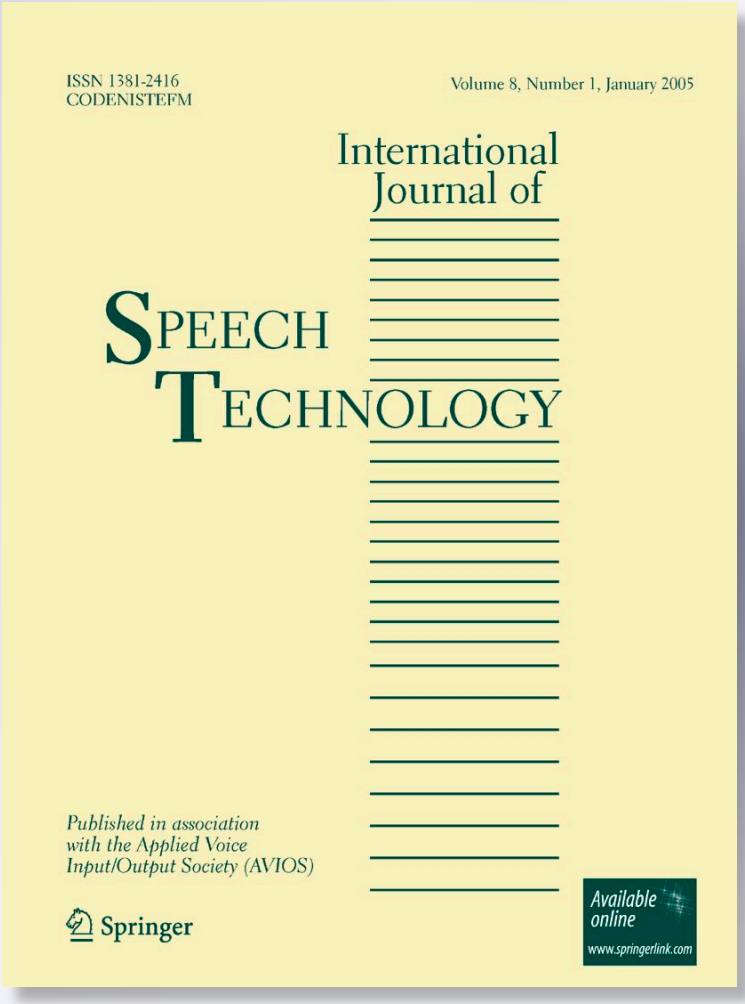
TEO-based speaker stress assessment using hybrid classification and tracking schemes

John H. L. Hansen, Evan Ruzanski, Hynek Bořil & James Meyerhoff

International Journal of Speech Technology

ISSN 1381-2416
Volume 15
Number 3

Int J Speech Technol (2012) 15:295-311
DOI 10.1007/s10772-012-9165-1



Springer

Int J Speech Technol (2012) 15:295–311
DOI 10.1007/s10772-012-9165-1

TEO-based speaker stress assessment using hybrid classification and tracking schemes

John H.L. Hansen · Evan Ruzanski · Hynek Bořil · James Meyerhoff

Received: 13 March 2012 / Accepted: 14 June 2012 / Published online: 29 June 2012

Abstract Speaker variability is known to have an adverse impact on speech systems that process linguistic content, such as speech and language recognition. However, speech production changes in individuals due to stress and emotions have similarly detrimental effect also on the task of speaker recognition as they introduce mismatch with the speaker models typically trained on modal speech. The focus of this study is on the analysis of stress-induced variations in speech and design of an automatic stress level assessment scheme that could be used in directing stress-dependent acoustic models or normalization strategies. Current stress detection methods typically employ a binary decision based on whether the speaker is or not under stress. In reality, the amount of stress in individuals varies and can change gradually. Using speech and biometric data collected in a real-world, variable-stress level law enforcement training scenario, this study considers two methods for stress level assessment. The first approach uses a nearest neighbor clustering scheme at the vowel token and sentence levels to classify speech data into three levels of stress. The second approach employs Euclidean distance metrics within the multi-dimensional feature space to provide real-time stress level tracking capability. Evaluations on audio data confirmed by biometric readings show both methods to be effective in assessment of stress level within a speaker (average accuracy of 55.6 % in a 3-way classification task). In addition, an impact of high-level stress on in-set speaker recognition is evaluated and shown to reduce the accuracy from 91.7 % (low/mid stress) to 21.4 % (high level stress).

J.H.L. Hansen (✉) · E. Ruzanski · H. Bořil · J. Meyerhoff
Center for Robust Speech Systems (CRSS), University of Texas at Dallas, 800 West Campbell Rd, EC33, Richardson, TX 75080-3021, USA
e-mail: john.hansen@utdallas.edu

Keywords Stress assessment from speech · FLETC Corpus · TEO operator

1 Introduction

The negative impact of stress and emotions on automatic speech recognition is well documented in the past literature (Rajasekaran et al. 1986; Hansen 1988; Womack and Hansen 1999). Intuitively, the focus of speech recognition (as well as dialect and language recognition) is on the linguistic content captured in the speech signal, and other signal components such as environmental characteristics, channel transfer function, or speaker traits are considered redundant and potentially harmful. In contrast to this, the task of speaker recognition fully relies on speaker traits captured in the signal. However, also here, speech production variability on the individual level may harm both human-based and automatic speaker identification accuracy if the provided training samples do not capture similar speech modality (Ikeno et al. 2007). Several approaches focused on increasing robustness of speaker recognition to stress have been proposed (Hansen and Varadarajan 2009; Patil and Hansen 2010; Hansen et al. 2012). Most of them require information about the level of stress (its presence or absence) to match the processed speech sample with corresponding stressed or modal speech models.

The focus of this study is on the analysis of stressed speech acquired in a real-world scenario and a design of a finer-resolution stress assessment scheme that could be used in unsupervised generation/selection of stress-level dependent speaker models or in directing stress-dependent speech normalization strategies. Additional applications include automatic assessment of stress levels of personnel in critical

task settings, such as pilots, air traffic controllers, and military and security personnel, and allowing decisions to be made regarding the suitability of such persons to adequately perform, or continue to perform, their duties and maintain the safety of others.

Current automatic stress detection methods for speech have focused on binary neutral/stress detection with no efforts to account for graduated stress levels of the speaker conveyed through speech (Cairns and Hansen 1994; Rahurkar et al. 2002; Zhou et al. 2001; Bořil et al. 2010, 2012). After motivating the feasibility of automatic stress detection in speech, this study considers two approaches to reliably assess stress level conveyed through speech. The first method is a classification-based scheme, whose goal is to accurately classify test tokens into one of 3 stress levels: *low*, *mid*, or *high*. The second method is an assessment-based scheme, tracking the stress level conveyed by a speaker in their speech in real-time relative to a stress level *anchor point* established by the speaker prior to assessment conditions.

Both methods use the Euclidean distance metric (Theodoridis and Koutroumbas 2003) to measure distances between vectors in an orthogonal three-dimensional feature space. This hybrid space consists of temporal, frequency, and Teager Energy Operator (TEO)-based features. Previous research has shown that mean pitch and vowel duration are directly correlated to the presence of stress in speech (Hansen 1988; Bořil et al. 2010). Specifically, mean pitch and vowel duration have been shown to increase when the speaker is under stressful conditions. Mean pitch and duration features are used as 2 dimensions in this space. A new non-linear TEO-based feature, the $\overline{\Delta_{TEO}}$, comprises the third dimension of this feature space. As will be shown, the average value of bands 3 and 9 of the Teager Energy Operator, Critical Band, Autocorrelation Envelope (TEO-CB-AutoEnv) feature show a consistent decrease in autocorrelation area as the level of stress increases within speech.

The paper is organized as follows. The Federal Law Enforcement Training Center simulated hostage scenario corpus is described, to include speech data and biometric data analysis. The TEO and the TEO-CB-AutoEnv feature are explained in detail in Sect. 3. We then motivate the feasibility of automatic stress level assessment by showing if the TEO-CB-AutoEnv feature can be used to track stress levels in speech under a specific set of test conditions. The development of a 3-dimensional orthogonal hybrid feature space using frequency-based and TEO-CB-AutoEnv-based features is presented next. We then illustrate the effectiveness of classification-type and assessment-type schemes based on this 3-dimensional hybrid feature space to automatically assess the level of stress conveyed in a test subject's speech. Finally, we study the impact of stress on an in-set speaker recognition. The paper concludes with a discussion of the results presented and future directions relating to this work.

2 The Federal Law Enforcement Training Center simulated hostage scenario corpus

2.1 Overview

The speech data used in this paper was taken from male trainees completing a simulated hostage scenario held at the Federal Law Enforcement Training Center (FLETC) in Glynco, GA. The FLETC scenario was designed to test the capacity of students to draw on their training and personal resources to survive in a novel, rapidly evolving, highly stressful, multi-task paradigm that realistically models lethal force situations often encountered in the line of duty. Male police trainees who had completed all associated training coursework and were ready to graduate from FLETC were recruited and enrolled in the study after giving their written, informed consent. Each of 10 male speakers ranging in age from 23–34 years completed the scenario with each scenario lasting approximately 5 minutes. The scenario was scripted by FLETC to include comparable actions at similarly organized intervals unknown to each trainee. A microphone was placed in the room to record speech data. In addition, speaker biometric data including heart rate (HR), systolic blood pressure (sBP), diastolic blood pressure (dBP), and secretion salivary cortisol (SC) were collected from the trainees completing the FLETC hostage scenario.

