An optimal two stage feature selection for speech emotion recognition using acoustic features

Swarna Kuchibhotla[1] · Hima Deepthi Vankayalapati[2] · Koteswara Rao Anne[2]

Received: 23 March 2016 / Accepted: 23 July 2016 / Published online: 2 August 2016

Abstract Feature Fusion plays an important role in speech emotion recognition to improve the classification accuracy by combining the most popular acoustic features for speech emotion recognition like energy, pitch and mel frequency cepstral coefficients. However the performance of the system is not optimal because of the computational complexity of the system, which occurs due to high dimensional correlated feature set after feature fusion. In this paper, a two stage feature selection method is proposed. In first stage feature selection, appropriate features are selected and fused together for speech emotion recognition. In second stage feature selection, optimal feature subset selection techniques [sequential forward selection (SFS) and sequential floating forward selection (SFFS)] are used to eliminate the curse of dimensionality problem due to high dimensional feature vector after feature fusion. Finally the emotions are classified by using several classifiers like Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA), Support Vector Machine (SVM) and K Nearest Neighbor (KNN). The performance of overall emotion recognition system is validated over Berlin and Spanish databases by considering classification rate. An optimal uncorrelated feature set is obtained by using SFS and SFFS individually. Results reveal that SFFS is a better choice as a feature subset selection method because SFS suffers from nesting problem i.e it is difficult to discard a feature after it is retained into the set. SFFS eliminates this nesting problem by making the set not to be fixed at any stage but floating up and down during the selection based on the objective function. Experimental results showed that the efficiency of the classifier is improved by 15–20 % with two stage feature selection method when compared with performance of the classifier with feature fusion.

Keywords Speech emotion recognition · Feature fusion · Optimal feature set selection · Classification

1 Introduction

Affective computing is a growing research area used to train the devices in such a way to detect and respond to human emotions in a more appropriate and empathic manner (Tao and Tan 2005). It is mainly used to enhance the communication between human and the machine by capturing and processing information effectively (Cowie et al. 2001; Murray and Arnott 1993). Through this a machine can respond to the user in a natural way like a human being. Even though an extensive development and usage of speech emotion recognition is done in certain applications like education, entertainment, multi media contents management, text to speech synthesis and medical diagnosis but there is still lack of development in speech understanding and recognition applications. For instance, In real time driving scenario application, detecting driver's emotion and make an alert from an accident is a difficult task because the driver's behavior changes with the emotion which occurs due to the communication with co-passengers, entertainment system and mobile devices as shown in Fig. 1.

✉ Swarna Kuchibhotla
swarna.kuchibhotla@gmail.com

Hima Deepthi Vankayalapati
nanideepthi@gmail.com

Koteswara Rao Anne
dac@veltechuniv.edu.in

[1] Acharya Nagarjuna University, Vijayawada, India

[2] Vel Tech Dr. RR & Dr. SR Technical University, Chennai, India

The features extracted from the pre processed speech samples carry most emotional information. Based on the literature survey, the acoustic features mainly classified as Prosody, Spectral and Voice quality features (Cowie et al. 2001). Energy, Pitch and Zero Crossing Rate (ZCR), etc., are comes under Prosody features, Linear Predictive Coding (LPC), Linear Prediction Cepestral Coefficients (LPCC) and Mel Frequency Cepstal Coefficients (MFCC) etc., are comes under Spectral features and tense, harsh, breathy etc., are comes under Voice quality features. From the literature (Luengo et al. 2010), the performance of the system is not good when these features used individually i.e either prosody or spectral, so to improve the performance of speech emotion recognition system, feature fusion technique is used by combining Prosody and Spectral features (Kuchibhotla et al. 2014a).

Even though the performance of the system is improved with feature fusion, it does n't reach to an optimal state. The disadvantage with this technique is the curse of dimensionality i.e., the number of features extracted are more when compared with the number of speech samples. So to improve the performance further, the innovative step in this direction is to use an optimal feature selection methods before classification. This paper concentrate more on implementation and analysis of results with both selection of an optimal feature set and fused feature set. All this is done by an application of various classification techniques viz., Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA), Support Vector Machine (SVM) and K Nearest Neighbor (KNN) on both Berlin and Spanish databases.

This paper is organized as follows, Sect. 2 describes the literature survey, Sect. 3 describes the basic concepts for optimal feature selection, Sect. 4 describes the proposed two stage feature selection method including feature fusion, optimal feature set selection methods and classification, Sect. 5 describes various speech corpora, Sect. 6 describes the experimental results with each feature selection method, Sect. 7 describes the conclusion and Sect. 8 describes future work.

2 Literature survey

Researchers proposed various features and classification techniques for the speech emotion recognition in literature. The selection of features is an important task to efficiently characterize the emotional state of the speech sample (El Ayadi et al. 2011). Even though there are several features explored for speech emotion recognition researchers could not identified which features are best for emotion recognition.

