

## Chapter 5

# Stress Classification and Detection

### 5.1 Introduction

In this chapter we discuss commercial stress classification and assessment systems, and summarize some recent research studies on voice stress classification. While a number of voice stress classification systems are available on the market today, the scientific basis on which they are built is not well understood. A systematic objective evaluation would be a good first step to determine their usefulness. The assessment of stress in a speaker's voice is an important issue in monitoring speaker state, especially in military environments where high physical task stress or fatigue induced stress can occur, or in forensic applications for law enforcement or security applications.

Today, commercial based speech recognition systems can achieve more than 95 % recognition rates for large vocabularies in restricted paradigms. However, their performance degrades greatly in stressful situations, such as a pilot in an emergency situation, a military operator under a heavy workload, or a medical team that is exhausted due to lack of rest. Similar losses in performance occur for other recognizers such as automatic speaker recognition systems. It is suggested that algorithms which are capable of detecting and classifying stress could be beneficial in improving automatic recognition system performance under stressful conditions. Furthermore, there are other applications for stress detection and classification. For example, a stress detector could be used to detect the physical and/or mental state of a pilot and that detection could put special procedures in place such as the rerouting of communications, the redirection of action, or the initiation of an emergency plan. To be able to detect and classify stress, it is necessary to understand the effects of stress on acoustical features. Thus far, differences in acoustical features between neutral and stressed speech brought on by a variety of emotions and the Lombard effect<sup>1</sup> have been the focus of a number of research investigations [1, 13, 54, 60, 85, 160, 146]. We have seen in Chapter 4 that many speech production features change when a person is speaking under stressful conditions.

This chapter is organized as follows. In Sec. 5.2, traditional methods for stress classification are discussed. Most commercial based systems fall into this area. In Sec. 5.3, methods proposed in the past few years using neural network concepts are presented. These methods employ speech features derived from a linear speech model, and typically features which are cepstral-based [73, 162]. Next, Sec. 5.4 considers more recent classification experiments which use linear based speech features and optimum Bayesian detection theory. These experiments were conducted on linear features such as duration, intensity, pitch, glottal source, and vocal tract spectrum for stress classification [167]. Next, it was shown in a previous study [26, 25] that the TEO-based (Teager Energy Operator) nonlinear speech feature has the potential to improve stress classification performance. In a recent USAF study, (verified by Hansen at RSPL, Univ.

---

<sup>1</sup>The Lombard effect occurs when a speaker, in either a conscious or sub-conscious manner, modifies his speech production in order to increase his communication ability in a noisy environment.

Colorado/Duke Univ.) several TEO-based nonlinear features were found to be very effective for both stress classification and stress assessment (Zhou, Hansen, Kaiser [165, 166, 168]). Therefore, in Sec. 5.5 several new nonlinear based features are summarized which have shown promise in both classification and assessment of speaker stress. When possible, results using speech data discussed in Chapter 3 (i.e., SUSAS and SUSC-0) are presented.

## 5.2 Traditional Methods for Commercial Voice Stress Analysis

Traditional methods for detecting the stress in a speaker's voice have evolved from the early interest by military and law enforcement agencies in the detection of deception. Military interrogators and law enforcement interviewers were anxious for a capability that would aid in determining whether a subject was making a truthful statement or lying. Such a voice stress analysis (VSA) capability would be extremely valuable in gathering information that could impact on the outcome of a battle or a trial. The reason for this interest lies not only in its information value but in the non-intrusive and efficient way that VSA equipment promises to obtain information that otherwise could not be obtained so quickly and conveniently. While military interest has continued to increase, law enforcement agency interest has grown immensely which has created a commercial market for VSA equipment.

An additional growing civilian application of VSA technology and equipment that is creating a market is pre-employment interviews. As a result, several commercial VSA systems are available in both hardware and software. The systems range in price from approximately \$100 to \$10 000. Many of the vendors offer training courses and some of these courses are intensive and require as much as a week or more to complete.

The basic assumption underlying the operation of these commercial systems is the belief that involuntary detectable changes in the voice characteristics of a speaker take place when the speaker is stressed during an act of deception. The systems in general detect inaudible and involuntary frequency modulations in the 8–12 Hz region. The frequency modulations, whose strength and pattern are inversely related to the degree of stress in a speaker, are believed to be the result of physiological tremor or microtremor (Lippoid, 1971 [97]) that accompanies voluntary contraction of the striated muscles involved in vocalization. The systems generally use filtering and discrimination techniques and display the result on a chart recorder. The determination of the degree of stress contained within a selected voice sample is made through the visual examination of the chart by a trained examiner [29]. The examiner looks for characteristic shapes related to amplitude, cyclic change, leading edge slope, and square waveform shapes called blocking.

## 5.3 Neural Networks with Linear Speech Model-based Features

### 5.3.1 Cepstral-based Features

In a previous study conducted at RSPL [73, 162, 161], a neural network based classification algorithm was considered for stress classification using cepstral-based features which have traditionally been employed for recognition. Five cepstral feature sets were investigated, which included Mel  $C_i$  (C-Mel), delta Mel  $DC_i$  (DC-Mel), delta-delta Mel  $D2C_i$  (D2C-Mel), auto-correlation Mel  $AC_i$  (AC-Mel), and cross-correlation Mel  $XC_{i,j}$  (XC-Mel) cepstral parameters. The first three cepstral features ( $C_i$ ,  $DC_i$ , and  $D2C_i$ ) had been shown to improve speech recognition performance in the presence of noise and Lombard effect [74]. The  $AC_i$  and  $XC_{i,j}$  features were new in that they provide a measure of the correlation between Mel-cepstral coefficients.

The Mel-cepstral (C-Mel) parameters are well known as features that represent the spectral variations of the acoustic speech signal. It is suggested that such parameters are useful for stress

classification since vocal tract and spectral structure vary due to stress. The C-Mel parameters are able to reflect these energy shifts.

The DC-Mel and D2C-Mel parameters provide a measure of the “velocity” and “acceleration” of movement of the C-Mel parameters. These features are calculated by performing polynomial fitting of the C-Mel parameters and taking the derivative of the polynomial itself. This may differ from other studies which use a first and second order difference method to estimate  $DC_i$  and  $D2C_i$  respectively. It appears that the reason delta parameters are more robust to stress variations is due to their reduced variance across stress conditions. This trait suggests that while these features are more useful for recognition, they may be less applicable to stress classification.

It is suggested that the two more recently derived feature representations (AC-Mel and XC-Mel) could be more successful in representing variations due to stress. The AC-Mel features are calculated as follows,

$$AC_i^{(\ell)}(k) = \sum_{m=k}^{m=k+L} [C_i(m) * C_i(m + \ell)] / \sup_k AC_i^{(\ell)}(k), \quad (5.1)$$

where  $k$  is the frame number,  $L$  is the correlation window length,  $\ell$  the number of correlation lags, and  $i$  the Mel coefficient index. When  $\ell = 0$ ,  $AC_i$  models the relative power between frequency bands. For  $\ell > 0$ ,  $AC_i$  models spectral slope and changes in the frame to frame correlation variation due to stress. The XC-Mel coefficients are similar to the AC-Mel coefficients except that the cross-correlation is found from one Mel coefficient  $C_i$  to another  $C_j$  across frames,

$$XC_{i,j}^{(\ell)}(k) = \sum_{m=k}^{m=k+L} [C_i(m) * C_j(m + \ell)] / \sup_k XC_{i,j}^{(\ell)}(k). \quad (5.2)$$

The XC-Mel parameters  $XC_{i,j}$  provide a quantitative measure of the relative change of broad versus fine spectral structure in energy bands. Since the correlation window length ( $L = 7$ ) and correlation lags ( $\ell = 2$ ) are fixed in this study, the correlation terms are a measure of how correlated adjacent frames are over a 72 ms window (24 ms/frame and 8 ms skip rate). It is apparent that both AC-Mel and XC-Mel parameters provide a measure of correlation and relative change in spectral band energies over an extended window frame. Feature analysis suggests that the AC-Mel parameters have similar properties to the XC-Mel parameters. In addition, the AC-Mel parameters can be directly compared with other selected feature sets since they are based on a single coefficient index  $i$ . Therefore, AC-Mel parameters appear to be a better choice for stress classification than the XC-Mel parameters.

