Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 9, No. 3, 2021, 269-282.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Multi-Task Feature Selection for Speech Emotion Recognition: Common Speaker-Independent Features Among Emotions

Elham Kalhor and Behzad Bakhtiari *

Faculty of Computer Engineering and IT, Sadjad University of Technology, Mashhad, Iran.

Article Info

Article History:
Received 04 July 2020
Revised 24 October 2020
Accepted 31 March 2021

DOI:10.22044/jadm.2021.9800.2118

Keywords:
Speech Emotion Recognition, Multi-Task Feature Selection, Speaker Independent Features, Cross-Corpus Feature Selection, Affective Processing.

**Corresponding author: bakhtiari@sadjad.ac.ir (B. Bakhtiari).*

Abstract

Feature selection is one of the most important steps in designing the speech emotion recognition systems. Since there is uncertainty as to which speech feature is related to which emotion, many features must be taken into account, and for this purpose, identifying the most discriminative features is necessary. In the interest of selecting the appropriate emotion-related speech features, in the current work, we focus on a multi-task approach. For this reason, we consider each speaker as a task, and propose a multi-task objective function in order to select the features. As a result, the proposed method chooses one set of speaker-independent features, of which the selected features are discriminative in all the emotion classes. Correspondingly, the multi-class classifiers are utilized directly or the binary classifications simply perform the multi-class classifications. In addition, we employ two well-known datasets, Berlin and Enterface. The experiments are also applied on the openSmile toolkit in order to extract more than 6500 features. After the feature selection phase, the results obtained illustrate that the proposed method selects the features that are common in different runs. Also the runtime of the proposed method is the lowest in comparison to the other methods. Finally, seven classifiers are employed; the best achieved performance is 73.76% for the Berlin dataset and 72.17% for the Enterface dataset in the face of a new speaker. These experimental results then show that the proposed method is superior to the existing state-of-the-art methods.

1. Introduction

Emotion is a type of phenomenon that has a direct relationship with each person's mood. One of the prominent areas of speech processing is to design a system which can accurately and simultaneously analyze speech, and recognize the emotions it conveys. Emotion is a complicated and internal phenomenon that is not always explicitly indicated. In most cases, a person's emotion is a set of different senses [1, 2]. Consequently, recognizing the emotions is not straightforward, and even the humans make mistakes in recognizing them [3]. In the recent years, many research works have been conducted in the field of automatic speech emotion recognition. Although a significant progress has been made, the work in this field still encounters various challenges.

There are many applications for speech emotion recognition. One of the main applications for emotion recognition is the emotional relationship between the robots and the humans [4]. Moreover, in the call centers and mobile communication operations such as those for fire departments and other emergency services, the emotion recognition systems assist the public. These systems can also be employed in order to recognize the degree of customer satisfaction in the customer relations systems [4]. Computer games are another application of emotion recognition [5, 6]. Furthermore, since depression and other stress-

related diseases affect the speech, there are medical applications for the emotion recognition systems [7, 8].

There are almost 300 different emotions. Despite their large number, the emotions are categorized into eight groups, namely happiness, anger, disgust, boredom, sadness, fear, surprise, and neutral [9]. Thus for the automatic speech emotion recognition, a large number of acoustic features should be extracted. However, there is a great variety of speech features when an individual expresses different emotions, and so the researchers must deal with a huge number of features extracted from the speakers. On the other hand, due to these numerous speech features, it is not known which ones are related to each emotion. Therefore, it is vital to select a set of features that is speaker-independent and common among all the emotion classes. As a result, a set of selected features can describe each emotion for all speakers.

The low number of training samples and the high dimensionality of features are the major challenges in this field. These challenges cause over-fitting. Consequently, the training process is not performed correctly, and the system performance drops. Meanwhile, the time complexity of the system is critical. As the number of features is high, the main objective is to have a system that has a low runtime and selects an appropriate set of features [10].

Much effort has been made by the emotion recognition systems in order to select the proper features. The current paper focuses on a multi-task approach. There has been little research works employing a multi-task method for the speaker-independent feature sub-space learning [11, 12]. In these works, a multi-class classification is performed by a set of binary classifiers with the feature selection process performed separately for each pair of classes. Therefore, the features selected for different pairs of classes may differ, and so a different set of features must be dealt with. Unfortunately, the sets of selected features cannot be directly used for multi-class classifiers. Furthermore, if a set of binary classifiers is to perform multi-class classifications, it is necessary to have a proper combiner of classifiers in order to achieve a reasonable performance. In other words, since different classifiers' input spaces vary, simple classifiers' output fusion methods such as majority voting cannot achieve satisfactory results.

The aim of the current paper is to propose a method to find a set of discriminative speaker-independent features that are common among the emotion classes. For this purpose, the present work considers the multi-task objective function, which selects common features among all the speakers and all the emotion classes. In this case, since the selected features are speaker-independent and common among all emotions, the multi-class classifiers can be employed directly or multi-class classifications can be performed with binary classifiers without complex classifier combiners. Hence, the proposed method is superior in terms of time and simplicity.

The second section of this work presents the related work, and Section 3 explains the preliminary to the multi-task feature selection. Section 4 discusses the proposed method in detail, while Section 5 reviews and analyzes the experiments and results. Finally, Section 6 provides the conclusion.

2. Related Works

In the recent years, many research works in the field of speech emotion recognition has attempted to resolve the challenge of the high dimensionality of features in the speaker-independent and speaker-dependent categories. For this reason, many methods for feature selection and dimensionality reduction have been explored. Some approaches are speaker-dependent systems [13-21]. For instance, the feature selection methods employ correlation-based methods [22-26]. Some works employ evolutionary methods [27-29]. Moreover, some are based on clustering [30].

Since a speaker-dependent system is learned with a few number of speakers, it may not perform appropriately when dealing with a new speaker. Therefore, when there is a large number of speakers, it is extremely important for the system to be independent from the speaker, and perform at an acceptable level in a reasonable amount of time [9]. Several papers have investigated speaker-independent systems, and these works are reviewed as what follows.

Some methods are based on the Sequential Floating Forward Selection (SFFS) [9, 31]. SFFS is a sequential method that adds a new feature to the selected feature space in each sequence; then a union of all features is considered. [10] employs the cascaded normalization method, which normalizes the features related to each speaker in

three steps. As presented in [32], normalization is carried out in the first step. The second step performs normalization in order to prevent sparsity, which is stated as $f(x) = sign(x)|x|^{\alpha}$, where $0 \leq \alpha \leq 1$. In the last step, L_2-norm normalizes each feature vector x. Finally, these three normalization steps remove the redundant features. [33] considers the feature space as a network, and accounts for each feature as a node. Then the best-first algorithm is employed for the feature selection. In this algorithm, the greedy hill-climbing method explores the feature space, and the Pearson correlation function handles the hill-climbing evaluation function. In this case, the Pearson correlation evaluation function considers the correlation between features and class labels. Finally, the features with the highest correlation are selected, and the selected features from each group are then combined. [34] introduces the binary tree structure, which is generated based on the three dimensions of emotion: negative valence and non-negative valence, negative-activation and positive-activation, and lower-stance and higher-stance. These emotions are then divided based on the emotion dimensionality. At each level of the tree and for each emotion, the Forward Feature Selection (FFS) approach performs the feature selection. Finally, the features obtained from all emotions are combined with each other.

