A Hierarchical Approach of Speech Emotion Recognition Based on Entropy of Enhanced Wavelet Coefficients

S. Sultana and C. Shahnaz
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
sharifa_ela@yahoo.com,celia.shahnaz@gmail.com

***Abstract*-This paper presents a hierarchical Speech Emotion Recognition method, where the speaker-independent emotional features are derived from the Teager energy (TE) operated wavelet coefficients of speech signal. The detail as well as approximate Wavelet coefficients enhanced by TE operation is used to compute entropy. Entropy values of TE operated detail and approximate wavelet coefficients reduce feature dimension, forming an effective feature vector for distinguishing different emotions when fed to a Euclidean distance based classifier in a hierarchical process. Detail simulations are carried on EMO-DB German speech emotion database containing four class emotions, such as angry, happy, sad and neutral. Simulation results show that the proposed hierarchical emotion recognition method gives quite satisfactory four-class emotion recognition performance, yet demonstrates a significant increase in versatility through its propensity for speaker independence with lower computation.**

***Keywords*— Wavelet, Teager Energy, Entropy, Euclidean Distance, Speaker-independent, Hierarchical.**

I. INTRODUCTION

Emotions are often considered irrational occurrences. Emotional intelligence is an important part of interpersonal communication. Speech emotion is considered as a method to analyze vocal behavior as a marker of affect (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech. Assuming some objectively measurable voice parameters those reflect the affective state a person is currently experiencing (or expressing for strategic purposes in social interaction). This assumption appears reasonable given that most affective states involve physiological reactions (e.g., changes in the autonomic and somatic nervous systems), which in turn modify different aspects of the voice production process. The field of emotion recognition from voice has recently gained considerable interest in Human-Machine Communication, Human-Robot Communication, and Multimedia Retrieval. Research on speech emotion recognition uses general qualitative acoustic correlations of emotion in speech and statistical properties of certain acoustic features [1], [2]. Most of the emotion recognition processes working in speaker dependent systems has recognition rates starting from 70% to 90% [3],[4]. Among different speaker-independent emotion recognition systems [5],[6], the method in [5] extracted 87 static features from the dynamic features of energy, pitch, and spectral features. The ratio of a spectral flatness measure to a spectral center (RSS) is proposed as a speaker-independent feature in [6]. Here, gender and emotion are hierarchically classified by using the feature RSS, pitch, energy, and mel-frequency cepstral coefficients. In general, speaker-independent systems show a lower accuracy rate compared with speaker dependent systems, as emotional feature values depend on the speaker and their gender.

This paper deals with hierarchical classification over an effective feature for speech emotion recognition with a strategy to decrease the speaker variation in the emotional features. The proposed method with no prior gender classification follows a classification process and provides higher accuracy with lesser complexity even with a simple Euclidean distance based classifier in comparison to an existing speaker independent feature based hierarchical method of emotion recognition.

II. PROPOSED METHOD

The proposed speech emotion recognition method consists of two major steps, namely feature extraction and hierarchical classification.

A. Feature Extraction

Emotion recognition accuracy persuasively depends upon the quality of the extracted features. The proposed feature extraction algorithm is shown in Fig. 1, where Discrete wavelet Transform (DWT) is first employed to each input speech frame. Both the detail and approximate DWT coefficients are subject to Teager Energy (TE) approximation. The TE operated enhanced coefficients are then employed to compute entropy that forms the feature vector.

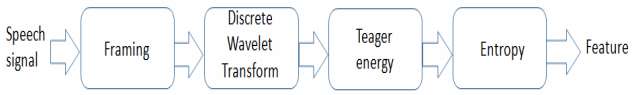

Fig.1: Proposed feature extraction algorithm

1) *Discrete Wavelet Transform*

Wavelet-transformation contains information similar to the short-time-Fourier-transformation, but with additional special properties of the wavelets, which show up at the resolution in time at higher analysis frequencies of the basis function. The difference in time resolution at ascending frequencies for the Fourier transform and the wavelet shows that wavelet transformation is good in time resolution of high frequencies, while for slowly varying functions, the frequency resolution is remarkable. For this reason, DWT can be considered as a potential feature in speech emotion recognition. For analyzing both the low and high frequency components of emotional speech, it is passed through a series of low-pass and high-pass filters with different cut-off frequencies through DWT. The filtering operations in DWT result in a change in the signal resolution, whereas sub sampling (down sampling/up sampling) causes change of the scale. Thus, DWT decomposes the speech signal into approximate and detail coefficients thereby helping in analyzing it at different frequency bands with different resolutions.

