Lang Resources & Evaluation (2019) 53:1–16
https://doi.org/10.1007/s10579-018-9427-x

ORIGINAL PAPER

ShEMO: a large-scale validated database for Persian speech emotion detection

Omid Mohamad Nezami[1] · **Paria Jamshid Lou**[2] · **Mansoureh Karami**[2]

Published online: 8 October 2018

Abstract This paper introduces a large-scale, validated database for Persian called *Sharif Emotional Speech Database* (*ShEMO*). The database includes 3000 semi-natural utterances, equivalent to 3 h and 25 min of speech data extracted from online radio plays. The ShEMO covers speech samples of 87 native-Persian speakers for five basic emotions including *anger*, *fear*, *happiness*, *sadness* and *surprise*, as well as neutral state. Twelve annotators label the underlying emotional state of utterances and majority voting is used to decide on the final labels. According to the kappa measure, the inter-annotator agreement is 64% which is interpreted as "substantial agreement". We also present benchmark results based on common classification methods in speech emotion detection task. According to the experiments, support vector machine achieves the best results for both gender-independent (58.2%) and gender-dependent models (female = 59.4%, male = 57.6%). The ShEMO will be available for academic purposes free of charge to provide a baseline for further research on Persian emotional speech.

Keywords Emotional speech · Speech database · Emotion detection · Benchmark · Persian

✉ Omid Mohamad Nezami
mnezami@iaubijar.ac.ir

[1] Bijar Branch, Islamic Azad University, Bijar, Iran

[2] Sharif University of Technology, Tehran, Iran

1 Introduction

Speech emotion detection systems aim at recognizing the underlying affective state of speakers from their speech signals. These systems have a wide range of applications from human–machine interactions to auto-supervision and control of safety systems (Huahu et al. 2010). For example, negative/positive experience of customers can be automatically detected in remote call centres to evaluate company services or attitude of staff towards customers (Batliner et al. 2003). These systems can also be used in health domains to monitor and detect the early signs of a depression episode (Dickerson et al. 2011) or help autistic children learn how to recognize more subtle social cues (Heni and Hamam 2016). Another application is crime detection where the psychological state of criminal suspects (i.e. whether or not they are lying) is discovered (Cowie et al. 2001). They are also useful for in-car board systems where information of the driver's emotion is extracted to increase their safety (Schuller et al. 2004). Furthermore, identifying the affective state of students in academic environments can help teachers or intelligent virtual agents to provide students with proper responses and improve teaching quality accordingly (Kort et al. 2001).

An important issue that should be considered before developing any speech emotion detection systems is the quality of database. In fact, the performance of these systems (like any statistical models) depends on the quality of training data (Busso et al. 2013). On the other hand, there is usually a lack of decent benchmark emotional speech database for non-English languages such as Persian. As some studies (Furnas et al. 1987; Feraru et al. 2015; Sagha et al. 2016) show, the relation between linguistic content and emotion is language dependent, so generalization from one language to another language is often difficult. That's why speech emotion detection systems are usually developed language-dependently.

A few studies have explored Persian speech emotion detection and introduced emotional databases (Esmaileyan and Marvi 2013; Savargiv and Bastanfard 2015; Keshtiari et al. 2015; Moosavian et al. 2007; Gharavian and Ahadi 2006; Mansoorizadeh 2009; Hamidi and Mansoorizade 2012). Persian Emotional Speech Database (Persian ESD) (Keshtiari et al. 2015) and Sahand Emotional Speech Database (SES) (Sedaaghi 2008) are two important datasets in Persian. Although its validity has been evaluated by a group of native speakers, Persian ESD covers emotional speech of only two speakers, which is not large enough for developing a robust system. SES covers the emotional speech of 10 speakers and is larger than Persian ESD but its reliability is relatively low according to the results of perception test.

In this paper, we present a large-scale validated dataset for Persian called *Sharif Emotional Speech Database* (*ShEMO*). The ShEMO is a semi-natural dataset which contains emotional (as well as neutral) speech samples of various Persian speakers. In addition to collecting the dataset, we benchmark the performance of standard classifiers on this dataset and compare the results to other languages to provide a baseline for further research. To the authors' best knowledge, this is the first systematic effort towards creating a large validated emotional speech dataset and

corresponding benchmark results for Persian. The ShEMO database will be publicly available to facilitate research on Persian emotional speech.[1]

The remainder of this paper is organized as follows. In Sect. 2, we review different types of emotional speech databases and explain the efforts made so far for designing/collecting a database for Persian. We introduce the ShEMO database and describe the process of data collection, annotation and validation in Sect. 3. We also discuss the baseline performance of standard classification methods on the ShEMO dataset and compare it to other databases in Persian, German and English. Finally in Sect. 4, we summarize our analysis and suggest future directions with our dataset.

2 Related work

Due to the vast amount of literature on emotional speech in general, this section will focus on reviewing different types of emotional speech database and the efforts made so far for data collection and validation for Persian language.

