Recognition of Stress in Speech using Wavelet Analysis and Teager Energy Operator

Ling He[1], *Margaret Lech*[1], *Sheeraz Memon*[1], *Nicholas Allen*[2]

[1]School of Electrical and Computer Engineering, RMIT, Australia
[2]ORYGEN Research Centre and Department of Psychology, University of Melbourne, Australia
ling.he@student.rmit.edu.au, margaret.lech@rmit.edu.au, nick.allen@unimelb.edu.au

Abstract

The automatic recognition and classification of speech under stress has applications in behavioural and mental health sciences, human to machine communication and robotics. The majority of recent studies are based on a linear model of the speech signal. In this study, the nonlinear Teager Energy Operator (TEO) analysis was used to derive the classification features. Moreover, the TEO analysis was combined with the Discrete Wavelet Transform, Wavelet Packet and Perceptual Wavelet Packet transforms to produce the Normalised TEO Autocorrelation Envelope Area coefficients for the classification process. The classification was performed using a Gaussian Mixture Model under speaker-independent conditions. The speech was classified into two classes: neutral and stressed. The best overall performance was observed for the features extracted using TEO analysis in combination with the Perceptual Wavelet Packet method. The accuracy in this case ranges from 94% to 96% depending on the type of mother wavelet.

Index Terms: stress in speech detection, Teager Energy Operator, wavelet analysis

1. Introduction

Prosodic features of speech produced under stress vary from features under the neutral condition. The most often observed changes include changes in the utterance duration, decrease or increase of pitch, and shift of formant frequencies. The presence of stress in speech makes the implementation of speech recognition algorithms more complicated compared to neutral speech. Stress recognition and classification is aimed to automatically detect stress in speech signals. It can be used to improve the robustness of speech and speaker recognition systems. Moreover, by assessing speaker's stress level, stress classification could support the automatic assessment of the mental state of people working in dangerous environments (e.g. chemicals, explosives) and people undertaking high levels of responsibility (e.g. pilots, surgeons). Automatic stress classification could support the medical diagnosis of depressive illness or sorting of emergency telephone messages. It could also improve human-computer interaction and help to develop more natural Virtual Reality environments.

A stress classification task consists of two major parts: feature extraction and feature classification. Many resent studies [1][2] focus on the acoustic features, such as pitch features, spectral features and intensity features. There are also studies proposing the use of features such as Linear Predictive coefficients (LPC) [5], and Mel Frequency Cepstral Coefficients (MFCC) [5]. The most often listed classifiers in the literature include neural network, k-nearest neighbors and Gaussian Mixture Model classifiers.

The majority of features listed in the literature are derived from linear models of a speech signal, such as the source-filter model [5]. The underlying assumption of the linear models is that there is a single excitation source with the fundamental frequency F0 (pitch). However, in their studies Teager [3] and Zhou [4] indicate that in the emotional state of anger or stress, the fast air flow causes vortices located near the false vocal folds providing additional excitation signals other than the pitch. The additional excitation signals appear in the speech spectrum as harmonics of fundamental frequencies not equal to F0, and cross-harmonics between the F0 source and additional sources. This implies that the vortex-flow interactions produce speech that can be treated as a multiple-source signal and the speech production process has a nonlinear character. The presence of additional harmonic series other than the F0-series indicates the emotional state of a speaker and can be used to derive characteristic features for the detection and classification of speech under stress.

In [4] Zhou proposed the use of the Teager Energy Operator for detection of additional harmonics produced due to emotion. The detection was performed at the vowel level; the additional harmonics were searched for around F0, and within Critical Bands. In this paper, a similar approach is proposed. However, the Teager Energy Operator is used to derive the classification feature at the voiced frame level and the search for the additional harmonics is done within the Critical Bands as well as the Wavelet and the Wavelet Packet bands.

2. Speech Analysis Using Teager Energy Operator

2.1. Nonlinear model of speech

In his work on nonlinear speech modelling, Teager[3] indicated modulation as a major process in the production of speech. He also noted the importance of analysing speech signals from the point of view of the energy required to generate them and derived a nonlinear energy-tracking operator known as the Teager Energy Operator (TEO). As indicated in [3], [4] the airflow in the vocal tract is separated into different tracts, each with its own energy. The different tracts include the main air flow through the vocal folds as well as additional vortex-flows generated due to specific emotional states (anger, fear, stress, etc).

