Multimedia Tools and Applications (2023) 82:36413–36430
https://doi.org/10.1007/s11042-023-15132-3

Speech emotion recognition in Persian based on stacked autoencoder by comparing local and global features

Azam Bastanfard[1] · Alireza Abbasian[2]

Received: 11 June 2022 / Revised: 9 August 2022 / Accepted: 13 March 2023 /
Published online: 30 March 2023

Abstract
Among the barriers to establishing effective human-machine interactions is the machines' inability to properly distinguish emotions from the human voice. The Speech Emotion Recognition (SER) systems have emerged to tackle this limitation. The accuracy of these systems depends on different factors such as the quantity and the types of emotions included in the database, feature extraction process including local and global features, feature selection method, and the type of classifier. This study presents a methodology for speech emotion recognition using an autoencoder neural network. It is shown that using a digit-level stacked autoencoder can be suitable for digit classification. The speech emotion recognition is done using the Persian emotional speech database (Persian ESD), which includes six emotional states: Happiness, Sadness, Fear, Disgust, Anger, and Neutral. Moreover, the popular, widely-used Berlin Emotional database (EMO-DB) is used to evaluate the effectiveness of the proposed approach. The experimental results show that the proposed method has significantly improved recognition accuracy.

Keywords Speech emotion recognition · Stacked autoencoder · Persian language · Deep learning

1 Introduction

Speech is the crucial yet easiest way of communication between humans. Contents transmitted through speech encompass a wide range of information that cannot be comprehensively written. For instance, the vocabulary content of the speech always comprises extra attributes

✉ Azam Bastanfard
bastanfard@kiau.ac.ir

Alireza Abbasian
a7abbasian@gmail.com

[1] Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran

[2] Faculty of Media Engineering, Islamic Republic of Iran Broadcasting University, Tehran, Iran

such as tone, accent, emotional states, etc., which contain a large amount of information to be transferred from one person to another. Since technological advancements and the growth of human-machine interactions are undeniable, the requirement of comfortable and inerrant communication between humans and machines is essential. Speech can be the best solution to address this requirement. However, it is necessary or the machine to be able to recognize emotions in speech. To this end, Speech Emotion Recognition (SER) systems have formed a new branch in speech processing studies [1].

Numerous objects can be considered for recognizing emotions from speech, including speech reconstruction, smartphones, dubbing, animation, computer games, text-to-speech, speech therapy applications [5], online call centers, speech classification, and many other applications.

Language and cultural differences are considered a challenge in these systems. In this study, the Persian language is chosen to analyze the proposed method, an Indo-European language spoken by more than 110 million people worldwide and is one of the most rhythmic languages [19, 30]. Although, there are limited researches on the Persian language in speech emotion recognition.

A speech emotion recognition system is generally defined as a set of methods that process and classify speech signals to identify their emotions [11]. There are three critical issues in recognizing emotion from the speech [15]. The first aspect is the existence of appropriate and comprehensive emotional datasets that can be well used to train and test the designed system. A major problem in designing an emotional database is labeling. Even humans cannot agree on the exact emotion in the spoken utterance [31]. This problem has alighted by using several judges to label each one of the utterances in the database. The second important aspect in speech emotion recognition systems is the impact of the extracted speech features on the classification efficiency of different emotions. Features are vital to a classification process. They reduce the original data to its most valuable characteristics. The third important aspect is choosing a classification system for recognizing speech emotions. Many articles and studies have addressed these three critical aspects of speech emotion recognition design. The method used in this research is developed based on deep learning regarding its application and importance [3].

1.1 Motivation

Communication between man and machine should be accessible, two-way, and practical. In order to achieve this goal, the machine needs to have the intelligence to recognize human speech and its emotions. Recognizing the speaker's emotion helps to understand the meaning better and increases the efficiency of speech processing systems. The recognition of speech emotions is done with almost suitable accuracies at present [55]. However, even though the Persian language is one of the most widely used languages in the world, fewer studies have been conducted on the recognition of emotions in the speech of this language [56]. In order to be on the same level as the world's living languages in this category, more studies must also take place in the Persian language.

Another motivation in designing a speech emotion recognition system is extracting suitable features that are effective for identifying different emotions. The purpose of extracting speech signal features is to obtain helpful information from the speech signal suitable for automatic speech processing [51]. In this case, the speech waveform becomes a parametric representation where the data rate is much lower than the original signal. So far, many features have been

used in the studies of emotion recognition from speech and have shown their effectiveness in this category. Using a large number of features to design a pattern recognition system causes a drop in the performance of the classification algorithm due to the increase in adjustment parameters, computational complexity, and inefficient training due to the high dimensions of the problem. Nevertheless, this problem can be significantly solved by using the autoencoder neural network, which can encode features into a set of features with smaller dimensions.

The features extracted from the speech are generally divided into global and local groups. Local features are obtained separately from each frame, while global features are obtained using statistical functions on local features in a speech utterance. It is expected that emotions have more opportunity to appear during an utterance than at the frame level. For this reason, global features are likely to perform better than local ones [21]. Demonstrating the correctness of this theory can be another motivation for this study.

1.2 Contribution

The contributions of this work are summarized below.

- Comparison of the efficiency of local and global features in speech emotion recognition systems.
- Achieving high accuracy in speech emotion recognition using Persian Emotional Speech Database (Persian ESD).
- Determining the accuracy of the proposed method using the Berlin emotional database (EMO-DB) for evaluating and comparing it with other studies.
- Deployment of the Autoencoder Neural Network to exploit the benefits of large feature sets by reducing their dimensions in two steps

1.3 Outline

The remainder of this paper is organized as follows: Section 2 discusses related works. Section 3 briefly introduces the chosen emotional speech databases and acoustic features used in this research. Also, the autoencoder-based method is presented in the same section. The experiments and results of the proposed system are demonstrated in Section 4. Finally, the conclusion is drawn in Section 6.

2 Related works

In previous works, various methods of production and use of emotional databases, feature extraction, and classification were used to enhance the recognition accuracy in speech emotion recognition systems. In general, there are helpful surveys to introduce the methods used in speech emotion recognition. Ayadi et al. [15] have prepared a comprehensive survey of studies conducted up to 2011 and examined the components used in common SER methods. More recent works are discussed in [46], where available literature on various databases, different features, and classifiers have been used to contemplate SER from diverse languages. The latest review of recent SER advancements has been published by Akçay et al. [1], an overview of emotional models, databases, pre-processing, features, supporting modalities, and classification.

