Applied Acoustics 146 (2019) 320–326

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A novel feature selection method for speech emotion recognition

Turgut Özseven

Department of Computer Engineering, Tokat Gaziosmanpaşa University, Tokat, Turkey

ARTICLE INFO

Article history:
Received 29 October 2018
Accepted 22 November 2018

Keywords:
Emotion recognition
Speech processing
Speech emotion recognition
Feature selection

ABSTRACT

Speech emotion recognition involves analyzing vocal changes caused by emotions with acoustic analysis and determining the features to be used for emotion recognition. The number of features obtained by acoustic analysis reaches very high values depending on the number of acoustic parameters used and statistical variations of these parameters. Not all of these features are effective for emotion recognition; in addition, different emotions may effect different vocal features. For this reason, feature selection methods are used to increase the emotional recognition success and reduce workload with fewer features. There is no certainty that existing feature selection methods increase/decrease emotion recognition success; some of these methods increase the total workload. In this study, a new statistical feature selection method is proposed based on the changes in emotions on acoustic features. The success of the proposed method is compared with other methods mostly used in literature. The comparison was made based on number of feature and emotion recognition success. According to the results obtained, the proposed method provides a significant reduction in the number of features, as well as increasing the classification success.

1. Introduction

The emotion that people are constantly experiencing and a physiological response have been investigated by many researchers both from engineering and medical fields for many years [1]. The first studies of emotion recognition have revealed vocal cues of emotional speech, and that these are different among people [1,2]. While "emotion" and "learning due to emotional reactions" are specific to the individual, the basic features are consistent with all people [3]. Emotion recognition has an important place in human-computer interaction studies, and speech and face expressions can be used to estimate emotional state. In recent years, Speech Emotion Recognition (SER) has been used in many areas such as medical emergency for stress and pain detection, interactions with robots, computer games and call centers [4].

SER systems are trained with recordings of people's emotional speech. The communicative influence of words pronounced in a sentence may differ according to the speakers' emotions [5]. Therefore, the effects of the emotion on the speech are investigated rather than the words spoken in the SER studies. SER is not a new field of study and was first implemented in the mid-1980s using statistical properties of some acoustic parameters. In the following years, with the development of computer architecture, more complex emotion recognition algorithms have begun to be used [6].

An important aspect of the SER is to obtain speech features that characterize the emotional content of the speech efficiently and at the same time are not dependent on the speaker or the word content [7]. For this purpose, speech is analyzed by acoustic analysis methods, and then acoustic properties are obtained. Together with acoustic features, language and discourse knowledge was used for emotion detection [8].

The number of features obtained after acoustic analysis reaches 1000; this increases both workload and negative consequences in SER performance [4]. For this reason, feature selection methods are used to speed up the learning process for SER systems and to reduce problems caused by the large number of features analyzed [9]. The main purpose of feature selection is to determine the properties that will achieve the best classification from the feature set. By selecting features, the size of the feature data set is reduced in an attempt to improve the classification performance and accuracy. Although a large number of feature selection methods are used for SER studies in the literature, some methods reduce SER success after size reduction. Furthermore, the methods used are not specific to SER and can be used in many fields of study.

Some feature selection methods used in the SER are [4] Forward Feature Selection (FFS) and Backward Feature Selection (BFS) [10], Sequential Floating Forward Selection (SFFS) [11], wrapper approach with forward selection [12], Principal Component

E-mail address: turgut.ozseven@gop.edu.tr

https://doi.org/10.1016/j.apacoust.2018.11.028
0003-682X/

Analysis (PCA), or Linear Discriminate Analysis (LDA) [13,14]. The PCA, Sequential Forward Selection (SFS), and Fast Correlation-Based Filter (FCBF) feature selection methods are at the forefront of the SER studies. Among these methods, PCA is the most used method and used in many studies [7,8,15–22].

There are many classifiers used in SER systems. These classifiers can be used independently or as a hybrid classifier containing a combination of multiple classifiers [14]. Support Vector Machine (SVM) is one of the most used classifiers in the literature in the literature and the highest success rate in EMO-DB studies (7 emotions) is 93.78% [23]. However, the number of data in the study was reduced from 535 to 494. In another study using SVM classifier, 535 data sets and 37 feature sets achieved 80.2% success [24]. In the study performed with Multilayer Perceptron (MLP) classifier and 493 data set, 89.62% success was achieved [25]. Schuller et al. (2005) achieved 86.5% success using 4 emotional states [26]. In another study, 66.8% success was achieved with MLP classifier and 7 emotional states [27]. In studies involving k-Nearest Neighbors (k-NN), the lowest success rate was found to be 51.7% [28] with 6 emotional states and, the highest success rate was found to be 78.9% [26] with 4 emotional states.