### 5.3.2 Neural Network Classifier

A neural network stress classifier was formulated using mono-partition features (i.e., a single phone class partition). Each partition of speech features was propagated through two hidden layers of the neural network to an output layer that estimates the stress probability scores. The neural network training method employed was the cascade correlation back-propagation network using the extended delta-bar-delta learning rule [106]. This method was selected due to its flexibility, and because it is capable of forming the complex contoured hypersurface decision boundaries needed for the stress classification problem. Fig. 5.1 shows the structure of this classification system.

### 5.3.3 Neural Network Stress Classification Evaluations with Cepstral Features

The neural network stress classification algorithm was evaluated using a collection of features from frame- and word-level features. Both fine and broad stress classes were evaluated. The fine (i.e., ungrouped) stress classes were the 11 stress conditions from the simulated portion of

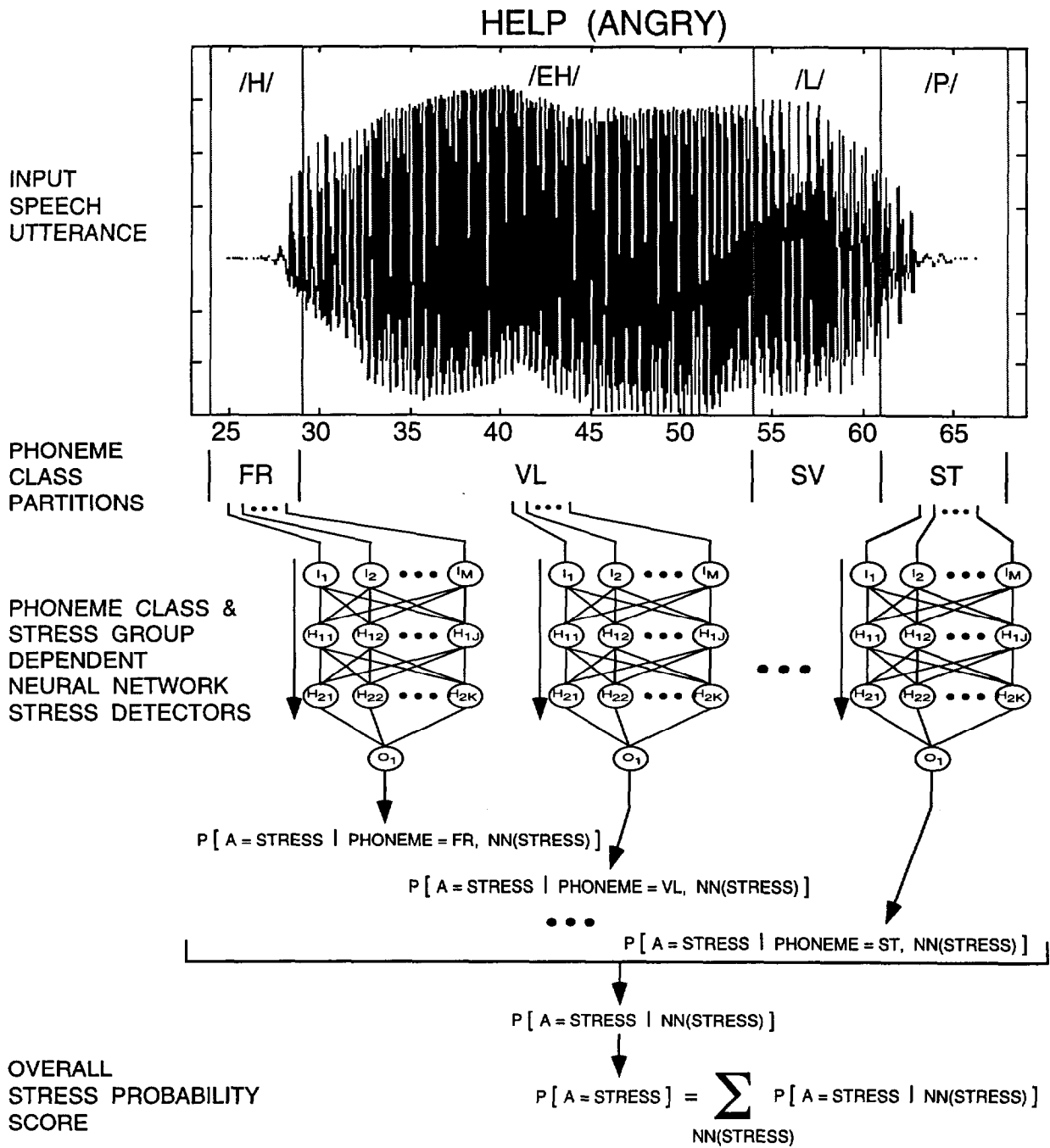


Figure 5.1: Stress classification method using phoneme based neural networks, with output scores combined for an overall stress score.

SUSAS. Ungrouped stress class neural network classifier performance is summarized in Table 5.1 using closed 35-word test sets. Classification rates ranged from 11–17 % for the 35 word test set, which is greater than chance (i.e., 9 %). It is clear that for some stress conditions reasonable classification performance is attained. A similar test using a 5 word vocabulary produced higher average classification rates (i.e., 32–67,%). It is also useful to point out that these results were for a 1 of  $N$  stress class test, versus a pairwise neutral versus particular stress style. Pairwise classification results are always significantly higher, since some stress conditions for SUSAS are similar (e.g., angry and loud, clear and Lombard, etc.).

STRESS CLASSIFICATION PERFORMANCE				
Single Speaker, 35 Words, Stress Ungrouped CLOSED VOCABULARY TEST SET				
STRESS CLASS	CLASSIFICATION RATE (%)			
	$C_i$	$DC_i$	$D2C_i$	$AC_i$
<i>Angry</i>	6.20	0.00	19.49	4.96
<i>Clear</i>	12.50	0.00	4.27	7.32
<i>Cond50/70</i>	42.47	59.76	56.08	14.61
<i>Fast</i>	7.26	1.65	44.53	68.10
<i>Lombard</i>	1.64	0.00	8.53	2.54
<i>Loud</i>	12.40	3.73	3.15	1.69
<i>Neutral</i>	22.31	0.00	2.61	1.72
<i>Question</i>	14.05	0.00	5.15	4.76
<i>Slow</i>	19.01	4.76	27.56	4.31
<i>Soft</i>	16.53	33.09	0.00	2.38
MEAN	15.44	10.30	17.14	11.24
STD. DEV.	11.33	20.12	19.62	20.35

Table 5.1: MPSC Performance for Ungrouped Closed 35 Word Test

### 5.3.4 Neural Network Stress Classification with Target Driven Features

Further classification studies have expanded on these neural network approaches using target driven features [162]. In this method, a wide selection of features are automatically extracted including articulatory measures, pitch, phone duration, spectral based, etc. Next, the most effective feature subset for each targeted stress condition is determined during a training phase. During classification, only those targeted features needed for a neural network stress classifier under test are employed. This allows the classifier to use the most discriminating features for classification of each stress style. A second study proposed an approach which combines stress classification and speech recognition functions into one algorithm [163]. This was accomplished by generalizing the one-dimensional hidden Markov model to an N-channel Hidden Markov Model (N-Channel HMM). Here, each stressed speech production style under consideration is allocated a dimension in the N-Channel HMM to model each perceptually induced stress condition. It is shown that this formulation better integrates perceptually induced stress effects for stress independent recognition and classification. This is due to the sub-phoneme (state level) stress classification that is implicitly performed by the algorithm. The N-channel stress HMM method was compared to a previously established 1-channel stress dependent isolated word recognition system yielding a **73.8 %** reduction in error rate.

