CLASSIFICATION OF STRESS IN SPEECH USING LINEAR AND NONLINEAR FEATURES

Tin Lay Nwe[1], Say Wei Foo[2] and Liyanage C. De Silva[1]

[1]Department of Electrical and Computer Engineering, 4 Engineering Drive 3, Singapore 117576.
[2]School of EEE, Nanyang Avenue, Nanyang Technological University, Singapore 639798.

ABSTRACT

In this paper, three systems for classification of stress in speech are proposed. The first system makes use of linear short time Log Frequency Power Coefficients (LFPC), the second employs Teager Energy Operator (TEO) based Nonlinear Frequency Domain LFPC features (NFD-LFPC) and the third uses TEO based Nonlinear Time Domain LFPC features (NTD-LFPC). The systems were tested using SUSAS (Speech Under Simulated and Actual Stress) database to categorize five stress conditions individually. Results show that, the system using LFPC gives the highest accuracy, followed by the system using NFD-LFPC features. While the system using NTD-LFPC features gives the worst performance. For the system using linear LFPC features, the average accuracy of 84% and the best accuracy of 95% were obtained in classifying five stress categories.

1. INTRODUCTION

Intra-speaker variability introduced by a speaker under stress degrades the performance of the recognizers trained with neutral tokens. A study conducted by Womack [1] showed that speech recognition could be made more robust if stress classification scores were integrated into it under multi-style training approach. A number of studies have been conducted to investigate acoustic indicators for stress in speech. The characteristics most often considered include fundamental frequency (F0) [2],[3], duration [2],[4],[3], intensity [2], spectral variation [4] and features derived from Teager Energy Operator (TEO)[5],[6]. Most of the studies on the analysis of stress focus on fundamental frequency F0. Hansen [3] made extensive statistical evaluations on pitch, glottal source, duration, intensity and vocal tract to characterize the stress on speech and identified pitch period to be one of the best stress discriminating parameters. However, these characteristics are not useful in discriminating stress arising from moderate versus high task workload conditions.

Distribution of spectral energy also varies on speech produced under stress [3]. Unvoiced speech is associated with low energy speech sections and voiced speech is associated with high-energy speech [4]. In the earlier research, features used were mostly derived from linear speech production models. In recent years, non-linear features derived from Teager Energy Operator (TEO) [5],[6] are also explored. TEO based features are recognized to reflect the nonlinear airflow structure of speech produced under stressful conditions. Although these TEO based features are able to distinguish well for pair-wise classification between 'Neutral' and stress [5], the classification performance decreases substantially when classifying stress styles individually [6].

In this paper, investigation is made to determine the set of acoustic features required for both pair-wise (stress/ neutral) classification and multi-style classification (classify each stress styles individually). The block diagram of the proposed system is shown in Figure 1.

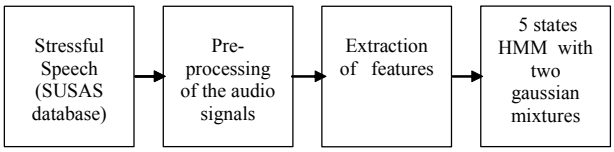

Figure 1. Block diagram of the proposed systems

The signal samples are segmented into frames. For each frame, a feature vector based on Log Frequency Power Coefficients and nonlinear TEO based LFPC feature parameters are obtained. Five-state HMM (Hidden Markov Model) based stress classifier with continuous Gaussian mixture distribution is employed for classification. Four stress styles, namely, 'Anger', 'Clear', 'Lombard' and 'Loud' together with 'Neutral', are selected for identification. The theory of HMM is well documented in [7]. Details of some of the other stages are presented in the subsections that follow.

2. ANALYSIS OF STRESS IN SPEECH

Normal speech may be regarded as speech made in a quiet room with no task obligations. Stress in speech, on the other hand, is a result of speech produced under emotional states, fatigue, heavy workload, environmental noise, and/or sleep loss. Some of the consequences of physiological stress are respiratory changes including increased respiration rate, irregular breathing and increased muscle tension of the vocal cords. These factors may result in irregular vocal fold movement and other vocal system modifications that ultimately affect the quality of the utterances [8]. The presence of stress in speech causes changes in phoneme production with respect to glottal source factors, pitch, intensity, duration, and spectral shape [3].

