Multimedia Tools and Applications (2021) 80:8127–8146
https://doi.org/10.1007/s11042-020-10119-w

Speaker independent feature selection for speech emotion recognition: A multi-task approach

Elham Kalhor[1] · Behzad Bakhtiari[1]

Received: 4 May 2019 / Revised: 25 August 2020 / Accepted: 19 October 2020 /
Published online: 31 October 2020

Abstract
Nowadays, automatic speech emotion recognition has numerous applications. One of the important steps of these systems is the feature selection step. Because it is not known which acoustic features of person's speech are related to speech emotion, much effort has been made to introduce several acoustic features. However, since employing all of these features will lower the learning efficiency of classifiers, it is necessary to select some features. Moreover, when there are several speakers, choosing speaker-independent features is required. For this reason, the present paper attempts to select features which are not only related to the emotion of speech, but are also speaker-independent. For this purpose, the current study proposes a multi-task approach which selects the proper speaker-independent features for each pair of classes. The selected features are then given to the classifier. Finally, the outputs of the classifiers are appropriately combined to achieve an output of a multi-class problem. Simulation results reveal that the proposed approach outperforms other methods and offers higher efficiency in terms of detection accuracy and runtime.

Keywords Speech emotion recognition · Multi-task feature selection · Speaker independent features

1 Introduction

In the formation of an individual's emotions, there is a set of emotions, such as happiness, sadness, anger, disgust, boredom, surprise, fear, and neutrality. Emotions play a prominent role in human communications. They are critical for exhibiting behavior under different conditions.

✉ Behzad Bakhtiari
bakhtiari@sadjad.ac.ir

Elham Kalhor
e.kalhor333@sadjad.ac.ir

[1] Department of Computer Engineering, Sadjad University of Technology, No. 64 Jalal Al Ahmad St, 9188148848 Mashhad, Iran

On one hand, emotions cause psychological changes, which form in the brain and manifest themselves in human reactions. In addition, emotions increase or decrease the physiological stimuli of the body and positively or negatively impact behavior and thoughts. Moreover, the emotion of speech depends on the speaker's language and culture, gender, age, speech content, and other factors [12, 20].

Speech emotion recognition offers numerous applications in human-machine communication systems. For example, there are different applications in the field of education, computer games, medicine, customer communication systems, telephone centers, and mobile communications [5, 31, 32, 44, 45]. Nevertheless, automatic speech emotion recognition requires an acoustic feature extraction. Since there is no information about which features are related to a speaker's emotions, many researchers have proposed several features. Unfortunately, employing all of these may pose two basic challenges. The first one is the low number of training samples and large number of features, which lead to data overfitting and a possible reduction in system efficiency. The second challenge is the increase in the algorithm runtime. As a consequence, it is necessary to select features which are related to emotion classes and which are also speaker-independent [42].

Despite the challenges mentioned, there are several methods for selecting features. One method is multi-task approach, which is the main focus of the present paper. A few studies have investigated the selection of speaker-independent features through a multi-task approach. For instance, multi-task learning is only employed in [50, 51] for learning a shared subspace for singing and speaking. In this regard, different tasks are separated in terms of male-female and speech-song. As a result, the selected features are not speaker-independent.

The main aim of the current paper is finding speaker-independent features. For this purpose, a multi-task objective function is considered by which common features among all speakers are obtained for each pair of emotion classes. Then, through a training of binary classifiers and a proper combination of their output, the predicted label is obtained.

In the present paper, Section 2 presents the related work. Section 3 explains some related preliminaries. Section 4 introduces the proposed method in detail. Section 5 discusses the experiments and analyzes the results. Finally, Section 6 provides the conclusion.

2 Related work

Speech emotion recognition systems are generally divided into two categories: speaker-dependent and speaker-independent. Numerous methods have been proposed to meet the challenges posed in both groups. Such methods attempt to improve efficiency by reducing the number of features. For example, speaker-dependent systems rely on a limited number of speakers and so may face low efficiency when new speakers are introduced. In these systems, various feature selection and dimensionality reduction methods have been used to overcome the high number of features. Some of these methods will be introduced in the following. The Locality Preserving Projections (LPP) [53], Diffusion Map (DM) [40], Isomap [49], and Kernel PCA [4, 16, 39] methods have been employed to reduce the dimensionality of the feature matrix.

Some works utilized correlation-based methods. For instance, [29] employs canonical correlation to compute the correlation between two groups of features and this results in the selection of features with the highest correlation. Moreover, canonical correlation and kernel methods are utilized for feature selection in two groups of voice and video [13, 15, 33]. In addition, [19] employs the correlation between features and class labels for feature selection. There are also unsupervised methods, such as Multi-Cluster Feature Selection (MCFS) [7], which are based on data clustering. In addition to the clustering method, evolutionary approaches, such as Particle Swarm Optimization (PSO) and Biogeography-based Optimization (BBO), are employed for feature selection in [41, 46–48].

Because there are numerous speakers in most speech emotion systems, the speaker-independent system is vital [42]. The main aim of some works is to meet this challenge by attempting to train systems which are suitably efficient for a new speaker. Research in speaker-independent feature selection can be generally categorized into five groups. Table 1 presents a summary of works on speaker-independent feature selection.

The first group utilizes baseline feature selection methods [21, 26, 42]. The second group proposes a cost function for feature selection and employs evolutionary methods to solve it [22, 24, 41]. The main drawback of the methods used by these two groups is their high runtime. The third group utilizes the relationship between features and class labels and selects features which have the highest correlation with class labels. For this reason, a different correlation function is used, such as the Canonical, Spearman, and Pearson [8, 25, 36]. Also, in [43], Mutual Information (MI) is employed to calculate the correlation between classes and features. This method shows how many features provide information about a class.

The fourth group has a multi-task approach to feature selection. However, a small amount of research has employed multi-task methods to find a speaker-independent sub-space [50, 51]. Such works simultaneously perform feature subspace learning and classifier training. These methods consider four tasks: female-speech, female-song, male-speech, and male-song. Also, four emotion classes are explored and a combination of six binary classifiers with one-against-one strategy is used. Each trained classifier receives test data and the class of the test data is obtained by majority voting on the output of the classifiers.

It should be noted that the methods in [50, 51] select features which are independent of male/female or song/speaking. The present study reveals that these methods can be employed for speaker-independent feature selection by applying some changes to multi-task systems [50, 51]. For this purpose, each speaker should be considered as a task. Moreover, only SVM classifiers can be used in these methods and, unfortunately, this is time-consuming in high dimensional problems. In contrast, the proposed method performs feature selection in two separate phases: feature selection and classifier training. As a result, different classifiers can be employed. Furthermore, with the proposed method's low runtime, the present study suggests a regression-based classifier output fuser.

