JIAS

Journal of Intelligent Automation Systems
Published online: 23 November 2013

Speech Emotion Recognition Using Radon and Discrete Cosine Transform-Based Features from Speech Spectrogram

Ali Harimi[a,*], Ali Shahzadi[b], Alireza Ahmadyfard[c] and Khashayar Yaghmaie[d]

Abstract: Speech Emotion Recognition (SER) is a multi-disciplinary research area that has received increased attention over the last years. The aim of a SER system is to recognize human emotion by analyzing acoustics of speech sound to improve the voice-based human-machine interactions. This study presents a new feature extraction technique for SER using Radon transform and discrete cosine transform (DCT). In order to capture the acoustic characteristics of the spectrogram, Radon projections are computed for seven orientations. DCT applied on Radon projections results in low dimensional feature vector. The extracted features are further reduced in dimensions using a filtering feature selection algorithm based on fisher discriminant ratio. The classification accuracy of the proposed SER system has been evaluated using the 10-fold cross-validation technique on the Berlin database. The average recognition rate of 87.38% and 85.04% were achieved for females and males, respectively. By considering the total number of males and females samples, the overall recognition rate of 86.36% was obtained.

Keywords: Speech emotion recognition; Radon transform; Discrete cosine transform; Cross validation

**Corresponding author. a Electrical & Computer Engineering Faculty, Semnan University, Semnan, Iran. phone:+989153596039; e-mail:a.harimi@gmail.com*
b Electrical & Computer Engineering Faculty, Semnan University, Semnan, Iran
c Electrical Engineering Department, Shahrood University, Shahrood, Iran
d Electrical & Computer Engineering Faculty, Semnan University, Semnan, Iran

1. Introduction

Development of efficient systems for man-machine interaction requires algorithms, which should be able to understand human emotions as well as possible. Since speaking is the fastest and most natural method of communication between human beings, recognizing human emotions by analyzing, this signal has been proved to be efficient in establishing interaction between a human being and a machine [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]. The data encoded by the speaker and transmitted through these two channels is decoded by the listener. An advanced target system should be able to decode the data in these two channels. A noticeable amount of research in automatic speech recognition with the aim of extracting the linguistic information from the speech signal has been reported [2]. 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 and challenging research fields in the last few years.

Speech Emotion Recognition (SER) is commonly treated as a statistical pattern recognition problem. It consists of two major steps, feature extraction and classification. Since the extraction of distinctive features from patterns is a highly empirical issue and depends strongly on the application and database on hand [2], it is the main challenge in most SER systems and the problem is open now [1].

Since the speech signal is non-stationary in nature, it is common in speech processing to divide a speech signal into small segments called frames in which the signal is considered to be approximately stationary [3]. The so called prosodic features such as pitch and energy and also spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) and formants are extracted from each frame and called local (frame-level) features [1, 4]. Global (utterance-level) features, on the other hand, are calculated as statistics of all local features extracted from an utterance [1, 4]. Most researchers believe that global features have several advantages over local ones; they reduce classification time and provide higher classification accuracy. Furthermore, the number of global features is much less, and so the feature selection and cross validation algorithms can be performed faster by global features than by local ones [5-8]. However, most global features suffer from loss of temporal information [1].

Classification is the last stage of a SER system. In 1990s, most SER systems were based on the simple Maximum Likelihood Bayes algorithm (MLB) and Linear Discriminant Classification (LDC) [9]. Around 2000, Neural Network (NN) based classifiers became popular for emotion recognition [10-12]. Since 2002, Support Vector Machine (SVM) [13-16] and Hidden Markov Model (HMM) [17-20] have received more attention. Each classifier has its advantages and shortfalls, and the researchers are still trying to find the better solution [9].

Recently, authors in [4] have introduced Modulation Spectral Features (MSFs) as effective features in SER. They reached the recognition rate of 85.4% using a SVM based classifier under the 10 fold cross validation technique on the Berlin database. In [21], we extract spectral patterns and harmonic energies from speech spectrogram. A SVM based hierarchical classifier was employed to classify emotional speech. To this end, we used 90% of the database for training and the remaining 10% was used for testing. We also tested these features using a linear multi class SVM under the 10 fold cross validation [22]. In [23], the speech production system was modeled by phase space reconstruction and non-linear dynamics features were extracted from this model. These features were tested by a multi class SVM classifier under the 10 fold cross validation technique and the accuracy of 82.72% and 85.9% was obtained for females and males, respectively.