## 5.4 Bayesian Stress Classification with Linear Speech Features

While neural network classifiers have shown promise, there is clearly a difference in performance based on the feature set used for stress classification. It has been shown that there are observable

differences in duration, intensity, pitch, glottal source information, and formant locations between neutral and stressed speech [54]. Therefore, it is worthwhile to evaluate their performance for stress classification, or stress detection. Here two terms, classification and detection, can be used interchangeably since only pairwise classification is considered. The methods employed for classification here are Bayesian hypothesis testing approach and distance measure.

#### 5.4.1 Feature Description

For linear feature based stress classification, only vowel sections are extracted from the simulated domain of the SUSAS database for evaluation. The length of each vowel in msec is used as the duration feature. The intensity feature is defined as,

$$Intens = \sqrt{\frac{1}{K} \sum_{i=1}^K s^2(i)} \quad (5.3)$$

where  $s(i)$  ( $i = 1, \dots, K$ ) represents the  $K$  individual samples in the vowel. Pitch, glottal source information, and formant locations are extracted on a frame basis with frame length being 32 ms and an overlap length between adjacent frames of 16 ms. The modified simple inverse filter tracking (MSIFT) algorithm [7] is employed to extract pitch frequencies from vowel speech waveforms. Spectral slope was used as the glottal source feature. It is difficult to obtain the glottal spectral slope from the raw vowel speech waveform due to the coupling effect between the sub-glottal structure and forward portion of the vocal tract. To avoid this effect, only data obtained during closed vocal fold periods was used. This unfortunately limits the usable data. Also, it is difficult to accurately locate the boundaries between vocal fold closing and opening periods. As an approximation, a frame based log average amplitude FFT was computed versus log frequency for each vowel section.

Next, a straight line is used to approximate its envelope, and the line's slope is considered as the glottal spectral slope. Only the first two formants are used for the evaluation since the remaining formants do not show much differences between neutral and stressed speech [54]. The HTK `xwaves` function "formant" was employed to extract formant locations for all vowels in the SUSAS database.

#### 5.4.2 Bayesian Hypothesis Testing versus Distance Measure Testing

A stress classifier is similar to a Bayesian hypothesis testing system. It has two hypotheses, that is,  $H_0$  and  $H_1$ . Under  $H_0$ , the speech is neutral; while under  $H_1$ , the speech is stressed. Given an input speech feature vector,  $\mathbf{x}$ , ( $\mathbf{x} = x_1, \dots, x_M$ ;  $M$  is the vector length), the following two conditional probability densities are calculated,  $p(\mathbf{x}|H_0)$  and  $p(\mathbf{x}|H_1)$ . The likelihood ratio,  $\lambda$ , is then defined as,

$$\lambda = \frac{p(\mathbf{x}|H_1)}{p(\mathbf{x}|H_0)}. \quad (5.4)$$

The decision of whether the input speech is neutral or stressed is made by comparing the likelihood or log likelihood ratio with a pre-defined threshold,  $\beta$ . If it is bigger than  $\beta$ , the input speech is labeled as stressed; otherwise it is classified as neutral. The value of  $\beta$  depends on what criterion is used for detection. In a stress classification system, a criterion should be selected so that the two important probabilities, the false acceptance rate (FAR) and the false rejection rate (FRR), should be as low as possible. Obviously, it is not possible to minimize both FAR and FRR, and hence, a compromise must be made between FA and FR. For some systems, the requirement for one probability is more important than the other. For a stress classification system, however, we are only interested in the overall accuracy and have no preference for either FAR or FRR. Therefore, the value of  $\beta$  corresponding to equal error (FAR = FRR) rate (EER)

is selected. In the experiments performed here, the values of FAR and FRR were calculated as the ratio of the number of falsely accepted vowels to the total number of vowels, and the ratio of the number falsely rejected vowels to the total number of vowels, respectively. By changing the threshold value, the value of  $\beta$  corresponding to EER can be found.

It is also possible to detect stressed speech from neutral by using a distance measure with prior trained feature distributions. Given an input speech feature vector,  $\mathbf{x} = x_1, x_2, \dots, x_M$ ;  $M$  is the vector length, two values, the distance between  $\mathbf{x}$  and the neutral speech feature distribution,  $d_n$ , and the distance between  $\mathbf{x}$  and the stressed speech feature distribution,  $d_s$ , are computed as follows,

$$d_n = \frac{|\hat{\mu} - \mu_n|}{\hat{\sigma}\sigma_n}, \quad (5.5)$$

$$d_s = \frac{|\hat{\mu} - \mu_s|}{\hat{\sigma}\sigma_s}, \quad (5.6)$$

where  $\mu_n, \sigma_n, \mu_s, \sigma_s$  are means and standard deviations for the neutral and stressed speech features, which are obtained from training data;  $\hat{\mu}$  and  $\hat{\sigma}$  are the sampled estimated mean and standard deviation of the components of the input vector,  $\mathbf{x}$ .

This distance measure reflects how close the input test speech feature vector is to the feature distributions of neutral and stressed speech data. If  $d_n$  is smaller than  $d_s$ , the input vector  $\mathbf{x}$  is labeled as neutral, otherwise, it is assigned as stressed. The distance scores can also be used to quantify the degree of stress content in the test data.

### 5.4.3 Linear Feature Based Evaluations

A 33 word vocabulary under neutral, angry, loud, and Lombard effect speaking styles from the simulated domain of the SUSAS database was employed for evaluations. From all identified vowels, duration, intensity, pitch, glottal spectral slope, and formant locations were extracted. For each feature, all extracted data was used to estimate the density function (*pdf*) of the feature distribution (Fig. 5.2 shows two examples, one for a Gaussian distribution for pitch and a second for Gamma distribution for glottal spectral slope for vowels under loud speaking style) to obtain ROC curves for the Bayesian hypothesis testing approach. To find average test results, the data was divided for each feature into 10 equal size sets. For each of the 10 sets, we test with one set and train with the other 9 to calculate the average EER threshold for the Bayesian hypothesis testing approach, and the mean and variance of the feature distribution for the distance measure approach.

Several testing feature vector lengths (1, 5, 10) were used to obtain ROC curves and error rates. Two of the many ROC curves obtained are shown in Fig. 5.3 for stress classification between neutral and loud for mean pitch and glottal spectral slope.

Table 5.2 shows an error analysis for all five feature domains using both the Bayesian hypothesis testing approach and distance measure approach. The pairwise errors for each detection technique and feature are given for the detection of three stress conditions (angry, loud, or Lombard) from neutral speech.

Based on Table 5.2, the following observations can be made: (1) that pitch is the best feature for stress classification among the five features, (2) error rates generally decrease as feature vector length increases, (3) performance differences exist between different stress styles, and (4) mean vowel formant locations are not suitable for stress classification. The results in this section have therefore established stress classification performance using linear speech production based features with two types of optimum detection methods. Further discussion of the evaluations presented here can be found in [63].