The scenario involved an officer trainee (subject) and three role players: a "partner", a "complainant" and a "suspect". The "partner" was in fact a confederate of the experimenters, a fact unknown to the trainee, and was scripted to make a critical mistake, putting the trainee in jeopardy. Both trainee and partner were armed with 9 mm semi-automatic sidearms, modified to accept FX marking cartridges (similar to "paint balls"), and 3 magazines of marking cartridges (color-coded to distinguish impact of subject's rounds from confederate's weapon's rounds). Trainee and partner were told by a supervisor to investigate a complaint of theft and directed to enter a building and interview the complainant. The trainee and the partner were wearing face shields and protective vests. Two instructors, unarmed but also wearing protective gear were inside the house and served as role players, one as a co-operative *complainant*, and the other as a belligerent *suspect*.

When the student and partner entered the building, the complainant was in the front room and they began taking a report from him. The data collected from the trainee during the initial informational interview portion of the cooperative party by the officers was classified as *low* stress-level data. Approximately 1 minute later, the suspect emerged from a back room and began shouting at the complainant. Amidst the argument, background music and barking from a dog were introduced into the environment at this time. The data collected from the trainee during this period was

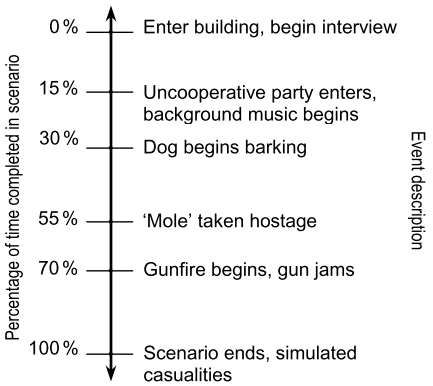


Fig. 1 Sequence of events in FLETC training scenario

classified as *mid* stress-level data. After this period of argument, the partner approached the suspect carelessly and had his holstered 9 mm weapon taken away by the suspect. The suspect shot the partner and took the complainant hostage, and used him as a shield. The suspect then shot the complainant with the partner's 9 mm weapon, ducked behind a wall, and re-emerged with a paint-ball shotgun with which he began firing at the student, who had minimal cover available. The suspect then resumed firing at the student with the 9 mm handgun. The student returned fire throughout this period with the third round in the student's weapon was *dud* (i.e., a round that failed to fire). This required the student to perform an action to clear the malfunction and further increased the stress level (Meyerhoff et al. 2004). The data collected during the hostage and simulated firefight portions of the scenario were classified as *high* stress data. The scenario concluded at a pre-arranged time when the suspect was scripted to fall to the floor, immobile and let the weapon fall from his hand. An instructor entered the room at this point to stop this segment of the scenario. The resulting corpus used for this experiment shows a definite build-up of speaker stress over time, which was confirmed by subject responses and biometric data analysis (Meyerhoff et al. 2004). The sequence of events in this training scenario is illustrated in Fig. 1.

One definition of stress is *the perception that situational demand exceeds resources* (Saunders et al. 1996). In the present study, the trainee relies on the partner as a resource, to perform as a partner, during the investigation of the simulated domestic dispute. The shortcomings of the partner constituted a significant deficit in resources and a major stressor for the trainee.

2.2 Speech data analysis

Recordings from ten male speakers completing the FLETC scenario as trainees are utilized in this study. Clearly, the number of subjects available in the FLETC corpus is relatively small, especially from the viewpoint of such tasks as speaker verification that often involve hundreds of subjects. However, the corpora utilized in these tasks capture routine conversations of naïve participants in modal conditions and are not suitable for the study of stress in speech. In this context, FLETC represents one of the first attempts to capture a broad range of stress levels in realistic conditions and provides a unique insight into how high-level stress (a hostage scenario) impacts speech production and in consequence, automated speech systems. In addition to the impact of stress itself, due to the adverse acoustic conditions in the FLETC scenarios (room reverberation, strong background music, cross-talks from partner, complainant, and suspect), the task of speaker recognition becomes significantly more complex compared to processing of office/telephone speech utilized in majority of current speaker recognition studies.

Table 1 Contents of FLETC speech corpus

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

It has been determined that vowels are an attractive class of phonemes to use as tokens in automatic speech recognition systems due to their definite quasi-periodic nature (Zhou et al. 2001). It has also been shown that there is little variance between vowel types used as tokens in an automatic TEO-based stress detection system (Ruzanski et al. 2005). A total of 756 vowel tokens of various types of suitable duration (Ruzanski et al. 2005) were extracted from the ten subject. The vowel set contains the phonemes /AA/, /AE/, /AO/, /AX/, /EH/, /EY/, /IH/, /IY/, /OW/, and /UW/. The vowels were manually extracted, as it was determined experimentally that automatic phoneme extraction schemes were adversely affected by the presence of stress in the speech.[1] Further details of the FLETC corpus are shown in Table 1.

Five tokens were extracted from each speaker during a *true neutral* condition, i.e., non-stressful conditions occurring approximately 30 minutes before and after the hostage scenario. While these tokens are not part of the FLETC hostage scenario, features extracted from these tokens are

[1] It has been shown that the performance of speech recognition systems trained with neutral speech data degrades significantly when testing with speech data collected while the speaker was under stress and that various feature-based spectral compensation schemes are needed to use automatic speech recognition for stressed speech (Hansen 1996; Bořil 2008; Bořil and Hansen 2010).

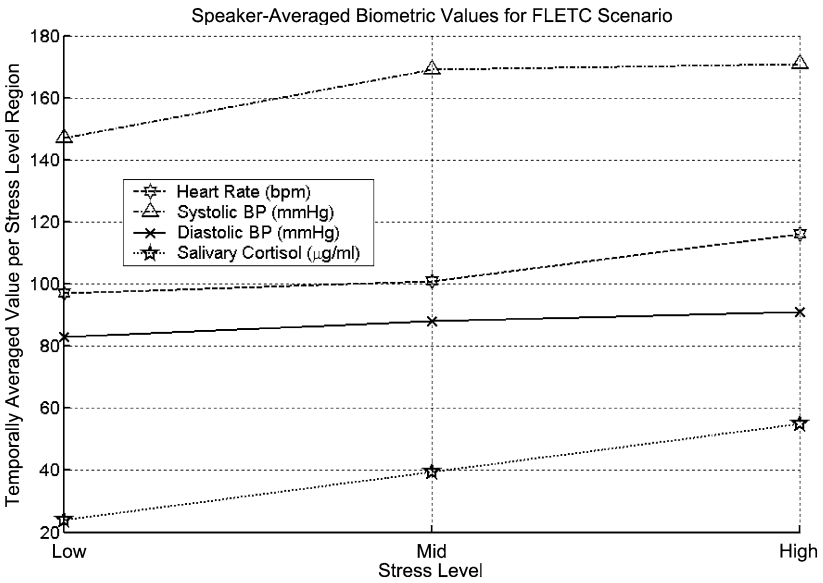


Fig. 2 Speaker biometric profile for FLETC hostage scenario

used for normalization of the feature dimensions described later in this study.

2.3 Biometric data analysis

Speaker biometric data, including HR, sBP, dBP and SC, is collected from each trainee completing the FLETC hostage scenario. This data is used in addition to the speech data collected from each speaker to independently assess the level of stress within a speaker during the scenario. It has been shown that biometric data is highly correlated with the level of stress within a speaker (Meyerhoff et al. 2004).

A SunTech Medical Corp. Accutracker II device was placed on the trainee to automatically record motion-tolerant systolic and diastolic BP readings approximately every minute throughout the scenario. A Polar Electro Oy Accurex Plus wireless HR monitor was used to record HR readings approximately every ten seconds throughout the scenario. Heart rate readings were transmitted wirelessly from a one-piece belt containing electrodes that was placed around the trainee's torso and recorded by an instrument placed on the trainee's wrist. Salivary cortisol was measured as the hormonal stress marker, since saliva is easily collected and changes in salivary levels reflect changes in plasma concentrations (Kirschbaum et al. 1996). Cortisol is released into the bloodstream from the cortex of the adrenal gland, and is one of several hormones that increase blood glucose as part of the normal response to exertion or psychological stress. Saliva samples were collected before and after the scenario after which cortisol analysis was accomplished. Saliva samples were collected using Salivette tubes. The Salivette cotton swab was chewed for 2 minutes to provide a sample at the start of baseline orientation, start of the scenario, immediately after the scenario, and 30 minutes after completion of the scenario. Saliva was frozen until assayed for cortisol by radioimmunoassay, as previously described in Meyerhoff et al. (1998). These biometric readings were used to assess the stress level in the speaker independently of speech data (Meyerhoff et al. 2004).

Biometric data for each of four biometric types, HR, sBP, dBP, and is taken from each trainee completing the FLETC hostage scenario. The results are temporally averaged within each of the three portions of the scenario, and these results are averaged across the 5 speakers. This averaging process reduces seasonality and illustrates the approximate linear trend in the biometric data (Brockwell and Davis 2002). These results are shown in Fig. 2. The results in Fig. 2 show a clear pseudo-linear increase in value for each type of biometric as stress level increases and verify the performance of the stress level classification scheme described in Sect. 6.1.