Table 1 A summary of works on speaker-independent feature selection for emotion recognition

	Category name	Description	Publication year
1	Baseline	Forward Feature Selection Method [21, 26, 42]	2007, 2010, 2012
2	Evolutionary algorithms	Best-First method [22], Particle Swarm Optimization [41], Genetic Algorithm [24]	2010, 2015, 2018
3	Correlation-based	Spearman Method [25], Canonical Correlation [36], Canonical-Based Feature Selection [8], Mutual Information [43]	2018, 2016, 2018, 2017, 2018
4	Multi-task	[50, 51]	2016, 2017
5	Others	Gaussian Kernel Technique [17], PLDA-Based [6], Three-Phase Feature Normalization [18]	2014, 2018, 2016

There are also other methods which are mentioned in Table 1. For instance, [17] divides data into n groups (the optimal value for n is obtained through different experiments). Then, a Gaussian kernel with different values of sigma is applied to each group and common features among each group are selected. Finally, the union of all features is obtained from the n groups.

Dang and et al. introduced speaker-related factors for each speaker [6] which employ the probabilistic linear discriminant analysis (PLDA) technique. This technique uses emotion factors related to each speaker and provides information about the features at each frame. In each step, information about the features of each speaker's whole frames is obtained. As a result, this technique finds the feature space containing information about all speakers.

In three steps, the cascaded normalization method in [18] normalizes features which are related to each speaker. The first step carries out normalization as presented in [34]. The second step performs normalization in order to prevent sparsity, which is stated as$f(x) = sig n(x)|x|^{\alpha}$, where $0 \leq \alpha \leq 1$. The last step normalizes each feature vector x by L_2-norm. Finally, these three normalization steps remove the redundant features.

3 Preliminary

3.1 Multi-task feature selection

Multi-task systems are useful for feature selection as they find common features among all tasks. As a result, the obtained space contains information about all tasks, thus improving the efficiency of the classifier [23, 30]. Although there are a high number of features in a multi-task system, a feature matrix must become sparse in order to achieve the desired features. Various methods may be applied for feature matrix sparsity. Usually, $L_{2,1}$-norm is utilized for sparsity and for selecting suitable and common features among different tasks [35, 37, 38]. The general objective function can be considered as:

$$WminLoss(W, X, Y) + \alpha \parallel W \parallel_{2,1} \tag{1}$$

where $Loss(W, X, Y)$ is a smooth convex loss function as least square or logistic loss. Also, $L_{2,1}$-norm is a non-smooth function and can be calculated as:

$$\parallel W \parallel_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{T} (w(i,j))^2} = \sum_{i=1}^{d} \parallel w^i \parallel_2 \tag{2}$$

where T is the number of tasks and d is the number of features. $W \in \mathbb{R}^{d \times T}$ is the weight matrix in which the i-th row is denoted by w^i and relates to the i-th feature and each W column relates to a certain task. Moreover, α is the regularization parameter which can be used to regularize the sparsity.

As seen, the objective function (1) contains two terms. The first term models the relationship between features and labels and the second term regulates the sparsity. Since $L_{2,1}$-norm performs the regularization, based on (2), the first L_2-norm is applied to the rows of W, followed by the application of the L_1-norm, which is common among all tasks. Consequently, after the $L_{2,1}$-norm application, if one row of the W matrix, corresponding to one feature, nears zero, then that feature will not be selected. Finally, selected features are common among all tasks. In this case, the value of α determines how many selected features are appropriate.

3.2 Combination of binary classifiers

The current study deals with a multi-class problem. There are two approaches for performing multi-class classification. The first is a multi-class method which is directly employed. The second approach employs a number of binary classifiers and performs multi-class classification by a combination of binary classifier outputs. It is worth mentioning that, if the features between different pairs of classes are not the same, then the features from each pair cannot be combined or a multi-class classifier cannot be directly utilized. This occurs, for example, when selected features for classifying the sadness and happiness classes differ from those features selected for classifying the fear and sadness classes.

Implementing a multi-class classifier with binary classification requires a suitable combination of binary classifiers. The majority voting method is the easiest combination for this purpose. In this method, each two-class classifier reports a vote on the label of the test data. The test data will then belong to the class receiving the highest vote. Majority voting may have low efficiency. Another method uses the scores reported by each two-class classifier. These scores represent the belonging degree of data to a class. Based on the number of classifiers and classes, the estimated belonging degree of data to class k, denoted by $\widehat{k}$, can be calculated as:

$$k\widehat{k} = argmin \frac{\sum_{j=1}^{L} \left| m_{kj} \right| g(m_{kj}, s_j)}{\sum_{j=1}^{L} \left| m_{kj} \right|} \tag{3}$$

where $k \in \{1, 2, , \mathrm{K}\}$ and K denotes the number of classes, L is the number of classifiers, g represents the binary loss function, and s_j stands for the score of the j-th classifier.$m_{kj} \in \{0, 1, -1\}$ represents whether class k is related to classifier $j$or not. If class k is not related to classifier j, then $m_{kj} = 0$. If class k is related to the first class of classifier j, $m_{kj} = 1$; otherwise, $m_{kj} = -1$ [1, 9, 14].

It is important to note that the methods discussed may have low efficiency if the feature space of classifiers is not the same and if the features obtained from each pair differ from each other. Low efficiency occurs when classifier inputs are different and their scores cannot be combined. Hence, if classifiers have a different input space, an efficient algorithm is needed. This algorithm should be able to combine classifier outputs and, for this reason, the present study suggests a regression-based method.

With the assumption that there are a number of trained binary classifiers, regression methods are appropriate for a combination of binary classifier outputs. In this regard, for the estimation of labels, the scores generated by each trained binary classifier can be given to the trained regressor model for estimating the label class.

To train the regressor, the outputs of the classifiers and true labels act as the training samples. Figure 1 demonstrates the regressor training process. Based on this figure, the features related to each pair of classes are extracted from all training data; generally, the features related to each pair of classes can differ from those of other pairs. Then, the trained binary classifiers receive these features and the scores obtained from each binary classifier, as well as the sample labels, are given to the regressor as inputs. Finally, the trained regressor model is produced, which can combine classifiers and estimate final labels.

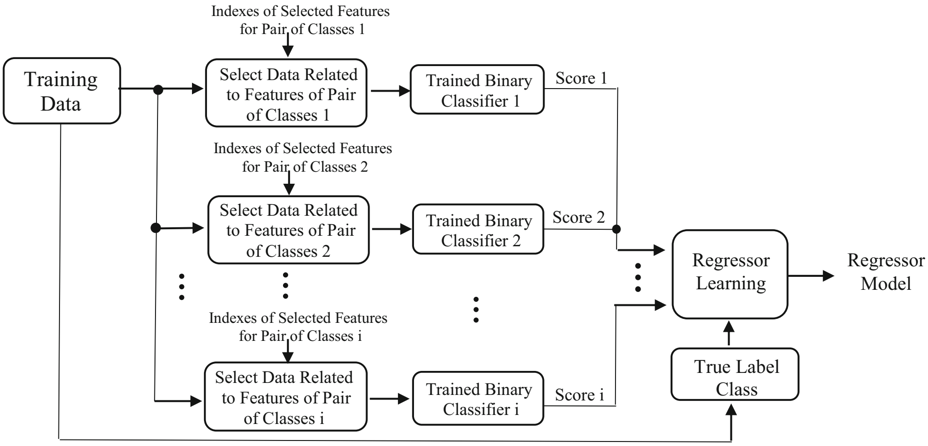
Indexes of Selected Features for Pair of Classes 1
Training Data
Select Data Related to Features of Pair of Classes 1
Trained Binary Classifier 1
Score 1
Indexes of Selected Features for Pair of Classes 2
Select Data Related to Features of Pair of Classes 2
Trained Binary Classifier 2
Score 2
Regressor Learning
Regressor Model
Indexes of Selected Features for Pair of Classes i
True Label Class
Select Data Related to Features of Pair of Classes i
Trained Binary Classifier i
Score i

Fig. 1 Regression training to combine the output of the binary classifiers

4 Proposed method

The current paper addresses a multi-class problem containing K classes and solves this with binary classifiers. Consequently, the data is considered as pair of classes and the feature selection process should be performed separately for each pair of classes. To solve the multi-class problem via binary classifiers, it is noteworthy that two strategies are used: one-against-one and one-against-all. In the one-against-one strategy, two different emotions can be perceived as a pair of classes. In the one-against-all strategy, the positive class includes one emotion and the other class includes other emotions.