Usage of feature selection methods does not always increase the success of classifiers. In [16], the effect on the classifier's success of feature selection was investigated and higher success was achieved in using all the features. In a study comparing PCA and LDA methods, it was found that the combined use of the two methods gave better results than the individual use. Because PCA is more effective on de-correlated data, LDA is more effective on low dimensional data [29]. Fischer is better than PCA in size reduction according to the results of the study where four comparative experiments were performed using two-feature reduction method (Fischer and PCA) and two classifiers [7]. Schuller et al., (2005) reduced the feature size from 276 to 75 using the SVM classifier and the SFFS feature selection method and increased the recognition of the emotion by 2.7% [30]. Altun and Polat (2009) presented two different frameworks for SER and used SFS, Least Squares (LS) bound, Mutual Information (MUTINF) and R^2W^2 feature selection methods. They reduced the size of the 58 attributes to 18, 31, 28 and 33, respectively, increasing the SER success by 4.5%, 3.0%, 2.2% and 0.7% respectively [31]. Luengo et al. (2010) reduced SER performance by 1.5% when Minimum Redundancy Maximum Relevance (mRMR) feature selection method reduced the feature size from 380 to 121 [32]. When Gharavian et al. (2013) reduced the feature size from 55 to 49, 45, 24 and 8 with the FCBF feature selection method, the change in the SER success was 0.9%, −1.1%, −2.3% and −3.4% [33]. Zhao et al. (2014) increased the SER success by 1.5% by reducing the feature size from 204 to 87 using the FCBF feature selection method and the SVM classifier [34].

Based on current studies, it cannot be said that any feature selection method increases or decreases the SER performance. The effect of feature selection methods on SER success depends on the classifier, data and size reduction ratio. In addition, existing feature selection methods use many areas, including SER. In this context, a detailed investigation of effective methods according to the classifier types and data sets as well as the creation of a feature selection method for SER will guide and contribute to the future work of researchers in the SER field.

In this study, a new feature selection method based on emotion-based changes of acoustic properties is proposed. The success of the proposed method has been examined comparatively with the existing methods. Methodology of working in Section 2 is given. In Section 3, the proposed method is given. Section 4 contains the results of the analysis. The results obtained in the last part of the study are also interpreted.

2. Material and methods

2.1. The used data

Four different data sets (EMO-DB, eNTERFACE05, EMOVO and SAVEE) were used to improve the generalization capabilities of the findings obtained in the study.

The Berlin Database of Emotional Speech (EMO-DB) is derived from the expression of different emotions by actors (anger, boredom, disgust, anxiety/fear, happiness, sadness, neutral). Audio recordings have 16 kHz sampling frequency and are 16 bit mono [35,36]. The eNTERFACE'05 is an audio-visual emotion database (anger, disgust, fear, happiness, sadness, surprise). The database contains 42 subjects, coming from 14 different nationalities [37]. Italian Emotional Speech Database (EMOVO) is a database built from the voices of up to 6 actors who played 14 sentences simulating emotional states (anger, disgust, fear, happiness, neutral, sadness, surprise) [36]. The recordings were performed with a sampling frequency of 48 kHz, 16 bit stereo, wave format [38]. Surrey Audio-Visual Expressed Emotion (SAVEE) Database recorded an audio-visual emotional database (anger, disgust, fear, happiness, sadness, neutral, surprise) from four native English male speakers, one of them was postgraduate student and rest were researchers at the University of Surrey [13]. The distribution of the data used in the study is given in Table 1.

2.2. Acoustic analysis

SER generally consists of 4 steps as shown in Fig. 1. Framing, windowing, feature extraction and classification steps are required. The signal processing methods of pre-processing and post-processing are used to reduce the workload or to increase the classification accuracy.

In the pre-processing step, digital signal processing methods such as noise reduction and filtering are applied to the speech signal. The speech signal is a non-stationary signal but is assumed to be stationary over short time intervals. In addition, since the speech signals are stable at small time intervals, the signal is processed by taking the sections at certain intervals. Framing is used for this process. When framing is performed, the frame size along with the overlap ratio are determined. The overlap rate is used to soften the transition from one frame to another and to prevent information loss. Windowing is used to arrange spectral infiltration in the signal and intersection due to overlap. Framing and windowing are part of the pre-processing process. In the feature extraction phase, features are extracted from each frame of the speech signal divided by frames. In the post-processing step, normalization and attribute selection methods are applied to the feature set. Unit variances in the feature set and the numerical size of the data directly affect the performance of the classifier; normalization techniques are used to overcome this problem. Feature selection methods are used to determine the features that will achieve the best classification from the feature set. By selecting features, the size of the feature data set is reduced in an attempt to improve the classification performance and accuracy. In the last step, emotion recognition is performed using the obtained feature set.

Framing and windowing were used for preprocessing methods when the SER process was performed. The frame size for framing is between 20 ms and 40 ms, and the overlap ratio is 50%. Hamming method is used for windowing. OpenSMILE v2.1.0 [39] software was used for framing, windowing and feature extraction. During the feature extraction step, the emobase2010 configuration file was used and 1582 features were obtained from each speech record. Please refer to the openSMILE book for detailed features

Table 1
Distribution of speech recordings used in the study.

