Nonlinear Feature Based Classification of Speech Under Stress

Guojun Zhou, *Member, IEEE*, John H. L. Hansen, *Senior Member, IEEE*, and James F. Kaiser, *Fellow, IEEE*

***Abstract*—Studies have shown that variability introduced by stress or emotion can severely reduce speech recognition accuracy. Techniques for detecting or assessing the presence of stress could help improve the robustness of speech recognition systems. Although some acoustic variables derived from linear speech production theory have been investigated as indicators of stress, they are not always consistent. In this paper, three new features derived from the nonlinear Teager energy operator (TEO) are investigated for stress classification. It is believed that the TEO based features are better able to reflect the nonlinear airflow structure of speech production under adverse stressful conditions. The features proposed include TEO-decomposed FM variation (TEO-FM-Var), normalized TEO autocorrelation envelope area (TEO-Auto-Env), and critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env). The proposed features are evaluated for the task of stress classification using simulated and actual stressed speech and it is shown that the TEO-CB-Auto-Env feature outperforms traditional pitch and mel-frequency cepstrum coefficients (MFCC) substantially. Performance for TEO based features are maintained in both text-dependent and text-independent models, while performance of traditional features degrades in text-independent models. Overall neutral versus stress classification rates are also shown to be more consistent across different stress styles.**

***Index Terms*—Human factors, nonlinear speech feature, speech analysis, speech recognition, stress classification, Teager energy operator (TEO).**

I. Introduction

STRESS and its effects on the acoustic speech signal have been the subject of many studies [1], [2]. Adverse environments, such as noisy backgrounds, emergency conditions, high workload stress, multitasking, fatigue due to sustained operation, physical environmental factors (G-force), emotional moods, etc., are some of the factors which introduce stress into the speech production process. When a speaker produces speech in the presence of background noise, Lombard effect [35] will also occur since the speaker must modify his/her speech in order to increase communication quality over the noisy environment. Numerous studies [6], [10], [16], [23], [24], [28], [29], [40], [45], [46] have shown distinctive differences in phonetic features between normal and speech produced under Lombard effect. Under emergency conditions such as that in aircraft pilot communications, speech normally is produced in a fast manner and can have aspects of emotional fear. High workload, multitasking, and/or fatigue could cause speech to sound slower, faster, softer, or louder than speech produced under neutral environments. The physical G-force movement, which a fighter cockpit pilot experiences during real maneuvers, or the movement a person might experience while riding a roller coaster, can disrupt the typical speech production process. A study by South [2] showed that pilots undergoing high G-force in a centerfuge resulted in a shrinking $F1$ versus $F2$ (first, second formant) vowel space. Moreover, emotional arousal can cause changes in respiration pattern and muscle tension in the vocal tract. Such changes in speech production brought on by a variety of emotions have been the focus of a number of research investigations [7], [16], [23], [53].

It is well-known that the performance of speech recognition algorithms is greatly influenced by the stressful conditions in which speech is produced. Workload task stress has been shown to significantly impact recognition performance [3], [4], [11], [16], [39], [41], [43], [54]. Adverse influence of the Lombard effect on speech recognition has been reported in [28], [46]. Effects of different stressful conditions on speech recognition and efforts to improve the performance of speech recognition algorithms under stressful conditions can be found in [3], [11], [16], [17], [19]–[22], [41].

For speech recognizers, a typical approach to improve recognition robustness under adverse conditions (e.g., varying communication channels, handset differences) is re-training reference models (i.e., train-test in matched conditions). A similar method, called multi-style training [34], has been used to improve speech recognition under stress, but at the expense of requiring the user to produce speech across a simulated range of stress styles. In a separate study, it was shown that multistyle training only works in speaker-dependent scenarios and that performance actually degrades below neutral training when applied in a speaker independent application [55]. The reason is that stressful conditions are too diverse to be represented by limited training data, and that speakers can at times use a nonuniform set of speech production adjustments to convey their stress state. A study by Bou-Ghazale and Hansen [8] explored this notion by developing perturbation models of neutral-to-stress using a hidden Markov model (HMM) framework. They were able to synthesize multi-style like speech recognition models by perturbing the neutral training tokens of an input speaker using perturbation models from a

Manuscript received December 22, 1997; revised May 6, 2000. This work was supported by a grant from the U.S. Air Force Research Laboratory, Rome, NY. The associate editors coordinating the review of this manuscript and approving it for publication were Dr. Gerard Chollet and Dr. B.-H. Juang.

G. Zhou was with the Robust Speech Processing Laboratory, Center for Spoken Language Research, University of Colorado, Boulder, CO 80309 USA. He is now with Intel's Architecture Labs, Intel Corporation, Hillsboro, OR 97124 USA.

J. H. L. Hansen and J. F. Kaiser are with the Robust Speech Processing Laboratory, Center for Spoken Language Research, University of Colorado, Boulder, CO 80309 USA (e-mail: John.Hansen@colorado.edu).

Publisher Item Identifier S 1063-6676(01)01323-2.

second set of speakers. Their results showed that recognition performance can be improved, but not to the same degree as seen for speaker-dependent stress models. It is suggested that algorithms which are capable of classifying stress could be used to classify stressed speech from neutral. Model adaptation techniques can be further used to adapt models so that stressed speech can be recognized well.

In fact, stress classification cannot only be used to improve the robustness of speech recognition systems, other scenarios can also benefit, such as telecommunications, military applications, medical applications, and law enforcement. In telecommunications, in addition to its potential to improve the telephone-based speech recognition performance, stress classification can be used to route 911 emergency call services for high priority emergency calls. Moreover, it can also be used to assess a caller's emotional state for telephone response services. The integration of speech recognition technology has already been seen in many military voice communication and control applications. Since many such applications involve stressful environments (e.g., aircraft cockpits, military peacekeeping/battlefield setting), stress classification and assessment become crucial to improve the system robustness in these applications [27]. Furthermore, computerized stress classification and assessment techniques can be employed by psychiatrists to aid in quantitative objective assessment of patients undergoing evaluation. Finally, stress classification can also be employed in forensic speech analysis by law enforcement to assess the state of telephone callers or as an aid in suspect interviews.

Although much research has been conducted on stressful conditions for speech recognition, there has been limited work performed in the area of stressed speech classification. The majority of studies in the field of speaker stress analysis have concentrated on pitch, with several considering spectral features derived from a linear model of speech production [23], [53], [16], [55], [57]. The number of studies in stress classification is much more limited. One recent study [24] considered stress classification using

1) estimated vocal tract area profiles;
2) acoustic tube area coefficients;
3) Mel-cepstral based parameters (MFCC [13]) including Mel-cepstral (MFCC), delta MFCC, delta-delta-MFCC, and a new feature based on the autocorrelation of the MFCCs (AC-mel).

Stress classification performance using these features were determined using separability distance metrics and neural network based classifiers. It was shown that stress classification performance varied significantly depending on the vocabulary size and speaker population. However, MFCC and AC-mel performed better than delta-MFCC and delta-delta-MFCC for vocabulary dependent tests. A later study showed that by using target driven features and context dependent phoneme neural networks, stress classification performance could be measurably improved [55]. Other acoustic features which have also been shown to be useful as indicators of speech under stress include fundamental frequency ($F0$), phoneme duration and intensity, glottal source structure (especially spectral slope), and vocal tract formant structure [23].

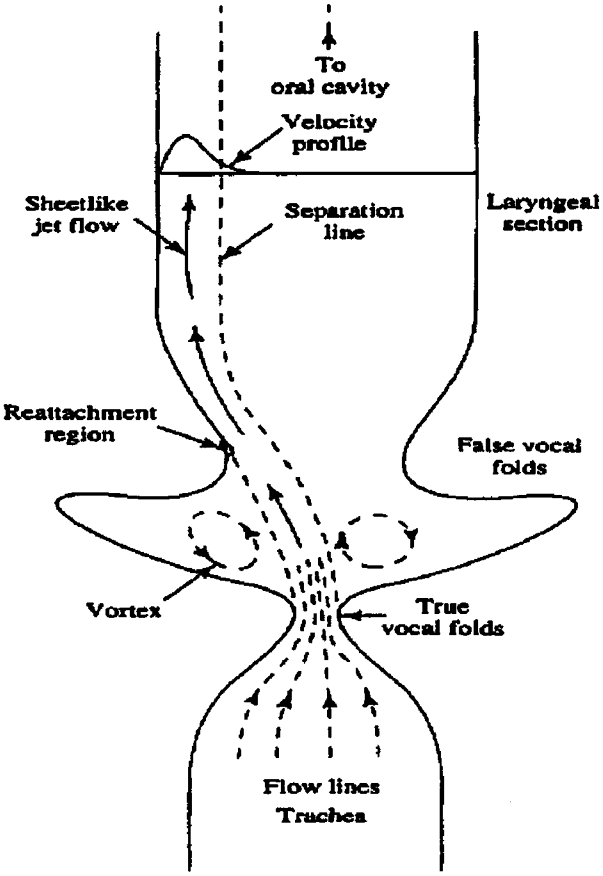


Fig. 1. Nonlinear model of sound propagation along the vocal tract.

All speech features used in [55], [23], which include the MFCC, are derived from a linear speech production models which assume that airflow propagates in the vocal tract as a plane wave. This pulsatile flow is considered the source of sound production. According to studies by Teager [49]–[51], however, this assumption may not hold since the flow is actually separate and concomitant vortices are distributed throughout the vocal tract (shown in Fig. 1 [30]).

Teager suggested that the true source of sound production is actually the vortex-flow interactions, which are nonlinear. This observation was supported by the theory in fluid mechanics [12] as well as by numerical simulation of Navier–Stokes equation [52]. It is believed that changes in vocal system physiology induced by stressful conditions such as muscle tension will affect the vortex-flow interaction patterns in the vocal tract. Therefore, nonlinear speech features are necessary to classify stressed speech from neutral.

It can be stated that there are two broad ways to model the human speech production process. One approach is to model the vocal tract structure using a source-filter model [15]. This approach assumes that the underlying source of phoneme identity comes from the vocal tract configuration of the articulators. Recent studies have explored the prospect of decomposing the system model characteristics for both vocal fold movement [5] and vocal tract structure [47]. An alternative way to characterize speech production is to model the airflow pattern in the vocal tract [52]. The underlying concept here, is that while the vocal tract articulators do move to configure the vocal tract shape, it is the resulting airflow properties which serve to excite those models which a listener will perceive as a particular phoneme. Studies by Teager emphasized this approach [49]–[51], with follow-up investigations by Kaiser [31]–[33] to support those concepts. Although the airflow pattern shown in Fig. 1 may be closer to that of the real speech production process, it is very difficult, if not impossible to model it mathematically, since complete Navier–Stokes solutions of airflow require accurate boundary conditions versus time. In an effort to reflect the instantaneous energy of nonlinear vortex-flow interactions, Teager developed an energy operator, with the supporting observation that hearing is the process of detecting the energy. The simple and elegant form of the operator was introduced by Kaiser [32] as

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

where $\Psi[\,\cdot\,]$ is the Teager energy operator (TEO), and $x(t)$ is a single component of the continuous speech signal.

One previous study [9] considered stress classification using a nonlinear feature based on properties of TEO, where the shape of a pitch normalized TEO profile was used. Good performance was obtained for speech produced under angry, loud, clear, and Lombard effect speaking conditions. That study, however, was limited to stress classification of extracted front and mid vowels.

