A comparative analysis of classifiers in emotion recognition through acoustic features

**Swarna Kuchibhotla,
H. D. Vankayalapati, R. S. Vaddi &
K. R. Anne**

International Journal of Speech Technology

ISSN 1381-2416
Volume 17
Number 4

Int J Speech Technol (2014) 17:401-408
DOI 10.1007/s10772-014-9239-3

Springer

Int J Speech Technol (2014) 17:401–408
DOI 10.1007/s10772-014-9239-3

A comparative analysis of classifiers in emotion recognition through acoustic features

Swarna Kuchibhotla · H. D. Vankayalapati · R. S. Vaddi · K. R. Anne

Received: 23 December 2013 / Accepted: 19 May 2014 / Published online: 15 June 2014

Abstract The most popular features used in speech emotion recognition are prosody and spectral. However the performance of the system degrades substantially, when these acoustic features employed individually i.e either prosody or spectral. In this paper a feature fusion method (combination of energy,pitch prosody features and MFCC spectral features) is proposed. The fused features are classified individually using linear discriminant analysis (LDA), regularized discriminant analysis (RDA), support vector machine (SVM) and k nearest neighbour (kNN). The results are validated over Berlin and Spanish emotional speech databases. Results showed that,the performance is improved by 20 % approximately for each classifier when compared with performance of each classifier with individual features. Results also reveal that RDA is a better choice as a classifier for emotion classification because LDA suffers from singularity problem, which occurs due to high dimensional and small sample size speech samples i.e the number of available training speech samples is small compared to the dimensionality of the sample space. RDA eliminates this singularity problem by using regularization criteria and give better results.

S. Kuchibhotla (✉)
Acharya Nagarjuna University, Namburu, Gunter Dt, Andhra Pradesh, India
e-mail: swarna.kuchibhotla@gmail.com

H. D. Vankayalapati · R. S. Vaddi · K. R. Anne
V. R. Siddhartha Engineering College, Kanuru, India
e-mail: nanideepthi@gmail.com

R. S. Vaddi
e-mail: syam.radhe@gmail.com

K. R. Anne
e-mail: raoanne@gmail.com

Keywords Emotion recognition · Feature fusion · Classification

1 Introduction

In recent years the need of communication interface between humans and machines is become essential. A lot of research work has been done since late 1950s by providing sufficient knowledge to the machine to recognize human voice (Cowie et al. 2001; Scherer 2003; Koolagudi and Rao 2012). Even though a great progress was made in speech and speaker recognition, still there is a lack of naturalness in identifying speaker emotions (Vogt et al. 2008; Ayadi et al. 2011). This has become challenging task to the researchers working in this field because the machine should interpret the emotional state of humans and respond properly without disturbing the human. The extensive application areas of speech emotion recognition are education, entertainment, medical diagnosis etc. In this paper, we are particularly interested in road safety by detecting the emotional state of the driver and alert him according to his emotion. This is all done by first selecting features from their speech samples. Different emotional speech samples of the drivers according to their emotional status are as shown in Fig. 1.

The features extracted from the speech sample contains most of the emotion specific information and are mainly classified as two types prosody and spectral. The differences between these two features are shown in Table 1.

The extracted features from the speech samples are given as input to the classifier. Different types of classifiers used by the researchers are Gaussian mixture models (GMM) by El Ayadi et al. (2007), hidden Markov models (HMM) by Nwe et al. (2003), Schuller et al. (2003), Neural Networks (NN) by joy Nicholson et al. (2000), k-nearest neighbour (kNN) by Ravikumar and Suresha (2013), regularized dis-

Fig. 1 speech samples of different emotions when the driver is in different states like (**a**) happy emotion, (**b**) neutral emotion, (**c**) anger emotion, (**d**) sad emotion

criminant analysis (RDA) by Ye et al. (2006) and support vector machine (SVM) by Koolagudi et al. (2011). Among these the classifiers considered in this paper are LDA, RDA, SVM and kNN.

