RECOGNITION OF EMOTION IN SPEECH USING VARIOGRAM BASED FEATURES

Zeynab Esmaileyan[1] , Hosein Marvi [2]

[1] Department of Electrical engineering, Science and Research branch, Islamic Azad University, Shahrood, IRAN
[2] Electrical Engineering Department, Shahrood University, Shahrood, IRAN

Email: [1] z.esmaileyan@gmail.com, [2] h.marvi@shahroodut.ac.ir

ABSTRACT

Speech Emotion Recognition (SER) is a relatively new and challenging branch in speech processing area. In this study, we propose new features derived from speech spectrogram using image processing techniques for emotion recognition. For this purpose, variogram graphs are calculated from speech spectrogram. The significant Discrete Cosine Transform (DCT) coefficients of variogram are used as proposed features. The contribution of these features as a complementary for the widely used prosodic and spectral features is also investigated. The feature selection is performed using Fisher Discriminant Ratio (FDR) filtering method. Finally, a linear Support Vector Machine (SVM) classifier is employed. All results are achieved under the 10 fold cross-validation on the Berlin and PDREC speech databases. Our results show that combining the proposed features with prosodic and spectral features significantly improves the classification accuracy. For Berlin database, when the proposed features were added to the prosodic and spectral ones, the recognition rates were improved from 83.18% and 89.36% to 86.82% and 90.43% for females and males, respectively. Also, on the PDREC, combining the proposed features with the prosodic and spectral features improve the recognition rate of females and males by 3.72% and 0.27%, respectively. For this database, the best classification accuracy of 63.18% and 57.37% were obtained for females and males, respectively.

Key words: Speech emotion recognition; speech spectrogram; variogram, Berlin speech database; PDREC speech database.

1.0. INTRODUCTION

As speech is vital for human communication, a large number of systems with the aim of recognizing human's speech have been developed. However, since machines are unable to recognize human's emotion, natural interaction between human and machine is not achievable. . This has introduced a relatively new and challenging field of research known as the Speech Emotion Recognition (SER). SER has a wide range of application in man-machine interaction. It can be used as a complementary part of Automatic Speech Recognition (ASR) systems to improve their performance [1]. It is also useful in medicine and psychology [2,3]. Furthermore, SER can be used in e-learning, computer games and other interactive systems [4].

Most acoustic features employed for SER can be grouped into two main categories: prosodic and spectral [5]. Prosodic features, which are commonly extracted based on statistics of pitch, energy and timing, have been extensively used in SER [6–11]. These features convey important hints about speaker's emotional state [6–11]. Spectral features, on the other hand, provide complementary information for prosodic features [5]. Features based on formants [12–15], Mel Frequency Cepstral Coefficients (MFCCs) [12,14,16,17], and Perceptual Linear Prediction (PLP) [18], have been reported as effective spectral features in SER.

Fig.1 shows a speech signal uttered by a male speaker under two different emotions: anger and boredom. Corresponding pitch and logarithmic energy tracking contours are also depicted in this figure. As can be seen from this figure, anger utterance has a higher average pitch and wider pitch range than boredom. Also, variation in the energy tracking contour for anger is more than boredom.

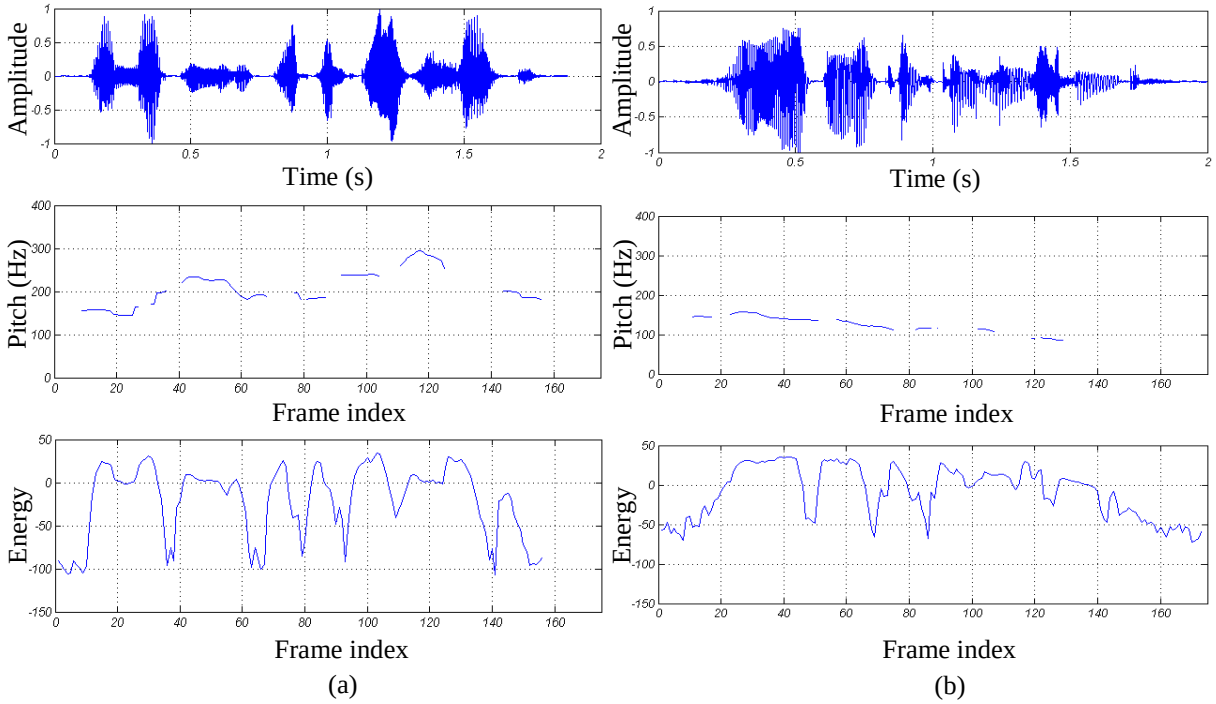

Fig.1. A speech signal uttered by a speaker under two different emotions (a) anger, (b) boredom, and their corresponding pitch and energy tracking contours.