2.1 Types of emotional speech database

Emotional speech databases can be categorized in terms of *naturalness*, *emotion*, *speaker*, *language*, *distribution* and so forth (Ayadi et al. 2011). Naturalness is one of the most important factors to be considered when designing or collecting a database. Based on the degree of naturalness, databases can be divided into three types of *natural*, *semi-natural* and *simulated* (Ayadi et al. 2011). In natural databases, speech data is collected from real-life situations to guarantee that the underlying emotions of utterances are naturally conveyed. Such databases are rarely used for research purposes due to the legal and ethical issues accompanied with data collection. To avoid the difficulty, most natural databases are built by recording the emotional speech of some volunteer or recruited participants whose emotions have been naturally evoked by a method. For instance, a natural database may cover speech samples of non-professional actors discussing emotional events of their lives. Belfast Induced Natural Emotion Database (Douglas-Cowie et al. 2000) is an example where individuals' discussion about emotive subjects and interactions among the audience in television shows induce emotional speech. Computer games can also be used to naturally elicit emotional speech since players usually react positively or negatively towards winning or losing a game (Johnstone et al. 2005). Another technique is *Wizard-of-Oz* scenario (Batliner et al. 2000) where a human, so-called *Wizard*, simulates a dialogue system to interact with the users in such a way that they believe they are speaking to a machine. For instance, FAU Aibo Emotion Corpus (Steidl 2009) contains the spontaneous emotional speech of children talking to a dog-like robot. The Aibo robot is controlled by a human wizard to show obedient and disobedient behaviors so that the emotional reactions of children can be induced.

[1] Upon publishing this paper, we release our database for academic purposes.

Another type of emotional speech database is semi-natural which is built using either scenario-based approach or acting-based one. In the scenario-based approach (Scherer et al. 1991), the affective state of speakers is first evoked by a method. For instance, speakers recall some memories or read given sentences describing a scenario to get emotional. Then, they are asked to read a pre-written text in a particular emotion which aligns their provoked affective state. Persian Emotional Speech Database (Keshtiari et al. 2015) is an instance of this type. In the acting-based approach, emotional utterances are extracted from movies or radio plays. To illustrate, Chinese Emotional Speech Database (Yu et al. 2001) includes 721 utterances extracted from teleplays. Giannakopoulos et al. (2009) also uses English movies to collect 1500 affective speech samples.

Emotional speech databases can be simulated. For collecting this type of databases, scripted texts including isolated words or sentences are used. The prompt texts are usually semantically neutral and interpretable to any given emotion.[2] For recording these databases, professional stage actors are recruited to express the pre-determined sentences or words in peculiar emotions. The utterances are usually recorded in acoustic studios with high quality microphones in order not to influence the spectral amplitude or phase characteristics of the speech signal. Berlin Database of Emotional Speech (Burkhardt et al. 2005) and Danish Emotional Speech Database (Engberg et al. 1997) are two examples of simulated data. The main disadvantage of simulated databases is that emotions are usually exaggerated and far from natural. To alleviate this problem, non-professional actors (such as academic students or employees) are hired to read the prompt.

In addition to degree of naturalness, the theoretical framework of emotional speech databases can be different from each other. Two important theories of emotion include categorical and dimensional approach. According to the categorical approach, there are a small number of basic emotions which are recognized universally. Ekman (1982) showed that *anger*, *fear*, *surprise*, *happiness*, *disgust*, *sadness* are six basic emotions which can be recognized universally (Wöllmer et al. 2013; McKeown et al. 2010; Nicolaou et al. 2011). According to dimensional approach, affective states are not independent from one another, but they are systematically related so that they can be demonstrated as broad dimensions of experience. In this approach, emotions are represented as continuous numerical values on two main, inter-correlated dimensions of valence and arousal. The valence dimension shows how positive or negative the emotion is, ranging from unpleasant to pleasant feelings. The arousal dimension indicates how active or passive the emotion is, ranging from boredom to frantic excitement (Russell 1980; Alvarado 1997; Lewis et al. 2007).

Emotional speech databases can also be differentiated in terms of speakers. In most cases, professional actors are recruited to read pre-written sentences in target emotions (e.g. Berlin Database of Emotional Speech). However, some databases use semi-professional actors (e.g. Danish Emotional Speech Database) or ordinary people (e.g. Sahand Emotional Speech Database) to avoid exaggerated emotion

[2] The prompt excludes any emotional contents in order not to intervene the expression and perception of emotional states.

expression. Furthermore, the utterances of some datasets (e.g. Berlin Database of Emotional Speech) are uniformly distributed over emotions while the distribution of emotions in other datasets are unbalanced and may reveal their frequency in the real world (e.g. Chinese Emotional Speech Database). Another important factor is availability of databases. While the majority of emotional speech databases are private [e.g. MPEG-4 (Schuller et al. 2005)], there are some datasets which are available for public use [e.g. FERMUS III (Schuller and Munchen 2002), RAVDESS (Livingstone et al. 2012)].