Most traditional systems in speech emotion recognition have been focused on either prosody or spectral features (Ververidis and Kotropoulos 2006; Tato et al. 2002). But the recognition accuracy is not good enough by using any one of them alone (Zhou et al. 2009). The innovative step in this direction is to fuse both the spectral and prosody features to improve the emotion recognition accuray (Zhou et al. 2009). This paper focuses more on implementation and analysis of results with both feature fusion and individual features. All this is done with those four classifiers specified earlier on both Berlin and Spanish databases.

This paper is organized as follows: Sect. 2 describes the description of databases used in the experiments. Section 3 describes the methodology and the extracted features followed by description of different classifier techniques used in this paper. Section 4 describes the experimental results and their extensive comparative study of the classifiers on different databases and Sect. 5 concludes the paper.

2 Speech databases

In this paper the experiments are conducted over Berlin and Spanish emotional speech data bases. The reason for choosing these databases is both are multi speaker databases, so it is possible to perform speaker independent tests. The databases contains the following emotions: anger, boredom, disgust, fear, happiness, sadness and neutral. The Berlin database contains 500 speech samples and are simulated by ten professional native German actors, five male and five female (Burkhardt et al. 2005). The number of speech files are as shown in Table 2.

Table 2 Description of number of files in Berlin and Spanish databases corresponding to each emotion

Emotion	No.of files in database	
	Berlin	Spanish
Anger	127	184
Boredom	81	184
Happy	71	184
Fear	69	184
Sad	62	184
Disgust	46	184
Neutral	79	184
Total	535	1288

The present work demonstrate on these databases

Spanish database contains 184 sentences for each emotion which include numbers, words, sentences etc. as shown in Table 2. The corpus comprises of recordings from two professional actors, one male and one female. Among 184 files, 1–100 are Affirmative sentences, 101–134 are Interrogative sentences, 135–150 Paragraphs, 151–160 Digits, 161–184 Isolated words.

3 Methodology

The Block diagram of overall process of Speech emotion recognition system is shown in Fig. 2. It shows the procedure of how the input speech sample is processed to get the correct emotion of the human. Test speech sample is classified by using one of the classifier and the information provided by the training speech samples.

3.1 Pre processing

Speech signals are normally pre-processed before features are extracted to enhance the accuracy and efficiency of the

Table 1 The two most efficient features of a speech signal are prosody and spectral

Prosody features	Spectral features
Occurs when we put sounds together in a connected speech	Extracted from spectral content of the speech signal
Influenced by vocal fold activity	Influenced by vocal tract activity
Discriminate high arousal emotions (happy, anger) with low arousal emotions (sadness, boredom) (Luengo et al. 2010; Scherer 2003)	Discriminate emotions in the same arousal state (high or low) (Luengo et al. 2010; Scherer 2003)
Examples: energy, pitch, speaking rate etc.,	Examples: MFCC, LPCC, LFPC, formants etc.,

The differences between these two features are described in this table

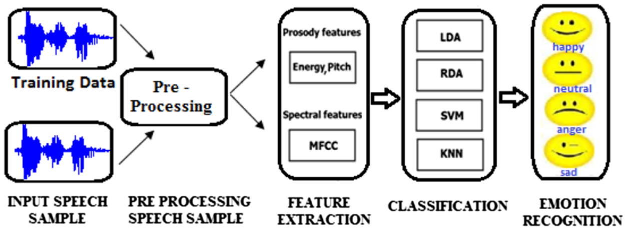

Fig. 2 Various stages in extracting emotions from a speech sample like pre processing, feature extraction, classification and emotion recognition

