Sindh Univ. Res. Jour. (Sci. Ser.) Vol.47 (1):103-106 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Speaker-Independent Speech Emotion Recognition using Gaussian and SVM Classifiers

S. HAQ, A. ALI*, M. ASIF*, T. JAN++, Y. KHAN**, Naveed AHMAD,***

Department of Electronics, University of Peshawar

Received 17th October, 2014 and Revised 10th December 2014

Abstract: This paper presents a novel technique of human emotion recognition using the audio modality for the speaker-independent task. In order to achieve a high emotion classification performance a standard set of audio features were extracted. The feature selection was performed using the Plus l-Take Away r algorithm based on Bhattacharyya distance criterion. The feature selection was followed by feature reduction using PCA and LDA, and classification using the Gaussian and SVM classifiers. The emotion classification performance better or comparable to state-of-the art techniques and humans were achieved on the standard Berlin emotional speech database.

Keywords: Audio Emotion Recognition, Feature Selection, Principal Component Analysis, Gaussian and SVM Classifiers

1. INTRODUCTION

In recent years, humans have developed advanced technologies which have made the human-computer interaction more attractive for humans. Primarily, human beings use speech to communicate with each other, but the complete message cannot be conveyed only through verbal content. Some additional information are required to completely understand the message, these includes vocalized emotions, facial expressions, hand gestures and body language (Kessous *et al.,* 2010). From human's point of view, the human-machine interaction will be more interesting if machines can recognize human emotions and are able to respond accordingly. On other hand, communication will be more reliable if machines are able to understand human emotions (Cowie *et al.,* 2001).

Automatic emotion recognition covers a wide range of applications including automobile systems, customer services systems, health care, and game and film industries. In recent years, researchers from different disciplines have taken interest in the field of human emotion recognition and significant progress has been made in several areas including recording of emotional databases, feature extraction, feature selection, and classification (Zeng *et al.,* 2009).

One of the factors upon which the reliability of an emotion recognizer depends is the quality of emotional data used to build the emotion recognition system. Popular audio databases include the AIBO Corpus, Berlin Emotional Speech Database (EMODB) and Danish Emotional Speech Database (Batliner *et al.,* 2004; Burkhardt *et al.,* 2005; Engberg and Hansen, 1996). The emotional databases are either acted or natural.

Audio features of various types have been analyzed to achieve better emotion classification results. Important acoustic features include pitch, energy, duration, spectral energy distribution, formants, and Mel frequency cepstral coefficients (MFCCs). These features have been extracted both at utterance-level (Haq and Jackson, 2009; Hassan and Damper, 2009), and at frame-level (Kao and Lee, 2006; Lin and Wei, 2005). Some studies have suggested the use of both acoustic and linguistic features to achieve improved classification performance (Seppi *et al.,* 2008). The automatic extraction of linguistic features is a difficult task, and for this reason most audio emotion recognizers are based on acoustic features. We used acoustic features for our analysis.

In the field of pattern recognition, feature selection and reduction techniques are used to remove the uninformative, redundant and noisy information. These techniques improve the classification accuracy and computational efficiency. The feature search techniques used in the field of emotion recognition include sequential floating forward selection (Haq *et al.,* 2008), genetic algorithms (Casale *et al.,* 2007) and best-first (Gunes and Piccardi, 2005). Feature reduction techniques include PCA and LDA (Busso *et al.,* 2004.

In any pattern recognition problem, the choice of classifier is very important. These classifiers include hidden Markov model (Lin and Wei, 2005), neural network (Petrushin, 1999), support vector machine (Ashraf *et al.,* 2007), and adaptive boosting (Bartlett *et al.,* 2005).

2. BERLIN EMOTIONAL SPEECH DATABASE

The Berlin Emotional Speech Database (EMODB) is a popular acted emotional database (Burkhardt *et al.,* 2005) recorded from 10 professional actors (5 male, 5 female) in German. The database consists of 10 sentences: 5 with one phrase and 5 with

++Corresponding Author: Tariqullah Jan, tariqullahjan@uetpeshawar.edu.pk, +92-91-9216498,

* Department of Electrical Engineering, Sarhad University of Science and IT, Peshawar.

** Department of Electrical Engineering, IQRA National University, Peshawar.

*** Department of Computer Science, University of Peshawar.

two phrases. The emotion categories consist of anger, boredom, disgust, fear, happiness, sadness and neutral. The data were recorded in an anechoic chamber with high quality equipment at a sampling rate of 48 kHz and later down sampled to 16 kHz. The recorded speech material consisted of 800 utterances. The database was evaluated by 20 subjects for recognition and naturalness of emotions. The final database consisted of 494 utterances with better than 80 % recognition and judged as natural by more than 60 % of the listeners. The classification accuracy in human perception experiments was 84.3 %.