Recently, time-frequency based speech processing methods such as spectrogram analysis have received increased attentions [19,20]. Spectrogram is a graphical display of the time-frequency energy distribution of speech. Methods based on spectrogram analysis have the advantage of preserving important underlying dependencies between different time and frequency contents [21], which can provide features conveying prosodic and spectral information simultaneously. Also, the dynamical behavior of speech can be effectively captured by analyzing the spectrogram. These facts have motivated researchers to use spectrogram as a practical speech analysis tool [22]. For example, Authors in [23] reported that the speech spectrogram contains rich information about energy, pitch, formants and timing that could be valuable in speech and speaker recognition. Spectrogram was also employed for speech perception [24]. Furthermore, features based on Radon and discrete cosine transforms derived from speech spectrogram have been reported to be efficient in speaker identification [21]. Spectral Patterns (SPs) and Harmonic Energies (HEs) derived from speech spectrogram have been also employed for emotion recognition [19].

The main contribution of this study is to propose features derived from Discrete Cosine Transform (DCT) of variogram curves extracted from speech spectrogram for emotion recognition. The proposed features are compared with the widely used prosodic and spectral features in terms of classification accuracy. We also evaluated the proposed variogram based features as complementary to the prosodic and spectral features.

The present study is organized as follows. Section 2 describes the proposed features as well as prosodic and spectral ones employed here as a benchmark. Section 3 introduces the databases employed. Section 4 presents the results of experimental evaluation. The paper ends with concluding remarks in section 5.

2.0. FEATURE EXTRACTION

In this section, we detail the proposed features extracted from the spectrogram of speech. Prosodic and spectral features considered here are also described. The prosodic and spectral features calculated in this work serve as a representative sampling of essential features in order to provide a benchmark, and also, to evaluate the proposed features as complements to the widely used prosodic and spectral features.

2.1. Pre-processing

The silent part of the speech signal is firstly removed by a Voice Activity Detection (VAD) algorithm [25]. Then, the speech signal x[n] is pre-emphasized as [26]:

$$\hat{x}[n] = x[n] - \alpha x[n-1], 0.9 \le \alpha \le 1 \quad (1)$$

Where $\hat{x}[n]$ represents the pre-emphasized speech signal. The most common value for α is around 0.95 [27]. Due to the non-stationary nature of speech, it is commonly divided into short duration frames (20–30 ms) wherein the speech signal remains approximately stationary [10]. Also a shift of 10 ms between consecutive frames is necessary to retain a good quality of the signal and to avoid loss of information [27,28]. Windowing is carried out to reduce the edge effects at the beginning and the end of the frames [26], and the most common windowing in SER is Hamming window multiplied with each frame.

2.2. Proposed features

Firstly, N=512 length Discrete Fourier Transform (DFT) of each windowed frame is computed to obtain the logarithmic power spectrum as:

$$S_i(k) = \log_{10}((\mathrm{Re}\{\tilde{X}_i(k)\})^2 + (\mathrm{Im}\{\tilde{X}_i(k)\})^2), k = 0,1,...,N-1, i = 1,2,...,M \quad (2)$$

Where *M* is the total number of frames. $\tilde{X}_i(k)$ denotes the kth component of DFT of $\tilde{x}_i(n)$ (ith windowed frame). Re{…} and Im{…} are real and imaginary parts, respectively. The spectrums of the frames, $S_i(k)$, are concatenated row-wise to construct the speech spectrogram, *f*, as:

$$f = \begin{bmatrix} S_1(0) & \dots & S_M(0) \\ \vdots & \ddots & \vdots \\ S_1(N-1) & \dots & S_M(N-1) \end{bmatrix} \quad (3)$$

Where *f(k,i)* (kth row and ith column of the spectrogram *f*) represents the logarithmic squared magnitude of the kth frequency component of the ith frame of speech signal. i and k indicate the time and frequency indexes respectively.

The size of speech samples are fixed to a predefined value of 3s, so every spectrogram image is made with the same size. To this end, shorter samples are repeated and longer ones are trimmed.

Fig.2 (a) to (c) show the spectrograms of 3 sentences expressed under 3 different emotional states. These sentences are chosen from the Berlin database [29]. As can be seen from these figures, characteristics of speech spectrogram deeply correlate with speaker's emotion. While in high activation emotions such as anger, higher harmonics are stronger, while in low arousal emotions such as sadness, most of the energy is concentrated in lower harmonics. Furthermore, in low arousal emotions, energy bands in spectrogram look more stable and closer to each other.

Variogram is a useful tool for analyzing texture patterns. Considering the spectrogram as a texture, we use variogram to extract emotional features from speech. Variogram is defined as the variance of the difference between 2 random variables [30,31]:

$$2\gamma(d) = E\left\{\left[z(x) - z(x+d)\right]^2\right\} \quad for\ all\ x, x+d \tag{4}$$

Where z(x) denotes the random variables, *x* and *d* are sampled location and the distance between sampled locations, respectively. E{…} denotes the expected value. In practice, the variogram is computed in units of distance (e.g. pixel in image processing).