The main contribution of this work is employing radon and discrete cosine transforms based features for classifying emotional speech samples. The essence of this technique lies in formulating the emotion recognition problem into pattern recognition of images and resolving it using machine learning tools [24, 25]. The technique computes the Radon projections of the speech spectrogram in different directions to extract the voice patterns of speech. Discrete Cosine Transform (DCT) of Radon projection reduces the feature vector dimension to derive effective and efficient features. The technique has been applied for other applications such as speaker identification [25] and reported to be computationally efficient, robust to session variations and insensitive to additive noise.

The reminder of the paper is organized as follows. Section 2 presents the conventional prosodic and spectral features. The proposed method along with preprocessing, Radon transform and discrete cosine transform is described in section 3. Section 4 introduces the database employed. Experimental results are presented and discussed in Section 5. Finally, Section 6 gives concluding remarks.

2. Conventional prosodic and spectral features

In this section, we describe the prosodic and spectral features considered in our experiments. The prosodic and spectral features calculated here are by no means exhaustive, but serve as a representative sampling of the essential features. These features are used here as a benchmark, and also, to verify whether the proposed features can serve useful additions to the widely used prosodic and spectral features.

2.1. Prosodic features

Prosodic features are the most widely used features in SER [1, 4]. In order to extract these types of features, statistical properties of pitch and energy tracking contours are commonly used. Here, 20 time domain functions are applied to capture the statistics of pitch and energy 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, average deviation, standard deviation, skewness and kurtosis [26, 27]. These functions are also applied to the first and second derivatives of the contours as a common practice [4]. So, we have in total 60 pitch-based features and 60 energy-based features.

The widely used Zero-Crossing Rate (ZCR) and the Teager Energy Operator (TEO) [28] 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 [29]. The TEO for a discrete-time signal x_n is defined as:

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

In order to extract ZCR and TEO related features, we apply the 20 statistical functions to ZCR and TEO curves and their deltas and double deltas. Finally, we have 240 prosodic features.

2.2. Spectral features

The Mel-Frequency Cepstral Coefficients (MFCCs) and formants are reported as effective spectral features for emotion recognition [30-33]. Here, the first 12 MFCCs and 4 formants are extracted from 20 ms Hamming-windowed speech frames every 10 ms, and so their contours are formed. Finally, the 20 functions described in section 2.3 are applied to extract spectral features from the extracted contours and their first and second derivatives. In total, 960 spectral features are extracted here.

3. Proposed method

The proposed feature extraction technique is schematically shown in Figure 1. In this scheme, the pre-processing block transforms the raw speech signal into a spectrogram. Then, seven Radon projections of the spectrogram are computed in different orientations. Due to excellent data compaction property of DCT, it is employed here to reduce feature vector dimension. Significant DCT coefficients of Radon projections are concatenated to form the final feature vector. These features have been successfully employed for speaker identification [25].

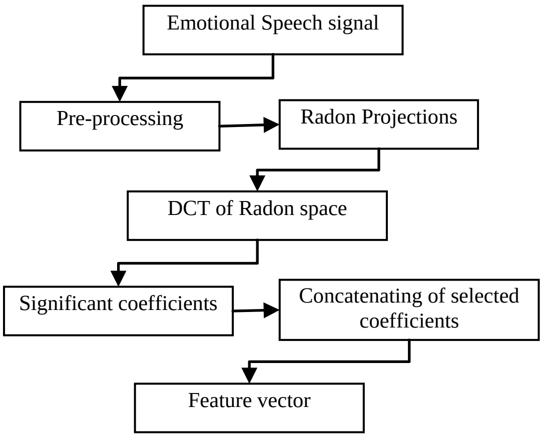


Fig. 1. The proposed feature extraction technique.