The emotion of a speech does not depend on the speech or the speaker. So, emotion should be a totally independent part of the speech apart from the information content of the speech that is represented by the approximate coefficients. It is found that details coefficients of emotional speech remain stationary compared to the approximation coefficients. Detail and approximate coefficients of speech signals with sad emotion are obtained and shown in plot of Fig. 2, where three emotional speech signals uttered by male and female speakers are considered. From this plot, it is clear that most of the detail coefficients remain almost compact for a particular class of emotion, being independent of speaker and speech, whereas the most approximate coefficients vary within a class of emotion. Thus, it is argued that detail coefficients play a dominant role over approximate coefficients in distinguishing different emotions. However, approximate DWT coefficients are also essential for speech-based emotion recognition. The reason behind this can be explained by considering the presence of emotional words in the speech. Positive emotional speech most of the time contains joyful words while the negative ones have joyless words in general. No matter what the word is, the emotional words are found to create an impact on approximate DWT coefficients. So, while forming feature vector for emotion recognition from speech, taking into account approximate DWT coefficients is expected to increase accuracy in the recognition process.

2) *Teager Energy Approximation*

The amplitude of detail DWT coefficients is very small as is vivid from the plots of Figs. 2. So, in particular, the energy of detail coefficients is needed to be enhanced.

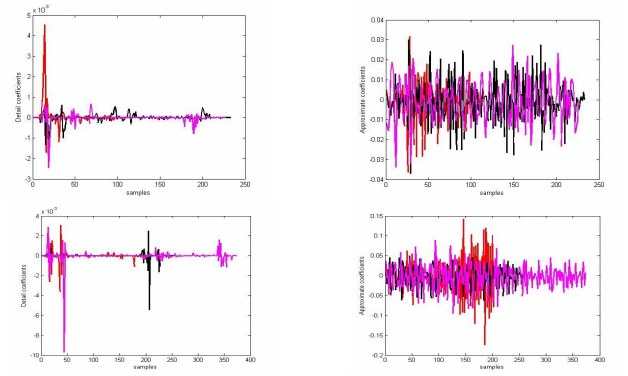

Fig. 2: Detail (left) and approximate (right) DWT coefficients of speech signals with sad emotion uttered by male (top) and female (bottom) speakers.

With a view to enhance the energy of DWT coefficients, TE operation on them is performed as

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

where a band-limited discrete signal $x[n]$ is approximated by TE operator [7]. In Fig. 3, TE of approximate and detail DWT coefficients of three speech signals with sad emotion are presented for male and female speakers, respectively. From these plots, it is clear that the amplitudes of detail and approximate DWT coefficients have been much enhanced in comparison to that of the DWT coefficients prior to TE operation as shown in Fig. 2.

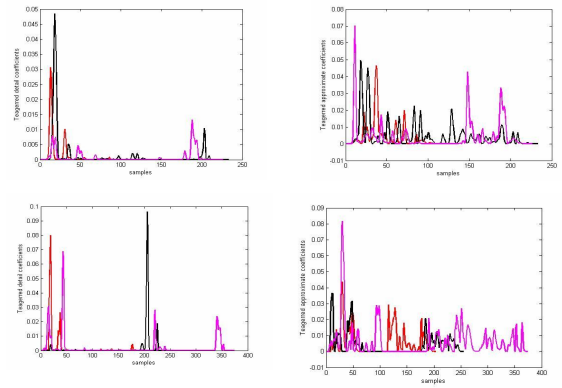

Fig. 3: TE of approximate and detail DWT coefficients of three speech signals uttered by male (top) and female (bottom) speaker with sad emotion.

The complexity of feature extraction and ease of its implementation is another important criterion while developing a feature vector for multiclass emotion recognition. In our case, if all the TE operated DWT coefficients are considered to form the feature vector, it would definitely result in a feature vector with a very large dimension. In view of reducing the feature dimension, we propose to utilize entropy of the detail and approximate DWT coefficients.