In linear acoustic theory, speech production process is described in terms of source/filter model [9]. This model assumes plane wave propagation in the vocal tract and neglects nonlinear terms. Linear acoustic theory suggests that frequency in vocal tract filter, intensity and duration of glottal signal may be assumed to change due to stressed speech production.

In this paper, linear acoustic features and nonlinear features in frequency domain have been investigated in stress classification

since the system performance deteriorate when using nonlinear features in time domain in characterizing different stress styles [5],[6].

3. SELECTION OF STRESS CLASSIFICATION FEATURES

Human auditory system is assumed to have the filtering system in which entire audible frequency range is partitioned into subbands. It is suggested that stress may affect different frequency bands differently and an improved stress classification features could be obtained by analyzing energy in different frequency bands. Based on these assumptions, a feature based on the distribution of energy in different log frequency bands is selected. By analyzing these feature data using an HMM recognizer, the effects of speaking rate and variation of tone are also taken care of.

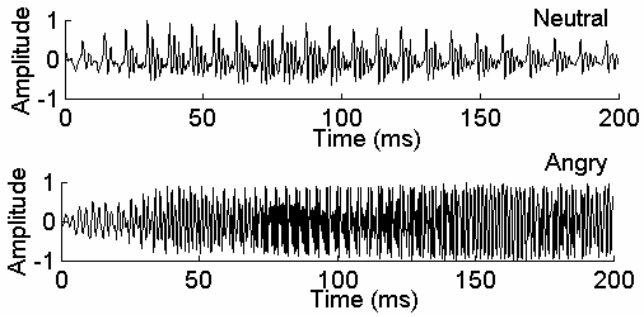

Figure 2. Waveforms of a segment of the speech signal produced under 'Neutral' and 'Anger' conditions of the word 'go' by a male speaker (200ms duration)

As can be seen from Figure 2, voiced speech spoken under stress is significantly different from voiced speech spoken under 'Neutral' or 'Normal' condition in both frequency and intensity variations. This observation suggests that features based on the distribution of energy in different frequency bands should be useful for stress classification.

4. COMPUTATION OF SPEECH FEATURES

4.1. Log Frequency Power Coefficients (LFPC)

In order to extract DFT based subband features, we make use of filter banks in different log frequency bands from 200Hz to 3.9kHz as represented in Figure 3. A Log frequency filter bank can be regarded as a model that follows the varying auditory resolving power of the human ear for various frequencies. The filter bank adopted in the study is designed to divide speech signal into 12 frequency bands that match the critical perceptual bands of the human ear. The center frequencies f_i and bandwidths b_i for a set of 12 bandpass filters are derived as in [7].

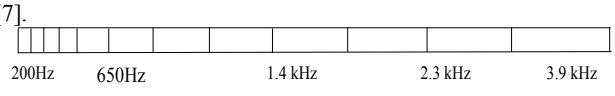

Figure 3. Subband frequency divisions

To compute the frame based energy variations in different Log Frequency bands, the signal samples are segmented into frames of 16ms each with 9ms overlap between consecutive frames. The samples of each frame are weighted with a Hamming window to reduce spectral leakage. This windowed speech is transformed to the frequency domain using the DFT (Discrete Fourier Transform) algorithm. The spectral components are separated into 12 bands. The m^{th} filter bank output is given by:

$$S_t(m) = \sum_{k=f_m-\frac{b_m}{2}}^{f_m+\frac{b_m}{2}} \left(X_t(k)\right)^2 \qquad m = 1,2,...,12 \tag{2}$$

where $X_t(k)$ = the k^{th} spectral component of the windowed signal, t = frame number, $S_t(m)$ = output of the m^{th} filter bank , f_m , b_m = center frequency and bandwidth of the m^{th} subband.