3.1. Speech Spectrogram

Here, all the samples have same length of 2.54s. Longer samples are cut and shorter ones are duplicated to reach the desirable length. In order to boost the higher frequencies of the speech signal, the input speech signal x[n] is commonly preemphasized by a first order filter with transfer function:

$$H(z)=1-\alpha z^{-1}, 0.9\le \alpha \le 1 \qquad (2)$$

Where α is set to 0.95, as it is suggested by [3]. Although the speech signal is non-stationary in nature, it can be assumed stationary over short period of time. Hence, frames of 320 samples length (20ms) with the frame shift of 160 samples (10 ms) are extracted here. The edge effects at the two ends of the frames are reduced using hamming windows multiplied with each frame. N=512 length Discrete Fourier Transform (DFT) of a windowed frame is computed to reach the power spectrum as:

$$S_i(k)=\log_{10}((\mathrm{Re}\{X_i(k)\})^2+(\mathrm{Im}\{X_i(k)\})^2), \qquad (3)$$
$$k=0,1,...,N-1, i=1,2,...,M$$

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). Re {…} and Im {…} indicate real and imaginary parts, respectively. The spectrums of these frames, $S_i(k)$, are concatenated row-wise to construct the speech spectrogram, f(k,i), as [25]:

$$f(k,i)=\begin{bmatrix} S_1(0) & ... & S_M(0) \\ . & . & . \\ . & . & . \\ S_1(N-1) & ... & S_M(N-1) \end{bmatrix} \qquad (4)$$

3.2. Radon transform

In this work, the speech spectrogram is treated as an image. In order to capture the directional features of the image f(x,y), the Radon transform is employed as [34]:

$$R(r,\theta)= \qquad (5)$$
$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\delta(r-x\cos\theta-y\sin\theta)dxdy$$

Where r is the distance to the origin of a line, $\theta\in[0,\pi]$ is the angle between the distance vector and x-axis and δ(.) is the Dirac function. According to equation (5), for a given image, the radon transform adds up the intensity values of the pixels along a straight line in a particular direction at a specific displacement [35].

Figure 2(a)-(f) shows the speech spectrogram of two sentences uttered by two speakers under three different emotional states: anger, neutral and boredom while Figure 2(g)-(l) shows their corresponding Radon projections at an angle of 90°.

As can be seen from Figure 2(a)-(f), harmonics, their position, stability and evolution are strongly related to the emotional state of the speaker. It has been reported that in high arousal emotions such as anger the resultant speech would be loud and fast with a higher pitch average and wider pitch range [1]. These conditions result in the presence of strong high-frequency energy in the corresponding speech signal [1]. In low arousal emotions such as boredom, on the other hand, the resultant speech would be slow, low pitched and with little high-frequency energy [1]. Figure 2(g)-(l) reveals that the shape of Radon projection curves are deeply depends on the emotional state of the speaker.

3.3. Discrete cosine transform

DCT is a well-known signal analysis tool employed in data compression due to its excellent energy compaction property for highly correlated data. The DCT of a signal x(n) is given as [36]:

$$X[k]=2\sum_{n=0}^{N-1}\alpha[n]x[n]\cos(\frac{\pi kn}{N-1}), \qquad (6)$$
$$0\le k\le N-1,$$
$$\alpha[n]=\begin{cases}0.5, n=0 and N-1\\ 1, 1\le n\le N-2\end{cases}$$