Our focus, here, is to remove phone or word level dependency in the stress classification task, and thereby concentrate on correlates of nonlinear excitation characteristics associated with stress. For this purpose, we propose three new features which incorporate TEO-based processing in this study. The features are entitled TEO-decomposed FM variation (TEO-FM-Var), normalized TEO autocorrelation envelope area (TEO-Auto-Env), and critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env). These features explore the prospects of variations in the energy of airflow characteristics within the vocal tract for speech under stress. We compare the performance of the proposed TEO-based features to traditional MFCC and pitch information for the task of stress classification using speech under simulated and actual stress from data provided by NATO IST/TG-01 (SUSAS, SUSC-0).[1]

The paper is organized as follows. In Section II, the background of the nonlinear Teager energy operator (TEO) is first described, followed by sections where we propose three new TEO-based stress classification features. An extensive set of evaluations and discussion are presented in Section III using speech under stress from several simulated and actual stress conditions. Finally, Section IV presents conclusions.

II. Stress Classification Features

A. Background of the Teager Energy Operator

The continuous from of the TEO is shown in (1). Since speech is represented in discrete form in most current speech processing systems, Kaiser [31], [33] derived the operator for discrete-time signals from its continuous form $\Psi_c[x(t)]$, as

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (2)$$

where $x(n)$ is the sampled speech signal. For example, the resulting continuous TEO response for $x(t) = A\cos\Omega t$ is a constant: $\Psi[x(t)] = A^2\Omega^2$; and the response for the discrete equivalent signal, $x(n) = A\cos\omega n$, is $\Psi[x(n)] = A^2\sin^2\omega$.

The TEO is typically applied to a bandpass filtered speech signal, since its intent is to reflect the energy of the nonlinear flow within the vocal tract for a single resonant frequency. Although the output of a bandpass filter still contains more than one frequency component, it can be considered as an AM–FM signal, $r(t) = a(t)\cos(2\pi f(t)t)$. The TEO output of $r(t)$ can be approximated as

$$\Psi[r(t)] \approx [a(t)2\pi f(t)]^2. \qquad (3)$$

This notion will be further explored during feature derivation in Section II-D.

In fact, the TEO profile can be used to decompose an AM–FM signal into its AM and FM components within a certain frequency band via

$$f(n) \approx \frac{1}{2\pi T}\arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right), \qquad (4)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{\left[1 - \left(1 - \dfrac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2\right]}} \qquad (5)$$

where

$y(n) = x(n) - x(n-1)$	time domain difference signal;
$\Psi[\,\cdot\,]$	TEO operator as shown in (2);
$f(n)$	FM component at sample n;
$a(n)$	AM component at sample n [36], [37].

On the basis of this work, Maragos *et al.* [37] proposed a nonlinear model which represents the speech signal $s(t)$ as

$$s(t) = \sum_{m=1}^{M} r_m(t) \qquad (6)$$

where

$$r_m(t) = a_m(t)\cos\left(2\pi\left(f_{cm}t + \int_0^t q_m(\tau)d\tau\right) + \theta\right) \qquad (7)$$

is a combined AM and FM structure representing a speech resonance at the mth formant with a center frequency $F_m = f_{cm}$. In this relation, $a_m(t)$ is the time-varying amplitude, and $q_m(\tau)$ is the frequency modulating signal at the mth formant.

Although TEO processing is intended to be used for a signal with a single resonant frequency, we will find in Section II-D that the TEO energy of a multi-frequency signal does not only reflects individual frequency components but also reflects interactions between them. This characteristic extends the use of TEO to speech signals filtered with wide bandwidth band-pass filters (BPF). These observations led us to propose the TEO-

[1] For further information on NATO IST/TG-01 efforts on stress, see their speech under stress web page at http://cslu.colorado.edu/rspl/stress.html.

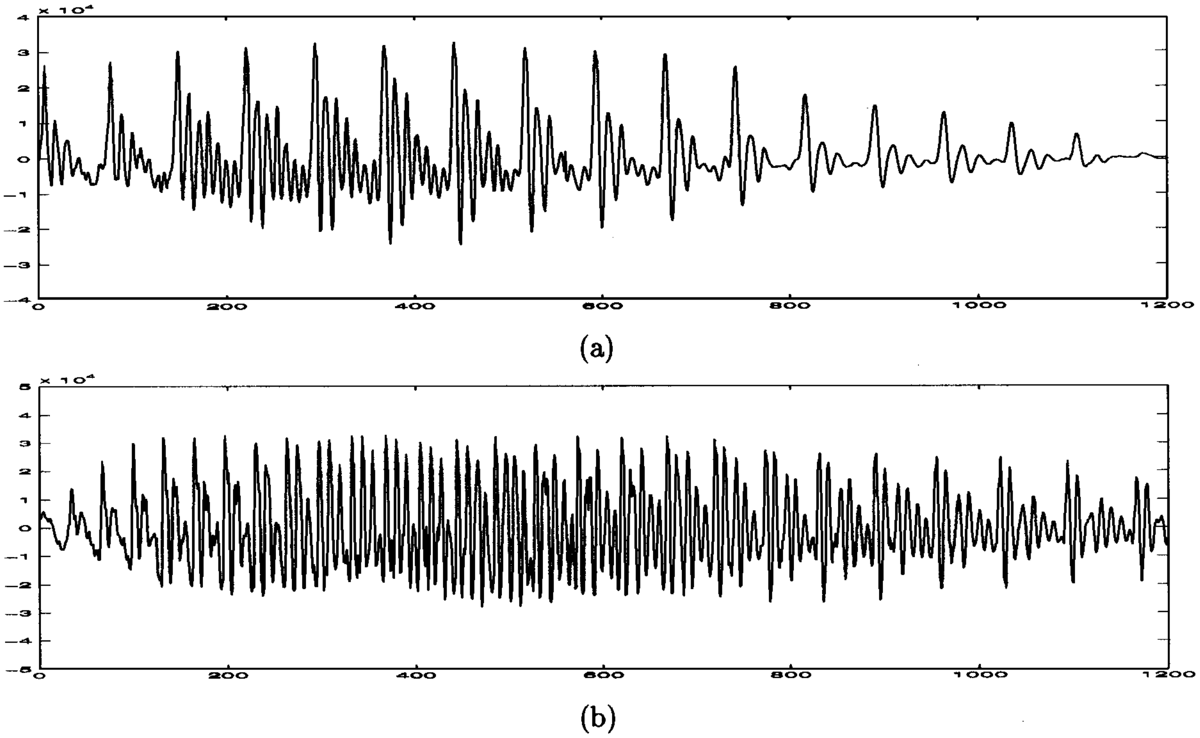


Fig. 2. Waveforms of 150-ms duration obtained from the voiced portion of word "help" spoken by the same male speaker under (a) neutral and (b) simulated angry conditions.

based stress classification features discussed in the following subsections.

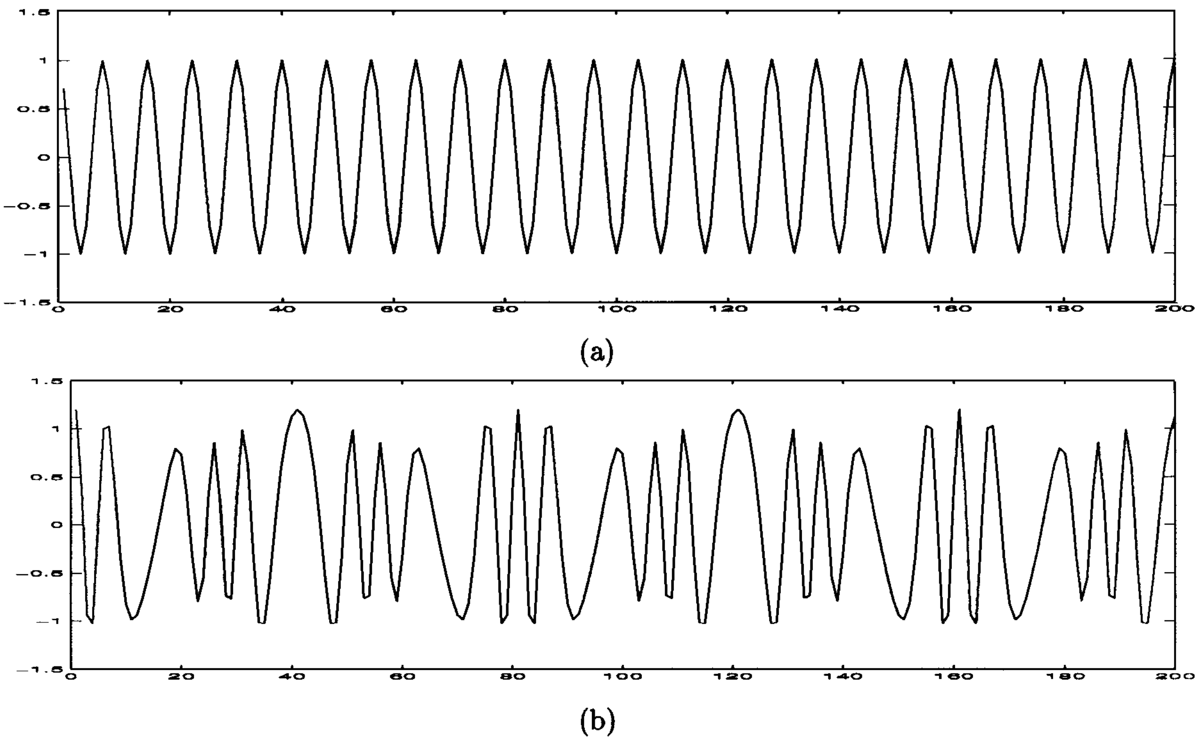
B. TEO-FM-Var: Variation of FM Component

Voiced speech spoken under stress generally has different instantaneous excitation variations from voiced speech spoken under neutral conditions. This can be verified by comparing voiced speech waveforms spoken under neutral and simulated angry conditions. For example, Fig. 2 shows sample waveforms from the voiced part of the word "help" in both neutral and angry conditions. The differences in pitch excitation is clearly evident. Therefore, features which represent fine excitation variations, should be useful for stress classification. This observation must also be verified across a range of voiced phonemes and speakers. We consider this later in the evaluation section. However, it is reasonable to believe that the fine excitation variations observed in the speech signal are due to the effects of modulations. This point is supported by comparing the waveforms of a pure steady-state sinusoidal signal and a slowly modulating AM–FM signal (shown in Fig. 3). We see that the AM and FM components cause measurable variations in the resulting waveform. It is believed that the modulation patterns observed in Fig. 3 are perhaps similar to the modulation variations due to stress in Fig. 2. Therefore, a stress classification feature is needed which reflects these modulation variations.