For the spectrogram image, f, the vertical variogram (90º variogram) is calculated as:

$$\gamma_{90}(d) = \frac{1}{N(d)} \sum_{j=1}^{M} \sum_{i} \left[f(i,j) - f(i+d,j) \right]^2 \tag{5}$$

Where f(i,j) is the i^{th} row and j^{th} column of spectrogram image. M is the number of columns in the spectrogram image (number of speech frames), and N(d) is the number of pairs within the spectrogram image contributed in the variogram calculation. Correspondingly, the horizontal variogram (0º variogram) is determined as:

$$\gamma_{0}(d) = \frac{1}{N(d)} \sum_{i=1}^{N} \sum_{j} \left[f(i,j) - f(i,j+d) \right]^2 \tag{6}$$

Where N denotes the number of rows in the spectrogram image (number of frequency components). In fact, the variogram can be considered as a tool to measure the difference between grade values relative to the distance separating them in a particular orientation. The variogram of the spectrogram images which are depicted in Fig.2 (a) to (c) are shown in Fig.2 (d) to (f), respectively.

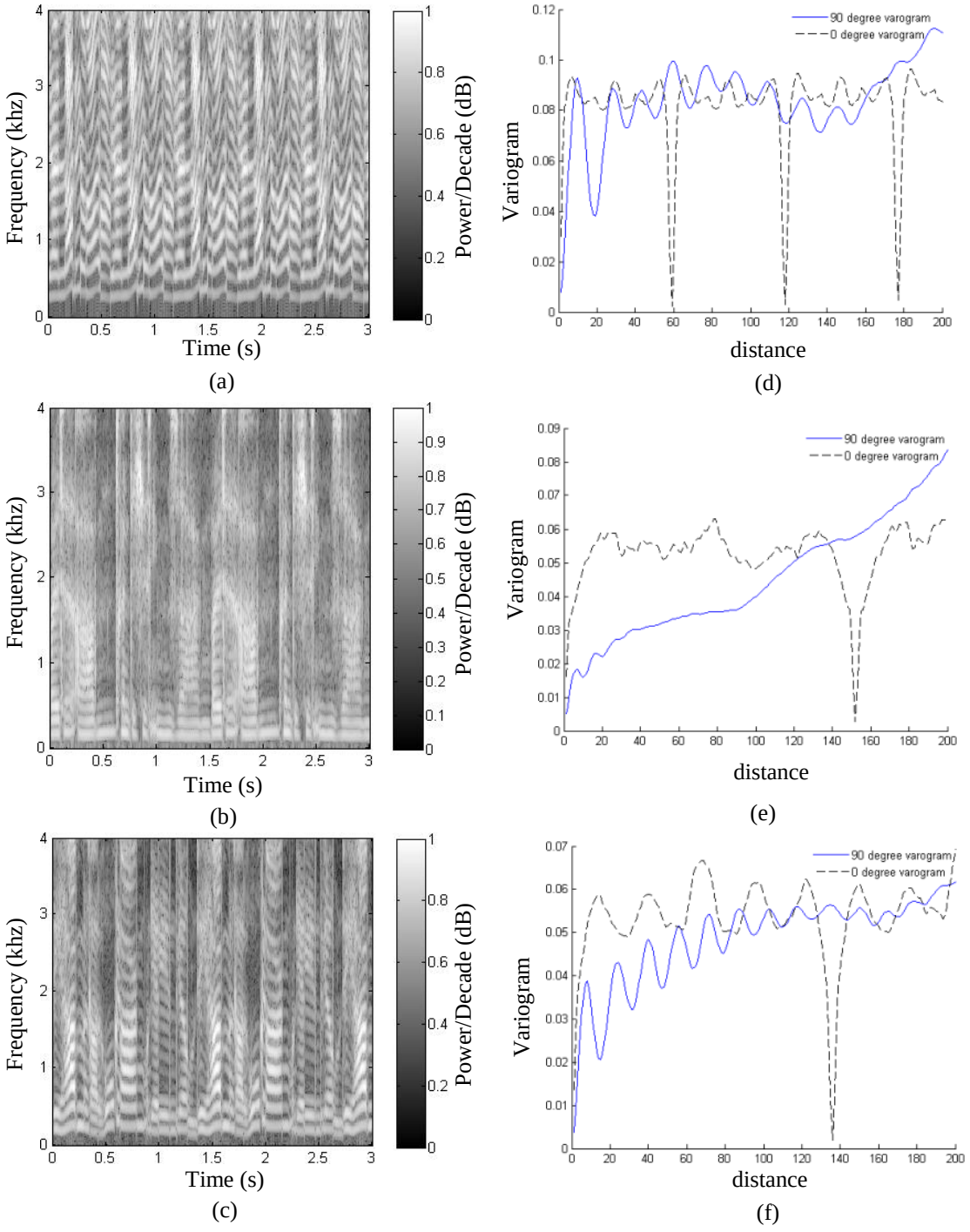

Fig.2. Spectrogram of 3 utterances expressed by emotions (a) anger, (b) sadness, and (c) neutral with the corresponding variograms in (d), (e) and (f), respectively.