Detection Method	Vector Length	Feature	Speaking Style of Submitted Test Speech					
			Neutral	Angry	Neutral	Loud	Neutral	Lombard
Bayesian Hypothesis Testing	1	Duration	45.13	45.38	38.21	38.72	40.77	40.26
		Intensity	40.26	37.44	34.87	32.82	40.77	39.49
		Pitch	18.95	18.57	11.94	11.63	24.08	24.18
		Glottal	33.33	36.78	41.38	41.72	42.76	42.07
		Formant 1	42.60	41.80	46.43	45.10	46.84	46.90
		Formant 2	51.48	50.88	58.20	54.51	52.98	49.88
	5	Duration	36.36	38.96	33.77	35.06	40.26	40.26
		Intensity	24.68	22.08	27.27	22.08	38.96	35.06
		Pitch	15.17	14.31	10.34	10.00	21.90	22.07
		Glottal	25.45	21.82	30.91	34.55	30.91	36.36
		Formant 1	40.60	40.30	46.12	45.82	47.91	46.87
		Formant 2	53.88	49.85	58.51	56.12	54.78	50.90
	10	Duration	41.03	35.90	38.46	35.90	38.46	46.15
		Intensity	23.08	17.95	28.21	17.95	35.90	35.90
		Pitch	12.76	11.72	7.24	8.28	20.69	19.31
Glottal		25.00	17.86	35.71	35.71	28.57	32.14	
Formant 1		38.79	40.91	43.03	44.24	47.58	47.88	
Formant 2		55.76	47.58	59.39	57.27	53.33	55.15	
Distance Measure	5	Duration	48.05	49.35	29.87	36.36	32.47	42.86
		Intensity	41.56	27.27	35.06	22.08	40.26	35.06
		Pitch	15.34	15.00	12.41	7.07	23.10	19.48
		Glottal	34.55	18.18	38.89	35.19	38.89	33.33
		Formant 1	43.58	37.76	44.63	45.97	45.82	46.72
		Formant 2	53.28	49.85	41.49	74.78	36.87	74.93
	10	Duration	43.59	53.85	28.21	41.03	30.77	46.15
		Intensity	30.77	33.33	25.64	23.08	41.03	33.33
		Pitch	14.48	12.76	12.07	4.83	21.38	17.59
		Glottal	35.71	17.86	44.44	25.93	44.44	25.93
		Formant 1	41.82	39.39	43.64	41.82	45.45	46.06
		Formant 2	54.55	49.09	40.00	74.85	38.48	76.06

Table 5.2: Detection Error Rates for Multiple Speaking Styles.

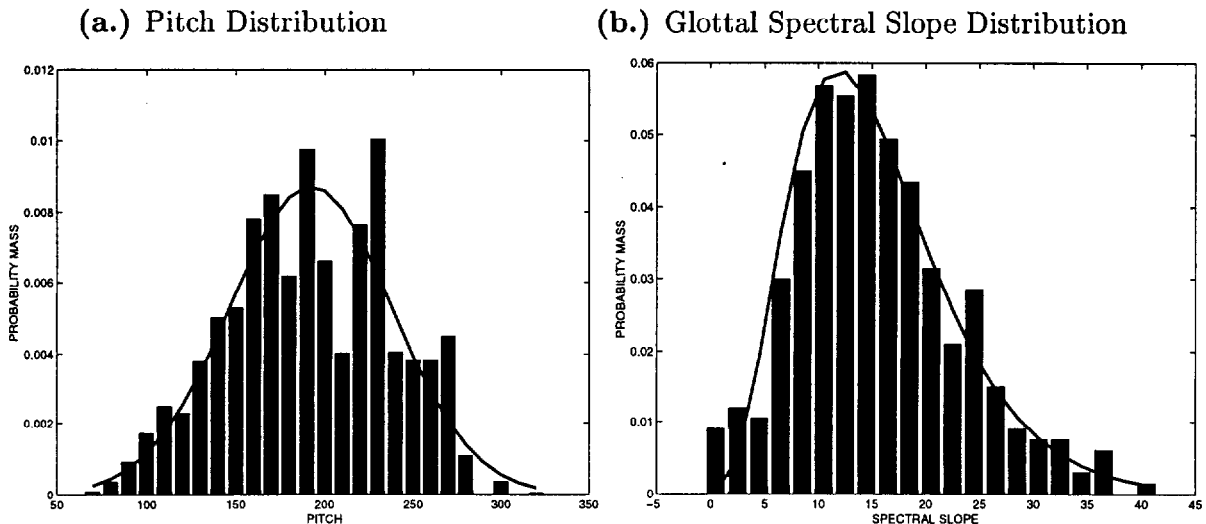


Figure 5.2: (a.) A conditional Gaussian *pdf* is used to approximate the pitch distribution of vowels under loud speaking style:  $N(\mu, \sigma^2 | X \geq 0)$  with  $(\mu = 192 \text{ Hz}, \sigma^2 = 2094)$ . (b.) A conditional Gamma *pdf* is used to approximate the distribution of glottal spectral slope for vowels under loud speaking style:  $\Gamma(\alpha, \beta)$  with  $(\alpha = 4.2329, \beta = 3.6612)$ .

## 5.5 Stress Classification Using Nonlinear Speech Features

In this section, recently proposed approaches to stress classification that employ on Teager Energy Operator based processing are considered. Four features are discussed, followed by evaluations using stressed speech data from SUSAS. These features have been shown to be more effective than many linear based features such as pitch and spectral structure (as reflected by MFCC parameters). Further details can be found in studies by Zhou, Hansen, and Kaiser [165, 166, 167, 168]. While some of the discussions in this section is more research oriented, the features discussed have potential as important processing tools in monitoring and assessing personnel in high stress military voice communication settings.

### 5.5.1 Teager Energy Operator

According to studies by Teager [152, 153, 154], the assumption that airflow propagates as a plane wave in the vocal tract may not hold, since the flow is actually separated and concomitant vortices are distributed throughout the vocal tract. Teager also suggested that hearing could be viewed as the process of detecting the energy. Based on the theory of the oscillation pattern of a simple spring-mass system, Teager developed an energy operator to measure the energy for simple sinusoids which can be believed as useful elements for speech. The simple and elegant form of the operator was introduced by Kaiser [86, 87] as,

$$\begin{aligned} \Psi_c[x(t)] &= \left( \frac{d}{dt}x(t) \right)^2 - x(t) \left( \frac{d^2}{dt^2}x(t) \right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \end{aligned} \quad (5.7)$$

where  $\Psi[\cdot]$  is the Teager Energy Operator (TEO), and  $x(t)$  is a single-frequency component of the continuous speech signal. Kaiser [86, 88] derived the operator for discrete-time signals from its continuous form  $\Psi_c[x(t)]$ , as,

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1), \quad (5.8)$$

where  $x(n)$  is the sampled speech signal.