While it might seem straightforward to apply a standard pitch estimation algorithm to estimate these variations, the large and erratic pitch changes under stress generally cause traditional estimation algorithms to fail, thus requiring human pitch label correction [16]. An alternative is to use the FM variation of each frame as the feature for stress classification. Since AM–FM signal analysis requires a carrier frequency which must be higher than the modulating frequencies within the signal, we filter the raw input speech through a Gabor bandpass filter [37] (BPF) centered at the median fundamental frequency, $F0$, with the root mean square (RMS) bandwidth of $F0/2$. The Gabor BPF is employed since it has excellent sidelobe cancellation. Here, we are only interested in fine excitation variations which are believed to reflect changing levels of speaker stress. The absolute magnitude difference function (AMDF) [42] is employed to automatically estimate the median fundamental frequency, $F0$, based on the TEO profile of the entire input. The reason to estimate $F0$ based on the TEO profile is that the TEO profile usually reflects better and more consistent period-to-period pitch information than that obtained in the original speech signal partly due to the square effect of the TEO. After the Gabor BPF, the TEO is applied and the resulting profile is used to separate the input speech signal into its AM and FM components using (4) and (5). The frame-based FM variations are further computed as the proposed feature. A flow diagram for extracting the first TEO-based feature (TEO-FM-Var) is shown in Fig. 4. Example waveforms are also shown at each stage of the feature extraction for neutral and stressed speech. We observe considerable differences in the final and intermediate feature response betweeen neutral and stressed speech.

C. TEO-Auto-Env: Normalized TEO Autocorrelation Envelope Area

The second TEO-based feature entitled TEO-Auto-Env also reflects the instantaneous excitation variations of speech. A flow diagram is shown in Fig. 5.



(a)

(b)

Fig. 3. Sample waveforms from (a) a single frequency (1 kHz) and (b) a modulated AM/FM response.

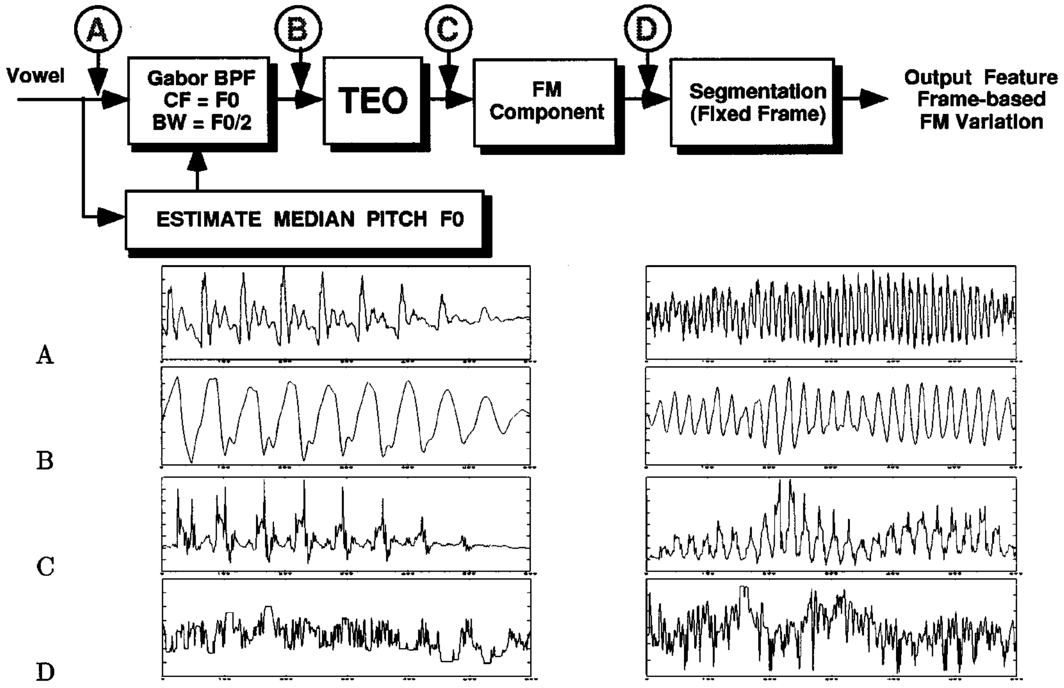


Fig. 4. TEO-FM-Var Feature Extraction [waveforms represents a segment of /IH/ sound in the word fix under neutral (left column) and stressed (right column) conditions].

The motivation for the TEO-FM-Var feature is to capture stress dependent information that may be present in changes within the FM component. Its processing is based on the entire band although the final FM variations are computed around the restricted frequency band. However, the presence of stress may affect modulation patterns across the entire speech frequency band. According to the nonlinear model proposed by Maragos *et al.* [36], [37], voiced speech can be modeled as the sum of AM–FM signals of which each is centered at a formant frequency [shown in (6)]. If a filter bank is used to bandpass filter voiced speech around each of its formant frequencies, the modulation pattern around each formant can be obtained using TEO

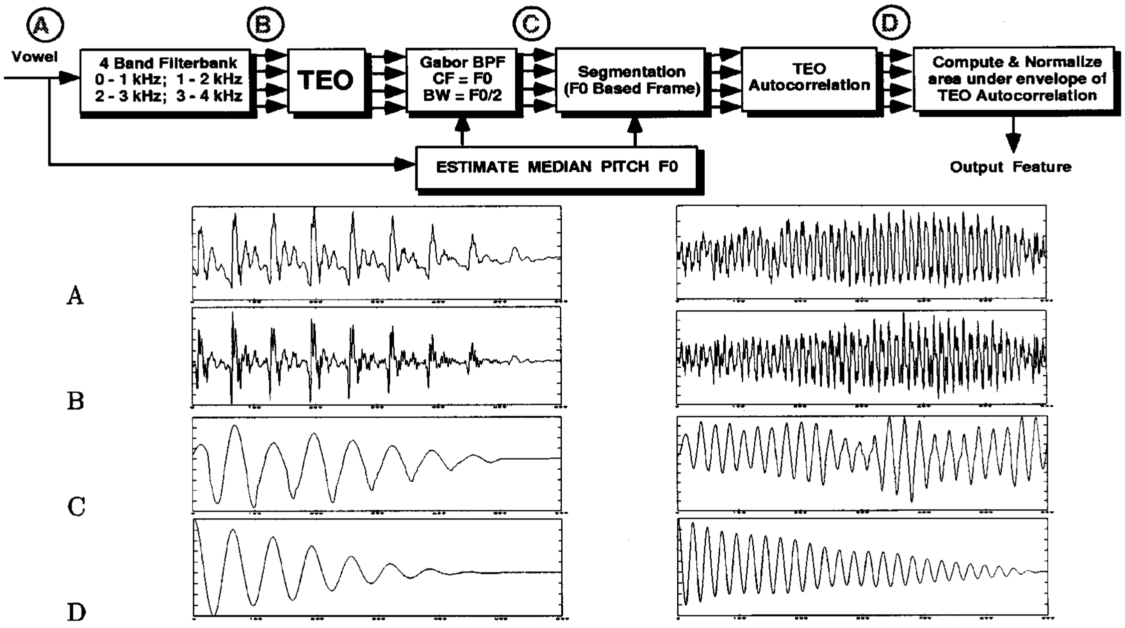


Fig. 5. TEO-Auto-Env Feature Extraction [all waveforms for B, C, and D are for the 2nd band, 1–2 kHz; waveforms represents a segment of /IH/ sound in the word fix under neutral (left column) and stressed (right column) conditions].

AM–FM decomposition, from which variations of modulation patterns across different frequency bands can be obtained. Such an approach, however, requires tracking all the formant frequencies, which could be difficult to estimate reliably since most traditional formant tracking algorithms fail when speech is spoken under stress, due to the large and erratic excitation variation [16], [23]. To avoid the difficulty of automatic formant tracking, four fixed bandpass filters are used with frequency ranges of (0–1 kHz), (1–2 kHz), (2–3 kHz), and (3–4 kHz), respectively. The number of formants which fall into each of the four frequency bands could range from 0 to 2 under neutral speaking conditions [14]. Under stressful conditions, however, the formants can shift their location in frequency, and therefore migrate into an adjacent filter (i.e., increase/decrease the location formants by as much as 6% (a 3%–6% change for F_1, F_2, and 0%–3% for F_3 and F_4) [16]). Different types, or varying degrees, of stress will influence the distribution of formant characteristics, and pitch structure and spectral based pitch harmonics from neutral conditions. As a side note, in addition to the primary issue of formant migration into adjacent filters, additional pitch harmonics would also occur. This concept is addressed in more detail in the following critical band based TEO feature (i.e., TEO-CB-Auto-Env).

The TEO-Auto-Env feature is obtained by passing the raw input speech through a filterbank consisting of 4 bandpass filters (BPF) (see Fig. 5). Each BPF output stream is processed to obtain an estimate of each TEO profile. Since the TEO output of a signal is roughly proportional to the square of both its amplitude and frequency as shown in (3), and the AM component for a single formant exhibits periodicity similar to the fundamental frequency, therefore, filtering the TEO profile with a filter centered at $F0$ captures variations around $F0$. A Gabor filter with a 3 dB bandwidth roughly equal to $F0/2$ can achieve this. $F0$ is obtained by using the same method as that used in the TEO-FM-Var feature extraction. Subsequently, each Gabor-filtered TEO stream is segmented into frames. In order to have equivalent averaging effects for the formant variations, the frame length is set to four times the median pitch period. Furthermore, the normalized autocorrelation function is computed for each frame. In the present formulation, if there is no pitch variation within a frame, the output TEO is a constant and its corresponding normalized autocorrelation function is a decaying straight line from (0, 1) to $(N, 0)$, where N is the frame length. The area under this ideal envelope (a straight line) for this frame should be $N/2$. In the case when pitch variation is present in a frame, its normalized autocorrelation envelope will not be an ideal straight line, and hence the area under the envelope will be less than $N/2$.[2] By computing the area under the normalized autocorrelation envelope and normalizing it by $N/2$, we can obtain four normalized TEO autocorrelation envelope area parameters for each time frame (i.e., one for each frequency band) which reflects the degree of excitation variability within each band. Fig. 5 also shows example waveforms extracted at points during TEO-Auto-Env feature processing for the second subband (1–2 kHz). By comparing the extracted waveforms for neutral and stressed speech, we see significant changes that we believe would allow the TEO-Auto-Env feature to respond favorably for a task in stress. Similar degrees of profile variation was also observed for the other subband frequencies.

D. TEO-CB-Auto-Env: Critical Band Based TEO Autocorrelation Envelope

The uniform partition of the entire speech frequency band for the TEO-Auto-Env was performed in an attempt to cap-

[2] Since the area under the envelope is obtained by tracking the autocorrelation peaks, its area can at most equal the autocorrelation response only if the autocorrelation function is a straight line.

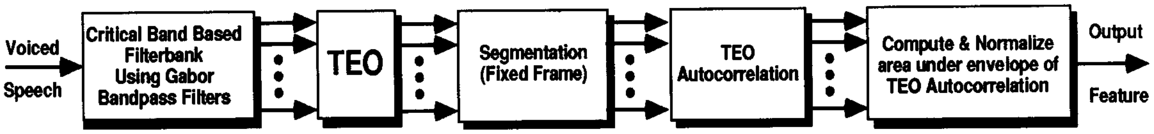

Voiced
Speech
Critical Band Based Filterbank Using Gabor Bandpass Filters
TEO
Segmentation (Fixed Frame)
TEO Autocorrelation
Compute & Normalize area under envelope of TEO Autocorrelation
Output
Feature

Fig. 6. TEO-CB-Auto-Env feature extraction.

ture stress sensitive changes outside the first formant. The TEO-Auto-Env feature allows us to probe nonlinear energy changes at higher frequencies. However, the frequency partition was coarse (i.e., 1 kHz bandwidth). A finer partition might help derive a more effective feature for stress classification. Empirically, the human auditory system is assumed to perform a filtering operation which partitions the entire audible frequency range into many critical bands [44], [56]. Based on this observation, the third proposed feature employs a critical band based filterbank to filter the speech signal followed by TEO processing (see Fig. 6). Each filter in the filterbank is a Gabor bandpass filter, with effective RMS bandwidth being the corresponding critical band. To extract the TEO-CB-Auto-Env feature, each TEO profile of a Gabor BPF output is segmented into 200-sample (25 ms) frames with 100-sample (12.5 ms) overlap between two adjacent frames. Similar to the extraction of the TEO-Auto-Env feature, M normalized TEO autocorrelation envelope area parameters are extracted for each time frame (i.e., one for each critical band), where M is the total number of critical bands. This is the TEO-CB-Auto-Env feature vector per frame. Fig. 6 shows the entire feature extraction procedure.