3 The Teager energy operator

Historically, most approaches to speech modeling have taken a linear plane wave point of view. While features derived from such analysis can be effective for speech coding, they are clearly removed from physical speech modeling. Teager (Teager 1980; Teager and Teager 1989) did extensive research on non-linear speech modeling and pioneered the importance of analyzing speech signals from an energy-based point-of-view. His studies showed that airflow is separated, with concomitant vortices distributed throughout the vocal tract. The differences in linear vs. non-linear vocal tract airflow modeling is illustrated in Fig. 3. It is believed that when a speaker is under stress, a change occurs in the vocal system physiology that further affects vortex-flow interaction patterns.

Teager devised a simple nonlinear energy-tracking operator that models the airflow through the vocal tract, shown mathematically for discrete-time signals as follows:

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

where $\Psi[\cdot]$ is the TEO. Kaiser first systematically introduced the TEO in Kaiser (1990a, 1990b).

The TEO-CB-AutoEnv feature employed here has previously been shown to reflect variations in excitation under stressful conditions (Zhou et al. 2001). A speech signal fundamental frequency (F_0) will change and hence the distribution pattern of pitch harmonics across critical bands will differ for speech under non-stressful conditions (Zhou et al. 2001). While it is possible to track the input F_0 over time and center a bandpass filter to extract a single TEO response, the accuracy of the response will be dependent on the F_0 tracking scheme, and could be influenced by noise (see Zhou et al. (2001) for such a TEO stress scheme). In this study, a fine frequency resolution is achieved using a critical band partitioning of entire audible frequency range into critical bands (Scharf 1970; Yost 1994), thus removing the need for F_0 tracking. The actual F_0 value from the speech signal will influence the number of harmonics in each frequency bin, which in turn will impact the cross-terms due to more than one harmonic being analyzed by the TEO. However, the process of finding the autocorrelation response followed by the envelope calculation significantly reduces the dependency on F_0 characteristics in the resulting TEO feature. We therefore extract the TEO and F_0 terms separately to allow for a data fusion strategy employing specific weighting of each domain for stress/neutral classification.

The TEO-CB-AutoEnv is extracted through a process shown in the flow diagram of Fig. 4 and illustrated mathematically using critical bandpass filters as,

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

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

The TEO-CB-AutoEnv feature is thus a 17-dimensional vector of normalized area coefficients, with each element representing the area under the TEO autocorrelation envelope in each of the 17 filter bands normalized by the energy across all bands collectively. While periodic structure related to the harmonics in the speech signal is exhibited in the TEO feature and the associated autocorrelation function, such structure is removed by taking the area under the envelope of the autocorrelation function thus creating a feature that is pitch resistant. It has also been observed that some area terms are smaller in value when speech under stress is evaluated (Zhou et al. 2001), suggesting a reduced correlation (i.e., regularity) of the signal versus neutral-condition speech within across these frequencies. For a speech signal with stationary excitation (constant pitch, regular glottal pulses), the envelope of the autocorrelation function could be interpolated by a line with a constant negative slope. Increased irregularity in the speech signal, that can be attributed to the elevated level of stress (Zhou et al. 2001), will result in a faster decaying autocorrelation envelope and in consequence, reduced area under the envelope.

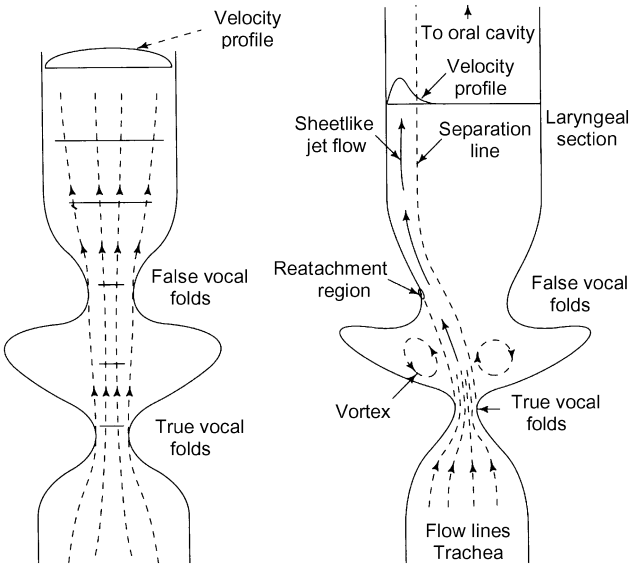


Fig. 3 Linear (*left*) vs. non-linear (*right*) vocal tract airflow model; after Zhou et al. (2001)

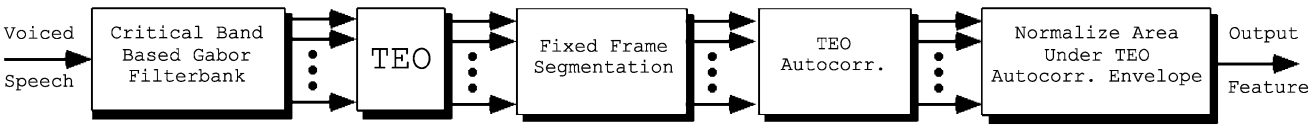


Fig. 4 TEO feature extraction flow diagram; after Zhou et al. (2001)

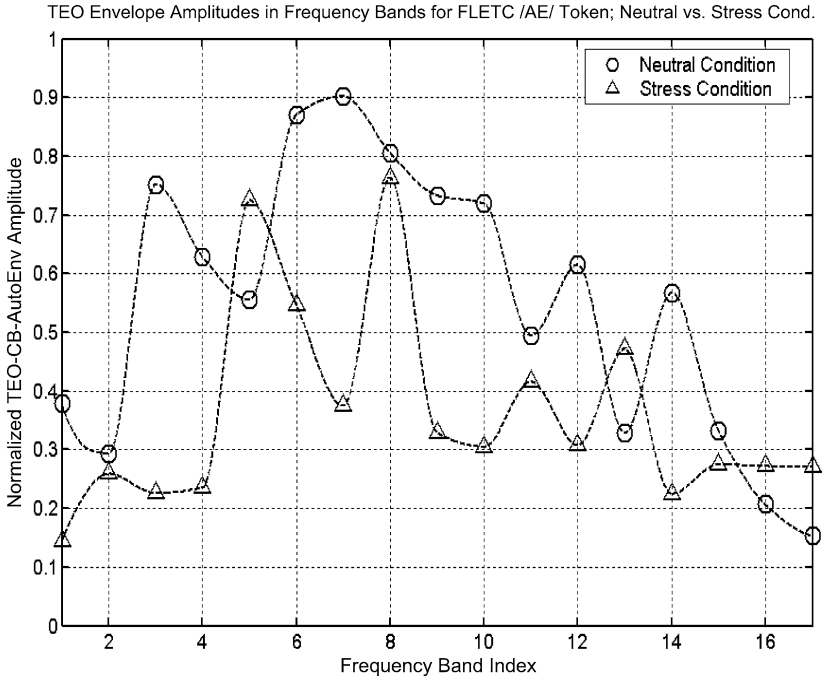


Fig. 5 TEO-CB-AutoEnv profile for /AE/ token, neutral vs. stress condition; single speaker experiment

4 Motivation for automatic stress level assessment using the TEO-CB-AutoEnv feature

The TEO-CB-AutoEnv was analyzed for an /AE/ vowel token under a clear neutral and stress condition taken from a single speaker. This comparison across frequency bands from 0–4 kHz is shown in Fig. 5. It can be seen that the normalized TEO-CB-AutoEnv values vary for each of the 17 frequencies bands (ranging from 0.035 to 0.527) and are generally higher for the token under the neutral condition. The frequency band information for the critical bands is shown in Table 2.

To investigate the behavior of the TEO-CB-AutoEnv feature over time in the FLETC variable-stress level scenario, we choose the four frequency bands that show the greatest difference between neutral and stress conditions from the analysis of Fig. 5. Here, these are bands 3, 7, 9, and 10, which correspond to frequencies centered at 250, 700, 1000, and 1170 Hz, respectively.

We extracted nine sentences at times distributed throughout the task in the FLETC scenario, from the same speaker, and extracted three vowel samples from each of these sentences. Our focus is on the development of a vowel-independent assessment scheme and hence, the samples were randomly selected from the pool of vowels listed in Sect. 2.2, whichever occurred in the sentences.

We found average values for the TEO-CB-AutoEnv for each of the four bands identified above across each of the three vowels in each sentence. Finally, we averaged the mean values for each of the three vowels to yield one value representative of the behavior of the TEO-CB-AutoEnv at each time across the duration of the scenario. The results are shown in Fig. 6. Figure 6 shows that the band-averaged TEO-CB-AutoEnv values across phonemes in sentences spoken at different times in a stressful scenario are inversely proportional to the level of stress the speaker is under at that particular time.

While these results show promise for automatically assessing and tracking stress levels in speech, it may not be feasible to acquire the four best bands from the comparison between neutral and stressful speech tokens. We employ additional speech feature parameters to assess stress level in speech and create a new hybrid multi-dimensional feature space in the following section.

