RECOGNITION OF EMOTION IN SPEECH USING SPECTRAL PATTERNS

Ali Shahzadi[1] , Alireza Ahmadyfard[2] , Khashayar Yaghmaie[3] , Ali Harimi[4]
[1] Electrical Engineering Department, Semnan University, Semnan, IRAN
[2] Electrical Engineering Department, Shahrood University of technology, Shahrood, IRAN
[3] Electrical Engineering Department, Semnan University, Semnan, IRAN
[4] Electrical Engineering Department, Semnan University, Semnan, IRAN
Email: [1] shahzadi@iust.ac.ir, [2] ahmadyfard@shahroodut.ac.ir, [3] khashayar.yaghmaie@gmail.com,
[4]a.harimi@gmail.com

ABSTRACT

Recent developments in man-machine interaction have intensified the need for recognizing human's emotion from speech. In this study we proposed using Spectral Pattern (SP) and Harmonic Energy (HE) features for the automatic recognition of human affective information from speech. These features were extracted from the spectrogram of the speech signal using image processing techniques. A filter and wrapper feature selection scheme was used to avoid the curse of dimensionality. Here, a hierarchical classifier is employed to classify speech signals according to their emotional states. This classifier is optimized by the Fisher Discriminant Ratio (FDR) to classify the most separable classes at the upper nodes, which can reduce the classification error. Moreover, a tandem classifier is employed to increase the recognition rate of highly confused emotions pairs. Our experimental results have demonstrated the potential and promise of SPs and HEs for emotion recognition. The proposed method was tested on the male and female speakers separately and the overall recognition rate of 86.9% is obtained for classifying seven emotion categories in the Berlin database.

Key words: speech emotion recognition; spectral patterns; harmonic energy; time-frequency speech analysis

1.0. INTRODUCTION

Speaking is the fastest and most natural method of communication between humans [1]. The speech communication consists of two explicit and implicit channels which carry linguistic content ("what was said") and paralinguistic information ("How it was said"), respectively [2]. A noticeable number of researches in automatic speech recognition (ASR) with the aim of extracting the linguistic information from the speech signal have been reported. However, there is still much research that needs to be done in order to decode the implicit channels to extract the paralinguistic information such as gender, age, emotion, quality and alcohol/ drug consumption of the speaker [2]. Among these, recognition of emotional state of the speaker has been one of the most attractive research fields in the last few years. Speech Emotion Recognition (SER) has a wide range of applications in interactive systems. It can improve the performance of speech recognition systems [3]. It can also be useful in E-learning, computer games, in-car boards and every other application that requires natural interaction [4]. Furthermore, SER can be used in medical science and psychology [5,6].

SER is commonly treated as a statistical pattern recognition problem which consists of three main stages: (1) feature extraction, (2) feature selection and (3) pattern classification [7]. It has been shown that acoustic features such as the pitch, timing, voice quality, and articulation of the speech signal highly correlate with the underlying emotion [8,9]. In emotions such as anger, fear and joy the sympathetic nervous system is aroused and consequently the resultant speech is loud and fast with a higher pitch average and wider pitch range [1,10]. On the other hand, with the arousal of the parasympathetic nervous system, as with emotions such as boredom and sadness, the resultant speech is slow, low pitched and with little high-frequency energy [1,10]. Despite of widespread efforts, specifying effective features

is still one of the main challenges in SER [1] and there are contradictory reports on the effect of emotions on some of acoustic features [1].

Most acoustic features used in SER can be studied in two main categories: prosodic features and spectral feature. Prosodic features, which are widely used in SER, have been shown to offer important emotional cues of the speaker [1,7,11-24]. These features are commonly based on information such as pitch, energy and timing. Spectral features, on the other hand, generally consist of information obtained from the speech spectrum. These features convey the frequency contents of the signal and provide complementary information for prosodic features [1]. In this way, the authors in [8,16, 17,21-30] used formants related features for emotion recognition. Also, Mell Frequency Cepstral Coefficients (MFCCs) [7,13,16-20,22,31], Weighted MFCC (WMFCC) and Linear Spectral Frequency (LSF) [22], Linear Prediction Coding (LPC) [13,18], Sub-Band features [13,16,18], Modulation Spectral Features (MSF) and Perceptual Linear Prediction (PLP) [12] have been reported as effective spectral features in SER.

It is necessary to select the most relevant subset of the whole candidate features to avoid the curse of dimensionality [7]. Various feature selection algorithms can be studied in three main categories: Filters, wrappers and embedded methods [32]. Filter based methods evaluate features individually, and independent of the utilized classifier, as a pre-processing step. Although filter based methods are efficient in terms of computational simplicity, utilizing these methods solely can be hazardous because they do not consider the combination effects of features and classifier properties. Fisher Discriminant Ratio (FDR) [33] and Information Gain Ratio (IGR) [17] are two filter based techniques used for selection of effective features in SER. Wrappers use the classifier in order to score subsets of features according to their classification accuracy. While these methods have the advantage of considering the combination effects of features and classifier properties, they suffer from computational complexity. Sequential Forward Selection (SFS) is one of the most popular wrapper feature selection algorithms employed for SER [2,12,13,18,23,34,35]. Authors in [12], employed a two stage filter and wrapper feature selection algorithm based on Fisher Discriminant Ratio (FDR) and SFS. Embedded methods perform feature selection in the process of training. They are usually specific to given learning machines [32]. Dimensionality reduction approaches such as Principle Component Analysis (PCA) [21] and Linear Discriminant Analysis (LDA) [12] are also employed in SER projects.

Classification is the last stage of a SER system. In 1990s, most of the SER systems were based on the simple Maximum Likelihood Bayes algorithm (MLB) and Linear Discriminant Classification (LDC) [7]. Around 2000, methods based on Neural Network (NN) classification became popular for emotion recognition applications [3,36,37]. Since 2002, Support Vector Machine (SVM) [4,38-40] and Hidden Markov Model (HMM) [41-44] have received more attention. Each classifier has its advantages and shortfalls, and the researchers are still trying to find the better solution [7].

Since the spectral distribution of speech signal changes over the time according to the speaker's emotional state, it seems that time-frequency analysis methods such as spectrogram can be efficient to fill the existing gap between time analysis methods and frequency analysis methods. A spectrogram is a graphical display of the squared magnitude of the time-varying spectral characteristics of speech. It can effectively capture the dynamical behavior of speech signal. Also, contextual variations in speech are better represented using a spectrogram [45]. Methods based on spectrogram analysis have the advantage of preserving the important underlying dependencies between different parameters. These facts have made the spectrogram as a useful tool for speech analysis [46]. It has been reported that the speech spectrogram contains rich information about energy, pitch, formants and timing that could be valuable in speech processing applications such as speech and speaker recognition [47]. In [48], spectrogram is used as a time-frequency representation technique for speech perception. The features based on Radon and discrete cosine transforms derived from speech spectrogram are reported to be efficient in speaker identification [45].

Fig.1 (a) to (c) show the spectrograms of an utterance expressed by a woman in 3 different emotions; "anger", "neutral" and "boredom", respectively. As can be seen from Fig.1, acoustic features such as the formants energy in the higher frequency bands, pitch, and formants vary under different emotions. This shows that the spectrograms contain important information that can be used to discriminate different emotions.

