IJE TRANSACTIONS A: Basics Vol. 27, No. 1, (January 2014) 79-90

International Journal of Engineering

Journal Homepage: www.ije.ir

A Database for Automatic Persian Speech Emotion Recognition: Collection, Processing and Evaluation

Z. Esmaileyan *[a], H. Marvi[b]

[a] Department of Electrical engineering, Science and Research branch, Islamic Azad University, Shahrood, Iran
[b] Department of Electrical engineering and Robotics, Shahrood University of technology, Shahrood, Iran

PAPER INFO

Paper history:
Received 13 March 2013
Received in revised form 17 June 2013
Accepted 20 June 2013

Keywords:
Persian Emotional Speech Database
PDREC
Speech Emotion Recognition

ABSTRACT

Recent developments in robotics automation have motivated researchers to improve the efficiency of interactive systems by making a natural man-machine interaction. Since speech is the most popular method of communication, recognizing human emotions from speech signal becomes a challenging research topic known as Speech Emotion Recognition (SER). In this study, we propose a Persian emotional speech corpus collected from emotional sentences of drama radio programs. Moreover, we propose a new automatic speech emotion recognition system which is used both for spectral and prosodic feature simultaneously. We compare the proposed database with the public and widely used Berlin database. The proposed SER system is developed for females and males separately. Then, irrelevant features are removed using Fisher Discriminant Ratio (FDR) filtering feature selection technique. The selected features are further reduced in dimensions using Linear Discriminant Analysis (LDA) embedding feature reduction scheme. Finally, the samples are classified by a LDA classifier. The overall recognition rate of 55.74% and 47.28% is achieved on proposed database for females and males, respectively. Also, the average recognition rate of 78.64% and 73.40% are obtained for Berlin database for females and males, respectively.

doi: *10.5829/idosi.ije.2014.27.01a.11*

1. INTRODUCTION

Speech is the most popular method of communication between humans. This fast has motivated researchers to use speech signal as an efficient method for human-computer interaction. Despite widespread efforts, we are still far away from natural interaction. It has been recognized that human emotions by machine can improve the efficiency of interactive systems. This has introduced a relatively new and challenging field of research in speech processing known as Speech Emotion Recognition (SER). SER has a wide range of applications in interactive systems.

Nicholson et al. [1] suggested that SER can improve the performance of speech recognition systems. Schuller et al. [2] introduced the usefulness of SER in e-learning, computer games, in-car boards and every other application that requires natural interaction. Furthermore, SER can be used in artificial intelligence, robotics, medical science and psychology as reported by France et al. [3-5]. Hansen and Carins [6] suggested SER to be employed in telephone center and mobile communication.

SER can be studied as a pattern recognition problem. From this point of view, it is composed of three main parts: (1) feature extraction, (2) feature selection and (3) classification. Extraction of efficient features is one of the main challenges of SER. It is due to the fact that most of emotional features employed for SER do not only depend on emotion, but also on factors such as: speakers, style of speaking and speaking rate, as reported by Ayadi et al. [7].

Fernandez [8, 9] represented various emotions in a two dimensional arousal-valence emotion space, as depicted in Figure 1. Arousal refers to required energy for expressing an emotion. When an emotion is attractive it is known to have positive valence, and on the contrary, when an emotion is aversive, it has a negative valence.

Since emotions are usually culture dependent, most of the related works are focused on one-lingual

*Corresponding Author Email: z.esmaileyan@gmail.com (Z. Esmaileyan)

database. This can reduce effects of culture in emotion expression. However, Hozjan and Kacic [10] addressed the multi- lingual SER.

Emotional databases can be grouped into two main categories: natural database and acted database. Natural databases are collected from the ordinary human conversations in daily life. But, in acted databases professional actors are usually wanted to express emotional sentences. The final goal of a SER system is to recognize human's emotions from natural speech, but developing natural databases are very expensive and they are commonly restricted. Acted database, on the other hand, are easier to develop. Furthermore, as it is reported by Ververidis and Kotropoulos [11], recognizing acted emotions is easier than natural emotions for machines as well as humans. These facts lead researchers to develop and employ acted database for emotion recognition.

In this study, we introduce a Persian emotional database. Various acoustic features are extracted from the samples of this dataset. Extracted features are reduced in dimension using a two stage feature selection scheme based on Fisher Discriminant Ratio (FDR) filtering and Linear Discriminant Analysis (LDA) embedding. Finally, samples are classified by a LDA classifier using 10-fold cross validation technique.

The present study is organized as following. In section 2, we briefly introduce several known speech databases. In section 3, the proposed database is described. Section 4 details main parts of the proposed SER system. Section 5 represents experimental results and discussions. Finally, the paper ends with concluding remarks in section 6.

2. COMMON EMOTIONAL SPEECH DATABASE

In this section, we present a brief review on several known emotional speech databases. Emotional speech corpuses are created for variety of purposes. Based on our studies, the majority of the databases have been used for automatic speech recognition and speech synthesis. However, emotional database are useful for medical applications, emotion perception by human and virtual teacher. Emotions can be grouped into two main categories: primary emotions and social emotions. Miller and Stoeckel [12] reported that primary emotions include "anger, fear, sadness, disgust, surprise (frightful-startle) and happy" which appear in the first six months of the human infant's development. These emotions are also present in both humans and animals. These are innate and may be assumed genetically hard-wired, so they have similarities in all cultures and languages, as reported by Ross et al. [13, 14]. Social emotions such as "pride, pity, jealousy and embarrassment", on the other hand, are learned by