5 Hybrid multi-dimensional feature selection and analysis

The features chosen for this experiment are both hybrid (i.e., temporal-based, frequency-domain-based, non-linear) and multi-dimensional. The dimensions were normalized using respective values for *true neutral* speech pooled for all ten speakers. This *true neutral* speech was collected from the same speakers who completed the training scenario but taken under controlled laboratory conditions 30 minutes before and after the scenario. The three dimensions chosen for stress level assessment in this experiment that include mean pitch, token duration, and $\overline{\Delta_{TEO}}$ are described below. The following analyses consider a development set of the 10 speakers in the FLETC corpus.

Table 2 Critical band frequency information; after Zhou et al. (2001)

	Critical Band Frequency Information (Hz)			
Band Number	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

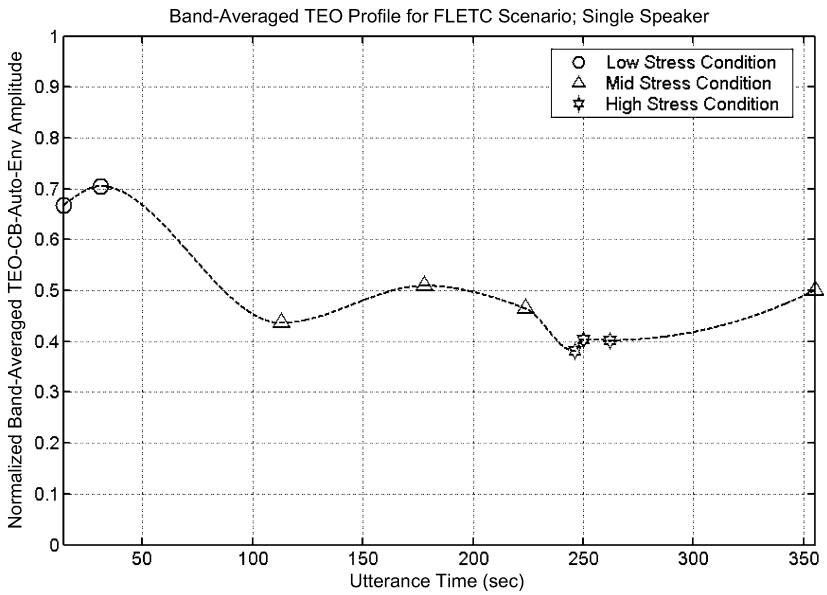


Fig. 6 Band-averaged TEO-CB-AutoEnv profile for FLETC scenario vs. time

5.1 Mean pitch

It has been shown that mean pitch increases significantly for a speaker under stress relative to a non-stress condition (Hansen 1988). In certain cases where the speaker is trying to conceal a stressful condition (e.g., polygraph examination), this may not be the case (Hansen 1988). Since speakers in the FLETC scenario are assumed to not attempt to conceal stress conveyed in their speech, we chose to explore the validity of using mean pitch as a feature for assessment of stress level in the FLETC scenario.

Frame-by-frame pitch values are extracted from each of the tokens using a sub-harmonic-to-harmonic ratio (SHR) algorithm. The SHR algorithm was previously demonstrated (Sun 2002) to provide a superior performance to a popular autocorrelation-based pitch tracker available in PRAAT (Boersma 1993) as well as to the cross-correlation based RAPT pitch tracker (Talkin 1995). One of the sources of

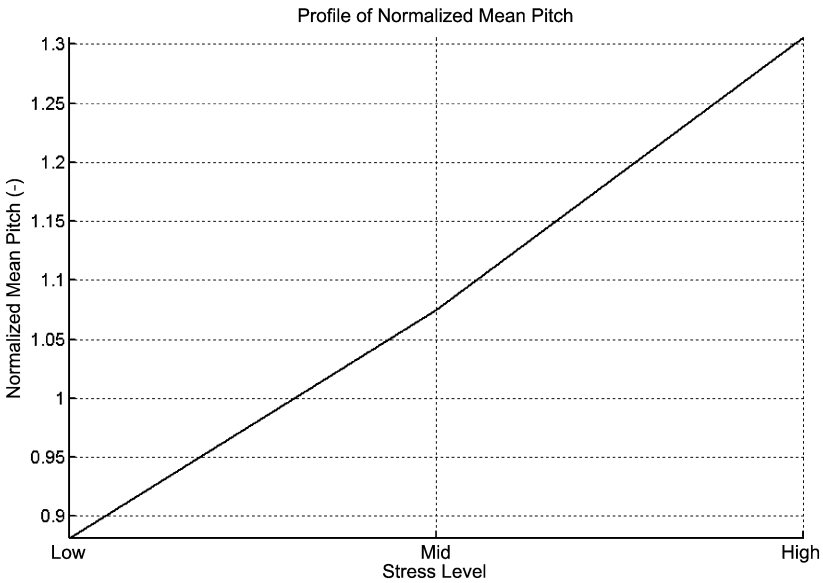


Fig. 7 Centroid plot for mean pitch dimension across stress levels

Table 3 Statistical analysis for mean pitch feature (normalized)

Stress Level	μ	σ^2	Min	Max
Low	0.8818	0.0437	0.5833	1.4896
Mid	1.0751	0.0230	0.6615	1.4948
High	1.3063	0.0838	0.6510	1.8594

Table 4 Statistical analysis for duration feature (normalized)

Stress Level	μ	σ^2	Min	Max
Low	0.9393	0.0667	0.5926	1.5635
Mid	1.0187	0.1264	0.4953	2.7958
High	1.1647	0.2025	0.6770	2.2976

pitch estimation errors in traditional pitch trackers is the presence of alternate cycles in speech signal. These cycles relate to voice quality and can be observed for example in creaky voices, voices with laryngealization, and pathological voices (Sun 2002). It can be assumed that the presence of stress will further impact the presence of the alternate cycles. The SHR algorithm accounts for the presence of these cycles and as such is a good candidate for pitch tracking in non-modal speech scenarios such as the one considered in this study. Normalized pitch values were determined by dividing frame pitch values by the overall mean pitch calculated over the whole speaker session. The resulting normalized pitch statistics for each of the three stress levels is shown in Table 3.

In Table 3, the μ column represents the average value taken across the collection of tokens of the normalized mean pitch values from each of the respective stress level regions (i.e., *low*, *mid*, and *high*). Similarly, σ^2, min, and max represent the variance, minimum value, and maximum value, respectively, of the normalized mean pitch values for tokens in each of the respective stress level regions.

The results from Table 2 indicate an increase in mean pitch with increasing stress level (e.g., a 30.6 % increase in mean for high stress). The small variance values confirm a relatively tight distribution about the mean of each region, which is desirable for a clustering-type classification scheme. The approximate linear relationship of the mean pitch behavior is shown in Fig. 7.

5.2 Vowel duration

It has been shown that vowel duration and the variability of vowel duration increases significantly for a speaker under stress (Hansen 1988). We chose vowel duration as a feature in the assessment of stress level in speech in the FLETC scenario. For the purpose of the speech production assessment presented here, boundaries of vowel segments were determined through manual labeling. For a fully automated processing, vowel segments can be tracked using automatic phone recognition (Schwarz 2009). Alternatively, the group of vowel segments can be extended to all voiced segments as their duration has been found to strongly correlate with the presence of stress (Bořil et al. 2010).

Duration is independent of mean pitch, assuming quasi-periodicity of the vowel tokens and the removal of the temporal component during the averaging operation. Pitch and duration dimensions are thus orthogonal. The normalized durations were obtained by dividing duration of each token by the mean duration extracted over the whole speaker session. The normalized duration statistics are shown in Table 4.

Again, under high stress conditions, there is a 16 % increase in mean vowel duration. Also note that the vowel

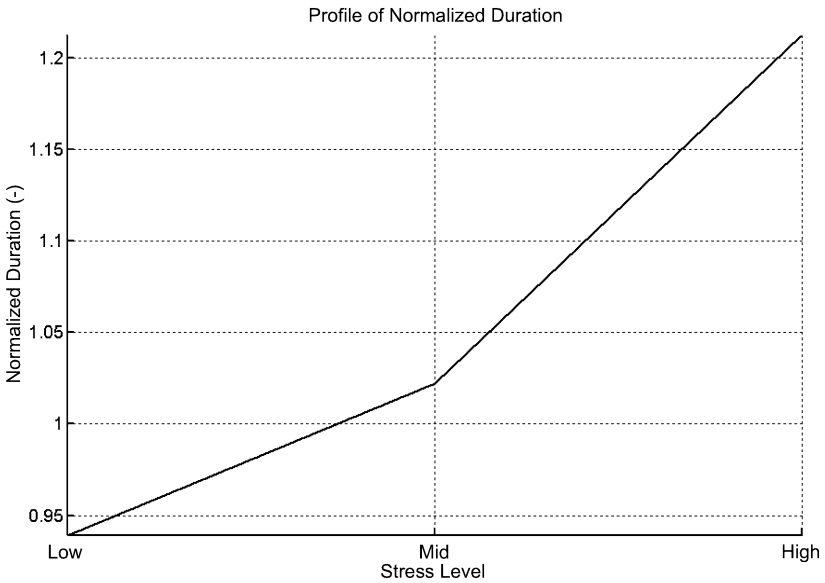


Fig. 8 Centroid plot for vowel duration dimension across stress levels

duration variability increases for increasing stress level as shown by the increasing variances in Table 4. The quasi-linear behavior related to stress level of the vowel duration dimension is illustrated in Fig. 8.