There are very few studies on SER in Persian. Emotion recognition in Persian also requires the presence of a suitable emotional database. Among the databases used in the Persian language can be mentioned Persian Emotional Speech Database (Persian ESD) [20], which is also used in [4, 45] in addition to this work. Savargiv and Bastanfard introduced Persian Audio Visual Corpus in [39, 40], which has also shown 76% accuracy using Hidden Markov Model (HMM). Moreover, Sharif Emotional Speech Database (ShEMO) is one of the large-scale and newest Persian databases comparable with the databases used in this research work [34]. Gharavian and Ahadi used a non-emotional database based on Farsi Speech Database (FARSDAT) that categorized only two emotional states [16]. Furthermore, Harimi and Esmaileyan introduced Persian Emotional Speech Database (PDREC) [17], and Sadeghi also provided the Sahand Emotional Speech database (SES), which covers ten different speakers [43]. The lack of studies on these databases makes comparing and evaluating their performance difficult.

Features are the information extracted from speech. Many features have been used for SER systems, but there is no agreement on which features are more efficient than others. Two types of features are extracted from the speech: local and global. Local features are obtained separately from each frame, while global features are obtained using statistical functions on local features in a speech utterance. Because the expression of emotion during speech is more comprehensive than one frame, global features give better information for speech emotion recognition [36, 38]. Features such as Energy, Pitch, Formants, and Mel Frequency Cepstral Coefficients (MFCCs) have been widely used in previous studies in speech emotion recognition. Furthermore, Langari et al. have achieved a very high recognition rate using modified feature extraction [22].

A wide range of classifiers is used for the classification task in SER. Several classifiers such as Hidden Markov Models (HMM) in [39, 42], Gaussian Mixture Model (GMM) in [23, 27], Support Vector Machine SVM in [18], and k Nearest Neighbours (kNN) in [23] have been utilized in order to recognize the speech emotion. Deep learning and deep neural networks have recently been ubiquitous in pattern recognition. SER classifications also have used some of these methods. For instance, Multilayer Perceptron (MLP) in [41], Recurrent Neural Network (RNN) in [25, 33, 53], and Convolutional Neural Network (CNN) in [57] have been used.

Furthermore, Autoencoder-based neural networks could be a suitable method in this field. In 2013 Cibau et al. used deep Autoencoder on the Berlin emotional database (EMO-DB), and the classifier's performance was over 70% [9]. Deng et al. [12] used a single sparse autoencoder to find a standard structure in small target data and then apply such structure to reconstruct source data to complete proper knowledge transfer from source data into a target task. In 2017, they proposed a Semi-supervised Autoencoder (SS-AE) for SER [13]. The efforts in [14, 24] can also be mentioned as more recent research works using Autoencoder in this field. It can be said there has been little exploration of SER using an autoencoder architecture.

So, it seems research work on the Persian language using a comprehensive feature set and deep learning methods is missing in the literature. Thus, the main contribution of this research is to propose an autoencoder-based neural network to train and test the system and classification in Persian speech emotion recognition.

3 Proposed methodology

Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her speech. It is based on an in-depth analysis of the generation mechanism

of the speech signal, extracting some features that contain emotional information from the speaker's voice and employing suitable pattern recognition methods to identify emotional states. Like typical pattern recognition systems, this work's speech emotion recognition system contains three main modules: speech input from an emotional database, pre-processing and feature extraction, and finally, autoencoder-based classification (Fig. 1).

3.1 Emotional speech database

Undoubtedly, one of the main requirements of emotion recognition systems is the existence of appropriate and comprehensive databases that can be used well to train and test the system. A low-quality database may affect the results adversely. Speech emotion databases differ in terms of the number of emotions, the number of speakers, the language and gender of speakers, spontaneous or acted emotion expression, and audio recording quality. The lack of a comprehensive database in speech-based processing could be a significant problem that is the case in Persian. However, one of the few databases available for emotional speech in the Persian language is the Persian emotional speech database (Persian ESD) [20], which is used in this research work. The corpus is created using two Persian actors (one man and one woman). They have spoken in six emotional states Anger (62 utterances), Happiness (58 utterances), Sadness (56 utterances), Fear (58 utterances), Disgust (58 utterances), and Neutral (180 utterances), under certain conditions in the three main categories of "congruent" (emotional lexical content expressed in a congruent emotional voice), "incongruent" (neutral utterances expressed in an emotional voice) and "baseline" (all emotional and neutral utterances expressed in neutral voice) in a specialized sound recording studio under the supervision of a linguist and an acoustic expert. In this research, only the audio file of these three groups is used, regardless of the categorization. In other words, the label of speakers' emotional expression is considered regardless of the lexical content of their speech. Persian emotional speech database (Persian ESD) is validated by a group of 1,126 native Persian speakers.

Since the mentioned Persian database is used in limited research works, another critical and widely used emotional database is considered support. The Berlin emotional database (EMO-DB) [8] is used to evaluate the performance of the designed system. The Berlin emotional database (EMO-DB) is a publicly available German emotional-based database. Because of its popularity and utility, this database is suitable for evaluating and comparing the results of the

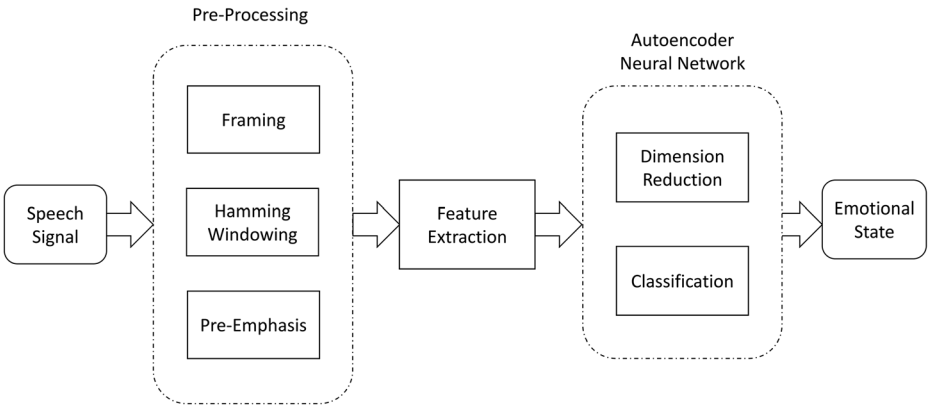

Fig. 1 The framework of the proposed SER system

proposed method with other studies. The Berlin emotional database consists of 5 male and five female speakers who spoke utterances of 1.5 to 5 s. Each one of the speakers is asked to speak ten different utterances in German. This database contains seven separate emotional classes, with the number of utterances for each class as follows: Anger (127), Happiness (71), Boredom (81), Sadness (62), Fear (69), Disgust (46), and Neutral state (79).