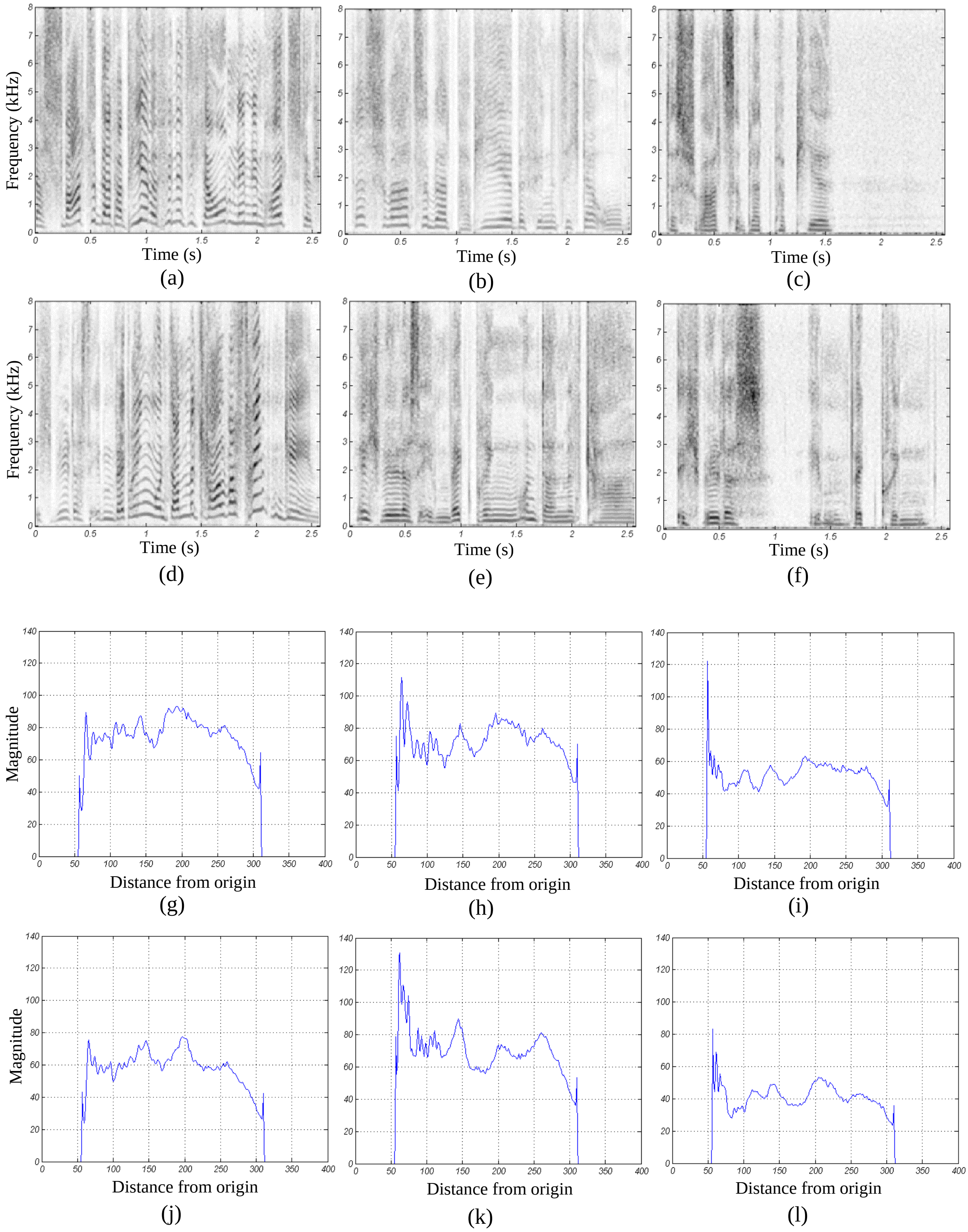
where k and N denote the frequency and length of x[n], respectively. DCT applied on Radon projections yields low dimensional feature vector.

4. Emotional speech data

The Berlin database of German emotional speech [37] is a well-known public database. The performance of many SER systems has been evaluated using this database [4, 37-41]. This database includes 535 utterances with 10 different contexts expressed by ten professional actors (5 males and 5 females) in 7 emotions. Table 1 lists the numbers of samples for the emotion categories.

Table 1. Number of samples in the Berlin database.

Emotion	Female	Male
Anger	67	60
Joy	44	27
Boredom	46	35
Neutral	40	39
Disgust	35	11
Fear	32	37
Sadness	37	25
All	301	234



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 2: Spectrograms and Radon projections of the same sentence uttered by two speakers under three different emotions: (a)-(c) spectrogram of anger, boredom and neutral emotional sentences uttered by speaker 1, (d)-(f) spectrogram of anger, boredom and neutral emotional sentences uttered by speaker 2, (g)-(i) Radon projections of spectrograms (a)-(c) at an angle 90°, (j)-(l) Radon projections of spectrograms (d)-(f) at an angle 90°.