Among the prosodic features specified earlier pitch and energy are the most commonly used features by the researchers because these features contain most of the emotional specific information of the speech sample (Fernandez 2003). According to the studies performed by Williams and Stevens (1981) the emotional state of the speech sample is characterized by the valence arousal space. The arousal state (high arousal vs low arousal) affects the overall energy, energy distribution across the frequency spectrum. Several studies confirmed this conclusion (Iohnstone and Scherer 2000; Cowie and Cornelius 2003). Even though the prosodic features effectively discriminate the emotions of different arousal states (high arousal emotion to low arousal) but there exists a confusion among the emotions of the same arousal state (Luengo et al. 2010). This confusion can be eliminated by using the spectral characteristics of the speech (Scherer et al. 1991; Nwe et al. 2003). According to Bou-Ghazale and Hansen (2000) among the available spectral features mel frequency cepstral coefficients yields better emotion recognition performance. Previous work suggests that the fused prosody and spectral features also significantly reduce the error rate by increasing the performance of the classifier when compared with individual features (Kim et al. 2007; Kwon et al. 2003; Kuchibhotlaa et al. 2015).

After feature extraction, selection classification is one of the important task in speech emotion recognition. Various traditional classifiers like HMM, GMM, SVM, kNN and artificial neural networks have been used for the task of speech emotion recognition. Similar to the features, there has been no agreement on which classifiers best classify the emotion of the speech sample. So in our work we considered an RDA classifier along with the traditional classifiers such as SVM and kNN. RDA effectively classifies the emotion of the speech sample when compared with these traditional classifiers on Berlin and Spanish databases (Kuchibhotla et al. 2014a).

An effective comparative study has already been done among these classifiers with feature fusion and without feature fusion (Kuchibhotla et al. 2014a). Even though the performance of the system increases with the feature fusion

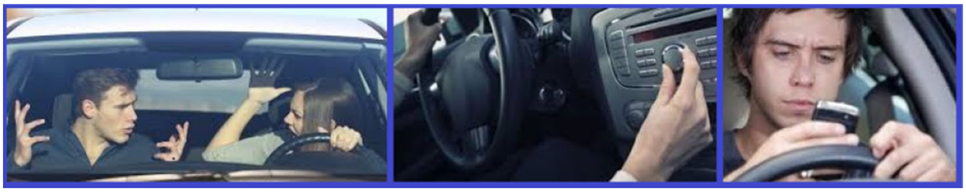

(a) (b) (c)

Fig. 1 Examples of different types of sounds that are produced in car. **a** Co-passenger's voice, **b** music system, **c** mobile phone communication

technique with same classifiers, it suffers with time complexity problem because of high dimensional fused features. In this work we proposed a two stage feature selection technique to reduce the time complexity. This technique selects the features in two stages, first is before the feature fusion, it selects the appropriate features for speech emotion recognition. Second selection stage is after fusion, it reduces the dimensionality and improves the performance of the system.

3 Basic concepts for optimal feature selection

In first stage feature selection, the energy, pitch and mel frequency cepstral coefficients are extracted and fused together to improve the classification performance. The dimensionality of the feature vector after fusion is very high so to reduce the dimensionality of the fused vector a second stage feature selection method is proposed. In this second stage feature selection, optimal feature selection techniques such as Sequential Forward Selection and Sequential Floating Forward Selection are used for dimensionality reduction which also enhances the performance of the classifier.

Optimal feature selection techniques can be distinguished as filters or wrappers based on the criterion function employed. Feature selection can be performed by using the properties such as orthogonality, mutual information, correlation etc. In filter approach, all the features are given a ranking by using some statistical criteria. Highest ranking features are selected and lowest ranking features are removed. But the disadvantage of this method is, they ignore the interdependency of the features and also ignore the interaction with the classifier (Sedaaghi et al. 2007). Due to this the performance of the classifier decreases. Even though the wrappers are slower than the filters, the selected features are more discriminative for specific classifier because wrappers train a classifier using the selected features and estimate the classification error using the validation set (John et al. 1994; Kohavi and John 1997). The most promising methods for wrappers are feature subset selection techniques (Pudil et al. 1994; Ververidis and Kotropoulos 2008) If the total number of features are n then the possible number of feature subset is 2^n in the search space. Now our task is to search in the space of possible feature subsets to find the best optimal feature subset which will classify the emotional states with low classification error rate (Ververidis and Kotropoulos 2006; Efron and Tibshirani 1994).

Each feature subset selection algorithm concentrate on search strategy and evaluation method. Search strategy is used to select the feature subsets and evaluation method is used to test their goodness and fitness based on some criterion function. Search strategies can be classified into exhaustive, sequential and random search (Ververidis and Kotropoulos 2008; Jain and Zongker 1997; Liu and Yu 2005). In exhaustive research the number of possible feature subsets grows exponentially so that this method is impractical even for small feature sets. In sequential search algorithms, insertion or deletion of features is done sequentially. Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are some examples of sequential algorithms which are simple to implement and their execution time is very low. These algorithms are proposed by Whitney (1971), Pohjalainen et al. (2013). SFS starts with an empty set and adds the best selected feature to the feature set in each iteration and in a similar way SBS starts with entire feature set and removes the worst performing feature from the entire set (Sedaaghi et al. 2007). SFS and SBS are suffers from nesting effect so to prevent this nesting of feature subsets another method called plus-l-minus-r is developed by stearns in 1976 (Somol et al. 1999). The drawback of this method is, there is no procedure to predict the values of l and r to achieve the best feature set. Instead of fixing these values let us keep them to float i.e to change the values flexibly to approximate the optimal solution (Pudil et al. 1994). These are called floating search methods which include and exclude features based on the direction of the search. The Sequential Floating Forward Selection (SFFS) is the floating method in the forward direction and Sequential Floating Backward Selection (SFBS) is the search method in the opposite direction. In random search algorithms, randomly selects subset and randomly insert or delete feature sets (Pantic and Rothkrantz 2003; Jaimes and Sebe 2007). Evolutionary algorithms are comes under random search algorithms which are used for feature selection (Sedaaghi et al. 2007).