3. MATERIAL AND METHODS

Proposed method

Proposed method

Emotion classification on EMODB was performed using a standard set of audio features. The feature extraction was followed by feature selection, feature reduction and classification using the Gaussian and SVM classifiers. The results were compared with the state-of-the-art techniques.

3.1 Feature Extraction

We extracted 384 audio features that were used in the Interspeech 2009 Emotion Challenge (Schuller *et al.*, 2009) using the openEAR toolkit (Eyben *et al.*, 2009). The feature set consisted of low-level descriptors (LLD) including fundamental frequency (f_0), energy, Mel frequency cepstral coefficients (MFCCs), zero crossing rate (ZCR) and line spectral frequencies (lsf). The delta coefficients were calculated for each of these LLD. In the next step, the following functions were applied to the LLD and their delta coefficients: mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, range, and 2 linear regression coefficients. Feature normalization was performed using the means and standard deviations calculated from the training data.

3.2 Feature Selection and Reduction

For the audio emotion classification, we adopted a two-step process: feature selection with Plus l-Take Away r algorithm, followed by feature reduction with PCA and LDA.

Feature Selection

The Plus l-Take Away r algorithm is a feature search method based on some criterion function (Kittler, 1978). It combines the sequential forward selection (SFS) and sequential backward selection (SBS) algorithms to achieve better results. At each step, l numbers of features are included to the current feature set and r numbers of features are discarded. The process continues until the required feature set size is achieved. The feature search was performed with $l = 2$ and $r = 1$, i.e., one feature was added at each step. We used this algorithm for feature selection based on Bhattacharyya distance criterion (Campbell, 1997).

Features were normalized prior to applying the feature selection by using the Z-norm (i.e., mean subtraction and division by standard deviation). The Bhattacharyya distance provides the separability between two classes. The distribution of classes was assumed to be normal. For m number of classes the separability measure is obtained by averaging it over all binary combinations of the classes.

Feature Reduction

Statistical methods can be used to reduce the dimensionality of a feature set. PCA technique (Shlens, 2005) is used to extract the important characteristics of high-dimensional data and to remove the uninformative and noisy data. The LDA method (Duda *et al.,* 2001) provides the separation between classes based on the ratio of between-class variance to within-class variance. For feature reduction, we applied both PCA and LDA as linear transformation .

3.3 Classification

The Gaussian classifier utilizes the Bayes decision theory for classification. It is assumed that the class-conditional probability $p(x|\omega_i)$ have Gaussian distribution for each class ω_i. The Bayes decision rule is described as

$$i_{Bayes} = \arg\max_i P(\omega_i|x) = \arg\max_i p(x|\omega_i)P(\omega_i) \quad (1)$$

where $P(\omega_i|x)$ denotes the posterior probability, and $P(\omega_i)$ defines the prior class probability. A single component Gaussian was used to model each emotion class ω_i using a diagonal covariance matrix.

The SVM and Gaussian classifiers were used to classify between seven emotion classes using three types of features: raw, PCA-transformed, and LDA-transformed.

4. RESULTS AND DISCUSSION

Experiments were conducted by dividing the data into 10 sets for leave-one-speaker-out (LOSO) cross-validation. Each set consisted of data from a single speaker. For each experiment, 9 sets were used for training and 1 set for testing. Finally, the results were averaged over 10 speakers' data.

Emotion classification was performed using three sets of features: 384, 170 and 105. Features were selected from the subgroups of full feature set. The set of 105 selected features consisted of 20 f_0 , 25 energy, 25 MFCC, 10 ZCR and 25 lsf features, whereas the set of 170 selected features included 30 f_0 , 40 energy, 40 MFCC, 20 ZCR and 40 lsf features.

4.1 Analysis of results

The classification accuracies achieved for seven emotion classes with the SVM and Gaussian classifiers are shown in **(Fig. 1)**. The results were achieved using the raw features without applying a linear transformation, and the features with PCA and LDA

transformation. The performance of the Gaussian classifier was lower with the raw features and linear transformation resulted in a significant improvement in the classification accuracy. Feature selection helped to improve the classification performance. For a feature set of 384 features, the performance was lower with the raw, PCA and LDA features.