5. Experimental results and discussions

In this study, it is assumed that a gender classifier with perfect classification accuracy, which is proposed by [42], is employed in the first stage, so the system is implemented completely separate for males and females. Features from training data are linearly scaled to [-1, 1] before applying linear Support Vector Machine (SVM). Features from test data are also scaled using the trained linear mapping function [4]. In order to avoid the curse of dimensionality [43], a filter-based feature selection scheme based on the Fisher Discriminant Ratio (FDR) is employed to remove irrelevant features. FDR evaluates features by means of measuring the inter-classes distance against the intra-class similarity as [4]:

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

$$1 \le c_1 < c_2 \le C$$

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, respectively. $i = 1, 2, \ldots, C$, and C is the total number of classes. Features with little discrimination ratio can then be removed by a thresholding process. The features proposed here are firstly compared to conventional prosodic and spectral features, using FDR criterion before applied to the classifier. To this end, features are ranked by their FDR values using all samples in the Berlin database and then, FDR values averaged over the top N_{fdr} FDR-ranked features, shown in Figures 3 and 4 for females and males, respectively.

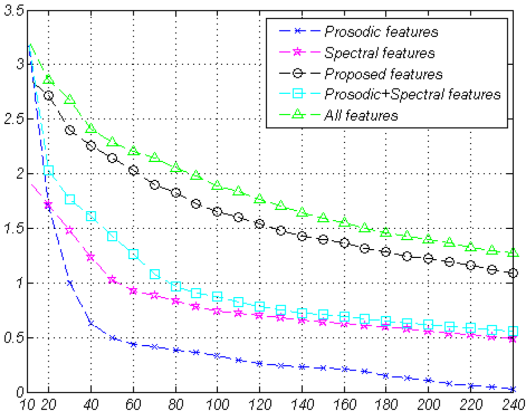


Fig. 3. FDR curves for different types of features (females)

Since there are only 240 prosodic features, all the curves depicted in Figures 3 and 4 computed for the same number of features. FDR curves can be considered as a rough indicator for discrimination power of features regardless of the employed classifier. As can be seen from Figures 3 and 4, the proposed features are superior to the conventional prosodic and spectral features in terms of FDR score. However, combining the prosodic and spectral features to the proposed ones can slightly upgrade the FDR curves. Interestingly, for all types of features, the average FDR of females are higher than males. This shows that the extracted features are more discriminative for females' emotions.

As mentioned earlier, FDR can only evaluate discriminative power of each feature individually. In order to evaluate power of a feature set in classification, we use directly classification accuracy as a criterion. The classification accuracy represents the performance of the employed classifier as a function of N_{top} features selected by the FDR-based feature selection algorithm. The classification accuracy is determined as the number of samples correctly recognized divided by the total number of samples. As a common practice for small sample size problems [44], results are produced using 10-fold cross-validation here. In this technique, each class has been randomly divided into 10 non-overlapping subsets approximately equal in size. In each validation trial, nine subsets from each class are taken for training, and the remaining one kept unseen until the testing phase. The overall recognition rate is achieved by averaging over the results of the 10 validation trials. We determine the accuracy curves for different types of features using a linear multi-class SVM. The computed curves are depicted in Figures 5 and 6 for females and males, respectively.

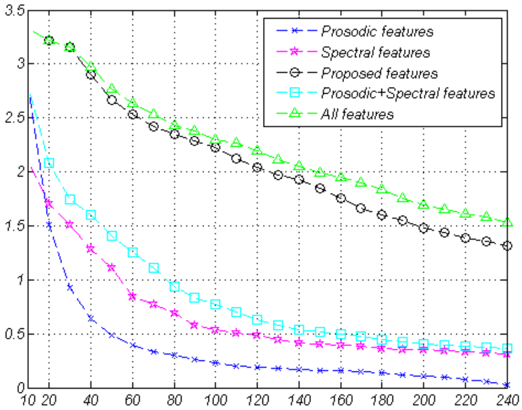


Fig. 4. FDR curves for different types of features (males)

According to Figures 5 and 6, for both females and males, spectral features are superior to prosodic and proposed features. However, when the proposed features are combined with the prosodic and spectral features, the maximum recognition rates of 87.38% and 85.04% are achievable using top 700 and 900 FDR features, for females and males, respectively. Interestingly, all accuracy curves suggest that females' emotions can be classified more accurately than males' emotions. This may be due to the fact that females are more emotionally perceptive and emotional stimuli than males [45].