1) Harmonic Analysis: The TEO-Auto-Env feature extraction is subject to the accuracy of median $F0$ extraction, which is not always reliable. The TEO-CB-Auto-Env extraction attempts to remove $F0$ estimation dependency. Although the TEO-CB-Auto-Env appears similar in structure to the TEO-Auto-Env feature, both features are actually representing very different aspects in the speech signal. The TEO-Auto-Env attempts to represent the variations around pitch caused by formant distribution variations across different frequency bands; while TEO-CB-Auto-Env is focused more on representing the variations of pitch harmonics since it has much higher frequency resolution than the TEO-Auto-Env. When spoken under stressful conditions, a speech signal's fundamental frequency will typically change so that the distribution pattern of pitch harmonics across critical bands will be different from that of speech spoken under neutral conditions. To verify this, we manually computed the average harmonic number in each critical band from 12 voiced tokens for each of the four speaking styles in the SUSAS (discussed in Section III-A) simulated stress domain (shown in Table I). For each voiced token, average pitch was calculated and the number of harmonics (based on averaged pitch) which fall in each critical band was obtained. From Table I, we can clearly see the differences in harmonic distribution across critical bands between neutral, angry, loud and Lombard speech. The difference in the number of harmonic terms within each band, as well as the regularity of each harmonic, both influence the resulting TEO features between neutral and stress conditions. Note that in the analysis for Table I, we did not attempt to quantify the number or form of the cross harmonic terms, due to their increased complexity; but clearly they will also influence the resulting feature response.

TABLE I
DISTRIBUTION OF PITCH HARMONICS ACROSS CRITICAL BANDS

Statistics of Pitch Harmonics across Critical Band
Data about Critical Band is from [34]

Band Number	Critical Band Frequency Information (Hz)				Average Harmonic Number (Obtained from 12 voiced tokens)			
	Lower	Center	Upper	Bandwidth	Neutral	Angry	Loud	Lombard
1	100	150	200	100	1.00	1.00	1.00	1.00
2	200	250	300	100	1.00	0.08	0.58	0.17
3	300	350	400	100	1.00	1.00	0.42	0.83
4	400	450	510	110	1.08	0.25	0.92	1.00
5	510	570	630	120	1.17	1.08	0.83	0.75
6	630	700	770	140	1.42	0.83	0.92	0.42
7	770	840	920	150	1.33	0.75	0.92	1.00
8	920	1000	1080	160	1.42	1.08	1.08	1.00
9	1080	1170	1270	190	1.75	1.08	1.33	1.67
10	1270	1370	1480	210	1.92	1.25	1.42	1.33
11	1480	1600	1720	240	2.42	1.42	1.42	1.33
12	1720	1850	2000	280	2.42	1.58	2.00	1.58
13	2000	2150	2320	320	3.08	2.08	2.08	2.08
14	2320	2500	2700	380	3.25	2.08	2.58	2.67
15	2700	2900	3150	450	4.25	2.67	2.92	3.00
16	3150	3400	3700	550	5.00	3.17	3.67	3.33

2) Quantitative Analysis: Next, we wish to quantitatively verify how the difference of pitch harmonic distributions across critical bands affect the TEO output from each critical band. We assume that two harmonics ω_{η_1} and ω_{η_2} exist in a critical band under neutral conditions, and that only one harmonic ω_{ζ_1} in the same critical band due to an increased fundamental frequency when the same speech is produced under stressful conditions. As a result, the TEO autocorrelation response from this critical band under neutral conditions will be different. Let us assume the output of a particular band i under neutral speech conditions can be written as $\eta^i(n)$, and under stress conditions as $\zeta^i(n)$. Since the fundamental frequency for neutral speech will be much lower, the critical band will typically possess more harmonic frequencies. If we assume a male speaker doubles his pitch under stress;[3] then we could assume that the output signal from the critical band possesses two harmonics for neutral, and one harmonic for stress as follows:

$$\eta^i(n) = A_{\eta_1}\cos(\omega_{\eta_1}n) + A_{\eta_2}\cos(\omega_{\eta_2}n) \tag{8}$$

$$\zeta^i(n) = A_{\zeta_1}\cos(\omega_{\zeta_1}n). \tag{9}$$

Here, the amplitudes A_{η_1}, A_{η_2} and A_{ζ_1} should be functions of time t, however, to simplify our discussion, we assume that they are all constants. Next, we apply the TEO to $\eta^i(n)$ and $\zeta^i(n)$,

[3] Previous analysis of one sample speaker from SUSAS showed a mean pitch for neutral speech of 121 Hz and 243 Hz for speech under angry conditions.

which produces the following relations:

For Neutral:

$$\begin{aligned}\Psi\left[\eta^i(n)\right] &= \left(\eta^i(n)\right)^2 - \eta^i(n-1)\eta^i(n+1)\\ &= (A_{\eta_1}\cos(\omega_{\eta_1}n) + A_{\eta_2}\cos(\omega_{\eta_2}n))^2\\ &\quad - (A_{\eta_1}\cos(\omega_{\eta_1}(n-1)) + A_{\eta_2}\cos(\omega_{\eta_2}(n-1)))\\ &\quad \times (A_{\eta_1}\cos(\omega_{\eta_1}(n+1)) + A_{\eta_2}\cos(\omega_{\eta_2}(n+1)))\\ &= A_{\eta_1}^2\sin^2(\omega_{\eta_1}) + A_{\eta_2}^2\sin^2(\omega_{\eta_2}) + 2A_{\eta_1}A_{\eta_2}\sin^2\\ &\quad \times \left(\frac{\omega_{\eta_1}-\omega_{\eta_2}}{2}\right)\cos((\omega_{\eta_1}+\omega_{\eta_2})n)\\ &\quad + 2A_{\eta_1}A_{\eta_2}\sin^2\left(\frac{\omega_{\eta_1}+\omega_{\eta_2}}{2}\right)\cos((\omega_{\eta_1}-\omega_{\eta_2})n)\end{aligned} \tag{10}$$

For Stress:

$$\begin{aligned}\Psi\left[\zeta^i(n)\right] &= \left(\zeta^i(n)\right)^2 - \zeta^i(n-1)\zeta^i(n+1)\\ &= (A_{\zeta_1}\cos(\omega_{\zeta_1}n))^2 - (A_{\zeta_1}\cos(\omega_{\zeta_1}(n-1)))\\ &\quad \times (A_{\zeta_1}\cos(\omega_{\zeta_1}(n+1)))\\ &= A_{\zeta_1}^2\sin^2(\omega_{\zeta_1}).\end{aligned} \tag{11}$$

If we compare $\Psi[\eta^i(n)]$ and $\Psi[\zeta^i(n)]$, we see that the TEO output of band i under stress is a constant, while the same output under the neutral speech condition is a function of time index n, consisting of two frequencies, $\omega_{\eta_1} + \omega_{\eta_2}$ and $|\omega_{\eta_1} - \omega_{\eta_2}|$. This difference in the TEO responses will subsequently influence their autocorrelation functions. Let us first derive the autocorrelation function for the neutral TEO. We begin with the basic simple autocorrelation function

$$R_{\Psi[\eta^i]}(k) = \lim_{M\to\infty}\frac{1}{2M+1}\sum_{n=-M}^{M}\Psi\left[\eta^i(n)\right]\Psi\left[\eta^i(n+k)\right]. \tag{12}$$

Next, we substitute the final result from (10), and finally we can obtain

$$\begin{aligned}R_{\Psi[\eta^i]}(k) &= \cos\left((\omega_{\eta_1}+\omega_{\eta_2})(n+k)\right) + \sin^2\left(\frac{\omega_{\eta_1}+\omega_{\eta_2}}{2}\right)\\ &\quad \times \cos\left((\omega_{\eta_1}-\omega_{\eta_2})(n+k)\right)\\ &= \left(A_{\eta_1}^2\sin^2(\omega_{\eta_1}) + A_{\eta_2}^2\sin^2(\omega_{\eta_2})\right)^2 + 2A_{\eta_1}^2A_{\eta_2}^2\\ &\quad \times \left\{\sin^4\left(\frac{\omega_{\eta_1}-\omega_{\eta_2}}{2}\right)\cos((\omega_{\eta_1}+\omega_{\eta_2})k)\right.\\ &\quad \left. + \sin^4\left(\frac{\omega_{\eta_1}+\omega_{\eta_2}}{2}\right)\cos((\omega_{\eta_1}-\omega_{\eta_2})k)\right\}.\end{aligned} \tag{13}$$

This final autocorrelation function for the neutral TEO response is complex, with frequency terms consisting of $\omega_{\eta_1} + \omega_{\eta_2}$ and $|\omega_{\eta_1} - \omega_{\eta_2}|$. Similarly, we can obtain the autocorrelation function for the stressed speech TEO response as follows:

$$R_{\Psi[\zeta^i]}(k) = A_{\zeta_1}^4\sin^4(\omega_{\zeta_1}). \tag{14}$$

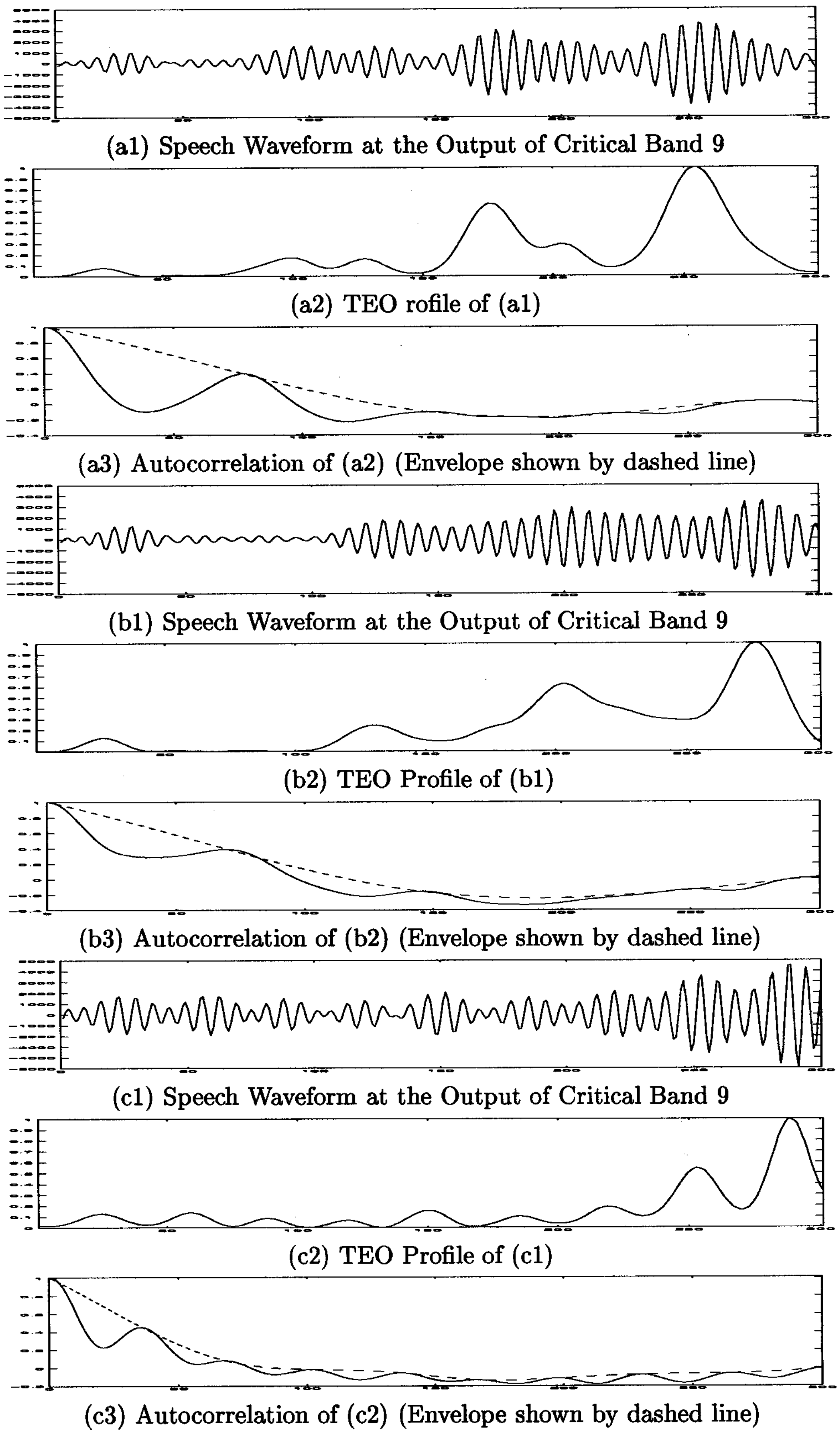
Clearly, the autocorrelation function for the stress case is a constant, independent of correlation lag k.

We again point out that the resulting autocorrelation functions in (13) and (14) resulted from the single and double harmonic outputs from a single critical band filter originally from (8) and (9). Although this mathematical derivation appears quite complex, this is in fact the simplest case since we are dealing with only a single or double harmonics. For this ideal case, one might suggest that calculating the TEO autocorrelation functions is unnecessary since they reflect the same variation trends as the TEO profile itself. In reality, however, critical band i may possess cross-harmonic terms in addition to the pure $F0$ harmonics. There may also be amplitude and/or frequency modulating terms corresponding to each harmonic or cross harmonic term. All of these factors can cause rapid changes in the TEO profile. The averaging effect of the autocorrelation calculation can suppress some of the fast-changing variations and still maintain those fluctuations which are believed to be due to stress. This process makes it easier to locate and track the upper envelope from the TEO autocorrelation function than from the TEO profile itself.

As a result, the constant TEO profile will be represented as the autocorrelation envelope which is a decaying straight line from (0, 1) to $(N, 0)$, where N is the frame length. Those variations caused by harmonic distribution differences as well as by modulations will be reflected by the change in the TEO autocorrelation envelopes.

3) Waveform Analysis: To further illustrate the output differences resulting from each critical band between neutral and stressed speech, waveform analysis for an arbitrary critical band was performed (band 9 was selected at random since it is a mid-frequency band). A segment with relatively stable pitch periods from the voiced section of "help" under the angry stress condition was employed for analysis. Accordingly, a corresponding segment from a neutral token of "help" was also extracted. For the example waveform analysis considered here, the pitch of the neutral segment was also artificially increased using a pitch-synchronous overlap-and-add (PSOLA) method [38] to the same pitch level of the segment under angry stress to obtain a new segment for the purpose of feature comparison. This step was performed so that the TEO-based features would reflect only the change in nonlinear speech or airflow characteristics. In effect, this allows us to separate the feature problem into two parts (i.e., suppress in impact of an increased pitch level. It is believed that the presence of stress causes an increase in the variability of airflow characteristics, due to differences in muscle tension of the vocal folds. This should cause changes in airflow patterns above the vocal folds, thus increasing the vortex interactions around the false vocal folds. The TEO is thus believed to represent a measure of the nonlinear energy present in this vortex airflow. However, under a stress condition such as anger, the rate of vocal fold movement is much higher. Therefore, while we believe the TEO output of each critical band filter will have increased variability under stress, the number of frequency harmonics in each frequency band will be less under stress (i.e., due to an increase in pitch). By adjusting the pitch of neutral to have the same mean as angry in this example, we can temporarily remove the impact of some of the resulting TEO cross-terms present in the given critical band filter.

Fig. 7 shows the output waveforms from critical band 9 (frequency between 1080 and 1270, Table I) for original neutral, pitch adjusted neutral, and angry. We plot the three speech



(a1) Speech Waveform at the Output of Critical Band 9

(a2) TEO rofile of (a1)

(a3) Autocorrelation of (a2) (Envelope shown by dashed line)

(b1) Speech Waveform at the Output of Critical Band 9

(b2) TEO Profile of (b1)

(b3) Autocorrelation of (b2) (Envelope shown by dashed line)

(c1) Speech Waveform at the Output of Critical Band 9

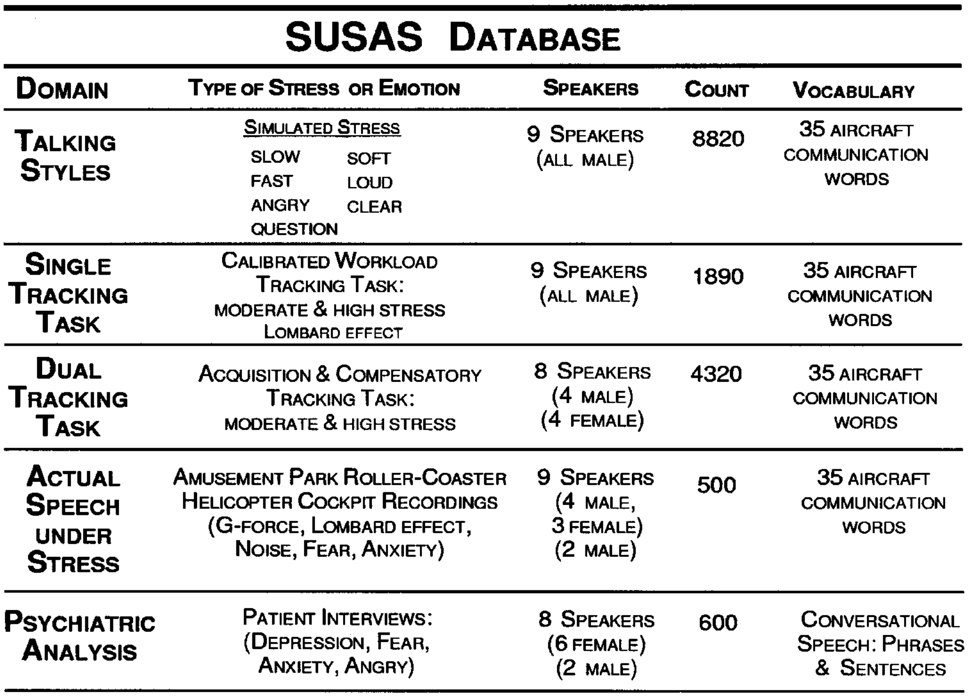
(c2) TEO Profile of (c1)

(c3) Autocorrelation of (c2) (Envelope shown by dashed line)

Fig. 7. Waveform analysis. (a) Neutral speech segment with average pitch $F0 = 111$ Hz, (b) pitch adjusted speech by increasing the pitch for neutral speech from (a) to 239 Hz, and (c) speech segment under angry stress with average pitch $F0 = 240$ Hz.

segments, their TEO profiles, and AM–FM energy components. Fourier transform analysis of this example showed that the output of critical band 9's neutral segment has two main peaks, which correspond to the main pitch harmonics in its spectrum;

TABLE II
DESCRIPTION OF SUSAS DATABASES



SUSAS DATABASE				
DOMAIN	TYPE OF STRESS OR EMOTION	SPEAKERS	COUNT	VOCABULARY
TALKING STYLES	SIMULATED STRESS SLOW SOFT FAST LOUD ANGRY CLEAR QUESTION	9 SPEAKERS (ALL MALE)	8820	35 AIRCRAFT COMMUNICATION WORDS
SINGLE TRACKING TASK	CALIBRATED WORKLOAD TRACKING TASK: MODERATE & HIGH STRESS LOMBARD EFFECT	9 SPEAKERS (ALL MALE)	1890	35 AIRCRAFT COMMUNICATION WORDS
DUAL TRACKING TASK	ACQUISITION & COMPENSATORY TRACKING TASK: MODERATE & HIGH STRESS	8 SPEAKERS (4 MALE) (4 FEMALE)	4320	35 AIRCRAFT COMMUNICATION WORDS
ACTUAL SPEECH UNDER STRESS	AMUSEMENT PARK ROLLER-COASTER HELICOPTER COCKPIT RECORDINGS (G-FORCE, LOMBARD EFFECT, NOISE, FEAR, ANXIETY)	9 SPEAKERS (4 MALE, 3 FEMALE) (2 MALE)	500	35 AIRCRAFT COMMUNICATION WORDS
PSYCHIATRIC ANALYSIS	PATIENT INTERVIEWS: (DEPRESSION, FEAR, ANXIETY, ANGRY)	8 SPEAKERS (6 FEMALE) (2 MALE)	600	CONVERSATIONAL SPEECH: PHRASES & SENTENCES

while the pitch-increased segment and the stressed segment showed one main peak (pitch harmonic) in their spectra. Distinctive differences in TEO profiles and corresponding autocorrelation functions are also shown between these three speech segments [e.g., compare autocorrelation responses for Fig. 7 (a3), (b3), (c3)]. From this evaluation, we can see that the angry speech is more than merely a pitch-increased version of its neutral counterpart, since there are many other factors which make it different from neutral. Further studies are needed to critically compare these factors across multiple speakers. We also note that the examples here are ideal cases, and in reality, there are cross-harmonic terms which make the output of each critical band response very complicated. In addition, the Gabor bandpass filter centered at each critical band will include those harmonics in neighboring critical bands due to the gradual change of filter's frequency response characteristics. However, the waveform analysis here has served to illustrate that under stress, there are measurable changes in the envelope of the autocorrelation of the TEO response, and that these changes are partly due to increases in fundamental frequency under stress, partly due to the variability in the harmonics present under stress, and partly due to nonlinear variations occurred in the airflow in the vocal tract.