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

The results in Sect. 4 suggest the TEO-CB-AutoEnv feature value per frequency band is inversely related to the level of stress conveyed by a speaker in a speech sample. The analysis shown in Sect. 4 requires the presence of both low and high stress tokens from the speaker. While low stress tokens can be collected easily from the speaker, as it is assumed a speaker will always be under some level of stress in a practical law enforcement situation, high stress tokens are more difficult to collect. For this reason, it is not feasible to find the 4 best bands for each speaker completing such a scenario to track stress levels for that speaker within the respective scenario. It has been shown in Bou-Ghazale and Hansen (1998) that it is possible to develop stress perturbation models from neutral-to-stress and stress-to-neutral from a core speaker set and apply these perturbation models to a new speaker set for which only neutral data exists for successful automatic speech recognition. We therefore took a similar approach by training models from open speakers for stress classification/assessment.

It has been shown in Zhou et al. (2001) that TEO-CB-AutoEnv feature values are independent of pitch. Therefore, any linear combination of such features will also yield an independent, thus orthogonal, dimension to pitch and duration values, as the TEO-CB-AutoEnv values used are temporally averaged across frames of the vowel tokens. This completes the creation of a Euclidean space, facilitating the use of the Euclidean distance metric used in the stress level assessment algorithm outlined later in this study.

The following experimental procedure is performed for stress level assessment in speech. Two different vowel tokens are selected for each of two speakers from the corpus. For the first speaker, vowels /OW/ and /AY/ were selected. For the second speaker, vowels /IH/ and /AO/ were selected (i.e., we wanted to explore cases where vowel mismatch exists). The vowels were selected such that the same vowel was spoken during low, mid, and high stress conditions. The normalized TEO-CB-AutoEnv values were extracted for each band from each of the tokens. The difference was taken from the low-to-high stress conditions. The results are shown in Table 5. From the analysis of the results in Table 5, the bands that exhibit ubiquitous decreasing behavior across the 4 tokens across the speakers from the low to high stress conditions are bands 1, 3, 4, 9, 11, and 14. We then examine the behavior of these bands for the low-to-mid stress condition. The results are shown in Table 6.

Table 6 shows that frequency bands 3 and 9 (200–300 Hz and 920–1080 Hz, respectively) exhibit a consistent decrease of the normalized TEO-CB-AutoEnv value. Further analysis of the TEO-CB-AutoEnv values across bands 3 and 9 is shown in Table 7. It can be seen that the majority of tokens display a consistent decrease in TEO-CB-AutoEnv values across all speakers.

By averaging the TEO-CB-AutoEnv values across bands 3 and 9, we create a new $\overline{\Delta_{TEO}}$ feature, shown mathematically as:

$$\overline{\Delta_{TEO}} \triangleq \frac{TCA_3 + TCA_9}{2}, \tag{3}$$

where TCA, $i = \{3, 9\}$ is the TEO-CB-AutoEnv value for the ith frequency band. The behavior of this feature is shown in Table 8. The results shown in Table 8 show a consistent decline in $\overline{\Delta_{TEO}}$ values with the level of stress across speakers and across tokens.

Table 5 TEO-CB-AutoEnv values for frequency bands across tokens; difference from low-to-high stress level

Frequency Band	TEO-CB-AutoEnv Difference (Low-to-High Stress)			
	Vowel Token			
	/OW/	/IH/	/AY/	/AO/
1	**-0.083**	**-0.642**	**-0.443**	**-0.190**
2	+0.147	-0.161	-0.429	-0.093
3	**-0.032**	**-0.567**	**-0.126**	**-0.370**
4	**-0.142**	**-0.281**	**-0.275**	**-0.496**
5	+0.164	+0.063	-0.167	+0.015
6	-0.068	-0.075	-0.106	+0.034
7	-0.426	-0.437	+0.306	-0.233
8	-0.099	+0.083	-0.161	-0.461
9	**-0.343**	**-0.295**	**-0.096**	**-0.254**
10	-0.085	-0.115	+0.083	-0.404
11	**-0.299**	**-0.165**	**-0.069**	**-0.078**
12	-0.022	-0.269	-0.122	+0.023
13	-0.037	-0.006	+0.180	-0.161
14	**-0.136**	**-0.021**	**-0.168**	**-0.012**
15	-0.210	+0.142	-0.095	+0.137
16	-0.633	-0.048	-0.191	+0.127
17	-0.581	+0.185	-0.023	+0.043

Table 6 TEO-CB-AutoEnv values for frequency bands across tokens; difference from low-to-high stress level

Frequency Band	TEO-CB-AutoEnv Difference (Low-to-Mid Stress)			
	Vowel Tokens			
	/OW/	/IH/	/AY/	/AO/
1	+0.114	-0.637	-0.026	-0.218
3	**-0.011**	**-0.207**	**-0.094**	**-0.186**
4	-0.229	+0.148	-0.140	-0.228
9	**-0.230**	**-0.155**	**-0.105**	**-0.217**
11	-0.262	-0.457	+0.002	-0.145
14	-0.076	-0.126	-0.123	+0.057

Table 7 TEO-CB-AutoEnv values for frequency bands across tokens

Vowel Token	Normalized TEO-CB-AutoEnv Value					
	Band 3			Band 9		
	Low	*Mid*	*High*	*Low*	*Mid*	*High*
/OW/	**0.529**	**0.591**	**0.497**	0.767	0.606	0.424
/IH/	0.875	0.668	0.308	0.603	0.448	0.308
/AY/	0.612	0.518	0.486	**0.574**	**0.469**	**0.478**
/AO/	0.665	0.479	0.295	0.847	0.630	0.593

Table 8 $\overline{\Delta_{TEO}}$ values across tokens

Vowel Token	$\overline{\Delta_{TEO}}$ Value		
	Low	*Mid*	*High*
/OW/	0.648	0.599	0.461
/IH/	0.739	0.558	0.308
/AY/	0.593	0.494	0.482
/AO/	0.756	0.555	0.444

Table 9 Statistical analysis for the '$1 - \overline{\Delta_{TEO}}$' feature (normalized)

Stress Level	μ	σ^2	Min	Max
Low	0.8283	0.0293	0.4770	1.1611
Mid	1.0187	0.0362	0.4834	1.4374
High	1.1647	0.0285	0.6852	1.5146

Finally, the $\overline{\Delta_{TEO}}$ samples are pooled together from all the tokens. To introduce similar behavior as seen in the mean pitch and duration dimensions (increasing trend with the level of stress), the resulting statistics of $(1 - \overline{\Delta_{TEO}})$ (normalized by dividing each sample by the overall mean) are presented in Table 9.

The approximate linear relationship of this increasing behavior is shown in Fig. 9. According to these results, frequency bands 3 and 9 exhibit similar sensitivity to stress. The lower TEO-CB-AutoEnv values suggest greater irregularity in speech within these 2 frequency bands. This irregularity implies a sharper decline in the TEO autocorrelation function over increasing lag and thereby a lower autocorrelation envelope area value (see also the discussion at the end of Sect. 3). The frequency band 3 (200–300 Hz) corresponds to the area of occurrence of the first two harmonics of the glottal spectrum (their spacing being equal to the fundamental frequency). The band 9 (920–1080 Hz) represents the upper area of the first formant (F_1) and bottom area of the second formant (F_2) occurrence. Past literature (Hansen 1996; Bořil et al. 2010) reports strong sensitivity of fundamental frequency and spectral slope of the glottal waveform spectrum, as well as of the $F_{1,2}$, to the presences of stress. The prominence of bands 3 and 9 seen in this section confirms this phenomenon also for the FLETC stress scenarios.

6 Automatic speech stress-level assessment schemes

In this section, we utilize the feature space described in the previous section in two automatic stress level assessment paradigms. In the first archetype, we consider the problem

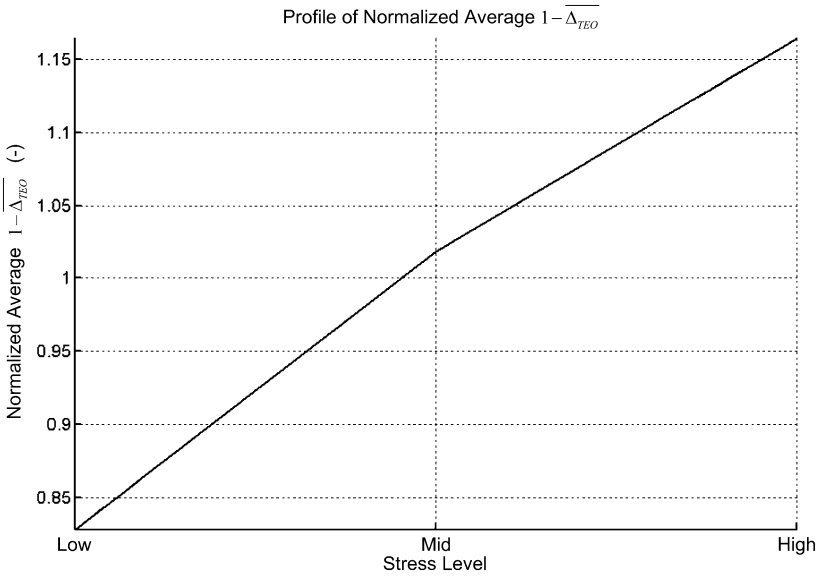


Fig. 9 Centroid plot for '$1-\overline{\Delta_{TEO}}$' dimension across stress levels

of automatic stress assessment in speech in the *classification* sense. We present error-rate classification results pertaining to test token classification into the appropriate stress level category, i.e., *low*, *mid*, *high*, explained previously. In the second archetype, we consider the problem of automatic stress assessment in speech in a relative measurement to a readily available low-stress level centroid. Since successive speech token distances are referenced to this anchor point, we can assess the stress level conveyed in speech by the creation of a real-time stress-level profile. We will then illustrate the effectiveness of median smoothing to enhance the visibility of this profile at the expense of the real-time property.