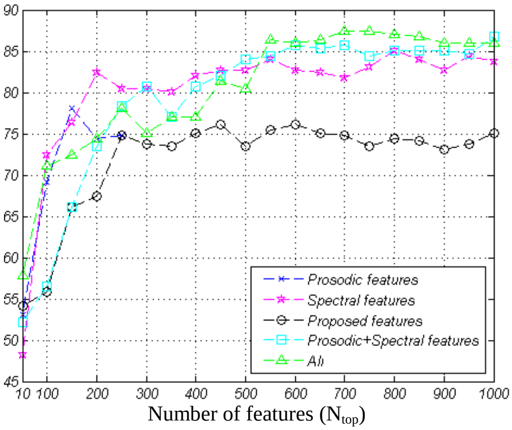


Fig. 5. Accuracy curves for different types of features (females)

The best results achieved by applying the proposed classifier for classifying 7 emotions using the combination of all types of features are represented in Tables 2 and 3 by two confusion matrices for females and males, respectively.

In these Tables, the left-most column is the true classes and the top row indicates the recognized classes. Furthermore, the rate column shows the average recognition rate for each class, which is determined as the total number of samples in the class divides the number of samples correctly recognized. The precision of each class is calculated as the number of samples is correctly classified divided by the total number of samples assigned to the class.

As can be seen from Tables 2 and 3, the ambiguity in classification of anger vs. joy and also boredom vs. neutral are responsible for major part of error in the proposed classifier. This may be due to the fact that most of the acoustic features employed for SER are related to arousal [46], and so they are not discriminative for valence related emotions such as anger and joy [4, 46].

By considering the total number of 301 female and 234 male samples, the overall recognition rate of 86.36% is obtained for the proposed SER system.

It can be also useful to review performance Figures reported on the Berlin database by other works. Although the numbers cannot be fairly compared due to different conditions of experiments such as different data partitioning, they can be useful for general benchmarking.

Table 2. Confusion matrix for classification of 7 emotions using combination of all types of features (females).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate (%)
Anger	**56**	0	1	2	8	0	0	83.58
Boredom	0	**41**	0	0	0	5	0	89.13
Disgust	0	0	**34**	1	0	0	0	97.14
Fear	2	0	1	**27**	1	0	1	84.38
Joy	7	1	1	0	**34**	1	0	77.27
Neutral	0	5	0	0	0	**35**	0	87.50
Sadness	0	0	0	0	0	1	**36**	97.30
Precision (%)	86.15	87.23	91.89	90.00	79.07	83.33	97.30	
Overall accuracy: 87.38%								

Table 3. Confusion matrix for classification of 7 emotions using combination of all types of features (males).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate (%)
Anger	**57**	0	0	1	2	0	0	95.00
Boredom	0	**27**	0	0	0	4	4	77.14
Disgust	0	0	**8**	3	0	0	0	72.73
Fear	2	0	1	**32**	1	1	0	86.49
Joy	5	0	0	2	**20**	0	0	74.07
Neutral	0	4	1	0	0	**33**	1	84.62
Sadness	0	3	0	0	0	0	**22**	88.00
Precision (%)	89.06	79.41	80.00	84.21	86.96	86.84	81.48	
Overall accuracy: 85.04%								

The recognition rate of 88.8% is reported by employing a three-stage classification scheme for recognizing only six emotions [39]. In [4], 85.6% accuracy is obtained under 10-fold cross-validation for classifying 7 emotions. In [47], the best average recognition rate of 85.5% is reported using a multi-class SVM classifier. In [21], by using spectral patterns and harmonic energies and a SVM based hierarchical classifier the best recognition rate of 86.9% was obtained. Similar features were also tested using a linear multi class SVM under the 10 fold cross validation and the best classification accuracy of 86.91% was achieved [22]. In [23], non-linear dynamics features were applied to a multi class SVM classifier and the 10 fold cross validation was performed. The accuracy of 82.72% and 85.9% was obtained for females and males, respectively.