In this paper the sequential forward selection and sequential floating forward selection are used for feature selection in order to maximize the emotion classification performance with a low dimensionality feature vector.

4 Two stage feature selection method for speech emotion recognition

The Block diagram of the proposed two stage feature selection method is shown in Fig. 2. It shows the overall process of how the speech samples are processed in different stages to recognize the emotional state of the person. Initially speech samples are preprocessed before the feature extraction. In the first stage of feature selection, the appropriate features which best classify the emotion are selected and are fused together to enhance the performance of the system. In the second stage, for further improvement

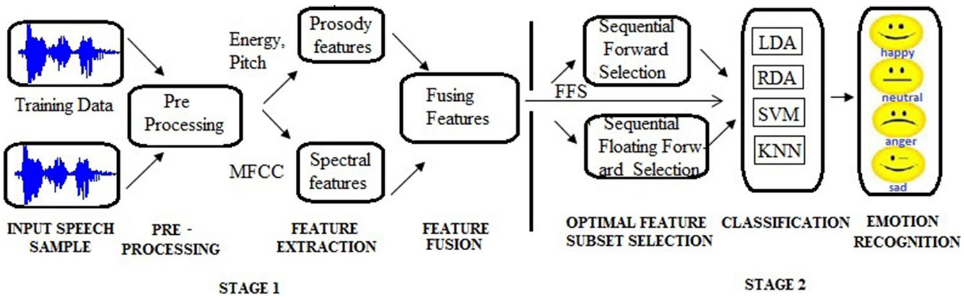


Fig. 2 Block diagram for speech emotion recognition system using two stage feature selection. STAGE1-Classification using feature fusion set STAGE2-Classification using optimal feature set selection

in the performance of the system, an optimal feature subset selection techniques are used which reduce the dimensionality of the fused feature vector by selecting an optimal feature set and are given as input to one of the classifiers to recognize the emotional state of the speech sample.

4.1 Pre processing

Filtering, framing and windowing are comes under the task of preprocessing of speech samples. A high pass filter is used to reduce the environmental noise while recording the speech sample. In framing, speech signal is split into several frames with 256 samples for each frame and an overlapping of 100 samples is done with a hamming window and a feature vector is extracted for each frame. This feature vector is used to classify the emotion of the speech sample based on some simple statistics like mean, variance, minimum, range, skewness and kurtosis because they are less sensitive to the linguistic data (Ververidis and Kotropoulos 2008). The detailed analysis of framing and windowing are given in Anne et al. (2015).

4.2 First stage of feature selection

An important module in the speech emotion recognition system is the selection of best features because there is no theoretical basis about the features which best classify the emotion. So the work is based on the features obtained from direct comparison speech signals portraying different emotions. This comparison is useful for identifying the best features which are useful for emotion identification. In this section the features like energy, pitch, and melfrequency cepstral coefficients which contains most of the emotion specific information based on the literature are used (Kuchibhotla et al. 2014a).

4.3 Feature extraction

Feature extraction includes extraction of feature vectors from the speech sample. The feature vectors are generated for each frame. Energy and Pitch are the most emotion specific prosodic features which are extracted from the speech sample. The information provided by the first and second order derivatives of these features are also considered. The amplitude variations of the speech signal are used to calculate the energy and the auto correlation method is used to calculate the pitch. The short time energy and pitch features for each frame are given in Eqs. 1 and 2

$$E = \sum_{i=1}^{N} x_i^2 \tag{1}$$

$$R_n(k) = \sum_{i=1}^{N} x(i)x(i+k) \tag{2}$$

where x_i is the speech sample for the ith frame, k is the time lag and N is the total number of frames. Totally 18 values are extracted for each feature including the corresponding feature and their first, second derivatives by using the statistics. As a final 36 values are extracted for both energy and pitch features.

In order to extract the correct emotional state of the speech sample the most efficient spectral representation of the speech sample are Mel Frequency Cepstral Coefficients (MFCC) (Luengo et al. 2010). The Mel-frequency scale is given in in Eq. 3

$$mel(f) = 2595 \times \log 10\left(1 + \frac{f}{100}\right) \tag{3}$$

where f is normal frequency and mel(f) is mel frequency for a given f. The procedure for implementing the MFCC are given in detail in Sato and Obuchi (2007), Vankayalapati HD (2011), Vankayalapati et al. (2010). Eighteen MFCC Coefficiens were estimated along with their first and second derivatives for each and every frame which gives a total of 54 spectral features. All these features are estimated over simple six statistics, so totally 324 spectral features are extracted for each speech sample.

Initially different experiments are conducted using this prosody and spectral features individually. But the performance of the system degrades with this technique. So feature fusion technique is applied to improve the performance of the speech emotion recognition system.