The production of a speech signal could be regarded as an effect of amplitude and frequency modulation of separate oscillatory waves in the vocal tract. Therefore, speech signals could be modelled as a combination of several amplitude and frequency modulated (AM-FM) oscillatory components. Maragos et al. [6] proposed a nonlinear model of speech, which represents the sampled speech signal s(n) as:

$$s(n) = \sum_{i=1}^{M} x_i(n) \tag{1}$$

where $x_i(n)$ is a single-component speech signal, and M is the number of speech components. Each component of speech can be modelled as an AM-FM sinewave given as:

$$x(n) = \alpha(n)cos(\Phi(n)) = \alpha(n)cos(\omega_c n+\omega_h \int_0^n q(k)dk+\theta) \tag{2}$$

where q(k) is the modulating signal, ω_c is the source frequency (carrier), $\omega_h \in [0;\omega_c]$ is the maximum frequency deviation, θ is a constant phase offset, and $\alpha(n)$ is the instantaneous amplitude.

2.2. Teager Energy Operator

The presence of the vortex-flow interactions can be detected using the Teager Energy Operator, which in the discrete time domain can be defined as:

$$\Psi[x(t)] = x^2(n) - x(n+1)x(n-1) \tag{3}$$

Substituting Eq(2) into Eq.(3), the Teager Energy Operator of the speech model described by Eq. (3) becomes:

$$\Psi[x(n)] = (\alpha(n))^2 sin(\omega_i^2(n)) \tag{4}$$

The instantaneous frequency $\omega(n)$ of the FM component can be then approximated in terms of the Teager Energy Operator $\Psi[x(n)]$ as:

$$\omega(n) = 2\pi f(n) \approx \frac{1}{2\pi T}\arcsin(\sqrt{\frac{\Psi[y(n)]}{4\Psi[x(n)]}}) \tag{5}$$

and the amplitude $\alpha(n)$ of the AM component can be approximated as:

$$\alpha(n) \approx \frac{2\Psi[x(n)]}{\sqrt{\Psi[y(n)]}} \tag{6}$$

where $y(n) = x(n+1) - x(n-1)$.

2.2.1. *Normalised TEO Autocorrelation Envelope*

If we assume that the speech signal $x(n)$ has a single harmonic with constant amplitude and instantaneous frequency, then the TEO contour $\Psi[x(n)]$ should be a constant number for all values of n. If the speech signal $x(n)$ consists of more than one harmonic, then the TEO contour changes in time and $\Psi[x(n)]$ is a function of n. In reality, speech signals always contain a number of harmonic components. If there is only one speech source or fundamental frequency F0, then there will be a whole harmonic series of integer multiples of F0. Additional sources (vortices) will generate their own harmonic series. However if we break speech into small bands, and calculate the TEO for each band, then we can more easily observe the presence or absence of a harmonic component within each band. This is done through calculation of the polynomial coefficients, which describe the normalised TEO autocorrelation envelope area [4]. The TEO autocorrelation envelope is described as:

$$R_{\Psi(x)}(k) = \frac{1}{2M+1}\sum_{n=-M}^{M} \Psi[x(n)]\Psi[x(n+k)] \tag{7}$$

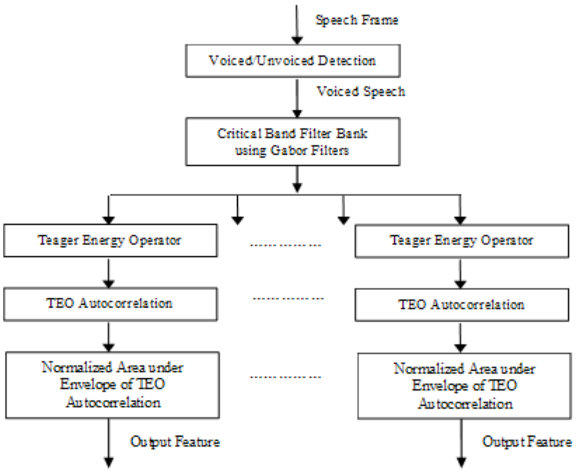
Speech Frame
Voiced/Unvoiced Detection
Voiced Speech
Critical Band Filter Bank using Gabor Filters
Teager Energy Operator
Teager Energy Operator
TEO Autocorrelation
TEO Autocorrelation
Normalized Area under Envelope of TEO Autocorrelation
Normalized Area under Envelope of TEO Autocorrelation
Output Feature
Output Feature

