ICICS-PCM 2003
15-18 December 2003
Singapore

Detection of Stress and Emotion in Speech Using Traditional and FFT Based Log Energy Features

T L Nwe[1]; S W Foo[2]; L C De Silva[3]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3]Institute of Information Sciences & Technology, Massey University, New Zealand

Abstract

In this paper, a novel system for detection of human stress and emotion in speech is proposed. The system makes use of FFT based linear short time Log Frequency Power Coefficients (LFPC) and TEO based nonlinear LFPC features in both time and frequency domains. The performance of the proposed system is compared with the traditional approaches which use features of LPCC and MFCC. The comparison of each approach is performed using SUSAS (Speech Under Simulated and Actual Stress) and ESMBS (Emotional Speech of Mandarin and Burmese Speakers) databases. It is observed that proposed system outperforms the traditional systems. Results show that, the system using LFPC gives the highest accuracy (87.8% for stress, 89.2% for emotion classification) followed by the system using NFD-LFPC feature. While the system using NTD-LFPC feature gives the lowest accuracy.

1. Introduction

Human speech recognition process uses a combination of sensory sources including facial expressions, gestures, non-verbal information such as emotion, stress as well as meaning conveyed in speech to respond to speaker's message accurately. In this study, stress in speech refers to speech made under environmental noise, emotions and high workload conditions. In human-computer interaction, it is more natural if one can try to interact with computers in the same way as human to human interaction. In order to facilitate human computer communication, human-computer interfaces should be sensitive to the user's expressed emotions or stress types. Research on human computer interaction considers detection of emotion or stress as important information for man-machine interactions.

A number of studies have been conducted to investigate acoustic indicators to detect stress and emotion in speech. The characteristics most often considered include fundamental frequency (F0) [1-4], duration [1, 2], intensity [1, 3], spectral variation [2, 5] and wavelet based subband features [6]. In these researches, features used are mostly derived from linear speech production models. However, in recent years, non-linear features derived from Teager Energy Operators (TEO) [7, 8] are 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 [8], the classification performance decreases substantially when classifying stress styles individually [7].

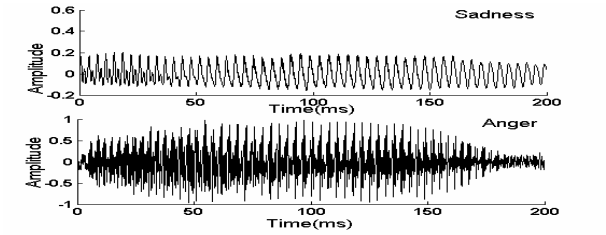
Form these studies, it appears that a number of basic emotions such as Anger, Disgust, Fear, Joy, Sadness, Surprise and the stress speech styles of Anger, Clear, Lombard, Loud, Neutral have been described in terms of changes in fundamental frequency F0, duration, intensity and spectral energy. However, certain emotion or stress utterances have very similar characteristics based on the above set of features. Hence, systems based on these features for emotion or stress classification are unable to accurately distinguish more than a couple of stress or emotion categories. Modification of the above parameters may obtain promising results for stress and emotion classification studies.

In this paper, investigation is made to determine the set of acoustic features required to classify among many categories of emotion and stress styles from the speech signals.

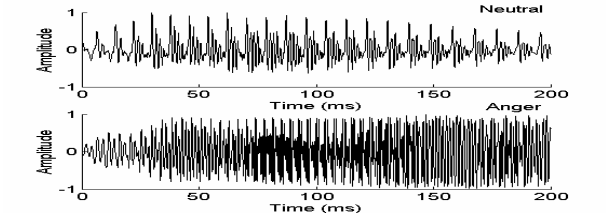
The signal samples are segmented into frames which cover approximately two periods of fundamental frequency as recommended in [8]. For each frame, a feature vector based on Log Frequency Power Coefficients (LFPC) and nonlinear TEO based LFPC feature parameters are obtained. Traditional features, MFCC and LPCC are also extracted for the purpose of comparison with proposed features. Four-state ergodic HMM (Hidden Markov Model) based stress or emotion 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 stress identification. Six emotion categories of Anger, Disgust, Fear, Joy, Sadness and Surprise are recorded for emotion classification. The theory of HMM is well documented in [9]. Details of the other stages are presented in the subsections that follow.