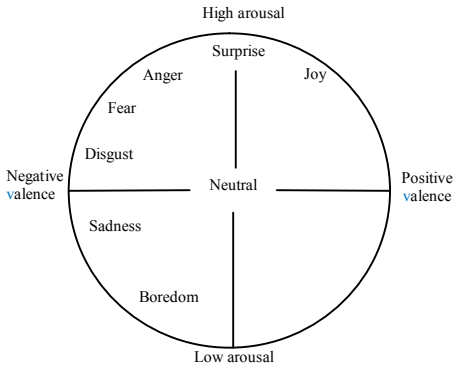

Figure 1. Two dimensional valence-arousal emotion space proposed by Fernandez [8,9].

human. Buck [13, 15-17] suggested that these emotions are culture dependent and follow special rules that vary across cultures. Social emotions could also be considered as a combination of the primary emotions, as suggested by Cosmides and Tooby [18]. Therefore, a noticeable number of speaker-language-independent SER researches focused on recognizing the primary emotions. In this regard, anger, sadness, happiness (joy), fear, disgust, surprise and boredom are the most common emotions presented in different databases. Table 1 summarizes characteristics of some databases commonly used in SER. In this table, the databases are characterized based on language, sample size and naturalness including emotions. As can be seen from this table, most of the databases are English, acted and include basic emotions.

3. THE PROPOSED PERSIAN DRAMA RADIO EMOTIONAL SPEECH CORPUS (PDREC)

Culture and language have a deep relationship with emotion expression. In this study, we develop a Persian emotional corpus in which the samples are collected from the several drama radio programs [1,2]. We named this database as Persian Drama Radio Emotional Corpus (PDREC). In PDREC**,** emotional sentences are cut from the radio programs using "Adobe Audition 3.0" and save as wav format. The proposed database includes 8 emotions: anger, boredom, disgust, fear, neutral, sadness, surprise and happiness (joy). The radio programs were recorded by the sampling rate of 44.1 kHZ. Since professional actors express emotional sentence of the utilized radio programs, the PDREC is an acted database. However, factors such as natural recording environments and background noise including car sound, sound of dripping water and children's playing sound have brought similarities to the natural database for the proposed database.

[1] http://www.radionamayesh.ir.

TABLE 1. Characteristics of common emotional speech databases

Database	Language	Size	Acted\ natural	Emotions
Natural, Morrison et al. [19]	Mandarin	388 utterances, 11 speakers, 2 emotions	natural	Anger, neutral
Baby Ears, Slaney and Roberts **[20]**	English	509 utterances, 12 actors (6 males + 6 females), 3 emotions	natural	Approval, attention, prohibition
Verbmobil [21]	German	29 male, 29 female	natural	Mainly anger, dissatisfaction
LDC Emotional Prosody Speech and Transcripts [22]	English	7actors *15 emotions *10 utterances	Acted	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
Berlin emotional database, Burkhardt et al**. [23]**	German	10actors *7 emotions *10 utterances+ some second version= 800 utterances	Acted	Anger, joy, sadness, fear, disgust, boredom, neutral
Danish emotional database, Engberg and Hansen **[24]**	Danish	4actors *5 emotions	Acted	Anger, joy, sadness, surprise, neutral
ESMBS, Nwe et al. **[25]**	Mandarin	720 utterances 12 speakers, 6 emotions	Acted	Anger, joy, sadness, disgust, fear, surprise
INTERFACE, Hozjan et al. **[26]**	English, Slovenian, Spanish, French	English 186 utterances Slovenian 190 utterances Spanish 184 utterances French 175 utterances	Acted	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral
KISMET, Breazeal and Aryananda **[27]**	American English	1002 utterances, 3 female speakers, 5 emotions	Acted	Approval, attention, prohibition, soothing, neutral
SUSAS, Hansen and Bou-Ghazale **[28]**	English	16,000 utterances, 32 actors (13 females + 19 males)	Acted	Four stress styles: Simulated Stress, Calibrated Workload Tracking Task, Acquisition and Compensatory Tracking Task, Amusement Park Roller-Coaster, Helicopter Cockpit Recordings
MPEG-4, Schuller et al. **[29]**	English	2440 utterances, 35 speakers	Acted	Joy, anger, disgust, fear, sadness, surprise, neutral
Beihang University, Fu et al**. [30]**	Mandarin	7actors *5 emotions *20 utterances	Acted	Anger, joy, sadness, disgust, surprise
FERMUS III, Schuller **[31]**	German, English	2829 utterances, 7 emotions, 13 actors	Acted	Anger, disgust, joy, neutral, sadness, surprise
SES, Petrushin **[32]**	Spanish	1 male actor* 4 emotion*30	Acted	sadness, happiness, anger, neutral
RUSSLANA, Makarova and Petrushin [33]	Russian	3660 sentences	Acted	Surprise, happy, anger, sad, fear and neutral
KES, Kim et al. **[34]**	Korean	5400 utterances 10 actors	Acted	Neutral, joy, sadness, anger
CLDC, Zhou et al. **[35]**	Chinese	1200 utterances, 4 actors	Acted	Joy, anger, surprise, fear, neutral, sadness
Hao Hu et al [36]	Chinese	8actors *5 emotions *40 utterances	Acted	Anger, fear, joy, sadness, neutral
Pereira, Hu et al. **[37]**	English	2actors *5 emotions *8 utterances	Acted	Hot anger, cold anger, joy, neutral, sadness

This database includes 748 utterances, which are expressed by 33 native speakers of Persian language. These speakers consist of 15 females and 18 males. Table 2 represents characteristics of the proposed database. In this table, the number of samples, average samples length, shortest samples length and longest samples length are presented for each emotion for females and males.

As seen from Table 2, the average samples lengths are 2.64 and 2.55 seconds for females and males,

respectively. The shortest sample length, with the duration of 0.58 second, belongs to surprise emotion of females. The longest sample length, with the duration of 10.35 second, belongs to boredom and neutral emotions of females.