3) *Entropy*

Entropy is the average unpredictability in a random variable, which is equivalent to its information content. Entropy of a signal can be defined as

$$E = -\sum_{i=1}^{N} p(x_i)\log_2 p(x_i), \qquad (2)$$

where $p(x_i)$ is the probability mass function of random variable x_i. Entropy depends crucially on the probabilistic model. On the other hand, energy depends on the magnitude of the signal only. The motivation behind computing entropy of DWT coefficients instead of energy of those coefficients can be justified in terms of compactness property for the same class emotion and separability property for the different class emotions.

In the left column of Fig. 4, entropy values of TE operated detail DWT coefficients of different speech signals of different emotions have been plotted. Since detail DWT coefficients are found significant in emotion recognition, they are considered for convenience of pictorial representation. From this plot, an area for every emotion can be estimated, which is almost non-overlapping among different classes of emotion. On the other hand, in the right column of Fig. 4, energy values of TE operated detail DWT coefficients of different speech signals of different emotions have been displayed. From this plot, no individual area for a particular emotion class can be estimated in terms of energy. Therefore, it is found that the entropy for different signals of a particular emotion shows better resembles in terms of their values, remains more clustered and provides stronger

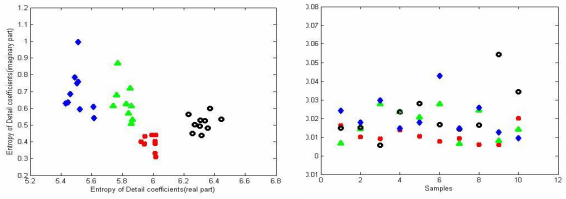

Fig. 4: (left column) Entropy and (right column) energy of TE operated detail DWT coefficients of 10 different speech signals with different emotions uttered by both female and male speakers. The green triangles denote happy emotion, red asterisks represent angry emotion, blue diamonds symbolize neutral emotion and black circle stands for sad emotion.

compactness as a feature while the energy for different signals of the same class emotion becomes scattered. This figure also attests that entropy is capable of providing the higher separability among different classes of emotion. Therefore, we prefer entropy over energy while computing the feature vector for multiclass emotion recognition. Considering the TE operated detail and approximate DWT coefficients, the proposed feature vector can be formed as

$$F = [A \quad B] \qquad (3)$$

where, A represents the entropy of TE operated approximate DWT coefficients, and B stands for the entropy of TE operated detail DWT coefficients.

B. *Hierarchical Classification:*

In this paper, speaker-independent emotion recognition is accomplished using the hierarchy method of classification rather than the direct nonhierarchical method which is commonly used. In general, humans have different vocal systems in terms of factors, such as shape and size. This causes variations of emotional features from speaker to speaker. For a speaker-independent system, the decrease of this variation is an important issue. However, in general, the difference of the vocal system between a male and a female is too large to ignore. Hence, in this paper, our feature being gender independent, there is no need of gender detection for the emotion recognition. Again, most emotional features are related to arousal. This suggests that the most emotional features confuse anger and happiness and sadness and neutrality as valence-related emotions. However, as is clear in Fig.4 that for the TE operated wavelet coefficients, anger and happiness or sadness and neutrality are located far from each other, while anger and sadness are close. Hence, the proposed feature is suitable for distinguishing sadness from neutrality and anger from happiness. Also, previous studies have shown that a hierarchical classification method achieves better performance than considering the features from all classes [8]-[10]. Thus, a hierarchical classification method is used in this paper as shown in Fig.5. First, the proposed feature is used to separate anger and neutrality (group 1) from happiness and sadness (group 2). Finally, group 1 is classified into anger and neutrality, and group 2 into happiness and sadness using the same feature. In both the cases, classification and speech emotion recognition is carried out based on a similarity measure defined as the Euclidean distance between the feature vectors of the training speech frames and those of the test speech frames.