[35] first divides the data into n groups (the optimal value of n is obtained based on several tests). A Gaussian kernel with different values for sigma is applied to each group. Then in each group, the common features of the data are selected, and the union of all the features obtained from n groups is calculated. Unfortunately, as the number of features increases, the time complexity of this method grows significantly.

The Z-score method [36] normalizes the features of each speaker, and then the feature selection is performed [37]. The purpose of this normalization is to reduce the difference among the speakers' speech features. After normalization and the removal of some of the less-related features, the Mutual Information (MI) method is applied to the feature selection. In this method, a correlation between features and class labels selects the proper features. Utilizing the Euclidean distance [38], first divides the features into four groups. Then the partial correlation is calculated for the features of each group. After that, the Spearman correlation obtains the correlation among the groups. Finally, the common features among the four groups are determined, and then the Fisher method performs the dimensionality reduction. Unfortunately, this method spends much time computing the partial correlation coefficients, and is, therefore, time-consuming, especially when the number of features is high. Dang *et al*. have introduced the speaker-related factors for each speaker [39], employing the Probabilistic Linear Discriminant Analysis (PLDA) technique [40]. This technique utilizes the emotion factors related to each speaker, and provides information about the features at every frame. In each step, the information about the features of each speaker's whole frames is obtained. This method identifies the feature space obtained, which contains information about the emotion of all speakers.

The other one is the end-to-end speech emotion recognition [41-45], which employs the neural networks and performs deep learning to classify the emotion from the raw wav file without any traditional speech processing extracted features. Indeed, the weight of the neural networks extracts the features automatically. Although some recent research works have performed SER in this manner, they have some drawbacks. For example, since they use raw data, the size of sample vector is very large. Therefore, the time complexity in the train and test step is very large. Also, the efficiency of these methods is not investigated in some conditions like different speakers, different languages or generally cross-corpus. Furthermore, these types of methods require a lot of speech sample for training, in which thereby the training time dramatically rises.

Very little research works have employed multi-task learning for the speaker-independent feature selection [11, 12]. These methods are a kind of wrapper feature selection method, which simultaneously performs feature selection and classifier training. In addition, these methods are proposed for binary classifiers, and therefore, feature selection is separately performed for each pair of classes. As a result, binary classifiers design the multi-class speech emotion recognition problem. In other words, this approach considers a set of different pairs of classes and then performs feature selection and classifier training for each pair of classes. Consequently, the sets of selected features for different classes may differ.

For example, it is possible that the selected features for the neutral and happiness classes differ from the selected features for the anger and disgust classes. In this way, the input spaces of different classifiers vary.

Table 1. A summary of speaker-independent feature selection methods for speech emotion recognition.

Reference	Description	Number of used features	Number of selected features
[31]	SFFS-based	200	8
[9]	SFFS-based	306	40, 20
[33]	Best-first algorithm and greedy hill climbing	1418	About 55 features
[34]	Binary tree on the emotion dimension + FFS	2286	75
[35]	Gaussian kernel	988	65
[37]	Mutual information	121	{20, 40, 60, 80, 100}
[38]	Spearman + Fisher	34	20
[39]	Emotion factors + PLDA	650	50
[10]	Feature normalization in three steps	6373	Not mentioned
[11]	Multi-task system	1170	Not mentioned
[12]	Multi-task system	6669	Not Mentioned
[41-45]	Deep learning	-	-
	Proposed method	6552	About 500 features and also {10, 20,…, 90}

Therefore, a complex classifier's output fusion is required to combine the outputs of the classifiers because a simple method such as majority voting cannot achieve an acceptable performance. Also since the methods of [11, 12] are only designed for the SVM classifier, they are not applicable to other classifiers. Unfortunately, as the number of data or features increases, the SVM's time complexity dramatically rises. As a result, these methods are not appropriate for a feature selection with a large number of features.

The currently proposed method selects a set of features that are common among the speakers and emotions. Due to the high number of features, it is expected that some common features describe all the emotions among all the speakers. Finally, the present work has one sub-space, and as a result, can be directly employed by the multi-class classifiers or can solve multi-class classifications with a simple combiner of binary classifiers. Furthermore, in contrast to [11, 12], the proposed method can utilize any classifier.

It should be noted that none of the research works mentioned in this section have performed the proposed method for the feature selection. Table 1 provides a summary of the works on speaker-independent emotion recognition systems. Table 1 indicates that many methods consider fewer features in comparison to the number taken into account by the present work. The exceptions are [10, 12], which study almost the same number of features as does the proposed method. However, it can be observed that the proposed method runs at a lower speed than do both methods [10, 12].

3. Preliminary to Multi-Task Feature Selection

In the multi-task system, the tasks can include different items such as datasets and hand-writing from different users. With the multi-task learning feature selection, there is a sub-space that contains common features among all the tasks, and as a result, this improves the efficiency of the classifier [46, 47]. For this purpose, $L_{2,1}$-norm can be employed for the feature selection, in which the redundant features are removed. As a result, only the common features among all tasks are selected [48-50]. For this purpose, the general objective function can be considered as (1):

$$\min_{W} Loss(W, X, Y) + \theta \|W\|_{2,1} \tag{1}$$

which includes two terms. The first term is the smooth convex loss function, $Loss(W, X, Y)$, which can be the least square loss or logistic loss. The second term is $L_{2,1}$-norm, which is non-smooth, and can be calculated based on (2):

$$\|W\|_{2,1} = \sum_{i=1}^{d} \left(\sum_{j=1}^{T} (w(i,j))^2 \right)^{1/2} = \sum_{i=1}^{d} \|w^i\|_2 \tag{2}$$

where T is the number of tasks, d is the number of features, and $W \in \mathbb{R}^{d \times T}$ is the weight matrix in which each row relates to one feature, each column relates to one task, and w^i is the i-th row of W. In addition, θ is the regularization parameter that can control sparsity.

4. Proposed Method

In the current work, we assume that there are T speakers in its data set $\{X_s, y_s\}_{s=1}^{T}$, where $X_s \in \mathbb{R}^{n_s \times d}$ is the feature matrix of the s-th speaker for n_s samples, and d is the number of features. Also $y_s \in \mathbb{R}^{n_s \times 1}$ is its label, where $y_s(i) \in \{1,2,\dots,k\}$ and k is the number of emotion classes (here $y_s(i)$ is the i-th element of y_s).The aim of the current work is to obtain a sub-space for the speaker-independent features, which relates to all emotions.