As mentioned, the proper feature space between pairs of classes may be different. For example, the proper features for classifying the emotions of sadness and disgust may have no relation to features related to the emotions of happiness and neutrality. Therefore, it is necessary to separately perform the selection and training processes for each pair of classes.

Assume the dataset contains N data from T speakers. In this case, for each pair of classes, assume that $x_s \in \mathbb{R}^{n_s \times d}$ is related to the data of the s-th speaker and n_s is the number of data. Moreover, $y_s \in R^{n_s \times 1}$ is related to the data labels of the s-th speaker, where $s \in \{1, 2, , \mathrm{T}\}$. Also, $x_{s,j}$ denotes the j-th data of the s-th speaker and its class label is $y_{s,j} \in \{+1, -1\}$, which represents the j-th element of y_s.

Since the current paper's main purpose is selecting common features among speakers, the objective function is proposed as:

$$W, c\min \sum_{s=1}^{T} \sum_{j=1}^{n_s} log\left(1 + exp\left(-y_{s,j}\left(w_s^T x_{s,j} + c_s\right)\right)\right) + \rho_1 \parallel W \parallel_F^2 + \rho_2 \parallel W \parallel_{2,1} \tag{4}$$

where $W \in \mathbb{R}^{d \times T}$ is the weight matrix, $w_s \in \mathbb{R}^{d \times 1}$ represents the s-th column of W and the weight matrix of the s-th speaker, and c_s denotes the bias term of the s-th speaker. In the first term of (4), there is a logistic regression-based classifier for each speaker and w_s is the weight vector of this classifier for the s-th speaker. Hence, if one element of w_s nears zero, this implies that the corresponding element has almost no impact on classifying the data of the s-th speaker. The second term of (4) attempts to control the complexity of each speaker's classifier, which is achieved by changing ρ_1. As described in Section 3.1, the third term of (4) plays the role of feature selector.

By applying the proper coefficients of ρ_1 and ρ_2, the current work searches W, which thus minimizes (4). In order to reach the optimal point, some rows of matrix W will become zero, which means that some features will be removed. As a result, the remaining features are common among all tasks (speakers) and can describe the class label (emotion) for any speaker.

Figure 2 provides a schematic of the training process. According to this figure, in the first step, training data related to each pair of classes are separated. In the feature selection step, the data of each pair of classes are given to the objective function (4) for selecting common features among all speakers in each pair. Figure 3 illustrates the process in detail. Based on Fig. 3, matrix W is calculated. Then, the indexes of the non-zero row of W are stored. These index features are related to a pair of classes and are thus common among speakers. It should be noted that the selected features for each pair of classes can differ.

For each pair of classes, the next step separately extracts the selected features from all data. Then, the selected features are given to the binary classifiers and the trained models for each pair of classes are obtained. Finally, to train the regressor and combine the classifier outputs, all steps explained in Section 3.2 are taken.

According to Fig. 4, when the test data are introduced, the features obtained from each pair of classes in the training step are extracted from the test data. Then, the set of obtained features are given to the trained models and the generated scores from each model are combined based on the trained regression model. As a result, the estimation of the test data label is calculated.

The objective function represented in (4) consists of two terms: the smooth term and non-smooth term, respectively. The smooth term is $F(W) = \sum_{s=1}^{T} \sum_{j=1}^{n_s} log \left(1 + exp\left(-y_{s,j}\left(w_s^T x_{s,j} + c_s\right)\right)\right) + \rho_{L2} \parallel W \parallel_F^2$, which is a logistic regression classifier. The non-smooth term is$G(W) = \rho_1 \parallel W \parallel_{2,1}$. Consequently, minimizing $F(W) + G(W)$ has no closed-form solution. However, several methods can solve this problem [1, 2, 28] and the present paper utilizes the algorithm represented in [2] to do so. The codes employed for this purpose are adopted from [52].

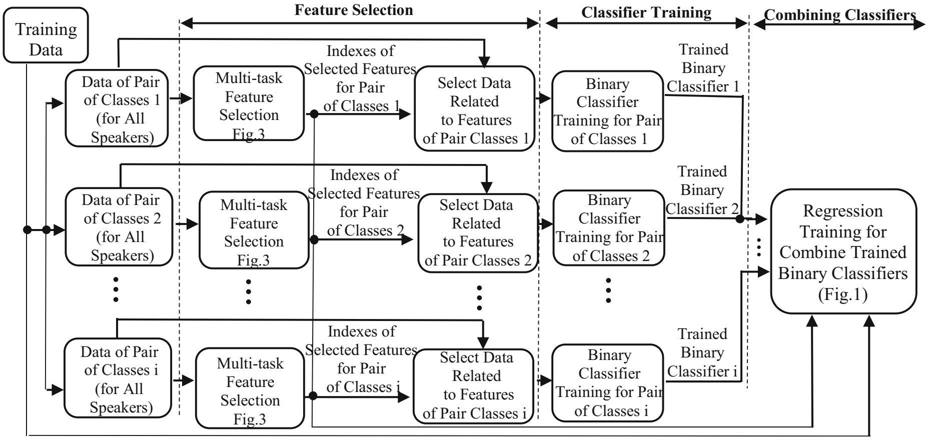
Training Data
Feature Selection
Classifier Training
Combining Classifiers
Data of Pair of Classes 1 (for All Speakers)
Multi-task Feature Selection Fig.3
Indexes of Selected Features for Pair of Classes 1
Select Data Related to Features of Pair Classes 1
Binary Classifier Training for Pair of Classes 1
Trained Binary Classifier 1
Data of Pair of Classes 2 (for All Speakers)
Multi-task Feature Selection Fig.3
Indexes of Selected Features for Pair of Classes 2
Select Data Related to Features of Pair Classes 2
Binary Classifier Training for Pair of Classes 2
Trained Binary Classifier 2
Regression Training for Combine Trained Binary Classifiers (Fig.1)
Data of Pair of Classes i (for All Speakers)
Multi-task Feature Selection Fig.3
Indexes of Selected Features for Pair of Classes i
Select Data Related to Features of Pair Classes i
Binary Classifier Training for Pair of Classes i
Trained Binary Classifier i