Data Set	AG	DG	FR	HP	SD	BD	SR	NR	Total
EMO-DB	127	46	69	71	62	81	–	79	535
eNTERFACE'05	176	190	196	198	182	–	205	–	1147
EMOVO	72	68	69	72	72	–	72	72	497
SAVEE	60	60	60	60	60	–	60	120	480

AG: anger, DG: disgust, FR: anxiety/fear, HP: happiness, SD: sadness, BD: boredom, SR: surprise, NR: neutral.

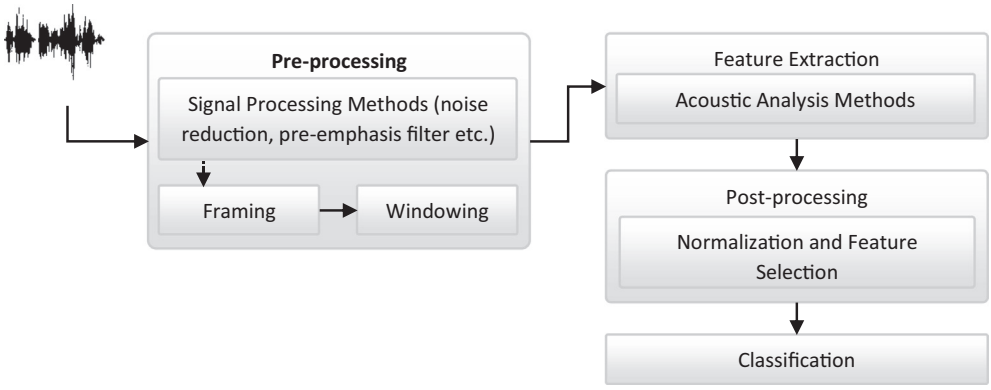


Fig. 1. SER flow diagram.

[40]. 20 ms and 40 ms framing is used depending on the feature to be obtained in the configuration file used.

2.3. Classifiers

SVM, MLP and k-NN classifiers and WEKA [41] packet program for classification were used. SVM is a method based on statistical learning theory. The main goal is based on the principle of defining the hyper-plane in other words the decision function that best separates the classes from each other. MLP are layered feedforward networks typically trained with static back propagation in order to classify static pattern [25]. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid [42]. However, MLP train slowly, and require lots of training data [25].

According to the existing studies, the classifier used, the selected feature set, the number of emotions contained in the data set, and the data set affect the results significantly. Linear kernel is preferred for SVM since the number of features is high. In the MLP model; the number of input neurons = number of features, number of hidden layers = (number of features + number of emotions)/2 and number of output neurons = number of emotions.

3. Proposed feature selection method

Feature selection which is used to reduce the size of the feature set before classification is used to speed up the learning process in the SER systems and reduce feature size-based problems [9]. Current methods are preferred in many fields of study and are not specific to the SER. Because the change in acoustic features is different by the emotional state. The proposed feature selection method is based on the changes of the feature set according to the emotions. The following definitions have been made for this purpose.

Definition 1. (*(mean of emotion)*). *The mean of each feature in the feature set is calculated on an emotion-based basis* (f_{mean}).

$$f_{mean}(c,i) = \frac{1}{|c|}\sum_{l\in c} f(l,i) \quad for \quad \forall i \tag{1}$$

In the Eq. (1); *f:* the feature set including the emotion classes, *c:* the emotions contained in the feature set, and *i:* each feature column contained in the feature set. An exemplary cross-section of the feature set is given in Table 2.

The *fmean (anger, F1)* in Table 2 is the average of F1 values for *anger* in the data set. The *fmean (fear, F1)* in Table 2 is the average of F1 values for *fear* in the data set.

Definition 2. (*(change between emotions)*). *The emotional exchange of features is calculated. The mean of the differences between the dual combinations of emotions is used for this process* ($f_{c\text{-}mean}$).

$$f_{c-mean}(i) = \frac{(n-2)!\times 2!}{n!}\sum_{j=1}^{n}\sum_{k=j+1}^{n}|f_{mean}(j,i) - f_{mean}(k,i)| \quad for \quad \forall i \tag{2}$$

In the Eq. (2); n: is the number of emotions contained in the feature set. The change between emotions for exemplary cross-section given in Table 2 is given in Table 3.

The *fc-mean* in Table 3 shows the average of the values of emotional changes for each feature. For example, the value 0.2498 is the average of the F1 values of "Anger-Fear, Anger-Happiness and Fear-Happiness" emotions.

Table 2
An exemplary cross-section from the feature set.

F1	F2	Class
0.088545	0.148411	Anger
0.462781	0.185746	Fear
0.209239	0.209966	Anger
0.071969	0.280642	Happiness
0.584426	0.153081	Fear
0.408545	0.256955	Happiness

c: {anger, fear, happiness}, i: {F1, F2}.
f_{mean} (anger, F1) = 0.1489.
f_{mean} (fear, F1) = 0.5236.