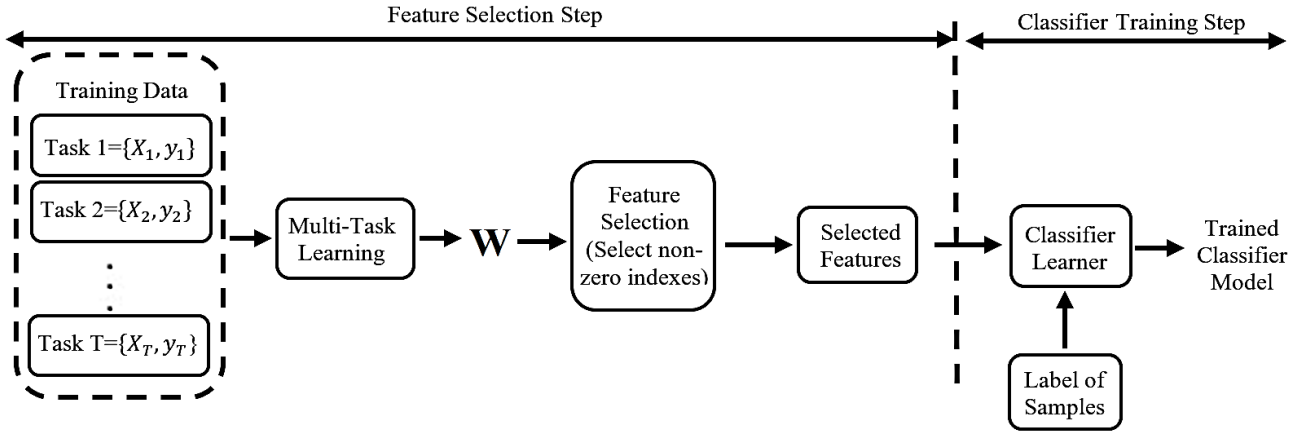

Figure 1. Feature selection and classifier training step.

Towards this aim, the proposed method considers the multi-task objective function in which each task is a speaker. In this case, the common features among the emotions and speakers are selected.

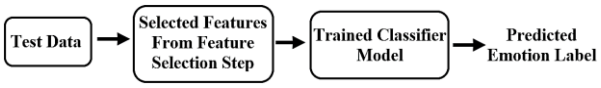

Figure 2. Testing step.

For this purpose, the objective function (3) is employed:

$$\min_{W} \frac{1}{2}\sum_{s=1}^{T}\left\|y_s - \frac{1}{2}X_s w_s\right\|_2^2 + \theta\|W\|_{2,1} \quad (3)$$

where $W \in \mathbb{R}^{d\times T}$ is the weight matrix, and $w_s \in \mathbb{R}^{d\times 1}$ is the weight vector of the s-th speaker and a column of W. In addition, θ is the regularization parameter that can control the sparsity of the weight matrix.

Figure 1 provides a block diagram of the proposed method and its two steps. The feature selection step separates the training data based on the speakers, and each speaker is considered as a task. After solving (3), W is obtained. Then the index of features with non-zero values is selected and stored. Also in the training phase, the multi-class classifiers are directly employed by the selected features or the binary classifiers are used in order to perform the multi-class classifications. Since this sub-space is the same for all the binary classifiers, it is quite simple to combine them and achieve a final label. In addition, it is obvious that after feature selection, any classifier can be applied. In the testing step, as shown in Figure 2, the result of the feature selection phase extracts the selected features from the test data. Then the features obtained are given to the trained classifier model and the label of the test data is predicted.

The first and second terms in (3) are the smooth and non-smooth, respectively. Since (3) contains a non-smooth term, there is no closed-form solution. Fortunately, several methods exist for solving the function with the smooth and non-smooth terms [51-53]. In the current work, we employ the algorithm proposed in [53]. Also the Microsoft website provides the online source code:(http://research.microsoft.com/apps/pubs/?id=264770).

5. Experiments

In this section, we explain the datasets, comparison methods, and experimental design. Finally, the results obtained are analyzed for each dataset.

5.1. Dataset

The first term of (3) is the least square regression, and is individually performed for each speaker's data. Also for each speaker, the aim of the first term of (3) is to predict the samples' labels.

In this case, for each speaker s, if the i-th feature is not appropriate and provides little useful information about labels, then the i-th element of w_s approaches zero. On the other hand, the second term of (3) is the regularization term, and considers the weight matrix of all the speakers. Thus as mentioned in Section 3, the common features among the speakers can be selected.

The present study's experiments employ the Berlin [54] and Enterface [55] datasets, about which Table 2 provides some brief information. The Berlin dataset contains seven emotions: happiness (HA), anger (AN), disgust (DI), boredom (BO), sadness (SA), fear (FE), and neutral (NE). The Enterface dataset consists of six emotions: HA, AN, SA, Fe, DI, and surprise (SU).

Table 2. Information datasets (S: speakers, F: Female, M: Male)

Database	Language	Size	#Classes	#S	#F	#M	Type
Berlin	German	535	7	10	5	5	Audio
Enterface	English	1287	6	43	9	34	Video

This dataset includes the video data, from which the voice is extracted and used. The current experiments utilize all the emotion classes in both datasets.

5.2. Comparison Method*s*

In the present work, we compare its method with those of some other methods related to the speaker-independent feature selection for speech emotion recognition. For this purpose, the following methods are used (and one can refer to Section 2 for more details):

Forward Feature Selection (FFS): a baseline feature selection method [9].

FFS+Binary Tree Structure (FFS+Tree): generates a binary tree structure based on the three dimensions of emotion, and then employs the FFS method [34].

FFS+Binary Tree Structure (FFS+Tree): generates a binary tree structure based on the three dimensions of emotion, and then employs the FFS method [34].

Mutual Information (MI): uses MI in order to compute the correlations between the features and labels [37].

Spearman: computes the correlations between two groups of features, based upon which the feature selection is then performed. Although dimensionality reduction is carried out by Fisher after the feature selection step in [38], for a fair comparison, in the present work, we only perform the feature selection step without the Fisher dimension reduction.

S-N (Speaker-normalization): performs three normalizations to select the speaker-independent features [10].

MTFS: a multi-task-based method [12] that separates the tasks based on singing, speech, and gender. In addition, the datasets are also considered as the tasks. However, in the current study's comparisons, the speaker-based tasks are considered.

5.3. Experimental Design

- Speaker-Independent Experiment Design

The aim of this experiment is to examine how many selected features are independent from the speaker's training data. For this reason, the classifier training step is performed after the feature selection step.

Feature Selection Step: In this step, one speaker is considered for the test, and so the other speakers are utilized for feature selection and the classifier training steps.

Classifier Training Step: In this step, a classifier is trained with the selected features, and so different classifiers are considered. Some are multi-class such as the Extreme Learning Machine (ELM), K-nearest Neighbors (KNN), Decision Tree, Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant classifiers. Furthermore, a two-class SVM is employed with two strategies: One-Against-One (OAO) and One-Against-All (OAA).

After the feature selection and classifier training steps, testing is performed with the speaker who plays no role in these two steps. This experiment is repeated by the number of speakers that undergo testing, i.e. one speaker from each run (the speaker who plays no role in either the feature selection or the classifier training steps).

- Cross-Corpus-Independent Experiment Design

This experiment attempts to answer the following question: how many selected features are appropriate for the unseen corpus data? For this reason, feature selection and classifier training steps are performed with a corpus. The testing step is then conducted with other corpus data.

5.4. Implementation Details

The OpenSMILE software [56] performs the feature extraction. With the '*emo_large*' config, 6,552 features are extracted from each voice file.

Table 3. Settings in the classifier training step.