4.4 Feature fusion

In this step the extracted 36 energy, pitch features and 324 mel frequency cepstral coefficient features are fused together and obtain 360 features totally. All these features are assigned to a classifier for recognition of emotion of speech samples. This will give better results than working with individual features (Kuchibhotla et al. 2014a). Even though it gives better results it should not lead to an optimal state because of high dimensional feature vector.

4.5 Second stage of feature selection

To reduce the curse of dimensionality problem that occur in first stage feature selection and to obtain better performance of speech emotion recognition system, second stage feature selection is required. This stage includes extraction of optimal feature subset from the fused feature set. Basically feature selection is the process of finding a subset of n features from a given set of N features i.e $n<N$ with out significantly degrading the performance of the classifier. In feature subset selection, each feature is assigned a value to reflect its usefulness. The detailed description of feature subset selection technique is explained below.

The feature vector extracted from the fusion of energy, pitch and melfrequency cepstral coefficients of the speech samples are given as input to the feature subset selection technique. Here we used two types of feature subset selection techniques viz., sequential forward selection (SFS) and sequential floating forward selection (SFFS) individually. The later is an extension of first technique. So the procedure for the selection of subset of features from the fused features using SFFS is shown in Fig. 3.

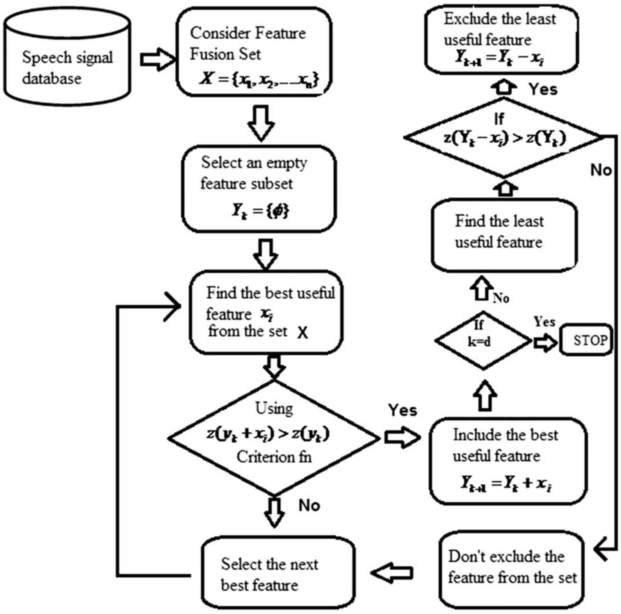

Speech signal database
Consider Feature Fusion Set $X=\{x_1, x_2, \ldots x_n\}$
Select an empty feature subset $Y_k=\{\phi\}$
Find the best useful feature x_i from the set X
Using $z(y_k+x_i)>z(y_k)$ Criterion fn
Yes
No
Select the next best feature
Include the best useful feature $Y_{k+1}=Y_k+x_i$
If k=d
Yes
STOP
No
Find the least useful feature
If $z(Y_k-x_i)>z(Y_k)$
Yes
No
Exclude the least useful feature $Y_{k+1}=Y_k-x_i$
Don't exclude the feature from the set

Fig. 3 Chart for sequential floating forward selection algorithm

4.5.1 *Sequential forward selection (SFS)*

SFS starts with an empty set, and the feature set is iteratively updated by adding the most significant feature by using a criterion function in each step. The criterion function used here is the unweighted average recall. If that selected feature satisfies the criterion function then it should be included into the updated set else it searches for the next best feature from the fused feature set. The functionality of the criterion function is, to check the performance of the classifier with the newly updated feature set with the old one. The efficiency of the classifier is done by using predicted labels and actual labels of the speech samples. If we get the better performance then only the most significant feature should be included into the feature set else it should not be added into the feature set. In this way to add most significant feature into the updated set this process repeats with each feature continuously till we get the optimal feature subset. But the drawback of this method is once a feature is added into the feature set it should not be possible to remove that feature from the set. This is called nesting effect. The problem with this nesting effect is eliminated by using sequential floating forward selection.

4.5.2 *Sequential Floating Forward Selection (SFFS)*

It includes new features by using basic SFS procedure followed by successive conditional deletion of least significant feature in the updated set which provides a better feature subset. Initially a feature vector is generated from fused features. In order to get the best useful features from this feature vector mainly three steps are used. First it starts with an empty set Y = 0 and the Sequential Forward Selection method is used to select the most significant feature from the feature vector and includes it into the set Y. If the newly added set satisfies the criterion function then keep that feature into the set else select the next best feature from the feature vector and add into the set. The second step is to find the least significant feature from the newly added set. If the deletion of the least significant feature satisfies the criterion function then exclude that feature from the set else continue with the forward selection. If the deletion does n't satisfies the criterion function then continue with conditional exclusion and is the final step in this procedure. The criterion function is calculated by using the formula given in Eq. 4 All this process is shown clearly with in the Fig. 3.

$$\text{Accuracy} = (\text{sum}(\text{N correct samples})/\text{sum}(\text{N instances})) \times 100. \quad (4)$$

4.6 Classification

In this paper various classification algorithms are used for classification namely Linear Discriminant Analysis, Regularized Discriminant Analysis, Support Vector Machine and K-Nearest Neighbour. The description of each algorithm is given in detail in Kuchibhotla et al. (2014a). The emotional classes used in this paper are happy, neutral, anger, sad, fear and disgust. The way the class label assigned to each test speech sample is depends on the minimum of the Euclidian distance of the training samples. Each classifier is given an enough training data, for better classification of test speech samples. LDA suffers from singularity problem because of high dimensional and low sample size. So it is difficult to get the accurate results with LDA. This singularity problem is eliminated by RDA with a regularization technique with which the performance of the classifier improves. KNN basically does not deal with feature relevance but SVM and RDA can better deal with high dimensionality and irrelevant features. These things justify that RDA and SVM better classifies the speech samples. The concept of two stage feature selection will plays a major role here which improves the performance of each classifier effectively. The performance of the classification method highly depends here on the quality of the feature set.