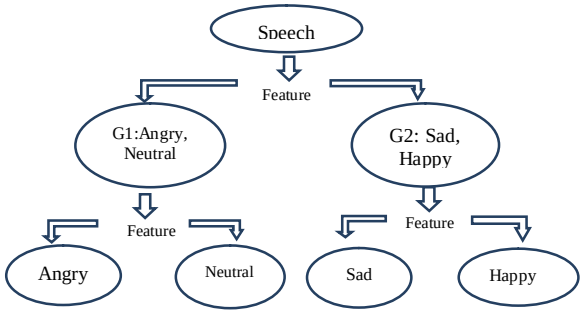

Fig.5: Hierarchical Speech Emotion Recognition Tree

For any particular class of emotion, q number of speech frames is considered for the purpose of recognition. Given the m-dimensional feature vector for the k^{th} frame of the j^{th} class of emotion be $\{\gamma_{jk}(1), \gamma_{jk}(2), ..., \gamma_{jk}(m)\}$ and the f^{th} test sample

with a feature vector $\{v_f(1), v_f(2), ..., v_f(m)\}$, a similarity measure between the test sample f of the unknown emotion class and the sample frame of the j^{th} emotion is defined as

$$D_j^f = \sum_{k=1}^{q} \sum_{i=1}^{m} |\gamma_{jk}(i) - v_f(i)|^2, \quad (4)$$

where, a particular class represents an emotion with q number of speech frames. Therefore, according to (4), given the f^{th} test frame of speech, the unknown emotion is classified as the emotion j among the r number of classes when

$$D_j^f \leq D_g^f, \forall j \neq g \text{ and } \forall g \in \{1,2,.....,r\}. \quad (5)$$

In this paper, we are interested to handle four types of emotion thus solving a four-class problem (r = 4). The use of Hierarchical Tree classifier instead of Euclidean distance based classifier is expected to improve the recognition accuracy.

III. SIMULATION RESULTS

A. Database and Simulation Conditions

Speech sentences from the German emotion database, known as EMO-DB are employed for the experiments, where the speech data is sampled at 16 KHz. From the EMO-DB database, 55 speech data of angry, 33 of happy, 29 of sad and 33 of neutral are randomly selected for training and 179 non overlapping speech data (none of them is used to form the training matrix) consisting of data from each class is then tested for emotion recognition. More analysis can be done with different emotion databases in English, which is a potential future work of this research.

In order to obtain overlapping analysis frames of the speech sentences from the German emotion database, hamming windowing operation is performed, where the size of each of the frame is 30 ms with 50% overlap between successive frames.

A 1-level wavelet decomposition with db2 bases function is applied on the speech frames. In this paper, the feature vector is of two component, component "A" (corresponds approximate coefficients) and "B" (corresponds detail coefficients), among which "B" is the dominating one. If the level of DWT increases, then the number of component in feature vector increases, which not only increases the computational burden but also diminishes the dominating factor of speech emotion recognition, resulting in a less accuracy in recognition.

B. Speaker Independency of the Proposed Feature

From detail analysis, it is found that for each gender group (male or female), in terms of mean of the proposed entropy values, the difference between the speakers is much smaller than the difference between the two emotions, namely neutrality vs. sadness, and anger vs. happy. Such analysis is summarized in Table I for different speakers from the database, where M1, M2, M3 etc. and F1, F2, F3 etc. denote male and female speakers, respectively. For example, for a male group, in case of neutrality, although the difference of mean of the proposed feature values among different speakers is 0.01-.02, such difference in values between neutrality and sadness ranges from 0.29 to 0.32, which is a 30 times increase compared with the difference in values in case of different speakers. Therefore, the proposed feature is found to be highly effective for speaker-independent emotion recognition.

Table I: SPEAKER INDEPENDENCY OF THE PROPOSED FEATURE

Mean	Neutral	Sad	Anger	Happy
M1	5.66	6.02	5.94	6.02
M2	5.64	6.13	6.06	6.04
M3	5.62	6.10	5.98	5.90
M4	5.64	6.06	5.99	5.93
F1	5.83	6.2	6.07	5.92
F2	5.85	6.15	6.00	6.06
F3	5.80	6.2	6.06	6.10
F4	5.80	6.2	5.96	5.93

C. Performance Evaluation:

For the purpose of comparison, we use state-of-the-art speaker independent feature RSS proposed in [6]. For the performance evaluation of the proposed and comparison methods, criterion considered in our simulation study is accuracy derived from confusion matrix. The rows in the confusion matrix stand for the actual emotion classes to be tested and columns provide the emotion class classified by a method.