Classification methods	Parameter settings
Support Vector Machine (SVM)	Kernel: Linear C parameter (Penalty term):16 values including $\{10^{-5}, 10^{-4}, \ldots, 3\times 10^{-2}, 3\times 10^{-3}\}$
Extreme Learning Machine (ELM) [57]	Activation function: Sigmoid, 10 values are considered for the number of neurons in the hidden layer including {10, 20, …, 100}
K-nearest neighbors (KNN)	Number of neighbors is considered in {2, 3, …, k}, (k indicates the number of emotion classes)
Decision tree	Number of father nodes = 10
Logistic Regression (LR)	Linear regression model
Linear Discriminant Analysis (LDA)	11 values are considered for regularization parameter including {0, 0.1, …, 1}
Quadratic Discriminant Analysis (QDA)	11 values are considered for regularization parameter including {0, 0.1, …, 1}

Also the following settings are considered for the parameters in all experiments:

1) All the experiments utilize seven classifiers. Table 3 provides the settings related to each classifier, and the Matlab software performs all the classifier implementations, except for ELM, which employs an implemented function.
2) For the parameter θ in (3), 50 values are considered in the range of $[10^7\ 10^{12}]$.
3) Different experiments are conducted so that the comparison methods can select different numbers of features, i.e. the number of selected features in the different experiments is equal to $\{100, 150, \dots, 600\}$ and $\{100, 200, \dots, 800\}$ for the speaker-independent and cross-corpus experiments, respectively.

It should be noted that the classifier training step is similarly performed for all the comparison and proposed methods. In other words, for the comparison methods and the proposed method, the same different parameters are examined for each classifier, and the best result is reported for each one. Also all the implementations are run on a PC with a Corei5 CPU and 8 GB RAM on the Windows operating system and with the Matlab software (version 2016b).

5.5. Speaker-independent Feature Selection Result

Feature selection step: According to the feature selection methods, there are two different ways to achieve the number of selected features. The first way provides the number of selected features to the feature selection method, and the second way obtains the number of selected features by changing some of the parameters. Based on this description, the proposed and the MTFS methods follow the second way and change some parameters in order to obtain the number of selected features. In contrast, the other methods are performed according to the first way. As a result, for each number of features determined by the first way described above, some of the parameters in the proposed and the MTFS methods change until the number of selected features approaches that of the first way. This process is performed for the purpose of a fair comparison. Thus as described above, the present work runs the second way of determining the number of selected features, and selects $\{100, 150, \dots, 600\}$ number of features. When the number of selected features rises, an increase in efficiency is observed, and the optimum number of features obtained is in the 500-600 range. As a result, with this second way of feature selection, the current work considers a number of selected features whose final performance in the classification turns out to be the best. The proposed and the MTFS methods also consider a regularization parameter in which the number of selected features achieved is in the 500-600 range. Tables 4 and 5 provide the results obtained from the feature selection step.

Meanwhile, since there are 10 speakers in the Berlin dataset, the experiments are repeated for 10 times. Also with 43 speakers in the Enterface dataset, the experiments are repeated for 43 times. In each run, one speaker is considered for the test, and the rest are used for the training step. Tables 4 and 5 present the mean feature selection time and the percentage of the similar selected features in the different runs. As it can be seen, the feature selection runtime in the proposed method is less than that in the other methods. Also the percentage of the commonly selected features implies that the proposed method has selected the same features in the all runs. Therefore, this indicates that the selected features are speaker-independent because in the case of one absent speaker, the same set of features is selected. Furthermore, the number of data in the Enterface dataset is larger than that of the Berlin dataset, and so more time is spent on the feature selection step. Moreover, in Tables 4 and 5, the results of the S-N method are close to those of the proposed method, i.e. S-N may also select the speaker-independent features as does the proposed method. Among the methods compared, the S-N method has the lowest runtime in the feature selection, and yet, it is higher than that of the proposed method. The other methods consume more time for feature selection. In addition, the low percentage of the commonly selected features demonstrates that the same features are not selected in different runs. Consequently, the selected features may not be speaker-independent.

Classifier training step: This step employs seven classifiers. Since the output of each classifier is dependent on the classifier setting, the best result is considered. Similar to the previous section, the experiments for each classifier are repeated for 10 times because there are 10 speakers in the Berlin dataset. In each run, one speaker is designated for the test, with the remaining nine speakers participating in the feature selection and classifier

learning steps.

Table 4. Feature selection runtime and percentage of the commonly selected features for the Berlin Dataset (′ indicates minutes and ″ indicates seconds). The results for 10 runs are reported.

Method	Mean of feature selection time	Percentage of common features in 10 runs	Number of features
Proposed method	30″	99.64%	535
FFS	55′	08.50%	600
FFS+Tree	25′	22.50%	600
MI	7′	69.50%	600
Spearman	18′	48.78%	600
S-N	2′:45″	99.50%	600
MTFS	21′*	53.00%	515

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

Table 5. Feature selection runtime and percentage of commonly selected features for the Enterface dataset (′ indicates minutes and ″ indicates seconds). The results for 43 runs are reported.

Method	Mean of feature selection time	Percentage of common features in 43 runs	Number of features
Proposed method	40″	99.27%	550
FFS	65′	10.50%	600
FFS+Tree	34′	24.66%	600
MI	9′	83.50%	600
Spearman	23′	61.00%	600
S-N	3′	99.42%	600
MTFS	30′*	59.33%	543

Table 6. SVM classifier results for the Berlin dataset (′ indicates minute and ″ indicates seconds). The results are reported for 10 runs. (Std: Standard Deviation).

Method	Efficiency±Std (OAO)	Efficiency±Std (OAA)	Mean of classifier learning runtime	Number of selected features
Proposed method	**72.57±6.45**	**73.76±6.45**	45″	535
FFS	60.08±9.70	63.09±9.23	55″	600
FFS+Tree	65.04±8.03	64.87±8.54	55″	600
MI	67.95±08.0	67.54±7.89	55″	600
Spearman	68.43±6.50	69.43±6.34	55″	600
S-N	65.54±7.02	66.19±7.39	55″	600
MTFS	61.59±7.98	67.91±8.09	21′*	515

Table 7. SVM classifier results for the Enterface dataset (′ indicates minute and ″ indicates seconds). The results are reported for 43 runs. (Std: Standard Deviation).

Method	Efficiency±Std (OAO)	Efficiency±Std (OAA)	Mean of classifier learning runtime	Number of selected features
Proposed method	**70.35±07.33**	**72.17±07.01**	48″	550
FFS	58.90±11.68	61.35±10.14	54″	600
FFS+Tree	60.43±08.12	63.87±08.43	54″	600
MI	67.54±08.48	69.00±08.19	54″	600
Spearman	65.09±09.12	67.98±09.31	54″	600
S-N	67.21±09.23	68.15±09.65	54″	600
MTFS	59.38±08.31	63.29±09.50	30′*	543

Finally, the mean result of 10 runs is reported. A similar task is performed for the Enterface dataset with its 43 repetitions.

Tables 6 and 7 illustrate the SVM classifier results in both strategies. These results obtained demonstrate that the proposed method outperforms the other methods in the two strategies. Moreover, in the proposed method, the classifier training time, classification error, and standard deviation of the results are almost all less than those of the other methods.

In the Berlin dataset, the results of the Spearman method in both strategies are also closer to those of the proposed method.

However, feature selection is very time-consuming in the Spearman method since the majority of time is spent in calculating the partial correlation coefficients. It is of note that the MTFS method performs feature selection for every pair of classes and the multi-class classification by majority voting on the outputs of the two-class SVMs. Therefore, the MTFS method's performance significantly decreases, especially with the OAO strategy.

Tables 8 and 9 provide the results of the multi-class classifiers in both datasets. According to these tables, among all the classifiers, the proposed method has the highest efficiency when compared with the other methods. Among all the classifiers, the performance of the S-N method is close to that of the proposed method. Also both the FFS and FFS+Tree methods perform at a much lower level than does the proposed method.