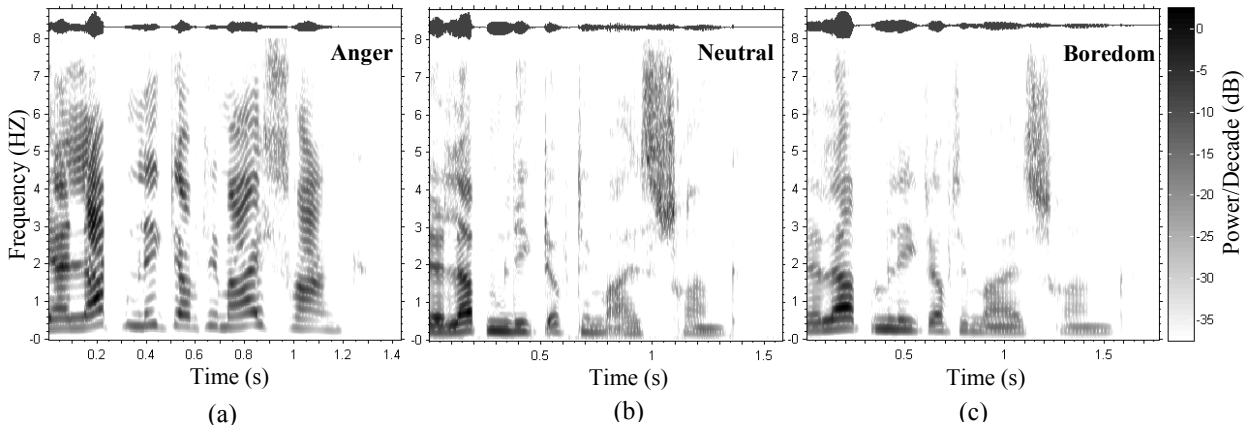

Fig. 1. Spectrograms of an utterance expressed by a woman in 3 different emotions; (a) anger, (b) neutral, and (c) boredom.

Our contribution in this work is to propose Spectral Pattern features (SPs) and Harmonic Energy features (HEs) extracted from the speech spectrogram using image processing techniques for SER. In this method, the spectrogram of speech signal is treated as a gray level image which represents the signal in time-frequency domain, and the proposed features are extracted from this image. The reminder of paper is organized as follows. Section 2 details the SPs and HEs proposed in this work, as well as prosodic and spectral features extracted for comparison purposes. Section 3 introduces the database employed. Experimental results are presented and discussed in Section 4. The paper finally ends with conclusion remarks in section 5.

2.0. FEATURE EXTRACTION

In this section, we detail the proposed SPs and HEs extracted from the time-frequency representation (spectrogram) of speech signal. Prosodic and spectral features considered in our experiments are also described here. The prosodic and spectral features calculated in this work are by no means exhaustive, but serve as a representative sampling of the essential features. These features are used here as a benchmark, and more importantly, to verify whether the SPs and HEs can serve as useful additions to the widely used prosodic and spectral features.

Since the speech signal is non-stationary in nature, it is common in speech processing to divide a speech signal into short duration of 20–30 ms called frames wherein the speech signal remains approximately stationary [10]. The so called local features (frame-level) such as pitch, energy, MFCCs and formants are extracted from each frame [1,12]. Global (utterance-level) features, on the other hand, are calculated as statistics (e.g. maximum, minimum, range, mean, median and variance) of the local features extracted from an utterance [1,12]. The majority of researchers believe that global features have several advantages over local ones [1]. These features reduce classification time and provide higher classification accuracy [49-52]. In this study, we use global features.

2.1. Time-Frequency Representation Of Speech

In order to extract the proposed Spectral Pattern features (Sps) and Harmonic Energy features (HEs), the speech signal should be firstly represented in time-frequency domain (spectrogram). As a pre-processing stage, the silent part of the speech signal is discarded by a Voice Activity Detection (VAD) algorithm [53], and then the input speech waveform x[n] is pre-emphasized as [27]:

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

Where $\hat{x}$ represents the pre-emphasized speech signal. The most common value for α is around 0.95 [54]. The pre-emphasized signal is segmented into frames of 20 ms duration with the shift of 10 ms between two consecutive frames to retain a good quality of the signal and to avoid loss of information [54,55]. Windowing is carried out to reduce the edge effects at the beginning and the end of the frame [27]. In this study, Hamming window is multiplied with each frame. In order to calculate the speech spectrogram, 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 and $\tilde{X}_i(k)$ is the k[th] component of DFT of $\tilde{x}_i(n)$ (i[th] windowed frame). $\mathrm{Re}\{\ldots\}$ and $\mathrm{Im}\{\ldots\}$ indicate real and imaginary parts, respectively. The spectrums of the frames, $S_i(k)$, are concatenated row-wise to construct the speech spectrogram, *f*, which represents the speech signal in time-frequency domain as:

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

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

2.2. Spectral Pattern Features

As a pre-processing stage, the contrast of the image is increased using histogram equalization to highlight details of the image [56]. Then, the image is decomposed to eight components by applying eight directional filters, H1 to H8, which have been shown in Fig.2 (a). These filters highlight different directional patterns (with the angle of θ) in the image. The eight resultant images are binarized, and then the morphological operators "cleaning" and "removing" are applied to remove the isolated pixels and interior pixels, respectively [56]. Finally, the desired patterns are detected using binary masking technique (matched filters). The applied binary masks, denoted by B1 to B8, have been shown in Fig. 2(b).

Since there are two different types of 63^o patterns in the image, the two resultant binary images which are obtained by applying H5 and H6 followed by binary masks, B5 and B6, are joint together using "and" operator. The process is repeated for -63^o patterns too. So, we have six different binary images representing six different directional patterns. In each of these images, if the value of a pixel is "1", the corresponding pattern exists at the same place, and in the contrary, if the value of the pixel is "0", the corresponding pattern is not present there. Since the characteristics of the spectrogram are different in various frequency bands for different emotions, we decompose the images into several sub bands. The simplest method is to divide the bandwidth equally. However, this does not seem appropriate, as it does not correspond to the human ear [57]. The spectral resolution of the human ear varies logarithmically along the frequency, with better resolution at lower frequencies [58]. The Mel-scale and the Bark-scale which are empirically determined using human subjects are two potential choices. In this study, we decompose the images into 17 critical bands (CB), according to Bark-scale with no overlapping filters. So, $6 \times 17 = 102$ sub-banded images are constructed. Fig. 3 shows the spectral pattern feature extraction process schematically.