The behaviors of the variogram graphs are meaningful and informative for speech emotion recognition. Since in the spectrogram image the horizontal and vertical axes represent time and frequency indexes respectively, 0º and 90º variograms convey timing-based and frequency-based information of speech signal, respectively. In 0º variogram, the valleys represent the time periodicity of speech which is originated from time normalization of speech described in section 2.1. In 90º variograms, the valleys show the frequency periodicity of speech signal (the vertical periodicity of spectrogram image). It can be considered as a rough indicator for average distance between harmonics which is highly correlated with mean value of fundamental frequency (pitch) [19]. In summary, variogram graphs convey important information about speaker's emotional state which makes them useful to extract features for emotion recognition.

The DCT is employed to compact the information of the variogram contours. Finally, significant DCT coefficients (the first 200 coefficients) of each variogram form the proposed features. So, there are 400 features for each input sample. The block diagram of proposed features extraction is shown in Fig.3.

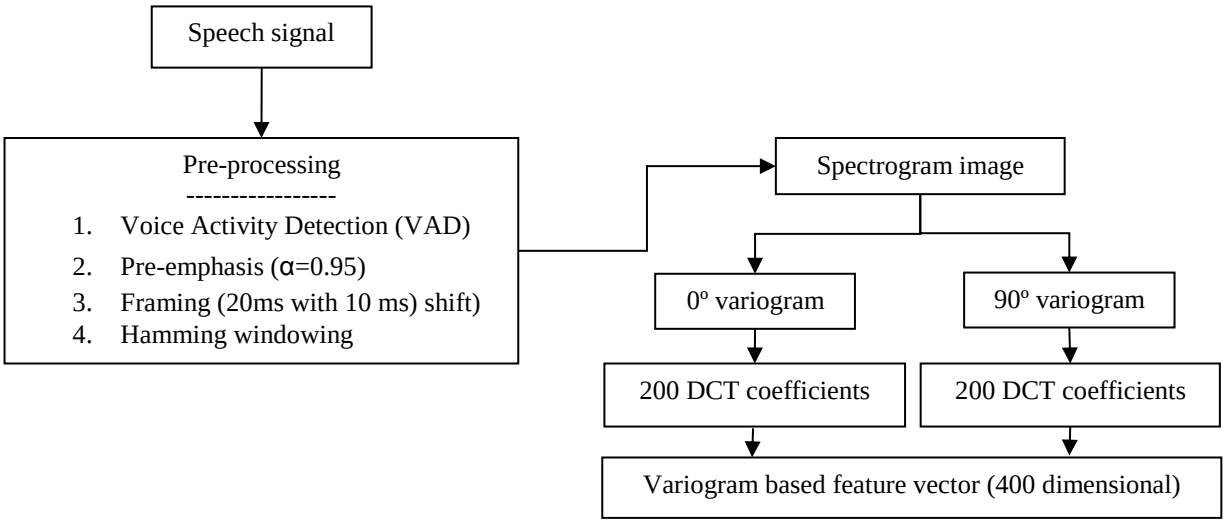

Fig.3. The process of extraction of variogram based features from speech signal.

2.3. Prosodic Features

Here, the 20 time domain statistical functions proposed by [5,32,33] are employed to extract features of pitch, energy and zero-crossing rate (ZCR) tracking contours. These functions include: min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentile, interquartile range, mean average deviation, standard deviation, skewness and kurtosis. As a common practice, the features are also extracted from the first and second derivatives of the contours [5,32,33]. Voiced to unvoiced duration ratio is also employed. In total, we extract $3\times 20\times 3+1=181$ prosodic features here.

2.4. Spectral Features

We employ 3 types of spectral features: The MFCC, PLP and formant. These are reported to be effective in SER [12–18]. After the pre-processing stage, the first 12 MFCCs, 13 PLPs and 4 formants are extracted from 20 ms Hamming-windowed speech frames every 10 ms. Statistical features are extracted using the 20 function described in

section 2.3. Features are also extracted from the first and second derivatives. In total, $29 \times 20 \times 3 = 1740$ spectral features are extracted here. Table 1 lists the features employed in this study.

Table 1. List of employed features.

Proposed features: (400 features)	*First 200 DCT coefficients of:* 0º variogram graph; 90º variogram graph;
Prosodic features: (241 features)	*Apply 20 statistical functions to:* Pitch, delta pitch, double-delta pitch; Energy, delta energy, double-delta energy; ZCR, ZCR delta, and ZCR-double delta; Ratio between the duration of voiced and unvoiced speech;
Spectral features: (960 features)	*Apply 20 statistical functions to:* 12 MFCCs, their deltas, and their double-deltas; 4 formants, their deltas, and their double-deltas; 13 PLP, their deltas, and their double-deltas;

The 20 statistical functions include: *min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentile, interquartile range, mean average deviation, standard deviation, skewness, kurtosis*.

3.0. EMOTIONAL SPEECH DATABASES

3.1. Berlin database

Berlin database of German emotional speech [29] have been extensively used in relative researches. It consists of 535 utterances uttered by 10 professional actors (5 males and 5 females) in 7 different emotional states. Table 2 details the number of speech utterances for each emotion category.

Table 2. Number of utterances in the Berlin database.

Emotional state	**anger**	**boredom**	**disgust**	**fear**	**joy**	**neutral**	**sadness**	**Total**
female	67	46	35	32	44	40	37	301
male	60	35	11	37	27	39	25	234
total	127	81	46	69	71	79	62	535

3.2. Persian Drama Radio Emotional Corpus (PDREC)

The Persian Drama Radio Emotional speech Corpus (PDREC) [34], which is introduced in our previous work, consists of 748 samples (339 female and 409 male samples) manually selected from the radio programs. The utterances expressed by 33 professional actors (15 females and 18 males) in 8 emotions. Factors such as natural recording environments, background noise such as car sound, sound of dripping water and children's playing sound have brought similarities to the naturalness of this database [34]. Table 3 describes the characteristics of this database.