Fig. 2 A schematic of feature selection, classifier training and combiner learning for different pairs of classes

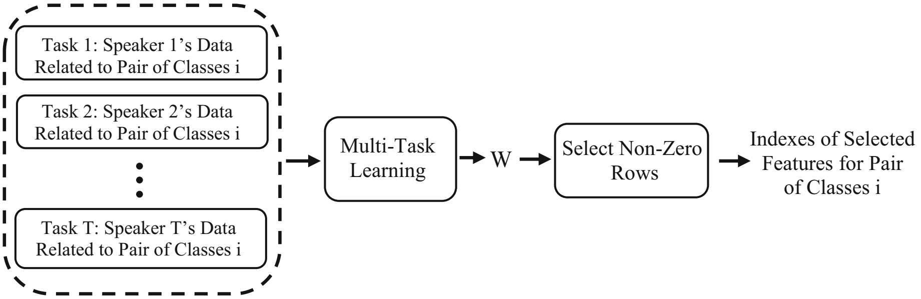
Task 1: Speaker 1's Data Related to Pair of Classes i
Task 2: Speaker 2's Data Related to Pair of Classes i
Task T: Speaker T's Data Related to Pair of Classes i
Multi-Task Learning
W
Select Non-Zero Rows
Indexes of Selected Features for Pair of Classes i

Fig. 3 Feature selection for i-th pair of classes

5 Experiments

In this section, the current paper's dataset is introduced. In addition, this section briefly explains the compared methods, provides the implementation details, and finally presents and discusses the results.

5.1 Datasets and compared methods

The current paper employs two datasets: EMO-DB[1] and ENTERFACE[2]. Information related to both datasets are represented in the following:

A. EMO-DB Dataset: This dataset features seven emotions: happiness (HA), anger (AN), disgust (DI), boredom (BO), sadness (SA), fear (FE), and neutrality (NE) [3]. Also, this dataset contains 535 voice files from 10 German speakers. The study employs all of these emotions and voice files.

B. ENTERFACE Dataset: This dataset features six emotions: happiness (HA), anger (AN), sadness (SA), fear (FE), disgust (DI), and surprise (SU) [27]. Additionally, this dataset contains 1,287 video files from 43 English speakers. It should be noted that the study uses all the emotions and voice files of each video.

In order to evaluate the proposed method and investigate its performance, the present paper compares it with some related approaches. These methods are summarized in the following. Section 2 provides more descriptions.

- FFS: A basic and common feature selection method [42].
- FFS + Tree: In this method, a binary tree is constructed based on the three dimensions of emotions [21].
- MI: First, the Z-Score technique normalizes the features of each speaker [11]. Through normalization, some of the less related features are removed. Then, the MI method calculates the correlation between these features and its labels. Finally, the proper features are selected [43].
- Spearman: This method first calculates the Euclidean distance and divides the features into four groups. Then, the partial correlation between these four groups is obtained and the Fisher method performs the dimensionality reduction [25]. In order to have the same work

[1] http://emodb.bilderbar.info/docu/.
[2] http://enterface.net/enterface05/main.php?frame=emotion.

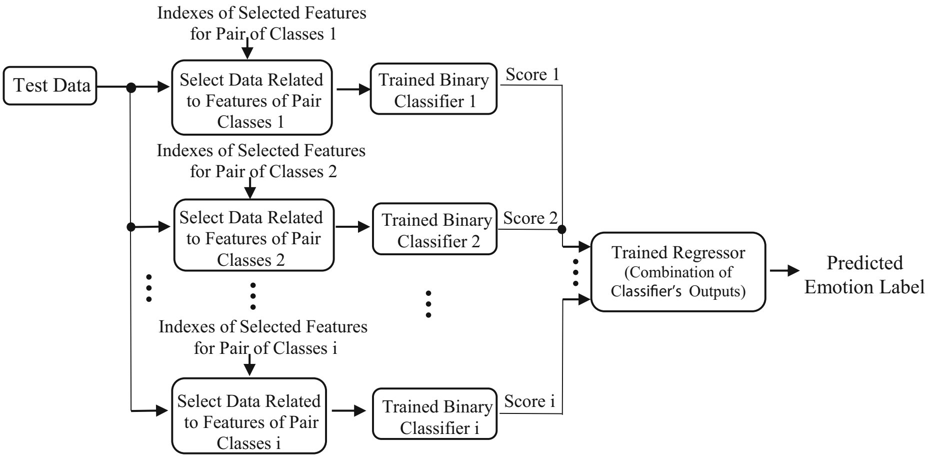
Indexes of Selected Features for Pair of Classes 1
Test Data
Select Data Related to Features of Pair Classes 1
Trained Binary Classifier 1
Score 1
Indexes of Selected Features for Pair of Classes 2
Select Data Related to Features of Pair Classes 2
Trained Binary Classifier 2
Score 2
Trained Regressor (Combination of Classifier's Outputs)
Predicted Emotion Label
Indexes of Selected Features for Pair of Classes i
Select Data Related to Features of Pair Classes i
Trained Binary Classifier i
Score i

Fig. 4 A schematic of the test process

performed in all of the compared methods, only the Spearman feature selection step is considered in the implementation.

- Speaker-Normalization: Three normalizations are performed [18].
- MTFS: This multi-task approach [51] selects corpus-independent features. However, the present study modifies this method to choose speaker-independent features.

5.2 Implementation details

The current work extracts features with openSMILE[3] software [10]. The "*emo_large*" config in this software extracts 6,552 features from each voice file. In addition, classification is achieved by a two-class SVM classifier employing one-against-one (OAO) and one-against-all (OAA) strategies. Since these classifiers have some parameters, the following provides the different intervals for parameters tuning in on all experiments:

- Denoting the sparsity of the feature matrix, parameter ρ_1 in (4) is considered within $\{10^1, 10^2, , 10^4\}$. Also, parameter ρ_2 in (4) is included in $\{10^4, 10^5, , 10^9\}$.
- The number of selected features in each compared method is considered within $\{100, 150, , 700\}$.
- All experiments utilize the SVM classifiers with a linear kernel, as well as the OAO and OAA strategies. Sixteen values are considered within $\{0.001, 0.003, 0.005, 0.02, 0.06, 0.08, 0.1, 0.2, , 1\}$ for parameter C (the penalty term).
- The Gaussian process-based regression with the *fitrgp* command in MATLAB software combines the output of the binary classifiers (see Fig. 2). Parameter σ in the regression is considered within $\{10^{-5}, 10^{-4}, , 10^2, 10^3\}$.

Therefore, the one-leave-out speaker cross validation performs parameter tuning. For this purpose, one speaker is considered for the test and so the remaining speakers perform the feature selection between the two classes. Then, the training of the two-class classifiers takes

[3] https://audeering.com/technology/opensmile.