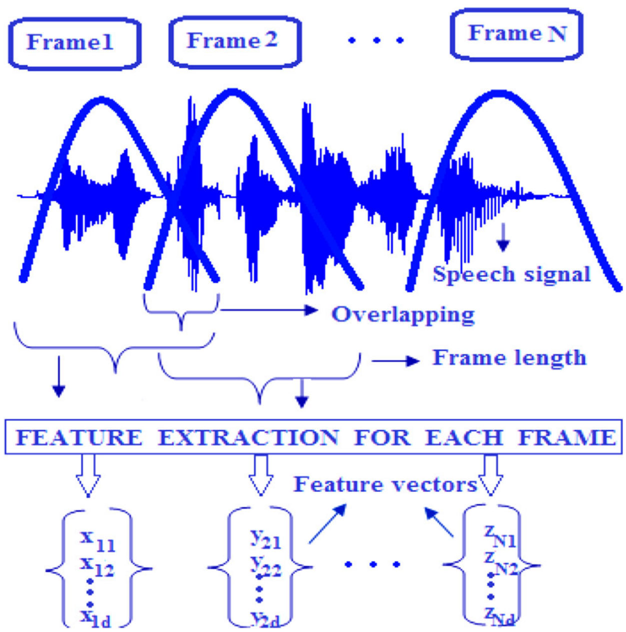

Fig. 3 The process of splitting the speech signal into several frames, applying an hamming window for each frame and its corresponding feature vector

feature extraction process. The modules filtering, framing and windowing are considered as steps under preprocessing.

The filtering technique is applied to reduce the noise, which is occurred due to the environmental conditions while recording the speech sample. This is done by using high pass filter.

Instead of analyzing the entire signal at once, split the signal into several frames with 256 samples for each frame. A feature vector is created for each frame. While dividing the continuous speech signal into frames some discontinuity in the speech sample is observed. To maintain the continuity of the speech sample an overlapping of 100 samples is applied to these frames with a hamming window. The procedure for implementing framing and windowing is shown in Fig. 3 .

3.2 Feature extraction

An important module in the design of speech emotion recognition system is the selection of features which best

Table 3 The statistics and their corresponding symbols used in this paper

Statistics	Symbol
Mean	E
Variance	V
Mininum	Min
Range	R
Skewness	Sk
Kurtosis	K

classify the emotions. In this paper some of the key prosody and spectral features are used. Those features distinguishes the emotions of different classes of speech samples. These are estimated over simple six statistics as shown in Table 3

The prosodic features extracted from the speech signal are Energy and Pitch. Their first and second order differentiation provides new useful information hence the information provided by their derivatives also considered (Luengo et al. 2005). Energy and pitch were estimated for each frame together with their first and second derivatives, providing six features per frame and applying statistics as shown in Table 3 giving a total of 36 prosodic features.

Mel frequency cepstral coefficients (MFCC) are the most widely used spectral representation of speech. These are most efficient in order to extract correct emotional state of the speech sample (Luengo et al. 2010). MFCC features represents the short term power spectrum of a speech signal, based on a linear cosine transform of a log power spectrum on a non linear melscale of frequency. The procedure for implementing MFCC is shown in Sato and Obuchi (2007), Vankayalapati and SVKK Anne (Vankayalapati and SVKK Anne). Similar to prosody statistics, spectral statistics are calculated using the statistics as shown in the Table 3. The extracted MFCC are 18 and the number of filter banks used are 24. Eighteen MFCC Coefficients and their first and second derivatives are estimated for each frame giving a total of 54 spectral features. The statistics as mentioned earlier are applied to these 54 values so totally $54\times6=324$ different features are calculated. The speech signal and its spectral and prosody features are shown in Fig. 4.