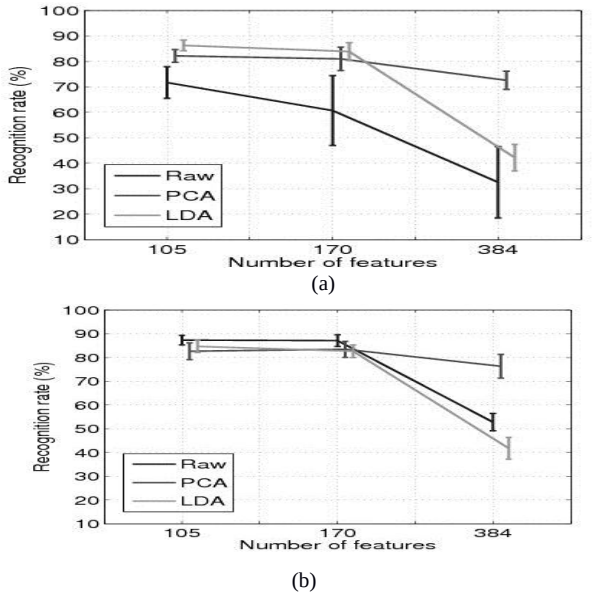


Fig. 1: Average classification accuracy (%) achieved for seven emotions on EMODB using 105, 170 and 384 Raw, PCA-transformed, and LDA-transformed features with (a) Gaussian, and (b) SVM classifiers.

Classification accuracy above 80 % was achieved for both the PCA and LDA with 170 and 105 features. On the other hand, the SVM classifier performed better with the raw features in comparison to the PCA and LDA features. The performance was lower with 384 features, and classification accuracy above 80 % was achieved for the raw, PCA and LDA with 170 and 105 features.

A comparison of the classification accuracies for the two classifiers indicated comparable performance for PCA and LDA applied to different numbers of features (105, 170, and 384). For 384 features, the classification performance of both classifiers was low. Feature selection improved the classification accuracy of both classifiers. In general, the Gaussian classifier performed better for LDA compared to PCA. The best performance achieved with the Gaussian classifier was 86 % (standard error (SE): 2.1) with 105 LDA-transformed features, and with the SVM classifier it was 87 % (SE: 2.0) using 105 and 170 raw features.

4.2 Comparison to other techniques

The EMODB has been used by many researchers for the investigation of audio emotion classification. A comparative analysis of various techniques is given in **(Table 1)** for seven emotion categories. Shami and Verhelst (2007) achieved a classification accuracy of 75.5 % using 200 features with an SVM classifier for 10-fold cross-validation. Hassan and Damper (2009) reported an average classification accuracy of 70.8 % by selecting 14 out of 166 audio features using SFS. The k-nearest neighbour classifier was used for classification with 10-fold cross-validation. Vlasenko *et al.* (2007) combined the frame and turn-level features and performed speaker normalization. An average classification accuracy of 81.6 % was achieved. The classification accuracy improved to 89.9 % (76 out of 1407 features) with the correlation-based feature selection. The classification was performed using an SVM classifier with leave-one-speaker-out cross-validation. Vlasenko *et al.* (2008) reported a classification accuracy of 83.2% with an SVM classifier by applying speaker normalization and feature selection to 1406 features. An average accuracy of 82.9 % was achieved with MFCC features and vocal tract length normalization.

Table 1: Comparison of average recognition rates achieved with different techniques for seven emotions on EMODB.

Reference	Method	Recognition rate (%)
Shami and Verhelst (2007)	SVM classifier	75.5
Hassan and Damper (2009)	Sequential forward selection, k-nearest neighbour classifier	70.8
Vlasenko *et al.* (2007a)	Correlation-based feature selection, SVM classifier	89.9
Vlasenko *et al.* (2008)	Speaker normalization, SVM classifier	83.2
	MFCC features, vocal tract length normalization	82.9
Our Method	Plus l-Take Away r selection, LDA, Gaussian classifier	86.3
	Plus l-Take Away r selection, SVM classifier	87.3

In comparison to other researchers, we achieved an average classification accuracy of 86.3 % with a Gaussian classifier by applying LDA to 105 features, and the best performance of 87.3 % was obtained with an SVM classifier using 105 features without any linear transformation. We achieved better performance than those reported by Shami and Verhelst (2007), Hassan and Damper (2009) and Vlasenko *et al.* (2008). On the other hand, Vlasenko *et al.* (2007) reported a slightly better performance with multiple level features and correlation-based feature selection.

To conclude, the results indicated better performance with the selected features. The SVM classifier performed better with the raw features as compared to the PCA- or LDA-transformed features, whereas the performance of the Gaussian classifier improved with the linear transformation.