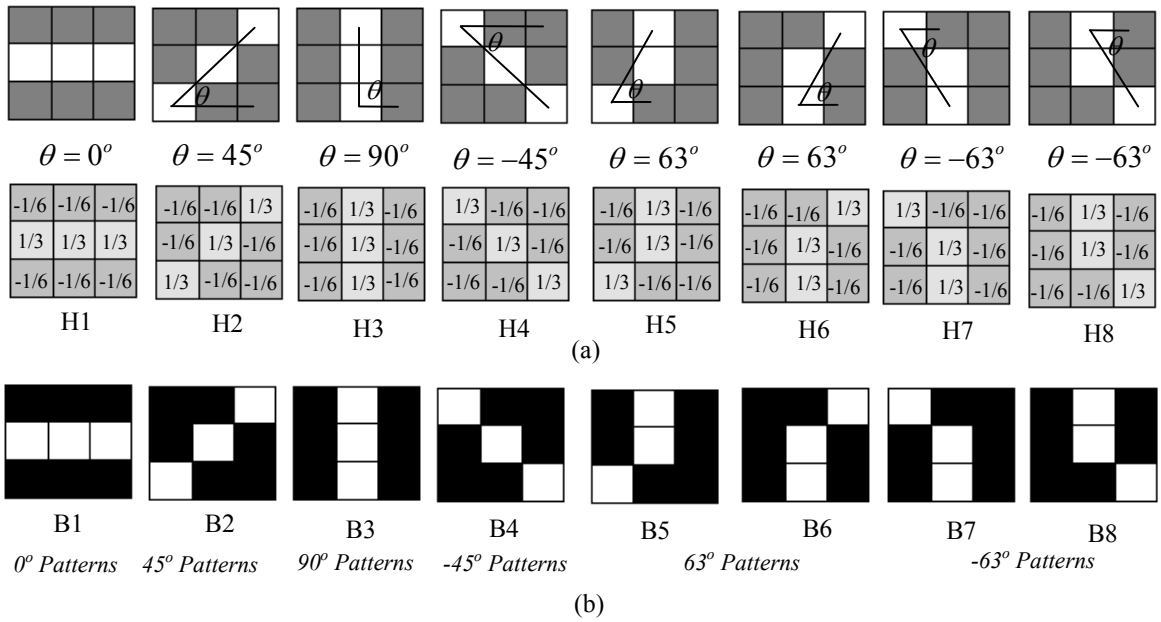

Fig. 2.. (a) Eight filters which are used to highlight eight directional patterns, (b) Eight binary masks (matched filters) which are used to detect eight directional patterns.

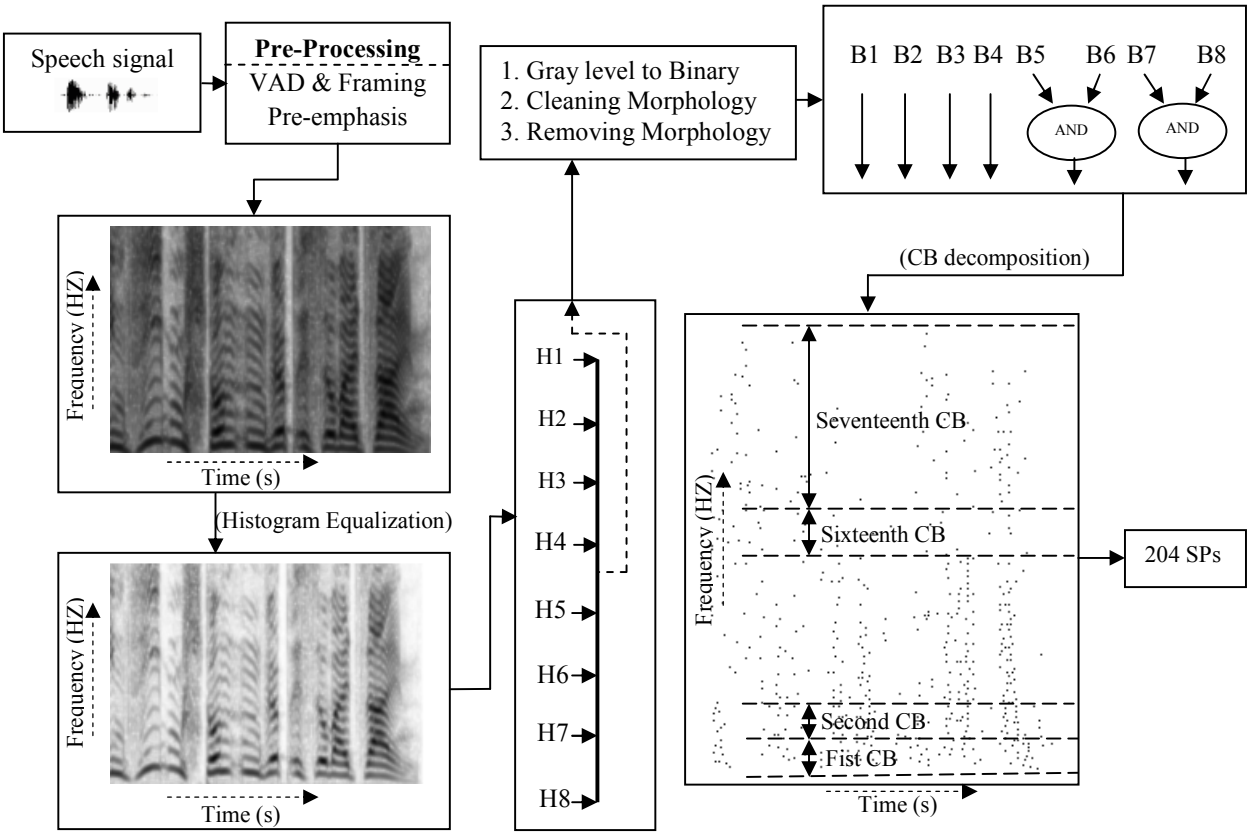

Fig. 3.. The block diagram of SP features extraction process.

We extract two different types of features from each of the *sub-banded images*. Denoting the q^{th} *sub-band* of the p^{th} *image* by $SI_{q,p}$, the extracted features are determined as follows:

1) *The average number of patterns per frame in each sub band;* These features form the first 102 SP features of the SP Feature Vector (*SP_FV)*. It could be formulated as follows;

$$SP\_FV(1:102) = \frac{1}{M}\sum_{r=1}^{R_q}\sum_{c=1}^{M} SI_{q,p}(r,c), 1 \le q \le 17, 1 \le p \le 6 \tag{4}$$

Where $SI_{q,p}$ is the q^{th} *sub-band* of the p^{th} binary *image,* R_q is the number of rows in the $SI_{q,p}$ and M is the number of columns in the $SI_{q,p}$ which is equal to number of frames in speech signal.

2) *The relative number of each pattern in each sub band*; These features form the second 102 features of the *SP_FV* and they could be determined as follows:

$$SP\_FV(103:204) = \frac{\sum_{r=1}^{R_q}\sum_{c=1}^{M} SI_{q,p}(r,c)}{\sum_{i=1}^{17}\sum_{r=1}^{R_i}\sum_{c=1}^{M} SI_{i,p}(r,c)}, 1 \le q \le 17, 1 \le p \le 6 \tag{5}$$