Table 2 Feature selection and classification in EMO-DB dataset, using OAO strategy (′ denotes minute and ″ denotes second)

Method	Efficiency ± Standard deviation	Feature selection time for each emotion pair	SVM classifier learning time	Number of features
Proposed Method	71.01 ± 10.28	2”	3’	505
Speaker-Normalization	69.12 ± 11.94	1’:40”	3’:30”	550
MI	66.85 ± 12.59	3’:13”	3’:30”	550
Spearman	63.01 ± 10.82	13’:22”	3’:30”	550
MTFS	61.59 ± 10.98	19′*		515
FFS + Tree	61.90 ± 12.76	18’:10”	4’	600
FFS	59.42 ± 11.44	21’:29	4’	600

*Since MTFS jointly performs feature selection and classifier training, the reported time is related to the sum of feature selection and classifier training times

place, followed by combining the results from classifiers and regression training. The testing process is performed with the speaker, who is specially selected for this task. This process repeats by the number of speakers and, in each repetition, one speaker is considered for the test.

It should be noted that parameter tuning for training classifiers and combiner regressors are considered the same by the proposed approach, as well as by the compared methods. Also, in the OAO strategy, the EMO-DB and ENTERFACE datasets contain 21 and 15 pairs of emotions, respectively; in the OAA strategy, they contain 7 and 6 pairs of emotion classes. All implementations are run on a PC with a Corei5 CPU with 8 GB RAM in the Windows operating system and MATLAB software, version 2016b.

5.3 Results

After the proposed objective function (4) performs feature selection, it is observed that the feature space between each pair of classes is not the same. For example, features obtained for sadness and disgust have no relation with the features for happiness and neutrality. Therefore, the current work examines different tasks to unify the feature space. First, the union of selected features from each pair of classes is calculated so that the obtained features for all pairs of classes are the same. Then, multi-class classifiers are employed. Despite this, all of the obtained results from each item are not necessarily acceptable.

Table 3 Feature selection and classification in ENTERFACE dataset, using OAO strategy (′ denotes minute and ″ denotes second)

Method	Efficiency ± Standard deviation	Feature selection time for each emotion pair	SVM classifier learning time	Number of features
Proposed Method	73.03 ± 07.56	5”	3’:13”	454
Speaker-Normalization	67.98 ± 10.94	2’:4”	3’:39”	550
MI	68.65 ± 12.07	5’:29”	3’:25”	500
Spearman	69.32 ± 08.76	14’:50”	3’:56”	600
MTFS	59.38 ± 08.31	21′*		530
FFS + Tree	65.38 ± 11.82	20’:58”	3’:39”	550
FFS	61.25 ± 12.40	24’:3”	3’:56”	600

*Since MTFS jointly performs feature selection and classifier training, the reported time is related to the sum of feature selection and classifier training times

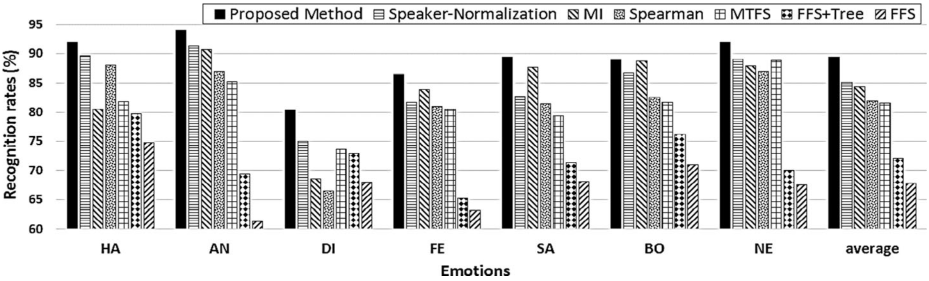

Fig. 5 The recognition rates of each emotion in OAO strategy for EMO-DB dataset

Simple combination methods, such as score-based and majority voting, are also examined and the results prove to be unacceptable. For this reason, the current study does not report these. Because the feature spaces between each pair of classes are different, such results demonstrate that an appropriate combination of classifier outputs should be utilized to achieve a multi-class classifier via binary classifiers. As explained in Section 3.2, the regression method is employed to combine classifiers and corresponding results are reported.

Feature selection and classification experiments are conducted in two ways. In the first feature selection experiment, the OAO strategy performs speaker-independent feature selection for each pair of emotions. In each step of this strategy, only one pair of emotion classes (two emotions) is considered for each speaker. In the second experiment, the OAA strategy selects between one emotion and other emotions. In this case, those features which discriminate one emotion from others are selected. In addition, the proposed and compared methods select a different number of features and so the numbers $\{100, 150, , 700\}$ are considered. Finally, those features with the highest efficiencies are reported.

In order to show that selected features are speaker-independent, the present study chooses one speaker for the test. While the data related to the test speaker are utilized for the test, the feature selection and training classifier procedures are performed with the data related to the rest of the speakers. For example, in the EMO-DB dataset, which consists of 10 speakers, nine are used for the training process; the experiments are repeated 10 times and, in each repetition, one speaker is chosen for the test. Finally, the average of these 10 results is reported. The ENTERFACE dataset, which contains 43 speakers, performs the same process.

Additionally, the OAO strategy calculates the intersection of the selected features among all the experiments. In fact, these features are common among all speakers and emotions. These

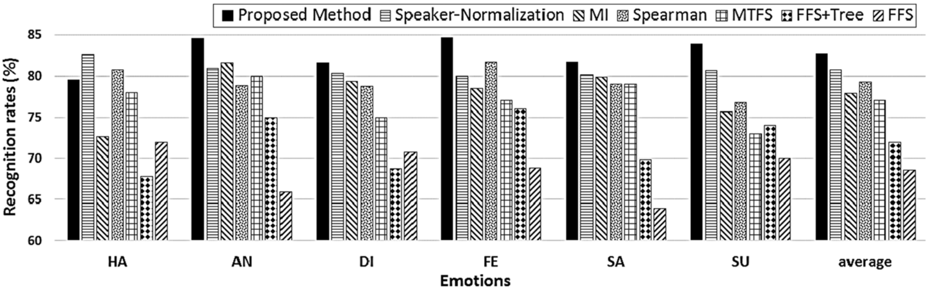

Fig. 6 The recognition rates of each emotion in OAO strategy for ENTERFACE dataset

Table 4 Feature selection and classification in EMO-DB dataset, using OAA strategy (′ denotes minute and ″ denotes second)

Method	Efficiency ± Standard deviation	Feature selection time for each emotion pair	SVM classifier learning time	Number of Features
Proposed Method	73.03 ± 07.04	7”	3’	530
Speaker-Normalization	70.129 ± 07.8	2’:7”	3’:30”	700
MI	68.45 ± 11.08	4’:50”	3’:30”	550
Spearman	65.89 ± 08.23	16’:12”	3’:30”	550
MTFS	67.91 ± 08.09	23′*		550
FFS + Tree	63.37 ± 07.35	20’:5”	4’	650
FFS	59.27 ± 08.27	22’:34”	4’	700

*Since MTFS jointly performs feature selection and classifier training, the reported time is related to the sum of feature selection and classifier training times

features are then employed to classify emotions and the results are reported. The following three sections report the results.

5.3.1 One-against-one strategy results

This experiment considers the data of one pair of emotions for each speaker. In this case, there are $\frac{K(K-1)}{2}$ pairs of emotions and classifiers and K is the number of classes. Tables 2 and 3 provide the average result of T reparations for both datasets (T is the number of speakers). The results reveal that the proposed method has a higher efficiency than the other methods. Moreover, the feature selection time, classifier learning time, and the number of selected features in the proposed method are lower than in the others. It should be noted that the best result is reported for each method.