5 Speech emotional databases

In general research with emotional speech samples deals with acted, induced and completely spontaneous database of emotions (Kuchibhotla et al. 2014b; Vogt et al. 2008). More number of emotional speech databases are designed. Some of them are Emo-DB (Berlin emotional speech database), Danish database of emotional speech (DES), Spanish emotional Speech database (SES), Chinese and English emotional speech databases. In this work the experiments are conducted over Berlin and Spanish emotional databases which comes under acted emotional speech databases.

Berlin database is an open source, simulated emotional speech database which contains 7 basic emotions anger, boredom, disgust, fear, happiness, sadness and neutral etc. There are totally 535 different German emotional speech samples. which are simulated by 10 professional native German actors (5 actors and 5 actresses) (Burkhardt et al. 2005) (Table 1).

Spanish database contains seven emotions anger, sadness, joy, fear, disgust surprise and neutral. There are 184 sentences for each emotion which include isolated words, sentences and a text passage. There are totally 1288 Spanish emotional speech samples. These emotional speech samples are recorded by one professional male speaker and one professional female speaker (Hozjan et al. 2002; Kuchibhotlaa et al. 2015).

6 Experimental evaluation

A two stage feature selection for speech emotion classification is applied individually for Berlin database and Spanish database with individual feature selection techniques like sequential forward selection (SFS) and sequential floating forward selection (SFFS), and the results are compared effectively. Databases are divided into training and testing set. 2/3rd of the whole data samples is used for training and 1/3rd of the samples are used for testing. Initially each classifier is trained with the data provided in training set. Test speech sample is classified by using a classifier and the information provided by the training speech samples. In order to validate the results of different emotional speech samples on various classifiers like LDA, RDA, SVM and KNN the experiments were conducted in two phases. First phase is Base line results with feature fusion and the second phase is results with optimal feature subset selection techniques.

6.1 Base line results

The task of feature fusion is to combine the pitch,energy and MFCC features extracted from emotional speech samples. These fused features are classified by several classification techniques and the results are shown in Table 2. According to these results the performance of LDA and KNN are in the range of 50–60 %. Even though the performance of the RDA and SVM classifiers is between 60–70 % it does not reach to an optimal state because of the high dimensional feature vector. So for

Table 1 Description of Type of files, No. of files and corresponding number of emotions in Berlin and Spanish databases

Type of files	No. of files in database	
	Berlin	Spanish
Training samples	357	859
Test samples	178	429
Total samples	535	1288
Emotions	7	7

Table 2 Emotion recognition percentage accuracy of various classifiers (LDA, RDA, SVM and KNN) over Berlin and Spanish databases using feature fusion

Classifier	Feature fusion set (FFS)	
	Berlin (%)	Spanish (%)
LDA	57.2	52
RDA	73.8	69.4
SVM	70.5	67.6
kNN	65	60.8

further improvement in the performance of all the classifiers, a two stage feature selection technique is needed.

6.2 Results with optimal two stage feature selection techniques

Initially a set of fused features are collected by combining the energy, pitch and mel frequency cepstral coefficient features extracted from speech sample. In order to get the best useful features from this fused features a feature subset selection algorithm is applied. This optimal feature set leads to an effective improvement in the performance of the classifiers when compared with base line results and are shown in Table 3. Here we used sequential forward selection (SFS) and sequential floating forward selection (SFFS) optimal feature subset selection techniques and LDA, RDA, SVM and KNN classification techniques and the experiments are evaluated on Berlin, Spanish emotional speech corpora and the results are compared effectively with each classifier. From the Table 3 it is observed that SFFS gives better emotional recognition performance than with SFS. It is also observed from the table that RDA and SVM performs considerably better when compared with the remaining classifiers.

The overall recognition performance of an RDA classifier with SFS and SFFS when compared with feature fusion set (FFS) is shown in Fig. 4. The horizontal axis represents the name of the feature selection method and the vertical axis represents the performance of the classifier. From the graph it is observed that the performance of the classifier is effectively improved by 20 % approximately with SFFS when compared with baseline results. This proves that two stage feature selection is a best technique for improving the performance of the classifier.