6.1 Stress level classification using nearest neighbor classification

This scheme begins with the establishment of centroids encompassing the 3 dimensions explained in Sect. 5. It was determined experimentally that use of the 3-dimension feature space yielded better results than any of the other 6 combinations of lower-dimensional class combinations. The available speech data is collected and the norms of the 3 dimensions are computed for each token. A subset of the lowest $1/3$ of this data is created and the resulting centroid becomes the low-stress level centroid. Likewise, the high-stress level centroid is taken from the subset comprised of the highest $1/3$ of the data. Based on the quasi-linear behavior of each of the 3 dimensions, the mid-stress level centroid is taken to be the mean of the low- and high-stress level centroids. The choice to select $1/3$ of the data was determined to give the best experimental results.

6.1.1 Token-level classification

The establishment of stress level centroids and test set is performed in a round robin fashion establishing a speaker-independent test scenario. A test token's class membership is determined by the minimum Euclidean distance (Theodoridis and Koutroumbas 2003) to the centroid of that class. The decision scheme for the token-level classification scheme is shown in Eq. (4):

$$M_T = \min\bigl\{\|\mathbf{x}-\mathbf{C}_{LOW}\|, \|\mathbf{x}-\mathbf{C}_{MID}\|, \|\mathbf{x}-\mathbf{C}_{HIGH}\|\bigr\},$$
$$\hat{D}_T = \begin{cases} \text{Low stress}, & \text{if } \|\mathbf{x}-\mathbf{C}_{LOW}\| = M_T, \\ \text{Mid stress}, & \text{if } \|\mathbf{x}-\mathbf{C}_{MID}\| = M_T, \\ \text{High stress}, & \text{if } \|\mathbf{x}-\mathbf{C}_{HIGH}\| = M_T, \end{cases} \tag{4}$$

where $\hat{D}_T$ is the token-level stress level classification decision, $\mathbf{x}$ is the 3-dimensional feature vector for the test token, $\mathbf{C}_{\{LOW,MID,HIGH\}}$ is the 3-dimensional (*low stress*, *mid stress*, *high Stress*) centroid, and $\|\cdot\|$ is the norm operator.

6.1.2 Sentence-level classification

It has been shown that utilizing a weighted majority rule decision algorithm yields considerable improvement for automatic stress detection over token-level algorithms (Saunders et al. 1996). We next employ a sentence-level classification scheme using accumulated Euclidean distances for each vowel token within a given sentence.

The stress level centroids are established and speaker-independent testing is conducted in a round robin fashion as was used with the token-level scheme. The stress classifi-

Table 10 Comparison of automatic stress-level detection schemes

Comparison of Automatic Stress-Level Detection Schemes (Percent-Error)				
Classification Scheme	Stress Level			
	Low	Mid	High	Overall
Token-Level	17.2	62.7	50.8	50.5
Sentence-Level	14.6	53.6	52.8	44.4

cation decision is made according to Eq. (5):

$$M_S = \min\left\{\sum_{i=1}^{N} W_{LOW,i}, \sum_{i=1}^{N} W_{MID,i}, \sum_{i=1}^{N} W_{HIGH,i}\right\},$$
$$\hat{D}_S = \begin{cases} \text{Low stress}, & \text{if } \sum_{i=1}^{N} W_{LOW,i} = M_S, \\ \text{Mid stress}, & \text{if } \sum_{i=1}^{N} W_{MID,i} = M_S, \\ \text{High stress}, & \text{if } \sum_{i=1}^{N} W_{HIGH,i} = M_S, \end{cases} \tag{5}$$

where $\hat{D}_S$ is the sentence-level stress level classification decision, $W_{\{LOW,MID,HIGH\},i} = \|\mathbf{x} - \mathbf{C}_{\{LOW,MID,HIGH\}}\|$ for the ith vowel token in the test sentence, $\mathbf{x}$ is the 3-dimensional feature vector for the test token, N is the number of vowel token per test sentence, $\mathbf{C}_{\{LOW,MID,HIGH\}}$ is the 3-dimensional *low stress*, *mid stress*, *high stress* centroid, and $\|\cdot\|$ is the norm operator.

6.1.3 Stress-level classification performance analysis

The performance of the token- and sentence-level stress level classification schemes is shown in Table 10. The results presented in Table 10 show effectiveness of both schemes in classifying a vowel token at each stress level, as a chance decision yields an asymptotic error rate of 2/3. Similar to the results presented in Saunders et al. (1996), a weighted majority decision scheme yields improved performance, here a relative improvement of 12.1 %.

Analysis of the results presented in Table 10 also raises other issues concerning the choice of paradigm for stress level assessment in speech. From the analysis of the results presented in Table 10, it is seen that the mid-stress level condition consistently yields worse performance than that for the low- and high-stress level conditions. It is believed that more stress level graduations within the mid-stress level region may improve performance. To investigate this stress-level *quantization error* effect, a speaker is chosen and an analysis of the errors in the mid-stress level region is performed. The results (see Table 11) show a migration of errors committed from low-stress level misclassification to high-stress level misclassification as the scenario, and subsequently stress stimulus, increases. The results presented in Table 11 show that as time progresses throughout the scenario, more errors are made through high-stress level misclassification versus low-stress level misclassification. This would suggest further segmentation of the mid-stress level for this speaker to enhance stress level detection performance. A response to the issue of optimal stress level segmentation is to change the experimental paradigm from a stress level *classification* method to a time-based stress level *assessment* method. We describe such an assessment scheme in the next subsection.

Table 11 Single-speaker mid-stress level region error analysis

Single-Speaker Mid-Stress Level Region Error Analysis			
Sentence #	Sentence Time (sec.)	"Low-Stress Level" Misclassifications	"High-Stress Level" Misclassifications
1	112–115	3/8	2/8
2	120–130	1/5	1/5
3	144–150	0/5	1/5
4	153–160	3/6	0/6
5	162–167	0/9	2/9
6	170–175	0/6	5/6
7	179–181	2/6	0/6
8	216–220	0/8	6/8

6.2 Stress level assessment using the Euclidean distance metric

In the following subsection, we outline two methods for stress level assessment based on considering the low-stress level centroid explained earlier as a stress level *anchor point* and the Euclidean distance metric from this centroid as a *ruler* to measure speaker stress level conveyed in speech.

6.2.1 Real-time stress level tracking algorithm

This algorithm requires the establishment of a stress level anchor point by extracting the three features explained earlier from speech data collected before a scenario over which stress levels desire to be tracked. As a scenario progresses, three features are extracted from each vowel token. We establish the stress level conveyed by the speaker in speech at that time by measuring the Euclidean distance from the token's feature set to the stress level anchor point previously established. This is shown in Eq. (6):

$$SL_t = \|\mathbf{x}_t - \mathbf{C}_{ANCHOR}\|, \tag{6}$$

where SL_t is the relative stress level assessed at time t, $\mathbf{x}_t$ is the 3-dimensional feature vector for the test token at time t, $\mathbf{C}_{ANCHOR}$ is the 3-dimensional centroid representing vowel tokens establishing a stress level reference point for stress level analysis over time, and $\|\cdot\|$ is the norm operator.

A stress level profile can be established for the speaker as time progresses. Profiles for the 10 speakers completing the

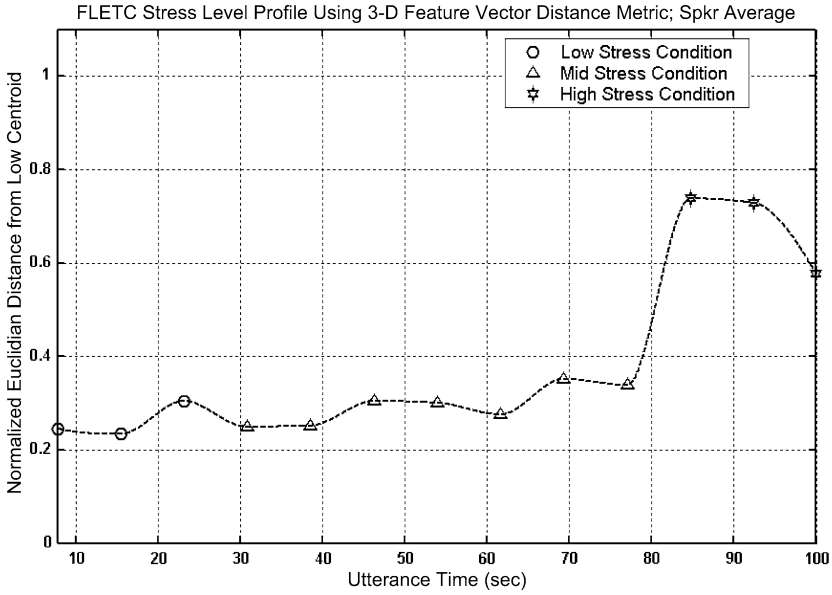


Fig. 10 FLETC stress level profile vs. time using 3-D vector distance metric, speaker average

FLETC simulated hostage scenario are shown in Fig. 11. Since the FLETC simulated hostage scenario is presented in a uniform structure for all speakers, the average profile of the 10 speakers is taken over a normalized time axis, representing increasing percentage of progression through the scenario. The results are shown in Fig. 10. Figure 10 shows a pseudo-linear increasing stress level trend within each speaker's speech as the scenario progresses. This result is supported by the speaker biometric analysis presented in Fig. 2.