3.2 Features extraction

This study performs three pre-processing stages before feature extraction: Framing, Hamming windowing, and pre-emphasis. For framing, the length of each frame is considered 20 milliseconds, and the overlap of adjacent frames is assumed to be ten milliseconds. After the framing step, each frame is multiplied in the Hamming window to reduce the effect of signal discontinuity at both the beginning and end of each frame. The last step of the pre-processing is applying a low-pass filter to eliminate the effects of the sudden change of the continuous-time signal. In other words, this filter smooths the signal and eliminates sudden changes in the signal due to ambient noise and the speech production system. The conversion function of this filter is defined as Eq. 1, and the α is assumed to equal 0.97.

$$H(z) = 1 - \alpha z^{-1} \tag{1}$$

Some studies have executed other pre-processing, such as voice activity detection, normalization, noise reduction, and so forth [28, 35, 42]. Low-quality, the length of silence, and the quality of speech are also effective in expressing emotions. For this reason, other pre-processing steps are ignored in this study to prevent harming recognition accuracy.

Many features can be extracted from the speech signal, and given these features, information is transmitted or processed from the speech signal. In this study, features such as Energy, Pitch, the First to Fourth Formants (F_1- F_4), Jitter, Shimmer, 13 Linear Predictive Coefficients (LPCs), 13 Mel Frequency Cepstral Coefficients (MFCCs), and Zero-Crossing Rate (ZCR) are extracted from each frame. All of these extracted features were benchmarked in the studies related to emotion recognition. The first and second derivatives of each of these features are also obtained to provide 101 local (frame-level) features for each speech frame in aggregate. After extracting these local features, we get global (utterance-level) features by applying twenty statistical functions, including maximum, minimum, range, mean, variance, standard deviation, median, 1st, 5th, 10th, 25th, 75th, 90th, 95th, 99th percentile, quadratic range, 10% adorned average, 25% adorned average, skewness, and kurtosis. Thereby 2020 global features are obtained. Table 1 shows the quantity of these features.

Table 1 Quantity of global and local extracted features

Features	Number of local features	Number of global features
Energy	3	60
Pitch	3	60
MFCCs	39	780
LPCs	39	780
Formants	12	240
Jitter	1	20
Shimmer	1	20
ZCR	3	60
Sum Total	101	2020

Because of the capabilities of Autoencoder in dimensions reduction, we have selected this vast number of features. Section IV shows that larger feature sets may appropriately train Autoencoder in classification accuracy. Also, both types of feature sets, including global and local, have been used in this study. However, there is no consensus on whether it is better to use local or global features. However, this research work, like other studies, has shown that the efficiency and accuracy of classification will increase with global features.

3.3 Stacked Autoencoder (SAE)

When a set contains so many features, the classification may be difficult and costly, and it is better to have a smaller set of features dependent on the primitive set of features. The autoencoder neural network can quickly implement this dimension reduction in feature vectors. This section proposes the architecture of Stacked Autoencoder (SAE) for emotion classification. A Stacked Autoencoder is a neural network consisting of multiple layers of sparse Autoencoders wherein the outputs of every layer have linked to the inputs of the subsequent layer [48]. The sparse Autoencoder is a three-layer neural network including an encoder and a decoder that output units are directly connected back to input units [49]. Sparse autoencoders use at least three layers:

- An input layer. For example, it could be the audio signal itself or its features.
- A few smaller hidden layers form the encryption.
- An output layer wherein every neuron has the same meaning as the input layer and is a reconstruction of the input layer.

The general structure of an autoencoder with a hidden layer is shown in Fig. 2.

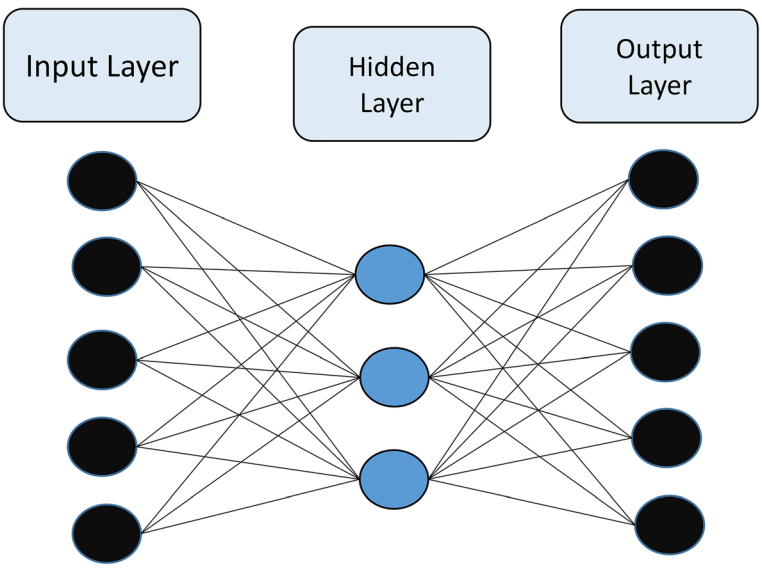

Fig. 2 Structure of a basic Autoencoder with a hidden layer

Autoencoders aim to learn a compressed data representation with minimum reconstruction loss [48]. The encoder in Autoencoder maps the input vector X_n to the hidden layer H_m with a non-linear function:

$$H_m = S\left(\sum_{i=0}^{n}(W_i{*}X_i) + b_m\right) \tag{2}$$

In this equation, n is the number of input or output neurons, m is the number of hidden neurons and W_i denotes the parameters (or weights) related to the connection between the input and hidden units. b_m are biases in the hidden layer. S(α) is the Sigmoid function. The sigmoid function is described as:

$$S(\alpha) = \frac{1}{1 + e^{-\alpha}} \tag{3}$$

The decoder maps the hidden layer H_m to the output layer Y_n that has an equally wide variety of units with the input layer:

$$Y_n = S\left(\sum_{j=0}^{m}\left(W_j{*}H_j\right) + b_n\right) \tag{4}$$

In Eq. (4) W_j denotes the parameters (or weights) related to the connection between the hidden and output units. b_n are biases in the output layer.

The error between X and Y must be minimized to train the Autoencoder and determine the optimized parameters [48].

$$\underset{w_i, w_j, b_n, b_m}{\arg\ \min}\ [error(X, Y)] \tag{5}$$

In this study, after pre-processing and extracting features from the database, two sparse autoencoders are used to classify emotions. The first sparse Autoencoder contains the input layer to learn the primary features of the raw input, illustrated in Fig. 3.