Definition 3. (*(threshold value)*). *The threshold is set for features to be cast from the feature set. Four statistical methods were used to determine the threshold value of the proposed method. These; the standard deviation (SD) of $f_{c\text{-}mean}$, the mean (MN) of $f_{c\text{-}mean}$, the median (MED) of $f_{c\text{-}mean}$ and the coefficient of variation (CV) of $f_{c\text{-}mean}$. The equations for these methods are given below.*

$$th_{SD} = \sqrt{\frac{1}{|i|}\sum_{m\in i}\left(f_{c-mean}(i) - \overline{f_{c-mean}}\right)^2} \tag{3}$$

$$th_{MN} = \frac{1}{|i|}\sum_{m\in i} f_{c-mean}(i) \tag{4}$$

$$f_{ord} = order(f_{c-mean}) \tag{5}$$

$$th_{MED} = \begin{cases} f_{ord}\left(\frac{|i|+1}{2}\right), & if\ |i|\ is\ odd \\ f_{c-mean}\left(\frac{f_{ord}\left(\frac{|i|}{2}\right)+f_{ord}\left(\frac{|i|}{2}+1\right)}{2}\right), & if\ |i|\ is\ even \end{cases} \tag{6}$$

$$th_{CV} = \frac{th_{SD}}{th_{MN}} \times 100 \tag{7}$$

Definition 4. (*(new feature set)*). *Features that are unimportant from the feature set are discarded and a new feature set is created. Features with an $f_{c\text{-}mean}$ value below the threshold value set for this process have been removed.*

$$f_{new}(c,i) = \begin{cases} f(c,i), & f_{c-mean} \geq threshold \\ 0, & otherwise \end{cases} \quad for\ \forall i \tag{8}$$

For the sample feature set given in Table 2, when the SD and MN threshold values are used, F2, F3, and F5 are removed from the feature set. When the MED threshold value is used, F3 and F5 are removed from the feature set.

For proposed method, the corresponding descriptive pseudocode is shown in Algorithm 1.

Algorithm 1. (proposed method)

Input: A sample data (f) with a full feature set
F={F_1, F_2, F_3,..,F_n) and the emotion classes C
A predefined threshold th_{SD}
Output: The selected feature subset f_{new}
1. $f_{new} \leftarrow \varnothing$;
2. Calculate the class count
$c_n \leftarrow |C|$;
3. For $x = 1$ to c_n do
4. Calculate mean of emotion
class_mean(: ,*x*) = *mean*(*f*(*find*(*f* (:,*end*)==*C*(*x*)),*1:end-1*));
5. End
6. *counter* ← 1;
7. For $y = 1$ to c_n do
8. For $z = y + 1$ to c_n do
9. Calculate change between emotions
mean_difference(:,*counter*) = abs(*class_mean*(: ,*y*) - *class_mean*(: ,*z*));
10. *counter* = *counter* + 1;
11. End
12. End
13. Calculate mean of change between emotions
feature_mean = mean(*mean_difference*,2);
14. f_{new} = *f*(: ,find(*feature_mean*>= th_{SD}));
15. End

Table 3
The change between emotions for a feature set with 4 emotions and 5 attributes.

	Anger-Fear	Anger-Happiness	Fear-Happiness
F1	0.3747115	0.0913650	0.2833465
F2	0.0097750	0.0896100	0.0993850

$f_{c\text{-}mean}$: {0.2498, 0.0663}.

4. Experimental results

1582 features have been extracted with openSMILE from each speech record in the 4 data sets used in the study. SVM, MLP and k-NN classifiers were used while performing the experimental study. PCA, SFS and FCBF feature selection methods were chosen to compare the effectiveness of the proposed feature selection method. All feature selection and classification processes were performed with WEKA [41]. To evaluate the classification success, the feature set is divided into a training and test set with 10-fold cross-validation. All analyzes were performed on a computer with an i7 2.7 GHz processor and 16 GB of RAM. The classification accuracies obtained with all feature set are given in Table 4.

As seen from the results in Table 4, emotion recognition success rates in SER studies vary depending on factors such as the data set, the way the data is collected, and the classifier used. It has been observed that the SVM classifier mostly achieves higher success in 4 different data sets. The MLP classifier only provides the highest success on eNTERFACE'05. However, when the workload is examined, there are many differences between; a similar situation is seen in all data sets for the MLP classifier.

The size of the new feature set obtained after feature selection varies according to the data set used. When PCA, SFS, FCBF, and proposed feature selection methods are applied, the feature size variation according to the data set and workload are given in Figs. 2 and 3.

Fig. 2 shows the number of features included in the new feature set obtained after applying feature selection methods. The methods that minimum and maximum the number of features are FCBF and PCA, respectively. When the variation of the feature number of the proposed method is examined, the results differ according to the threshold value. In the proposed method, the threshold values that decrease the number of features the most and the least are th_{CV} and th_{MED}, respectively.