6.2.2 *Stress level tracking algorithm with median smoothing*

Analysis of the real-time stress level profile for Speaker 10 in Fig. 11 shows considerable variation, especially in the low and high stress regions. This might be undesirable for certain applications. At the expense of the real-time aspect of the algorithm by the introduction of a one-sample delay, we illustrate the effectiveness of 3-point median filtering on the stress profile of speaker 10.

The output $y[n]$ of the 3-point median filter for an input sequence $x[n]$ is shown in Eq. (7) (Mitra 2001):

$$y[n] = med\big\{x[n-1], x[n], x[n+1]\big\}. \tag{7}$$

Figure 12 shows the effectiveness of 3-point median smoothing to clean up the stress level profile for speaker 10.

7 Impact of stress on automatic in-set speaker recognition

Previous sections have demonstrated the impact of stress on speech signal and exploited the stress-induced speech production variability in the design of automatic stress assessment schemes. In this section, the impact of stress on an automatic speaker identification is studied.

A Gaussian Mixture Model-Universal Background Model (GMM-UBM) (Reynolds et al. 2000) speaker recognition system is evaluated in the task of in-set speaker recognition (pick one out of ten speakers). The system utilizes a Mel frequency cepstral coefficients (MFCC) (Davis and Mermelstein 1980) feature extraction front-end. Twenty static cepstral coefficients (c_0–c_{19}) and their first and second order time derivatives are extracted from a 25 ms window shifted with a 10 ms step. During the system training stage, a 32-mixture UBM is first trained on pooled training samples from all target speakers. Subsequently, speaker-dependent models are derived by maximum a posteriori (MAP) adaptation of the UBM towards speaker-dependent training samples. Speech samples from approximately first half or two thirds of each speaker session, representing *low* and *mid* stress conditions, are used for the UBM training and speaker-dependent model adaptation. A subset of low/mid stress level samples is excluded from the training set to form a *low*/*mid* stress open test set. The speech data acquired starting from the instance when the suspect took away the gun from the trainee's partner is used to form the *high* stress open test set. While the speaker recognition task may seem simple here (ten target classes), it is noted that the processed speech samples are corrupted by the room reverberation, background music, and cross-talks from the other subjects in the room (partner, complainant, suspect). Most of those represent non-stationary sources of distortion and are quite adverse to the classification task.

The evaluation results together with the average test sample durations are shown in Table 12. The first row of the results represents a closed test set evaluation where the

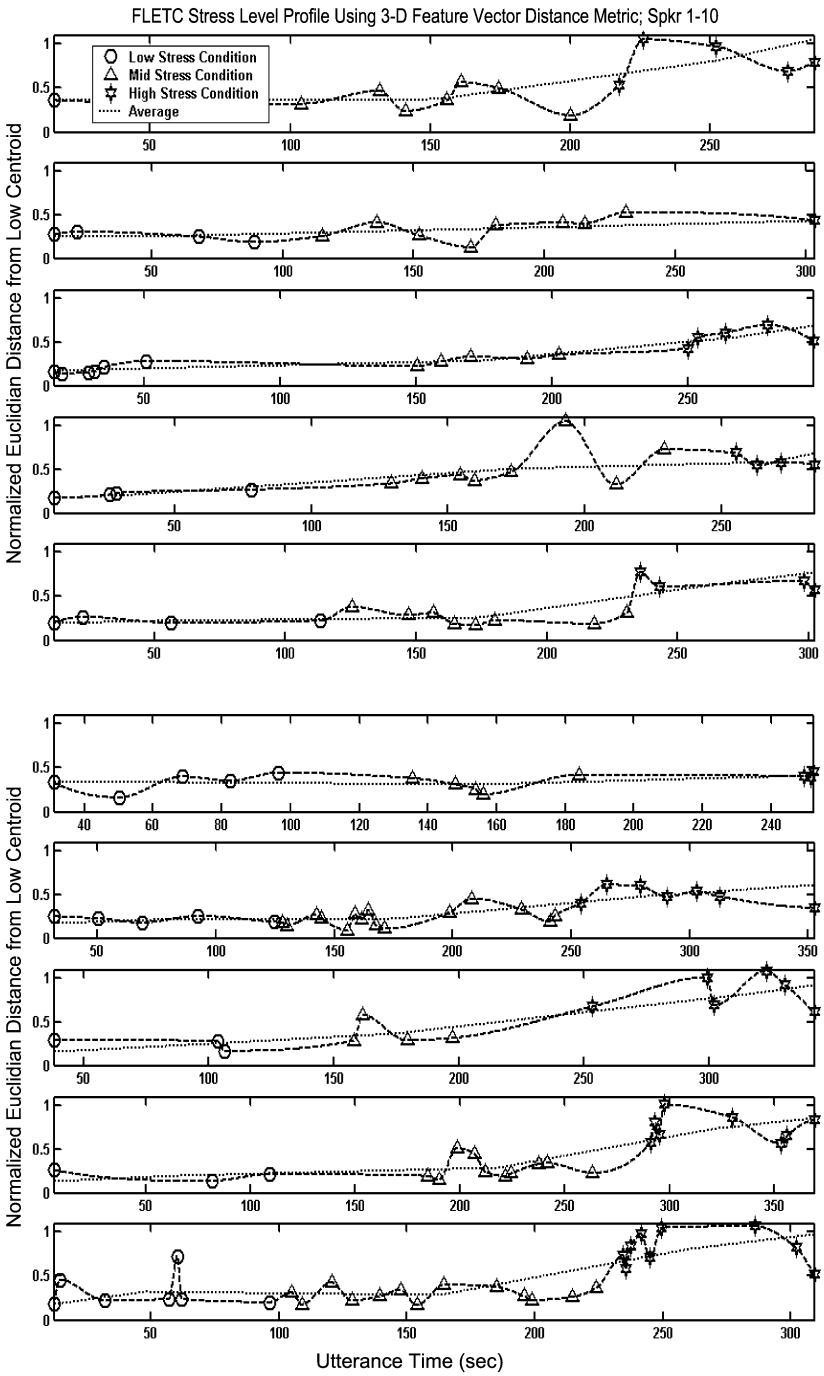


Fig. 11 FLETC stress level profile vs. time using 3-D vector distance metric

Table 12 In-set speaker recognition task

Evaluation Task	Stress Level	Avg. Sample Duration (sec)	In-Set Speaker ID Accuracy (%)
Closed	Low/Mid	1.9	100.0
Open	Low/Mid	1.7	91.7
	High	1.7	21.4

low/mid speech samples drawn from the training set are processed by the classifier. The following two rows represent open test set evaluations where speech samples unseen during the classifier training are processed. It can be seen that while the open *low/mid* test set yields a speaker recognition accuracy of 91.7 %, the performance drops down by 70 % when exposed to the *high* level stress samples. This result only confirms observations on the impact of stress on speech systems presented in past literature. It is expected that availability of automatic stress assessment schemes such as those proposed in Sect. 6 can be exploited in the design of stress-robust speaker recognition systems; e.g., through stress-dependent model selection or stress-adaptive feature normalizations.

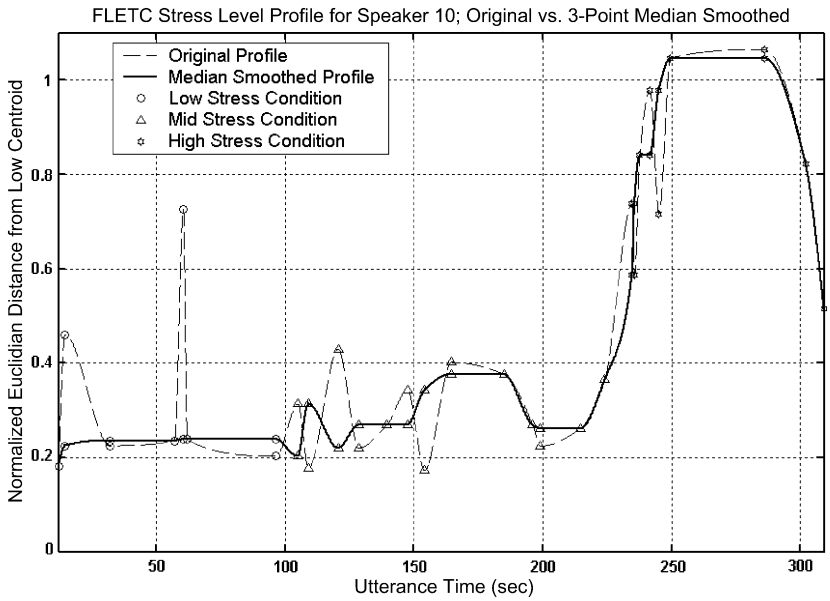


Fig. 12 FLETC stress level profile for speaker 10 (3-point median smoothed)

8 Discussion and conclusions

The ability to assess the stress level in speech plays an important role in the design of speech-enabled systems with increased robustness to speech variability. In particular, the detected level of stress in individuals can be used in unsupervised generation of stress level-specific acoustic models or for directing adequate speech normalization strategies. Equally important, automatic stress classification can be utilized in automated call centers for redirecting calls to human operators, or in security and law enforcement applications.

