Linear and Nonlinear Speech Feature Analysis for Stress Classification*

Guojun Zhou, John H.L. Hansen, and *James F. Kaiser*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech gzhou@ee.duke.edu jhlh@ee.duke.edu

ABSTRACT

There are many stressful environments which deteriorate the performance of speech recognition systems. Examples include aircraft cockpits, 911 emergency telephone response, high workload task stress, or emotional situations. To address this, we investigate a number of linear and nonlinear features and processing methods for stressed speech classification. The linear features include properties of pitch, duration, intensity, glottal source, and the vocal tract spectrum. Nonlinear processing is based on our newly proposed Teager Energy Operator (TEO) speech feature which incorporates frequency domain critical band filters and properties of the resulting TEO autocorrelation envelope. In this study, we employ a Bayesian hypothesis testing approach and a hidden Markov model (HMM) processor as classification methods. Evaluations focused on speech under loud, angry, and the Lombard effect[1] from the SUSAS database. Results using receiver operating characteristic (ROC) curves and EER (equal error rate) based detection show that pitch is the best of the five linear features for stress classification; while the new nonlinear TEO-based feature outperforms the best linear feature by +5.2%, with a reduction in classification rate variability from 8.66 to 3.90.

1 Introduction

Current speech recognition and speaker identification techniques are not very successful in adverse conditions where speech technology is needed. There have been limited success in addressing issues related to varying communication channels, handset differences, and increased vocabulary sets. The ability to address such issues has been achieved largely by merely collecting speech data from the same adverse environments, and thereby re-training reference models (i.e., train-test matched conditions). Issues such as handset and channel responses are, to some degree, easier to address, since their influence is generally fixed over a voice transmission. The variability introduced by speech under stress is much more challenging since it has been shown that stress impacts phone classes in a non-uniform manner [7]. Alternate training methods such as multi-style training [10] can improve speech recognition under stress, but at the expense of requiring the user to produce speech across a simulate range of stress styles. Studies have also shown that multi-style training only works in speaker-dependent scenarios, and that performance actually degrades below neutral training if applied in a speaker independent application [13]. The primary reason for this is that stressful conditions are too diverse to be represented by limited training data, and that speakers can at times use a non-uniform set of speech production adjustments to convey their stress state. One approach for robust speech recognition which has shown promise is to first classify input speech as being neutral/stressed [13]. Special processing could then be applied once non-neutral stress states are detected. In this study, only the first task of stress classification is considered.

Speech analysis plays an important role for successful stress classification. Past studies [4, 5, 12] showed that speech characteristics in areas such as duration, intensity, pitch, glottal source, and vocal tract spectrum can be useful as indicators of speech under stress. One experiment in [4] demonstrated useful variation patterns of several features under loud speaking style. For example, duration and intensity are increased for vowels but decreased for semi-vowels and consonants, and both the mean and variance of pitch are increased. These features, however, have never been employed for stress classification.

Our previous study [1, 2] proposed the following three new nonlinear TEO based processing features: TEO-decomposed FM Variation (TEO-FM-Var), normalized TEO Autocorrelation Envelope area (TEO-Auto-Env), and TEO based Pitch (TEO-Pitch). They successfully explored the prospects of variations in the energy of airflow characteristics within the vocal tract for speech under stress.

In this paper, we first consider five linear features from the domains: duration, intensity, pitch, glottal source, and vocal tract spectrum, to classify stressed speech from neutral. Bayesian hypothesis testing is employed for the classification. Second, we introduce our newly-proposed TEO-based feature, which is derived from our previously-proposed 4-band based TEO-Auto-Env feature. This new TEO-based feature employs critical bands to partition the frequency range. Sec. 4 presents stress classification evaluations, with conclusions presented in Sec. 5.

2 Linear Stress Classification

It has been shown that there are observable differences in duration, intensity, pitch, glottal source, and formant locations between neutral and stressed speech [4]. Therefore, it would be useful to evaluate the performance of features from such domains for stress classification. Since

*This work was supported in part by a grant from the U.S. Air Force Research Laboratory/IFEC, Rome NY.

[1]The Lombard effect occurs when a speaker modifies his/her speech in order to increase communication quality when producing speech in the presence of acoustic background noise.

pitch, glottal source information, formant locations are meaningful for vowels, we extracted all five features only from vowel sections of speech so that they can be compared under the same scenario.