Figure 2 (a) and (b) represent the amount of recordings from each emotion for the proposed PDREC database for females and males, respectively. As seen from Figure 2, while boredom, disgust and surprise form the smallest parts of the proposed database, anger, neutral and sadness construct the major parts of it. It is because of the conceptual properties of the dialogues of the drama radio programs. In this study, we use the 5 larger size emotions: anger, fear, neutral, sadness and joy.

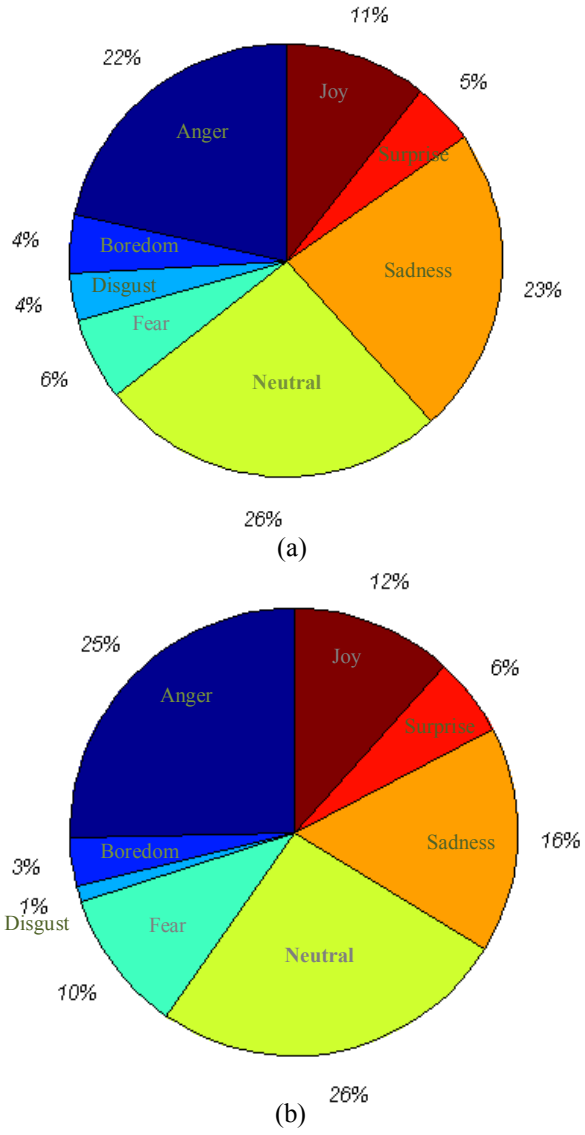

Figure 2. Amount of recordings from each emotion for the proposed PDREC database for (a) females and (b) males.

4. PROPOSED SPEECH EMOTION RECOGNITION SYSTEM

In order to evaluate the proposed database and compare it with the public and widely used Berlin database, we design a SER system which is schematically shown in Figure 3. As it is depicted in this figure, in the first stage of the proposed SER system, the signal is filtered using a pre-emphasis filter to accentuate the high-frequency content. The transfer function of the filter is as:

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

where α is set to 0.95 as it is suggested by Rabiner and Juang [38].

Then, the prosodic and spectral features are extracted from speech signal. Since a complete coverage of prosodic and spectral features is infeasible, we calculate here a representative sampling of the essential prosodic and spectral feature. As it is suggested by Wu et al. [39-43], statistics of pitch, energy and zero crossing rate tracking contours are used as prosodic features here.

Also, we employ spectral features extracted from Mel Frequency Cepstral Coefficients (MFCC) suggested by Wu et al. [39-44], Perceptual Linear Prediction (PLP) studied by Kim et al. [39, 41], Linear Prediction Coefficients (LPC) by Marvi et al. [45] and formants by Wu et al. [39, 41, 44]. In total, 181 prosodic and 2280 spectral features are employed here.

Then, irrelevant and noisy features are removed using a two stage filter and embedded feature selection algorithm. In this technique, firstly the features are individually ranked by FDR and features with less discriminant FDR score are removed from the raw feature vector.

Then, the features which are selected by the FDR filtering scheme are reduced in dimensions using LDA feature selection algorithm. Finally, the emotional speech samples are classified into the 5 following categories: anger, fear, joy, sadness and neutral using a LDA classifier. More details of the proposed algorithm are described in next subsections.

4. 1. Feature Extraction Steidl et al. [46] categorized most acoustic features, which have been employed for SER in two groups: prosodic features and spectral features.

4. 1. 1. Prosodic Features Krajewski and Kroger [46, 47] introduced the prosodic features as the most widely used features in SER. These features are usually extracted from the pitch and energy tracking contours. Lee et al. [42-44] showed that these contours convey important emotional information. Rong et al. [48, 49] used the statistics of these contours as prosodic features.

TABLE 2. Characteristics of the proposed Persian Drama Radio Emotional Database (PDREC).

Emotions	Number of samples		Average samples length (s)		Shortest sample length (s)		Longest sample length (s)	
	Female	Male	Female	Male	Female	Male	Female	Male
Anger	73	104	2.3	2.51	0.8	0.77	6.64	6.04
Boredom	15	14	4.2	2.78	1.56	1.36	10.35	5.32
Disgust	12	5	3.25	2.8	1.52	1.57	6	3.93
Fear	21	42	2.09	2.21	0.98	0.76	4.3	4.66
Neutral	88	106	2.81	2.83	1	0.71	10.35	7.3
Sadness	78	67	2.72	2.51	1.02	0.86	7.2	6.13
Surprise	16	23	2.59	2.22	0.58	0.7	8.36	4.36
Joy	36	48	2.11	2.45	0.85	1.34	6.58	6
All	339	409	2.64	2.55	0.58	0.7	10.35	7.3