One of the main objectives of feature selection methods is to reduce workload. However, it should be noted that the application of the feature selection method will create additional workload. For this reason, the workload of these methods is examined and the results are given in Fig. 3.

The application of feature selection methods is mostly based on the SFS method, with the highest workload according to the resulting workload results. Upon examination of the workload results of the proposed method, it is observed to be less than one second for all threshold values. For this reason, the workload that will be created for the feature selection of the proposed method will not have much effect on the overall workload. Table 5 presents the total workload for the feature selection method and classification operation.

Based on the total workloads given in Table 5, the proposed method generally had the lowest total workload. In all data sets, PCA and SFS increased workload with SVM and k-NN, while MLP reduced workload. FCBF decreased the workload with SVM and MLP, while workload increased with k-NN. The major factor in these changes in total workload is the workload required to implement the feature selection method. Because the feature selection method reduces the number of features, the classification workload decreased. However, the workload of the feature selection

Table 4
The classification accuracies obtained with all feature set.

Data Set	SVM		k-NN		MLP	
	Acc. (%)	El.Time (s)	Acc. (%)	El.Time (s)	Acc. (%)	El.Time (s)
EMO-DB	84.62	4.56	63.74	0.01	81.32	628.40
eNTERFACE'05	59.74	11.39	39.74	0.01	69.23	1648.60
EMOVO	60.40	3.17	39.05	0.01	58.58	621.65
SAVEE	72.39	2.30	53.37	0.01	71.17	527.28

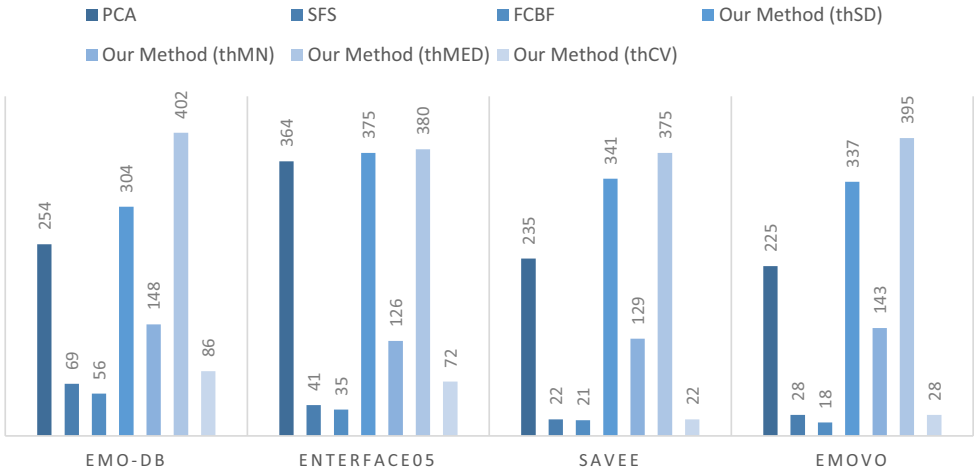
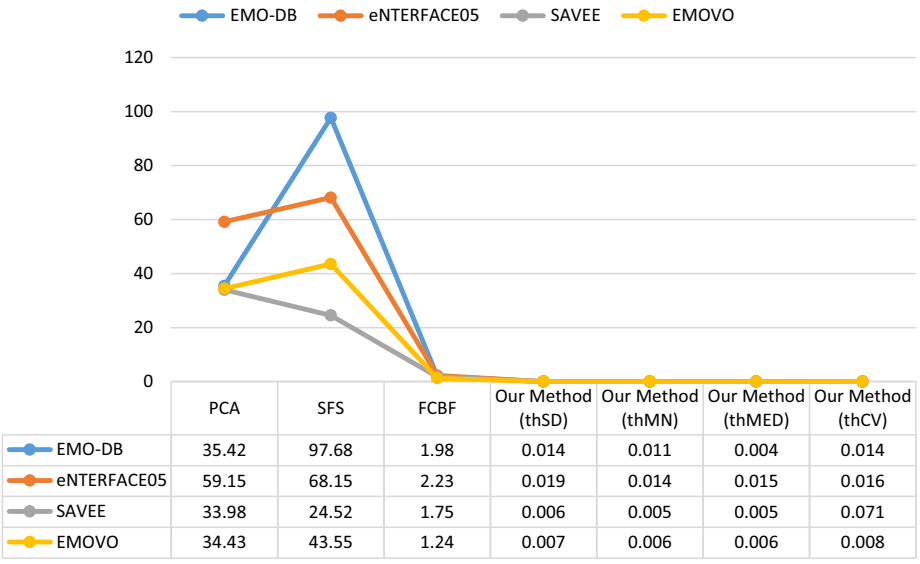


Fig. 2. Change in feature dimensions after feature selection.