Another experiment is conducted for selecting less than 100 features. In this experiment, the number of selected features in each comparison method is accounted for in $\{10, 20, \ldots, 90\}$. Based on these numbers, the number of selected features in the proposed method is considered as it approaches these numbers. As a result, the proposed method and the other methods' number of selected features are almost the same. For this experiment, the feature selection times are similar, as shown in Tables 4 and 5. Since the SVM classifier proved to have the highest efficiency in the previous experiments, only the SVM classifier was tested with the two strategies in this part: one-against-one and one-against-all. Figures 3 and 4 present the results obtained from the two datasets. In these datasets, the proposed method in both strategies shows the highest efficiency. The S-N method's performance comes close to that of the proposed method, while the other methods show a lower

efficiency than the proposed method. The experiment's results reveal that even when selecting less than 100 features, the proposed method chooses the speaker-independent and emotion-related features. The feature selection time of the proposed method is much less than that of the other methods.

5.6. Cross-corpus Feature Selection Experiment Results

This experiment considers one dataset for training and another dataset for the test. Also, the SVM classifier was utilized with both strategies since it had the highest efficiency in the previous experiments. For the comparison methods, the feature selection is performed with different numbers of features: $\{100, 200, \ldots, 800\}$. For this reason, the experiment considers those parameters of the proposed method that have the most similar number of features selected. Then, the SVM classifier in both strategies is trained. Figures 5 and 6 illustrate the results obtained. The proposed method has the highest recognition rates in both strategies, and also its performance slope rises higher than that of the other methods.

Table 8. Results of classifiers for the Berlin dataset (′ indicates minutes and ″ indicates seconds). The results are reported for 10 runs. (Std: Standard Deviation).

Classifier	Method	Efficiency ± Std	Mean of classifier learning time	Number of features	Classifier	Efficiency ± Std	Mean of classifier Learning time	Number of features
ELM	Proposed method	**71.67±10.50**	1′	525	KNN	**68.81±12.56**	32″	578
	FFS	59.98±15.98	1′:20″	600		54.47±13.56	40″	600
	FFS+Tree	67.54±13.00	1′:20″	600		57.12±13.40	40″	600
	MI	68.45±08.51	1′:10″	550		61.00±11.45	40″	600
	Spearman	67.00±10.76	1′:10″	600		65.54±13.00	40″	600
	S-N	69.17±10.66	1′:20″	600		65.10±13.65	40″	600
Decision tree	Proposed method	**69.23±10.20**	30″	560	LR	**71.31±9.78**	40″	569
	FFS	56.98±13.60	45″	600		55.00±12.8	45″	600
	FFS+Tree	59.90±12.40	45″	600		61.00±11.4	45″	600
	MI	65.30±10.90	32″	550		68.30±11.0	45″	600
	Spearman	65.98±11.40	1′	600		66.49±10.6	45″	600
	S-N	67.10±09.43	1′	600		69.50±10.0	45″	600
LDA	Proposed method	**68.78±09.45**	54″	533	QDA	**69.87±11.0**	49″	570
	FFS	53.45±12.00	1′	600		52.12±13.0	57″	600
	FFS+Tree	55.00±12.56	58″	550		56.00±11.6	57″	600
	MI	65.49±10.00	1′	600		64.00±11.0	57″	600
	Spearman	62.87±10.40	58″	600		65.33±9.43	57″	600
	S-N	64.89±10.67	1′	600		68.94±11.4	57″	600

Table 9. Results of classifiers for the Enterface dataset (′ indicates minutes and ″ indicates seconds). The results are reported for 43 runs. (Std: Standard Deviation).

Classifier	Method	Efficiency ± Std	Mean of classifier learning time	Number of features	Classifier	Efficiency ± Std	Mean of classifier Learning time	Number of features
ELM	Proposed method	**71.12±10.09**	1′:5″	580	KNN	**64.89±10.34**	50″	600
	FFS	55.43±12.56	1′:15″	600		51.11±14.10	50″	600
	FFS+Tree	65.29±11.00	1′:15″	600		54.10±12.40	50″	600
	MI	66.76±**0**6.45	1′:15″	600		59.39±10.65	40″	550
	Spearman	68.07±10.40	1′	600		62.65±10.30	50″	600
	S-N	68.23±09.54	1′:15″	600		61.70±10.87	50″	600
Decision tree	Proposed Method	**68.93±10.67**	42″	530	LR	**70.02±08.12**	45″	570
	FFS	53.65±13.00	48″	600		58.51±12.85	49″	600
	FFS+Tree	58.09±12.40	48″	600		61.81±11.90	49″	600
	MI	59.39±10.65	48″	600		64.03±10.01	49″	600
	Spearman	65.76±10.40	45″	600		67.70±10.40	49″	600
	S-N	64.02±11.65	48″	600		67.64±11.00	49″	600
LDA	Proposed method	**70.98±08.01**	1′	560	QDA	**69.50±09.50**	56″	540
	FFS	55.2**0**±13.80	1′:5″	600		57.45±13.50	1′:3″	600
	FFS+Tree	59.40±12.00	1′:5″	600		60.34±11.90	1′:3″	600
	MI	65.32±09.56	1′:5″	600		67.00±10.65	1′	550
	Spearman	67.12±09.10	1′:5″	600		66.54±10.30	1′	600
	S-N	68.56±12.00	1′:5″	600		67.32±11.65	1′:3″	600

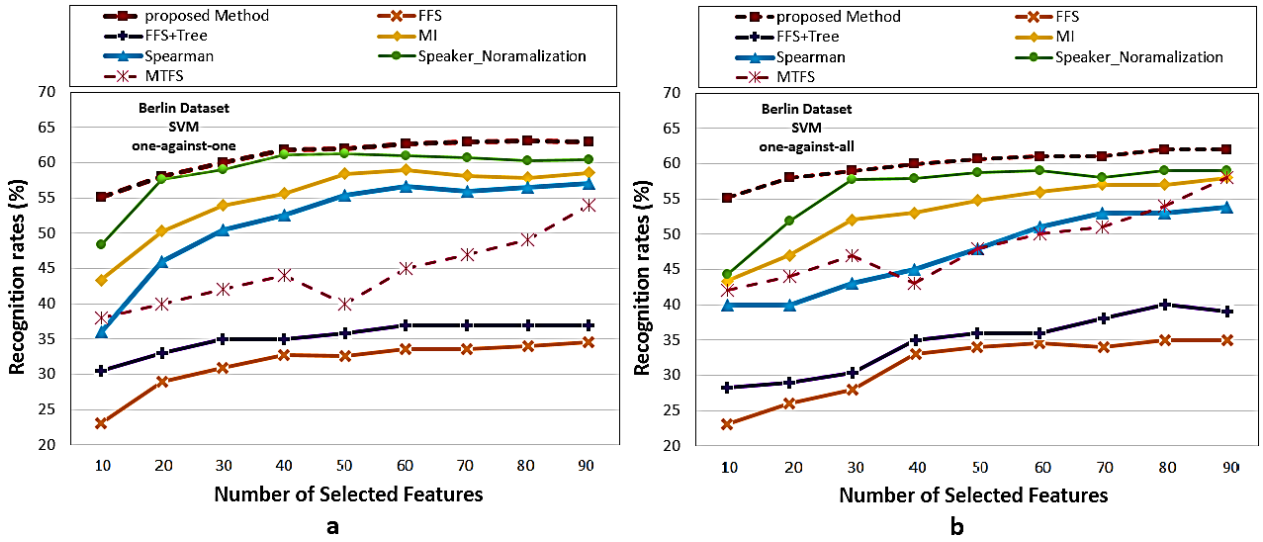

Figure 3. Selecting 10 to 90 features for Berlin dataset. a) SVM with the one-against-one strategy; b) SVM with the one-against-all strategy. (Speaker-Normalization: S-N).