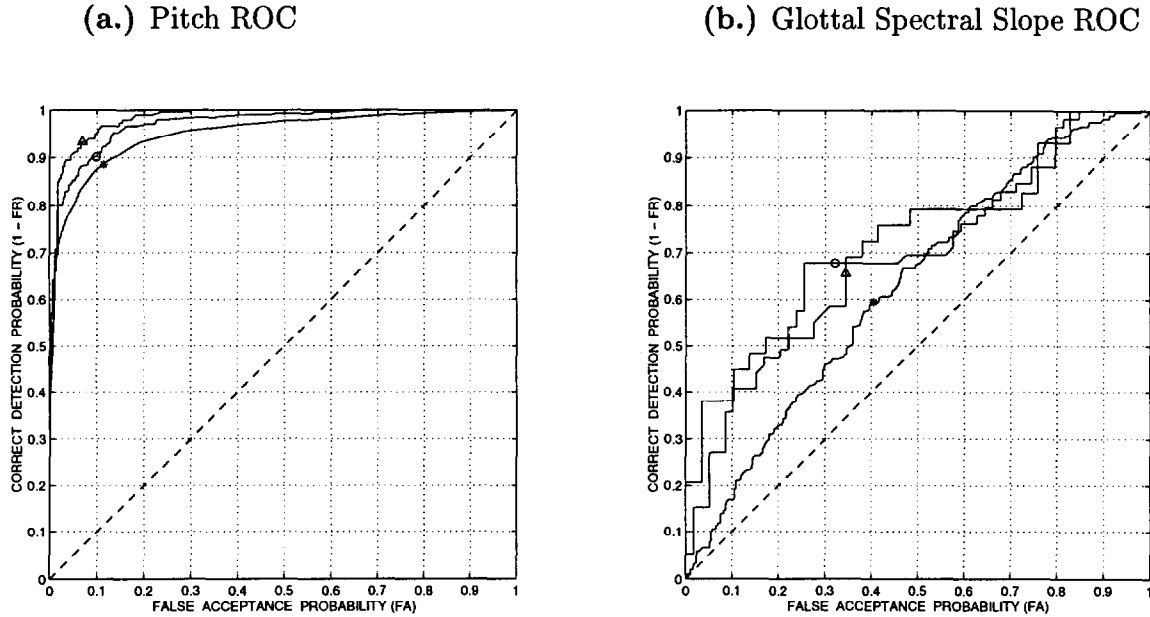


Figure 5.3: ROC detection curves for “loud” versus neutral speech using: (a.) pitch (line with \*: input vector length is 1,  $EER(*) = 11.47\%$ ; line with  $\circ$ : input vector length is 5,  $EER(\circ) = 9.86\%$ ; line with  $\Delta$ : input vector length is 10,  $EER(\Delta) = 6.80\%$ ). (b.) spectral slope (line with \*: input vector length is 1,  $EER(*) = 40.51\%$ ; line with  $\circ$ : input vector length is 5,  $EER(\circ) = 32.22\%$ ; line with  $\Delta$ : input vector length is 10,  $EER(\Delta) = 34.48\%$ ).

The TEO is typically applied to a bandpass filtered speech signal, since its intent is to reflect the energy of the nonlinear energy flow within the vocal tract for a single resonant frequency. Under this condition, the resulting TEO profile can be used to decompose a speech signal into its AM and FM components within a certain frequency band via,

$$f(n) \approx \frac{1}{2\pi T} \arccos \left( 1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]} \right), \quad (5.9)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{\left[ 1 - \left( 1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]} \right)^2 \right]}}, \quad (5.10)$$

where  $y(n) = x(n) - x(n-1)$ ,  $\Psi[\cdot]$  is the TEO operator as shown in Eq. 5.8,  $f(n)$  is the FM component at sample  $n$ , and  $a(n)$  is the AM component at sample  $n$  [103, 104]. On the basis of this work, Maragos, Kaiser, and Quatieri [104] proposed a nonlinear model which represents the speech signal  $s(t)$  as,

$$s(t) = \sum_{m=1}^M r_m(t), \quad (5.11)$$

where

$$r_m(t) = a_m(t) \cos \left( 2\pi(f_{cm}t + \int_0^t q_m(\tau)d\tau) + \theta \right) \quad (5.12)$$

is a combined AM and FM structure representing a speech resonance at the  $m$ th 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  $m$ th formant.

Although the TEO is formulated for single-frequency signals or signals with a single resonant frequency, previous studies have shown that the TEO energy of a multi-frequency signal is not only different from that of single-frequency signal but also reflects interactions between different

frequency components [165, 166]. This characteristic extends the use of TEO to speech signals filtered with wide bandwidth band-pass filters (BPF).

### 5.5.2 TEO-FM-Var: FM Variation

Previous studies have shown that vowels spoken under stress generally have more instantaneous pitch variations than vowels spoken under neutral conditions. This suggests that features which represent fine excitation variations, would be useful for stress classification. To some extent, these variations are believed to be due to the effects of modulations. According to work by Maragos, Kaiser, and Quatieri [103, 104], the TEO is a nonlinear differential operator that can detect modulations in the speech signal and further decompose the signal into its AM and FM components. It is not difficult to understand that the AM-FM decomposition of a speech signal over a wide frequency band will not provide correct estimation of the real modulations. AM-FM signal analysis requires a carrier frequency which must be higher than the modulating frequencies within the signal. Because of interest in the fine excitation variations, the raw input speech is filtered using a Gabor bandpass filter (BPF) centered at the median fundamental frequency,  $F_0$ , with a root mean square (RMS) bandwidth of  $F_0/2$  based on the TEO profile of the entire input. The  $F_0$  is estimated using the average magnitude difference function (AMDF). After the Gabor BPF, TEO analysis is performed and the resulting profile is used to separate the input speech signal into its AM and FM components using Eq. 5.9 and Eq. 5.10. A flow diagram for extracting the TEO-FM-Var feature is shown in Fig. 5.4.

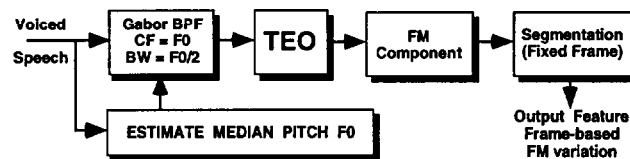


Figure 5.4: TEO-FM-Var Feature Extraction

### 5.5.3 TEO-Pitch: TEO based Pitch

Unlike the feature presented in 5.5.2, or the feature to be presented in 5.5.4, the TEO-Pitch feature is a direct estimate of the pitch itself. Since it is difficult for currently available techniques to correctly detect the pitch of speech under stress, especially under extreme stress, TEO processing is first applied to the raw vowel waveform. As will be explained in Sec. 5.5.4, the TEO profile has the same periodicity as pitch. Furthermore, experiments determined that it generally showed better periodicity than the raw stressed speech partly because of the square effect of TEO. Since we found that pitch usually falls within the extreme range of 50 Hz to 750 Hz (female speech from actual high stress can have pitch as high as 700 Hz), the TEO profile is bandpass filtered over (50:750 Hz) [165]. As shown in Fig. 5.5, after the BPF and segmentation, a normalized cross-correlation function (NCCF) and dynamic programming [151] is applied to detect the pitch structure. Here the waveform is first down-sampled, and candidate peaks in the NCCF are selected. Subsequently, the peaks are fine-tuned by using the NCCF of the original waveform (before down-sampling). The candidate frame-based pitch periods are determined by the average distance of two neighboring peaks within that frame. Finally, dynamic programming is employed to decide the pitch period of each frame.

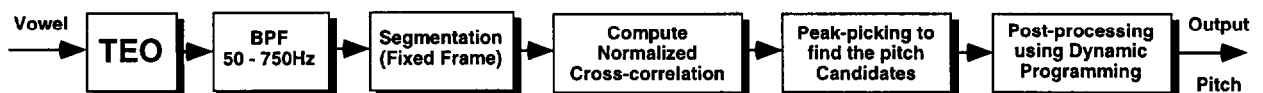


Figure 5.5: TEO-Pitch Feature Extraction

### 5.5.4 TEO-Auto-Env: Normalized TEO Autocorrelation Envelope Area

The third feature, named TEO-Auto-Env, also reflects the instantaneous excitation variations of speech. A flow diagram is shown in Fig. 5.6. This feature is based on the idea that the presence of stress may affect modulation patterns within the frequency bands of speech differently. It is obtained by passing the raw input speech through a filterbank consisting of 4 bandpass filters (BPF).

Each BPF output stream is processed by a TEO to estimate each profile. Our experiments show that the TEO profile of an AM-FM signal has the same periodicity as the modulating signals. Furthermore, the TEO profile periodicity is generally dominated by an amplitude modulating signal frequency. This explains why the TEO profile reflects the same periodicity as the pitch profile since both are affected by amplitude modulations. Therefore, we obtain a feature representing the fine pitch variation by analyzing the TEO autocorrelation envelope.