The first sparse Autoencoder produces the primary feature that contains 100 features (Feature Vector 1). The primary feature feeds the input layer into the second trained sparse Autoencoder that produces the secondary features that contain 50 features (Feature Vector 2). Figure 4 shows the primary features used as the raw input to the next sparse Autoencoder to learn secondary features.

The secondary feature is then handled as an input layer to a Softmax classifier to map secondary features to emotion labels, as shown in Fig. 5.

Finally, the first and second sparse Autoencoder has been mixed with Softmax classifier to provide three layers of stacked Autoencoder shown in Fig. 6. The proposed stacked Autoencoder contains two hidden layers (primary and secondary features) and the output layer (Softmax classifier) capable of classifying emotions from speech.

4 Experiments and results

The proposed method of this study is run on the Persian emotional speech database (Persian ESD) and Berlin emotional database (EMO-DB), and its results are presented in this section. After pre-processing and extracting the mentioned features from these databases, many feature sets are formed to be used as the input of the autoencoder neural network. As shown in

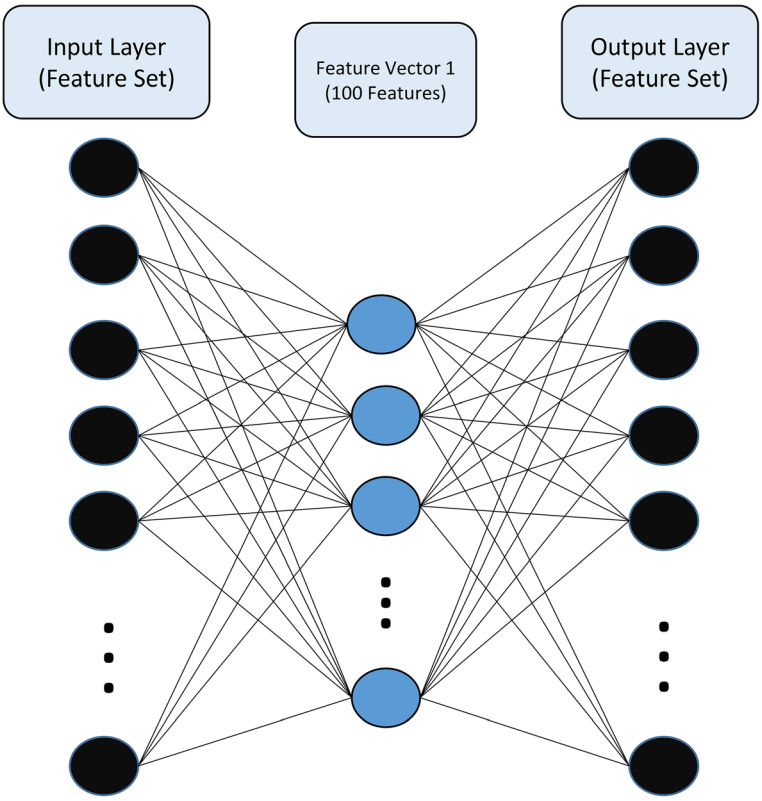

Fig. 3 Outline of the first sparse Autoencoder

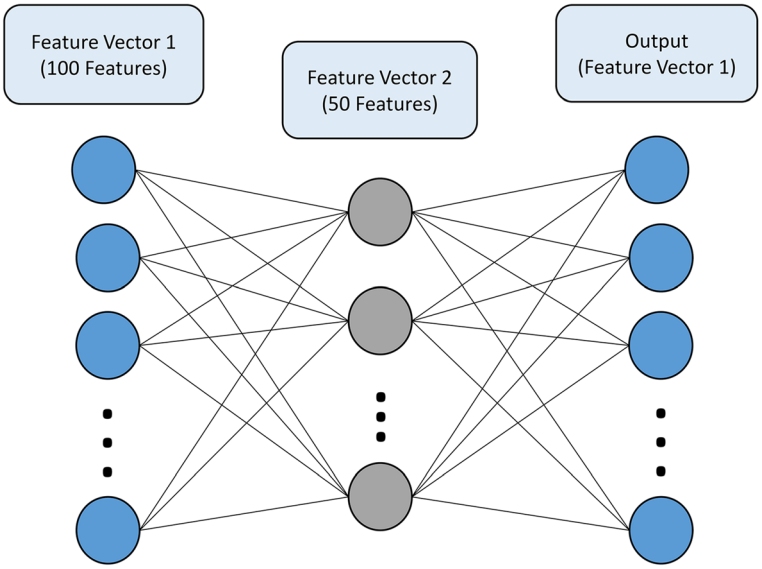

Fig. 4 Outline of the second sparse Autoencoder

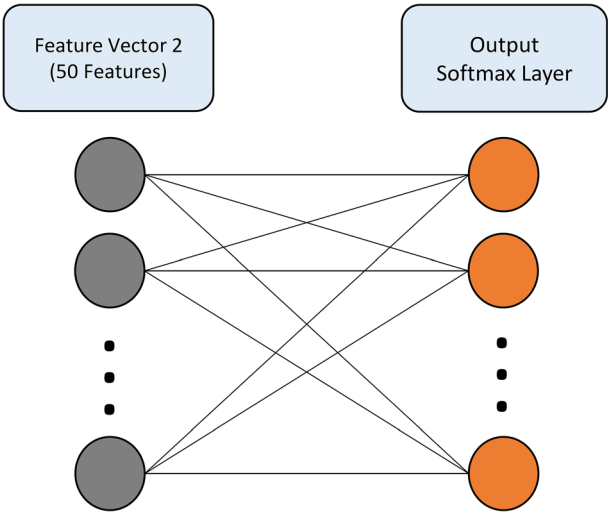

Fig. 5 Softmax classifier

Table 2, in Feature Set 1, all the extracted global features are used. Then, the experiments were repeated by reducing a few extracted features, which decreased emotion recognition accuracy. To show the importance of using this number of features, in Feature Set 2, some of these features, including LPCs and ZCR, have been omitted, and in Feature Set 3, addition to them, MFCCs have also been removed. Also, to show the priority and superiority of global features over local features, Feature Set 4 is assigned to local features in SER.