	PCA	SFS	FCBF	Our Method (thSD)	Our Method (thMN)	Our Method (thMED)	Our Method (thCV)
EMO-DB	35.42	97.68	1.98	0.014	0.011	0.004	0.014
eNTERFACE05	59.15	68.15	2.23	0.019	0.014	0.015	0.016
SAVEE	33.98	24.52	1.75	0.006	0.005	0.005	0.071
EMOVO	34.43	43.55	1.24	0.007	0.006	0.006	0.008

Fig. 3. The workloads for feature selection (second).

method causes the total workload to change. The emotion recognition successes of the new feature sets obtained after feature selection are given in Table 6.

The success rate obtained without a feature selection method as the reference value is used to determine the effect of the feature selection methods. The change in the results obtained after feature selection is determined according to this reference value. After the feature selection method, cases in which the success rate is higher, lower or equal to the reference value are indicated by ↑, ↓ and ↔, respectively. One of the main objectives of the feature selection method is to achieve higher success with fewer features. Another goal is to reduce the workload even if the success rate is not increased. Table 6 contains the results of 84 experimental studies performed after feature selection. According to these results, the success rate decreased in 44 experiments, increased in 36 experiments and did not change in four experiments. To examine the results in detail based on the data set and classifier; in EMO-DB + SVM, all feature selection methods reduced classification success. In EMO-DB + MLP, PCA and FCBF decreased the success rate, and the proposed method and SFS classification increased the success rate. In EMO-DB + k-NN, all methods except for PCA increased the success rate, and the highest achievements were obtained with the proposed method. In eNTERFACE05 + SVM, SFS and FCBF decreased the success rate while other methods increased. In

Table 5
Total workload for feature selection method and classification.

Data Sets	Classifier	Total workload (second)							
		All Features	PCA	SFS	FCBF	OM (th_{SD})	OM (th_{MN})	OM (th_{MED})	OM (th_{CV})
EMO-DB	SVM	4.56	35.85	97.86	2.14	3.34	1.23	2.12	0.46
	MLP	628.40	47.95	98.78	2.76	433.24	119.26	177.29	8.85
	k-NN	0.01	35.43	97.68	1.98	0.02	0.01	0.01	0.01
eNTERFACE05	SVM	11.39	62.63	68.43	2.55	10.40	5.64	6.28	0.71
	MLP	1648.56	129.73	69.31	3.05	870.65	235.49	359.15	3.02
	k-NN	0.01	59.16	68.16	2.23	0.01	0.03	0.01	0.02
SAVEE	SVM	2.30	34.55	24.58	1.84	2.99	1.57	2.18	0.16
	MLP	527.28	46.60	24.71	1.94	380.81	114.78	151.20	2.37
	k-NN	0.01	33.99	24.52	1.75	0.01	0.02	0.01	0.071
EMOVO	SVM	3.17	34.12	43.68	1.34	4.17	1.43	2.23	0.10
	MLP	621.65	45.50	43.83	1.40	346.59	112.60	149.81	0.29
	k-NN	0.01	34.44	43.56	1.241	0.01	0.02	0.01	0.01

Table 6
Classification successes after feature selection.

Data Set	Classifier	Classification accuracy (%)							
		All	PCA	SFS	FCBF	OM (th_{SD})	OM (th_{MN})	OM (th_{MED})	OM (th_{CV})
EMO-DB	SVM	84.62	↓ 81.71	↓ 82.93	↓ 81.32	↓ 84.07	↓ 83.00	↓ 81.87	↓ 78.57
	MLP	81.32	↓ 60.99	↑ 85.71	↓ 80.77	↑ 82.42	↑ 85.71	↑ 82.97	↑ 82.97
	k-NN	63.74	↓ 30.77	↑ 71.43	↑ 69.78	↑ 68.13	↑ 73.08	↑ 72.53	↑ 71.98
eNTERFACE05	SVM	59.74	↑ 62.31	↓ 49.49	↓ 54.62	↑ 60.51	↑ 60.77	↑ 60.00	↓ 48.72
	MLP	69.23	↓ 49.74	↓ 57.69	↓ 57.69	↓ 67.18	↓ 68.46	↓ 67.95	↓ 48.97
	k-NN	39.74	↓ 23.59	↑ 43.85	↓ 38.46	↑ 41.03	↑ 41.03	↑ 41.79	↓ 33.85
SAVEE	SVM	72.39	↔72.39	↓ 66.26	↓ 55.83	↑ 73.62	↑ 77.92	↑ 74.85	↓ 57.67
	MLP	71.17	↓ 49.69	↓ 65.03	↓ 60.12	↑ 73.62	↔71.17	↔71.17	↓ 54.60
	k-NN	53.37	↓ 22.70	↑ 58.28	↓ 49.69	↑ 57.06	↑ 55.83	↓ 52.76	↓ 47.24
EMOVO	SVM	60.40	↑ 63.91	↓ 51.48	↓ 50.89	↑ 61.54	↓ 59.17	↑ 60.95	↓ 43.20
	MLP	58.58	↓ 42.60	↓ 48.52	↓ 48.52	↔58.58	↓ 56.21	↑ 59.17	↓ 43.79
	k-NN	39.05	↓ 23.67	↑ 59.17	↑ 41.42	↑ 42.60	↑ 46.15	↑ 43.20	↑ 43.79