If we consider the fact that pitch is a slow-changing variable, we can bandpass-filter each TEO output stream through a Gabor BPF centered at the median fundamental frequency ( $F_0$ ), with the 3 dB bandwidth being roughly  $F_0/2$ .  $F_0$  is obtained using the AMDF based pitch detection method on the TEO profile instead of the raw speech. Subsequently, each Gabor-filtered TEO stream is segmented into frames. In order to have equivalent averaging effects, the frame length is set to 4 times the median pitch period. Furthermore, the normalized autocorrelation function is computed for each frame. If there is no pitch variation within a frame, its normalized autocorrelation function should be a damped sinusoidal response with a straight line envelope. The area under the ideal envelope (without pitch variation) should be the same for each frame for a specified vowel, that is,  $N/2$ , where  $N$  is the frame length. 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 not be  $N/2$ . By computing the area under the normalized autocorrelation envelope and normalizing by  $N/2$ , it is possible to obtain 4 normalized TEO autocorrelation envelope area parameters for each time frame (i.e., one for each frequency band). This 4 parameter vector represents the TEO-Auto-Env feature per frame.

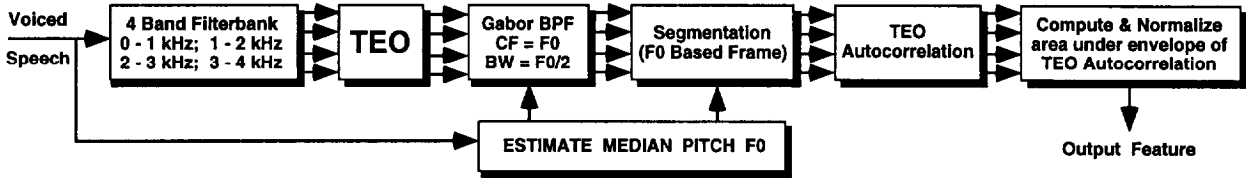


Figure 5.6: TEO-Auto-Env Feature Extraction

### 5.5.5 TEO-CB-Auto-Env: Critical Band Based TEO Autocorrelation Envelope

Empirically, the human auditory system is assumed to be a filtering process which partitions the entire audible frequency range into many critical bands [164]. Based on this assumption, the last proposed nonlinear feature employs a critical band based filterbank to filter the speech signal followed by TEO processing (see Fig. 5.7). Each filter in the filterbank is a Gabor bandpass filter, with the 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 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

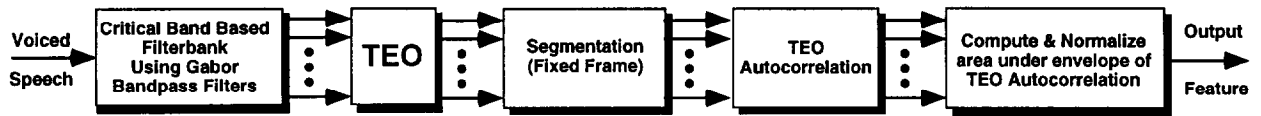


Figure 5.7: TEO-CB-Auto-Env Feature Extraction

band), where  $M$  is the total number of critical bands. This is the TEO-CB-Auto-Env feature vector per frame. Fig. 5.7 shows the entire feature extraction procedure. Since each critical band possesses a much narrower bandwidth than the 1 kHz bandwidth used for BPFs in the TEO-Auto-Env feature, post Gabor bandpass filtering centered at median  $F_0$  is not needed in TEO-CB-Auto-Env extraction. This makes the new feature independent of the accuracy of median  $F_0$  estimation.

In practice, all TEO profiles are segmented into many frames and all autocorrelation functions are normalized. As a result, the constant autocorrelation function is represented as a decaying straight line from  $(0, 1)$  to  $(N, 0)$ , where  $N$  is the frame length. Those variations caused by harmonic distribution as well as by modulations from stress are expected to be reflected by the change in the TEO autocorrelation envelopes.

### 5.5.6 Evaluations

Evaluations were conducted using the SUSAS, *Speech Under Simulated and Actual Stress* database. SUSAS consists of five domains spoken under a wide range of stresses and emotions. In experiments discussed here, angry, loud and Lombard effect styles were used from SUSAS for simulated stress (speakers were requested to speak in that style; 85 dB SPL pink noise played through headphones was used to simulate the Lombard effect). Data for SUSAS actual stress was selected from the subject motion-fear domain. In the actual domain, a series of controlled speech data collection experiments were performed with speakers riding an amusement park roller coaster.

Since the TEO is more applicable for the voiced sound than for the unvoiced sound, only voiced sections of all word utterances were used for the evaluation. A baseline 5-state HMM-based stress classifier with continuous Gaussian mixture distributions was employed for the evaluations. For the purposes of comparison, the traditional pitch feature tracked by the algorithm proposed in [151] and the MFCC feature [42] were used.

The evaluation results are shown in Fig. 5.8. In general, TEO based features are effective in classifying stressed speech from neutral for both simulated and actual stress situations. Among them, the TEO-Auto-Env feature has very consistent performance across different styles of stress, but the accuracy is not as high as the TEO-CB-Auto-TEO because of fewer frequency band partitions. The TEO-CB-Auto-Env feature with fine frequency partitions, however, provide the most effective and consistent level of stress classification performance compared with MFCC and pitch information.

The evaluations in this section have shown that recently proposed nonlinear based features can be effective in the classification of speech under stress in both simulated and actual stress settings [165, 166, 167]. This assumes that the goal is to detect the presence of stress. In some military or law enforcement settings, it is also necessary to assess the level of stress in an operator's voice. The next section considers both linear and nonlinear based features for the task of stress assessment using actual emergency military voice communications between aircraft pilots from the SUSC-0 stress database.

## 5.6 Stress Assessment

In many military and civilian applications, it is necessary to assess whether or not a speaker is under stress. To evaluate the techniques discussed and their ability to detect real stress,