The parameters, $SE_t(m)$, which provide an indication of energy distribution among sub-bands, are calculated as follows.

$$SE_t(m) = \frac{10\log_{10}(S_t(m))}{N_m} \tag{3}$$

where N_m = the number of spectral components in the m^{th} filter bank. For each speech frame, 12 Log Frequency Power Coefficients are obtained.

4.2. Nonlinear Time/Frequency Domain LFPC

The study by Douglas [5], suggested that Teager Energy profile alone is not sufficient to reliably separate 'Lombard' effect speech from 'Neutral' speech. They recommended that the features relating to spectral shape should be incorporated into TEO based features to separate these two speaking conditions. In this paper, TEO based nonlinear properties in combination with the LFPC are also investigated. TEO is commonly applied in the time domain [5],[6]. In this paper, TEO in both time and frequency domain are considered.

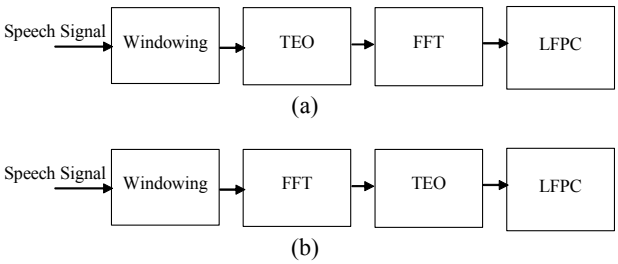

Figure 4. (a) Nonlinear time domain LFPC feature extraction (b) nonlinear frequency domain LFPC feature extraction

The process of feature extraction for Nonlinear Time Domain LFPC (NTD-LFPC) and Nonlinear Frequency Domain LFPC (NFD-LFPC) are shown in Figures 4(a) and 4(b) respectively. The same window size and frame rate are employed as for LFPC.

For NTD-LFPC, Teager Energy Operator (TEO) described in Kaiser [10] is applied to the time domain windowed speech signal as described in the equation below.

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \quad (4)$$

In the above equation, $x(n)$ is sampled speech component in the time domain and $\Psi[x(n)]$ is the TEO operator. Fast Fourier Transform is then applied to obtain the LFPCs.

For NFD-LFPC, time domain windowed speech signal is converted to frequency domain using DFT (Discrete Fourier Transform) and the following TEO operation is then applied.

$$\Psi[x(f)] = x^2(f) - x(f+1)x(f-1) \quad (5)$$

In (5), $x(f)$ is sampled speech component in the frequency domain. As an illustration, the time domain and frequency domain representations together with the results after the TEO operation of a segment of the neutral and 'Anger' speech signals of the word 'destination' are shown in Figure 5.

5. EXPERIMENTS AND RESULTS

5.1. Conduct of experiments

The proposed system is evaluated using the simulated portion of SUSAS (Speech Under Simulated and Actual Stress) database. SUSAS has been employed extensively in the study of the effect on speech production and recognition when speaking under stressed conditions [1],[2],[3],[5],[6]. The stress classifier consists of five-state continuous density HMM model with two Gaussian mixtures per states for each stress style.

For pair-wise classification, the stress style for the model that gave the higher score was taken as the style to be identified. For multi-style stress classification, the HMM model with the highest score was selected among five models with five speaking styles.

Table 1. Average classification performance

Speaker	FEA1 (%)		FEA2 (%)		FEA3 (%)	
	Mul	Pw	Mul	Pw	Mul	Pw
S1	95	99	86.4	94	76.4	89.6
S2	82.9	96.4	78.6	89.6	70.7	89.5
S3	84.3	95.7	78.6	92.5	71.4	92
S4	89.3	97.1	87.9	95.9	81.4	92.3
S5	82.1	97	72.9	96.1	62.9	89.3
S6	74.3	90.4	74.3	93.2	67.1	86.6
S7	85	95.7	77.9	91.8	62.1	83.9
S8	90	96.6	81.4	95.2	65.7	76.1
S9	77.1	95.5	70.7	93.2	62.9	85.2
Mean	84.4	96	78.7	93.5	69	87.2