2.2 Persian emotional speech databases

In recent years, some efforts have been made to record and collect validated datasets in Persian. In this section, we will elaborate two important ones including Sahand Emotional Speech (SES) Database (Sedaaghi 2008) and Persian Emotional Speech Database (Persian ESD) (Keshtiari et al. 2015).

Persian ESD (Keshtiari et al. 2015) is a semi-natural database which includes 470 utterances in five basic emotions including *anger*, *disgust*, *fear*, *happiness*, *sadness*, as well as neutral state. For collecting this database, 90 sentences were evaluated by a large group of native Persian-speakers to make sure that they were emotionally neutral. Two native speakers of Persian, a 50-year old man and a 49-year old woman, were asked to articulate the sentences in target emotions. The speakers were semi-professional actors who had participated in acting classes for a while. Prior to recording sessions, the speakers were asked to read a scenario and imagine experiencing the situation. Five scenarios (each corresponding to an emotional state) were used in this project. Thirty-four native speakers validated the database by recognizing the underlying emotion of each utterance in a 7-point nominal scale. According to the perceptual study, they achieved an accuracy of 71.4% which is five times chance performance.

SES (Sedaaghi 2008) is a simulated dataset which includes 1200 utterances or 50 min of speech data. To record SES, 10 university students (5 males and 5 females) were asked to read 10 single words, 12 sentences and 2 passages in four basic emotions of *surprise*, *happiness*, *sadness*, *anger* plus neutral mode. After recording the database, 24 annotators listened to the utterances only once and recognized the conveyed emotional state on a 5-point nominal scale. The annotators achieved 42.66% accuracy in classifying emotions which was twice what would be expected by chance.

Compared to the SES in terms of linguistic structure, the Persian ESD contains sentences with a single grammatical structure (subject + object + prepositional phrase + verb). The SES, however, covers various linguistic forms including word, sentence and passage. The questionnaire used in the SES perceptual study has a 5-point nominal scale which only includes the target emotions. Therefore, the participants were forced to choose an option from the given short list of emotions. James (1994) argues that not allowing listeners to label emotions freely results in forging agreement. As a solution, Frank and Stennett (2001) suggest adding *none of the above* to the response option. As a result, a part of the recognition accuracy reported in the SES can be artifact because of excluding *none of the above* option.

Moreover, neither the SES nor the Persian ESD has provided standard phonetic transcriptions, so it is difficult to extract the linguistic content from their utterances. Table 1 summarizes the Persian emotional speech databases in terms of accessibility, number of utterance, number of speakers, type of emotions, naturalness, pre-written text (scripted/unscripted), audio/visual mode and validation.

3 Sharif Emotional Speech Database

Sharif Emotional Speech Database (ShEMO) is a large-scale semi-natural database for Persian which contains 3 h and 25 min of speech data from 87 native-Persian speakers (31 females, 56 males). There are 3000 utterances in *.wav* format, 16 bit, 44.1 kHz and mono which cover five basic emotions of *anger*, *fear*, *happiness*, *sadness* and *surprise*, as well as neutral state. The utterances are extracted from radio plays which are broadcast online.[3] In the following subsections, we elaborate different phases of developing ShEMO, including pre-processing, annotation and measuring reliability.

3.1 Pre-processing, annotation and reliability

We selected 50 radio plays of various genres including comedy, romantic, crime, thrilled and drama as potential sources of emotional speech. We balanced out the differences of the audio streams using a free open-source audio editor software application, named *Audacity*. Since most streams (about 90% of them) had a sampling frequency of 44.1 kHz, we upsampled the streams which had a lower sampling rate using cubic interpolation technique. We also converted the stereo-recorded streams to mono. Mono channel is commonly used in speech communications where there is only one source of audio whereas stereo channel is usually applied when there are more than one source of audio. For example, using stereo channel in music production applications (where sound is generated from different instruments) leads to high-quality extraction and separation of multiple instruments from a single purely monophonic audio recording. Therefore, using stereo channel in speech applications only results in increasing bandwidth and storage space.

We segmented each stream into smaller parts such that each segment would cover the speech sample of only one speaker without any background noise or effect. We recruited 12 annotators (6 males, 6 females) to label the affective state of the utterances on a 7-point scale (including *anger*, *fear*, *neutrality*, *happiness*, *sadness*, *surprise*, and *none of the above*). The annotators were all native speakers of Persian with no hearing impairment or psychological problems. The mean age of the annotators was 24.25 years (SD = 5.25 years), ranging from 17 to 33 years. Tables 2 and 3 highlight the detailed information of anonymous annotators.

The utterances were randomly played in a quiet environment. The annotators were instructed to select *none of the above* where more than one emotion was conveyed from an utterance or the underlying emotion was not among the specified

[3] www.radionamayesh.ir.