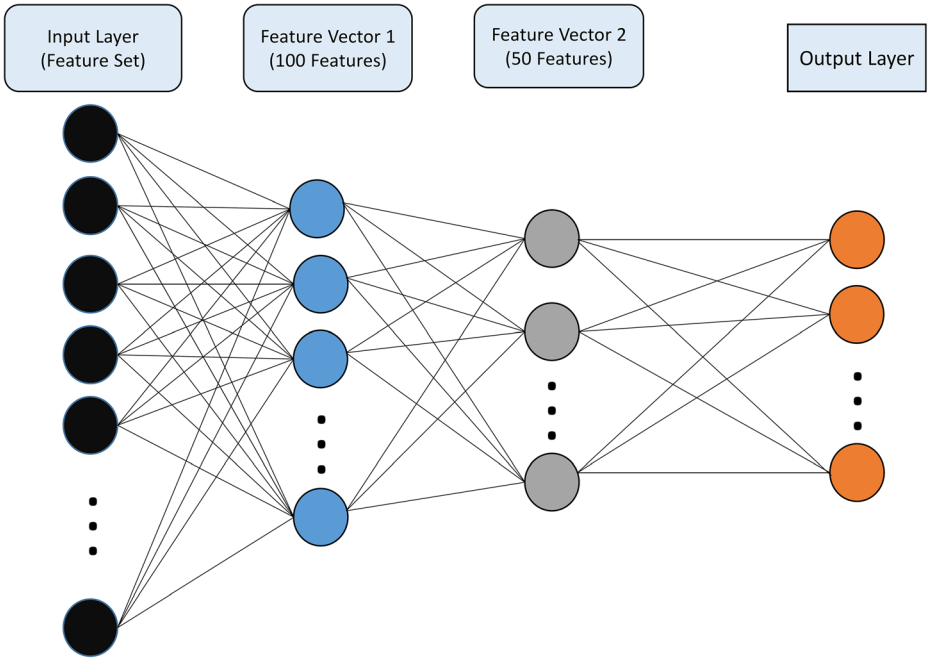

Fig. 6 Outline of the proposed stacked Autoencoder

Table 2 Properties of feature sets

Feature Set	Type	Used Features	Dimensions in Persian ESD
Feature Set 1	Global	Energy and its first and second derivatives (3), Pitch and its first and second derivatives (3), ZCR and its first and second derivatives (3), Jitter (1), Shimmer (1), First to Fourth Formants and Their first and second derivatives (12), LPCs and their first and second derivatives (39), MFCCs and their first and second derivatives (39)	2020×472
Feature Set 2	Global	Energy and its first and second derivatives (3), Pitch and its first and second derivatives (3), Jitter (1), Shimmer (1), First to Fourth Formants and Their first and second derivatives (12), MFCCs and their first and second derivatives (39)	1180×472
Feature Set 3	Global	Energy and its first and second derivatives (3), Pitch and its first and second derivatives (3), Jitter (1), Shimmer (1), First to Fourth Formants and Their first and second derivatives (12)	400×472
Feature Set 4	Local	Energy and its first and second derivatives (3), Pitch and its first and second derivatives (3), ZCR and its first and second derivatives (3), Jitter (1), Shimmer (1), First to Fourth Formants and Their first and second derivatives (12), LPCs and their first and second derivatives (39), MFCCs and their first and second derivatives (39)	101×231,982

To further explain Table 2, it is necessary to mention that in the column "Used Features," the number of features extracted from each frame is written in front of it. The total of these features in the feature sets 1 and 4, which have all the extracted features, is 101. Column "Dimensions in Persian ESD" also shows the input dimensions of the system for each feature set. Since in feature set 1, 101 features have been extracted for each frame, by applying the 20 statistical functions introduced in the paper, there will be a total of 2020 global features for 472 audio files in the Persian ESD. In feature sets 2 and 3, the input dimensions will be smaller due to reduced features. In feature set 4, since the used features are local, the input dimensions were 101 features for the number of Persian ESD frames.

It is needed to be highlighted that the experiments are performed in the MATLAB programming environment. The stacked Autoencoder consisted of an encoder with input layers of the dimensions of the Feature Set. The first sparse Autoencoder is constructed through an input layer of length 2020 in Feature Set 1, a hidden layer of length 100, and an output layer of the same length as the input layer. The first sparse Autoencoder is trained to generate 100 features. It should be noted that the maximum number of epochs is assumed to equal 1000, and the sparsity proportion is 0.05 in the first sparse Autoencoder. The subsequent sparse Autoencoder is designed with the aid of using an input layer of length 100, a hidden layer of length 50, and an output layer of length 100. The second sparse Autoencoder is trained to produce 50 features.

Moreover, in the next sparse Autoencoder, the maximum epochs and sparsity proportion are 400 and 0.05, respectively. Finally, those 50 features feed the Softmax layer. The Softmax layer is trained to generate the output class. Furthermore, cross-validation is used for validation in this study [37]. According to this validation, the data set is divided into N equal parts, and the classification is done with N. In each experiment, the N-1 part of each data set is used for training and one part for classification testing. This method allows all data to participate in training and testing. The final recognition accuracy will also be the average N accuracy of the calculated experiment. In this research work, as in [29, 52], N is considered equal to 10, meaning that 90% of the data are considered for training, and 10% are considered for testing in

Table 3 The confusion matrix of the proposed method for Persian ESD by Feature Set 1

	Sadness	Happiness	Neutral	Disgust	Fear	Anger
Sadness	**100%**	0	0	0	0	0
Happiness	0	**90%**	0	0	0	10%
Neutral	2.7%	0	**97.3%**	0	0	0
Disgust	0	0	0	**100%**	0	0
Fear	0	0	0	0	**100%**	0
Anger	0	0	12.5%	0	12.5%	**75.0%**
			95.6%			

each experiment. Simulations are performed based on the mentioned assumptions, and their results are as follows.

Tables 3, 4, 5 and 6 summarize the confusion matrix by using mentioned Feature Sets for the Persian emotional speech database (Persian ESD). As can be seen, for Feature Set 1, 95.6% of the predictions are correct.

The first thing to notice from the results of the experiments is that the lowest recognition accuracy usually occurs in Feature Set 4, where frame-based local features were used as autoencoder input. The Feature Set 4 experiment always takes longer than the other experiments, which leads to the conclusion that global features are far more efficient than local features and reduce computational time and cost. In global features, it is also observed that the best result is usually obtained in the Feature Set 1 experiment, where the feature vectors are fully used. Furthermore, according to Tables 3, 4, 5 and 6, misclassifications often occur between emotions with similar arousal. For example, in Tables 3 and 10% of Happiness labeled data are misclassified as anger because both these emotions are considered high arousal emotions.