FEA1= LFPC, FEA2=NFD-LFPC, FEA3=NTD-LFPC
Mul=Multi-style, Pw=Pair-wise

First, the performance of the system in classifying stress/neutral (pair-wise) speech was assessed. From the results shown in Table 1 it is observed that all four stress styles 'Anger', 'Clear', 'Lombard' and 'Loud' can be well differentiated from 'Neutral'

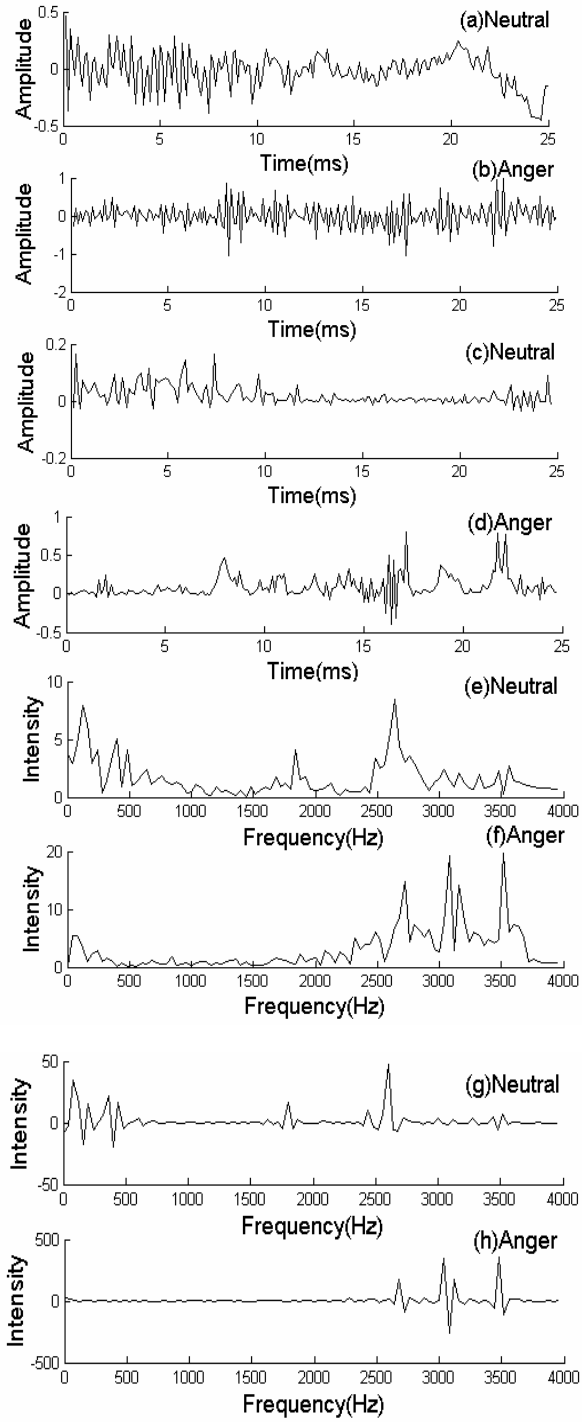

Figure 5. (a),(b) Wave forms of a segment of the word 'destination' spoken by a male speaker under 'Neutral' and 'Angry' conditions respectively . (c),(d) Teager Energy operation of the respective signals in the time domain. . (e),(f) Intensity variation of respective signals in the frequency domain (g),(h) Teager Energy Operation of the respective signals in the frequency domain

style. The classification performances are also consistent across all stress styles. The mean pair-wise classification rates by LFPC and NFD-LFPC are higher than the mean pair-wise classification accuracy (89.1%) reported by Cairns and Hansen [5] where the same database was used to evaluate the same set of speaking styles and where TEO based nonlinear features in time domain was used. Furthermore, the performance of their system degrades when classifying 'Clear' speech from neutral [5].

From the results summarized in Table 1 for the multi-style classification, the system using LFPC gives the highest accuracy, followed by the system using NFD-LFPC features. While the system using NTD-LFPC features gives the worst performance. The classification accuracies are also consistent across all speaking styles using LFPC and NFD-LFPC features.