3.3 Classification

The features extracted from the feature extraction module are given as input to the classifier. The purpose of classifier is to optimally classify the emotional state of a speech sample. The speech samples considered in this paper belongs to four(happy, neutral, anger and sad) emotional classes of two (Berlin, Spanish) different databases. Among the speech

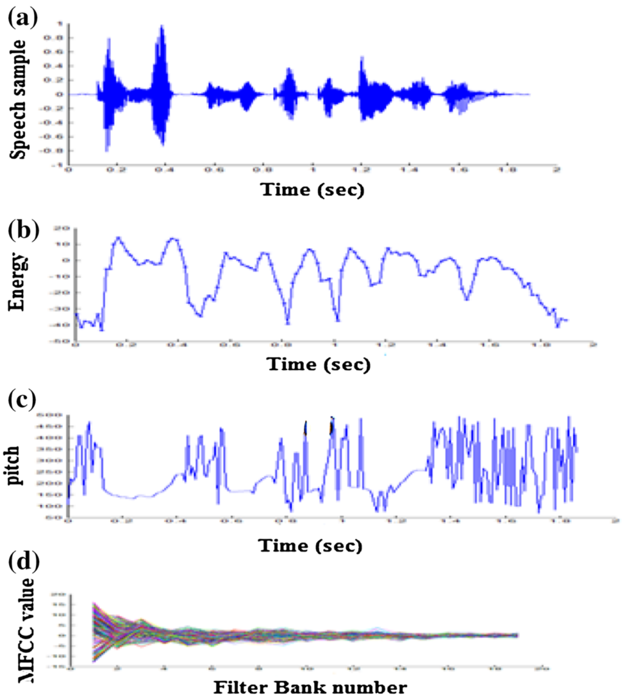

Fig. 4 Prosody and spectral features of the speech signal. **a** Original speech signal, **b** Energy contour, **c** Pitch track, **d** Mel frequency cepstral coefficients

samples $\frac{2}{3}$rd of the samples are used for training and $\frac{1}{3}$rd of the samples are used for testing for each database. The classifiers are trained using training set and the classification accuracy is estimated on test set. The training and test samples are chosen randomly. Once the testing is completed on different test samples the results are gathered and are used to calculate the emotion recognition accuracy of each classifier. The way the classifiers used in this paper are described as follows.

3.3.1 Linear discriminant analysis (LDA)

LDA is a well known statistical method for classification as well as dimensionality reduction of speech samples. It computes an optimal transformation matrix by maximizing the between class covariance matrix and minimizing the within class covariance matrix of the speech samples. In other words, it groups the wav files belongs to one class and separates the wav files belongs to different emotional classes. A class is a collection of speech samples belonging to the same emotion. When dealing with high dimensional and low sample size speech data, LDA suffers from singularity problem, Ji and Ye (2008). The consequences of this problem are we may not get the accurate results. To improve the results Regularized Discriminant Analysis is applied on these speech samples.

3.3.2 Regularized discriminant analysis(RDA)

The key idea behind the regularized discriminant analysis is to add a constant to the diagonal elements of with in class scatter matrix S_w of the speech samples is shown in Eq. 1.

$$S_w = S_w + \lambda I \tag{1}$$

where λ is regularized parameter which is relatively small such that S_w is positive definite. In our paper the value of λ is 0.001. It is somehow difficult in estimation of regularization parameter value in RDA as higher values of λ will disturb the information in the within class scatter matrix and lower values of λ will not solve the singularity problem LDA (Vankayalapati et al. 2010; Ye et al. 2006).

The transformation matrix W is calculated by using ${S_w}^{-1} S_b W = W\lambda$. Once the transformation matrix W is given, the speech samples are projected on to this W. After projection, the Euclidian distance between each train speech sample and the test speech sample are calculated, the minimum value among them will classify the result.

3.3.3 Support vector machine (SVM)

SVM is used for speech emotion classification as it can handle linearly non separable classes. For multi class classification of speech samples one against all approach is used. A set of prosody and spectral features are extracted from the speech sample and given as input to train the SVM. Among many possible hyper planes the SVM classifier selects a hyper plane which optimally separates the training speech samples with a maximum margin. To construct an optimal hyper plane support vectors are calculated for speech samples in training phase which determines the margin (Milton et al. 2013). As higher the margin, more the speech signal classification performance and vice versa.