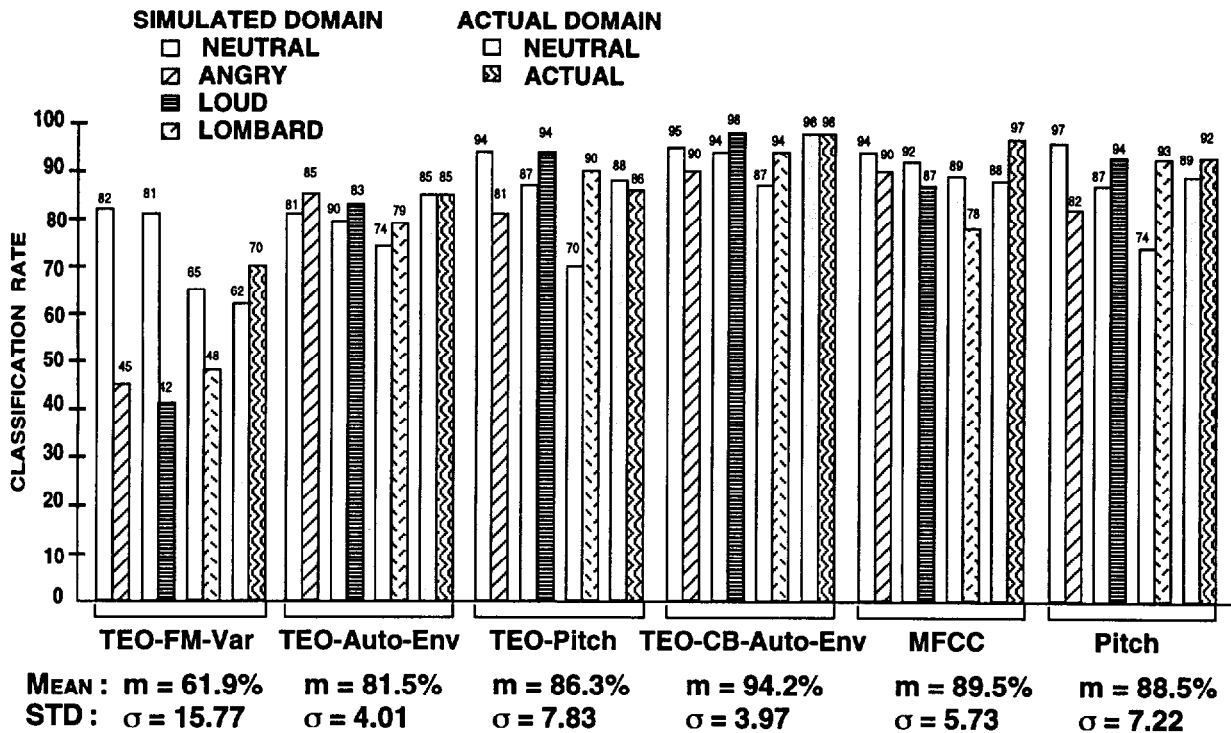


Figure 5.8: Pairwise Stress Classification Results (Mean and standard deviation of overall neutral/stress classification rates are shown; Different speaker groups were used for simulated and actual stress conditions)

the SUSC-O database containing speech of pilots under stress was processed [63]. The SUSC-0 database is from NATO IST-TG01, which consists of actual aircraft pilot communications under emergency situations. Specifically, *Mayday2* domain in SUSC-0 was used. *Mayday2* contains speech data between a pilot and controller collected from the initial ground aircraft system check, through preliminary discovery of engine emergency, until safe resolution of the emergency. The different stress degrees experienced by the pilot are reflected by his speech in *Mayday2*. A second database entitled TORONTO-AIR was also considered. This tape recording consisted of voice communications between a pilot and the controller before a fatal aircraft crashed. Since the pilot was generally unaware that an emergency was taking place until it was too late, there is only mild levels of uncertainty in his voice. Also, this tape is an ‘air traffic control’ (ATC) tape recorded from the ground, so high levels of noise were present. Due to these issues, the results discussed here focus on SUSC-0. Further details of the second database results are in [63]. Twelve (12) sentences from each database were extracted to represent different speaking styles for the assessment evaluation. Table 5.3 shows the 12 sentences from SUSC-0.

A baseline HMM-based stress assessor with continuous Gaussian mixture distributions was used for the evaluation. Two reference HMM models, one representing neutral speech and the other representing stressed speech, were trained. All voiced segments of word “help” under neutral conditions in SUSAS database were used to train the neutral HMM reference model. For the stressed HMM reference model, two different sets were trained, one from the simulated angry, loud, and Lombard stress conditions, and one from that actual stress roller coaster and free fall ride data, respectively. If a speech feature can assess the degree of stress regardless of text, the log likelihood ratio of the unknown speech generated by the stressed HMM model versus the neutral HMM model should be able to indicate whether it is more likely under stress or neutral. Since TEO-based autocorrelation envelope features (TEO-Auto-Env, and TEO-CB-Auto-Env), MFCC, and frame-based pitch information were shown to be very effective for stress classification, they were used to assess the stress for SUSC-0 database. Since both TEO-based

features and pitch information are only useful for voiced speech, the assessment is based on the extracted voiced portions from each utterance. To consider the variations within each utterance, 4 voiced portions per utterance (shown in Table 5.3) are extracted for the assessment.

Sentences from Mayday2 Domain of SUSC-0	
No.	Sentence
1	avionics <b>IIGHt</b> hydr <b>AULic</b> oil pressure <b>IIGHt</b> engine indications <b>ARE</b> ...
2	<b>AND</b> you'er g <b>ONNA</b> declare an em <b>ER</b> gency or am <b>I</b>
3	... checklist <b>OIL</b> pressure malfunction <b>G</b> one-hundred ... cruise altitude st <b>ORe</b> jett ... throttle minimize m <b>OV</b> ement ...
4	roger that <b>OIL</b> indic <b>Ation</b> is n <b>OW</b> z <b>ER</b> O
5	... <b>ALRIGHt</b> newt ... engine fault <b>IIGHt</b> still lit ... hydr <b>AULics</b> are ... total p <b>OUN</b> ds six ...
6	and I'm going there and I'm there I'm des <b>EN</b> ding down to ten gr <b>AN</b> d right I'm n <b>Ot</b> picking up a t <b>A</b> can lock
7	no I'M doing <b>ALRIGHt</b> now and the r <b>AD</b> ial is wh <b>At</b>
8	ok <b>AY</b> give me imm <b>ED</b> iate vectors this is an em <b>ER</b> gency I'm engine <b>OUt</b>
9	g <b>I</b> ve me h <b>EAD</b> ings I <b>NEED</b> headings n <b>OW</b>
10	put the c <b>AB</b> le d <b>OW</b> N p <b>U</b> t the c <b>AB</b> le down
11	I'm h <b>Ot</b> I <b>NEED</b> the c <b>AB</b> le ...
12	m <b>AN</b> I th <b>OU</b> GHt I w <b>AS</b> g <b>ON</b> e

Table 5.3: Sentences from SUSC-0 for Stress Assessment Evaluation. Note that bold uppercase characters represent voiced section which were used for overall stress assessment of that sentence.

The assessment results are shown in Fig. 5.9 for SUSC-0. Here, one score is obtained by finding an average output score across the four extracted voiced sections per sentence. Generally speaking, the recordings begin in a neutral relaxed setting (sentence numbers 1–2), then move into concern while pilot begins to determine the cause of the problem (sentence numbers 3–7). Finally, the pilot determines that the emergency is serious must land the aircraft without power (sentence numbers 8–11). Sentence number 12 indicates his relief after a safe landing.

Both figures ((a) and (b) in Fig. 5.9) show that the general assessment score trend is similar regardless of which anchor stress HMM reference model is used. However, the stress HMM reference model trained from actual SUSAS stress results in larger fluctuations among assessment scores. This may be because that model represents an extreme case of stress. It is noted that SUSC-0 recording can at times have high levels of background noise, so the stress assessment can be affected. We believe that it is the background noise which impacts the stress assessment because both simulated and actual stress HMM reference models produced similar results, while the actual stress HMM reference model was trained from very noisy data.

## 5.7 Stress Assessment and Classification Issues

We have seen that the problem of stress classification is a problem which is becoming increasingly important for military and security in the field of multi-national communications between operators/personnel. Past methods for voice stress analysis have focused on what is believed to be microtremors in the muscles for voice production. More recent methods using digital speech processing have suggested alternative methods which offer the promise of better system integration within speech/speaker recognition or voice communications equipment for military scenarios.

