dependent on stress or emotion dependent. In this paper, a simple scheme for the speaker-independent weights was employed to address a limited amount of data, however, other conventional approaches (i.e., SVD (Singular Vector Decomposition), PCA (Principle Component Analysis), etc.) could also be explored to identify the anchor frequency components.

6. Effects of Phoneme Duration on Stress Detection Performance

In this section, we discuss the effect of phoneme duration on performance of stress detection. Similar to the experiments discussed in previous sections, the 8-kHz digitized speech data from the SOQ board was processed for isolation of the vowel /OW/ from each of the original 42 sentences from each of the original corpus of 6 speakers. Each extracted sample of the vowel /OW/ was verified to be valid, and manual processing was done to ensure a complete and accurate representation of the vowel /OW/. This requires removal of audible instances of the phoneme /N/ from the complete word "no" as well as removal of trailing silence after /OW/. These manipulations are critical for phoneme duration testing. These extracted vowels were considered to be the "100%-duration" vowels. The probability density functions (PDF) depicting the time duration probability distributions for the neutral and stress sets are shown in Figure 5, with mean values of 214 msec, and 203 msec and standard deviation values of 63 msec and 63 msec for the neutral and stress set, respectively. The similar mean

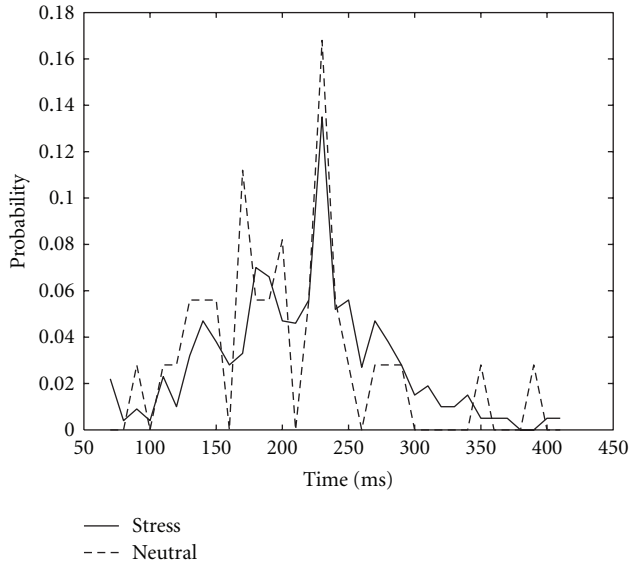

Figure 5: Probability density functions of duration of vowel /OW/ tokens.

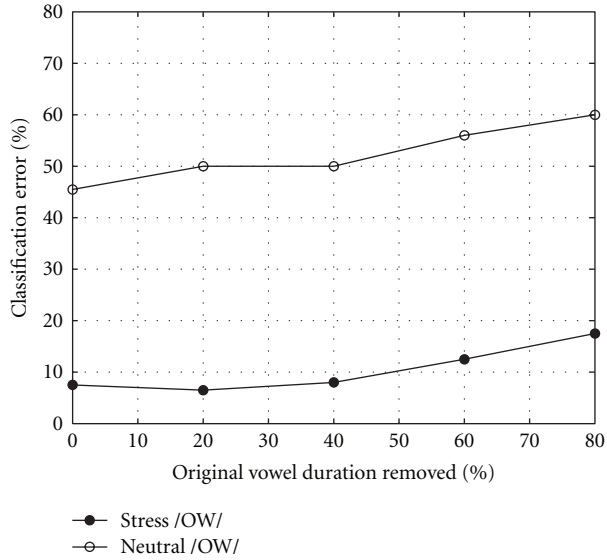

Figure 6: Vowel duration effect on baseline TEO stress detection performance.

and standard deviation values suggest the duration of the vowel /OW/ is not affected by stress content in the SOQ data.

Each of the "100%-duration" vowel samples was truncated (by taking samples from the back-end) to specified durations of 80, 60, 40, and 20% of the 100% vowel duration length. The TEO feature was extracted on a frame-by-frame basis from each vowel instance in each of these five duration sets. Similar to the studies performed in previous sections, the frame lengths for the feature extraction were fixed to 200 samples. The amount of shift of this 200-sample frame was set to the values of 100 samples and 25 samples, constant across the full set of vowel tokens. The hypothesis was that in setting the frame shift to a lower value (100 samples were used in the testing accomplished in the previous sections) more TEO feature values could be computed in a given duration sample. This fact will become critical in the HMM testing and scoring of the lower (i.e., 40 and 20%) vowel durations.

For the vowel duration study, the critical bands in the TEO extracted features were equally weighted (i.e., baseline tests were performed). The band-weighting scheme was not used in the duration testing. The goal of this experiment was to determine the effect of vowel duration on stress detection performance and so we have deferred introducing the variable band-weighting scheme, as that would introduce an additional analysis dimension. This explains the relatively high error percentages for these experiments versus those reported above. The extracted features from the "100%-duration" vowel /OW/ were used to train the HMMs used for the round robin testing. As above, 3-state, 2-Gaussian mixture HMMs were trained and used in the testing phase. Only the "100%-duration" vowels were used in the HMM training and testing was performed with 100% and progressively shorter duration vowel test material.