Table 1 Persian emotional speech databases

Data	Access	Size	Speaker	Emotion
Moosavian et al. (2007)	Private	40	One with patois	Anger, happiness, sadness, neutral mode
Gharavian and Ahadi (2006)	Private	116 + 1800 neutral utterances selected from Farsdat (Bijankhan et al. 1994)	One male	Sadness, anger, neutral mode
Sedaaghi (2008)	Commercially available	1200	5 females, 5 males	Sadness, happiness, surprise, anger, neutral mode
Gharavian and Ahadi (2006)	Private	252	22	Happiness, anger, interrogative, neutral mode
Mansoorizadeh (2009)	Private	26	Not reported	Anger, fear, disgust, sadness, happiness, surprise
Hamidi and Mansoorizade (2012)	Private	2400	330 actors and actresses	Happiness, fear, sadness, anger,disgust, neutral mode
Esmaileyan and Marvi (2013)	Private	748	33 professional actors (18 males, 15 females)	Anger, fear,sadness, happiness, boredom, disgust, surprise, neutral mode
Keshtiari et al. (2015)	Public and free	470	One actor, one actress	Anger, disgust, fear,happiness, sadness, neutral mode
Savargiv and Bastanfard (2015)	Private	6720 (3–7 s each)	10 males, 10 females	Anger, fear, sadness,happiness, boredom, disgust, surprise, neutral mode
ShEMO[a]	Public and free	3000	31 females, 56 males	Anger, fear, happiness, sadness, surprise, neutral mode

Table 1 continued

Data	Naturality	Scripted	Mode	Validation
Moosavian et al. (2007)	Simulated (recorded in a non-acoustic environment and includes background noise)	Yes	Audio	Not reported
Gharavian and Ahadi (2006)	Simulated	Yes	Audio	Conducted but not reported
Sedaaghi (2008)	Simulated	Yes	Audio	24 students evaluated the utterances
Gharavian and Ahadi (2006)	Simulated	Yes	Audio	Not reported
Mansoorizadeh (2009)	Semi-natural (speakers were given pre-written scenarios and asked to imagine the situation)	Yes	Audio and visual	Not reported
Hamidi and Mansoorizade (2012)	Semi-natural (utterances were extracted from more than 60 Persian movies)	No	Audio	Not reported
Esmaileyan and Marvi (2013)	Semi-natural (collected from radio plays)	No	Audio	Not reported
Keshtiari et al. (2015)	Semi-natural (actors were given scripted scenarios for each emotion)	Yes	Audio	3 different evaluations were performed
Savargiv and Bastanfard (2015)	Simulated	Yes	Audio	Not reported
ShEMO[a]	Semi-natural	Yes	Audio	Yes

[a]ShEMO is added for the sake of comparison

Table 2 Detailed information of annotators

Code	Gender	Age	Education
01	Male	23	Undergraduate student
02	Female	18	Associate student
03	Female	20	Associate student
04	Male	22	Undergraduate student
05	Male	31	PhD candidate
06	Male	33	Master degree
07	Female	31	Master degree
08	Male	21	Undergraduate student
09	Female	23	Undergraduate student
10	Female	17	High school student
11	Male	25	Master degree
12	Female	27	Master degree

Table 3 Statistical values of annotators' age in year

Gender	Mean	SD
Female	22.66	4.99
Male	25.83	5.46
Total	24.25	5.25

Table 4 Number and duration of utterances per each gender and affective state (SD = standard deviation)

Affective state	Number			Duration (s)			
	Female	Male	Total	Min.	Max.	Mean	SD
Anger	455	604	1059	0.44	22.42	3.61	2.63
Fear	22	16	38	0.76	8.97	3.17	1.84
Happiness	111	90	201	0.82	13.39	3.81	2.36
Neutral	284	744	1028	0.56	33.32	4.89	4.1
Sadness	271	178	449	0.69	27.89	4.84	3.7
Surprise	120	105	225	0.35	10.95	1.79	1.45
Total	1263	1737	3000	0.35	33.32	4.11	3.41

emotional states. Since the utterances were extracted from radio plays, there was no guarantee that the lexical contents would be emotionally neutral. Therefore, there might be some cases where the affective state of the speaker implied from their speech would be in a stark contrast with the lexical content of the utterance. To resolve this ambiguity and avoid any confusion, the annotators were intentionally asked to label the emotional state of utterances only based on the ways it had been portrayed in speech, regardless to the lexical contents. To make a final decision on the labels of the utterances, we considered majority voting (Mower et al. 2009a, b;

شما چرا وقتی آقای سریا ناپدید شد با کازن تماس نگرفتین و این جریانو بهش نگفتین

ʃomɑ ʧerɑ væqti ʔɑqɑye seriyɑ nɑpædid ʃod bɑ kɑzen tæmɑs nægereftin væ ʔin dʒæryɑno beheʃ nægoftin

Why didn't you call Kazen and let him know the issue when Mr. Seriya got disappeared

Fig. 1 Orthographic, phonetic and English translation of an utterance conveying anger

Audhkhasi and Narayanan 2010). The utterances for which the majority voting decided *none of the above* were discarded from the database as they probably reflected multiple emotions or an emotion which our database did not cover.