eNTERFACE05 + MLP, all methods reduced the success rate. In eNTERFACE05 + k-NN, PCA and FCBF decreased the success rate while other methods increased, the success rate. In SAVEE + SVM, the proposed method increased the success rate, and other methods decreased the success rate. In SAVEE + MLP, the current methods reduced the success rate, and the proposed method contributed to the workload and success rate. In SAVEE + k-NN, the proposed method and SFS increased the success rate, while other methods decreased the success rate. In EMOVO + SVM, PCA and the proposed method increased the success rate, and other methods decreased the success rate. In EMOVO + MLP, the current methods reduced the success rate, but the proposed method contributed to the success rate based on the selected threshold value. In EMOVO + k-NN, other methods except for PCA increased the success rate. The highest success in the EMO-DB data set was obtained with a combination of MLP + SFS and MLP + OM (th_{SD}). The highest success in the eNTERFACE05 data set was obtained with MLP + OM (th_{MN}). The highest success in the SAVEE data set was obtained with SVM + OM (th_{MED}). The highest success in the EMOVO data set was obtained with SVM + PCA.

5. Conclusion

One of the main objectives of SER studies is to achieve high success with a small feature set. For this purpose, classifiers, preprocessing methods and post-processing methods used for the SER are important. However, in real-time applications, the success rate can be decreased slightly to reduce the workload. One of the methods used to reduce the workload or increase the success rate is normalization; this removes measurement differences from the data set and contributes to classifier performance. However, normalization is often not sufficient by itself and feature selection methods should also be used. The main purpose of feature selection methods is to remove unnecessary data from the feature set to speed up the learning process and reduce problems caused by feature size.

In this study, a new feature selection method based on emotional differences for the SER studies is proposed. The validity of the proposed method is experimentally compared to PCA, SFS and FCBF feature selection methods with three different classifiers on different data sets. When the experimental results were examined in the case for which the feature selection method was not used, the success rate was 84.62% on EMO-DB, 69.23% on eNTERFACE'05, 60.40% on EMOVO and 72.39% on SAVEE. When the workloads were examined, the lowest workload related to the k-NN classifier; however, the success rate was low for the k-NN classifier. For the MLP classifier, the workload was too high. A researcher using the eNTERFACE'05 data set should choose the priority in this case; if success rate is considered more important, the MLP classifier should be chosen, while the SVM classifier would be preferred if the calculation load is more important. One of the ways to reduce workload independently of the classifier is to reduce the size of the feature set.

There is also a workload involved in implementing feature selection methods. It would not make sense to use this method if the resulting workload impacts the overall workload to heavily. In this context, the workloads of the present feature selection methods and the proposed method were compared. According to the results, there was a significant increase in the workload with

the preferred method. However, the workload of the proposed method is below one second; this suggests that the proposed method will not put an additional burden on the overall workload.

According to the effects on classification performance of the four selection sets, three classifiers and feature selection methods, these methods caused a significant decrease in feature size. The feature size of 1582 was reduced between 77% and 86% with PCA, between 95.6% and 98.6% with SFS, between 96.5% and 98.8% with FCBF and between 75% and 98.6% with the proposed method, depending on the data set. These reductions in the feature size were reflected in the success of the classification. Feature selection increased the success rate in 42.85% of 84 experimental works. In addition, as the feature size decreased, the workload declined significantly. In this context, there is no definite provision for effect on the success rate of feature selection methods; many factors, such as the method used, the data set and the feature set affect the rate of emotion recognition.

The proposed feature selection method produced successful results in general terms; depending on the selected threshold value, the size of the feature set decreased, the workload decreased, and the classification success mostly increased in all the experimental results. In the data sets (except for the EMOVO data set), the highest classification success was achieved by the proposed method. The success rate in the EMOVO data set also increased, but this increase was 2.37% lower than with PCA. These results prove that the proposed method can be used in SER studies.