The decision function of a SVM to classify the speech samples is given by Eq. 2

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i yik(x_i, x) + b\right) \tag{2}$$

where $k(x_i, x) = x_i.x$ is a linear kernel and N is the number of support vectors. The dot product is applied between each test speech sample. The support vectors, alpha and bias values are obtained during the training phase. As a result we obtain four models, one for each emotional class in training phase. Basing on these models and the generated support vectors we will classify the emotion of the test speech sample.

3.3.4 KNearestNeighbor (KNN)

KNN is a non-parametric method for classifying speech samples based on closest training samples in the feature space

Table 4 Emotion recognition percentage accuracy of various classifiers (lda, rda, svm and knn) over Berlin and Spanish databases using prosody and spectral features

Classifier	Berlin		Spanish	
	Prosody (%)	Spectral (%)	Prosody (%)	Spectral (%)
LDA	42.0	51.0	40.0	49.0
RDA	67.0	78.75	44.0	61.0
SVM	55.0	68.75	60.25	64.5
kNN	57.0	60.7	58.25	67.75

Table 5 Emotion recognition percentage accuracy of various classifiers (lda, rda, svm and knn) over Berlin and Spanish databases using feature fusion

Classifier	Feature fusion	
	Berlin	Spanish (%)
LDA	62	60
RDA	80.7	74
SVM	75.5	73
kNN	72	71.5

(Ravikumar and Suresha 2013). Similar to SVM and RDA the speech samples are given as input to the KNN classifier. Nearest Neighbor classification uses training samples directly rather than that of a model derived from those samples. It represents each speech sample in a d-dimensional space where d is the number of features. The similarity of the test samples with the training samples is compared using Euclidian distance. Once the nearest neighbor speech samples list is obtained the input speech sample is classified based on the majority class of nearest neighbors.

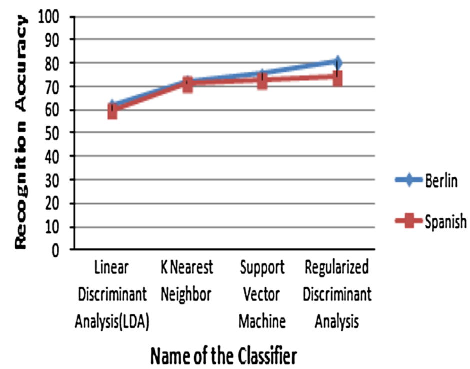

Fig. 5 Comparison of emotion recognition accuracy performance of different classifiers using Berlin database

4 Experimental evaluation

In order to validate the results of different classifiers on happy, neutral, anger and sad emotional classes, recognition tests were carried out in two phases like baseline results and feature fusion results.

4.1 Baseline results

In the first phase all the classifiers have been trained using prosody and spectral features of Berlin and Spanish emotional speech samples and the results are shown in Table 4. It seems that spectral features provide higher recognition accuracies than prosody ones. According to these results, it is observed that even though prosody parameters show very low class separability it does not perform so badly. It shows an accuracy in the range of 40– 67 % for both the databases. Results with spectral statistics seems rather modest reaching an accuracy of 50–70 % for all the classifiers, except for RDA it reaches 78.75 %. For further improvement in the performance of all the classifiers, feature fusion technique is used in the next phase.

4.2 Feature fusion results

Feature fusion is done by combining both prosody and spectral features. By doing this, the emotion recognition performance of the classifiers is effectively improved when compared with baseline results and are shown in Table 5. Among these RDA and SVM performs considerably better when compared with remaining classifiers.

The overall recognition performances of these four classifiers for Berlin and Spanish databases is shown in Fig. 5. Horizontal axis represents the name of the classifier and vertical axis represents the accuracy rate. The efficiency of each classifier is improved by 20 % approximately for each classifier when compared with base line results.This proves that feature fusion is a best technique for improving the efficiency of the classifier.