We calculated Cohen's kappa statistics (Cohen 1960) as a measurement of inter-rater reliability.[4] According to the kappa statistics, there was 64% agreement on the labels which means there is "substantial agreement" (Landis and Koch 1977) among the annotators.[5] We discarded the utterances for which a low reliability was reported.

The mean length of utterances is 4.11 s (SD = 3.41), ranging from 0.35 to 33 s. The detailed information of utterances are illustrated in Table 4.

As shown in Table 4, *anger* and *fear* have the highest and lowest number of utterances in the database. For female and male speakers, the maximum number of utterances belongs to *neutral* mode and *anger*, respectively. Mean length of the utterances conveying *surprise* (mean = 1.79, SD = 1.45) is remarkably shorter than other emotions (total mean = 4.11, total SD = 3.41). On the other hand, the highest duration has been reported for *sadness* (mean = 4.84, SD = 3.7). It can be due to the fact that people usually use frequent silences and stops within their speech when conveying *sadness*.

The ShEMO database is also orthographically and phonetically transcribed according to the International Phonetic Alphabet (IPA),[6] which can be useful for extracting linguistic features. A sample of orthographic and phonetic transcription, along with its English translation for *anger* is illustrated in Fig. 1.

3.2 Benchmark results

We provide baseline results of common classification methods on the ShEMO database. As features, we use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al. 2016). The eGeMAPS[7] refers to a basic standard acoustic parameter set including spectral (balance/shape/dynamics), frequency- and

[4] Cohen's kappa ranges generally from 0 to 1, where large numbers indicate higher reliability and values near zero suggest that agreement is attributable to chance alone.

[5] As Landis and Koch (1977) explain, $0.61 < kappa < 0.80$ is interpreted as "substantial agreement" among the judges.

[6] The IPA was devised by the International Phonetic Association as a standardized representation of the sounds of oral language.

[7] It contains 88 different parameters. For further information, please refer to Eyben et al. (2016).

energy/amplitude-related parameters selected based on their potentiality for indexing affective physiological modifications in voice production, their automatic extractability, their proven value in previous studies and their theoretical importance (Eyben et al. 2016). We use the Munich Versatile and Fast Open-Source Audio Feature Extractor called openSMILE (Eyben et al. 2010) to extract the eGeMAPS. To eliminate speaker variabilities, the features are normalized using z-score.

We use three classifiers, namely support vector machine (SVM), *k*-nearest neighbour (*k*-NN), and decision tree (DT) for classification. According to Eyben et al. (2016), SVM is the most widely used static classifier in the field of speech emotion detection. Decision tree and *k*-NN have also been applied for detecting the underlying emotional state of speech (Grimm et al. 2007; Lee et al. 2011). We use SVM with Radial Basis Function (RBF) kernel and for decision tree model, we use random forest algorithm (Breiman 2001) which are typical approaches in classification tasks. We apply nested cross-validation using one-vs-one multi-class strategy. The nested cross validation effectively uses a series of train/validation/test set splits to optimize each classifier's parameters and unbiasedly measure the generalization capabilities of classifier (Cawley and Talbot 2010). In our work, first an inner 10-fold cross validation uses Bayesian optimization (Snoek et al. 2012) to tune the parameters of each classifier and select the best model. Then, an outer 5-fold cross validation is used to evaluate the model selected by the inner cross validation. Finally, we report the Unweighted Average Recall (UAR), which is popular in this field (Schuller et al. 2009, 2010, 2011, 2016; Metze et al. 2011; Deng et al. 2012), averaged over the evaluation results for each classifier. Table 5 demonstrates the performance of each classifier. The range of parameters used in Bayesian optimization is given below:

- *k*-NN: Number of neighbours is set from 1 to 30,
- *k*-NN: Distance metrics include euclidean, cosine, chebychev and cubic.
- SVM: Sigma (the kernel scale) and box (a cost to the misclassification) values are chosen between 0.00001 and 100,000.
- DT: Minimum observations number per leaf node is selected from 1 to 20.
- DT: Number of predictors (features) at each node is chosen from 1 to the number of feature variables.

As presented in Table 5, SVM outperforms *k*-NN and decision tree in both gender-dependent and -independent experiments. Decision tree has a better performance in comparison with *k*-NN in gender-dependent experiment; however, it is slightly worse in gender-independent case. The best result is achieved for the SVM model trained on the female subsection of the data. Although the number of utterances is lower for female speakers (1263 vs. 1737 for male), they have a higher inter-rater reliability (kappa = 0.67 vs. 0.61 for male). In other words, the annotators had a stronger level of consensus on the underlying affective state of the female utterances. This may be the reason why SVM and *k*-NN have better performance on the female subsection of the data. The confusion matrix of the performance in gender-independent mode (i.e. SVM = 58.2) is shown in Table 6. It

Table 5 Mean UAR obtained for SVM, *k*-NN and decision tree using female, male and all utterances of the ShEMO

	SVM	*k*-NN	DT
Female	59.4	47.4	49.0
Male	57.6	45.6	46.6
All	58.2	47.6	47.4

Table 6 Confusion matrix of the best performance in gender-independent mode

	Anger	Happiness	Neutrality	Sadness	Surprise
Anger	**911**	18	85	23	22
Happiness	**68**	37	59	25	12
Neutrality	69	12	**902**	33	12
Sadness	37	13	107	**263**	29
Surprise	31	13	56	35	**90**

The maximum value of each class is bold

should be mentioned that we excluded the fear utterances in our classification experiment because there was a small number of them in the database (38 in total).