Figure 6 illustrates the effect of vowel duration on the performance of the TEO stress detection scheme using TEO features extracted at a frame shift rate of 25 samples. From the examination of Figure 6, it is apparent that stress detection performance remains nearly constant up to removal of between 40 and 60% of the original vowel duration for both neutral and stress condition speech. This suggests the necessary mean duration values for speech tokens to yield desirable stress detection performance are approximately 85.6–128.4 msec for neutral and 81.2–121.8 msec for stress condition speech, based on mean durations shown in Figure 5. This exercise therefore suggests that a phoneme duration threshold should be set for effective stress detection.

A relatively large number of TEO feature samples is necessary to ensure proper amounts of feature data are used to train the HMMs. In the 100-sample frame shift case, at 40% original vowel duration, 32/216 neutral test tokens did not contain enough information to be scored properly by the HMMs. At 20% vowel duration, this number increased to 105/216. In the 25-sample frame shift case, at 40% vowel duration, all tokens were properly scored by the HMM and at 20% vowel duration, only 16/216 of the tokens did not contain enough information to be properly scored. No further analysis of the frame shift was deemed necessary, as at 20% vowel duration in the 25-sample frame shift case, only 7.4% of the tokens were not useful for the test. The 25-sample frame shift will be used for the remainder of the TEO feature extractions in this study.

7. Analysis of Vowel Type Difference on Stress Detection Performance

As with the duration experiment above, the vowel type is isolated in the following way. First, we employed the baseline stress detection scheme to eliminate the effects of vowel

Table 8: Multistyle HMM stress detection performance across vowel types.

Test vowel types	Error in neutral (%)	Error in stress (%)
/AE/	31.25	37.14
/AX/	44.17	54.28
/IY/	33.90	38.89
/OW/	33.10	50.00
Overall (Std.)	**35.60 (5.82)**	**45.06 (8.37)**

type on the critical band weighting. Next, the duration of each token in the set is fixed to 50 msec. We note that since the word "no" is monosyllabic and a keyword in the sentence response, the /OW/ phoneme was typically longer in duration than the other vowels. This value is chosen to allow for a reasonable number of tokens across the vowel types thus allowing a broad test set among vowel types.

The vowels used for the type testing include /AE/, /AX/, /IY/, and /OW/. The collection of full-length vowels among these vowel types across the sentences in the test set was used to train a 3-mixture, 3-state multistyle HMM. The multistyle HMM was employed to account for the various vowel types in one model, and the number of Gaussian mixtures was increased to three for better resolution between vowel types. The results are summarized in Table 8.

The results in Table 8, namely, the standard deviation values, show small variability in stress detection across vowel types when using the multistyle HMM trained on the full-duration vowel tokens, with some vowel types performing slightly better under neutral and stress conditions. It is noted that test materials were 50 msec in duration, which would include at least one test block for most vowel sections extracted (i.e., from Figure 5, on average there would be between 2–8 test blocks available from a single /OW/ vowel section). It can be seen from the results in Table 8, the error rates average 35.60% for neutral and 45.06% for the stress condition.

8. Conclusion

In this paper, we have proposed a novel algorithm for stressed speech detection. This approach was based on nonlinear analysis using features derived from the TEO. Speech data obtained from an SOQ paradigm developed at WRAIR independently showed a statistically significant change in blood pressure, heart rate, and salivary hormone levels between neutral and stress conditions. Although this corpus is small, it is the first and presently only corpus of speech under stress with independent biometric data to suggest the presence of physiological stress.

Individual stress detection experiments across critical subband frequencies showed some bands to be more sensitive for stress detection, while others were sensitive to neutral speech. Objective evaluations showed that this novel scheme leads to a substantial improvement in stress detection performance. One of the main drawbacks in most studies on emotion recognition is the lack of a benchmark database to test different algorithms. Also, the knowledge of ground truth regarding the presence of stress or emotion is typically lacking. However in our case, statistical analysis of the biometric data suggests that test subjects were in a significantly different state, and it is reasonable to assume, by experimental design, that emotional stress caused this change in state.

The experimental results using a classifier which weighted the top 4 frequency bands showed a substantial improvement over a classifier that used the entire TEO-CB-AutoEnv feature vector. This motivated the formulation of an automatic critical band weighting scheme for closed-speaker and open-speaker stress classification. The closed-speaker-set algorithm produced a stress/neutral classification error rates of 4.7/4.6% versus the baseline rates of 22.5/13.0%. The open-set speaker system gave stress/neutral rates of 13.6/4.0% versus the baseline rates of 69.7/16.2%. The results here strongly suggest that a frequency-band weighting approach is more effective for stress classification and makes progress in helping reduce speaker dependencies in open-speaker stress classification.