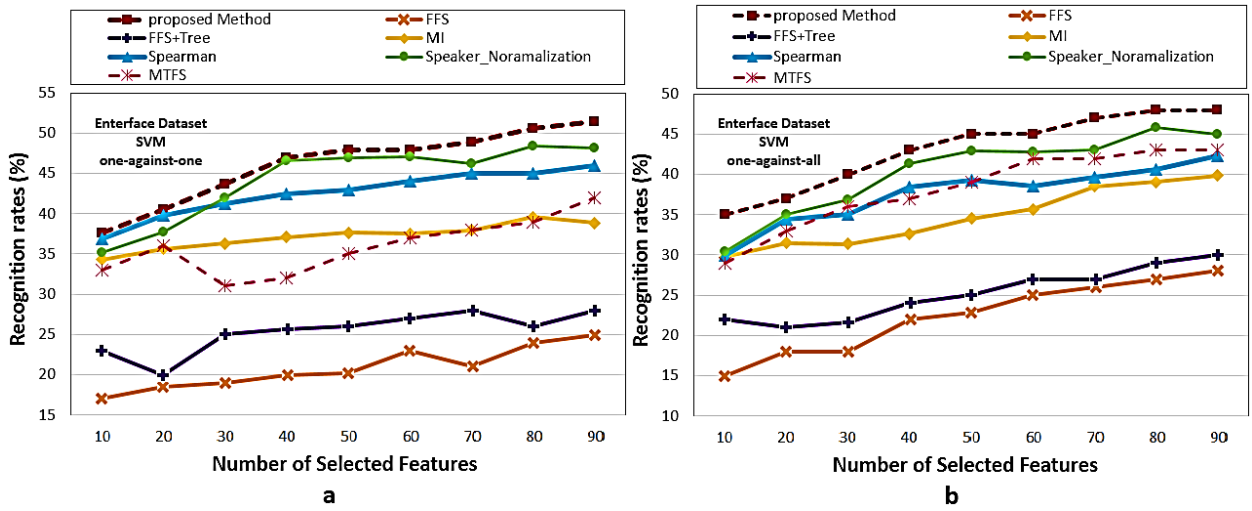

Figure 4. Selecting 10 to 90 features for Enterface dataset. a) SVM with the one-against-one strategy; b) SVM with the one-against-all strategy. (Speaker-Normalization: S-N).

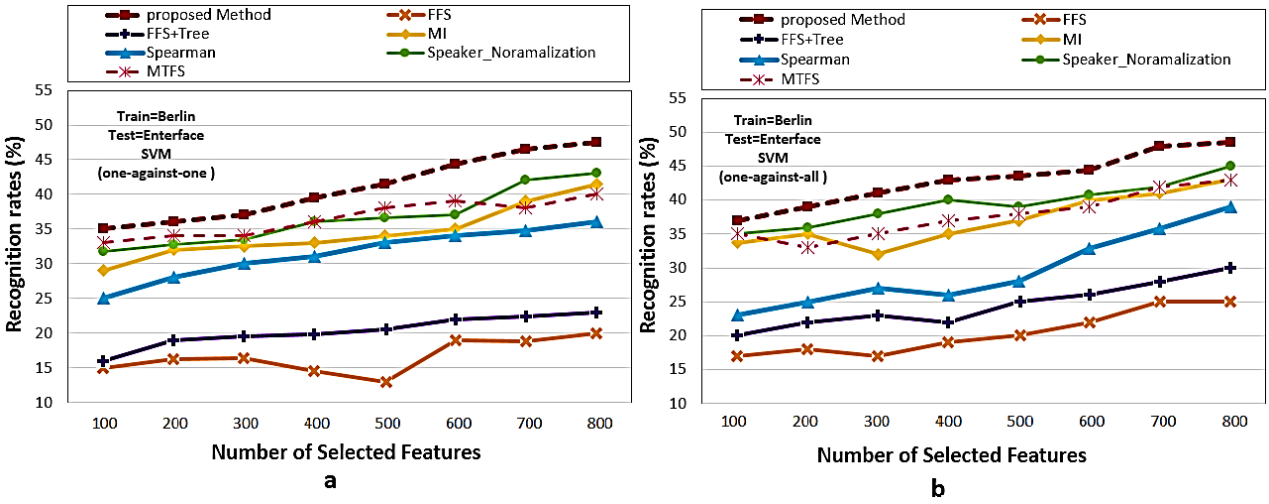

Figure 5. Selecting 100 to 800 features (Berlin dataset for training and Enterface dataset for testing). a) SVM with the one-against-one strategy; b) SVM with the one-against-all strategy. (Speaker-Normalization: S-N).

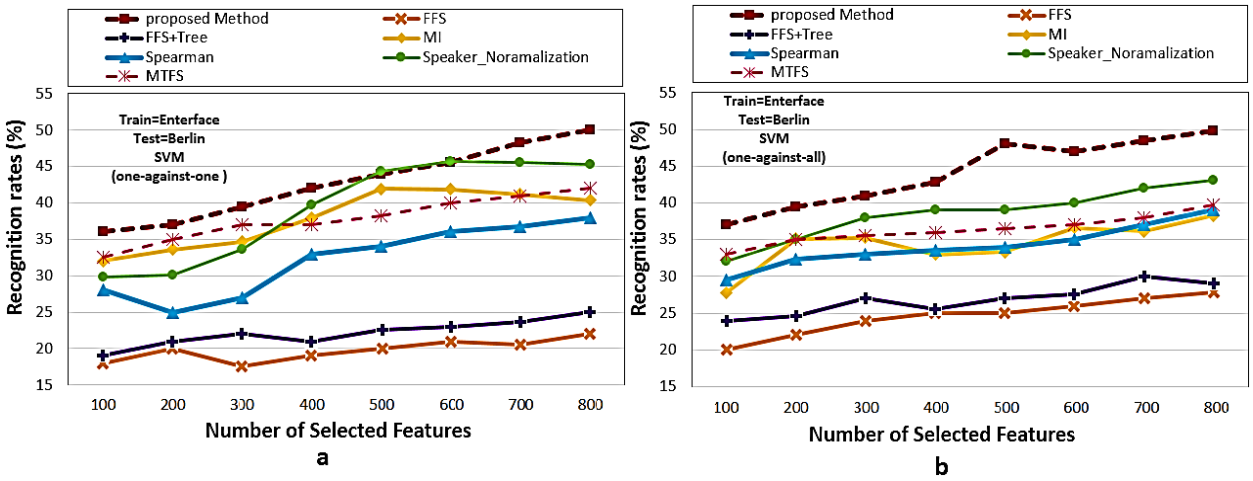

Figure 6. Selecting 100 to 800 features (Enterface dataset for training and Berlin dataset for testing). a) SVM with the one-against-one strategy; b) SVM with the one-against-all strategy. (Speaker-Normalization: S-N).