Table 3. Number of utterances in the Persian Drama Radio Emotional speech Corpus (PDREC).

Emotional state	**anger**	**boredom**	**disgust**	**fear**	**joy**	**neutral**	**sadness**	**Total**
female	73	15	12	21	36	88	78	16
male	104	14	5	42	40	106	67	23
total	177	29	17	63	76	194	145	39

4.0. EXPERIMENTS

Linear Support Vector Machine [35] is employed to classify 5 emotions categories: anger, fear, joy, neutral and sadness of the Berlin and PDREC databases. A perfect accuracy gender classifier proposed by [36], is assumed to be utilized beforehand [5]. Thus, the experiments are performed separately for males and females. Before the classification, training features are linearly scaled to (- 1, 1), and then testing features are scaled using the trained linear mapping function [11,37].

Irrelevant features are removed using Fisher Discriminant Ratio (FDR) criterion. This can quickly filter the features to avoid the curse of dimensionality [38]. The FDR for the uth feature can be computed as [11]:

$$FDR(u) = \frac{2}{C(C-1)} \sum_{c_1} \sum_{c_2} \frac{(\mu_{c_1,u} - \mu_{c_2,u})^2}{\sigma^2_{c_1,u} + \sigma^2_{c_2,u}}, 1 \le c_1 < c_2 \le C \quad (7)$$

Where $\mu_{c_i,u}$ and $\sigma^2_{c_i,u}$ denote the mean and variance of the u^{th} feature of the i^{th} class, $i = 1, 2, \ldots, C$, and C is the total number of classes. The less discriminative features (with small FDR values) are removed during thresholding process. While the filter based methods such as FDR have the advantage of computational simplicity, they suffer from shortfalls to take into account the features correlation and classifier properties [11,5]. However, we apply only the FDR feature selection algorithm in our experiments as it satisfactorily performs the selection task. In this work, all results are achieved using 10 fold cross-validation. In this technique, samples in each class are randomly divided into 10 non-overlapping subsets which roughly have the same size. 1 subset from every class are retained for testing (testing dataset), and the remaining 9 are kept for training (training dataset) [11,39]. The process iterated 10 times to contribute all the 10 subsets in evaluation. The proposed SER system is schematically shown in Fig.4.

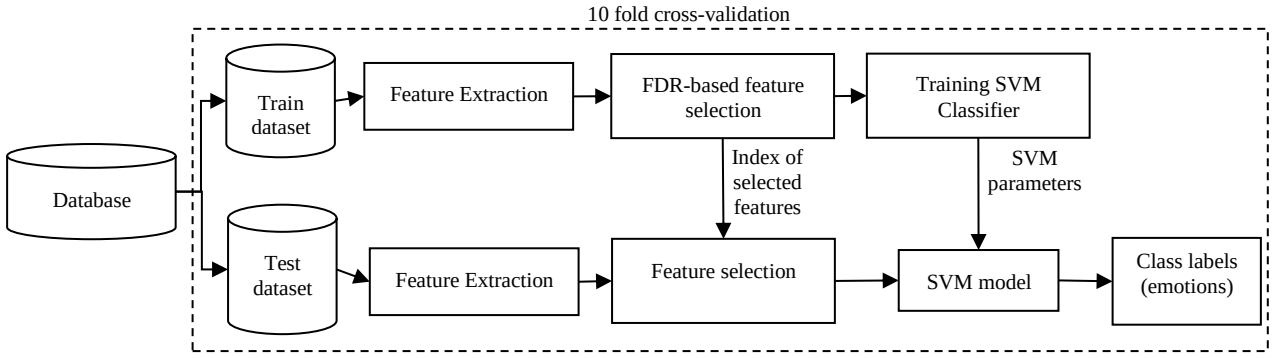

Fig.4. The block diagram of the proposed SER system.

In the following, the proposed features are compared with prosodic and spectral features by means of classification accuracy. Then, their contribution as supplementary features to prosodic and spectral features is investigated.

Table 4 represents the results of classification using different feature sets on the Berlin and PDREC databases. Although the proposed features are not better than the prosodic and spectral ones, they can improve the recognition rate when used to augment them. As seen from Table 4, by using the combination of prosodic and spectral features, the best accuracy of 83.18% and 89.36% were achieved on the Berlin database for females and males, respectively. However, when the proposed features were added to the prosodic and spectral ones, the recognition rate improved to 86.82% and 90.43% for females and males, respectively. Also, on the PDREC, combining the proposed features with the prosodic and spectral features can improve the recognition rate of females and males by 3.72% and 0.27%,

respectively. Correspondingly, the best classification accuracy of 63.18% and 57.37% are obtained on PDREC for females and males, respectively.

Fig.5 (a) and (b) show the average recognition rates of different emotions using combination of all types of features. As it can be seen from Fig.5 (a) and (b), the average recognition rates obtained on PDREC are smaller than those of achieved on Berlin database. It may be due to the similarities of PDREC to the naturalness of the databases, as pointed in section 3.2.