4.3 Analysis of results with each emotion

The confusion matrices of RDA classifier for both the databases is shown in the Table 6. By observing the results, we can say that the emotions happy is confused with anger and the emotion neutral is confused with sad. This occurs in all the features but the amount of variation is more by using individual features and is less by using feature fusion. The reason for this confusion is explained by using valence arousal space. Prosodic features are able to discriminating the emotions (happy, anger) from high arousal space to emotions (neutral, sad) from low arousal space, but there exists a confusion among the emotions in the same arousal state. By using

Table 6 Emotion classification performance in percentage using featurefusion technique (a) using Berlin database speech utterances, (b) using Spanish database speech utterances

Berlin	Happy	Neutral	Anger	Sad
(a)				
Happy	73	4	20	3
Neutral	9	70	–	21
Anger	14	–	83	3
Sad	–	3	–	97
Spanish	**Happy**	**Neutral**	**Anger**	**Sad**
(b)				
Happy	71	5	15	9
Neutral	–	60	14	26
Anger	22	–	67	11
Sad	1	2	–	97

Table 7 Recognition accuracy percentage for emotions(happy, neutral, anger and sad) with various classifiers using both the data bases (a) Berlin database, (b) Spanish database

Berlin	Feature fusion			
Algorithm	Happy	Neutral	Anger	Sad
(a)				
LDA	49	59	68	72
RDA	73	70	83	97
SVM	70	65	74	93
kNN	55	63	93	77
Spanish	**Feature fusion**			
(b)				
LDA	49	56	65	70
RDA	71	60	67	97
SVM	67	77	65	83
kNN	63	70	88	65

spectral features and feature fusion the confusion among the emotions in the same arousal state is reduced.

The recognition accuracy of each emotion are examined separately with all the classifiers and are shown in Table 7. The left column of the Table 4 shows the classifiers and top row shows the emotion. Each cell represents the recognition accuracy of the emotion by the corresponding classifier. With Berlin database, the emotion recognition rate of the classifiers RDA and SVM are comparable with each other for all the emotions. The emotion anger is identified more with kNN classifier for both databases.

The graphical representation of efficiencies of these classifiers is shown in the Fig. 6. The Blue, Red, Green and violet bars or bars from front to back comprises the efficiencies of LDA, kNN, SVM and RDA respectively. Analysis of emotions happy, neutral, anger and sad are represented by the bars from left to right.

4.4 Analysis of results by ROC curves

The shape of the ROC Curve and the area under the curve helps to estimate the discriminative power of a classifier. The area under the curve can have any value between 0 and 1 and it is a good indicator of the goodness of the test. From the literature survey the relationship between Area under curve and its diagnostic accuracy are shown in the Table 8. From the Table 8 it is observed that as the area under curve (AUC) reaches 1.0, the performance of the classification technique is excellent in classifying emotions, if AUC is less than 0.6 the performance of classification technique is poor, if the AUC is in between 0.6 and 0.9 then the results obtained are satisfactory if it is less than 0.5 the classification technique is not useful.

To plot an ROC curve to our four emotional classes, we first divide speech samples into positive and negative emotions. Happy and Neutral speech samples are positive emotions and speech samples of Anger, Sad are negative emotions. Sensitivity evaluates how the test is good at detecting positive emotions and Specificity evaluates how good the negatives emotions are discarded from positive emotions. ROC curve is a graphical representation of the relationship between both sensitivity and specificity.

The Fig. 7. shows the performance comparison of different classifiers. The Table 9 shows the values extracted from ROC plot for different classification Algorithms over Berlin and Spanish databases. The results obtained using ROC curve are comparable with each other. The AUC is greater than 0.8 for RDA and SVM which means the diagnostic accuracy of these classifiers is very good, similarly the AUC is greater than 0.6 for kNN and LDA which means the diagnostic accuracy of these classifiers is sufficient from Table 8.