Table 3 Emotion recognition percentage accuracy of various classifiers (lda, rda, svm and knn) over Berlin and Spanish databases using SFS and SFFS feature selection methods

Classifier	Berlin		Spanish	
	SFS (%)	SFFS (%)	SFS (%)	SFFS (%)
LDA	65	72.5	63.3	69.4
RDA	85.2	92.6	83.6	90.5
SVM	80.4	88.1	77.3	86.2
KNN	73.8	81.1	74.5	78.03

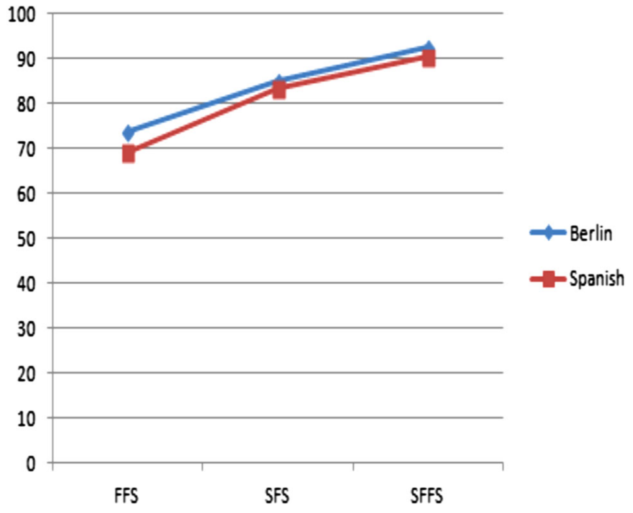


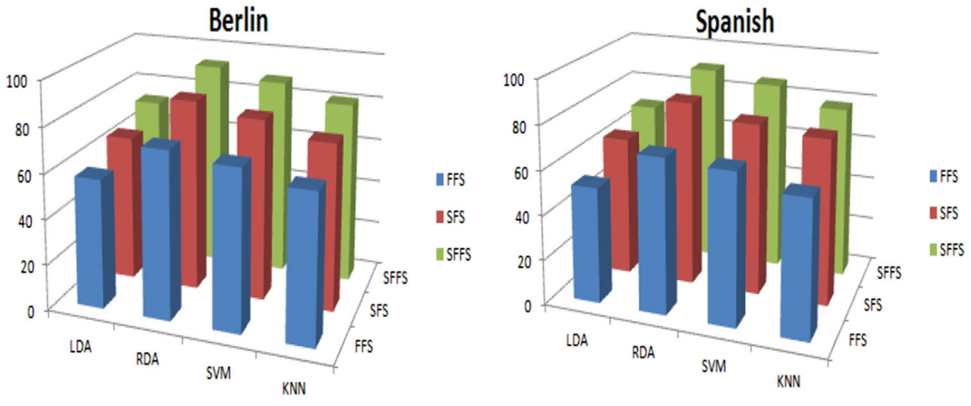
Fig. 4 Comparison of emotion recognition performances of feature subset selection algorithms using Berlin and Spanish databases

6.3 Analysis of results with each emotion using various classifiers

The emotions and the classifiers considered in this paper are happy, neutral, anger, sad, fear, disgust and LDA, RDA, SVM and KNN. The results are analysed with each emotion using various classifiers individually in both Berlin and Spanish databases. The recognition accuracy of each classifiers on each emotion using SFS and SFFS are shown in Tables 4 and 5. The left column of both the tables represent the name of the classifier and the title of the row represents the name of the feature selection method and the name of the emotion. Each cell represents the emotion recognition accuracy of the corresponding classifier. Even though LDA suffers with singularity problem i.e., the number of speech samples is less than that of the dimension of the feature set it does not perform so poor and its recognition accuracy is nearly reached 70 %. This problem is eliminated by RDA by using a regularization technique with which it reaches an accuracy of 90 %. In a similar way the next highest emotion recognition accuracy is with SVM. KNN is also gives a very good emotion recognition performance with an accuracy of nearly 80 %.

It is also observed from the Tables 4 and 5 that SFFS shows the highest recognition accuracy than that of the SFS which is because SFS suffers from nesting problem i.e ones the feature is recovered it should not be eliminated. This nesting problem is eliminated with SFFS by removing the unwanted features with which it leads to an effective improvement in the performance of the classifier.

Fig. 5 Comparison of emotion recognition performances of various feature subset selection algorithms using **a** Berlin database **b** Spanish database



(a) (b)

Table 4 Recognition accuracy percentage for emotions with various classifiers in Berlin database using feature selection algorithms

Berlin	Emotion					
Algorithm	H	N	A	S	F	D
(a) SFS-sequential forward selection						
LDA	65	63	72	63	65	62
RDA	87	80	88	92	81	83
SVM	82	79	83	78	80	88
kNN	77	84	67	77	73	67
(b) SFFS-sequential floating forward selection						
LDA	72	82	67	75	71	68
RDA	95	93	96	92	89	91
SVM	90	88	91	90	84	86
kNN	90	87	83	70	80	77

H happy, *N* neutral, *A* anger, *S* sad, *F* fear and *D* disgust

We obtain an optimal feature set with SFS and SFFS which are selected from a set of 360 features which is obtained by fusing energy, pitch and MFCC features extracted from the speech sample. The Table 6 shows the best feature combination with SFFS for both databases. A set of 12 features are obtained for Berlin and 18 features are obtained for Spanish.

The graphical representation of efficiency of each classifier with each and every feature selection technique is shown in Fig. 5. The blue, red and green bars represent the results with full feature set (FFS), sequential forward selection (SFS) and sequential floating forward selection (SFFS)respectively. Among all the bars the green bars belongs to the classifiers RDA and SVM shows the efficient performance by using SFFS in both the databases.