III. EVALUATIONS

A. Database

In this study, evaluations for stress classification were conducted using *speech under simulated and actual stress* (SUSAS) [16], [23], [25] database which is now available through LDC. Table II summarizes the main features of SUSAS. Two domains of SUSAS (simulated stress from "talking styles" and actual stress from "amusement park roller-coaster") were utilized for the evaluation. The following subset of SUSAS words were used: "freeze," "help," "mark," "nav," "oh," and "zero." Angry, loud and Lombard styles were used for simulated stress (speakers were requested to speak in that style, and 85 dB SPL pink noise played through headphones was used to simulate Lombard effect). Data for actual stress was selected from the subject motion-fear "actual speech under stress" domain. In the actual domain, a series of controlled speech data collection experiments were performed with speakers riding amusement park roller coaster. Background noise levels and stress levels were monitored during the completion of each ride. Since the TEO is more applicable for voiced sounds than for unvoiced sounds, only high-energy voiced sections (i.e., vowels, diphthongs, liquids, glides, nasals) were automatically extracted from the word utterances. All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8 kHz. A baseline five-state HMM-based stress classifier with continuous distributions, each with two Gaussian mixtures, was employed for the evaluations.

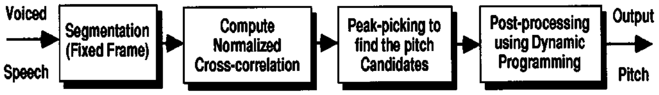

Voiced
Speech
Segmentation (Fixed Frame)
Compute Normalized Cross-correlation
Peak-picking to find the pitch Candidates
Post-processing using Dynamic Programming
Output
Pitch

Fig. 8. Pitch tracking.

B. Traditional Features

Since all three proposed features are based on nonlinear excitation information, it was determined that it would be useful to compare their performance to the traditional pitch feature and the MFCC [13] feature. The pitch feature is obtained using the pitch tracking method proposed in [48] (flow diagram shown in Fig. 8). MFCCs have been widely used for speech recognition

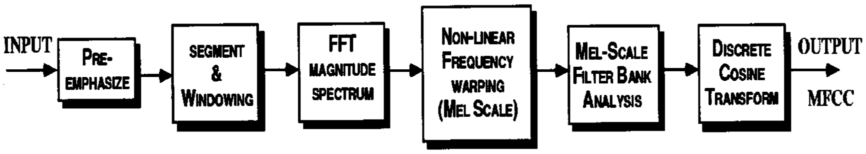

INPUT
PRE-EMPHASIZE
SEGMENT & WINDOWING
FFT MAGNITUDE SPECTRUM
NON-LINEAR FREQUENCY WARPING (MEL SCALE)
MEL-SCALE FILTER BANK ANALYSIS
DISCRETE COSINE TRANSFORM
OUTPUT
MFCC

Fig. 9. MFCC extraction.

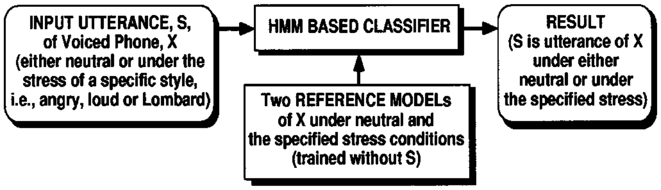

INPUT UTTERANCE, S, of Voiced Phone, X (either neutral or under the stress of a specific style, i.e., angry, loud or Lombard)
HMM BASED CLASSIFIER
RESULT (S is utterance of X under either neutral or under the specified stress)
Two REFERENCE MODELs of X under neutral and the specified stress conditions (trained without S)

Fig. 10. Evaluation flowchart of text-dependent pairwise stress classification.

due to their effectiveness in representing the spectral variations of speech. Fig. 9 shows the extraction procedure of the MFCC feature. Pitch and MFCC have also been used previously for stress classification evaluations [24], [55]. Therefore, these two features represent a good basis of comparison for the new proposed features.

C. Stress Classification Results

To determine which features are better for stress classification, we performed three different evaluations. First, text-dependent pairwise stress classification was evaluated to pre-select good features from the proposed TEO features, and MFCC and pitch. Based on results from the first evaluation, we selected the top three features and conducted a second evaluation for text-independent pairwise stress classification. Finally, a text-independent multi-style stress classification evaluation was performed for the same three features used in the second evaluation.

1) Text-Dependent Pairwise Stress Classification: As the first step, the task was constrained to be a text-dependent pairwise stress classification. We trained an HMM model for the voiced portion of each word using 18 tokens from nine speakers for each stress style, from the SUSAS simulated stressed speech domain. One neutral HMM model per voiced portion of each word was trained using 18 neutral tokens; and 90 neutral tokens per word were used for pairwise testing between neutral and stress style trained HMMs. Since only 18 stressed tokens per word for each style are available, a round-robin method (i.e., for each of 18 tokens, we use the remaining 17 tokens for training, and test on this token) was employed for training and scoring. A total of 648 tokens were used for open test evaluation. For actual speech under stress, we used seven speakers producing 20 tokens of "freeze," nine tokens of "help," 16 tokens of "mark," 16 tokens of "nav," 15 tokens of "oh," and 18 tokens of "zero" for neutral and actual stressed conditions. A total of 188 tokens were used for open test evaluations. Since the speech data from the actual stress domain contains increased levels of background noise, a previously formulated single-channel speech enhancement method was first applied as a preprocessing phase [18] for all feature extraction methods. Informal listening evaluations suggest that the enhanced speech sounds much cleaner than the original, but a small level of perceived background noise is still present. Round-robin training and scoring were employed for both neutral and actual data. Fig. 10 shows the diagram of the stress classification evaluation procedure for this evaluation.

The results of the first evaluation, text-dependent pairwise classification, are shown in Fig. 11 . For simulated stressed speech, the results show that the TEO-FM-Var feature can classify neutral speech from their stress counterparts well (rates are in the range: 65.0%–82.2%), but it is not as successful in classifying stressed speech from neutral (rates are in the range: 41.6%–48.2%). The TEO-Auto-Env feature is very consistent for stress classification across different stress styles (rates fall in the range: 73.9%–85.2%); while the TEO-CB-Auto-Env feature keeps the consistency of TEO-Auto-Env but improves the performance by +13.5% in terms of average classification accuracy (rates range from 87.4% to 98.2%). The two traditional features, pitch information and MFCC have better average classification accuracy than the TEO-FM-Var and TEO-Auto-Env features. However, they seem to have difficulty in differentiating neutral speech and speech with Lombard effect, and thus are less consistent across different stress styles than the TEO-Auto-Env and TEO-CB-Auto-Env features.

For speech from the SUSAS actual stress domain, since the stress level of speech from roller-coaster rides is far more severe, stress classification rates were generally higher. The results for the three nonlinear TEO-based features performed better than under simulated stress, with the TEO-CB-Auto-Env feature performing best. The result here, as seen in the simulated case, is that the TEO-CB-Auto-Env feature performed substantially better than the traditional MFCC and pitch features. These results suggest the consistency of the TEO features from simulated to actual speech under stress domains. Furthermore, human interaction (manual pitch correction) is needed to improve the pitch estimation accuracy from traditional algorithms for actual stressed speech, thus making automatic stressed speech classification difficult.

During the extraction of the TEO-FM-Var and TEO-Auto-Env features, pitch information is utilized. For convenience, a simple absolute magnitude difference function (AMDF) method was used. Because of its simplicity, this method results in lower accuracy than other more sophisticated pitch-tracking algorithms. Therefore, the relatively lower classification accuracy by these two features could have been caused by less accurate pitch estimation. As we observed, however, even the sophisticated pitch-tracking algorithm as shown in Fig. 8 cannot give an accurate pitch estimation when speech is produced under stressful conditions. It is reasonable to try a new feature which does not depend on the accuracy of pitch estimation. This partly explains why we proposed the TEO-CB-Auto-Env feature.

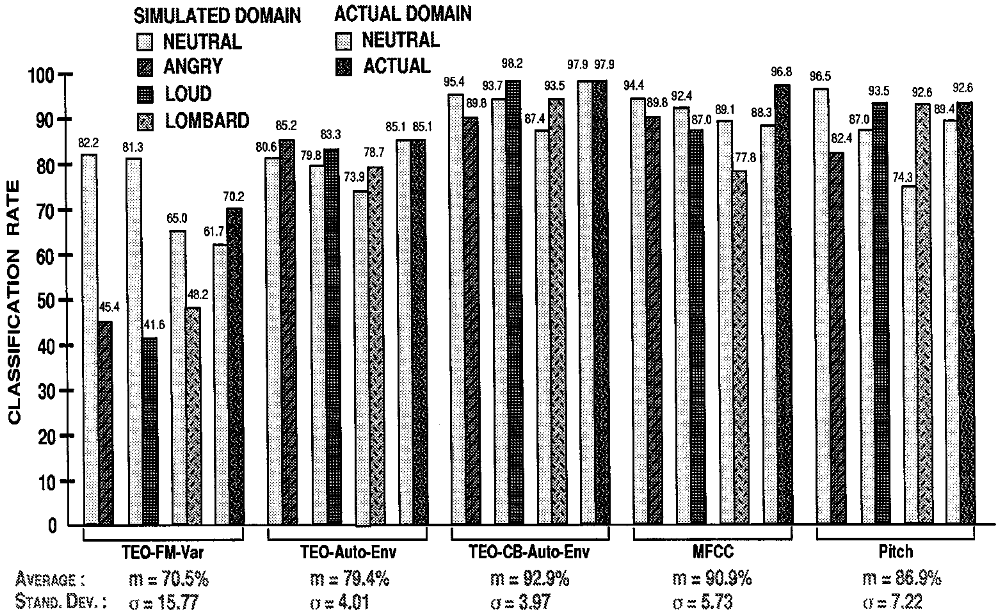


Fig. 11. Text-dependent pairwise stress classification results using SUSAS database (in-vocabulary test).

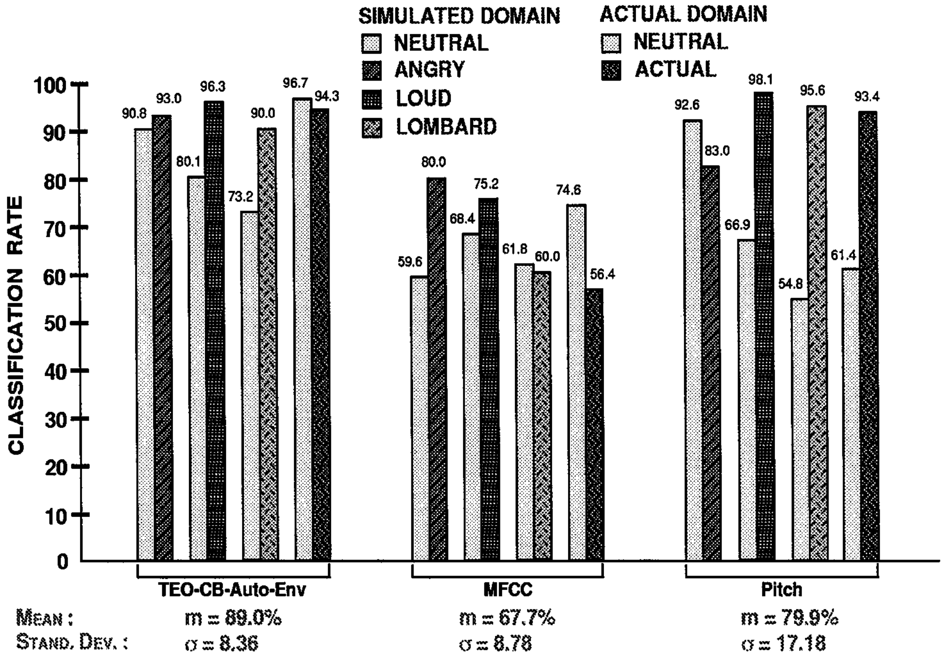


Fig. 12. Text-independent pairwise stress classification results using SUSAS database (out-of-vocabulary test).