6. Conclusion

The aim of this study was to evaluate the proposed features extracted from speech spectrogram for the recognition of human emotions from speech. These features have also been compared to conventional prosodic and spectral features in terms of FDR score and classification accuracy. This paper has demonstrated the potential and promise of the proposed features for emotion recognition. The following conclusions can be drawn from the present study.

The first major finding was that, harmonics, their position, stability and evolution are mostly related to the emotional state of the speaker. This affects the behavior of energy bands and spectral patterns on the spectrogram image. Moreover, the properties of radon projections of speech spectrogram are deeply related to emotional state of the speaker.

The second major finding was that, although the proposed features are superior to conventional prosodic and spectral ones in term of FDR score, they are not the best in classification performance. However, these features boost the classification accuracy to 86.36% when used to augment the conventional prosodic and spectral features.

Also, our experiments reveal that females' emotions can be recognized more accurately (about 2.34%) than males' emotions. Moreover, most acoustic features employed for SER are discriminative for classifying emotions based on arousal level, while they are ineffective for classification of valence related emotions [4, 46]. This fact results in the ambiguity in classification of anger vs. joy and also boredom vs. neutral which is responsible for major part of error in most SER systems.

In order to improve the performance of the proposed SER systems, the structure of the classifier can be optimized. To this end, tandem classifiers can be employed for classification of valence related emotions. Also, finding effective features for classifying valence related emotions can be a beneficial research. Moreover, as ultimate aim of a speech emotion recognition system is to recognize emotions for real work data, evaluating the proposed system under different conditions such as in presence of noise and chatter is useful.

References

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

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

[3] L.R. Rabiner and R.W. Schafer, Digital processing of speech signals: Prentice-Hall, 1978.

[4] S. Wu, T.H. Falk and W-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Communication, **53**, 768–785, (2011).

[5] D. Ververidis and C. Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication, **48**, 1162–1181, (2006).

[6] M.T. Shami and M.S. Kamel, Segment-based approach to the recognition of emotions in speech, presented at the IEEE International Conference on Multimedia and Expo, 2005.

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

[8] H. Hu, M.X. Xu and W. Wu, Fusion of global statistical and segmental spectral features for speech emotion recognition, presented at the International Speech Communication Association—8th Annual Conference of the International Speech Communication Association, 2007.

[9] J. Rong, G. Li and Y-P.P. Chen, Acoustic feature selection for automatic emotion recognition from

speech, Information Processing & Management, **45**, 315–328, (2009).

[10] V.A. Petrushin, Emotion recognition in speech signal: experimental study, development, and application, presented at the Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 2000.

[11] C.H. Park and K.B. Sim, Emotion recognition and acoustic analysis from speech signal, presented at the international joint conference on neural networks (IJCNN'03), 2003.

[12] J. Nicholson, K. Takahashi and R. Nakatsu, Emotion recognition in speech using neural networks, presented at the sixth international conference on neural information processing (ICONIP'99), 1999.

[13] B. Schuller, G. Rigoll and M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, presented at the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP'04), 2004.

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

[15] S. Hoch, F. Althoff, G. McGlaun and G. Rigoll, Bimodal fusion of emotional data in an automotive environment, presented at the IEEE international conference on acoustics, speech, and signal processing (ICASSP'05), 2005.

[16] Z.J. Chuang and C.H. Wu, Emotion recognition using acoustic features and textual content, presented at the IEEE international conference on multimedia and expo (ICME'04), 2004.

[17] M. Song, J.C. Bu C. and N. Li, Audio-visual based emotion recognition-a new approach, presented at the IEEE computer society conference on computer vision and pattern recognition (CVPR0´ 4), 2004.

[18] L. Shafran, M. Riley and M. Mohri, Voice signatures, presented at the The eighth IEEE automatic speech recognition and understanding workshop (ASRU 2003), 2003.

[19] B. Schuller, G. Rigoll, and M. Lang, Hidden markov model-based speech emotion recognition, presented at the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP'03), 2003.