6. Conclusions

The result of the current work and the previous works [11, 12] indicate that the multi-task approach is the most appropriate for feature selection. However, since [11, 12] fall into the category of the wrapper feature selection method, they pose some drawbacks. For instance, these previous works performed a multi-class problem by combining the output of two-class SVMs. Therefore, they considered all pairs of emotion classes and, for each one, jointly performed feature selection and SVM training.

Unfortunately, the selected features of different pairs of classes may differ since the feature selection is individually performed for each pair of classes. Consequently, the multi-class classifier cannot be used, and it is necessary to design a complex combining method to fuse the output of the binary classifiers. Furthermore, since the SVM training time increases significantly by expanding the input dimensionality, the time for feature selection and SVM training in [11, 12] is very long for the high dimensional problems (such as for more than 6,500 features as examined in the current work's experiments).

In comparison to [11, 12], the proposed method's runtime is very low in both the feature selection and classifier training steps because the feature selection runtime is low and the classifier training is performed by a smaller number of selected features. Also the proposed method achieves one set of speaker-independent features that are common among all emotion classes. As a result, the multi-class classifier can be directly employed.

Finally, the experimental results prove the superiority of the proposed method over the other methods, especially in the case of time and the selection of appropriate speaker-independent features.

Moreover, for future work, the current authors shall attempt to answer the question: "*How can speaker-independent features be selected according to their source?*" Each feature is known to have a special source. For example, some feature sources are prosody or acoustic in speech signal. As a result, in addition to considering different speakers (multi-tasks), future research works should take into account the sources of features (multi-view). In this case, the groups of same features are considered in which the features in a group have the same effect in the classification process [58].

References

[1] France, D.J. et al.(2000). "Acoustical properties of speech as indicators of depression and suicidal risk". *IEEE transactions on Biomedical Engineering*. vol. 47, no. 7, pp. 829-837.

[2] Chenchah, F. and Z. Lachiri.(2017). "A bio-inspired emotion recognition system under real-life conditions". Applied Acoustics. 115: pp. 6-14.

[3] Harimi, A., Shahzadi, A., Ahmadyfard, A., & Yaghmaie, K. (2014). "Classification of emotional speech using spectral pattern features*". Journal of AI and Data Mining*, vol. 2, no. 1, pp. 53-61.

[4] Yuanchao, L. et al. (2017). “Emotion Recognition by Combining Prosody with Text Information and Assessment Selection for Human-Robot Interaction. SIG-SLUD”. 5(03): pp. 43-48.

[5] Bahreini, K., R. Nadolski, and W. Westera. (2014). "Improved multimodal emotion recognition for better game-based learning". in *International Conference on Games and Learning Alliance*. Springer.

[6] Poria, S. et al. (2017). "A review of affective computing: From unimodal analysis to multi-modal fusion". Information Fusion. 37: pp. 98-125.

[7] Yogesh, C. et al. (2017). "A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signa"l. Expert Systems with Applications. 69: pp. 149-158.

[8] Cummins, N. et al. "Enhancing Speech-based Depression Detection through Gender Dependent Vowel-Level Formant Features". in *Conference on Artificial Intelligence in Medicine in Europe*. Springer(2017).

[9] Yang, B. and M. Lugger. (2010). "Emotion recognition from speech signals using new harmony features". *signal processing*. vol. 90, no. 5, pp. 1415-1423.

[10] Kaya, H. and A.A. Karpov. (2018). "Efficient and effective strategies for cross-corpus acoustic emotion recognition”. Neurocomputing. 275: pp. 1028-1034.

[11] Zhang, B., E.M. Provost, and G. Essl. “Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach”. in ICASSP. , vol. 10, pp. 564-579, June 2016.

[12] Zhang, B., E.M. Provost, and G. Essl, (2017). “Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences”. *IEEE Transactions on Affective Computing*,(1): pp. 1-1.

[13] Zou, D. and J. Wang. (2015). “Speech Recognition Using Locality Preserving Projection Based on Multi Kernel Learning Supervision”. , vol. 10, pp. 20-27.

[14] Xu, X. *et al*. (2016). “Locally Discriminant Diffusion Projection and its Application in Speech Emotion Recognition. Automatika”. vol. 57, no. 1, pp. 37-45.

[15] Zhang, S., X. Zhao, and B. Lei. (2013). “Speech emotion recognition using an enhanced kernel isomap for human-robot interaction”. *International Journal of Advanced Robotic Systems*. vol. 10, no. 2, pp. 114.

[16] Charoendee, M., A. Suchato, and P. Punyabukkana. (2017). “Speech emotion recognition using derived features from speech segment and kernel principal component analysis”. in *Computer Science and Software Engineering (JCSSE)*, 14th International Joint Conference on. IEEE. , vol. 10, pp. 10-6, June 2017.

[17] Xie, Z. and L. Guan. (2013). “Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis”. *International Journal of Semantic Computing*. 7(01): pp. 25-42.

[18] Gao, L. et al. (2014). “A fisher discriminant framework based on Kernel Entropy Component Analysis for feature extraction and emotion recognition”. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE. , vol. 5, pp. 17-27.

[19] Özseven, T. (2019). “A novel feature selection method for speech emotion recognition. Applied Acoustics”. 146: pp. 320-326.

[20] Özseven, T. (2018). “Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition”. Applied Acoustics. 142: pp. 70-77.

[21] Peng, Z. et al. (2017). “Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on Gammatone Auditory Filterbank”. in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, vol. 9, pp. 98-108.

[22] Nicolaou, M.A. *et al*. (2014). “Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest”. in *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on. IEEE.

[23] Fu, J. et al. (2017). “Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. Multimedia Systems”. vol. 31, no. 2, pp. 1-11.

[24] Sarvestani, R.R. and R. Boostani,. (2017). “FF-SKPCCA: Kernel probabilistic canonical correlation analysis”. Applied Intelligence. vol. 46, no. 2, pp. 438-454.

[25] Štruc, V. and F. Mihelic. (2010). “Multi-modal emotion recognition using canonical correlations and acoustic features”. in *Pattern Recognition (ICPR)*, 2010 20th International Conference on. IEEE.

[26] Kaya, H. et al. (2014). CCA-based feature selection with application to continuous depression recognition from acoustic speech features. in Proceedings 39th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014, Florence, Italy.

[27] Yogesh, C. et al. (2017). Bispectral features and mean shift clustering for stress and emotion recognition from natural speech. *Computers & Electrical Engineering*. 62: pp. 676-691.

[28] Yogesh, C. et al. (2017). “Hybrid BBO_PSO and higher order spectral features for emotion and stress

recognition from natural speech". Applied Soft Computing. 56: pp. 217-232.

[29] Yaacob, S., H. Muthusamy, and K. Polat. (2015). "Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals". *Applied Soft Computing*. 58: pp. 287-295.

[30] Sun, Y. and G. Wen. (2015). "Emotion recognition using semi-supervised feature selection with speaker normalization*". International Journal of Speech Technology*. vol. 18, no. 3, pp. 317-331.

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

[32] Schuller, B. *et al*. (2010). "Cross-corpus acoustic emotion recognition: Variances and strategies". IEEE Transactions on Affective Computing. vol. 1, no. 2, pp. 119-131.

[33] Kotti, M., F. Paterno, and C. Kotropoulos. (2010). "Speaker-independent negative emotion recognition". in Cognitive Information Processing (CIP), 2010 2nd International Workshop on. IEEE.