From the Table 9 it is observed that the AUC is above 0.8 for RDA and above 0.7 for SVM which means the diagnostic accuracy of the classifiers are very good and good respectively, in the same way the AUC is more than 0.6 for kNN and LDA which means the diagnostic accuracies of these classifiers is sufficient from Table 8.

4.5 Analysis of results for driver applications

In driving task, emotion of the driver has an impact on the driving performance, which can be improved if the automobile actively respond to the emotional state of the driver. It is important for a smart driver support system to accurately monitor the driver state in a robust manner.

The analysis presented in this paper has been performed using two acted speech emotional databases. Emotions from these databases are extracted up to moderate level. This

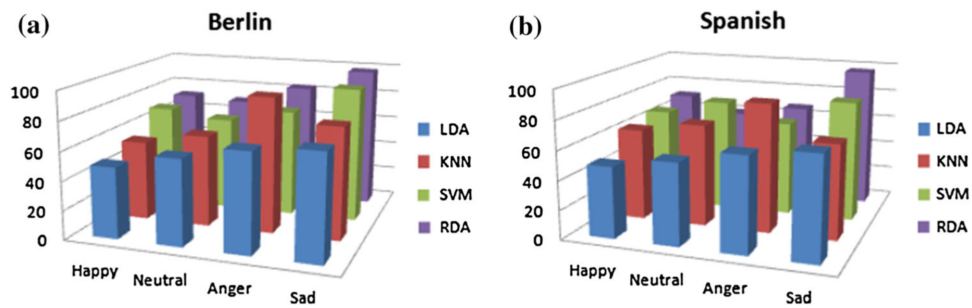

Fig. 6 Comparison of emotion recognition performances of different classifiers using (**a**) Berlin database, (**b**) Spanish database

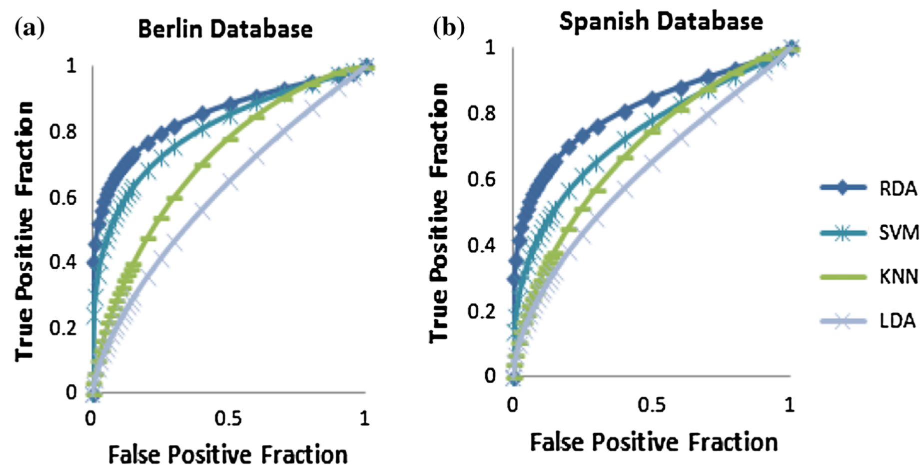

Fig. 7 Shows the comparison of performances of different classifiers using (**a**) Berlin database, (**b**) Spanish database

Table 8 The relationship between area under curve and its diagnostic accuracy

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

Table 9 Shows the values (Accu: Accuracy, Sens: Sensitivity, Spec: Specificity, AUC: Area Under Curve) extracted from roc plot for different classifiers for (a) Spanish database and (b) Berlin database