The next point in Fig. 7 is that the general results of the Persian ESD are better than the Berlin emotional database (EMO-DB). The justification is that the Persian ESD is produced in 6 emotional classes by two speakers (a man and a woman). This dependence on the speaker can lead to higher recognition results and accuracy. Whereas in the Berlin emotional database (EMO-DB), there are ten different speakers and seven different emotional classes, which shows the higher quality of these data.

Table 7 compares classification and feature selection methods in other research works for speech emotion recognition. It is noticeable that all of these methods are implemented on the Berlin emotional database (EMO-DB). As it is shown, the proposed method has obtained a suitable accuracy rate.

Table 4 The confusion matrix of the proposed method for Persian ESD by Feature Set 2

	Sadness	Happiness	Neutral	Disgust	Fear	Anger
Sadness	**71.4%**	28.6%	0	0	0	0
Happiness	0	**83.3%**	0	0	16.7%	0
Neutral	0	0	**100%**	0	0	0
Disgust	0	0	0	**100%**	0	0
Fear	0	23.1%	0	0	**53.8%**	23.1%
Anger	0	6.3%	0	12.5%	18.7%	**62.5%**
			80.9%			

Table 5 The confusion matrix of the proposed method for Persian ESD by Feature Set 3

	Sadness	Happiness	Neutral	Disgust	Fear	Anger
Sadness	**66.7%**	0	0	16.7%	0	16.7%
Happiness	0	**100%**	0	0	0	0
Neutral	0	0	**100%**	0	0	0
Disgust	14.3%	0	0	**85.7%**	0	0
Fear	0	0	6.2%	0	**81.3%**	12.5%
Anger	18.2%	0	0	0	9.1%	**72.7%**
			89.4%			

Table 6 The confusion matrix of the proposed method for Persian ESD by Feature Set 4

	Sadness	Happiness	Neutral	Disgust	Fear	Anger
Sadness	**58.8%**	9.4%	12.3%	8.5%	3.3%	7.7%
Happiness	9.4%	**56.2%**	2.5%	13.5%	6.9%	11.5%
Neutral	5.5%	2.6%	**80.4%**	4.1%	5.3%	2.1%
Disgust	16.6%	9.8%	4.3%	**45.6%**	8.4%	15.3%
Fear	8.0%	6.1%	5.7%	10.0%	**59.7%**	10.4%
Anger	2.4%	7.9%	37.2%	9.8%	4.5%	**38.2%**
			45.6%			

Figure 8 illustrates the recognition accuracy rate using Feature Set 1 for the Persian ESD and Berlin emotional database (EMO-DB) for each emotion separately. As can be seen, the average accuracy of the Persian emotional speech database (Persian ESD) is more than the Berlin emotional database (EMO-DB). In addition to linguistic justifications, the number of speakers in the Persian emotional speech database (Persian ESD) is limited, and the Berlin emotional database (EMO-DB) is larger in dimensions. However, as seen in Fig. 8, more misclassification can occur in high arousal emotions such as Happiness and anger.

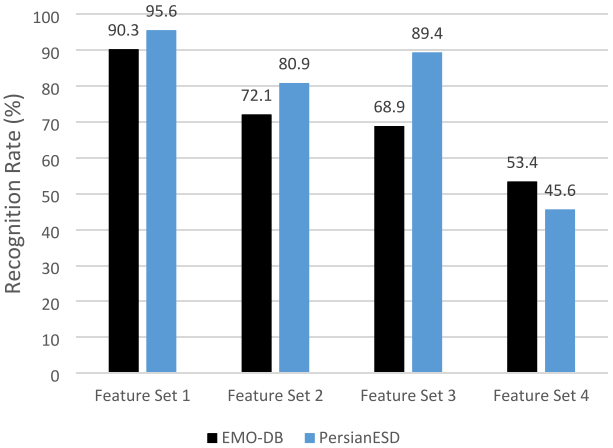

Fig. 7 Comparison of feature sets by the proposed method for Persian ESD and Berlin emotional database (EMO-DB)

Table 7 Comparative table of past works and proposed method on Berlin emotional database (EMO-DB).

Reference	Year	Classifier	Features	Recognition Accuracy
Borchert et al. [7]	2005	SVM, J48	Formants, Energy, HNR, Jitter, Shimmer.	70%
Yang et al. [54]	2010	Bayesian classifier	Prosodic, Spectral, and Voice Quality Features.	73.5%
Bitouk et al. [6]	2010	SVM	Spectral Features	81.3%
Albornoz et al. [2]	2011	Hierarchical classifier using HMM, GMM, and MLP	Log Spectrum, MFCC, and Prosodic Features.	71.5%
Shen et al. [44]	2011	SVM	Energy, Pitch, LPCC, MFCC, LPCMCC.	82.5%
Wu et al. [52]	2011	SVM	Prosodic Features, Speaking Rate, ZCR, TEO-based Features.	91.3%
Deng et al. [12]	2013	Denoising Autoencoder	ZCR, RMS, Energy, Pitch, HNR, MFCC.	57.9%
Li et al. [26]	2013	DNN-HMM	MFCC	77.9%
Cibau et al. [9]	2013	Deep Autoencoder	MFCC, MLS, Pitch, Energy.	70%
Mao et al. [32]	2014	CNN	Automatically Learned by CNN	85.2%
Wang et al. [50]	2015	SVM	Fourier Parameters, MFCC.	88.88%
Vasuki et al. [47]	2015	SVM	MFCC, PLP, Prosodic.	74.7%
Lanjewar et al. [23]	2015	GMM kNN	MFCC, Pitch, Wavelet.	66% 51%
Savargiv et al. [39]	2015	HMM	Energy, Pitch, MFCC.	79.6%
Zhao et al. [57]	2019	CNN	Local Feature Learning Blocks (LFLB).	95.33%
Daneshfar et al. [10]	2020	pQPSO, GMM	MFCC, PLPC, PMVDR, Pitch.	82.8%
Dissanayake et al. [14]	2020	CNN-LSTM	MFCC	69.7%
Proposed method	2022	Stacked Autoencoder	Feature Set 1	90.3%

5 Discussion

Persian emotional database (Persian ESD) has been used in this study, which has already been used as a benchmark in other studies. Persian ESD is the most appropriate and available emotional database in the Persian language. However, this database has problems such as being small for use in deep neural networks. Nevertheless, using features globally causes augmentation in the feature vector. However, producing an emotional speech database in Persian can be one of the future works of this study.