2.3. Harmonic Energy Features

Harmonics of a speech signal are multiples of its fundamental frequency, F0, which is created by the vibration of the vocal folds. Since F0 varies over time, a filter bank consists of time varying band-pass filters, is required to extract the harmonics of speech signal. In this filter bank, the central frequency for the h^{th} band-pass filter at a certain time is h^{th} multiple of the fundamental frequency at that time, and the bandwidth is equal to the fundamental frequency [59]. In this study, the filter bank is implemented on the spectrogram image. The central frequency and cutoff frequencies of each of the sub-band filters on the image is determined as follows: The frequency range which is covered by each pixel in the vertical direction could be determined as follows:

$$FR = f_s / 2R \tag{6}$$

Where *fs* and *R* are the sampling rate and the number of rows in the image, respectively. In spectrogram image, the position of central frequency for the h^{th} filter in the i^{th} column (i^{th} frame) can be determined as:

$$Fc_h(i) = h \times F0(i) / FR, 1 \le i \le M \tag{7}$$

Where *F0(i)* is the fundamental frequency for the i^{th} frame which is calculated by the auto correlation method [27]. Since the bandwidth of each sub-band filter in the i^{th} frame is *F0(i)*, the first and second cutoff frequencies in that frame on the image can be determined as follows:

$$Fs_{h,1}(i) = Fc_h(i) - F0(i) / 2FR \tag{8}$$

and

$$Fs_{h,2}(i) = Fc_h(i) + F0(i) / 2FR \tag{9}$$

Where $Fs_{h,1}(i)$ and $Fs_{h,2}(i)$ are the lower and upper cutoff frequencies of the h^{th} sub-band filter in the i^{th} column of image, respectively. It is clear that the obtained values should be rounded to be used for digital images. The energy of the h^{th} harmonic in the i^{th} frame can be computed as:

$$E_h(i) = \sum_{k=Fs_{h,1}(i)}^{Fs_{h,2}(i)} f(k,i), 1 \le i \le M \tag{10}$$

Where $f(k,i)$ is the k^{th} row and i^{th} column of the spectrogram image, f, which is determined by equation (3). Fig. 4 shows the 1^{th}, 5^{th}, 9^{th} and 13^{th} sub-band filters on the spectrogram image. In this figure, the $F0$ contour and its 5^{th}, 9^{th} and 13^{th} harmonics are indicated with black dashed lines. Each of these curves can be considered as the center frequency of the corresponding sub-band filter. The cutoff frequencies of each filter are also depicted around the center frequencies by solid lines. In summary, the log energy contour of each harmonic could be computed by adding the gray level of the pixels within the corresponding pass band around the corresponding center frequency.

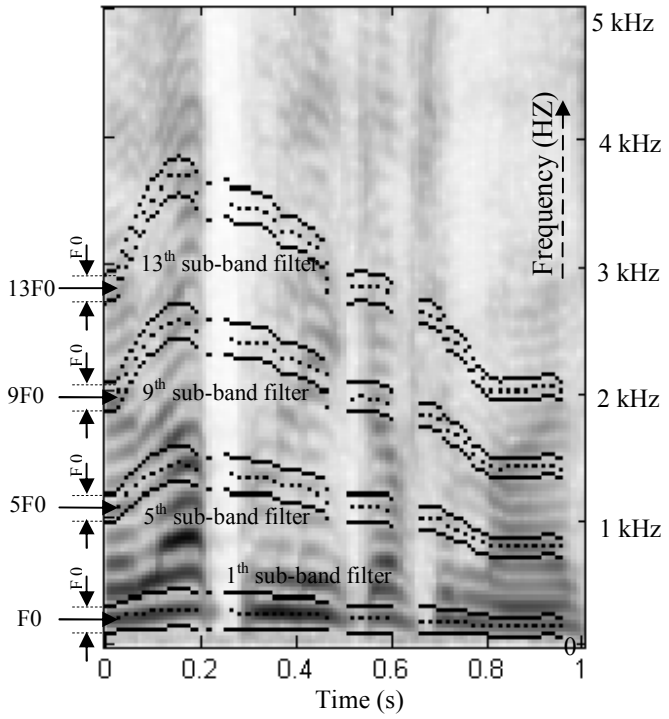

Fig. 4.. Implementation of the band pass filters on spectrogram image to determine harmonic energy contours.

In this work, we determine the energy contours of the 13 first harmonics. Then, we apply 20 statistical functions to these contours to extract HEs. The statistical functions are also applied to their first and second derivatives (velocity and acceleration) to extract the dynamical information of the curves. 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. So, in total, $3 \times 20 \times 13 = 780$ HE features are extracted in this work.

2.4. Prosodic Features

Prosodic features are the most widely used features in SER. They are commonly based on pitch, energy and speaking rate. In this study, pitch and energy tracking contours are firstly determined (frame-level features). The statistics of these trajectories are shown to offer important emotional cues of the speaker [1]. Here, the 20 statistical function described in section 2.3 are used to extract the utterance-level features from these contours and their first and second derivatives. Ratio between the duration of voiced and unvoiced speech is also used as a timing-based feature.

The traditional Zero-Crossing Rate (ZCR) and the Teager Energy Operator (TEO) [1] of the speech signal are also examined here. These features do not directly relate to prosody but in this work we evaluate their performance along with prosodic features. TEO conveys information about the nonlinear airflow structure of speech production [1]. The TEO for a discrete-time signal x_n is defined as:

$$TEO(x_n) = x_n^2 - x_{n-1}x_{n+1} \tag{11}$$

We also apply the aforementioned 20 statistical functions to ZCR and TEO curves and their deltas and double deltas to extract their features. In total, 241 prosodic features are extracted in this work.

2.5. Spectral Features

In this paper we employ two types of spectral features: The Mel-Frequency Cepstral Coefficients (MFCCs) [27] and formants [27]. These features are successfully applied to automatic speech recognition and reported as effective spectral features for emotion recognition [12,16,17]. The first 12 MFCCs and 4 formants are extracted from 20 ms Hamming-windowed speech frames every 10 ms. After that, their contours are formed and then the 20 statistical functions described in section 2.3, are applied to extract utterance level features from the contours. As a common practice, the features are also extracted from the delta and double-delta contours [12]. In total, $16 \times 20 \times 3 = 960$ spectral features are extracted here. All the extracted features are listed in Table 1.

Table 1. List of extracted features.

SP features: (204 features)	The average number of patterns per frame in each sub band; The relative number of each pattern in each sub band;
HE features: (780 features)	*Apply 20 statistical functions to:* H_1, H_2, …, H_{13}, Their delta and double-delta;
Prosodic features: (241 features)	*Apply 20 statistical functions to:* Pitch, delta pitch, double-delta pitch; Energy, delta energy, double-delta energy; Ratio between the duration of voiced and unvoiced speech; ZCR, ZCR delta, and ZCR-double delta; TEO, TEO delta, and TEO-double delta;
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;
Note! 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 DATA

Berlin database of German emotional speech [60] is a well-known public database. There are many papers which test their algorithms with this database [11-15]. This database consists of 535 utterances with 10 different contexts, which are expressed by 10 professional actors (5 male and 5 female) in seven emotions: "anger, boredom, disgust, fear, joy, neutral and sadness". The number of speech samples for each emotion is presented in Table 2.

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

We test the proposed SER system using this database. The non-overlapping training and testing sets are chosen by random sampling of utterances from a pool wherein all speakers are mixed to ensure the context and speaker independency.

4.0. EXPERIMENTS

In this section, results of the experimental evaluation are presented. Hierarchical Support Vector Machines (SVMs)-based classifier is employed to recognize emotions from speech. In this paper, it is assumed that a gender classifier with perfect classification accuracy, which is proposed by [61], is utilized in the first level of classification. So, the proposed system is implemented separately for males and females. Features from training data are linearly scaled to [-1, 1] before applying the classifier. As suggested in [12], features from test data are also scaled using the trained linear mapping function.