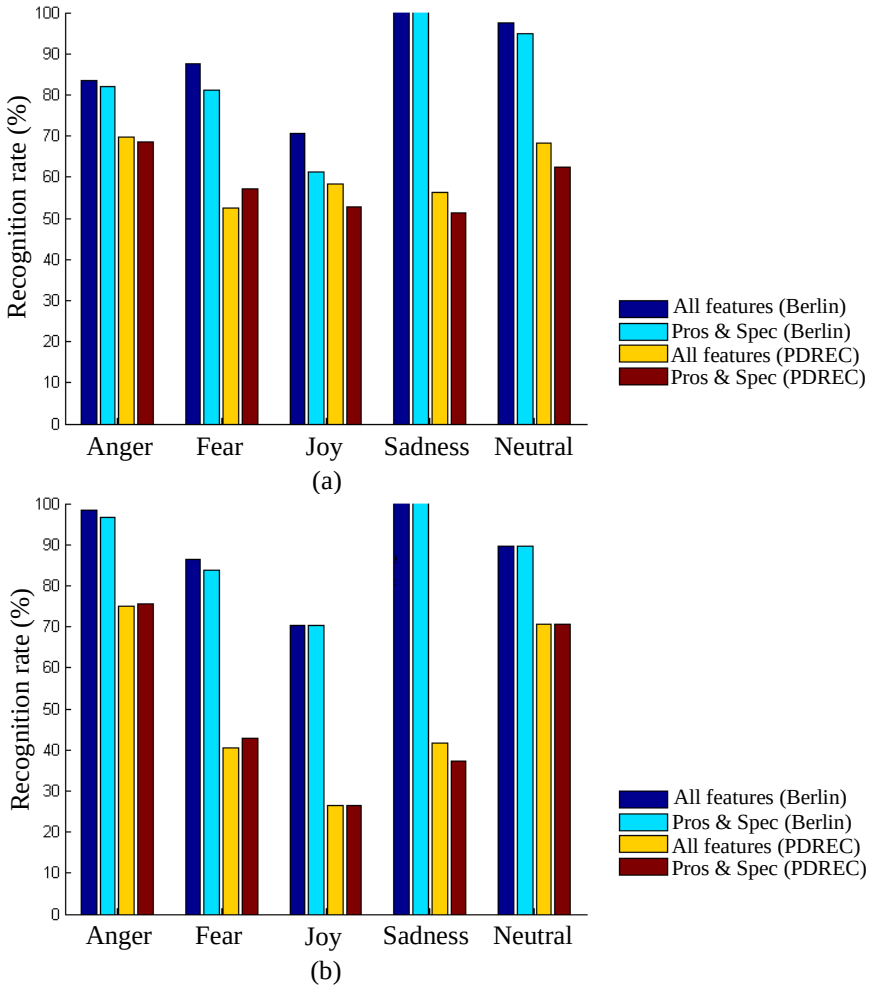

Fig.5. The best average recognition rates of different emotions on Berlin and PDREC databases for (a) females, and (b) males.

Table 4. Classification results using different types of features.

	Recognition rate on PDREC (%)		Recognition rate on Berlin (%)	
Feature type	Female	Male	Female	Male
Prosodic	52.03	51.90	74.09	77.66
Spectral	56.76	52.72	81.82	87.77
Variogram	53.04	46.76	66.36	68.09
Prosodic& Spectral	59.46	57.07	83.18	89.36
All features	**63.18**	**57.34**	**86.82**	**90.43**

The classification results achieved on Berlin database using combination of prosodic and spectral features are shown by 2 confusion matrices in Table 5 and 6 for females and males, respectively. The confusion matrices achieved using combination of prosodic and spectral features on PDREC are presented in Table 7 and 8 for females and males, respectively. In these tables, the left-most column and the top row indicate the true and recognized emotions, respectively. The "Rate" column represents the average recognition rate for each class. It is determined as the number of samples correctly recognized divided by the total number of samples in the class. The "Pre" row denotes the precision of each class. It can be calculated as the number of samples correctly classified divided by the total number of samples assigned to the class.

Table 5. Confusion matrix using combination of prosodic & spectral features (Berlin-females).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**55**	4	8	0	0	82.09
Fear	6	**26**	0	0	0	81.25
Joy	14	1	**27**	0	2	61.36
Sadness	0	0	0	**37**	0	100
Neutral	0	0	2	0	**38**	95.00
Precision (%)	73.33	83.87	72.97	100	95.00	
Average recognition rate: 83.18						

Table 6. Confusion matrix using combination of prosodic & spectral features (Berlin-males).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**58**	1	1	0	0	96.66
Fear	1	**31**	3	1	1	83.78
Joy	5	3	**19**	0	0	70.37
Sadness	0	0	0	**25**	0	100
Neutral	1	2	0	1	**35**	89.74
Precision (%)	89.36	83.78	82.61	92.60	97.22	
Average recognition rate: 89.36						

Table 7. Confusion matrix using combination of prosodic & spectral features (PDREC-females).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**50**	1	11	3	8	68.49
Fear	1	**12**	0	8	0	57.14
Joy	8	0	**19**	4	5	52.77
Sadness	9	7	6	**40**	16	51.28
Neutral	11	3	5	14	**55**	62.50
Precision (%)	63.30	52.17	34.46	57.97	65.48	
Average recognition rate: 59.46						

Table 8. Confusion matrix using combination of prosodic & spectral features (PDREC-males).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**79**	10	9	4	2	75.96
Fear	12	**18**	0	5	7	42.86
Joy	14	0	**13**	5	17	26.53
Sadness	6	8	2	**25**	26	37.31
Neutral	5	4	9	13	**75**	70.75
Precision (%)	68.10	45.00	39.40	48.08	59.05	
Average recognition rate: 57.07						

Confusion matrices for Berlin database using the combination of all types of features are shown in Table 9 and 10 for females and males, respectively. Table 11 and 12 also show the confusion matrices for PDREC using the combination of all types of features.