6.4 Analysis of results with Receiver Operating Characteristic (ROC) Curves

Table 5 Recognition accuracy percentage for emotions with various classifiers in Spanish database using feature selection algorithms

Spanish	Emotion					
Algorithm	H	N	A	S	F	D
(a) SFS-sequential forward selection						
LDA	66	60	69	58	59	68
RDA	85	80	87	89	79	82
SVM	80	75	80	73	77	79
kNN	70	70	73	83	90	60
(b) SFFS-sequential floating forward selection						
LDA	73	77	70	69	71	68
RDA	92	88	94	89	95	85
SVM	89	83	89	92	83	82
kNN	83	77	70	69	71	60

H happy, *N* neutral, *A* anger, *S* sad, *F* fear and *D* disgust

Table 6 Optimal feature subset extracted by using SFS and SFFS for both the databases

SFFS	Best feature combination
Berlin (12 features)	12 23 50 84 30 326 34 281 17 156 189 4

Spanish (18 features)	21 9 78 92 58 24
	43 22 98 301 252 120
	301 15 260 19 4 56

The Experimental results are analysed with receiver operating characteristic curves. The results obtained with these ROC curves shows that, the performance of the classifier with feature subset selection techniques is approximately equal with that of the results obtained using ROC Curves. The accuracy, sensitivity, specificity and Area Under Curve are the elements extracted from the ROC curves. The detailed analysis of these ROC curves are given in Fawcett (2006), Kuchibhotla et al. (2014a). The shape of the curve estimates power of the feature selection algorithm. The value of area under curve should lies between 0 and 1. The more the area under curve, the more the

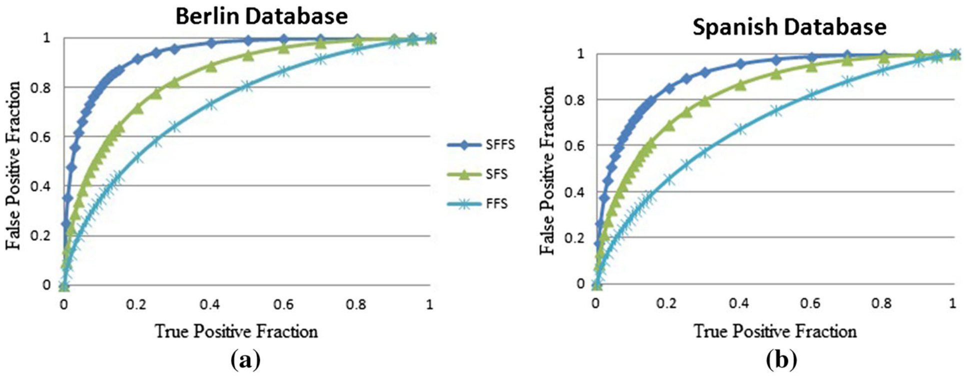


Fig. 6 Comparison of emotion recognition performances of various feature subset selection algorithms using **a** Berlin database **b** Spanish database

Table 7 Shows the values (Accu:Accuracy, Sens:Sensitivity, Spec:Specificity, AUC: Area Under Curve)extracted from ROC plot for different feature subset selection algorithms for (a) Berlin database and (b) Spanish database

Database	Accu (%)	Sens (%)	Spec (%)	AUC
a				
FFS	73.8	74.2	73.4	0.734
SFS	85.2	85	85.3	0.845
SFFS	92.7	94.4	91	0.937
b				
FFS	69.5	70.1	68.9	0.692
SFS	83.7	83.9	83.4	0.831
SFFS	90.5	90.1	89.8	0.908

performance of the technique. The ROC curves are drawn for each feature selection method including both the databases for all the classifiers. Here we are giving a set of ROC curves which are drawn for the classifier RDA and the corresponding graphs are shown in Fig. 6. The area under curve is more for Sequential floating forward selection algorithm for both the databases in all the classifiers.

The values extracted from the ROC curves are given in Table 7. The area under curve for SFFS is 0.824 for Berlin database which shows that the performance of the classifier is very good by using SFFS as feature selection algorithm. Similarly the area under curve for SFFS is 0.774 for Spanish database which shows that the performance of the classifier is good. From the literature survey (Šimundić 2008) the area under curve and its diagnostic accuracy is shown in Table 8.

Table 8 "The relationship between area under curve and its diagnostic accuracy" Table 2 in Šimundić (2008)

Area under curve	Diagnostic accuracy
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Sufficient
0.5–0.6	Bad
<0.5	Test not useful

7 Conclusion

The main objective of the proposed two stage feature selection method is to reduce the dimension of the fused feature vector and to improve the performance of the classifier. The high dimensional fused feature vector contains some irrelevant features which have very less emotion specific information. This two stage feature selection method eliminate such type of features and generates a new feature vector with features, which are less in number and are more in emotion specific content. The emotion recognition accuracy of the classifier is effectively improved with this feature vector. The experiments are conducted over emotional speech samples of Berlin and Spanish databases and are systematically evaluated by using various feature selection techniques and several classification methods. An effective comparative study has also been done with feature selection and without feature selection using each and every classifier used in this work. The experimental results showed that the classifiers RDA and SVM with SFFS gives best emotion recognition performance and also the recognition accuracy of KNN and LDA are improved when compared with base line results. The results also shows that the recognition accuracy is improved by 15–20 % approximately with each classifier and they also reveal that SFFS is a better choice as a feature subset selection technique because it eliminates the nesting problem that occurred in SFS.