2) Text-Independent Pairwise Stress Classification: In the second evaluation, we selected the top three features, which are the TEO-CB-Auto-Env, MFCC, and Pitch, based on their performance in the first evaluation, and conducted a text-independent pairwise classification. The purpose here is to verify whether these features are dependent on text or phoneme information when performing stress classification. For this purpose, only one HMM model for each stress style (i.e., angry, loud, Lombard, and actual) was trained from all tokens available for that stress style; that is, 108 training tokens for angry, loud or Lombard HMM model, and 94 training tokens for actual stress model. Two neutral models, one for the simulated stress domain trained from 108 tokens and one for the actual stress domain trained from 94 tokens, were used. For simulated stress domain, a set of 270 voiced tokens other than those used for training were extracted automatically for test from the SUSAS database for each stress style; for actual stress, a set 140 out-of-vocabulary voiced tokens were extracted automatically for test from the SUSAS actual stress domain. The neutral test set for both simulated and actual stress domain consists of 272 out-of-vocabulary voiced tokens extracted from the SUSAS database.

The results, shown in Fig. 12, indicate that the same three features have slight-to-measurably lower classification accuracy for out-of-vocabulary test tokens than those in-vocabulary test tokens (results shown in Fig. 11).

It is expected that the MFCC feature would have the largest performance decrease (average loss in classification rate: from 90.9% to 67.7%) because it is dependent on vocal tract spectral structure and mainly designed for speech recognition. and thus relies on text sequence information. The pitch information, in general, can classify stressed speech from neutral very well, but does not do as well in classifying neutral speech from stressed. This is could be due to the lack in the pitch-tracking algorithm's ability to provide accurate pitch estimation. In this test, we did not perform hand correction for pitch estimation results of actual stressed speech (as was performed in the first evaluation). Although the performance of the TEO-CB-Auto-Env feature is reduced, the decrease is the smallest. Its average classification rate only decreases by 3.9%; while the average classification rate of pitch decreases by 7.0%, and MFCC by 23.2%. Moreover, the TEO-CB-Auto-Env feature still remains the most consistent across different stress styles compared to the other two features [standard deviation: TEO-CB-Auto-Env (8.36), MFCC (8.78), Pitch (17.18)]. If we examine the performance decrease of the TEO-CB-Auto-Env feature for each stress versus neutral pair, we can see that the major decrease occurs for the simulated domain, especially for the two pairs, neutral versus loud and neutral versus Lombard. As we know, simulated speech under stress is not as easily identified as actual speech under stress and it is likely that some acoustic confusion or overlap between different stress styles exist. Also we should note that many more test tokens were used for the second evaluation. It is reasonable to conclude that the results here are more reliable statistically compared with those shown in Fig. 11, and that these performance values would be realized in real voice communication systems where stress classification is to be employed.

3) Text-Independent Multistyle Stress Classification: After conducting the text-dependent and text-independent pairwise stress classification evaluations, we considered a more ambitious set of evaluations for text-independent multi-style stress classification. The same features (TEO-CB-Env, MFCC, Pitch) as in the second evaluation were used. The goal of this evaluation is first to find out how accurate these features are in detecting neutral versus stressed speech, and further, to see how well they can classify stressed speech into different stress styles. We performed our evaluation on the SUSAS simulated domain. The reason for leaving the actual stress domain out is that actual stress represents an extreme stressed condition (collected while speakers were riding roller-coasters) and can be more easily singled out. The same four HMM models (neutral, angry, loud, Lombard) and vocabulary-test sets as used in the second evaluation were employed.

Results are shown in Tables III–V. In each table, we first report correct neutral and stress detection rates [part (a) in each table]. For this part, the three stress models (angry, loud, Lombard) were grouped together for an overall decision of "stress." Therefore, if a neutral test token is submitted, correct detection occurs only if the neutral model is selected [e.g., 70.6% of neutral test tokens detected as neutral for the TEO-CB-Auto-Env feature (Table V)]. For a stressed token, if any of the three models are selected, then we say the token was correctly identified as being under stress [e.g., for the TEO-CB-Auto-Env feature (Table V), 96.3% of angry test tokens detected as stressed speech, where either angry, loud or Lombard model picked over neutral]. In part (b) of each table, we report the individual stress classification rates, assuming we achieved correct detection (e.g., for the TEO-CB-Auto-Env feature (Table V), after correctly detecting angry speech as being stress 96.3% of time, we see that the angry model was actually selected 65% of the time, with loud and Lombard selected 29.2% and 5.8% of the time). Finally, when the neutral model is selected for neutral test tokens, we have correct detection. When neutral tokens are detected as stress, we have detection error, and we therefore wish to identify which stress models are selected in error. The stress classification rates reported for neutral test speech for part (b) in each table reflect the error classification rates, e.g., for those neutral tokens incorrectly detected 29.4% of the time as stress for the TEO-CB-Auto-Env feature (Table V), the majority were selected as Lombard (68.8%), while a smaller percentage for the other two possible stress styles.

It is clear that the MFCC feature (Table III) does not perform as well as either pitch (Table IV) or the TEO-CB-Auto-Env feature (Table V) for text-independent multistyle stress classification. The performance of TEO-CB-Auto-Env and pitch does vary, with the TEO-CB-Auto-Env feature performing better for detection of neutral from stressed, while pitch performs better for detection of stressed from neutral. This suggests that a combination of pitch and TEO based features could improve stress

TABLE III
TEXT-INDEPENDENT MULTISTYLE STRESS CLASSIFICATION RESULTS USING MFCC

Test Speech Style	(a) Correct Detection Rate (%)		(b) Distribution of **STRESS** Detection Rate across 3 Stress Styles (%)		
	Neutral	Stressed	Angry	Loud	Lombard
Neutral	46.3		45.2	14.4	40.4
Angry		85.9	58.6	18.5	22.8
Loud		89.3	48.9	20.7	19.6
Lombard		71.9	45.9	18.5	35.1

TABLE IV
TEXT-INDEPENDENT MULTISTYLE STRESS CLASSIFICATION RESULTS USING PITCH

Test Speech Style	(a) Correct Detection Rate (%)		(b) Distribution of **STRESS** Detection Rate across 3 Stress Styles (%)		
	Neutral	Stressed	Angry	Loud	Lombard
Neutral	52.2		0.8	16.9	82.3
Angry		96.7	44.4	34.1	21.5
Loud		100	16.7	53.3	30.0
Lombard		95.6	0.8	9.7	89.5

TABLE V
TEXT-INDEPENDENT MULTISTYLE STRESS CLASSIFICATION RESULTS USING TEO-CB-AUTO-ENV

Test Speech Style	(a) Correct Detection Rate (%)		(b) Distribution of **STRESS** Detection Rate across 3 Stress Styles (%)		
	Neutral	Stressed	Angry	Loud	Lombard
Neutral	70.6		3.8	27.5	68.8
Angry		96.3	65.0	29.2	5.8
Loud		97.0	34.0	51.9	14.1
Lombard		91.5	22.3	32.8	44.9

TABLE VI
EVALUATION RESULTS FOR INFLUENCE OF SPEECH RECOGNITION ON STRESS CLASSIFICATION

Test Speech Style	Correct Rate (%) for Both Speech Recognition and Stress Classification			Correct Rate (%) for Only Stress Classification		
	MFCC	Pitch	TEO-CB-Auto-Env	MFCC	Pitch	TEO-CB-Auto-Env
Neutral	84.0	8.5	0	88.3	92.6	97.9
Actual Stress	70.2	22.3	0	96.8	89.4	97.9

classification performance. If we examine the distribution of the stress detection rate across the three stress styles, the most confusing pairs are (angry, loud) and (neutral, Lombard). As we commented earlier, all training and test data for stressed speech are from the simulated domain of SUSAS. Some speakers might be better at simulating speech on a particular emotion or style. Even though every speaker simulated each stressed style, there is still overlap between different styles acoustically such as angry and loud (e.g., sometimes people show their anger by speaking louder).

We further conducted a final evaluation in the actual domain to determine how the speech recognition aspect of these three features contributes to stress classification performance. MFCC is currently one of the most successful features for speech recognition; pitch can be combined with other features for speech recognition; while TEO-CB-Auto-Env was proposed mainly to characterize the nonlinear airflow excitation during speech production and therefore should not be as good at speech recognition. To verify this, we used 12 text-dependent HMM models (six for neutral, six for stressed) trained during the first evaluation (see Section III-C1). While training tokens were also used as test tokens, the round-robin method was employed to ensure open-set testing. During testing, each token was submitted to all 12 HMM models. Based on the resulting HMM scores, two rates were computed, that is, the correct rate for both speech recognition and stress classification, and the correct rate for only stress classification. Table VI shows these results, which indicate what we might expect, that pitch and TEO-CB-Auto-Env are not effective for combined speech recognition and stress classification, but that TEO-CB-Auto-Env outperforms the others for stress classification. Combined with results from the first and second evaluations, we can say that the performance of MFCC for stress classification heavily depends on its ability to first achieve reliable speech recognition performance. The performance of pitch for stress classification can at times benefit from its speech recognition ability, but only in a limited sense. The TEO-CB-Auto-Env feature, however, captures factors independent of text information during speech production for effective stress classification. This final evaluation therefore suggests that the TEO-CB-Auto-Env should be used for stress classification, and thereby provide useful information which could be employed in an MFCC feature based speech recognition system to improve speech recognition under stress.

IV. CONCLUSIONS

In this study, we proposed the following three new TEO-based nonlinear features: TEO-FM-Var, TEO-Auto-Env, and TEO-CB-Auto-Env, for stress classification. TEO-based features strive to reflect what is believed to be the variation in nonlinear airflow excitation during speech production under stress. Evaluation results using the SUSAS database for speech under stress showed that the TEO-FM-Var and TEO-Auto-Env features are not as effective for stress classification because they depend on pitch estimation accuracy. The traditional MFCC feature heavily depends on its speech recognition ability, and thus works well for text-dependent pairwise stress classification but degrades rapidly for text-independent stress classification. Pitch can be a useful feature for stress classification, but lacks consistency and reliability partly because user input correction is needed to repair its estimation accuracy for speech under high degrees of stress. The TEO-CB-Auto-Env feature, however, is the best feature evaluated for stress classification in terms of both accuracy and reliability. Furthermore, evaluation results showed that this new feature does not depend on text information, but is capable of capturing those factors, which we believe, are nonlinear airflow excitation changes which cause listeners to perceive stressed speech sounding different from neutral.

ACKNOWLEDGMENT

The authors would like to thank Dr. H. Steeneken (Chair of NATO IST/TG-01, formerly RSG.10) of TNO Human Factors Research Institute, The Netherlands, for providing the SUSC-0 stressed speech corpus. They would also like to thank the reviewers for their helpful comments, and Dr. B.-H. Juang for his generous time in overseeing the review process of this paper.

REFERENCES

[1] *Speech Commun., Special Issue on Speech Under Stress*, vol. 20, Nov. 1996.