According to the confusion matrix, the model has the best performance in detecting *anger* and *neutrality*. The reason is that both *anger* and *neutral* mode have the highest number of utterances in the database, so the model properly learns the parameters associated with these two emotional states. On the other hand, the worst classification performance is reported for *happiness* which has the lowest number of utterances in our data.[8] According to Table 6, *happiness* is mostly confused with *anger*. *Anger* and *happiness* are categorized into high-arousal emotions; this can be the reason why the model has a poor performance in discriminating these two. As Scherer (1986) argues, emotions which are in the same category in terms of valence and arousal are usually confused with each other. Moreover, *anger*, *sadness* and *surprise* are confused with *neutrality*. It seems that this happens for the utterances with lower emotional strength. Moreover, the utterances conveying *surprise* are relatively short; therefore, it can be challenging for the model to differentiate *surprise* from other emotional states based on a short context.

In order to compare our baseline model and see how it works on other databases and languages, we train and test the mentioned classifiers on Persian ESD (Keshtiari et al. 2015), Berlin Emotional Speech database (EMO-DB) (Burkhardt et al. 2005) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone et al. 2012). The EMO-DB covers 535 speech samples of 10 speakers (5 males, 5 females) in 6 emotional states of *anger*, *boredom*, *disgust*, *anxiety/fear*, *happiness* and *sadness*, as well as neutral mode. The EMO-DB is a balanced, simulated database with pre-written texts.[9] The RAVDESS database

[8] *Happiness* has the lowest number of utterances after *fear*. As mentioned before, fear utterances were ignored in the classification experiments.

[9] Actors were asked to read 10 short emotionally neutral sentences.

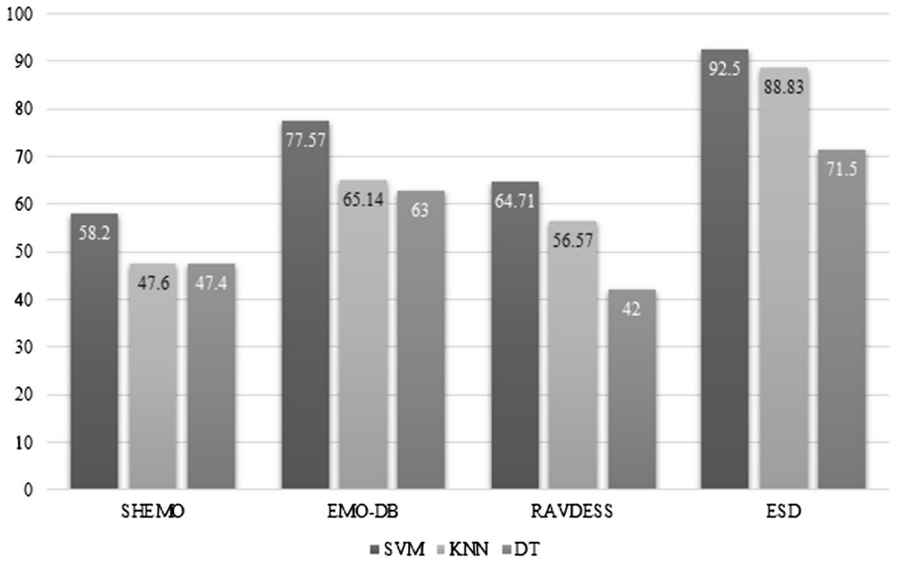

Fig. 2 Comparison of SVM, *k*-NN and decision tree on the ShEMO (Persian), EMO-DB (German), RAVDESS (English) and Persian ESD datasets. The vertical axis indicates UAR

includes 1440 utterances[10] articulated by 24 (12 males, 12 females) speakers in a north American English accent. It includes 7 emotional states: *happiness*, *sadness*, *anger*, *calmness*, *fear*, *surprise* and *disgust*, as well as neutral mode. The RAVDESS is a balanced, simulated database and uses professional actors to record the utterances. Figure 2 illustrates the results of comparison.

As shown in Fig. 2, the classifiers trained on the Persian ESD result in the highest UAR. On the contrary, the decision tree trained on the RAVDESS dataset and SVM and *k*-NN trained on the ShEMO have the lowest performance. All databases, except for the ShEMO, are balanced and have a fixed prompt for all speakers and emotions. On the other hand, the ShEMO is unbalanced and more realistic data, so detecting the underlying affective state of the utterances is a harder task in this case. As the results indicate the ShEMO would provide the research community with challenges in developing proper classification techniques for emotion detection in more realistic environments.