2.1 Linear Feature Description

Five features are considered here. The length of each vowel in msec is used as the duration feature. The intensity feature is defined as,

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

where $s(i)$ $(i = 1, \cdots, K)$ represents the K individual samples in the vowel. Pitch, glottal source, and formant locations are extracted on a frame basis with frame length being 32 msec and an overlap between adjacent frames of 16 msec. The modified simple inverse filter tracking (MSIFT) algorithm [3] is employed to extract pitch frequencies from vowel portions. We use spectral slope as the glottal source feature. It is not easy to obtain the glottal spectral slope from the raw vowel speech waveform due to the coupling effect between the sub-glottal structure and the forward portion of the vocal tract. To avoid this effect, we use only data obtained during closed vocal fold periods. Since it is not easy to accurately locate the boundaries between vocal fold closing and opening periods, we compute a frame-based log average amplitude FFT versus log frequency for each vowel section. Next, a straight line is used to approximate the envelope, and the line's slope is considered as the glottal spectral slope. Finally, the first two formant locations are used as the vocal tract spectral information.

2.2 Bayesian Hypothesis Testing for Stress Classification

In our study, we consider pairwise (neutral/stressed) classification. Here, the stress classifier is similar to a Bayesian hypothesis testing system (i.e., two hypotheses: $H0$ and $H1$). Under $H0$, the speech is neutral; while under $H1$, the speech is stressful. When there is an input speech feature vector, $\mathbf{x}$, ($\mathbf{x} = x_1, \cdots, x_M$; M is the vector length), two conditional probability densities have to be calculated, that is, $p(\mathbf{x}|H0)$ and $p(\mathbf{x}|H1)$. The likelihood ratio, λ, is then defined as,

$$\lambda = \frac{p(\mathbf{x}|H1)}{p(\mathbf{x}|H0)}. \quad (2)$$

To obtain $p(\mathbf{x}|H0)$ and $p(\mathbf{x}|H1)$, we need to know the probability density functions (*pdf*) of both neutral and stressed speech features. The *pdf* is obtained from training data by plotting the feature and then fitting a *pdf* to the histogram. If we assume all components ($x_1, x_2, \cdots, x_M$; M is the vector length) of the input speech feature vector, $\mathbf{x}$, have independent identical Gaussian distributions, with mean, μ_n, and variance, σ_n^2, under neutral conditions; and with mean, μ_s, and variance, σ_s^2, under stressed conditions, i.e.,

$$f(x_i|H0) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(x_i-\mu_n)^2}{2\sigma_n^2}\right), \quad (3)$$

$$f(x_i|H1) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(x_i-\mu_s)^2}{2\sigma_s^2}\right), \quad (4)$$

then, $p(\mathbf{x}|H0)$ and $p(\mathbf{x}|H1)$ can be computed as follows,

$$p(\mathbf{x}|H0) = (2\pi\sigma_n^2)^{-\frac{M}{2}} \exp\left(-\frac{1}{2\sigma_n^2}\sum_{i=1}^{M}(x_i-\mu_n)^2\right), \quad (5)$$

$$p(\mathbf{x}|H1) = (2\pi\sigma_s^2)^{-\frac{M}{2}} \exp\left(-\frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(x_i-\mu_s)^2\right). \quad (6)$$

Thus, the log likelihood ratio (LLR) is found as follows,

$$\ln\lambda = M\ln(\frac{\sigma_n}{\sigma_s}) + \frac{1}{2\sigma_n^2}\sum_{i=1}^{M}(x_i-\mu_n)^2 - \frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(x_i-\mu_s)^2. \quad (7)$$

The following form of the LLR can be obtained from Eq. 7 for easy computation,

$$\ln\lambda = M\ln(\frac{\sigma_n}{\sigma_s}) + \frac{M}{2\sigma_n^2}(\hat{\sigma}^2+(\hat{\mu}-\mu_n)^2) - \frac{M}{2\sigma_s^2}(\hat{\sigma}^2+(\hat{\mu}-\mu_s)^2), \quad (8)$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the estimated mean and variance of input vector, $\mathbf{x}$, which are defined as,

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i, \quad \hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu})^2. \quad (9)$$

The decision is made by comparing the LLR with a pre-defined threshold, β. If the LLR is larger than β, the input speech is labeled as stressed; otherwise it is classified as neutral. The value of β 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 probability (FA) and the false rejection probability (FR), should be as low as possible. Obviously, it is not possible to minimize both FA and FR, and hence, a compromise must be made. For some systems, the requirement for one probability is more important than the other. For a stress classification system, however, we are interested in overall accuracy and have no preference. Therefore, the value of β corresponding to equal error (FA=FR) rate (EER) is selected. In our experiment, we calculate the value of FA as the ratio of the number of falsely accepted (neutral/stressed) vowels to the total number of vowels, and the value of FR as the ratio of the number falsely rejected (neutral/stressed) vowels to the total number of vowels. By changing the threshold value, the value of β corresponding to EER can be found.