We have also shown the effects of phoneme duration and type on automatic TEO feature-based stress detection performance. It was shown that both phoneme duration and type mismatch affect the stress detection performance. In the case of vowel duration, shortening the vowel duration was shown to adversely affect stress detection performance if the vowel duration is less than a threshold of about 50% of the original duration in the case of the vowel /OW/ (i.e., the duration needs to be about 85–128 msec). In the case of vowel type, it was shown that stress detection performance varies among and between the different vowel types whose durations were sufficiently long for use in the HMM-based stress detection scheme.

Acknowledgments

The authors thank George Saviolakis and Michael Koenig (WRAIR) for their efforts on the data collection for this study. This research was funded in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen, DARPA through SPAWAR under Grant no. N66001-00-2-8906, and the U.S. Army Medical Research and Materiel Command. Volunteers participated in this research only after having given their free and informed consent under protocol WRAIR #1116 and #131 (HSRRB Log #A-4256). Investigators adhered to AR 70–25 and USAMRDC Reg 70–50 on the use of volunteers in research. The views of the author(s) do not purport to reflect the position of Department of the Army or the Department of Defense, (para 4-3, AR 360-5).

References

[1] J. H. L. Hansen, *Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition*, Ph.D. Thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA, USA, 1998.

[2] J. H. L. Hansen and M. A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 407–415, 1995.

[3] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 1-2, pp. 151–173, 1996.

[4] J. H. L. Hansen, C. Swail, A. J. South et al., "The impact of speech under 'stress' on military speech technology," Tech. Rep. NATO IST/TG-01, 2000.

[5] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[6] R. Sarikaya and J. H. L. Hansen, "High resolution speech feature parameterization for monophone based stressed speech recognition," *IEEE Signal Processing Letters*, vol. 7, no. 7, pp. 182–185, 2000.

[7] B. D. Womack and J. H. L. Hansen, "Classification of speech under stress using target driven features," *Speech Communication*, vol. 20, no. 1-2, pp. 131–150, 1996.

[8] J. H. L. Hansen and S. E. Bou-Ghazale, "Robust speech recognition training via duration and spectral-based stress token generation," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 415–421, 1995.

[9] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[10] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *Journal of the Acoustical Society of America*, vol. 96, no. 6, pp. 3392–3400, 1994.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[12] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, pp. 801–804, September 2006.

[13] V. Sethu, E. Ambikairajah, and J. Epps, "Empirical mode decomposition based weighted frequency feature for speech-based emotion classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 5017–5020, April 2008.

[14] W. Kim and J. H. L. Hansen, "Angry emotion detection from real-life conversational speech by leveraging content structure," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '10)*, pp. 5166–5169, 2010.

[15] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 307–313, 1996.

[16] C. E. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates," *Journal of the Acoustical Society of America*, vol. 52, no. 4, pp. 1238–1250, 1972.

[17] M. A. Rahurkar, J. H. L. Hansen, M. A. Oleshansky, J. L. Meyerhoff, and M. Koenig, "Frequency band analysis for stress detection using a teager energy operator-based feature," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 2021–2024, Denver, Colo, USA, September 2002.

[18] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[19] H. Teager and S. Teager, "A phenomenological model for vowel production in the vocal tract," in *Speech Science: Recent Advances*, R. G. Daniloff, Ed., pp. 73–109, College-Hill, San Diego, Calif, USA, 1983.

[20] H. Teager and S. Teager, "Evidence for nonlinear production mechanisms in the vocal tract," *Speech Production and Speech Modeling*, vol. 55, pp. 241–261, 1990.

[21] S. E. Bou-Ghazale and J. H. L. Hansen, "Stress perturbation of neutral speech for synthesis based on hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 201–216, 1998.

[22] J. L. Flanangan, *Speech Analysis, Synthesis and Perception*, Springer, New York, NY, USA, 1965.

[23] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, Calif, USA, 1978.

[24] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '90)*, pp. 381–384, April 1990.

[25] J. F. Kaiser, "Teager's energy algorithm, its generalization to continuous signals," in *Proceedings of the 4th IEEE Digital Signal Processing Workshop*, 1990.

[26] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio: a new measure for describing pathological voices," *Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[27] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for description of pathologic voices," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[28] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 300–313, 1998.

[29] J. L. Meyerhoff, M. A. Oleshansky, and E. H. Mougey, "Psychologic stress increases plasma levels of prolactin, cortisol, and POMC-derived peptides in man," *Psychosomatic Medicine*, vol. 50, no. 3, pp. 295–303, 1988.

[30] M. A. Oleshansky and J. L. Meyerhoff, "Acute catecholaminergic responses to mental and physical stressors in man," *Stress Medicine*, vol. 8, no. 3, pp. 175–179, 1992.