Table 9. Confusion matrix using combination all types of features (Berlin-females).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**56**	3	7	0	1	83.58
Fear	4	**28**	0	0	0	87.50
Joy	10	0	**31**	0	3	70.54
Sadness	0	0	0	**37**	**0**	100
Neutral	0	1	0	0	**39**	97.50
Precision (%)	80.00	87.50	81.58	100	90.70	
Average recognition rate: 86.82						

Table 10. Confusion matrix using combination all types of features (Berlin-males).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**59**	1	0	0	0	98.33
Fear	2	**32**	3	0	0	86.49
Joy	6	2	**19**	0	0	70.37
Sadness	0	0	0	**25**	0	100
Neutral	1	2	0	1	**35**	89.74
Precision (%)	86.76	86.49	86.36	96.15	100	
Average recognition rate: 90.43						

Table 11. Confusion matrix using combination all types of features (PDREC-females).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**51**	3	7	5	7	69.86
Fear	2	**11**	0	8	0	52.38
Joy	7	0	**21**	4	4	58.33
Sadness	11	5	4	**44**	14	56.41
Neutral	10	1	2	15	**60**	68.18
Precision (%)	62.96	55.00	61.76	57.89	70.59	
Average recognition rate: 63.18						

Table 12. Confusion matrix using combination all types of features (PDREC-males).

	Anger	Fear	Joy	Sadness	Neutral	Rate (%)
Anger	**78**	10	9	4	3	75.00
Fear	13	**17**	0	5	7	40.48
Joy	14	0	**13**	5	17	26.53
Sadness	6	7	2	**28**	24	41.79
Neutral	6	4	8	13	**75**	70.75
Precision (%)	66.66	44.74	40.62	50.91	59.52	
Average recognition rate: 57.34						

The analysis of the confusion matrices shows that the confusion of joy and anger contributes to a major part of classification error. Classification of this valence related emotions pair reported as a serious challenge in many SER systems [5,11,40]. It is due to the correlation of acoustic features employed for SER with the activation of emotions. Although the proposed features could not significantly reduce the confusion of these emotions, they improve the average recognition rate on both Berlin and PDREC databases.

Due to different conditions of experiments such as data partitioning and number of categories, it is hard to compare the performance of this research with other related works. However, we have presented the results of existing studies as a benchmark. Authors in [5] classified 7 emotions on the Berlin database using Non-Linear Dynamics features (NLDs) combined with prosodic and spectral features under the 10 fold cross-validation. They achieved the average recognition rate of 82.72% and 85.90% for females and males, respectively. The accuracy of 93.78% is obtained for a 3 emotions (neutral, fear and anger emotions) using features based on nonlinear dynamics and a neural network classifier [41]. The best average recognition rate of 85.6% is obtained under 10 fold cross-validation for classifying 7 emotions using a multi-class SVM classifier and Modulation Spectral Features (MSFs) [11]. The accuracy of 88.8% is obtained for recognition of 6 emotions using a 3 stage classification scheme and a large set of voice quality parameters [42].

5.0. CONCLUSION

This study aims to assess the proposed variogram-based features in recognition of human's emotions from speech. This paper has demonstrated the effectiveness of proposed features for SER. Our experiments show that the spectrogram of speech contains important hints about speaker's emotion. Moreover, the variogram graphs derived from speech spectrogram have shown to be useful for extracting effective emotional features from speech. These curves contain important information about timing and frequency characteristics of speech signal. Although the

proposed features are not as discriminative as the widely used prosodic and spectral features, they can serve as useful complements for them.

Future works include extraction of new features from spectrogram image which are discriminative for valence related emotions: joy and anger. Also, since a practical SER system works with real data under different conditions such as in presence of noise , it is useful to evaluate various feature types in such conditions.

REFERENCES

[1] J. Nicholson, K. Takahashi, R. Nakatsu. Emotion recognition in speech using neural networks. Neural Comput 2000; 9: 290–296.

[2] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE T BIO-MED ENG 2007; 47: 829–837.

[3] Raj, R.G. and Abdul-Kareem. Information Dissemination And Storage For Tele-Text Based Conversational Systems' Learning. MALAYAS J COMPUT SCI, Dec 2009; Vol. 22(2): 138-159.

[4] Raj, R.G. and Abdul-Kareem, S. A Pattern Based Approach for The Derivation Of Base Forms Of Verbs From Participles And Tenses For Flexible NLP. MALAYAS J COMPUT SCI, Jun 2011; Vol. 24(2): pp 63-72.

[5] A. Shahzadi, A. R. Ahmadyfard, A. Harimi, K. Yaghmaie. Speech emotion recognition using non-linear daynamic feature. TURK J ELECTR ENG CO 2013; in press.

[6] M. ElAyadi, M. S. Kamel, F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 2011; 44: 572–587.

[7] H. Altun, G. Polat. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst Appl 2009; 36: 8197–8203.

[8] D. Bitouk, R. Verma, A. Nenkova. Class-level spectral features for emotion recognition. Speech Commun 2010; 52: 613–625.

[9] N. Kamaruddin, A. Wahab, C. Quek. Cultural dependency analysis for understanding speech emotion. Expert Syst Appl 2011; 39:5115–5133.

[10] J. Rong, G. Li, Y. P. Phoebe Chen. Acoustic feature selection for automatic emotion recognition from speech. Inform Process Manag 2009; 45: 315–328.

[11] S. Wu, T.H. Falk, W.Y. Chan. Automatic speech emotion recognition using modulation spectral features. Speech commun 2011; 53: 768–785.