3 A New TEO-based Feature

Based on Teager's studies [11], the airflow from the vocal folds does not propagate uniformly as a plane wave in the vocal tract, but is separated with energy concentrated near the walls, and thus vortices are generated throughout the vocal tract. To measure the energy from speech which is produced by such a nonlinear process, Teager developed an energy operator, which is known as Teager Energy Operator (TEO) as follows,

$$\Psi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \quad (10)$$

where $\Psi[\cdot]$ is the TEO, and $x(t)$ is a single component of the continuous speech signal. Kaiser [8, 9] later derived the operator for discrete-time signals from its continuous form $\Psi_c[x(t)]$, as,

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1), \quad (11)$$

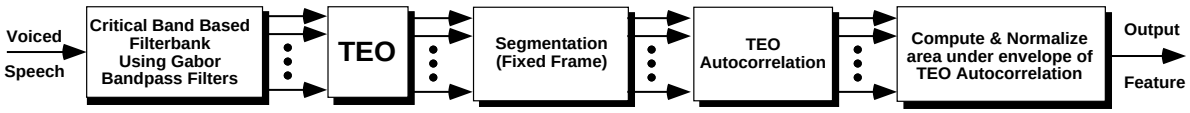

Figure 1: TEO-CB-Auto-Env Feature Extraction

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

Although TEO processing is intended for a signal with a single resonant frequency, our previous work [1, 2] showed that the TEO energy of a multi-frequency signal is not only different, but also reflects interactions between frequency components. This characteristic extends the use of TEO to speech signals filtered with wide bandwidth band-pass filters (BPF).

In [1], we proposed three TEO-based nonlinear features for stress classification. The 4-band filterbank-based feature, TEO-Auto-Env, was promising and consistent. It is suggested that stress may affect different frequency bands differently and an improved stress classification feature could be obtained by increasing the number of filterbank partitions to better reflect subtle energy changes across frequency. Empirically, the human auditory system is assumed to be a filtering process which partitions the entire audible frequency range into many critical bands. Based on this assumption, the new proposed feature employs a critical band filterbank to partition the speech signal followed by TEO processing (see Fig. 1). Each filter in the filterbank is a Gabor bandpass filter, with the effective root mean square (RMS) bandwidth being the corresponding critical band. To extract the TEO-CB-Auto-Env feature vector, each TEO profile is segmented into 25 msec frames with 12.5 msec overlap between adjacent frames in time. Similar to the extraction of the TEO-Auto-Env feature, M normalized areas under the TEO autocorrelation upper envelope 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.1 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 $F0$ is not needed in the TEO-CB-Auto-Env extraction. This makes the new feature independent of the accuracy of the median $F0$ estimation.

4 Evaluations

A 33-word vocabulary under neutral, angry, loud, and the Lombard effect speaking styles from the simulated domain of the SUSAS (*Speech Under Simulated and Actual Stress*) database [6] is employed for our evaluations. All data was obtained by requesting speakers to speak in the corresponding style (Lombard effect simulated with 85dB SPL pink noise played through headphones). Only vowel sections of each vocabulary word were used for evaluations.

4.1 Linear Features

From all identified vowels, duration, intensity, pitch, glottal spectral slope, and formant locations are extracted using the methods described in Sec. 2.1. For each feature, we use all available data to estimate the *pdf* (Fig. 2 shows one example positive conditional Gaussian pdf for pitch under loud stress style). These *pdf*s are then used to obtain the ROC curves for the Bayesian hypothesis testing approach. In order to achieve open-set test results, in the test phase we first divide the data of each feature into 10 equal size sets. For each of the 10 sets, we test with 1 subset and train with the other 9 in order to obtain the EER threshold for Bayesian hypothesis testing. The final error rate is obtained by accumulating all error rates from the 10 open-set tests. Different testing feature vector lengths $(1, 5, 10)$ are used to obtain ROC curves and error rates. An example of one of the many ROC curves is shown in Fig. 3 for stress classification between neutral and loud for mean pitch information. Table 1 shows detection results for all five feature domains using the Bayesian hypothesis testing approach.