5. CONCLUSION AND FUTURE WORK

For the speaker-independent task using the Berlin database, the average recognition rates

comparable to humans and state-of-the-art techniques were achieved for the audio modality. An emotion recognition system consisting of feature selection using the Plus l-Take Away r algorithm, feature reduction using the PCA and LDA, and classification using Gaussian and SVM classifiers was tested and high performance was achieved. The best classification achieved for seven emotions on EMODB was 86.3 % with a Gaussian classifier using 105 LDA-transformed features, and it was 87.3 % with an SVM classifier using 105 raw features. In comparison to other techniques, we achieved a better or comparable performance on the EMODB.

REFERENCES:

Ashraf, A. B., S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon and B. J. Theobald (2007) The Painful Face – Pain Expression Recognition Using Active Appearance Models. Proc. ACM Int'l Conf. on Multimodal Interfaces, Nagoya, Japan, 9–14.

Bartlett, M.S., I. Fasel and J. Movellan (2005) Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. Proc. IEEE Int'l Conf. Computer Vision, San Diego, CA, USA, 568–573.

Batliner, A., C. Hacker, S. Steidl, E. Noth, M. Russel and M. Wong (2004) "You Stupid Tin Box" - Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech Corpus. Proc. Int'l Conf. on Language Resources and Evaluation, Lisbon, Portugal, 171–174.

Busso, C., Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, and S. Narayanan (2004) Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. Proc. Int'l Conf. on Multimodal Interfaces, State College, PA, USA, 205–211.

Casale, S., A. Russo and S. Serranoa (2007) Multistyle Classification of Speech Under Stress using Feature Subset Selection Based on Genetic Algorithms. Speech Communication 49(10): 801–810.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor (2001) Emotion Recognition in Human-Computer Interaction. IEEE Signal Processing Magazine 18(1):32–80.

Eyben, F., and B. Schuller, (2009). Open EAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. Proc. IEEE 4th Int'l HUMAINE Association Conference on Affective Computing and Intelligent Interaction, Amsterdam, The Netherlands, 576–581.

Gunes, H. and M. Piccardi (2005) Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics, Hawaii, USA, 3437–3443.

Haq, S. and P.J.B. Jackson (2009) Speaker-Dependent Audio-Visual Emotion Recognition. Proc. Int'l Conf. Auditory-Visual Speech Processing, Norwich, UK, 53–58.

Haq, S. and P.J.B. Jackson (2010) Multimodal Emotion Recognition, Machine Audition: Principles, Algorithms and Systems, IGI Global Publisher, USA.

Haq, S., P.J.B. Jackson and J. D. Edge (2008) Audio-Visual Feature Selection and Reduction for Emotion Classification. Proc. Int'l Conf. on Auditory-Visual Speech Processing, Tangalooma, Australia, 185–190.

Kessous, L., and G. Caridakis (2010) Multimodal Emotion Recognition in Speech-Based Interaction using Facial Expression, Body Gesture and Acoustic Analysis, Journal on Multimodal User Interfaces 3(1):33–48.

Kim, E. H., K. H. Hyun and Y. K. Kwak (2006) Improvement of Emotion Recognition from Voice by Separating of Obstruents. Proc. IEEE Int'l Symposium on Robot and Human Interactive Communication, Hatfield, UK, 564–568.

Kittler, J. (1978) Feature Set Search Algorithms, Pattern Recognition and Signal Processing, Sijthoff & Noordoff International Pub., The Netherlands, 41–60.

Lee, C. and S. Narayanan, (2005) Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing, 13(2):293–303.

Pantic, M., N. Sebe, and T. Huang (2005) Affective Multimodal Human Computer Interaction. Proc. ACM Int'l Conf. on Multimedia, Hilton, Singapore, 669–676.

Schuller, B., S. Steidl, and A. Batliner, (2009) The Interspeech 2009 Emotion Challenge. Proc. Interspeech, Brighton, UK, 312–315.

Sebe, N., I. Cohen and T. S. Huang (2005) Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision, World Scientific, Singapore.

Seppi, D., A. Batliner, B. Schuller, S. Steidle, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, (2008) Patterns, Prototypes, Performance: Classifying Emotional User States. Proc. Interspeech, Brisbane, Australia, 601–604.

Vlasenko, B., B. Schuller, A. Wendemuth, and G. Rigoll, (2007) Combining Frame and Turn-Level Information for Robust Recognition of Emotions Within Speech. Proc. Interspeech, Antwerp, Belgium, 2249–2252.

Vlasenko, B., B. Schuller, K. Mengistu, G. Rigoll, and A. Wendemuth, (2008) Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest. Proc. Interspeech, Brisbane, Australia, 805–808.

Zeng, Z., M. G. I. Pantic, Roisman and T. S. Huang (2009) A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1): 39–58.