2. Selection of Stress and Emotion Classification Features

As can be seen from Figure 1(a), voiced speech spoken under Anger emotion is significantly different from voiced speech spoken under Sadness emotion in both frequency and intensity domains. The similar trend can also be observed between Anger and Neutral conditions in Figure 1(b). In view of these factors, spectral structure is altered when speech is under emotion or workload stress. One possible measure of the stress and emotional content of speech is thus the distribution of the spectral energy across the speech range of frequencies.



(a)



(b)

Figure 1: (a) Sadness and Anger emotions spoken by Burmese female speaker (200*ms* duration, ESMBS database) (b) Waveforms of a segment of the speech signal produced under Neutral and Anger conditions of the word 'go' by a male speaker (200*ms* duration, SUSAS database)

Human auditory system is assumed to have a filtering system in which the entire audible frequency range is partitioned into frequency bands [9]. According to Fletcher [10], speech sounds are preprocessed by the peripheral auditory system through a bank of bandpass filters. These auditory filters perform the process of frequency weighing for frequency selectivity of ear. Another important aspect in human auditory perception is loudness. In terms of perception of loudness, speech sound can be ranked on a scale extending from quiet to loud.

As shown in Figure 1, different types of stress and emotion may affect different frequency bands differently and an improved stress classification features could be obtained by analyzing energy in several different frequency bands. By extending the subbands to fundamental frequency, information of fundamental frequency can be included on this feature. By analyzing these feature data using an HMM recognizer, the effects of speaking rate and variation of tone are also taken care of. Based on all these assumptions, a feature based on the distribution of energy in different log-frequency bands is selected.

In [6], wavelet based subband features are proposed as an important stress speech relayers. However, wavelet based subband decomposition provides time dependent spectral features which may be more suitable for speech recognition [11] than stress classification. The reason is that specific phoneme sequence variation in time is important in recognizing words. However, stress speech such as Anger cannot be assumed to contain specific sequential events in the signal. For example, if loudness is associated with the Anger stress, there is no fixed time in the utterance for loudness to occur. It can be an event at the beginning, the middle or the end of the utterance. As long as loudness occurs, Anger may be considered [12]. For this reason, FFT (Fast Fourier Transform) based subband features together with HMM are more suitable for stress or emotion classification since Fourier analysis preserves linearity in frequency resolution with no time dependency.

2.1 Computation of FFT based log energy features

Log-Frequency Power Coefficients (LFPC): FFT based Log-Frequency Power Coefficients (LFPC) are designed to simulate logarithmic filtering characteristics of human auditory system by measuring spectral band energies. First, the signal is segmented into short-time windows which are 16*ms* and 20*ms* for emotion and stress speech samples respectively. The reason for using short frame length for emotion database is that it includes female speech utterances which have shorter pitch period than male speech and frame size needs to cover two pitch period of fundamental frequency [8]. Then, the window is moved with the frame rate (overlapping portion between consecutive frames) 9*ms* for emotion and 13*ms* for stress speech samples. And the frequency content is calculated in each frame using Fast Fourier Transform (FFT) method. Then, this power spectrum is accumulated into a bank of log-frequency filters. The filterbank splits input speech signal into multiple outputs by passing through the parallel set of bandpass filters which range from 100*Hz* to Nyquist frequency (half of the sampling frequency). Energy in the m^{th} filter bank output is calculated by the following equation.

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

where, $X_t(k)$ is the k^{th} spectral component of the windowed signal, $W_m(k)$ is a rectangular window, t is the frame number, $S_t(m)$ is the output of the m^{th} filter bank, and f_m, b_m is center frequency and bandwidth of the m^{th} subband respectively. The parameters, $SE_t(m)$, which

provide an indication of energy distribution among sub-bands, are calculated as follows.

$$SE_t(m) = \frac{10\log_{10}(S_t(m))}{N_m} \quad (2)$$

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

Nonlinear Time/Frequency Domain LFPC (NTD-LFPC and NFD-LFPC): In this study, Teager Energy Operator (TEO) nonlinear FFT based features which could improve stress and emotion classification performance are also studied. It is suggested that Teager Energy profile alone is not sufficient to reliably separate Lombard effect speech from Neutral speech [8]. The features relating to spectral shape should be incorporated into Teager Energy Operation (TEO) based features to separate these two speaking conditions [8]. For this reason, TEO based nonlinear properties in combination with the LFPC are investigated. TEO is commonly applied in the time domain [7, 8]. In this paper, TEO in both time and frequency domain are considered.