4 Summary, conclusion and future work

This paper introduces a large-scale validated dataset for Persian which contains semi-natural emotional, as well as neutral speech of a wide variety of native-Persian speakers. In addition to the database, we present the benchmark results of common classification methods to be shared among the researchers of this field.

[10] We trained the models on the audio (not video), speech (not song) files of the database.

Our immediate future work includes increasing the number of utterances for fear. We also intend to extend the benchmark results to include other classification methods such as hidden Markov models and deep neural networks as the state-of-the-art technique in speech emotion detection. Labelling the data in terms of arousal and valence is another potential future extension. Moreover, it would be interesting to study the frequency of neutral and emotional speech among native-Persian speakers and see whether the distribution of utterances in the ShEMO conforms to the standard distribution of emotions in Persian. In future, we can also annotate the emotional strength of utterances.

Acknowledgements We would like to thank the anonymous reviewers for their insightful comments and suggestions. We also gratefully thank Dr. Steve Cassidy for his helpful points.

References

Alvarado, N. (1997). Arousal and valence in the direct scaling of emotional response to film clips. *Motivation and Emotion*, *21*, 323–348.

Audhkhasi, K., & Narayanan, S. (2010). Data-dependent evaluator modeling and its application to emotional valence classification from speech. In *Proceedings of INTERSPEECH* (pp. 2366–2369), Makuhari, Japan.

Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Noth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. In *Proceedings of ISCA workshop on speech and emotion* (pp. 195–200).

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Noth, E. (2003). How to find trouble in communication. *Speech Communication*, *40*(1–2), 117–143.

Bijankhan, M., Sheikhzadegan, J., Roohani, M., & Samareh, Y. (1994). FARSDAT—The speech database of Farsi spoken language. In *Proceedings of Australian conference on speech science and technology* (pp. 826–831), Perth, Australia.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Proceedings of INTERSPEECH* (pp. 1517–1520), Lissabon, Portugal. ISCA.

Busso, C., Bulut, M., & Narayanan, S. (2013). Toward effective automatic recognition systems of emotion in speech. In J. Gratch & S. Marsella (Eds.), *Social emotions in nature and artifact: Emotions in human and human–computer interaction* (pp. 110–127). New York, NY: Oxford University Press.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Deng, J., Han, W., & Schuller, B. (2012). Confidence measures for speech emotion recognition: A start. In *Proceedings of speech communication* (pp. 1–4), Braunschweig, Germany.

Dickerson, R., Gorlin, E., & Stankovic, J. (2011). Empath: A continuous remote emotional health monitoring system for depressive illness. In *Proceedings of the 2nd conference on wireless health* (pp. 1–10), New York, NY, USA.

Douglas-Cowie, E., Cowie, R., & Schroeder, M. (2000). A new emotion database: Considerations, sources and scope. In *Proceedings of ISCA workshop on speech and emotion* (pp. 39–44).

Ekman, P. (1982). *Emotion in the human face*. Cambridge: Cambridge University Press.

Engberg, I., Hansen, A., Andersen, O., & Dalsgaard, P. (1997). Design, recording and verification of a Danish emotional speech database. In *Proceedings of EUROSPEECH* (Vol. 4, pp. 1695–1698).

Esmaileyan, Z., & Marvi, H. (2013). A database for automatic Persian speech emotion recognition: Collection, processing and evaluation. *International Journal of Engineering*, *27*, 79–90.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202.

Eyben, F., Wollmer, M., & Schuller, B. (2010). openSMILE—The Munich versatile and fast open-source audio feature extractor. In *Proceedings of ACM multimedia* (pp. 1459–1462), Florance, Italy.

Feraru, S. M., Schuller, D., & Schuller, B. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *Proceedings of the 6th international conference on affective computing and intelligent interaction (ACII)* (pp. 125–131), Xi'an, China.

Frank, M., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Personality and Social Psychology*, *80*(1), 75–85.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human–system communication. *Communications of the ACM*, *30*(11), 964–971.

Gharavian, D., & Ahadi, S. (2006). Recognition of emotional speech and speech emotion in Farsi. In *Proceedings of international symposiumon chinese spoken language processing* (Vol. 2, pp. 299–308).

Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2009). A dimensional approach to emotion recognition of speech from movies. In *Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 65–68).

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, *49*(10–11), 787–800.

Hamidi, M., & Mansoorizade, M. (2012). Emotion recognition from Persian speech with neural network. *Artificial Intelligence and Applications*, *3*(5), 107–112.

Heni, N., & Hamam, H. (2016). Design of emotional education system mobile games for autistic children. In *Proceedings of the 2nd international conference on advanced technologies for signal and image processing (ATSIP)*.

Huahu, X., Jue, G., & Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. In *Proceedings of international conference on artificial intelligence and computational intelligence* (Vol. 1, pp. 537–541).

James, A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102–141.