Figure 1: *Flowchart of the TEO-CB analysis.*

where N is the number of samples within the analysed speech frame. By computing the TEO autocorrelation envelope and normalising it by the maximum value of N/2, the normalised TEO autocorrelation envelope area parameters (TEO-Auto-Env) can be obtained for each of the analysed frequency bands. The maximum value of N/2 corresponds to the simplest case with a single harmonic and no amplitude or frequency modulation. The normalised TEO autocorrelation envelope area values reflect the degree of excitation sources variability within each frequency band.

3. Feature Extraction Methods

3.1. Critical Bands Based TEO Analysis (TEO-CB)

It is known that the human auditory system decomposes the whole audible frequency range into several critical bands (see Table 2). The width of these bands increases logarithmically with frequency. Observations of the changes in the numbers of harmonics within critical bands can provide cues for the recognition of stress in speech [1],[4]. The flowchart of the critical band based TEO analysis is illustrated in Fig.1. After the voiced/unvoiced detection, the voiced speech was filtered using a bank of Gabor band pass filters. The filters' centre frequencies were set to the centre frequencies of the critical bands, and the effective RMS bandwidth of each filter was set to the width of the corresponding critical band. For each band the Teager Energy Operator and the area under the normalised TEO autocorrelation envelope were calculated. The analysis was performed on a frame-by-frame basis, with the sampling frequency of 8 kHz and 256 samples per frame with 50% overlap between frames.

3.2. Discrete Wavelet Transform Based TEO Analysis (TEO-DWT)

The Discrete Wavelet Transform (DWT) [8] can be computed by successive lowpass and highpass filtering of the discrete time-domain signal. The DWT was used to analyse the speech signal into a number of dyadic frequency bands listed in Table 1. Four decomposition levels were generated. At each decomposition level, the half band filters produced signals spanning only half the frequency band.

With this approach, very high time resolution of the signal was achieved at high frequencies, while the frequency resolution became very high at the low frequency bands. The analysis was done on the frame-by-frame basis and only the voiced speech was analysed. At each level of analysis, the high pass filters produced detail information, while the low pass filter associated with scaling function produced coarse approximations. The detail information for the decomposition levels 0-3 and the approximation information for level 3 were used to calculate the area under the normalised autocorrelation envelope of the Teager Energy Operator (TEO-DWT-Auto-Env). Frequency components of speech from 300-3000 Hz are known to be the most important for speech intelligibility [5]. It was therefore expected that the high resolution DWT analysis achieved at low frequencies could provide important features representing the emotional aspect of speech.

The TEO-DWT-Auto-Env features were calculated using 8 different types of mother wavelets: haar (db1), db2, db3, db5, bior2.4, bior3.1, bior6.8 and coif1.

Table 1: *DWT bands used in TEO-DWT analysis and WP bands used in TEO-WP analysis.*

Band	Discrete Wavelet Transform			Wavelet Packet		
	Lower	Upper	BW	Lower	Upper	BW
1	0	250	250	0	500	500
2	250	500	250	500	1000	500
3	500	1000	500	1000	1500	500
4	1000	2000	1000	1500	2000	500
5	2000	4000	2000	2000	2500	500
6				2500	3000	500
7				3000	3500	500
8				3500	4000	500

3.3. Wavelet Packet Based TEO Analysis (TEO-WP)

In the search for the optimal sub-band decomposition for the TEO based feature extraction the Wavelet Packet (WP) analysis was tested. The Wavelet Packet method is a modified form of the Discrete Wavelet Transform where the signal is passed iteratively through a larger number of filters than in the DWT. In the DWT, each level is calculated by passing the previous approximation coefficients through high and low pass filters. However, in the WP decomposition both the detail and approximation signals are decomposed. The WP analysis could provide both low and high frequency affect cues. The TEO-WP-Auto-Env features were calculated for the outputs from 8 bands. The corresponding frequency ranges of these bands are listed in Table 1. Like for DWT, the TEO-feature extraction was performed with 8 different types of mother wavelets: haar (db1), db2, db3, db5, bior2.4, bior3.1, bior6.8 and coif1.