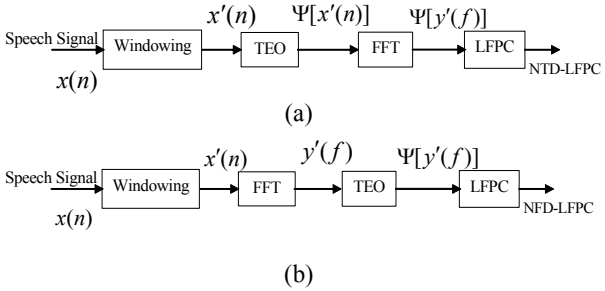


Figure 2. (a) Nonlinear time domain LFPC feature extraction (b) Nonlinear frequency domain LFPC feature extraction

The processes of feature extractions for Nonlinear Time Domain LFPC (NTD-LFPC) and Nonlinear Frequency Domain LFPC (NFD-LFPC) are shown in Figures 2(a) and (b) respectively. The same window size and frame rate are employed as for LFPC.

For NTD-LFPC, Teager Energy Operator (TEO) described in Kaiser [13] is applied to the time domain windowed speech signal as described in the equation below.

$$\Psi[x'(n)] = x'^2(n) - x'(n+1)x'(n-1) \quad (3)$$

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

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

$$\Psi[y'(f)] = y'^2(f) - y'(f+1)y'(f-1) \quad (4)$$

In Equation (4), $y'(f)$ is the sampled speech component in the frequency domain. Then, follow the LFPC feature extraction process to obtain NFD-LFPC coefficients.

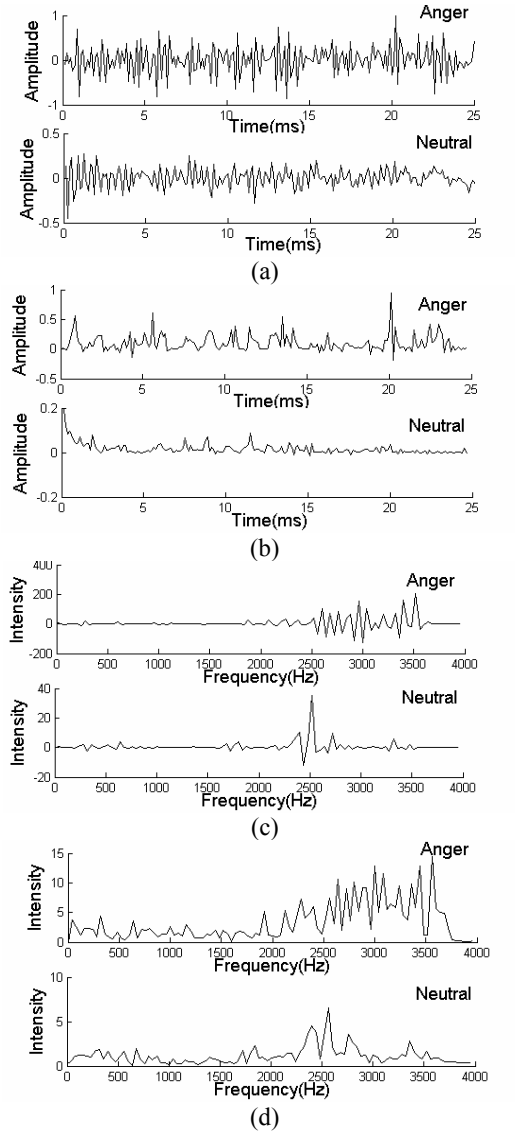


Figure 3. (a)Wave forms of 25*ms* segment of the word 'destination' spoken by a male speaker under Anger and Neutral conditions (SUSAS database) (b) Teager Energy operation of the respective signals in the time domain. (c) Teager Energy operation of the respective signals in the frequency domain (d) Intensity variation of the respective signals in the frequency domain.