In the EMO-DB dataset, the speaker-normalization method has the highest efficiency in comparison with the other methods and is closer to the result obtained by the proposed method. Since MTFS simultaneously performs feature selection and classifier training, it reports the sum of the feature selection time and classifier training time. However, the results obtained by this method are not desirable.

Since the number of speakers in the ENTERFACE is more than that of the EMO-DB dataset, the ENTERFACE dataset’s feature selection time and classifier training time are

Table 5 Feature selection and classification in ENTERFACE dataset, using OAA strategy (′ denotes minute and ″ denotes second)

Method	Efficiency ± Standard deviation	Feature selection time for each emotion pair	SVM classifier learning time	Number of features
Proposed Method	69.43 ± 08.85	15”	3’:13”	507
Speaker-Normalization	68.01 ± 10.18	3’:23”	3’:19”	700
MI	65.48 ± 11.53	7’:18”	3’:25”	550
Spearman	66.34 ± 09.12	17’:48”	3’:56”	600
MTFS	63.29 ± 09.50	27′*		543
FFS + Tree	63.34 ± 13.41	21’:5”	3’:39”	600
FFS	58.64 ± 14.07	25’:32”	3’:56”	650

*Since MTFS jointly performs feature selection and classifier training, the reported time is related to the sum of feature selection and classifier training times

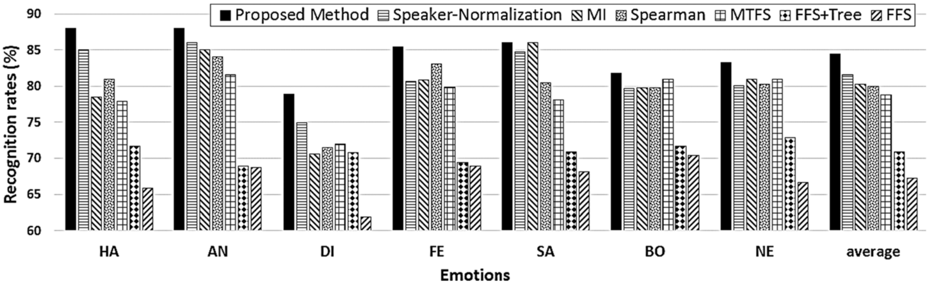

Fig. 7 The recognition rates of each emotion in OAA strategy for EMO-DB dataset

longer. In this dataset, the Spearman method's efficiency nears that of the proposed method, but Spearman's feature selection time is longer. The Spearman method spends much time calculating partial correlation coefficients during the feature selection step.

The recognition rates of each emotion are separately presented in Figs. 5 and 6. Figure 5 illustrates the EMO-DB dataset results. Based on Fig. 5, only the MI method's result for the emotion of boredom is close to that of the proposed method. Figure 6 provides results obtained by the ENTERFACE dataset. As seen, the Spearman and speaker-normalization methods outperform the proposed method only in the emotion of happiness. According to Figs. 5 and 6, the proposed method detects almost each emotion with the highest accuracy.

5.3.2 One-against-all strategy results

In this experiment, the positive class contains one emotion and the other class contains other emotions. In this case, the current work will have K pairs of classes and classifiers (K = the number of emotions). Tables 4 and 5 present the feature selection and classification results, which are related to an average of T repetitions (T = the number of speakers in each dataset). The results of the OAO strategy demonstrate that the proposed method has a higher efficiency than the other methods. Similar to the OAO experiment, the proposed method's feature selection time, classifier learning time, and the number of selected features in the OAA experiment are lower than in the other methods. Since MTFS simultaneously performs feature selection and classifier training, the sum of the feature selection time and classifier training time is reported. However, the results obtained by this method are not desirable.

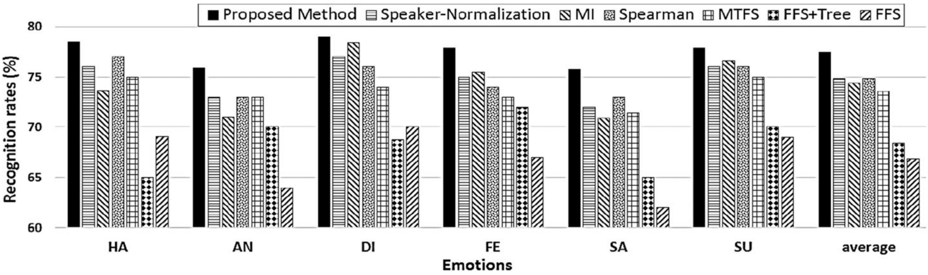

Fig. 8 The recognition rates of each emotion in OAA strategy for ENTERFACE dataset

Table 6 Results of common selected features using SVM with OAO strategy, for EMO-DB dataset (Results relate to 10 repetitions)

Method	Average percent of common features in all pairs of emotions	Number of common features	Efficiency ± Standard deviation	Efficiency ± Standard deviation (LDA)
Proposed Method	89.34%	387	73.65 ± 05.48	79.14 ± 4.30
Speaker-Normalization	81.00%	358	70.20 ± 08.30	72.10 ± 6.40
MI	65.45%	306	67.12 ± 08.20	69.70 ± 5.40
Spearman	69.50%	341	62.30 ± 07.30	65.30 ± 9.40
MTFS	69.68%	376	63.59 ± 09.45	65. 98 ± 5.60
FFS + Tree	43.00%	290	60.12 ± 10.39	62.15 ± 8.37
FFS	37.58%	281	59.10 ± 10.00	61.23 ± 9.23

Similar to OAO, the OAA experiment separately calculates the recognition rate for each emotion. Figures 7 and 8 provide the results for the EMO-DB and ENTERFACE datasets. The proposed method shows the highest efficiency for all emotions.

5.3.3 Results for sharing selected features

In previous experiments, one speaker is considered for the test and for each pair of emotions, feature selection and classifier learning were performed with the other speakers. In the case that each speaker is considered for a test in each repetition and the sharing of selected features in total repetitions is calculated, then sharing selected features will be common among all classes and all speakers.

The main aim, therefore, is investigating how much these common features describe different emotions when a new speaker is introduced. To this end, common features are considered and two-class SVM classifiers with the OAO strategy are trained. Tables 6 and 7 report the results. These tables indicate that the proposed method can more effectively select features related to all emotions than is possible by the other methods. In addition, comparing the results of Tables 6 and 7 with those of Tables 2 and 3 demonstrates that the efficiency of the proposed method rises as the number of selected features falls. Furthermore, the average percent of common features in all pairs of classes is reported, which indicates that the proposed method, in the absence of one speaker, will select almost the same features.