8 Future work

As a future work, the Advanced Driver Assistance Road Safety system (ADARS) is considered. It is a primitive level of developing application in which the safety of the driver is given the highest priority. The main objective of this ADARS system is to reduce the number of accidents by detecting the emotion of the driver and to make him an alert from an accident at an appropriate time. In this process of capturing the real time driving information and speech samples of the driver, pre processing is needed because the driver's voice is mixed with the noise around the system which is due to music system and communication with mobile and co-passengers. The driving behaviour changes with the emotions that occur during driving. Finally the future work includes the collection and extraction of the emotional state of the driver speech sample during driving and alert him through an alarm at an appropriate time which definitely avoids an accident and save the driver's life.

References

Anne, K. R., Kuchibhotla, S., & Vankayalapati, H. D. (2015). *Acoustic modeling for emotion recognition*. Berlin: Springer.

Bou-Ghazale, S. E., & Hansen, J. H. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *8*(4), 429–442.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of german emotional speech. *Interspeech*, *5*, 1517–1520.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1), 5–32.

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

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* (Vol. 57). Boca Raton, FL: CRC Press.

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

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Fernandez, R. (2003). A computational model for the automatic recognition of affect in speech. PhD thesis, Massachusetts Institute of Technology

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., & Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. In *LREC*

Iohnstone, T., & Scherer, K. (2000). *Vocal communication of emotion. Handbook of emotion* (pp. 220–235). New York: Guilford.

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1), 116–134.

Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158.

John, G. H., Kohavi, R., Pfleger, K., et al. (1994). Irrelevant features and the subset selection problem. *ICML*, *94*, 121–129.

Kim, S., Georgiou, P. G., Lee, S., & Narayanan, S. (2007). Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *MMSP 2007, IEEE 9th Workshop on Multimedia Signal Processing* (pp. 48–51). IEEE.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1), 273–324.

Kuchibhotla, S., Vankayalapati, H., Vaddi, R., & Anne, K. (2014a). A comparative analysis of classifiers in emotion recognition through acoustic features. *International Journal of Speech Technology*, *17*(4), 401–408.

Kuchibhotla, S., Yalamanchili, B., Vankayalapati, H., & Anne, K. (2014b). Speech emotion recognition using regularized discriminant analysis. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013* (pp. 363–369). Springer.

Kuchibhotlaa, S., Vankayalapati, H. D., Yalamanchili, B., & Anne, K. R. (2015). Analysis and evaluation of discriminant analysis techniques for multiclass classification of human vocal emotions. In *Advances in Intelligent Informatics* (pp. 325–333). Springer.

Kwon, O. W., Chan, K., Hao, J., & Lee, T. W. (2003). Emotion recognition by speech signals. In *INTERSPEECH*. Beijing: Citeseer.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491–502.

Luengo, I., Navas, E., & Hernáez, I. (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, *12*(6), 490–501.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*, 1097.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech Communication*, *41*(4), 603–623.

Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390.

Pohjalainen, J., Räsänen, O., & Kadioglu, S. (2013). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language*, *29*(1), 145–171.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125.

Sato, N., & Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, *2*(3), 835–848.

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

Sedaaghi, M. H., Ververidis, D., & Kotropoulos, C. (2007). Improving speech emotion recognition using adaptive genetic algorithms. In *Proceedings of European Signal Processing Conference (EUSIPCO)*.

Šimundić, A. M. (2008). Measures of diagnostic accuracy: Basic definitions. *Medical and Biological Sciences*, *22*(4), 61–65.

Somol, P., Pudil, P., Novovičová, J., & Paclık, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, *20*(11), 1157–1163.

Tao, J., & Tan, T. (2005). Affective computing: A review. In *International Conference on Affective computing and intelligent interaction* (pp. 981–995). Springer.

Vankayalapati, H., Anne, K., & Kyamakya, K. (2010). Extraction of visual and acoustic features of the driver for monitoring driver ergonomics applied to extended driver assistance systems. In *Data and Mobility* (pp. 83–94). Springer.

Vankayalapati, H. D., Siddha, V. R., Kyamakya, K., & Anne, K. R. (2011). Driver emotion detection from the acoustic features of the driver for real-time assessment of driving ergonomics process. *International Society for Advanced Science and*

Technology (ISAST) Transactions on Computers and Intelligent Systems, *3*(1), 65–73.

Ververidis, D., & Kotropoulos, C. (2006). Fast sequential floating forward selection applied to emotional speech features estimated on des and susas data collections. In textit2006 14th European on Signal Processing Conference

Ververidis, D., & Kotropoulos, C. (2008). Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition. *Signal Processing*, *88*(12), 2956–2970.

Vogt, T., André, E, & Wagner, J. (2008). Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. In *Affect and Emotion in Human-Computer Interaction* (pp. 75–91). Springer.

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, *100*(9), 1100–1103.

Williams, C. E., & Stevens, K. N. (1981). Vocal correlates of emotional states. In J. K. Darby (Ed.), *Speech evaluation in psychiatry* (pp. 221–240). New York: Grune & Stratton.