The time domain and frequency domain representations together with the results after the TEO operation of segments of Anger stress and Neutral speaking styles are shown in Figure 3. Comparing the LFPC representations of two stress styles shown in Figure 3(d), it can be observed that the difference is the most conspicuous between Anger (high arousal stress) and Neutral (low arousal) speaking

conditions among all the figures grouped under Figure 3. Anger has the higher frequency content than Neutral. Furthermore, as can be seen from Figure 3(c) for Anger stress, TEO operation suppresses certain intensity values in the frequency range $3kHz$ to $3.2kHz$ down to near zero because of nonlinear property analysis. These result in loss of important information on high frequency energy, which is an essential feature of Anger [6]. Between NFD-LFPC (Figure 3(c)) and NTD-LFPC (Figure 3(b)), it can also be observed that nonlinear energy variations in frequency domain present more significant discrimination among different speaking conditions. Anger has high intensity in higher frequency regions. Neutral 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. The same trend has been observed between Anger (high arousal) and Sadness (low arousal) emotions. However, the graphical representations using emotion samples are omitted due to space limitation and can be found in [14].

2.2 Traditional features

For speech recognition, LPC based Cepstral coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) are the popular choices as features representing the phonetic content of speech. To observe the classification performance of stress and emotion utterances using phonetic features and to compare their performances with subband based features, LPCC and MFCC [9] features are extracted using the same window size and frame rate as for LFPC.

3. Stress and Emotion Databases

For stress classification, a subset of SUSAS (Speech Under Simulated and Actual Stress) database including Neutral speech and speech from four simulated stressed styles of Anger, Clear, Lombard and Loud is used. SUSAS has been employed extensively in the study of how speech production and recognition varies when speaking under stressed conditions [6, 15, 16]. The database comprises a highly confusable vocabulary set of 35 aircraft communication words. The training set used consists of 21 words under Neutral condition and four Stress styles from 9 male speakers. The testing set consists of the remaining 14 words. Different sets of words are used for training and testing to assess text independent stress classification. A total of 210 utterances are included in the training set and 140 utterances are included in the testing set for each speaker. A total of 3150 utterances are used.

For emotion classification, a database is specifically designed and set up for text-independent emotion classification. The database includes short utterances covering the six archetypal emotions, namely Anger, Disgust, Fear, Joy, Sadness and Surprise. A total of six native Burmese language speakers (3 males and 3 females) are employed to generate 720 utterances. Sixty different utterances, ten each for each emotional mode, are recorded for each speaker. The recording is done in a quiet environment using a mouthpiece microphone.

4. Experiments and Results

Experiments are conducted to evaluate the performance of the proposed system by the use of three Log-Frequency Power Coefficients (LFPC) based features and 4 state continuous HMM classifier with two Gaussian mixtures per state. In addition to the proposed system, experiments using traditional features of LPCC and MFCC are also conducted for the purpose of comparison.

First, the performance of the system in classifying all the six basic emotions and five stress conditions individually is assessed and it is named as multi-style stress/emotion classification. As stated above, the utterances of certain emotions have similar acoustic features. Hence, experiments are conducted to classify between high arousal emotion group (Anger, Surprise, Joy) and low arousal group (Fear, Disgust, Sadness) and it is referred to as reduced-set classification. For the case of stress utterances, the experiments are conducted to classify between each Stress type and Neutral style and it is named as pair-wise stress classification. The recognition rates of all classification experiments using utterances reserved for testing are shown in Table 1 for all feature sets.

Table 1: Average stress and emotion classification accuracy

Features	Stress (%)		Emotion (%)	
	Mul	Pw	Mul	Rs
LFPC	87.8	97.6	89.2	93.4
NFD-LFPC	86.7	97.6	85.8	93.6
NTD-LFPC	74.6	89.9	78.9	91.8
MFCC	66.6	91.9	77	90.5
LPCC	68.6	88.6	67.3	85.3

Mul=Multi-style, Pw=Pair-wise, Rs= Reduced-set

For multi-style stress classification, it can be seen that accuracies of 87.8% and 86.7% are achieved using LFPC and NFD-LFPC features respectively. However, the performances using NTD-LFPC, MFCC and LPCC features are not as good with the average classification accuracies of 74.6%, 66.6% and 68.6% respectively. MFCC features are formulated to simulate acoustic phonetic perception of human ear. LPCC provides a good approximation to vocal tract spectral envelope. They are mainly designed to extract phonetic content of the speech signal for speech recognition [9]. Therefore, these features do not perform well on stress classification task.