From Table 1, we can see that (1) pitch is the best feature for stress classification among the five features, (2) error rates generally decrease as feature vector length increases, (3) there are performance differences between different styles of stress, and (4) mean vowel formant locations are not suitable for reliable stress classification.

4.2 TEO-based Nonlinear Feature

For the Bayesian hypothesis stress classification approach using the above linear features, we made an assumption that all feature vector components were independent(i.e., we did not consider the correlation between feature components). This assumption is actually not always true (e.g., pitch values of two adjacent frames are usually dependent). To consider the correlation issue, Eq. 7, 8 must be re-derived. An alternate approach, however, is to choose an HMM framework since HMMs are capable of modeling temporal transitions across frames. For this evaluation, a baseline 5-state HMM-based stress classifier with continuous Gaussian mixture distribution was employed. Only two HMM models (neutral and stressed) were trained for each pairwise classification. For purposes of comparison, we also used the HMM classifier to evaluate the pitch information. The evaluation results are shown in Fig. 4. As expected, pitch achieves better performance with HMM vs. the Bayesian hypothesis testing approach. Furthermore, it is shown that the new TEO-CB-Auto-Env feature not only performs better than our previous TEO-Auto-Env feature, but also outperforms the pitch feature in terms of both average accuracy and consistency across different stress styles.

5 Conclusions

In this paper, we investigated linear and nonlinear speech features for the classification of speech under stress. Both Bayesian hypothesis and HMM frameworks are employed for stress classification. Evaluation results show that pitch is the best stress classification feature among the five linear features. Although formant locations show observable difference under neutral and stressful conditions, they were not suitable for stress classification. Motivated by our previous TEO-based features reported in [1], we

Vector Length	Feature	Speaking Style of Submitted Test Speech						OVERALL ERROR RATES	
		Neutral	Angry	Neutral	Loud	Neutral	Lombard	Mean $m_{\bar{A}LL}$	stand. dev. σ_{ALL}
1	Duration	45.13	45.38	38.21	38.72	40.77	40.26	41.41%	3.12
	Intensity	40.26	37.44	34.87	32.82	40.77	39.49	37.61%	3.19
	Pitch	18.95	18.57	11.94	11.63	24.08	24.18	18.23%	5.54
	Glottal	33.33	36.78	41.38	41.72	42.76	42.07	39.67%	3.77
	Formant 1	42.60	41.80	46.43	45.10	46.84	46.90	44.95%	2.24
	Formant 2	51.48	50.88	58.20	54.51	52.98	49.88	52.99%	3.03
5	Duration	36.36	38.96	33.77	35.06	40.26	40.26	37.45%	2.78
	Intensity	24.68	22.08	27.27	22.08	38.96	35.06	28.36%	7.08
	Pitch	15.17	14.31	10.34	10.00	21.90	22.07	15.63%	5.34
	Glottal	25.45	21.82	30.91	34.55	30.91	36.36	30.00%	5.49
	Formant 1	40.60	40.30	46.12	45.82	47.91	46.87	44.61%	3.31
	Formant 2	53.88	49.85	58.51	56.12	54.78	50.90	54.00%	3.23
10	Duration	41.03	35.90	38.46	35.90	38.46	46.15	39.32%	3.86
	Intensity	23.08	17.95	28.21	17.95	35.90	35.90	26.50%	8.22
	Pitch	12.76	11.72	7.24	8.28	20.69	19.31	13.33%	5.58
	Glottal	25.00	17.86	35.71	35.71	28.57	32.14	29.19%	6.93
	Formant 1	38.79	40.91	43.03	44.24	47.58	47.88	43.74%	3.61
	Formant 2	55.76	47.58	59.39	57.27	53.33	55.15	54.75%	4.06

Table 1: Error Rates (percentage) of Open-set Pairwise Stress Classification Using Five Linear Features