[20] Z. Inanoglu and R. Caneel, Emotive alert: Hmm-based motion detection in voicemail messages, presented at the 10th international conference on intelligent user interfaces (IUI'05), San Diego, California, 2005.

[21] A. Shahzadi, A.R. Ahmadyfard, A. Harimi and K. Yaghmaie, Rrcognition of emotion in speech using spectral patternS, Malaysian Journal of Computer Science, **26**, 140–158, (2013).

[22] A. Shahzadi, A.R. Ahmadyfard, A. Harimi and K. Yaghmaie, Classification of emotional speech using spectral pattern features, Journal of AI and Data Mining (JAIDM), In press, (2013).

[23] A. Shahzadi, A.R. Ahmadyfard, A. Harimi and K. Yaghmaie, Speech emotion recognition using non-linear dynamics features, Turkish Journal of Electrical Engineering and Computer Sciences, In press, (2013).

[24] K. Saeed and M.K. Nammous, A speech-and-speaker identification system: feature extraction, description, and classification of speech-signal image, IEEE Transactions on Industrial Electronics, **54**, 887–897, (2007).

[25] P.K. Ajmera, D.V. Jadhav and R.S. Holambe, Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram, Pattern Recognition, **44**, 2749–2759, (2011).

[26] B. Schuller, M. Wimmer, L.M. osenlechner, C. Kern and G. Rigoll, Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?, presented at the Proceedings International Conference on Acoustics, Speech, and Signal Processing, 2008.

[27] J. Krajewski, A. Batliner and M. Golz, Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach, Behavior Research Methods, **41**, 795–804, (2009).

[28] J. Kaiser, On a simple algorithm to calculate the 'energy' of a signal, presented at the Internat. Conf. on Acoustics, Speech and Signal Processing, 1990.

[29] G. Zhou, J.H.L. Hansen and J.F. Kaiser, Nonlinear feature based classification of speech under stress, Speech and Audio Processing, IEEE Transactions on, **9**, 201–216, (2001).

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

[31] C.-C. Lee, E. Mower, C. Busso, S. Lee and S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Communication, **53**, 1162–1171, (2011).

[32] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson and K. Elenius, Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation, Computer Speech & Language, **25**, 84–104, (2011).

[33] L. He, M. Lech, N.C. Maddage and 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**, 139–146, 2011.

[34] K. Jafari-Khouzani and H. Soltanian-Zadeh, Rotation invariant multi- resolution texture analysis using radon and wavelet transform, IEEE Transactions on Image Processing, **14**, 783–794, (2005).

[35] G. Beylkin, Discrete radon transform, IEEE Transactions on Acoustics Speech and Signal Processing, **35**, 162–172, (1987).

[36] A.V. Openheim, R.W. Schafer and J.R. Buck, Discrete-time signal processing. New Jersey: Prentice-Hall, 1999.

[37] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, A database of German emotional speech, presented at the Interspeech, 2005.

[38] B. Schuller, D. Seppi, A. Batliner, A. Maier and S. Steidl, Emotion recognition in the noise applying large acoustic feature sets, presented at the Speech Prosody, 2006.

[39] M. Lugger and B. Yang, The relevance of voice quality features in speaker independent emotion recognition, presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007.

[40] N. Kamaruddin, A. Wahab, and C. Quek, Cultural dependency analysis for understanding speech emotion, Expert Systems with Applications, **39**, 5115–5133, (2012).

[41] H. Abarbanel, Analysis of Observed Chaotic Data: Springer, 1996.

[42] M. Kotti and C. Kotropoulos, Gender classification in two Emotional Speech databases, presented at the 19th International Conference on Pattern Recognition (ICPR), 2008.

[43] C. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.

[44] J.R. Raudays and A.K. Jain, Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence, **13**, 252–264, (1991).

[45] S. Whittle, M. Yücel, M.B.H. Yap and N.B. Allen, Sex differences in the neural correlates of emotion: Evidence from neuroimaging, Biological Psychology, **87**, 319–333, (2011).

[46] E.H. Kim, K.H. Hyun, S.H. Kim and Y.K. Kwak, Improved Emotion Recognition With a Novel Speaker-Independent Feature, IEEE/ASME Transactions on Mechatronics, **14**, 317–325, (2009).

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