For emotion classification, the best average rates of multi-style classification are 89.2% and 85.8% respectively. Both are obtained when using LFPC and NFD-LFPC features. The average accuracies obtained by using the other features are 78.9%, 77% and 67.3%. This shows that LFPC

and NFD-LFPC are good selections as feature parameters for stress and emotion classification in speech.

The results of pair-wise classification between Neutral and each Stress condition also confirm the superiority of LFPC and NFD-LFPC features over the others. 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 [8] where the same database was used to evaluate the same speaking styles.

When similar emotions are grouped together, the percentage recognition improves dramatically. The best classification accuracies are obtained when using LFPC and NFD-LFPC features.

All the results prove that NFD-LFPC features outperform NTD-LFPC features for all stress and emotion classification experiments. This shows that Teager Energy operation in frequency domain is more capable than in time domain to classify stress.

5. Conclusion

In this paper, a novel system for stress and emotion classification is proposed. Linear acoustic feature LFPC and nonlinear acoustic features NTD-LFPC which is in time domain and NFD-LFPC which is in frequency domain are investigated. It is found that linear Log Frequency Power Coefficients (LFPC) and nonlinear acoustic feature in frequency domain are important in representing speaking styles. Comparing the two approaches for TEO operation, it is observed that nonlinear variation of energy distribution in the frequency domain provides a better representation than that in the time domain. When comparing LFPC based features and two traditional features (MFCC and LPCC), it is found that proposed features (LFPC, NFD-LFPC and NTD-LFPC) perform well over the two traditional features.

References

[1] 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.

[2] S. Bou-Ghazale and J.H.L. Hansen, "A novel training approach for improving speech recognition under adverse stressful environments," EUROSPEECH-97, Sept. 1997, Rhodes, Greece, Vol. 5, pp. 2387-2390.

[3] S .McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk and S. Stroeve. "Approaching automatic recognition of emotion from voice: a rough benchmark," ISCA Workshop on Speech and Emotion, 2000, Belfast.

[4] J.E. Cahn, "The generation of affect in synthesized speech," Journal of the American Voice I/O Society, Vol. 8, pp.1-19, 1990.

[5] S.E. Bou-Ghazale, and J.H.L. Hansen, "Stress perturbation of neutral speech for synthesis based on hidden Markov models," IEEE Transactions on Speech & Audio Processing, Vol. 6, no. 3, pp. 201-216, May 1998.

[6] R. Sarikaya, and J.N. Gowdy, "Subband based classification of speech under stress," Proceedings of the IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing, Vol, 1, pp. 569 -572, 1998.

[7] 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. 3, pp. 201-216, March 2001.

[8] D. Cairns and J.H.L. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," J. Acoust. Soc. Am, Vol. 96, no. 6, pp. 3392-3400, December 1994.

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

[10] H. Fletcher, Auditory Patterns, Review of Modern Physics, Vol.12, pp 47-65,1940.

[11] D.M. Nadeu, and J. Hernando. "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," Speech Communication, Vol. 34, no.1-2, pp. 93-114, April, 2001.

[12] T. Yamada, H. Hashimoto and N. Tosa. "Pattern recognition of emotion with neural network," Proc. IEEE 1995 IECON 21st Inter. Conf. on Industrial Electronics, Control, and Instrumentation, Vol. 1, pp. 183 –187, 1995.

[13] J.F. Kaiser, "Some useful properties of Teager's energy operator," in Proc, Inter. Conf. Acoustic, Speech, Signal Processing '93, Vol. 3, pp. 149-152, 1993.

[14] T.L. Nwe, Analysis and Detection of Human Stress and Emotion from Speech Signals, Ph. D. Thesis, National University of Singapore, 2003.

[15] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," IEEE Transactions on Speech and Signal Processing, Vol 36, no. 4, April 1988.

[16] S. Bou-Ghazale, and J.H.L. Hansen. "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Transactions on Speech and Audio Processing, Vol 8, no. 4, July 2000.