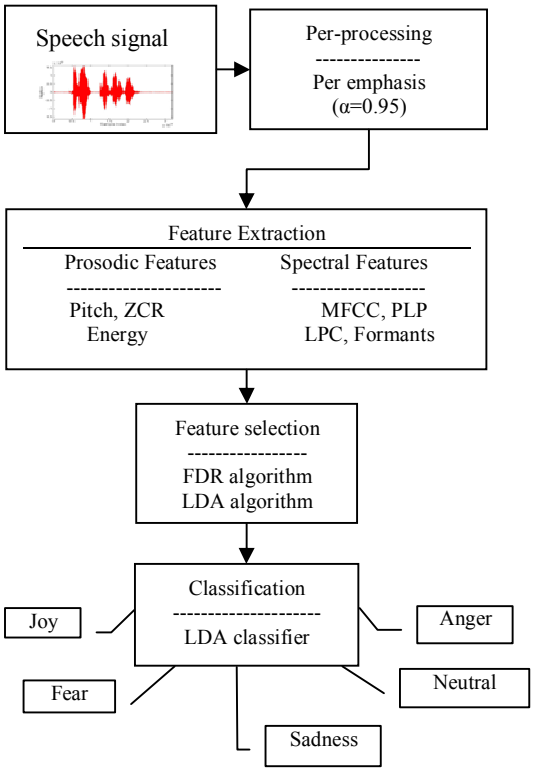

Figure 3. The proposed SER system.

In this paper, we employ 20 time-domain statistical functions employed by Wu et al. [39, 48, 50] to extract prosodic features from pitch and energy 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, these functions are also applied to the first and second derivatives of pitch and energy tracking contours. Also, Zero Crossing Rate (ZCR) contour is utilized here. ZCR does not directly relate to prosody but it is employed for SER by Rong et al. [39, 48].

4. 1. 2. Spectral Features Spectral features are usually extracted from the speech spectrum. Unlike the prosodic features which are determined based on time-domain analysis of the speech signal, these features are computed based on frequency-domain analysis techniques.

The spectral features employed in this work include: MFCC employed by Wu et al. [39-44], PLP studied by Kim et al. [39, 41], LPC by Marvi et al. [45] and Formants by Bozkurt et al. [39, 41, 44]. To this end, the first 12 MFCCs, the first 13 PLP coefficients, the 9 LPCs and the first 4 formants are calculated for 20 ms frames every 10 ms. Then, the corresponding contours are formed, and finally spectral features are extracted from these contours using 20 statistical functions described in section 4.1.1. These functions are also applied to the first and second derivatives of these contours. Tables 3 lists the emotional features employed here.

TABLE 3. List of extracted features.

Feature types and numbers
Prosodic features: (181)
Apply 20 statistical functions to:
Pitch, delta pitch, double-delta pitch,(61);
Energy, delta energy, double-delta energy,(60);
ZCR, ZCR delta, and ZCR-double delta,(60);
Spectral features: (2280)
Apply 20 statistical functions to:
12MFCCs, their deltas, and their double-deltas,(720);
4Formants, their deltas, and their double-deltas,(240);
9 LPC, their deltas, and their double-deltas,(540);
13 PLP, their deltas, and their double-deltas,(780);

4. 2. Feature Selection To avoid the curse of dimensionality, removing irrelevant and noisy features is an important part of most pattern recognition systems, as it is noted by Bishop [51]. To this end, we employ a feature selection algorithm based on FDR as a filter approach. FDR can evaluate individual features regardless of the classifier properties by means of measuring inter-class distance and intra-class similarity for each feature. For a C class's problem, the FDR of the U-th feature determined by Wu et al. [39] as:

$$FDR(u) = \frac{2}{c(c-1)} \sum_{c_1} \sum_{c_2} \frac{(\mu_{c_{1\mu}} - \mu_{c_{2\mu}})^2}{\sigma^2_{c_{1\mu}} + \sigma^2_{c_{2\mu}}} \quad (2)$$

where $1 \le c_1 \le c_2 \le C$, $\mu_{c_{i\mu}}$ and $\sigma^2_{c_{i\mu}}$ are the mean and variance of the u-th feature for the i-th class, respectively. Also, C is the total number of classes. According to Equation (2), FDR can be considered as a rough indicator for discrimination power of a feature. In this method, all of the features are ranked by their FDR valves. Then features with the less discriminative power are removed by a thresholding process. In this paper, the threshold is empirically set to 0.25.

Since effects such as features correlation and classifier properties cannot be accounted for the filter based feature selection algorithm such as FDR scheme, it is common to employ a second feature selection algorithm as organized by the LDA feature reduction algorithm here. LDA which is employed by Ye et al. [38, 39, 52] is an embedded based method that transforms features to a new space with smaller dimensions.

4. 3. Classification Classification is the final stage of a pattern recognition system. As it is suggested by Laukka et al. [53], we use a LDA classifier to categorize speech samples according to their emotional states.

5. EXPERIMENTAL RESULTS

Sarunas and Jain [54] introduced cross validation as a useful method for small sample size problems. To this end, we evaluate our SER system on the proposed PDREC and Berlin database using 10-fold cross validation approach. In this technique, all of the classes are divided into 10 non-overlapping subsets approximately equal in size. In each validation trial, 9 subsets are employed as the training set and the remaining one are used for testing. The overall recognition rate is achievable by averaging over the 10 validation results, as performed by Wu et al. [39].

In our experiments, the samples of the Berlin database and PDREC are classified into only 5 emotions: anger, fear, neutral, sadness and joy, duo to the leakage of sample size in other emotion categories. Table 4 represents the result of classification for these two databases using different types of features individually and in combination for females and males separately. In this table, prosodic features are composed of pitch, energy and ZCR based features. Also, spectral features include MFCC, PLP, LPC and formants.