[2] *Proc. Int. Conf. Acoustics, Speech, Signal Processing '99: Special Session on Speech Under Stress*, vol. 4, Mar. 1999, pp. 2079–2098.

[3] C. Baber, B. Mellor, R. Graham, J. M. Noyes, and C. Tunley, "Workload and the use of automatic speech recognition: The effects of time and resource demands," *Speech Commun.*, vol. 20, no. –12, pp. 37–54, Nov. 1996.

[4] E. G. Bard, C. Sotillo, A. H. Anderson, H. S. Thompson, and M. M. Taylor, "The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment," *Speech Commun.*, vol. 20, pp. 71–84, Nov. 1996.

[5] D. A. Berry, H. Herzel, I. R. Titze, and K. Krischer, "Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions," *J. Acoust. Soc. Amer.*, vol. 6, pp. 3595–3604, 1995.

[6] Z. S. Bond and T. J. Moore, "A note on loud and lombard speech," in *Int. Conf. Speech Language Processing '90*, 1990, pp. 969–972.

[7] S. E. Bou-Ghazale and J. H. L. Hansen, "Generating stressed speech from neutral speech using a modified CELP vocoder," *Speech Commun.*, vol. 20, pp. 93–110, Nov. 1996.

[8] ——, "Stress perturbation of neutral speech for synthesis based on hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 201–216, May 1998.

[9] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," *J. Acoust. Soc. Amer.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[10] A. Castellanos, J. M. Benedi, and F. Casacuberta, "An analysis of general acoustic-phonetic features for spanish speech produced with lombard effect," *Speech Commun.*, vol. 20, pp. 23–36, Nov. 1996.

[11] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 433–439, 1988.

[12] A. J. Chorin and J. E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, 2nd ed. Berlin, Germany: Springer-Verlag, 1990.

[13] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.

[14] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2000.

[15] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Berlin, Germany: Springer-Verlag, 1983.

[16] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, 1988.

[17] J. H. L. Hansen and O. N. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," in *Proc. Int. Conf. Speech Language Processing '90*, Kobe, Japan, 1990, pp. 1125–1128.

[18] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, Apr. 1991.

[19] J. H. L. Hansen, "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '93*, 1993, pp. 95–98.

[20] J. H. L. Hansen, "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 598–614, July 1994.

[21] J. H. L. Hansen and D. A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments," *Speech Commun.*, vol. 16, no. 4, pp. 403–406, 1995.

[22] 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 Trans. Speech Audio Processing*, vol. 3, pp. 407–415, Sept. 1995.

[23] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, pp. 151–173, Nov. 1996.

[24] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network based classification of speech under stress," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 307–313, July 1996.

[25] J. H. L. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. EUROSPEECH '97*, vol. 4, Rhodes, Greece, Sept. 1997, http://www.ldc.upenn.edu; http://cslu.colorado.edu/rspl/stress.html, pp. 1743–1746.

[26] 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 Trans. Biomed. Eng.*, vol. 45, pp. 300–313, Mar. 1998.

[27] J. H. L. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. A. Vloeberghs, I. Trancoso, and P. Verlinde, *The Impact of Speech Under Stress on Military Speech Technology*: NATO Research & Technology Organization RTO-TR-10, Mar. 2000, vol. AC/323(IST)TP/5 IST/TG-01.

[28] J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 1, pp. 510–524, 1993.

[29] ——, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as lombard reflex," *Speech Commun.*, vol. 20, no. 1–2, pp. 13–22, 1996.

[30] J. F. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view," *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, 1983.

[31] ——, "On a simple algorithm to calculate the "energy" of a signal," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '90*, 1990, pp. 381–384.

[32] ——, "On Teager's energy algorithm, its generalization to continuous signals," in *Proc. 4th IEEE Digital Signal Processing Workshop*. New Paltz, NY, Sept. 1990.

[33] ——, "Some useful properties of Teager's energy operator," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '93*, vol. 3, 1993, pp. 149–152.

[34] R. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '87*, 1987, pp. 705–708.

[35] E. Lombard, "Le Signe de l'Elevation de la Voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[36] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Processing*, vol. 41, pp. 1532–1550, Apr. 1993.

[37] ——, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, pp. 3025–3051, Oct. 1993.

[38] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, 1995.

[39] I. R. Murray, C. Baber, and A. South, "Toward a definition and working model of stress and its effects on speech," *Speech Commun.*, vol. 20, pp. 3–12, Nov. 1996.

[40] I. R. Murray, J. L. Arnott, and E. A. Rohwer, "Emotional stress in synthetic speech: Progress and future directions," *Speech Commun.*, vol. 20, pp. 85–92, Nov. 1996.

[41] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '87*, 1987, pp. 713–716.

[42] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[43] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, "Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions," *Speech Commun.*, vol. 20, pp. 111–130, Nov. 1996.

[44] B. Scharf, "Critical bands," in *Foundations of Modern Auditory Theory*, J. V. Tobias, Ed. New York: Academic, 1970, vol. 1, pp. 157–202.

[45] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '88*, 1988, pp. 331–334.

[46] ——, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '89*, 1989, pp. 675–678.

[47] B. H. Story and I. R. Titze, "Parameterization of vocal tract area functions by emperical orthogonal modes," *J. Phonetics*, vol. 26, pp. 223–260, 1998.

[48] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 497–518, 1995.

[49] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-28, no. 5, pp. 599–601, Oct. 1980.

[50] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract," *Speech Science: Recent Advances*, pp. 73–109, 1983.

[51] ——, "Evidence for nonlinear production mechanisms in the vocal tract," in *Speech Production and Speech Modeling*. Norwell, MA: Kluwer, 1989, vol. 55, pp. 241–261.

[52] T. J. Thomas, "A finite element model of fluid flow in the vocal tract," *Comput. Speech Lang.*, vol. 1, pp. 131–151, 1986.

[53] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *J. Acoust. Soc. Amer.*, vol. 52, no. 4, pp. 1238–1250, 1972.

[54] J. Whitmore and S. Fisher, "Speech during sustained operations," *Speech Commun.*, vol. 20, pp. 55–70, Nov. 1996.

[55] B. D. Womack and J. H. L. Hansen, "Classification of speech under stress using target driven features," *Speech Commun.*, vol. 20, pp. 131–150, Nov. 1996.

[56] W. A. Yost, *Fundamentals of Hearing*, 3rd ed. New York: Academic, 1994, pp. 153–167.

[57] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Classification of speech under stress based on features from the nonlinear teager energy operator," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing '98*, Seattle, WA, May 12–15, 1998, pp. 549–552.

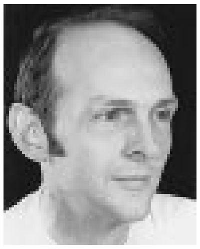
Guojun Zhou (M'00) received the B.S. degree from Southeast University, Nanjing, China, in 1988, the M.S. degree from Tsinghua University, Beijing, China in 1993, and the Ph.D. degree from Duke University, Durham, NC, in 1999, all in electrical engineering.

From 1988 to 1990, he was with Xinhe Electronic Audio Equipment, Inc., Guangdong Province, China, as a System Circuit Design Engineer. He continued to work in speech recognition and audio/video system circuit design while he was in Singapore from 1994 to 1996. In August 1996, he joined the Robust Speech Processing Laboratory (RSPL), Duke University. From August 1996 to December 1999, he worked at RSPL on robust issues in speech recognition as well as in speech enhancement and speaker verification. During the Summer of 1998, he was a visiting researcher at Nuance Communications, Inc., working on speech recognition. During 1999, he continued to work at RSPL, but was focused on problems in large vocabulary continuous speech recognition at the Center for Spoken Language Research (CSLR), University of Colorado, Boulder. He joined Intel Corporation, Hillsboro, OR, in December 1999. He is currently working on natural language understanding and dialogue systems at Intel's Architecture Labs. His interests include speech processing, digital signal processing, natural language processing, and dialogue system design. He is also interested in building real-world application systems using speech recognition and natural language understanding technologies. He has published several papers in IEEE ICASSP, ICSLP and other speech-related conferences.

John H. L. Hansen (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ in 1982, and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

He is presently an Associate Professor with the Departments of Speech, Language, and Hearing Sciences, and Electrical and Computer Engineering, University of Colorado, Boulder. In 1988, he established and has since directed the Robust Speech Processing Laboratory (RSPL). He serves as Associate Director for the Center for Spoken Language Research (CSLR), and directs the research activities of RSPL at CSLR. He was a Faculty Member at the Departments of Electrical and Biomedical Engineering, Duke University, Durham, NC, for 11 years before joining the University of Colorado in 1999. He has served as a technical consultant to industry and the U.S. Government, including AT&T Bell Laboratories, IBM, Sparta, Signalscape, ASEC, BAE Systems, VeriVoice, and DoD in the areas of voice communications, wireless telephone, robust speech recognition, and forensic speech/speaker analysis. He is the author of more than 100 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (Piscataway, NJ: IEEE Press, 2000), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000, ISBN: 92–837–1027–4). His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech pathology, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust recognition and training methods for topic spotting in noise, accent, stress, and Lombard effect, and speech feature enhancement in hands-free environments for human-computer interaction.

Dr. Hansen was an Invited Tutorial Speaker for the IEEE International Conference on Acoustics, Speech, and Signal Processing '95 and the 1995 ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal. He has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01: Research Study Group on Speech Processing, 1996–1998), Chairman for the IEEE Communications and Signal Processing Society of North Carolina (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), Tutorials Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing '96, Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), and Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000). He also served as guest editor of the October 1994 Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was the recipient of a Whitaker Foundation Biomedical Research Award, a National Science Foundation's Research Initiation Award, and has been named a Lilly Foundation Teaching Fellow for "contributions to the advancement of engineering education." He will be serving as General Chair for the International Conference on Spoken Language Processing in October 2002.



James F. Kaiser (S'50–A'52–SM'70–F'73) was born in Piqua, OH, in 1929. He received the electrical engineering degree from the University of Cincinnati, Cincinnati, OH, in 1952 and the S.M. and Sc.D. degrees in 1954 and 1959, respectively, from the Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

Currently, he is a Visiting Professor with the Department of Electrical and Computer Engineering, Duke University, Durham, NC. He was formerly a Distinguished Member of Technical Staff with the Speech and Image Processing Research Division, Bell Communications Research, Inc., which he joined in 1984 at its formation. Prior to that, he was a Distinguished Member of Technical Staff, Bell Laboratories, Murray Hill, NJ, for 25 years, where he worked in the areas of speech processing, system simulation, digital signal processing, computer graphics, and computer-aided design. He is the author of more than 65 research papers and the coauthor and editor of eight books in the signal processing and automatic control areas.

Dr. Kaiser is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He received the Technical Achievement Award of the IEEE Signal Processing Society (SPS) in 1978, its Meritorious Service Award in 1979, its Society Award in 1982, and the IEEE Centennial Medal in 1984. In 1970, he was presented with the Distinguished Engineering Alumnus Award by the College of Engineering, University of Cincinnati, and, in 1980, the Eta Kappa Nu Award of Merit, also from the University of Cincinnati. He has served in a number of positions in both the SPS and the IEEE Circuits and Systems Society. He is a Registered Professional Engineer in Massachusetts, a member of the Acoustical Society of America, AAAS, EURASIP, and SIAM.