The confusion matrix derived for the speaker-independent RSS feature based hierarchical method using Euclidean distance classifier is represented in Table II. In this table, the diagonal entries stand for the number of cases when a particular class of emotion is correctly classified. It can be seen from Table II that RSS based method is able to recognize anger, happy, sad and neutrality with accuracies of 79.7%, 35.3%, 57.6% and 81.3%, respectively. No gender detection is performed and in this hierarchical process of classification, the overall recognition accuracy of 67.6% is obtained. For nonhierarchical classification, the overall accuracy based on this feature is obtained as 56.98%.

Table III shows the confusion matrix derived for the speaker-independent proposed hierarchical method using Euclidean distance classifier. It is demonstrated from Table III that the proposed method is able to recognize anger, happy, sad and neutrality with accuracies of 84.05%, 51.439%, 72.73% and 74.42%, respectively. Here, overall recognition accuracy of 73.33% is achieved with no prior gender classification, which is higher compared to the method using RSS feature. For nonhierarchical approach, the overall accuracy based on the proposed feature is 63.63%, which shows the effectiveness of employing hierarchical method for classification.

Table II: CONFUSION MATRIX FOR THE RSS FEATURE BASED HIERARCHICAL METHOD USING EUCLIDEAN DISTANCE CLASSIFIER

	Angry (%)	Happy (%)	Sad (%)	Neutral (%)
Angry	79.7	11.6	5.8	2.9
Happy	44.1	35.3	14.7	5.8
Sad	9.1	15.2	57.6	18.2
Neutral	2.3	6.9	9.3	81.3

Overall Accuracy=67.6%

TABLE III: CONFUSION MATRIX FOR THE PROPOSED FEATURE BASED HIERARCHICAL METHOD USING EUCLIDEAN DISTANCE CLASSIFIER

	Angry (%)	Happy (%)	Sad (%)	Neutral (%)
Angry	84.05	7.2	5.8	2.9
Happy	25.7	51.43	5.7	17.1
Sad	9.09	18.18	72.73	0
Neutral	25.58	0	0	74.42

Overall accuracy = 73.33%

IV. CONCLUSION

In this paper, a method is proposed with a strategy to obtain greater accuracy of speech emotion recognition along with a view to decrease the speaker variation in the emotional features. The proposed feature based on entropy of enhanced wavelet coefficients is not only speaker-independent but also capable of distinguishing different emotions even using simple nonhierarchical method with Euclidean distance based classifier. But the classification accuracy improves to a very significant level if the classification is done in a hierarchical process while comparing to one of the state-of-the art method using speaker-independent feature and hierarchical approach of classification as shown in detail simulation results.

REFERENCES

[1] U. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates," *JASA*, vol. 52, no. 4, pp. 1238–1250, 1972.

[2] F. J. Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameter," *J. Exp. Psychol.: Hum. Percept. Perform.*, vol. 12 no. 3, pp. 302–313, Aug. 1986.

[3] P. Y. Oudeyer, "The production and recognition of emotions in speech: feature and algorithms," *Int. J. Hum.Comput. Stud.*, vol. 59, no. 1, pp. 157–183, 2003.

[4] B. Schuller, R. J. Villar, G. Riqoll, and M. Lang, "Meta-classification in acoustic and linguistic feature fusion-based affect recognition," *in Proc.* IEEE Int. Conf. Acoust., Speech Signal Process, Mar. 2005, vol. 1, pp. 325–328.

[5] D. Ververidis and C. Kotropolos, "Automatic emotional speech classification," *in Proc.* IEEE int. Conf. Acoust., Speech Signal Process., May 2004, vol. 1, pp. 593–596.

[6] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved emotion recognition with a novel speaker-independent feature", *IEEE/ASME Trans. on Mechatronics*, vol. 14, no. 3, pp. 317-325, 2009.

[7] T. F. Sanam, and C. Shahnaz , "Noisy speech enhancement based on an adaptive threshold and a modified hard thresholding function in wavelet packet domain," *Digital Signal Processing, Elsevier*, vol.23, pp. 941-951, 2013.

[8] Communicative intent in robot-directed speech," Auton. Robots, vol. 12, no. 1, pp. 83–104, Jan. 2002.

[9] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "A hierarchical framework for speech emotion recognition," IEEE Int. Symp. Ind. Electron.,vol.1, pp. 515–519, Jul. 2006.

[10] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech emotion recognition based on a fusion of all-class and pairwise-class feature selection,"Lecture Notes Comput. Sci., vol. 4487, pp. 1611–3349, Jul. 2007.