While research and progress have been made in the areas of stress classification and assessment, a number of important research areas require further investigation. Here, we briefly

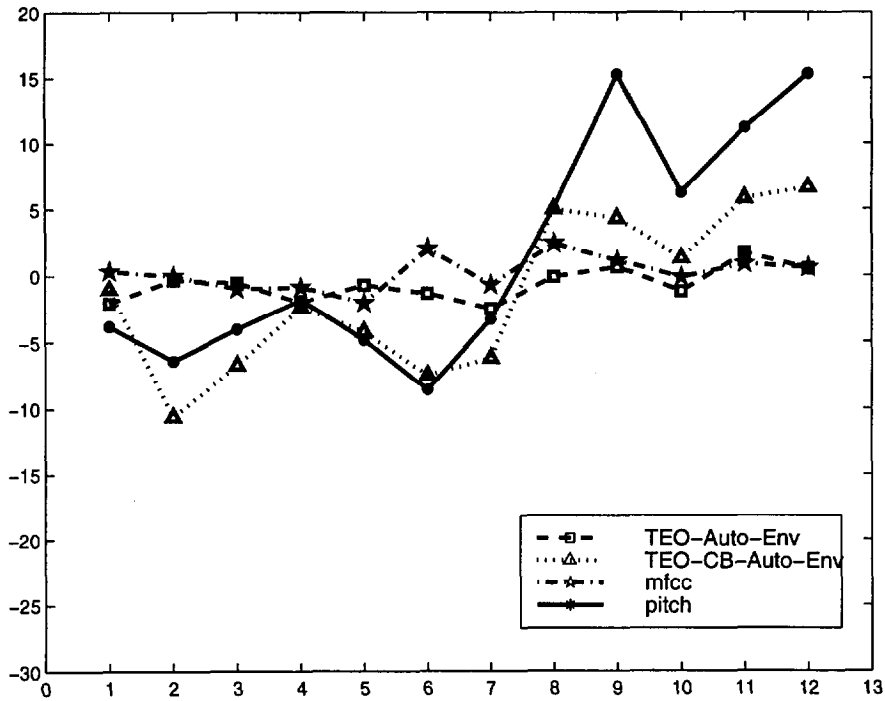
consider four points. First, in order to perform stress classification or assessment, two anchor models are needed (one for neutral and one for stress). These models should be trained using speech obtained from the actual stressful environments in which we wish to assess operators (i.e., aircraft pilot recordings if pilots are to be assessed). The type of stress which is displayed in one setting (aircraft cockpit), may not reflect the same workload conditions an operator may experience in another (army tank operator). Second, further research is needed to assess the consistency of stress assessment/classification for a given speaker and for unseen speakers (i.e., explore the impact of using other training data to assess new speakers). Third, there is clearly a range of emotions and workload factors which all contribute to operator 'stress.' In military scenarios an operator may experience a combination of fear, anxiety, fatigue, etc. at the same time. The ability to classify/assess this mixture of speaker traits is important in determining the stress state of the speaker/operator. Finally, there exists an unknown relationship between how computer based speech systems are able to classify stress and how humans perform stress classification. This issue is important in the collection of future databases so that better stress anchor models can be used with speech technology. From the research conducted here, it is suggested that speakers often vary how they convey stress in their speech, and that several speech features may be needed to capture the subtle differences in how speakers convey their stress state in military voice communications.

## 5.8 Selected References of Interest:

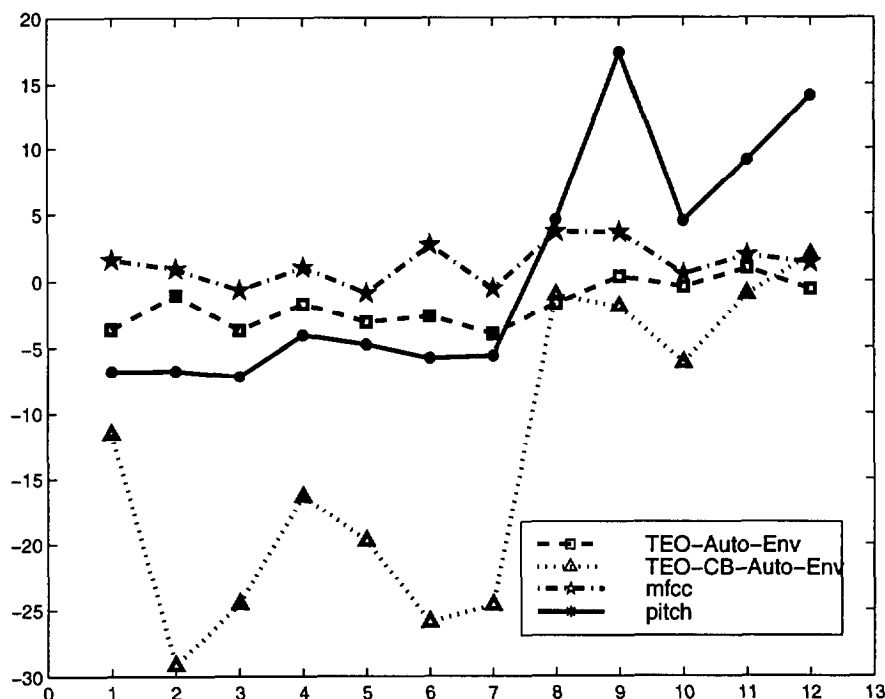
Here, we summarize several references which have considered classification or features related to classification of speech under stress. The reference section at the end of this report contains all references cited in this chapter.

1. P. Benson, "Analysis of the Acoustic Correlates of Stress from an Operational Aviation Emergency," Proc. ESCA-NATO Tutorial and Research Workshop on Speech Under Stress, Lisbon, Portugal, pp. 61-64, 1995.
2. D.A. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions," *Journal of the Acoustical Society of America*, vol.96, no.6, pp. 3392-3400, Dec. 1994.
3. D.A. Cairns, J.H.L. Hansen, "Nonlinear Speech Analysis using the Teager Energy Operator with Application to Speech Classification under Stress," *ICSLP-94: Inter. Conf. on Spoken Lang. Proc.*, vol. II, vol. 3, pp. 1035-1038, Yokohama, Japan, Sept. 1994.
4. V.L. Cestaro, "A Comparison between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy", Tech. Report, DoD Polygraph Inst., Aug. 1995.
5. J.H.L. Hansen, B.D. Womack, "Feature Analysis and Neural Network based Classification of Speech under Stress," *IEEE Trans. Speech and Audio Proc.*, vol. 4, no. 4, pp. 307-313, July 1996.
6. O. Lippold, "Physiological Tremor," *Scientific American*, vol. 224, no. 3, pp. 65-73, Mar. 1971.
7. R. Sarikaya, J.N. Gowdy, "Subband Based Classification of Speech under Stress", *IEEE 1998 ICASSP*, pp. 569-573, 1998.
8. B.J. Stanton, L.H. Jamieson, G.D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," *IEEE 1988 ICASSP*, pp. 331-334, 1988.

9. C.E. Williams, K.N. Stevens, "On Determining the Emotional State of Pilots During Flight: An Exploratory Study," *Aerospace Medicine*, **40** 1369–1372, 1969.
10. B.D. Womack and J.H.L. Hansen, "Classification of Speech under Stress Using Target Driven Features," *Speech Communication*, Vol. 20, Nos. 1–2, pp. 131–150, Nov. 1996.
11. B.D. Womack, J.H.L. Hansen, "N-Channel Hidden Markov Models for Combined Stress Speech Classification and Recognition," accepted to *IEEE Trans. Speech & Audio Proc.*, Jan. 1999.
12. G. Zhou, J.H.L. Hansen and J.F. Kaiser, "Classification of Speech under Stress Based on Features from the Nonlinear Teager Energy Operator," *ICASSP'98*, vol. 1, pp. 549–552, Seattle, WA, 1998.
13. G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Linear and Nonlinear Speech Feature Analysis for Stress Classification," *ICSLP-98: Inter. Conf. Spoken Lang. Proc.*, vol. 3, pp. 883–886, Sydney, Australia.



(a)



(b)

Figure 5.9: Assessment results for pilot's speech from Mayday2 domain of SUSC-0 database (Log likelihood ratio is shown along Y-axis while sentence number is shown along X-axis): (a) Neutral vs Simulated stress (Loud, Angry and Lombard) HMM reference models; (b) Neutral vs Actual stress HMM reference models