According to the results of Table 4, the best recognition rates are achieved using all types of features, in both Berlin database and PDREC, for females and males. Figure 4 depicts the average recognition rates of different emotions using combination of all types of features.

As it can be seen from Figure 4, the average recognition rates obtained for PDREC are smaller than the classification performances achieved by the Berlin database. It may be due to the effects such as background noise and other similarities of the PDREC to the natural database described in section 3.

TABLE 4. Classification results using different types of features.

	Recognition rate for PDREC (%)		Recognition rate for Berlin (%)	
Feature type	**Female**	**Male**	**Female**	**Male**
MFCC	51.01	43.74	55.00	53.38
PLP	23.65	38.04	41.36	49.32
LPC	30.07	24.46	28.18	27.13
FORMANTS	46.28	44.57	54.09	50.68
ZCR	41.89	27.72	56.82	64.86
Energy	37.16	38.59	56.36	58.78
Pitch	42.23	39.95	66.82	56.08
Prosodic	48.65	42.66	71.36	68.09
Spectral	54.05	43.21	71.36	70.74
All features	**55.74**	**47.28**	**78.64**	**73.40**

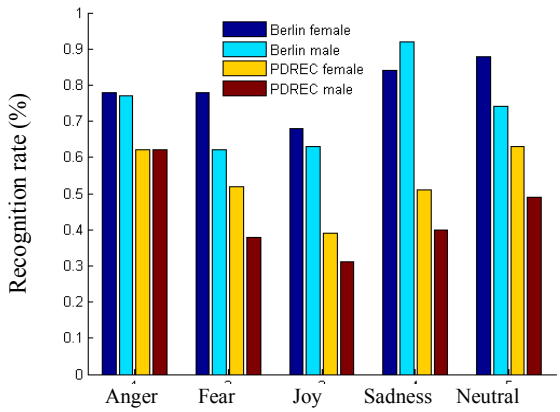

Figure 4. Average recognition rates on PDREC and Berlin databases achieved by combination of all types of features.

The confusion matrices of classifying 5 emotions on PDREC and Berlin databases using prosodic features are represented in Table 5 to 8 for females and males, respectively. In these tables the leftmost column is the true emotions and the top row indicates the recognized emotions. The "Rate" column, which shows the average recognition rate for each class, is determined as the number of samples correctly recognized divided by the total number of samples in the class. Here, each class is abbreviated by the first letter of the corresponding emotion (i.e. anger, fear, joy, sadness and neutral are represented by A, F, J, S and N, respectively).

The "Pre" row, which represents the precision of each class, is calculated as the number of samples correctly classified, divided by the total number of samples assigned to the class.

Comparing Tables 5 to 8 show that using prosodic features, for both PDREC and Berlin database, females' emotions can be recognized more accurately than males' emotions. Also, the average accuracies achieved by classifying Berlin database are higher than those obtained by classifying PDREC. Wu et al. [39] reported that the SER systems which use acted databases commonly achieved higher recognition rate than systems which employed natural databases.

Table 9 to 12 represent confusion matrices achieved for classifying 5 emotions on PDREC and Berlin databases using spectral features for females and males, respectively.

TABLE 5. Confusion matrix of classifying PDREC using prosodic features for females.

	A	F	J	S	N	Rate %
A	**36**	4	19	8	6	49.31
F	5	**6**	2	8	0	28.57
J	10	1	**15**	4	6	41.66
S	10	12	9	**32**	15	41.02
N	5	2	7	19	**55**	62.5
Pre.(%):	54.54	24.00	28.84	45.07	62.07	
Average recognition rate: 48.65						

TABLE 6. Confusion matrix of classifying Berlin using prosodic features for females.

	A	F	J	S	N	Rate %
A	**41**	11	14	0	1	61.19
F	5	**20**	4	3	0	62.5
J	8	2	**30**	1	3	68.18
S	0	1	0	**31**	5	83.78
N	0	0	3	2	**35**	87.5
Pre.(%):	75.92	58.82	58.82	83.78	79.54	
Average recognition rate: 71.36						

TABLE 7. Confusion matrix of classifying PDREC using prosodic features for males.

	A	F	J	S	N	Rate %
A	**48**	21	14	16	5	46.15
F	8	**17**	1	7	9	40.48
J	10	3	**14**	9	13	28.57
S	3	13	11	**21**	19	31.34
N	1	5	19	24	**57**	53.77
Pre.(%):	68.57	28.81	23.73	27.27	55.34	
Average recognition rate: 42.66						

TABLE 8. Confusion matrix of classifying Berlin using prosodic features for males.

	A	F	J	S	N	Rate %
A	**39**	6	14	0	1	65.00
F	2	**26**	0	1	8	70.27
J	12	2	**12**	0	1	44.44
S	0	0	0	**22**	3	88.00
N	0	3	0	7	**29**	74.36
Pre.(%):	73.58	70.27	46.15	73.33	69.04	
Average recognition rate: 68.09						

TABLE 9. Confusion matrix of classifying PDREC using spectral features for females.

	A	F	J	S	N	Rate %
A	**42**	4	14	6	7	57.35
F	1	**11**	1	6	2	52.38
J	11	1	**15**	8	1	41.67
S	5	9	6	**40**	18	51.28
N	10	2	6	18	**52**	59.09
Pre.(%):	60.67	40.74	35.71	51.28	65.00	
Average recognition rate: 54.05						

TABLE 10. Confusion matrix of classifying Berlin using spectral features for females.

	A	F	J	S	N	Rate %
A	**44**	5	9	5	4	65.67
F	4	**22**	2	3	1	70.97
J	9	0	**29**	1	5	65.91
S	1	1	6	**29**	0	78.38
N	6	0	0	1	**33**	82.50
Pre.(%):	68.75	78.57	63.04	74.36	76.74	
Average recognition rate: 71.36						

TABLE 11. Confusion matrix of classifying PDREC using spectral features for males.