The employed two-stage filter and wrapper feature selection scheme is described in Section 4.1. In Section 4.2, the proposed features are initially compared to prosodic and spectral features, and then their contribution as supplementary features to them is investigated.

4.1. The Filter And Wrapper Feature Selection Technique

Firstly, the Fisher Discriminant Ratio (FDR) is applied to evaluate each of the extracted features according to inter class distance and intra class similarity. The normalized multi class FDR for the u^{th} feature could be written as [12]:

$$FDR(u) = \frac{2}{C(C-1)} \sum_{c_1} \sum_{c_2} \frac{(\mu_{c_1,u} - \mu_{c_2,u})^2}{\sigma^2_{c_1,u} + \sigma^2_{c_2,u}}, 1 \le c_1 < c_2 \le C \tag{12}$$

Where $\mu_{c_i,u}$ and $\sigma^2_{c_i,u}$ are mean and variance of the u^{th} feature for the i^{th} class, $i=1,2,...,C$, and C is the total number of classes. When the FDR is calculated for all of the extracted features, they are ranked based on their FDR values. At the next step, as it is proposed by [12], the features with the FDR values higher than 0.15 are chosen. This method can quickly filter the irrelevant features to reduce the computational cost.

Since FDR does not consider the combination effects of features, utilizing this method solely can be hazardous. So, the wrapper Sequential Forward feature Selection (SFS) algorithm is employed as a complementary stage. In this method each feature would be added to the selected feature subset if it causes to improve the recognition rate. Generally, comparing to filter methods, the wrapper based methods have the advantage of considering combination effects of features and classifier properties while they suffer from high computational cost [32].

4.2. Classification

The proposed classifier is shown in Fig. 5. We named the binary classifiers as "x_CbLy_z" where "x" is "F" and "M" for female and male speakers, respectively, "b" is 1 for the first classifier and 2 for the second classifier, "y" indicates the classification level and "z" is the binary classifier index in each level. In this figure, anger, boredom, disgust, fear, joy, neutral and sadness emotions are indicated by the numbers of 1 to 7, respectively. The choice of binary classifiers in the hierarchical classifier is flexible and largely dependent on the feature selection technique [20]. Such classifiers are usually designed empirically [20], but they can be further improved by optimizing the choice of binary classifiers along with the appropriate feature selection method at each classification level. Here, the choice of binary classifiers in the hierarchical structure is optimized using FDR score. In this method, the separability of each two classes is measured by means of FDR score. The most separable classes are separated at the upper nodes, which can reduce classification error. For example, at the first level of classification, there are 35 possible choices to classify seven emotions into two separate classes; one class consists of 3 and the other consists of 4 emotions. The best four choices with the higher FDR values for F_C1L1 and M_C1L1, and their corresponding FDR values are shown in Table 3.

Table 3. The best four choices for F_C1L1 and M_C1L1 and their corresponding maximum FDR**.**

The best Possible choice	Class 1	Class 2	Maximum FDR value	
			F_C1L1	M_C1L1
1	1, 3, 4, 5	2, 6, 7	**5.3**	**3.5**
2	2, 3, 6, 7	1, 4, 5	4	3.5
3	2, 4, 6, 7	1, 3, 5	2.2	2.2
4	1, 4, 5, 6	2, 3, 7	1.6	1.6

From this Table, it is clear that the best choice for both F_C1L1 and M_C1L1 is to consider "anger, disgust, fear and joy" in one class and "boredom, neutral and sadness" in another class. Designing the first classifier is finalized by repeating this algorithm until all of the emotions have been classified.

In order to improve the classification accuracy, we employed a tandem classifier (the second classifier). As it is shown in Fig. 5, this classifier consists of three binary classifiers. The methodology of designing the second classifier is described in section 4.3. Also, it is investigated that how this classifier can reduce the error rate.

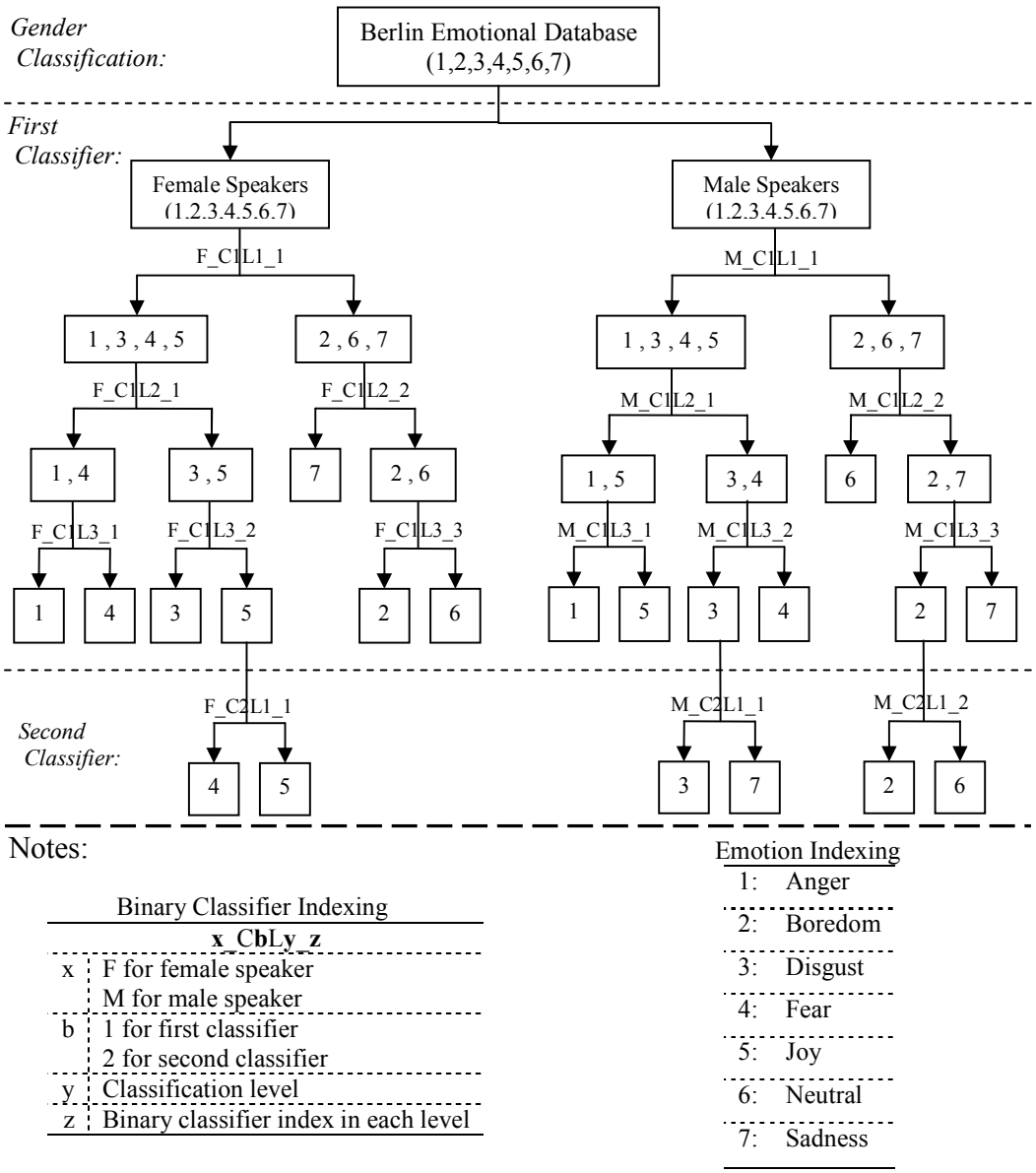

Fig. 5.. The proposed hierarchical SVM-based classifier.