Table 7 Results of common selected features and using SVM with OAO strategy, for ENTERFACE dataset (Results relate to 43 repetitions)

Method	Average percent of common features in all pairs of emotions	Number of common features	Efficiency ± Standard deviation	Efficiency ± Standard deviation (LDA)
Proposed Method	82.10%	349	73.65 ± 05.48	80.63 ± 4.31
Speaker-Normalization	74.59%	319	67.00 ± 08.48	70.32 ± 7.35
MI	58.39%	287	66.84 ± 09.30	68.27 ± 8.43
Spearman	61.12%	304	68.92 ± 06.21	70.39 ± 6.40
MTFS	65.87%	321	60.69 ± 06.20	63.09 ± 5.00
FFS + Tree	58.12%	265	64.71 ± 10.30	66.31 ± 8.10
FFS	42.51%	221	59.39 ± 10.11	61.85 ± 8.13

Table 8 Results of common selected features with other classifiers, using common features for EMO-DB dataset (Results relate to 10 repetitions)

Method	Efficiency ± Standard deviation (ELM)	Efficiency ± Standard deviation (ELM_LDA)	Efficiency ± Standard deviation (Decision tree)	Efficiency ± Standard deviation (Decision Tree_LDA)
Proposed Method	69.27 ± 06.39	76.68 ± 5.87	67.23 ± 6.59	72.33 ± 5.67
Speaker-Normalization	67.87 ± 09.34	73.01 ± 6.56	65.17 ± 7.49	67.13 ± 6.45
MI	66.00 ± 09.01	69.12 ± 8.23	61.26 ± 7.45	65.34 ± 7.89
Spearman	65.23 ± 07.49	68.87 ± 6.65	63.14 ± 6.87	66.97 ± 6.45
MTFS	-	-	-	-
FFS + Tree	58.00 ± 09.34	61.56 ± 7.49	57.16 ± 8.40	60.29 ± 6.48
FFS	55.19 ± 09.00	58.63 ± 8.31	55.54 ± 9.65	58.45 ± 8.43

Moreover, with the application of the Linear Discriminant Analysis (LDA) method on the common feature space and the training of the SVM classifier, the efficiency of the proposed method appears to considerably increase. It should be noted that if LDA is first applied to all features and is followed by classifier training, the EMO-DB and ENTERFACE datasets achieve an efficiency of 55.67% and 51.34%, respectively. This reveals that better mapping to the new space and increased efficiency of the classifier occur when the number of features is lower and more appropriate features are available.

In addition to the SVM classifier, the remainder of the experiment employs the extreme learning machine (ELM) and decision tree classifiers for common features and reduced feature space. The sigmoid activation function, as well as some hidden layers within $\{10, 20, 100\}$, is utilized for the ELM classifier. Ten father nodes are also considered for the decision tree classifier. Tables 8 and 9 report the experiments' results. In comparison with the other methods, the efficiency of the proposed method is higher.

All experiments in this section demonstrate that, by considering common features between each pair of emotions, the proposed method can select the features most related to emotions. This will significantly improve the efficiency of each classifier. Additionally, when considering Tables 6, 7, 8 and 9 and comparing them with Tables 2, 3, 4 and 5, the present study observed that the standard deviations obtained from the different experiments significantly decreased. This drop is most significant in the proposed method, which points to its higher accuracy in selecting common features.

6 Conclusion

The results of the current work demonstrate that multi-task methods are suitable for the purpose of feature selection. The multi-task methods utilized in [50] and [51] simultaneously perform feature selection and classifier training. In these methods, the presented objective function contains a multi-task SVM objective function and an additional term for the classifier's sparsity of weight vector. One important issue is the rise in the number of data or features that significantly increases the runtime order of the SVM classifier solution. As a result, these methods are not suitable when there are a high number of features and the task is time-consuming. Also, although the features obtained for each pair of classes differ in [50] and [51], no proper combination is presented for them and they have low efficiencies. In contrast, the proposed method is performed in two phases. First, feature selection is performed followed by classification. Hence, in the proposed method, each of classifiers can be utilized. Also, the speed of feature selection in the proposed method is very high.

Moreover, in all the methods compared in this study, there are some tuning parameters. Consequently, to find the proper parameters for each method, trial and error must be employed. In this case, if a method is time-consuming, then its process will be tedious. However, with the proposed method, the speed is acceptable and suitable parameters can be found quickly. Consequently, designing multi-class classifiers takes little time. Finally, the proposed method presents a fast approach for selecting the speaker-independent features for each pair of classes. In addition, every classifier can be used in the training phase and these results can be efficiently combined. Furthermore, the experiments indicated that the obtained common features can improve the results of all experiments. In addition, the application of dimensionality reduction methods, such as LDA, after feature selection will considerably

Table 9 Results of common selected features with other classifiers, using common features for ENTERFACE dataset (Results relate to 43 repetitions)

Method	Efficiency ± Standard deviation (ELM)	Efficiency ± Standard deviation (ELM_LDA)	Efficiency ± Standard deviation (Decision tree)	Efficiency ± Standard deviation (Decision Tree_LDA)
Proposed Method	71.03 ± 6.00	78.62 ± 5.36	69.00 ± 6.12	75.13 ± 5.25
Speaker-Normalization	68.00 ± 8.64	71.11 ± 7.45	66.49 ± 7.20	70.23 ± 7.32
MI	65.76 ± 8.34	68.42 ± 6.75	65.16 ± 6.87	69.92 ± 6.34
Spearman	68.00 ± 7.48	70.13 ± 6.40	62.11 ± 6.01	66.45 ± 6.13
MTFS	-	-	-	-
FFS + Tree	64.71 ± 8.30	67.39 ± 8.11	60.18 ± 8.54	63.41 ± 7.73
FFS	59.39 ± 8.11	61.70 ± 9.61	58.23 ± 8.10	60.38 ± 7.31

improve efficiency, especially when the proper features are selected in the feature selection phase.

Compliance with ethical standards

Conflict of interest The authors have no potential conflict of interest.