References

[1] Rong J, Li G, Chen Y-PP. Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag May 2009;45(3):315–28.
[2] Fairbanks G, Hoaglin LW. An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monogr 1941;8(1):85–90.
[3] Sethu V. Automatic emotion recognition: an investigation of acoustic and prosodic parameters. The University of New South Wales; 2009.
[4] Gharavian D, Sheikhan M, Nazerieh A, Garoucy S. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput Appl 2011;21(8):2115–26.
[5] Mencattini A et al. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowl-Based Syst 2014;63:68–81.
[6] Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Commun Sep. 2006;48(9):1162–81.
[7] Chen L, Mao X, Xue Y, Cheng LL. Speech emotion recognition: features and classification models. Digit Signal Process Dec. 2012;22(6):1154–60.
[8] Lee Chul Min, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process Mar. 2005;13(2):293–303.
[9] Anagnostopoulos C-N, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev Feb. 2015;43(2):155–77.
[10] Pao T-L, Chen Y-T, Yeh J-H, Chang Y-H. Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification. In: Proceedings of the 17th Conference on Computational Linguistics and Speech Processing. p. 203–12.
[11] Ververidis D, Kotropoulos C. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. 2006 14th European Signal Processing Conference 2006:1–5.
[12] Sidorova J. Speech emotion recognition with TGI+. 2 classifier. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. p. 54–60.
[13] Jackson P, Haq S, Edge JD. Audio-visual feature selection and reduction for emotion classification. In: Proc. Int'l Conf. on Auditory-Visual Speech Processing. p. 185–90.
[14] Özseven T, Düğenci M, Durmuşoğlu A. A content analysis of the research approaches in speech emotion recognition. 7(1); 2018: 1–26.
[15] Erickson D, Yoshida K, Menezes C, Fujino A, Mochida T, Shibuya Y. Exploratory study of some acoustic and articulatory characteristics of sad speech. Phonetica 2006;63(1):1–25.
[16] Grimm M, Kroschel K, Mower E, Narayanan S. Primitives-based evaluation and estimation of emotions in speech. Speech Commun Oct. 2007;49(10–11):787–800.
[17] Busso C, Narayanan SS. Interrelation between speech and facial gestures in emotional utterances: a single subject study. IEEE Trans Audio Speech Lang Process Nov. 2007;15(8):2331–47.
[18] Goudbeek M, Scherer K. Beyond arousal: valence and potency/control cues in the vocal expression of emotion. J Acoust Soc Am 2010;128(3):1322.
[19] Patel S, Scherer KR, Björkner E, Sundberg J. Mapping emotions into acoustic space: the role of voice production. Biol Psychol Apr. 2011;87(1):93–8.
[20] Laukka P, Neiberg D, Forsell M, Karlsson I, Elenius K. Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation. Comput Speech Lang Jan. 2011;25(1):84–104.
[21] Ntalampiras S, Fakotakis N. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. IEEE Trans Affect Comput 2012;3(1):116–25.
[22] Scherer KR, Sundberg J, Tamarit L, Salomão GL. Comparing the acoustic expression of emotion in the speaking and the singing voice. Comput Speech Lang Jan. 2015;29(1):218–35.
[23] Ivanov A, Riccardi G. Kolmogorov-Smirnov test for feature selection in emotion recognition from speech. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. p. 5125–8.
[24] Chiou B-C, Chen C-P. Feature space dimension reduction in speech emotion recognition using support vector machine. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific. p. 1–6.
[25] Khanchandani KB, Hussain MA. Emotion recognition using multilayer perceptron and generalized feed forward neural network. J Sci Ind Res 2009;68(5):367.
[26] Schuller B, Arsić D, Wallhoff F, Lang M, Rigoll G. Bioanalog acoustic emotion recognition by genetic feature generation based on low-level-descriptors. In: Computer as a Tool, 2005. EUROCON 2005. The International Conference on. p. 1292–5.
[27] Albornoz EM, Milone DH, Rufiner HL. Spoken emotion recognition using hierarchical classifiers. Comput Speech Lang Jul. 2011;25(3):556–70.
[28] Lanjewar RB, Mathurkar S, Patel N. Implementation and comparison of speech emotion recognition system using gaussian mixture model (GMM) and K-nearest neighbor (K-NN) techniques. Procedia Comput Sci 2015;49:50–7.
[29] Hoque ME, Yeasin M, Louwerse MM. Robust recognition of emotion from speech. In: Intelligent virtual agents. p. 42–53.
[30] Schuller B, Müller R, Lang MK, Rigoll G. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: INTERSPEECH. p. 805–8.
[31] Altun H, Polat G. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst Appl 2009;36(4):8197–203.
[32] Luengo I, Navas E, Hernaez I. Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans Multimed 2010;12(6):490–501.
[33] Gharavian D, Sheikhan M, Ashoftedel F. Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model. Neural Comput Appl 2013;22(6):1181–91.
[34] Zhao X, Zhang S, Lei B. Robust emotion recognition in noisy speech via sparse representation. Neural Comput Appl Jun. 2014;24(7–8):1539–53.
[35] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. Interspeech 2005;5:1517–20.
[36] Özseven T. The acoustic cues of fear: investigation of acoustic parameters of speech containing fear. Arch Acoust 2018;43(2):245–51.
[37] Martin O, Kotsia I, Macq B, Pitas I. The eNTERFACE'05 audio-visual emotion database. Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on, 2006. pp. 8–8.
[38] Costantini G, Iaderola I, Paoloni A, Todisco M. EMOVO corpus: an Italian emotional speech database. In: LREC. p. 3501–4.
[39] Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia. p. 835–8.
[40] Eyben F, Woellmer M, Schuller B. The Munich open speech and music interpretation by large space extraction toolkit. IEEE Netw 2010;24(2):36–41.
[41] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 2009;11(1):10–8.
[42] Boersma P, Weenink D. Praat: doing phonetics by computer [Computer program]. Version 5.1. 44; 2010.