The high accuracy of classification by LFPC suggests that variation of intensity values across subbands in DFT based log frequency scale provides essential information for distinguishing stress and neutral speech. The comparative performance between NTD-LFPC and NFD-LFPC also shows that nonlinear variation of energy distribution in frequency domain is more significant than that in time domain for discriminating stressed speech.

Comparing Figure 5(e) and Figure 5(f), which show the LFPC representations of 'Anger' and 'Neutral' respectively, it can be observed that the difference is the most conspicuous among all the figures grouped under Figure 5. Furthermore, as can be seen from Figures 5(f) and (h) for 'Anger' stress, TEO operation suppresses certain intensity values in the frequency range 3.2kHz to 3.5kHz down to near zero because of nonlinear property analysis. This results in the loss of important information on high frequency energy, which is an essential feature of 'Anger' stress style [11]. This further explains the superior performance of LFPC over the TEO-based LFPC.

Between NFD-LFPC (Figure 5(g) and Figure 5(h)) and NTD-LFPC (Figure 5(c) and Figure 5(d)), it can also be observed that nonlinear energy variations in frequency domain present more significant discrimination between 'Anger' and 'Neutral' conditions. 'Anger' has higher intensity in higher frequency scales and 'Neutral' style has higher intensity values in lower frequency scales. This shows that Teager Energy operation in frequency domain is more capable than in time domain to detect stress.

6. CONCLUSION

In this paper, a novel system for stress classification is proposed that focus on the application of linear Log Frequency Power Coefficients (LFPC) and nonlinear acoustic features (Teager Energy Operation based LFPC) in both time and frequncy domain to represent speaking styles. The system is evaluated for both pair-wise and multi-style stress classification. Results show that very high classification rates can be achieved using the LFPC and NFD-LFPC as features for both pair-wise and multi-style stress classification. The average accuracy of classification using LFPC is higher than that using nonlinear LFPC. It can therefore be said that energy distribution of the signal in the different log frequency bands provides a good representation of the stress styles. Comparing the two approaches for TEO operation, nonlinear variation of energy distribution in frequency domain provides a better representation than that in time domain.

7. REFERENCES

[1] B.D. Womack, J.H.L. Hansen, "Classification of Speech Under Stress Using Target Driven Features", Speech Communications, Special Issue on Speech Under Stress, vol. 20(1-2), pp. 131-150, November 1996.

[2] J.H.L. Hansen, "Analysis and Compensation of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition," Speech Communications, Special Issue on Speech Under Stress, vol. 20(2), pp. 151-170, November 1996.

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

[4] J.H.L. Hansen, S. Bou-Ghazale, "Duration and Spectral Based Stress Token Generation for Keyword Recognition Using Hidden Markov Models", IEEE Transactions on Speech & Audio Processing, vol. 3, no. 5, pp. 415-421, September 1995.

[5] D. Cairns, J.H.L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions", The Journal of the Acoustical Society of America, vol. 96, no. 6, pp. 3392-3400, December 1994.

[6] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech Under Stress", IEEE Transactions on Speech & Audio Processing, vol. 9, no. 2, pp. 201-216, March 2001.

[7] L.R. Rabiner, and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, N.J, 1993.

[8] H.J.M. Steeneken, J.H.L. Hansen, "Speech Under Stress Conditions: Overview of The Effect of Speech Production and on System Performance," IEEE ICASSP-99: Inter. Conf. on Acoustics, Speech, and Signal Processing, vol. 4, pp. 2079-2082,Phoenix, Arizona, March 1999.

[9] G. Fant, Acoustic Theory of Speech Production, (Mouton, La Hauge), 1960.

[10] J.F. Kaiser, "Some Useful Properties of Teager's Energy Operator", in Proc, Int. Conf. Acoustic, Speech, Signal Processing '93, vol. 3, pp. 149-152, 1993

[11] R. Sarikaya, and J.N. Gowdy, "Subband Based Classification of Speech Under Stress", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol, 1, pp. 569 -572, 1998.