[12] T. Polzehl, A. Schmitt, F. Metze, M. Wagner. Anger recognition in speech using acoustic and linguistic cues. Speech Commun 2011; 53: 1198–1209.

[13] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, K. Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. Comput Speech Lang 2011; 25: 84–104.

[14] E. Bozkurt, E. Erzin, C. E. Erdem, A.T. Erdem. Formant position based weighted spectral features for emotion recognition. Speech Commun 2011; 53: 1186–1197.

[15] M. B. Goudbeek, J. P. Goldman, K. Scherer. Emotion dimensions and formant position. In: 10th Annual Conference of the International Speech Communication Association; September 6–10, 2009; Brighton, United Kingdom: Curran Associates. pp. 1575–1578.

[16] L. He, M. Lech, N. C. Maddage, N. B. Allen. Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomed Signal Proces 2011; 6: 139–146.

[17] C. C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun 2011; 53: 1162–1171.

[18] S. Wu, T. H. Falk, W. Y. Chan. Automatic speech emotion recognition using modulation spectral features. Speech commun 2011; 53: 768–785.

[19] A. Shahzadi, A. R. Ahmadyfard, K. Yaghmaie, A. Harimi. Recognition Of Emotion In Speech Using Spectral Patterns. MALAYAS J COMPUT SCI 2013; 26.

[20] P. Go´mez-Vilda, J. M. Ferra´ndez-Vicente, V. Rodellar-Biarge, R. Ferna´ ndez-Baı´llo. Time-frequency representations in speech perception. NEUROCOMPUTING 2009; 72: 820–830.

[21] P. K. Ajmera, D. V. Jadhav, R. S. Holambe. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. PATTERN RECOGN 2011; 44: 2749–2759.

[22] R. A. Cole, A. I. Rudnicky, V. M. Zue. Performance of an expert spectrogram reader. Journal of Acoustic Society of America 1979; 65: 81–87.

[23] V. W. Zue. An expert spectrogram reader: a knowledge-based approach to speech recognition. In: Proceedings of International Conference on Acoustic Speech and Signal Processing; 7–11 April 1986; Japan: IEEE. pp. 1197–1200.

[24] P. G. Vilda, J. M. Ferrandez-Vicent, V. Rodellar-Biarge, R. Fernandez-Baillo. Time-frequency representations in speech perception, NEUROCOMPUTING 2009; 72 : 820–830.

[25] J. Sohn, N. S. Kim, W. Sung. A statistical model-based voice activity detection. IEEE SIGNAL PROC LET 1999; 6: 1–3.

[26] L. Rabiner, R. Schafer. Digital Processing of Speech Signals. 1nd ed. United States: Pearson Education,1978.

[27] L. Rabiner, B. H. Juang. Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, 1993.

[28] T. Kinnunen, H. Li. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 2010; 52: 12–40.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss. A database of German emotional speech. In: 9th European Conference on Speech Communication and Technology; 4–8 September 2005; Lisbon, Portugal: Interspeech 2005. pp. 1517–1520.

[30] N. A. C. Cressie. Statistics for spatial data. Revised ed. New York, NY, USA: Wiley-Interscience,1993

[31] G. Matheron. Principles of geostatistics. ECON GEOL 1963; 58: 1246–1266.

[32] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, B. Schuller. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. Neurocomputing 2012; 84: 65–75.

[33] B. Schuller, M. Wimmer, L. Mosenlechner, C. Kern, D. Arsic, G. Rigoll. Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?. In: IEEE 2008 International Conference on Acoustics, Speech, and Signal Processing; 31 March– 4 April 2008; Las Vegas, Nevada. New York, NY, USA: IEEE. pp. 4501–4504.

[34] Z. Esmaileyan, H. Marvi. A database for automatic Persian speech emotion recognition: collection, processing and evaluation. International Journal of Engineering; Vol. 27 - 1 - Transactions A: Basics, January 2014, pp. 79-90.

[35] Raj, R.G. and Balakrishnan, V. A Model For Determining The Degree Of Contradictions In Information. MALAYAS J COMPUT SCI, September 2011; Vol. 24(3): pp 160-167.

[36] M. Kotti, C. Kotropoulos. Gender classification in two Emotional Speech databases. In: 19th International Conference on Pattern Recognition; 8–11 Dec 2008; Tampa, Florida, USA. New York, NY, USA: IEEE. pp. 1–4.

[37] C. C. Hsu, C. C. Chang, C. J. Lin. A practical guide to support vector classification. Tech rep, Department of Computer Science, National Taiwan University, Taiwan, 2007.

[38] C. M. Bishop. Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[39] C. Y. Fook, H. Muthusamy, L. S. Chee, S. B. Yaacob, A. H. Bin Adom. Comparison of speech parameterization techniques for classification of speech dysfluencies. Turk J Elec & Comp Sci 2012; in press.

[40] E. H. Kim, K. H. Hyun, S. H. Kim, Y. K. Kwak. Improved Emotion Recognition With a Novel Speaker-Independent Feature. IEEE-ASME T MECH 2009; 14: 317–325.

[41] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. R. Orozco-Arroyave. Application of Nonlinear Dynamics Characterization to Emotional Speech. In: Travieso-González CM, Alonso-Hernández JB, editors. Advances in Nonlinear Speech Processing. Springer Berlin Heidelberg, 2011. pp. 127–136.

[42] M. Lugger, B. Yang. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In: IEEE 2008 International Conference on Acoustics, Speech, and Signal Processing; 31 March– 4 April 2008; Las Vegas, Nevada. New York, NY, USA: IEEE. pp. 4945–4948.