Johnstone, T., Van Reekum, C., Hird, K., Kirsner, K., & Scherer, K. (2005). Affective speech elicited with a computer game. *Emotion*, *5*(4), 513–518.

Keshtiari, N., Kuhlmann, M., Eslami, M., & Klann-Delius, G. (2015). Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD). *Behavior Research Methods*, *47*(1), 275–294.

Kort, B., Reilly, R., & Picard, R. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings of the IEEE international conference on advanced learning technologies (ICALT)* (pp. 43–46), Washington, DC, USA.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*(9–10), 1162–1171.

Lewis, P. A., Critchley, H. D., Rotshtein, P., & Dolan, J. R. (2007). Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, *17*(3), 742–748.

Livingstone, S., Peck, K., & Russo, F. (2012). RAVDESS: The Ryerson audio-visual database of emotional speech and song. In *Proceedings of the 22nd annual meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS)*, ON, Canada.

Mansoorizadeh, M. (2009). *Human emotion recognition using facial expression and speech features fusion*. PhD thesis, Tarbiat Modares University, Tehran, Iran **(in Persian)**.

McKeown, G., Valstar, M., Cowie, R., & Pantic, M. (2010). The semaine corpus of emotionally coloured character interactions. In *Proceedings of IEEE international conference on multimedia and expo*

(ICME'10) (pp. 1079–1084), Singapore, Singapore. IEEE Computer Society. https://doi.org/10.1109/ICME.2010.5583006.

Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., & Steidl, S. (2011). Emotion recognition using imperfect speech recognition. In *Proceedings of INTERSPEECH* (pp. 478–481), Makuhari, Japan.

Moosavian, A., Norasteh, R., & Rahati, S. (2007). Speech emotion recognition using adaptive neuro-fuzzy inference systems. In *Proceedings of the 8th conference on intelligent systems***(in Persian)**.

Mower, E., Mataric, M., & Narayanan, S. (2009b). Evaluating evaluators: A case study in understanding the benefits and pitfalls of multi-evaluator modeling. In *Proceedings of INTERSPEECH* (pp. 1583–1586), Brighton, UK.

Mower, E., Metallinou, A., Lee, C., Kazemzadeh, A., Busso, C., Lee, S., & Narayanan, S. (2009a). Interpreting ambiguous emotional expressions. In *Proceedings of the 3rd international conference on affective computing and intelligent interaction and workshops (ACII)* (pp. 662–669), Amsterdam, The Netherlands.

Nicolaou, M., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valance–arousal space. *IEEE Transactions on Affective Computing*, *2*(2), 92–105. eemcs-eprint-21287.

Russell, J. A. (1980). A circumplex model of affect. *Personality and Social Psychology*, *39*(6), 1161–1178.

Sagha, H., Matejka, P., Gavryukova, M., Povolny, F., Marchi, E., & Schuller, B. (2016). Enhancing multilingual recognition of emotion in speech by language identification. In *Proceedings of INTERSPEECH* (pp. 2949–2953).

Savargiv, M., & Bastanfard, A. (2015). Persian speech emotion recognition. In *Proceedings of the 7th international conference on information and knowledge technology (IKT)* (pp. 1–5).

Scherer, K. (1986). Vocal affect expression: A review and a model for future research. *Psychol Bull*, *99*(2), 143–165.

Scherer, K., Banse, R., Wallbott, H., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*(2), 123–148.

Schuller, B., Batliner, A., Steidl, S., Schiel, F., & Krajewski, J. (2011). The INTERSPEECH 2011 speaker state challenge. In *Proceedings of INTERSPEECH* (pp. 3201–3204), Florence, Italy. ISCA.

Schuller, B., & Munchen, T. U. (2002). Towards intuitive speech interaction by the integration of emotional aspects. In *Proceedings of IEEE international conference on systems, man and cybernetics (SMC)* (Vol. 1, pp. 6–11).

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. In *Proceedings of IEEE international conference on multimedia and expo (ICME)* (pp. 864–867).

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (Vol. 1, pp. 577–580).

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proceedings of INTERSPEECH* (pp. 312–315), Brighton, UK. ISCA.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Muller, C., et al. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Proceedings of INTERSPEECH* (pp. 2794–2797), Makuhari, Japan. ISCA.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J., Baird, A., et al. (2016). The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Proceedings of INTERSPEECH* (pp. 2001–2005), San Francisco, USA. ISCA.

Sedaaghi, M. (2008). *Documentation of the Sahand Emotional Speech Database (SES)*. Technical report, Department of engineering, Sahand University of Technology.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951–2959).

Steidl, S. (2009). *Automatic classification of emotion related user states in spontaneous children's speech*. Ph.D. thesis, University of Erlangen-Nuremberg Erlangena, Bavaria, Germany.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, *31*(2), 153–163.

Yu, F., Chang, E., Xu, Y., & Shum, H. (2001). Emotion detection from speech to enrich multimedia content. In *Proceedings of the 2nd IEEE Pacific Rim conference on multimedia: Advances in multimedia information processing* (pp. 550–557), London, UK. Springer.