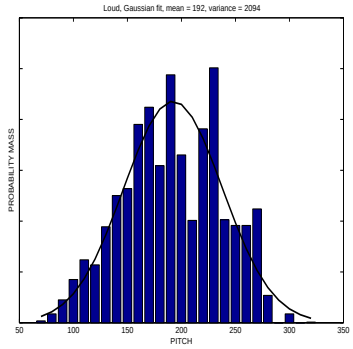

Fig. 2. Conditional Gaussian *pdf*, $N(\mu, \sigma^2|X \geq 0)$ (here $\mu = 192$ Hz, $\sigma^2 = 2094$) to approximate the pitch distribution of vowels under loud speaking style

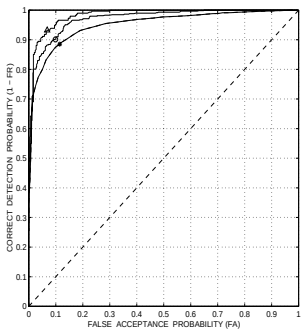

Fig. 3. ROC curves for detection of "loud" speech from neutral using pitch (input vector lengths: * = 1; o = 5; △ = 10)

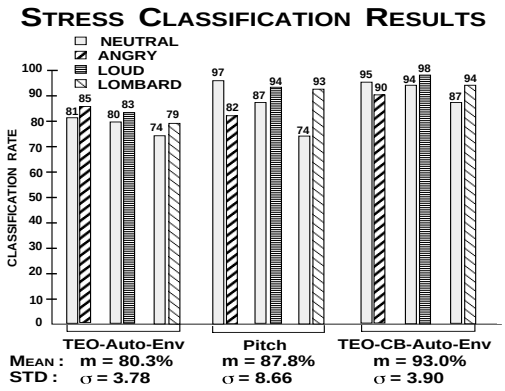

Fig. 4. Pairwise Stress Classification Results (Mean and standard deviation of overall neutral/stress classification rates are shown)

proposed a new TEO-based nonlinear feature, TEO-CB-Auto-Env. The new feature was shown to have better performance for stress classification than our previous features. Moreover, it outperforms the linear pitch feature by +5.2%, and is more consistent across different stress styles for stress classification. Even though we may conclude here that the nonlinear TEO-CB-Auto-Env feature using an HMM structure is very effective for stress classification, we suggest that it may be possible to integrate HMM and Bayesian hypothesis frameworks for multiple feature types. Since speakers can differ in how they adjust speech features to convey stress, a combination of linear and nonlinear features may be needed for universal speaker stress classification.

References

[1] G. Zhou, J.H.L. Hansen, J.F. Kaiser, "Classification of Speech under Stress Based on Features from the Nonlinear Teager Energy Operator," *IEEE ICASSP-98*, **1**:549-52, Seattle, WA, 1998.

[2] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress", submitted to *IEEE Trans. Speech & Audio Process.*, Dec. 1997.

[3] L. Arslan, "Foreign Accent Classification", Ph.D. Thesis, Duke Univ., Dept. Elect. & Comp. Eng., July, 1996.

[4] J.H.L. Hansen, "Analysis and Compensation of Stressed & Noisy Speech with Application to Robust Automatic Recognition", Ph.D. Thesis, Georgia Inst. of Tech., Atlanta, GA, 1988.

[5] J.H.L. Hansen, "Analysis and Comparison of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communication*, **20**:151–173, 1996.

[6] J.H.L. Hansen, S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database", *EUROSPEECH–97*, **4**:1743-46, Rhodes, Greece, Sept. 1997.

[7] J.H.L. Hansen, D.A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful & Lombard effect environments," *Speech Comm.*, **16**:391-422, 1995.

[8] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *IEEE ICASSP–90*, pp. 381–384, 1990.

[9] J.F. Kaiser, "Some Useful Properties of Teager's Energy Operator," *IEEE ICASSP–93*, Vol. 3, pp. 149 – 152, 1993.

[10] R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multi-style Training for Robust Isolated-word Speech Recognition", *ICASSP–87*, pp. 705 – 708, Dallas, TX, USA, 1987.

[11] H. Teager, S. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract," *Speech Production & Speech Modeling*, NATO Advanced Study Inst., Vol. 55, Bonas, France, (Boston: Kluwer Acad. Pub.), pp. 241-61, 1990.

[12] C. Williams, K. Stevens, "Emotions and Speech: Some Acoustic Correlates", *J. Acoust. Soc. Am.*, **52**(4):1238-50, 1972.

[13] B.D. Womack, J.H.L. Hansen, "Classification of Speech under Stress Using Target Driven Features", *Speech Communication*, **20**(1/2):131-150, 1996.