In general, the proposed method in this study is presented on two emotional databases: the Persian emotional speech database (Persian ESD) as the main database in the Persian language, and the German database (EMO-DB) to evaluate the proposed method with other studies. The reasons for choosing these two databases are given below.

- The Persian Emotional Speech Database (Persian ESD) is the most appropriate and available Persian database and includes all six primary emotions.
- The Berlin database is also used to evaluate the method since it is prevalent, and we could make an effective comparison between our method and others.
- Both databases have already been used in other studies as benchmarks [8, 20].
- The type and number of emotions used in both databases are close and comparable. Both databases have six primary emotions.

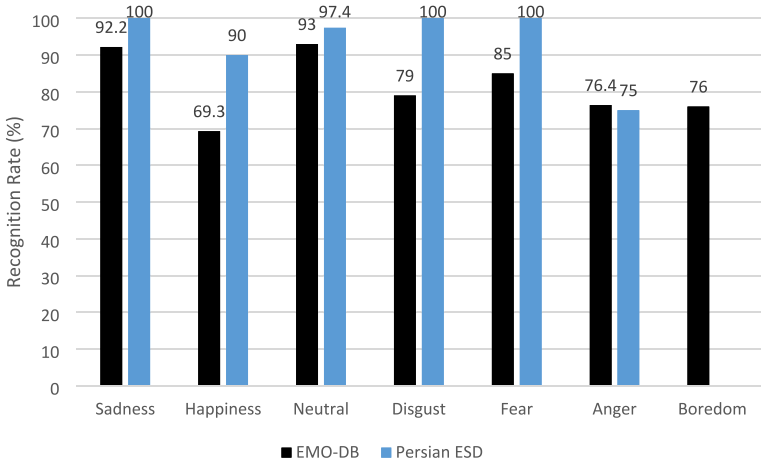

Fig. 8 Comparison of recognition accuracy of different classes in Berlin emotional database (EMO-DB) and Persian ESD with proposed autoencoder classification

In Fig. 9, a graph has been prepared for the statistical comparison of the type and number of different emotions in the databases used. As seen in Fig. 9, the Persian database is more unbalanced than the German one regarding the number of labeled audio files.

Moreover, the findings of the studies presented in the paper suggest that the Autoencoder-based methodology in two steps of feature reduction by sparse autoencoders has improved the accuracy of the speech emotion recognition system. Autoencoder performs both feature reduction and classification.

Generally, when an utterance is expressed in total, it is better to understand its emotions. Hearing a frame or word in the middle of an utterance makes it very difficult to judge its emotions. For this reason, it was expected from the beginning that global features extracted from the sentence level would be more effective than local features. The results of the experiments of this study also show the truth of this claim. Another advantage of global features is that they can augment input vectors in limited data, especially in the Autoencoder neural network where this augmentation is needed.

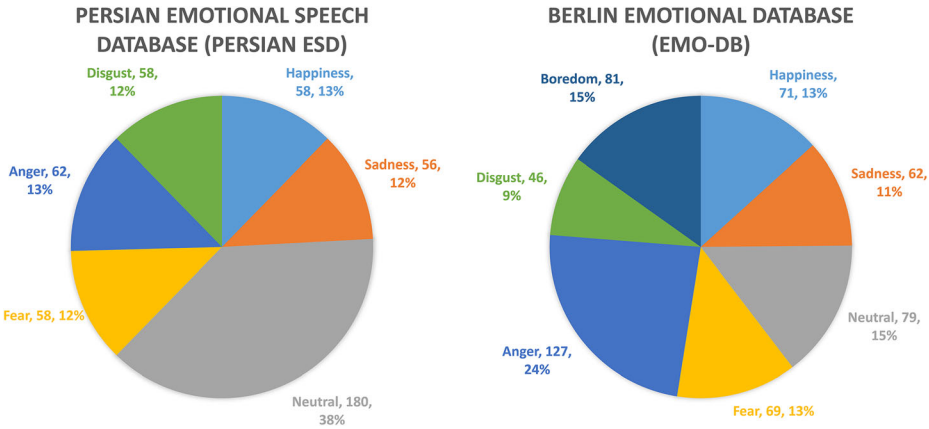

Fig. 9 Statistical comparison between used databases

6 Conclusion

As mentioned, speech emotion recognition plays an essential role in human-machine interaction development. This study discussed the impact and efficiency of using a stacked autoencoder neural network in SER. A Stacked Autoencoder was formed with two sparse autoencoders and a Softmax layer to reduce the dimension of features and classification. For the input of this stacked Autoencoder, features including Energy, Pitch, LPCs, Mel-Frequency Cepstral Coefficients (MFCCs), first to fourth Formants, Jitter, Shimmer, ZCR, and their first and second derivatives that were extracted as global features and local features were used. The performance of these features was also compared against each other, showing that global features are more efficient in speech emotion recognition. It should also be noted that in this study, a database in the Persian language (Persian ESD) was used on which the recognition accuracy of the proposed method is 95.6%. The proposed method was also presented on the popular Berlin emotional database (EMO-DB) and compared to other methods and studies, resulting in an accuracy of 90.3%, which indicates the efficiency of the stacked autoencoder neural network in speech emotion recognition.

Data availability Data sharing not applicable to this article as no datasets were generated during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76
2. Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. Comput Speech Lang 25(3):556–570
3. Badshah AM, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee MY, Baik SW (2019) Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications 78(5):5571–5589
4. Bashirpour M, Geravanchizadeh M (2018) Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments. EURASIP J Audio Speech Music Process 2018(1):1–13
5. Bastanfard A, Aghaahmadi M, Fazel M, Moghadam M (2009) Persian viseme classification for developing visual speech training application. In: Pacific-rim conference on multimedia. Springer, Berlin, pp 1080–1085
6. Bitouk D, Verma R, Nenkova A (2010) Class-level spectral features for emotion recognition. Speech Commun 52(7–8):613–625
7. Borchert M, Dusterhoft A (2005) Emotions in speech experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: 2005 international conference on natural language processing and knowledge engineering. IEEE, pp 147–151
8. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: Interspeech, vol 5, pp 1517–1520
9. Cibau NE, Albornoz EM, Rufiner HL (2013) Speech emotion recognition using a deep autoencoder. Anales de la XV Reunion de Procesamiento de la Informacion y Control, vol 16, pp 934–939
10. Daneshfar F, Kabudian SJ (2020) Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. Multimed Tools Appl 79(1):1261–1289
11. Dangol R, Alsadoon A, Prasad PWC, Seher I, Alsadoon OH (2020) Speech emotion recognition UsingConvolutional neural network and long-short TermMemory. Multimedia Tools and Applications 79(43):32917–32934

12. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: 2013 humaine association conference on affective computing and intelligent interaction. IEEE, pp 511–516
13. Deng J, Xu X, Zhang Z, Frühholz S, Schuller B (2017) Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Signal Process Lett 24(4):500–504
14. Dissanayake V, Zhang H, Billinghurst M, Nanayakkara S (2020) Speech emotion recognition 'in the wild' using an autoencoder. In: Interspeech, pp 526–530
15. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44(3):572–587
16. Gharavian D, Ahadi SM (2006) Recognition of emotional speech and speech emotion in Farsi. In: The Proceedings of international symposium on Chinese spoken language processing, vol 2, pp 299–308
17. Harimi A, Esmaileyan Z (2014) A database for automatic persian speech emotion recognition: collection, processing and evaluation. Int J Eng 27(1):79–90
18. Javidi MM, Roshan EF (2013) Speech emotion recognition by using combinations of C5. 0, neural network (NN), and support vector machines (SVM) classification methods. J Math Comput Sci 6(3):191–200
19. Keshtiari N, Kuhlmann M (2016) The effects of culture and gender on the recognition of emotional speech: evidence from persian speakers living in a collectivist society. Int J Soc Cult Lang 4(2):71–86
20. Keshtiari N, Kuhlmann M, Eslami M, Klann-Delius G (2015) Recognizing emotional speech in Persian: a validated database of Persian emotional speech (Persian ESD). Behav Res Methods 47(1):275–294
21. Kwon S (2021) MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Syst Appl 167:114177
22. Langari S, Marvi H, Zahedi M (2020) Efficient speech emotion recognition using modified feature extraction. Inf Med Unlocked 20:100424
23. Lanjewar RB, Mathurkar S, Patel N (2015) Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest neighbor (K-NN) techniques. Procedia Comput Sci 49:50–57
24. Latif S, Rana R, Khalifa S, Jurdak R, Epps J, Schuller BW (2020) Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Trans Affect Comput 13:992–1004
25. Lee J, Tashev I (2015) High-level feature representation using recurrent neural network for speech emotion recognition. In: Interspeech 2015
26. Li L, Zhao Y, Jiang D, Zhang Y, Wang F, Gonzalez I … Sahli H (2013) Hybrid deep neural network–hidden Markov model (dnn-hmm) based speech emotion recognition. In: 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, pp 312–317
27. Low LSA, Maddage NC, Lech M, Sheeber LB, Allen NB (2010) Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans Biomed Eng 58(3):574–586
28. Luengo I, Navas E, Hernáez I, Sánchez J (2005) Automatic emotion recognition using prosodic parameters. In: Ninth European conference on speech communication and technology
29. Lugger M, Yang B (2007) The relevance of voice quality features in speaker-independent emotion recognition. In: 2007 IEEE international conference on acoustics, speech and signal processing-ICASSP'07, vol 4. IEEE, pp IV-17
30. Mahdavi R, Bastanfard A, Amirkhani D (2020) Persian accents identification using modeling of speech articulatory features. In: 2020 25th international computer conference, computer society of Iran (CSICC). IEEE, pp 1–9
31. Mansoorizadeh M, Moghaddam Charkari N (2010) Multimodal information fusion application to human emotion recognition from face and speech. Multimedia Tools and Applications 49(2):277–297
32. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimed 16(8):2203–2213
33. Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2227–2231
34. Mohamad Nezami O, Jamshid Lou P, Karami M (2019) ShEMO: a large-scale validated database for persian speech emotion detection. Lang Resour Eval 53(1):1–16
35. Pohjalainen J, Fabien Ringeval F, Zhang Z, Schuller B (2016) Spectral and cepstral audio noise reduction techniques in speech emotion recognition. In: Proceedings of the 24th ACM international conference on multimedia, pp 670–674
36. Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. Int J Speech Technol 16(2):143–160
37. Raudys SJ, Jain AK (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 13(3):252–264

38. Savargiv M, Bastanfard A (2014) Study on unit selection and statistical parametric speech synthesis techniques. J Comput Rob 7(1):19–25
39. Savargiv M, Bastanfard A (2015) Persian speech emotion recognition. In: 2015 7th Conference on Information and Knowledge Technology (IKT). IEEE, pp 1–5
40. Savargiv M, Bastanfard A (2016) Real-time speech emotion recognition by minimum number of features. 2016 Artificial Intelligence and Robotics (IRANOPEN). IEEE, pp 72–76
41. Schuller B, Reiter S, Muller R, Al-Hames M, Lang M, Rigoll G (2005) Speaker independent speech emotion recognition by ensemble classification. In: 2005 IEEE international conference on multimedia and expo. IEEE, pp 864–867
42. Schuller B, Müller R, Lang M, Rigoll G (2005) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In: Proc. of Interspeech 2005-Proc. Europ. Conf. on Speech Communication and Technology, Lisbon, Portugal
43. Sedaaghi M (2008) Documentation of the sahand emotional speech database (SES). Department of engineering, Sahand University of Technology
44. Shen P, Changjun Z, Chen X (2011) Automatic speech emotion recognition using support vector machine. In: Proceedings of 2011 international conference on electronic & mechanical engineering and information technology, vol 2. IEEE, pp 621–625
45. Shirani A, Nilchi ARN (2016) Speech emotion recognition based on SVM as both feature selector and classifier. Int J Image Graphics Signal Process 8(4):39
46. Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 21(1):93–120
47. Vasuki P (2015) Speech emotion recognition using adaptive ensemble of class specific classifiers. Res J Appl Sci Eng Technol 9(12):1105–1114
48. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning, pp 1096–1103
49. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA, Bottou L (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11(12):3371–3408
50. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using Fourier parameters. IEEE Trans Affect Comput 6(1):69–75
51. Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E (2021) A comprehensive review of speech emotion recognition systems. IEEE Access 9:47795–47814
52. Wu S, Falk TH, Chan WY (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53(5):768–785
53. Yadav SP, Zaidi S, Mishra A, Yadav V (2022) Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Arch Comput Methods Eng 29(3):1753–1770
54. Yang B, Lugger M (2010) Emotion recognition from speech signals using new harmony features. Sig Process 90(5):1415–1423
55. Yang Y, Xu F (2022) Review of research on speech emotion recognition. In: International conference on machine learning and intelligent communications. Springer, Cham, pp 315–326
56. Yazdani A, Simchi H, Shekofteh Y (2021) Emotion recognition in persian speech using deep neural networks. In: 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE). IEEE, pp 374–378
57. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Process Control 47:312–323