	A	F	J	S	N	Rate %
A	**60**	13	17	5	9	57.69
F	8	**16**	3	9	6	38.09
J	13	1	**13**	9	13	26.53
S	2	10	12	**25**	18	37.31
N	9	18	16	28	**45**	42.45
Pre.(%):	62.22	33.33	21.31	32.89	49.45	
Average recognition rate: 43.21						

TABLE 12. Confusion matrix of classifying Berlin using spectral features for males.

	A	F	J	S	N	Rate %
A	**47**	4	2	2	5	78.33
F	7	**20**	3	2	5	54.05
J	6	2	**16**	1	2	59.26
S	3	0	0	**22**	0	88.00
N	4	5	1	1	**28**	71.79
Pre.(%):	70.15	64.51	69.56	78.57	70.00	
Average recognition rate: 70.74						

TABLE 13. Confusion matrix of classifying PDREC using combination of all types of features for females.

	A	F	J	S	N	Rate %
A	**45**	3	15	5	5	61.64
F	1	**11**	1	6	2	52.38
J	13	1	**14**	5	3	38.88
S	7	11	8	**40**	12	51.28
N	5	5	7	16	**55**	62.50
Pre.(%):	63.38	35.48	31.11	55.55	71.43	
Average recognition rate: 55.74						

TABLE 14. Confusion matrix of classifying Berlin using combination of all types of features for females.

	A	F	J	S	N	Rate %
A	**52**	4	7	0	4	77.61
F	2	**25**	1	3	1	78.12
J	12	0	**30**	0	2	68.18
S	1	1	4	**31**	0	83.78
N	4	0	1	0	**35**	87.50
Pre.(%):	73.24	83.33	68.18	91.18	83.33	
Average recognition rate: 78.64						

TABLE 15. Confusion matrix of classifying PDREC using combination of all types of features for males.

	A	F	J	S	N	Rate %
A	**64**	12	19	3	6	61.54
F	8	**16**	4	10	4	38.09
J	8	2	**15**	10	14	30.61
S	2	8	14	**27**	16	40.30
N	10	2	19	23	**52**	49.06
Pre.(%):	69.56	40.00	21.13	36.97	56.52	
Average recognition rate: 47.28						

TABLE 16. Confusion matrix of classifying Berlin using combination of all types of features for males.

	A	F	J	S	N	Rate %
A	**46**	5	3	1	5	76.67
F	5	**23**	1	1	7	62.16
J	1	4	**17**	1	4	62.96
S	1	1	0	**23**	0	92.00
N	5	3	1	1	**29**	74.36
Pre.(%):	79.31	63.88	72.27	85.18	64.44	
Average recognition rate: 73.40						

Comparing Tables 9 to 12 reveal that by using spectral features, Berlin database is classified more accurately than PDREC. Interestingly, for both Berlin database and PDREC, especially for females, spectral features are more effective than prosodic ones in terms of classification accuracy. Table 13 to 16 represent confusion matrices achieved on PDREC and Berlin databases using combination of all types of features.

According to the results of Tables 13 to 16, combining prosodic and spectral features improves the recognition rate for both females and males on both Berlin database and PDREC. Exploring the results of these confusion matrices reveals that in both databases, emotions with similar arousal level such as anger, fear and joy are responsible for majority parts of classification error. It is due to the fact that, most of the acoustic features are related to arousal, as suggested by Kim et al. [41]. Thus, arousal-related emotions are easier to classify than valence-related emotions, as reported by Wu et al. [39].

6. CONCLUSION

The purpose of the current study was to propose and evaluate a new Persian emotional speech database named as PDREC. The samples of PDREC were collected from emotional sentences of drama radio

programs. The proposed study was compared to the public and widely used Berlin emotional speech corpus. This paper has demonstrated that the proposed database can be useful in SER researches on Persian language. The following conclusions can be drawn from the present study.

Firstly, in both Berlin database and PDREC, females' emotions can be recognized more accurately than males' emotions when prosodic and spectral features are employed individually and in combination.

Secondly, for both Berlin database and PDREC, especially for females, spectral features are more effective than prosodic features in terms of classification performance. However, the best classification performance is achievable when the combination of prosodic and spectral features is used.

Moreover, the samples of the Berlin database can be classified more accurately than those of PDREC. This may be due to the effects such as background noise and other similarities of the PDREC to the natural databases, which are commonly harder to classify than acted databases.

As further work, we plan to improve the performance of the proposed system by investigating more effective features that can classify noisy speech such as samples presented in PDREC. Furthermore, investigating for more effective feature selection and classification techniques can be a useful research.