4.3. Comparison Of Proposed SPs And HEs With Prosodic And Spectral Features

The proposed features are first compared to prosodic and spectral features, by means of FDR scores before applied to SVMs. To this end, the features are ranked by their FDR values using all instances in the Berlin database before data partitioning and then, FDR values averaged over the top N_{fdr} FDR-ranked features. Fig .6 (a) and (b) show the average FDR curves as a function of N_{fdr} for prosodic, spectral and proposed features, for females and males, respectively. These average FDR curves can offer rough indicators for discrimination power of the three feature types independent of the utilized classifier.

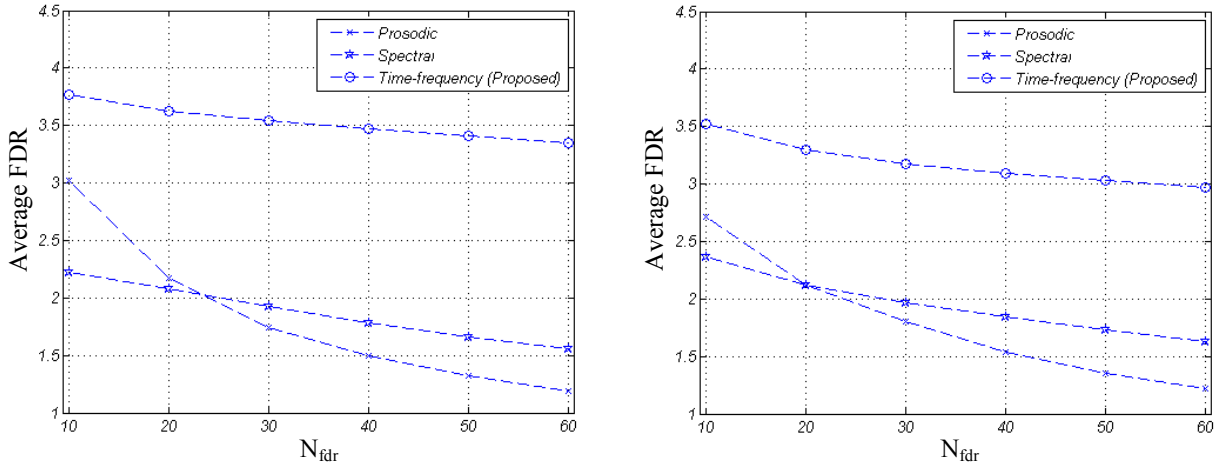

Fig. 6. Average FDR curves for the prosodic, spectral and proposed time-frequency features, (a) Females and (b) Males.

Based on Fig.6, the proposed SPs and HEs offer higher average FDR score than both prosodic and spectral features. It is promising for us to employ these features in our SER system. We combine these three feature sets (prosodic, spectral and proposed features) to form the final feature vector for the classification task. This feature vector is applied to the first classifier. The confusion matrices of classification for females and males are shown in Table 4 and 5, respectively.

Table 4. Confusion matrix of the first classifier for female speakers.

CM%	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**100%**	0	0	0	0	0	0
Boredom	0	**77.8%**	11.1%	0	11.1%	0	0
Disgust	0	0	**100%**	0	0	0	0
Fear	0	0	0	**33.3%**	**66.7%**	0	0
Joy	0	0	0	0	**100%**	0	0
Neutral	0	12.5%	0	0	0	**87.5%**	0
Sadness	0	28.6%	14.3%	0	0	0	**57.1%**

Table 5. Confusion matrix of the first classifier for male speakers.

CM%	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**100%**	0	0	0	0	0	0
Boredom	0	**85.7%**	0	0	0	0	14.3%
Disgust	0	0	**50%**	50%	0	0	0
Fear	0	0	0	**100%**	0	0	0
Joy	20%	0	0	0	**80%**	0	0
Neutral	0	**25%**	0	12.5%	0	**62.5%**	0
Sadness	0	0	**20%**	0	0	0	**80%**

The average recognition rate which will be the number of samples from all emotions correctly recognized divided by the total number of samples, is 83% for both female and male speakers. Considering the results of Tables 4 and 5 and at the same time the structure of the first classifier depicted in Fig.5, reveals that two types of errors are appeared in the confusion matrices. The first type of errors arises from confused pairs which are directly classified by specific binary classifiers (e.g. the 12.5% error rate resulted from the classification of "neutral" versus "boredom"

for female speakers using F_C1L3_3). These types of errors are unavoidable because a specific binary classifier is trained to classify such two emotions but some samples could not be recognized correctly. The second type of errors arises from indirect classification of emotions. For example, considering the two binary classifies M_C1L2_2 and M_C1L3_3, the first one classifies "neutral" versus "boredom and sadness" and the second one classifies "boredom" versus "sadness". Thus, these three emotions are classified but, "neutral" and "boredom" are classified indirectly and there is no specific binary classifier to classify these two emotions. Our experiments showed that these types of errors can be reduced using a tandem classifier.

To this end, all confused pairs of emotions which have not been directly classified by specific binary classifiers in the first classifier, can be re-classified using tandem binary classifiers in the second classifier. Since this approach results in a large number of binary classifiers in the second stage (5 binary classifiers for females and 3 binary classifiers for males), we have empirically chosen the most effective ones as the second classifier. As seen from Fig. 5, the second classifier is constructed using three binary classifiers: F_C2L1_1, M_C2L1 and M_C2L1_2, while using further binary classifiers in this structure does not significantly improve the classification accuracy. The confusion matrices obtained after applying the second classifier are shown in Tables 6 and 7 respectively.

Table 6. Confusion matrix after applying the second classifier for female speakers.

CM%	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**100%**	0	0	0	0	0	0
Boredom	0	**77.8%**	11.1%	0	11.1%	0	0
Disgust	0	0	**100%**	0	0	0	0
Fear	0	0	0	**66.7%**	**33.3%**	0	0
Joy	0	0	0	0	**100%**	0	0
Neutral	0	12.5%	0	0	0	**87.5%**	0
Sadness	0	28.6%	14.3%	0	0	0	**57.1%**

Table 7. Confusion matrix after applying the second classifier for male speakers.

CM%	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**100%**	0	0	0	0	0	0
Boredom	0	**85.7%**	0	0	0	0	14.3%
Disgust	0	0	**50%**	50%	0	0	0
Fear	0	0	0	**100%**	0	0	0
Joy	20%	0	0	0	**80%**	0	0
Neutral	0	**12.5%**	0	12.5%	0	**75%**	0
Sadness	0	0	**0**	0	0	0	**100%**

In Fig. 7 (a) and (b) we compare the average recognition rate of each emotion obtained before and after applying the second classifier for females and males, respectively.