Algorithm	Accu (%)	Sens (%)	Spec (%)	AUC (%)
(a)				
RDA	74.0	70.7	78.4	0.814
SVM	72.0	72.2	71.8	0.734
kNN	71.8	75.1	69.2	0.688
LDA	62.0	60.3	64.3	0.619
(b)				
RDA	80.8	75.7	88.2	0.854
SVM	75.5	72.0	80.4	0.806
kNN	72.0	67.5	79.7	0.709
LDA	60.0	57.7	60.8	0.601

success is obtained due to implementation of fused features. In order to see whether these results are applicable to real life driver emotion identification system we have to create a database of speech samples collected from a driver naturally. The speech samples from Berlin and Spanish databases are collected in a noise free environment so they do not contain noise but the speech samples that are collected in real life contain noise due to vehicle, environment and co-passengers voice. To get the noise free samples of a driver the technique called voice activity detection (VAD) is used as a preprocessing step to eliminate noise. After extracting the noise free samples features are extracted and combined using feature fusion which gives the emotion specific information of the driver. Finally by using this information and a classifier will classify the emotional state of the driver. Basing on these results the smart driver support system will alert the driver.

5 Conclusion

The emotion recognition accuracy using Acoustic information generated by the driver is systematically evaluated by using Berlin and Spanish emotional databases. This has been implemented by a variety of classifiers including LDA, RDA, SVM and KNN using various prosody and spectral features. The emotion recognition performance of the classifiers is obtained effectively with feature fusion technique. An extensive comparative study has been made on these classifiers. The results of the evaluation have showed that RDA yields better recognition performance and SVM also gives good recognition performance compared with other classifiers. The use of feature fusion technique instead of using individual features enhances the recognition performance. The experimental results suggests that recognition accuracy is improved by 20 % approximately for each classifier with feature fusion.

Acknowledgments This work was supported by Research Project on "Non-intrusive real time driving process ergonomics monitoring system to improve road safety in a car – pc environment" funded by DST, New Delhi.

References

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

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

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.

El Ayadi, M. M., Kamel, M. S., & Karray, F. (2007). Speech emotion recognition using gaussian mixture vector autoregressive models. In: *Acoustics, Speech and Signal Processing, 2007*. ICASSP 2007. IEEE International Conference on, IEEE, vol. 4, pp. IV-957.

Ji, S., & Ye, J. (2008). Generalized linear discriminant analysis: A unified framework and efficient model selection. *IEEE Transactions on Neural Networks*, *19*(10), 1768–1782.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology*, *15*(2), 99–117.

Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011). Speech emotion recognition using segmental level prosodic analysis. In: *Devices and communications (ICDeCom), 2011 International Conference on, IEEE*, pp. 1–5.

Luengo, I., Navas, E., Hernáez, I., & Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. In: *Interspeech*, pp. 493–496.

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.

Milton, A., Roy, S. S., & Selvi, S. (2013). Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, *69*(9), 34–39.

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing and Applications*, *9*(4), 290–296.

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

Ravikumar, M., & Suresha, M. (2013). Dimensionality reduction and classification of color features data using svm and knn. *International Journal of Image Processing and Visual Communication, 1*(4), 16–21.

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

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden markov model-based speech emotion recognition. In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on, IEEE*, vol. 2, pp. II-1.

Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002). Emotional space improves emotion recognition. In: *Interspeech*.

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*, Springer, pp. 83–94.

Vankayalapati, H. D., & SVKK 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 journal*, *3*(1), 65–73.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181.

Vogt, T., André, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In C. Peter & R. Beale (Eds.), *Affect and emotion in human–computer interaction* (pp. 75–91). Springer.

Ye, J., Xiong, T., Li, Q., Janardan, R., Bi, J., Cherkassky, V., & Kambhamettu, C. (2006). Efficient model selection for regularized linear discriminant analysis. In: *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, pp. 532–539.

Zhou, Y., Sun, Y., Zhang, J., & Yan, Y. (2009). Speech emotion recognition using both spectral and prosodic features. In: *Information engineering and computer science, 2009. ICIECS 2009*. International Conference on, IEEE, pp. 1–4.