References

1. Argyriou A, Evgeniou T, Pontil M (2007) Multi-task feature learning. In: Advances in neural information processing systems, pp 41–48
2. Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imaging Sci 2(1):183–202
3. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology, pp 1516–1520
4. Charoendee M, Suchato A, Punyabukkana P (2017) Speech emotion recognition using derived features from speech segment and kernel principal component analysis. In: Computer Science and Software Engineering (JCSSE), 2017 14th International Joint Conference on IEEE, pp 1–6
5. Chen L, Wu M, Zhou M, Liu Z, She J, Hirota K (2017) Dynamic emotion understanding in human-robot interaction based on two-layer fuzzy SVR-TS model. IEEE Trans Syst Man Cybern Syst 50(99):1–12
6. Dang T, Sethu V, Ambikairajah E (2016) Factor analysis based speaker normalisation for continuous emotion prediction. In: INTERSPEECH, pp 913–917
7. Demircan S, Kahramanli HJNC, Applications, (2018) Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. Neural Comput Appl 29(8):59–66
8. Dibeklioğlu H, Hammal Z, Cohn JF (2018) Dynamic multimodal measurement of depression severity using deep autoencoding. IEEE J Biomed Health Inf 22(2):525–536
9. Escalera S, Pujol O, Radeva P (2010) On the decoding process in ternary error-correcting output codes. IEEE Trans Pattern Anal Mach Intell 32(1):120–134
10. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. ACM, New York, pp 1459–1462
11. Farrús M, Ejarque P, Temko A, Hernando J (2007) Histogram equalization in svm multimodal person verification. In: International Conference on Biometrics. Springer, Berlin, pp 819–827
12. Fredrickson BL (2001) The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. Am Psychol 56(3):218
13. Fu J, Mao Q, Tu J, Zhan Y (2019) Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. Multimed Syst 25(5):451–461
14. Fürnkranz J (2002) Round robin classification. J Mach Learn Res 2(Mar):721–747
15. Gajsek R, Štruc V, Mihelič F (2010) Multi-modal emotion recognition using canonical correlations and acoustic features. In: 2010 20th International Conference on Pattern Recognition. IEEE, pp 4133–4136
16. Gao L, Qi L, Chen E, Guan L (2014) A fisher discriminant framework based on Kernel Entropy Component Analysis for feature extraction and emotion recognition. In: 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) IEEE, pp 1–6
17. Jin Y, Song P, Zheng W, Zhao L (2014) A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In: Acoustics, Speech and Signal Processing (ICASSP) (2014) IEEE International Conference on. IEEE, pp 4808–4812
18. Kaya H, Karpov AA (2018) Efficient and effective strategies for cross-corpus acoustic emotion recognition. Neurocomputing 275:1028–1034
19. Kaya H, Eyben F, Salah AA, Schuller B (2014) CCA based feature selection with application to continuous depression recognition from acoustic speech features. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, pp 3729–3733
20. Kok BE, Coffey KA, Cohn MA, Catalino LI, Vacharkulksemsuk T, Algoe SB, Brantley M, Fredrickson BL (2016) How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone: Corrigendum. Psychol Sci 27(6):931
21. Kotti M, Paternò F (2012) Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. Int J Speech Technol 15(2):131–150

22. Kotti M, Paterno F, Kotropoulos C (2010) Speaker-independent negative emotion recognition. In: 2010 2nd International Workshop on Cognitive Information Processing IEEE, pp 417–422
23. Liu J, Ji S, Ye J (2012) Multi-task feature learning via efficient l2, 1-norm minimization. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp 339–338
24. Liu Z-T, Xie Q, Wu M, Cao W-H, Mei Y, Mao J-W (2018) Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing 309:145–156
25. Liu Z-T, Wu M, Cao W-H, Mao J-W, Xu J-P, Tan G-Z (2018) Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 273:271–280
26. 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, IEEE, pp 17–20
27. Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface'05 audio-visual emotion database. In: Data Engineering Workshops (2006) Proceedings. 22nd International Conference on, IEEE, pp 8–8
28. Nemirovskii A, Nesterov Y (1994) Interior point polynomial algorithms in convex programming. SIAM 36(4):682–683
29. Nicolaou MA, Panagakis Y, Zafeiriou S, Pantic M (2014) Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, pp 1522–1526
30. Obozinski G, Taskar B, Jordan M (2006) Multi-task feature selection. Statistics Department, Berkeley UC, Tech Rep 2 (2.2):2
31. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
32. Rottenberg J (2017) Emotions in depression: What do we really know? Annu Rev Clin Psychol 13:241–263
33. Sarvestani RR, Boostani R (2017) FF-SKPCCA: Kernel probabilistic canonical correlation analysis. Appl Intell 46(2):438–454
34. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Trans Affect Comput 1(2):119–131
35. Shi C, Ruan Q, An G, Zhao R (2014) Hessian semi-supervised sparse feature selection based on L2, 1/2 -matrix norm. IEEE Trans Multimed 17(1):16–28
36. Shirani A, Nilchi ARN (2016) Speech emotion recognition based on SVM as both feature selector and classifier. Int J Image Graph Sig Process 8(4):39–45
37. Song X, Zhang J, Han Y, Jiang J (2016) Semi-supervised feature selection via hierarchical regression for web image classification. Multimed Syst 22(1):41–49
38. Tang J, Liu H (2012) Unsupervised feature selection for linked social media data. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, pp 904–912
39. Xie Z, Guan L (2013) Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis. Int J Semant Comput 7(01):25–42
40. Xu X, Huang C, Wu C, Zhao L (2016) Locally discriminant diffusion projection and its application in speech emotion recognition. Automatika 57(1):37–45
41. Yaacob S, Muthusamy H, Polat K (2015) Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals. PLoS One 10(3):1–20
42. Yang B, Lugger M (2010) Emotion recognition from speech signals using new harmony features. Signal Process 90(5):1415–1423
43. Yang N, Yuan J, Zhou Y, Demirkol I, Duan Z, Heinzelman W, Sturge-Apple M (2017) Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification. Int J Speech Technol 20(1):27–41
44. Yang X, Garcia KM, Jung Y, Whitlow CT, McRae K, Waugh CE (2018) vmPFC activation during a stressor predicts positive emotions during stress recovery. Soc Cognit Affect Neurosci 13(3):256–268
45. Yeh Y-c, Lai G-J, Lin CF, Lin C-W, Sun H-C (2015) How stress influences creativity in game-based situations: Analysis of stress hormones, negative emotions, and working memory. Comput Educ 81:143–153
46. Yogesh C, Hariharan M, Ngadiran R, Adom AH, Yaacob S, Polat K (2017) Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech. Appl Soft Comput 56:217–232
47. Yogesh C, Hariharan M, Yuvaraj R, Ngadiran R, Yaacob S, Polat K (2017) Bispectral features and mean shift clustering for stress and emotion recognition from natural speech. Comput Electr Eng 62(2):676–691
48. Yogesh C, Hariharan M, Ngadiran R, Adom AH, Yaacob S, Berkai C, Polat K (2017) A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. Expert Syst Appl 69(1):149–158
49. Zhang S, Zhao X, Lei B (2013) Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. Int J Adv Rob Syst 10(2):114

50. Zhang B, Provost EM, Essl G (2016) Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Piscataway, pp 5805–5809
51. Zhang B, Provost EM, Essl G (2017) Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. IEEE Trans Affect Comput 10(1):85–99
52. Zhou J, Chen J, Ye J (2011) Malsar: Multi-task learning via structural regularization. Arizona State University, Tempe, 21
53. Zou D, Wang J (2015) Speech recognition using locality preserving projection based on multi kernel learning supervision. In: 2015 International Symposium on Computers & Informatics, vol 2352-538X. Atlantis Press, Amsterdam, pp 1508–1516

Elham Kalhor received a B.S. degree in Computer Software Engineering from Sadjad University of Technology, Mashhad, Iran, in 2015 and an M.S. degree in Artificial Intelligence at Sadjad University of Technology, Mashhad, Iran, in 2019. Her research interests are Speech Emotion Recognition, Speech Processing, and Signal Processing.

Behzad Bakhtiari earned a B.S. degree in Computer Software Engineering from Islamic Azad University, Mashhad, Iran, in 2003, an M.S. degree in Artificial Intelligence from Amirkabir University of Technology, Tehran, Iran, in 2005, and a Ph.D. degree in Software Engineering at Ferdowsi University of Mashhad, Iran, in 2018. He has been a faculty member of the Computer Engineering Department at Sadjad University of Technology since 2008. His research interests are Machine Learning, Pattern Recognition, Speech Processing, Bayesian Filter, and Crowdsourcing.