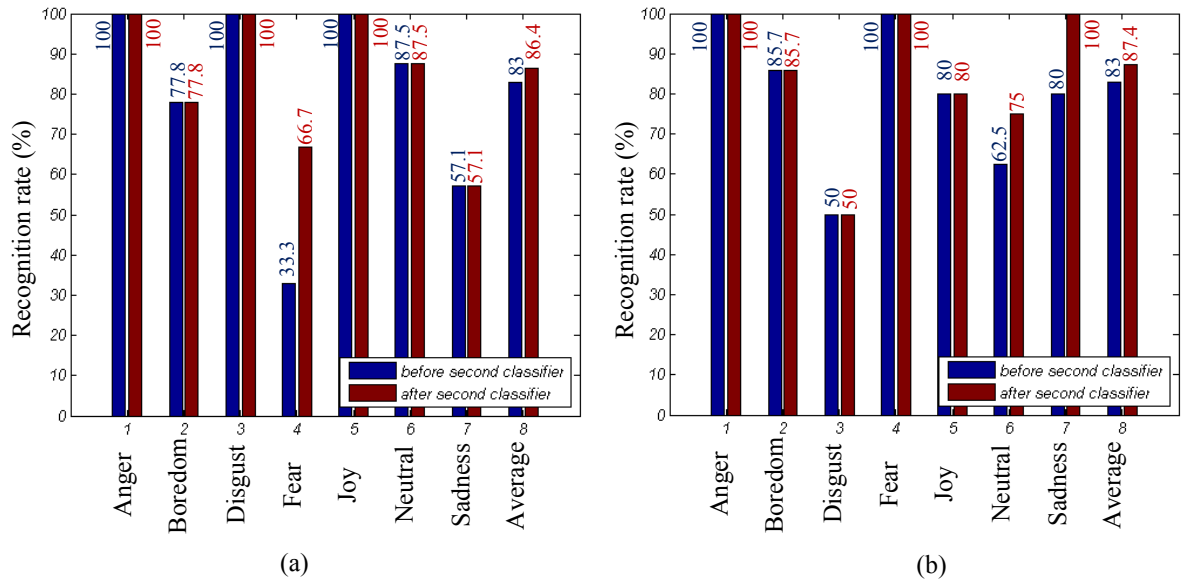

Fig. 7. Average recognition rate for each emotion before and after applying the second classifier, (a) Females and (b) Males.

As seen from Fig. 7, the second classifier can considerably improve the recognition rate. For female speakers, the recognition rate of "fear" is increased from 33% to 67%, and for male speakers, the recognition rates of "neutral" and "sadness" are increased from 62.5% and 80% to 75% and 100%, respectively. This improves the average recognition rates of females and males by 3.4% and 4.4%, respectively.

It is also useful to briefly review performance figures reported on the Berlin database by other works. Although the numbers cannot be directly compared due to factors such as different data partitioning, employed classifier and testing strategies, they are still useful for general benchmarking. The maximum average recognition rate is reported 71.75% in [14], 84.8% in [13], 84.6% in [7] and 85.6% in [11]. In this work, we achieved the average recognition rate of 86.9%.

5.0. CONCLUSION

The aim of this study was to evaluate Spectral Pattern features (SPs) and Harmonic Energy features (HEs) for the recognition of human's emotions from speech signal. The proposed features were extracted from the spectrogram of speech signal using image processing techniques. These features were also compared to traditional prosodic and spectral features. This paper has demonstrated the potential and promise of SPs and HEs for emotion recognition. The following conclusions can be drawn from the present study.

The first major finding of this paper was that, for speech signal, distribution of patterns in the spectrogram related to the emotional state of the speaker. Our experiments show that, the change rate of frequency directly related to the arousal of the corresponding emotion. For high arousal emotions such as joy and anger, the change rate of frequency is rapid, while for low arousal emotions such as boredom and sadness the frequency contents seems to be more stable. Moreover, for low arousal emotions, significant part of energy is concentrated in low frequencies. By increasing the arousal level of emotions, the higher harmonics becomes stronger.

The second major finding was that, the proposed time-frequency based features, SPs and HEs, are superior to traditional prosodic and spectral features in term of FDR score. However, their combination can significantly increase the classification accuracy.

Furthermore, the classification accuracy can be improved by employing a tandem classifier. Such a classifier can reduce the misclassified samples resulted from indirect classification of emotions pairs.

Since most of the prosodic and spectral features extracted based on time and frequency based methods, respectively, extracting features from spectrogram can be efficient to fill the gap by providing time-frequency features. According to our experiments these features can provide useful complementary information for prosodic and spectral features.

With possible refinement in future works, the performance of extracted features from the spectrogram may be further improved. Since most of the acoustic features related to arousal, further research on finding effective features to classify valence related emotions can improve the SER systems.

REFERENCES

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

[2] B. Yang, M. Lugger, "Emotion recognition from speech signals using new harmony features". Signal Processing, 90, 2010, 1415–1423.

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

[4] B. Schuller, G. Rigoll & M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", in: Proceedings of the ICASSP 2004, vol. 1, pp. 577–580, 2004.

[5] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Trans. Biomedical Eng. 47 (7) ,829–837,2007.

[6] M. Hariharan1, M. P. Paulraj, S. Yaacob, Time-Domain Features And Probabilistic Neural Network For The Detection Of Vocal Fold Pathology, Malaysian Journal of Computer Science, pp 60-67. 2010.

[7] J. Rong, G. Li, Y.P. Phoebe Chen, "Acoustic feature selection for automatic emotion recognition from speech". Information Processing and Management, 45, 2009, 315–328.

[8] J. Cahn, "The generation of affect in synthesized speech", J. Am. Voice Input/ Output Soc. 8, 1990, 1–19.

[9] M.B. Mustafa, R.N. Ainon1, R. Zainuddin, Z.M. Don, G. Knowles, S. Mokhtar, Prosodic Analysis And Modelling For Malay Emotional Speech Synthesis, Malaysian Journal of Computer Science, pp 102-110. 2010.

[10] L. Sherwood, "HUMAN PHISIOLOGY From Cells to Systems", eight edition., Cengage Learning. Library of Congress Control Number: 2011939366, 2010.

[11] N. Kamaruddin, A. Wahab, C. Quek, "Cultural dependency analysis for understanding speech emotion". Expert Systems with Applications, 28, 2011.

[12] S. Wu, T.H. Falk, W.Y. Chan, "Automatic speech emotion recognition using modulation spectral features". Speech communication, 53, 2011, 768–785.

[13] H. Altun, G. Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection". Expert Systems with Applications, 36, 2009, 8197–8203.

[14] D. Bitouk, R. Verma, A. Nenkova, "Class-level spectral features for emotion recognition". Speech Communication, 52, 2010, 613–625.

[15] E.M. Albornoz, D.H. Milone, H.L. Rufiner, "Spoken emotion recognition using hierarchical classifiers". Computer Speech and Language, 25, 2011, 556–570.

[16] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette,"Fear-type emotion recognition for future audio-based surveillance systems". Speech Communication, 50, 2008, 487–503.

[17] T. Polzehl, A. Schmitt, F. Metze, M. Wagner, "Anger recognition in speech using acoustic and linguistic cues". Speech Communication, 53, 2011, 1198–1209.

[18] H. Pérez-Espinosa, C.A. Reyes-García, L. Villasenor-Pineda, "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model". Biomedical Signal Processing and Control, 2011, 02.008.