3.4. Perceptual Wavelets Packet Based TEO Analysis (TEO-PWP)

The WP decomposition provides a wide selection of bands covering the entire frequency range of speech. Such wide selection could provide a lot of redundant information from the perceptive point of view. The human auditory system shows the highest sensitivity within Critical Bands; therefore selection of WP bands corresponding to the Critical Bands could improve the efficiency of affect detection in speech.

The Wavelet Packet decomposition algorithm was used to select 17 bands with frequency ranges close to the Critical Bands. This type of decomposition was called the Perceptual Wavelet Packet (PWP). The TEO-PWP-Auto-Env features were calculated for the 17 terminal tree outputs with frequency bands listed in Table 2. The analysis was performed for 8 different types of mother wavelets: haar (db1), db2, db3, db5, bior2.4, bior3.1, bior6.8 and coif1.

4. Feature Classifications Using Gaussian Mixture Model (GMM)

The Gaussian Mixture Model [5] method was used to classify the speech samples into two classes: neutral and stressed.

For two classes: neutral and stressed, the class models $\lambda_i$for j=1,2 were estimated using the Expectation Maximization (EM) algorithm [5]. Then for each test utterance, sets $\{\bar{x}_n\}$ of feature vectors were calculated. The probability of each speaker model given the feature sets, i.e., $P(\{\bar{x}_n\}|\lambda_j)$ was calculated using known Gaussian Mixture Model (GMM) pdfs. The class with the highest probability $\max P(\{\bar{x}_n\}|\lambda_j)$ was then identified as a source of the test utterance.

Table 2: *Critical Bands used in TEO-CB analysis and PWP bands used in TEO-PWP analysis.*

Band	Critical Bands			Perceptual Wavelet Packet		
	Lower	Upper	BW	Lower	Upper	BW
1	100	200	100	0	125	125
2	200	300	100	125	250	125
3	300	400	100	250	375	125
4	400	510	110	375	500	125
5	510	630	120	500	625	125
6	630	770	140	625	750	125
7	770	920	150	750	875	125
8	920	1080	160	875	1000	125
9	1080	1270	190	1000	1250	250
10	1270	1480	210	1250	1500	250
11	1480	1720	240	1500	1750	250
12	1720	2000	280	1750	2000	250
13	2000	2320	320	2000	2250	250
14	2320	2700	380	2250	2500	250
15	2700	3150	450	2500	3000	500
16	3150	3700	550	3000	3500	500
17				3500	4000	500

5. Training and Testing Data

The training and testing data were selected from the ”Single Tracking Task” domain of SUSAS database [7]. The speech recordings were made by 9 different speakers and included 35 aircraft communication words. Every word was repeated twice by each speaker, under simulated stressed and neutral conditions, thus generating a total of 1260 speech recordings. Before each run of the classification process, the entire set of 1260 recordings was randomly divided into the training set (1034 recordings) and testing set (126 recordings). Both sets contained 50% of neutral speech recordings and 50% of stressed speech recordings.

6. Subjective Listening Test

A subjective listening test was performed to estimate the human classification level for the purpose of comparison. The test was performed using testing sets of speech recordings selected from the same data that was used in the objective classification. Four listeners (2 male and 2 female) were asked to listen to each of the test words presented in random order and decide if it represented neutral speech or speech produced under stress. The computer played the utterance with 3 second intervals between every word. The test could be paused whenever the subjects wanted to rest. No training program was offered before the test, and no feedback about the classification results was provided during the test. Table 3 contains the confusion matrix resulting from the listening test.

Table 3: *Confusion matrix produced by the listening test.*

Actual Emotion	Classified Emotion	
	Neutral	Stressed
Neutral	86.03%	13.97%
Stressed	14.60%	85.40%

7. Objective Speaker-Independent Classification

The Gaussian Mixture Model classifier was used to classify the test set of words into two classes, neutral and stressed, using 4 different feature selection methods. The classification was performed using feature vectors obtained from TEO-CB-Auto-Env, TEO-DWT-Auto-Env, TEO-WP-Auto-Env and TEO-PWP-Auto-Env. The results are summarized in Tables 4-5. For each method, the classification was run 10 times, each time with different randomly selected training and testing sets. The percentage of correct classifications was calculated as an average of the 10 runs. The classification process had speaker-independent character, the speaker identities were not taken into account.