This study has illustrated the effectiveness of two stress level assessment schemes using the FLETC simulated hostage scenario corpus, a real-world development constructed to present increasing stress stimulus to the participant over time as verified by biometric data. Two algorithms were developed which utilize vectors located in an orthogonal three-dimensional space extracted from vowel tokens from the FLETC simulated hostage scenario. The dimensions comprising this space include mean pitch, vowel token duration, and a new TEO-based non-linear feature, the $\overline{\Delta_{TEO}}$ feature. It was shown that these features exhibit ubiquitous and monotonic quasi-linear behavior over increasing stress levels, and independently confirmed using biometric (i.e., heart rate, systolic and diastolic blood pressure) data.

The first classification paradigm is based on the ability to categorize test tokens into one of three stress levels, *low*, *mid*, or *high*. This paper showed that this method was effective relative to chance decisions, using both token- and sentence-level classification schemes based on minimum Euclidean distance metrics and accumulated Euclidean distance-weighted metrics, respectively. It was also shown that the issue of *quantization error* would be present when deciding between stress levels. The ability to accurately determine the locations and sizes for stress levels motivated the second archetype, an *assessment* paradigm directly relating the level of stress to the Euclidean distance between a test token's 3-dimensional feature vector a vector extracted from a token set collected under a condition of a known stress level. The results shown are independently confirmed using speaker biometric analysis and present a significant step toward optimal stress level assessment in spontaneous, unrestricted speech.

This study brings about several issues that could be investigated to further explore and improve stress level assessment in speech, namely relating to feature selection for the development of stress level assessment algorithms. The experimental observations leading to the selection of the $\overline{\Delta_{TEO}}$ feature for the feature space should be substantiated by further physiological experimental and theoretical analysis of airflow dynamics relating to the TEO-profile irregularity in frequency bands 3 and 9. Inquiry into the robustness to noise of the features chosen in this paper might lead to improved performance of stress level classification and assessment schemes. Furthermore, given the consistent quasi-linear behavior of the features selected in this study over increasing stress level, methods to predict stress levels should be explored.

Finally, the impact of stress in speech on an in-set speaker recognition task is evaluated. It is shown that exposure of the low/mid stress level trained system to high level stress samples results in a complete performance breakdown (accuracy drops from 91.7 % to 21.4 %). It is expected that availability of automatic stress assessment schemes such as those proposed in this study can be exploited in the design of stress-robust speaker recognition systems; e.g., through stress-dependent model selection or stress-adaptive feature normalizations.

Acknowledgements This project was funded by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

References

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings - Instituut Voor Fonetische Wetenschappen, Universiteit Van Amsterdam*, *17*, 97–110.

Bořil, H. (2008). *Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora*. Ph.D. thesis, Czech Technical University in Prague, Czech Republic. http://www.utdallas.edu/~hxb076000.

Bořil, H., & Hansen, J. H. L. (2010). Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*, 1379–1393.

Bořil, H., Sadjadi, O., Kleinschmidt, T., & Hansen, J. H. L. (2010). Analysis and detection of cognitive load and frustration in drivers' speech. In *Interspeech'10*, Makuhari, Chiba, Japan (pp. 502–505).

Bořil, H., Boyraz, P., & Hansen, J. H. L. (2012). Towards multi-modal driver's stress detection. In J. Hansen, P. Boyraz, K. Takeda & H. Abut (Eds.), *Digital signal processing for in-vehicle systems and safety* (pp. 3–20). New York: Springer.

Bou-Ghazale, S., & Hansen, J. (1998). HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. *IEEE Transactions on Speech and Audio Processing*, *6*, 201–216.

Brockwell, P., & Davis, R. (2002). *Introduction to time series and forecasting*. New York: Springer.

Cairns, D. A., & Hansen, J. H. L. (1994). Nonlinear analysis and classification of speech under stressed conditions. *The Journal of the Acoustical Society of America*, *96*, 3392–3400.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 357–366.

Hansen, J. H. L. (1988). *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*. Ph.D. thesis, Dept. of Elect. Eng., Georgia Institute of Technology, Atlanta, GA.

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

Hansen, J., & Varadarajan, V. (2009). Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*, 366–378.

Hansen, J., Sangwan, A., & Kim, W. (2012). Speech under stress and Lombard effect: impact and solutions for forensic speaker recognition. In H. Patil & A. Neustein (Eds.), *Forensic speaker recognition: law enforcement and counter-terrorism* (pp. 103–123). New York: Springer.

Ikeno, A., Varadarajan, V., Patil, S., & Hansen, J. (2007). UT-Scope: speech under Lombard effect and cognitive stress. In *IEEE aerospace conference* (pp. 1–7).

Kaiser, J. (1990a). On a simple algorithm to calculate the 'energy' of a signal. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing 1990 (ICASSP'90)*, Albuquerque, NM (Vol. 1, pp. 381–384).

Kaiser, J. (1990b). On Teager's energy algorithm and its generalization to continuous signals. In *Proc. 4th IEEE digital signal processing workshop*, Mohonk, NY.

Kirschbaum, C., Wolf, O. T., May, M., Wippich, W., & Hellhammer, D. H. (1996). Stress- and treatment-induced elevations of cortisol levels associated with impaired declarative memory in healthy adults. *Life Sciences*, *58*, 1475–1483.

Meyerhoff, J. L., Oleshansky, M. A., & Mougey, E. H. (1998). Psychological stress increases plasma levels of prolactin, cortisol and POMC-derived peptides in man. *Psychosomatische Medizin*, *50*, 295–303.

Meyerhoff, J. L., Norris, W., Saviolakis, G., Wollert, T., Burge, B., Atkins, V., & Spielberger, C. (2004). Evaluating performance of federal law enforcement personnel during a stressful training scenario. *Annals of the New York Academy of Sciences*, *1032*, 250–253.

Mitra, S. K. (2001). *Digital signal processing: a computer-based approach* (2nd ed.). New York: McGraw Hill.

Patil, S. A., & Hansen, J. H. (2010). The physiological microphone (PMIC): a competitive alternative for speaker assessment in stress detection and speaker verification. *Speech Communication*, 327–340.

Rahurkar, M., Hansen, J., Oleshansky, M., Meyerhoff, J., & Koenig, M. (2002). Frequency band analysis for stress detection using a Teager energy operator-based feature. In *ICSLP–02*, Denver, CO (pp. 2021–2024).

Rajasekaran, P., Doddington, G., & Picone, J. (1986). Recognition of speech under stress and in noise. In *Proc. of ICASSP'86*, Tokyo, Japan (Vol. 11, pp. 733–736).

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41.

Ruzanski, E., Hansen, J., Meyerhoff, J., Saviolakis, G., & Koenig, M. (2005). Effects of phoneme characteristics on teo feature-based automatic stress detection in speech. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing 2005 (ICASSP'05)*, Philadelphia, PA (Vol. 1, pp. 357–360).

Saunders, T., Driskell, J., Johnston, J., & Salas, E. (1996). The effect of stress inoculation training on anxiety and performance. *Journal of Occupational Health Psychology*, *1*, 170–186.

Scharf, B. (1970). Critical bands. In V. Tobias (Ed.), *Foundation of modern auditory theory*. New York: Academic Press.

Schwarz, P. (2009). *Phoneme recognition based on long temporal context*. Ph.D. thesis, Brno University of Technology, Czech Republic.

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing 2002 (ICASSP'02)*, Orlando, FL (Vol. 1, pp. 333–336).

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech coding and synthesis* (pp. 495–518). Amsterdam: Elsevier.

Teager, H. M. (1980). Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 599–601.

Teager, H., & Teager, S. (1989). Evidence for nonlinear production mechanisms in the vocal tract. *Speech Production and Speech Modelling*, *55*, 241–261.

Theodoridis, S., & Koutroumbas, K. (2003). *Pattern recognition* (2nd ed.). Amsterdam: Elsevier.

Womack, B., & Hansen, J. (1999). N-channel hidden Markov models for combined stress speech classification and recognition. *IEEE Transactions on Speech and Audio Processing*, *7*, 668–677.

Yost, W. (1994). *Fundamentals of hearing*. New York: Academic Press.

Zhou, G., Hansen, J., & Kaiser, J. (2001). Nonlinear feature-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *9*, 201–216.