7. REFERENCES

1. Nicholson, J., Takahashi, K. and Nakatsu, R., "Emotion recognition in speech using neural networks", ***Neural Computing & Applications***, Vol. 9, No. 4, (2000), 290-296.
2. Schuller, B., Rigoll, G. and Lang, M., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", in Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP'04). IEEE International Conference on, Vol. 1, (2004), 577-580.
3. France, D. J., Shiavi, R. G., Silverman, S., Silverman, M. and Wilkes, M., "Acoustical properties of speech as indicators of depression and suicidal risk", ***Biomedical Engineering, IEEE Transactions on***, Vol. 47, No. 7, (2000), 829-837.
4. Abdi, J., Khalili, G. F., Fatourechi, M., Lucas, C. and Sedigh, A. K., "Control of multivariable systems based on emotional temporal difference learning controller", ***International Journal of Engineering-Transactions A: Basics***, Vol. 17, No. 4, (2004), 363.
5. Mirmomeni, M. and Yazdanpanah, M., "An unsupervised learning method for an attacker agent in robot soccer competitions based on the kohonen neural network", ***International Journal of Engineering) IJE Transactions A: Basics***, Vol. 21, No. 3, (2008), 255-268.
6. Hansen, J. H. and Cairns, D. A., "Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments☆", ***Speech Communication***, Vol. 16, No. 4, (1995), 391-422.
7. El Ayadi, M., Kamel, M. S. and Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases", ***Pattern Recognition***, Vol. 44, No. 3, (2011), 572-587.
8. Fernandez, R., *A* computational model for the automatic recognition of affect in speech., Massachusetts Institute of Technology (2004).
9. Russell, J. A. and Barrett, L. F., "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant", ***Journal of Personality and Social Psychology***, Vol. 76, No. 5, (1999), 805.
10. Hozjan, V. and Kacic, Z., "Context-independent multilingual emotion recognition from speech signals", ***International Journal of Speech Technology***, Vol. 6, No. 3, (2003), 311-320.
11. Ververidis, D. and Kotropoulos, C., "A review of emotional speech databases", in Proc. Panhellenic Conference on Informatics (PCI). (2003), 560-574.
12. Babcock, D. E. and Miller, M. A., "Client education: Theory & practice", ***Gastroenterology Nursing***, Vol. 18, No. 4, (1995), 157.
13. Buck, R., "Human motivation and emotion", John Wiley & Sons, (1988).
14. Ross, N., Medin, D. and Cox, D., "Epistemological models and culture conflict: Menominee and euro-american hunters in wisconsin", ***ETHOS***, Vol. 35, No. 4, (2007), 478-515.
15. Lewis, M. and Michalson, L., "Children's emotions and moods: Developmental theory and measurement", Plenum Press New York:, (1983).
16. Malatesta, C. Z. and Kalnok, M., "Emotional experience in younger and older adults", ***Journal of Gerontology***, Vol. 39, No. 3, (1984), 301-308.
17. de Albornoz, J. C., Plaza, L., Gervás, P. and Díaz, A., A joint model of feature mining and sentiment analysis for product review rating, in Advances in information retrieval. Springer. (2011) 55-66.
18. Cosmides, L. and Tooby, J., "Evolutionary psychology, moral heuristics, and the law", Dahlem University Press, (2006).
19. Morrison, D., Wang, R. and De Silva, L. C., "Ensemble methods for spoken emotion recognition in call-centres", ***Speech Communication***, Vol. 49, No. 2, (2007), 98-112.
20. Slaney, M. and McRoberts, G., "Baby ears: A recognition system for affective vocalizations", in Acoustics, Speech and Signal Processing, 1998. Proceedings of the IEEE International Conference on, Vol. 2, (1998), 985-988.
21. "Bavarian archive for speech signals, http://www.Bas.Uni-muenchen.De/bas/".
22. Liberman, M., Davis, K., Grossman, M., Martey, N. and Bell, J., "Emotional prosody speech and transcripts", ***Linguistic Data Consortium, Philadelphia***, (2002).
23. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. and Weiss, B., "A database of german emotional speech", in Interspeech. (2005), 1517-1520.
24. Engberg, I. S. and Hansen, A. V., "Documentation of the danish emotional speech database des", ***Internal AAU report, Center for Person Kommunikation, Denmark***, (1996).
25. Nwe, T. L., Foo, S. W. and De Silva, L. C., "Speech emotion recognition using hidden markov models", ***Speech Communication***, Vol. 41, No. 4, (2003), 603-623.
26. Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A. and Nogueiras, A., "Interface databases: Design and collection of a multilingual emotional speech database", in LREC, (2002).
27. Breazeal, C. and Aryananda, L., "Recognition of affective communicative intent in robot-directed speech", ***Autonomous Robots***, Vol. 12, No. 1, (2002), 83-104.

28. Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R. and Pellom, B., "Getting started with susas: A speech under simulated and actual stress database", in ***Eurospeech***. Vol. 97, (1997), 1743-46.
29. Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., and Rigoll, G., "Speaker independent speech emotion recognition by ensemble classification", in Multimedia and Expo,. ICME International Conference on, IEEE. (2005), 864-867.
30. Fu, L., Mao, X. and Chen, L., "Speaker independent emotion recognition based on svm/hmms fusion system", in Audio, Language and Image Processing, ICALIP International Conference on, IEEE. (2008), 61-65.
31. Schuller, B., "Towards intuitive speech interaction by the integration of emotional aspects", in Systems, Man and Cybernetics, International Conference on, IEEE. Vol. 6, (2002), 6 -11
32. Petrushin, V., "Emotion in speech: Recognition and application to call centers", in Proceedings of Artificial Neural Networks in Engineering. (1999), 7-10.
33. Makarova, V., "A database of russian emotional utterances", in ICSLP 2002. (2002).
34. Kim, E. H., Hyun, K. H., Kim, S. H. and Kwak, Y. K., "Speech emotion recognition using eigen-fft in clean and noisy environments", in Robot and Human interactive Communication, RO-MAN. The 16th International Symposium on, IEEE. (2007), 689-694.
35. Zhou, J., Wang, G., Yang, Y. and Chen, P., "Speech emotion recognition based on rough set and svm", in Cognitive Informatics, ICCI. 5th International Conference on, IEEE. Vol. 1,., (2006), 53-61.
36. Hu, H., Xu, M.-X. and Wu, W., "Gmm supervector based svm with spectral features for speech emotion recognition", in Acoustics, Speech and Signal Processing,. ICASSP. International Conference on, IEEE. Vol. 4, (2007), IV-413-IV-416.
37. Pereira, C., "Dimensions of emotional meaning in speech", in ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion., (2000).
38. Rabiner, L. and Juang, B.-H., "Fundamentals of speech recognition", (1993).
39. Wu, S., Falk, T. H. and Chan, W.-Y., "Automatic speech emotion recognition using modulation spectral features", ***Speech Communication***, Vol. 53, No. 5, (2011), 768-785.
40. Ntalampiras, S. and Fakotakis, N., "Modeling the temporal evolution of acoustic parameters for speech emotion recognition", ***Affective Computing, IEEE Transactions on***, Vol. 3, No. 1, (2012), 116-125.
41. Kim, E. H., Hyun, K. H., Kim, S. H. and Kwak, Y. K., "Improved emotion recognition with a novel speaker-independent feature", ***Mechatronics, IEEE/ASME Transactions on***, Vol. 14, No. 3, (2009), 317-325.
42. Lee, C.-C., Mower, E., Busso, C., Lee, S. and Narayanan, S., "Emotion recognition using a hierarchical binary decision tree approach", ***Speech Communication***, Vol. 53, No. 9, (2011), 1162-1171.
43. Pérez-Espinosa, H., Reyes-García, C. A. and Villaseñor-Pineda, L., "Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model", ***Biomedical Signal Processing and Control***, Vol. 7, No. 1, (2012), 79-87.
44. Bozkurt, E., Erzin, E., Erdem, Ç. E. and Erdem, A. T., "Formant position based weighted spectral features for emotion recognition", ***Speech Communication***, Vol. 53, No. 9, (2011), 1186-1197.
45. Harimi, A., Marvi, H. and Esmaileyan, Z., "Estimation of lpc coefficients using evolutionary algorithms", ***Journal of AI and Data MiningJournal of AI and Data Mining***, Vol. 1, No. 2, (2013), 111-118.
46. Steidl, S., Batliner, A., Noth, E. and Hornegger, J., Quantization of segmentation and f0 errors and their effect on emotion recognition, in 11th international conference on Text, Speech and Dialogue.: Heidelberg: Springer-Verlag. (2008) 525-534.
47. Krajewski, J. and Kröger, B. J., "Using prosodic and spectral characteristics for sleepiness detection", in ***Interspeech***., (2007), 1841-1844.
48. Rong, J., Li, G. and Chen, Y.-P. P., "Acoustic feature selection for automatic emotion recognition from speech", ***Information Processing & Management***, Vol. 45, No. 3, (2009), 315-328.
49. Schuller, B., Batliner, A., Steidl, S. and Seppi, D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge", ***Speech Communication***, Vol. 53, No. 9, (2011), 1062-1087.
50. Altun, H. and Polat, G., "Boosting selection of speech related features to improve performance of multi-class svms in emotion detection", ***Expert Systems with Applications***, Vol. 36, No. 4, (2009), 8197-8203.
51. Bishop, C. M. and Nasrabadi, N. M., "Pattern recognition and machine learning", springer New York, Vol. 1, (2006).
52. Ye, J., Janardan, R., Li, Q. and Park, H., "Feature extraction via generalized uncorrelated linear discriminant analysis", in Proceedings of the twenty-first international conference on Machine learning, ACM. (2004), 113.
53. Laukka, P., Neiberg, D., Forsell, M., Karlsson, I. and Elenius, K., "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation", ***Computer Speech & Language***, Vol. 25, No. 1, (2011), 84-104.
54. Raudys, S. J. and Jain, A. K., "Small sample size effects in statistical pattern recognition: Recommendations for practitioners", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 13, No. 3, (1991), 252-264.

A Database for Automatic Persian Speech Emotion Recognition: Collection, Processing and Evaluation

Z. Esmaileyan[a], H. Marvi[b]

[a] *Department of Electrical engineering, Science and Research branch, Islamic Azad University, Shahrood, Iran*
[b] *Department of Electrical engineering and Robotics, Shahrood University of technology, Shahrood, Iran*

P A P E R I N F O

Paper history:
Received 13 March 2013
Received in revised form 17 June 2013
Accepted 20 June 2013

Keywords:
Persian Emotional Speech Database
PDREC
Speech Emotion Recognition

چکیده

پیشرفت روزافزون در سیستم های اتوماتیک و رباتیک موجب شده است که محققان تلاش های زیادی در جهت افزایش کیفیت این ارتباط انجام دهند. از آنجا که گفتار متداول ترین روش ارتباط میان انسان هاست، تشخیص احساس انسان از روی گفتار به یکی از موضوعات چالش برانگیز در این حوزه تبدیل شده است. ما در این تحقیق یک پایگاه داده احساسی فارسی تدوین نموده ایم. جملات این پایگاه داده از نمایش های رادیویی موجود در وب سایت رسمی رادیو نمایش گرفته شده است. علاوه بر آن یک سیستم تشخیص احساس از روی گفتار فارسی طراحی نموده ایم. بدین منظور از ویژگی های عروضی و طیفی سیگنال گفتار استفاده گردیده است. نتایج حاصل از انجام آزمایشات بدست آمده از پایگاه داده ی پیشنهادی با پایگاه داده ی معروف برلین مقایسه شده است. سیستم مورد نظر برای گویندگان زن و مرد بصورت جداگانه طراحی شده است. در این سیستم ویژگی های غیر مرتبط و نویزی بوسیله ی الگوریتم انتخاب ویژگی فیشر حذف می شوند. ویژگی های انتخاب شده توسط الگوریتم فیشر، در یک مرحله ی دیگر توسط الگوریتم جداساز خطی کاهش می یابند. سپس داده ها با استفاده از کلاسه بند جداساز خطی کلاسه بندی می شوند. متوسط نرخ تشخیص بدست آمده برای زنان و مردان در پایگاه داده پیشنهادی ٥٥/٧٤٪ و ٤٧/٨٩٪ می باشد. همچنین متوسط نرخ تشخیص بدست آمده برای زنان و مردان در پایگاه داده برلین ٧٨/٦٤٪ و ٧٣/٤٠٪ می باشد.

doi: *10.5829/idosi.ije.2014.27.01a.11*.