[34] Kotti, M. and F. Paternò. (2012). "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema". *International journal of speech technology*. vol. 15, no. 2, pp. 131-150.

[35] Jin, Y. *et al*. (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.

[36] Farrús, M. et al. (2007)." Histogram equalization in svm multimodal person verification. in International Conference on Biometrics". *Springer*, vol. 10, no. 7, pp. 411-426.

[37] Jiang, X. et al. (2017). "Emotion Recognition from Noisy Mandarin Speech Preprocessed by Compressed Sensing". in International Conference on Intelligent Computing. Springer. vol. 14, no. 7, pp. 100-119

[38] Liu, Z.-T. et al. (2018). "Speech emotion recognition based on feature selection and extreme learning machine decision tree". *Neurocomputing*. 273: pp. 271-280.

[39] Dang, T., V. Sethu, and E. Ambikairajah. (2016). Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction. in INTERSPEECH.

[40] Prince, S.J. and J.H. Elder. (2007). Probabilistic linear discriminant analysis for inferences about identity. in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE.

[41] Zhang, Z., B. Wu, and B.r. Schuller. (2019). "Attention-augmented end-to-end multi-task learning for emotion prediction from speech". in ICASSP 2019-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

[42] Li, Y., T. Zhao, and T. Kawahara. (2019). "Improved End-to-End Speech Emotion Recognition using Self-Attention Mechanism and Multi-task Learning". in Interspeech. vol. 19, no. 3, pp. 317-322.

[43] Yao, Z. et al. (2020). "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN". *Speech Communication*. vol. 10, no. 4, pp. 202-215.

[44] Wang, C. et al. (2019). "Multi-Task Learning of Emotion Recognition and Facial Action Unit Detection with Adaptively Weights Sharing Network". in 2019 IEEE International Conference on Image Processing (ICIP). IEEE. vol. 10, no. 2, pp. 317-331.

[45] Atmaja, B.T. and M. Akagi. (2020). "Dimensional speech emotion recognition from speech features and word embeddings by using multi-task learning". APSIPA Transactions on Signal and Information Processing. vol. 15, no. 2, pp. 10-19.

[46] Obozinski, G., B. Taskar, and M. Jordan. (2006). "Multi-task feature selection". Statistics Department, UC Berkeley, Tech. Rep. 2, vol. 11, no. 3, pp. 12-24.

[47] Liu, J., S. Ji, and J. Ye. (2009). "Multi-task feature learning via efficient l 2, 1-norm minimization". in Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. 2009. AUAI Press.

[48] Sun, L. et al. (2009). "Efficient recovery of jointly sparse vectors". in Advances in Neural Information Processing Systems. 19(5): p. 110-118.

[49] Nie, F. et al. "Efficient and robust feature selection via joint ℓ2, 1-norms minimization". in Advances in neural information processing systems. 2010, vol. 5, no. 1, pp. 14-26.

[50] Tang, J. and H. Liu. (2012). "Unsupervised feature selection for linked social media data". in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, vol. 8, no. 7, pp. 65-71.

[51] Argyriou, A., T. (2007). Evgeniou, and M. Pontil. "Multi-task feature learning". in Advances in neural information processing systems, vol. 8, no. 9, pp. 254-268.

[52] Nemirovski, A. (2004)." Interior point polynomial time methods in convex programming". *Lecture notes*. vol. 5, no. 3, pp. 542-553.

[53] Beck, A. and M. Teboulle. (2009)." A fast iterative shrinkage-thresholding algorithm for linear inverse problems". *SIAM journal on imaging sciences*. 2(1): p. 183-202.

[54] Burkhardt, F. et al. (2005). "A database of German emotional speech". in *Ninth European Conference on*

Speech Communication and Technology. vol. 3, no. 4, pp. 59-64.

[55] Martin, O. et al. (2006). "The enterface'05 audio-visual emotion database". in Data Engineering Workshops, 2006. Proceedings. *22nd International Conference on. IEEE*. vol. 4, no. 1, pp. 68-75.

[56] Eyben, F., M. Wöllmer, and B. Schuller. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor". in Proceedings of the 18th ACM international conference on Multimedia.

[57] Huang, G.-B. et al. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. vol. 42, no. 2, pp. 513-529.

[58] Zhang, R. et al. (2019). "Feature selection with multi-view data: A survey". *Information Fusion*. 50: pp. 158-167.

انتخاب ویژگی چند وظیفه‌ای برای شناسایی احساس گفتار: ویژگی‌های مستقل از گوینده در بین تمام احساسات

الهام کلهر و بهزاد بختیاری*

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد، ایران.

ارسال ۲۰۲۰/۰۷/۰۴؛ بازنگری ۲۰۲۰/۱۰/۲۴؛ پذیرش ۲۰۲۱/۰۳/۳۱

چکیده:

انتخاب ویژگی‌های مناسب یکی از مهم‌ترین مراحل سیستم‌های شناسایی احساس فرد از روی گفتار وی می‌باشد. لذا از آنجایی که معلوم نیست کدام ویژگی از گفتار با کدام احساس رابطه دارد و به خوبی آن احساس را توصیف می‌کند، بایستی تعداد زیادی ویژگی در نظر گرفت. بنابراین، انتخاب ویژگی‌های مناسب و تمییز دهنده احساسات امری ضروری است. برای همین منظور در این مقاله برای انتخاب ویژگی‌های مناسب گفتاری مرتبط با احساسات از روش چند وظیفه‌ای استفاده شده است. بدین صورت که تابع هدفی چند وظیفه‌ای ارایه شده که در آن هر فرد به عنوان یک وظیفه در نظر گرفته می‌شود. در نتیجه با حل این تابع هدف یک مجموعه ویژگی مستقل از گوینده انتخاب می‌شود به نحوی که این ویژگی‌ها تمییزدهنده خوبی در بین تمام احساسات مختلف باشند. چون این ویژگی‌ها در بین تمام کلاس‌ها (احساسات) مشترک هستند، این ویژگی‌ها را می‌توان به طور مستقیم برای طبقه‌بندهای چند کلاسه استفاده نمود و یا با استفاده از طبقه‌بندهای دودویی به سادگی طبقه‌بندی چند کلاس را انجام داد. برای ارزیابی روش پیشنهادی، دو دادگان معروف Berlin و Enterface مورد استفاده قرار گرفت. همچنین، استخراج ویژگی توسط ابزار OpenSmile صورت گرفت و بیش از ۶۵۰۰ ویژگی استخراج شد. نتایج آزمایشات نشان می‌دهند که روش پیشنهادی کارایی بهتری نسبت به روش‌های موجود دارد و زمان اجرای آن نسبت به سایر روش‌ها کمتر است. در این مقاله، هفت طبقه‌بند استفاده شد و نتایج نشان داد که در برابر گوینده جدید، بهترین کارایی برای دادگان Berlin ۷۶.۷۳ درصد و برای دادگان Enterface ۷۲.۱۷ می‌باشد. نتایج نشان می‌دهد که روش پیشنهادی بهتر از سایر روش‌های موجود است.

کلمات کلیدی: شناسایی حالت احساس گفتار، انتخاب ویژگی چندوظیفه‌ای، ویژگی‌های مستقل از گوینده، انتخاب ویژگی مستقل از دادگان، پردازش احساس.