[19] M. Kockmann, L. Burget, J. Cernocky, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition". Speech Communication, 53, 2011, 1172–1185.

[20] C.C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach". Speech Communication, 53, 2011, 1162–1171.

[21] 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". Computer Speech and Language, 25, 2011, 84–104.

[22] E. Bozkurt, E. Erzin, C.E. Erdem, A.T. Erdem, "Formant position based weighted spectral features for emotion recognition". Speech Communication, 53, 2011, 1186–1197.

[23] E. Vayrynen, J. Toivanen, T. Seppanen, "Classification of emotion in spoken Finnish using vowel-length segments: Increasing reliability with a fusion technique". Speech Communication, 53, 2011, 269–282.

[24] C.T. Ishi, H. Ishiguro, N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality". Speech Communication, 50, 2008, 531–543.

[25] I. Murray, J. Arnott, "Toward a simulation of emotions in synthetic speech: A review of the literature on human vocal emotion". J. Acoust. Soc. Am. 93, 2, 1993, 1097–1108.

[26] L.C.D. Silva, T. Miyasato, R. Nakatsu, "Facial emotion recognition using multimodal information. In: Proceedings of the IEEE International Conference on Information". Communications and Signal Processing (ICICS'97), 1997, 397–401.

[27] L. Rabiner, R. Schafer, "Digital Processing of Speech Signals", first ed., Pearson Education, 1978.

[28] L. Leinonen, T. Hiltunen, "Expression of emotional-motivational connotations with a one-word utterance". J. Acoust. Soc. Am. 102, 3, 1997, 1853–1863.

[29] T. New, S. Foo, L.D. Silva, "Speech emotion recognition using hidden Markov models". Speech Commun, 41, 2003, 603–623.

[30] M.B. Goudbeek, J.P. Goldman, K. Scherer, "Emotion dimensions and formant position". 10th Annual Conference of the International Speech Communication Association, 2009, 1575-1578.

[31] 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". Biomedical Signal Processing and Control, 6, 2011, 139–146.

[32] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection". Journal of Machine Learning Research, 3, 2003, 1157-1182.

[33] Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2005). Features extraction and selection for emotional speech classification. In Proceedings of IEEE conference on advanced video and signal based surveillance (AVSS) (pp. 411–416).

[34] Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In Proceedings of the 2004 international symposium on circuits and systems (ISCAS'04) (Vol. 2). Canada: Vancouver.

[35] D. Ververidis, C. Kotropoulos, Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal Processing 88 (2008) 2956–2970.

[36] R.G. Raj, V. Balakrishnan. "A Model for Determining The Degree of Contradictions in Information". Malaysian Journal of Computer Science, 2011, 24(3): 160 – 167.

[37] Park, C.-H., & Sim, K.-B. (2003). Emotion recognition and acoustic analysis from speech signal. In Proceedings of the international joint conference on neural networks (IJCNN'03) (Vol. 4, pp. 2594–2598).

[38] Chuang, Z.-J., & Wu, C.-H. (2004). Emotion recognition using acoustic features and textual content. In Proceedings of IEEE international conference on multimedia and expo (ICME'04) (Vol. 1, pp. 53–56). IEEE Computer Society.

[39] Hoch, S., Althoff, F., McGlaun, G., & Rigoll, G. (2005). Bimodal fusion of emotional data in an automotive environment. In Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP'05) (Vol. 2, pp. 1085–1088). IEEE Computer Society.

[40] Lee, C. M., Narayanan, S., & Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. In Proceedings of the seventh international conference on spoken language processing (ICSLP'02). Denver, CO, USA.

[41] Inanoglu, Z., & Caneel, R. (2005). Emotive alert: Hmm-based emotion detection in voicemail messages. In Proceedings of the 10th international conference on intelligent user interfaces (IUI'05), no. 585. San Diego, California, USA: ACM Press.

[42] Shafran, L., Riley, M., & Mohri, M. (2003). Voice signatures. In Proceedings of The eighth IEEE automatic speech recognition and understanding workshop (ASRU 2003).

[43] Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden markov model-based speech emotion recognition. In Proceedings of the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP'03) (Vol. 2, pp. 1–4). IEEE Computer Society.

[44] Song, M., Chen, C., & You, M. (2004). Audio-visual based emotion recognition using tripled hidden markov model. In Proceedings of IEEE international conference on acoustic, speech and signal processing (ICASSP'04) (Vol. 5, pp. 877–880). IEEE Computer Society.

[45] 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 Recognition 44 (2011) 2749–2759.

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

[47] 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, Japan, 1986, pp. 1197–1200.

[48] P.G. Vilda, J..M. Ferrandez-Vicent, V. Rodellar-Biarge, R. Fernandez-Baillo, Time-frequency representations in speech perception, Neurocomputing 72 (2009) 820–830.

[49] D. Ververidis, C. Kotropoulos, Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, July 2005, pp. 1500–1503.

[50] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[51] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.

[52] M.T. Shami, M.S. Kamel, Segment-based approach to the recognition of emotions in speech, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, 2005, 4pp.

[53] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. IEEE Signal Processing Lett., 6 (1): 1–3, 1999.

[54] L. Rabiner, B.H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, Englewood Cliffs, NJ, 1993.

[55] T. Kinnunen, H. Li, “An overview of text-independent speaker recognition: from features to supervectors”, Speech Communication, 52, 1, 2010, 12–40.

[56] R.C. Gonzalez, R.E. Woods, Digital Image Processing, second ed., Prentice Hall, NJ, 2001.

[57] K. M. Indrebo, R. J. Povinelli, M. T. Johnson, “Sub-banded reconstructed phase space for speech recognition”, Speech Communication, 48, 2006, 750-774.

[58] B. Gold, N. Morgan, “Speech and Audio Signal Processing”. John Wiley and Sons, New York, 2000.

[59] Z. Xiao, E. Dellandrea, W. Dou, L. Chen, “Automatic Hierarchical Classification of Emotional Speech”. Ninth IEEE International Symposium on Multimedia Workshops, ISMW '07, 2007, 56.

[60] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, “A database of German emotional speech”. Interspeech, 2005, 1517–1520.

[61] M. Kotti, C. Kotropoulos, “Gender classification in two Emotional Speech databases”. 19th International Conference on Pattern Recognition, ICPR 2008, 2008, 4761624.

BIOGRAPHY

Ali Shahzadi holds a PhD from Iran University of Science and Technology. At the present he is Faculty Member of the Semnan University. His work deals with communication systems, signal processing and pattern recognition.

Alireza Ahmadyfard obtained a PhD in computer vision from Center for Vision and Signal Processing (CVSSP) at the university of surrey in UK in 2003. He is currently a Faculty Member of the Shahrood Univercity of Technology. His current research interests are pattern recognition, image processing and in general, signal processing.

Khashayar Yaghmaie, 55, died from heart attack in July 2012. He held a PhD from the University of Surrey in UK in 1997. He was also Faculty Member of the Semnan University. And his research was mostly related to speech processing and watermarking.

Ali Harimi received a Master of Science degree from Shahrood University of Technology in 2009. Following that, He accepted to continue his graduate studies in Semnan University. His work deals with speech emotion recognition, pattern recognition and image processing.