Table 4: *The percentage of correct classificaiton .*

Method	Correct percentage
TEO-CB-Auto-Env	93.59%

Table 5: *The percentage of correct classificaiton.*

Mother Wavelet	Methods		
	TEO-DWT	TEO-WP	TEO-PWP
db1	90.63%	88.10%	94.13%
db2	94.76%	87.38%	95.95%
db3	92.54%	88.73%	94.68%
db5	90.46%	92.06%	95.48%
bior2.4	94.05%	86.51%	94.52%
bior3.1	91.19%	90.24%	95.32%
bior6.8	89.13%	92.30%	95.79%
coif1	94.44%	90.16%	94.68%

8. Conclusions

Assuming that a speech signal has multi-component character, and the speech production can be modelled as a nonlinear process, characteristic features were derived for the purpose of detecting the presence of stress in speech. One previously known method [4] TEO-CB-Auto-Env, and three newly proposed here feature extraction methods, TEO-DWT-Auto-Env, TEO-WP-Auto-Env and TEO-PWP-Auto-Env, were tested.

The results show a 93.6% correct performance rate for the TEO-CB-Auto-Env method. The newly introduced methods, TEO-DWT-Auto-Env, TEO-WP-Auto-Env and TEO-PWP-Auto-Env, in general show very comparable performance to TEO-CB-Auto-Env with correct classification rates ranging from 86.5% to 95.8%.

As illustrated in Tables 4-5, the worse results were obtained when using the DWT based feature extraction TEO-DWT-Auto-Env. The performance in this case was below the TEO-CB-Auto-Env level for all types of mother wavelets. When the DWT was replaced by WP in TEO-WP-Auto-Env, the performance for some types of mother wavelets become better than for TEO-CB-Auto-Env. Further improvement of performance was achieved by replacing the WP by PWP in TEO-PWP-Auto-Env. The TEO-PWP-Auto-Env method clearly outperformed the TEO-CB-Auto-Env method for all types of mother wavelets and showed consistently the best overall performance out of all four methods tested in this experiment.

It can also be observed that the subjective classification level of about 86% was at the lowest performance level achieved by any of the tested automatic classification methods.

The results indicate that the selection of characteristic features from Critical Bands covering both the low and the high frequency ranges is crucial for the purpose of stress recognition in speech.

9. Acknowledgements

This research was partially supported by the Australian Research Council Linkage Grant LP0776235.

10. References

[1] Ververidis, D. and Kotropoulos, C., ”Emotional speech recognition: Resources, features, and methods”, Speech Communication, 48:1162-1181, 2006.

[2] Ramamohan, S., ”Sinusoidal model-based analysis and classification of stressed speech Sinusoidal model-based analysis and classification of stressed speech”, Audio, Speech, and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on] 14(3): 737-746, 2006.

[3] Teager, H., ”Some observations on oral air flow during phonation”, Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on 28(5): 599-601, 1980.

[4] Zhou, G. and Hanson J.H.L., ”Nonlinear feature based classification of speech under stress”, Speech and Audio Processing, IEEE Transactions on 9(3): 201-216, 2001.

[5] Quatieri T. F., Discrete Time Speech Signal Processing, Principles and Practice, Prentice Hall, 2002.

[6] Maragos P., Keiser J.F. and Quatieri T.F., ”Energy separation in signal modulations with application to speech analysis”, IEEE Trans. Signal Processing, 41:3025-3051, 1993.

[7] Hansen J.H.L. and Sahar E.B-G., ”Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database”, EUROSPEECH-97: Inter. Conf. On Speech Communication and Technology, Rhodes, Greece. 4: 1743-1746, 1997.

[8] Agbinya, J. I., ”Discrete wavelet transform techniques in speech processing”, TENCON ’96. Proceedings. 1996 IEEE TENCON. Digital Signal Processing Applications, 2: 514-519, 1996.