Research Journal of Applied Sciences, Engineering and Technology 9(12): 1105-1114, 2015
DOI:10.19026/rjaset.9.2604
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 10, 2014 Accepted: November 10, 2014 Published: April 25, 2015

**Research Article**

# Speech Emotion Recognition Using Adaptive Ensemble of Class Specific Classifiers

P. Vasuki
Department of Information Technology, SSN College of Engineering, Chennai 603110, India

**Abstract:** Emotion recognition plays a significant role in Human Computer Interaction (HCI) field for effective communication. The aim of this study is to built a generic emotion recognition system to face the challenges of recognition in resolving confusion among acoustical characteristics of emotions, identifying dominating emotion from mixed emotions etc. When there is confusion among the perception of emotion by human, the understanding of it by machine is a real challenge. Due to these reasons, it is very hard to produce highly accurate emotion recognition system in real time. Researchers are working to improve the performance of emotion recognition task by designing different classifiers and also using different ensemble methodologies at data level, feature level or decision levels to recognize emotion. We have built a generic SVM based emotion recognition system, which models emotions using given features. Out of given acoustical features, for every emotional class, class specific best features are identified based on $f_1$ measure. The responses of the systems, built based on these best features are combined using new smart additive ensemble techniques. Decision logic is employed to decode the responses into an emotional class, the class which produces maximum value among all emotional classes. A rejection framework is also designed to reject a noisy and weak input file. We have tested the framework with 12 acoustical features on Berlin emotional corpus EMO-DB. The accuracy obtained from our generic emotion recognition system is 74.70% which is better than classifiers reported in the literature.

**Keywords:** Adaptive ensemble, ensemble classifier, speech emotion recognition, SVM classifier

## INTRODUCTION

In the growing robotic age, the need for automatic understanding of emotions is very much needed for effective communication. Emotion maybe recognized from facial reaction, from speech and even from EEG signal (Yuen *et al*., 2013). Speech is a natural way of communication between human and machine. Speech Emotion Recognition (SER) has been used in various applications like evaluation of employees of call center by analyzing emotions of responses of the clients (Valery, 1999). Speech emotion recognition may also used to enhance accuracy of automatic speech recognition. Same statements when stated with different intonation leads to different meaning. Understanding emotions may provide exact meaning of speech uttered.

Recognition of emotion is difficult as there are large variations in emotions expressions due to various reasons ranging from social cultural background to recording environment variations. These variations affect the performance of emotion recognition. Various research works have been carried out to improve the performance of SER, using different methodologies.

In this study, we have built a system with frequency and time domain features and prosodic features. The experiments are carried out using the corpus EMO-DB. The corpus consists of seven different emotions namely anger, boredom, disgust, fear, happiness, neutral and sadness. We have extracted twelve different features using the tool opensmile. The features are analyzed and the suitable feature for classifying every individual emotion is identified. From the observation it is found that either the Perceptual Linear Prediction (PLP) coefficients (PLP_0_D_A) or Mel Frequency Cepstral Coefficients (MFCC_E_D_A and MFCC_0_D_A) is the best features for different emotional classes. We ensemble the responses of the classifiers built on these features using adaptive addition technique. The adaptive addition learns the weight of different classifiers using our algorithm. We have also constructed a rejection framework which rejects the noisy input utterance based on the responses of the first level classifiers.

## LITERATURE REVIEW

Performance of speech emotion recognition may be improved by using different input features. Various input features (Fulmare *et al*., 2013) in frequency domain (Wu *et al*., 2011), time domain (Koolagudi and Krothapalli, 2011) andprosodic (Dellaert *et al*., 1996; Rao and Koolagudi, 2013) and linguistic features (Polzhehl *et al*., 2011) are extracted from input speech and used as input parameters of an emotion classifier. Some of the scientists have worked in class specific features; (Milton and Tamil Selvi, 2013; Bitouk *et al*., 2010) where different feature sets are involved in

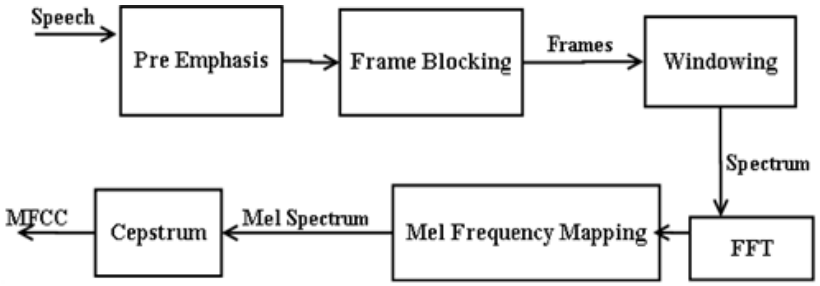


Fig. 1: MFCC extraction from speech signal

identifying different classes (Chen *et al*., 2012). Scientists also have worked in the direction of deriving optimal feature set from a pool of features using algorithms like sequential forward selections and genetic algorithms (Böck *et al*., 2010; Schuller *et al*., 2006). There are many classifiers like SVM (Support Vector Machines) (Schuller *et al*., 2009), HMM (Hidden Marcov Model) (Böck *et al*., 2010; Tin *et al*., 2003), (GMM) Gaussian Mixer Model (Neiberg *et al*., 2006), Artificial Neural Network (ANN) (Mehmet *et al*., 2009; Böck *et al*., 2010) are involved in classifying emotion. Researchers also have carried out researches in fine tuning the process of input parameter extractions. As the utterance length considered in training and testing is varying, the performance of speech emotion recognition may get affected. Researchers also have carried out work to fix the length of input speech utterance using methods like contraction and elongation method and intermediate matching kernel technique where the reference patterns are generated for inputs, every input utterance will be mapped to an appropriate reference pattern which is closer to the given input pattern (Dileep and Chandra Sekar, 2014). Research works have also been carried out in extraction of features in different ways by changing input frame length using wavelets (Krishna Kishore and Krishna Satish, 2013), by changing frequency range of input filters (Trabelsi *et al*., 2013) and changing number of mixtures in GMM and using new kernels in SVM (Maaoui and Pruski, 2008) etc. Specific features to identify emotions in a particular language are also having been studied.

The performance of the system may also be improved by various ensemble methodologies (Kobayashi and Calag, 2013). Ensemble methods are the learning algorithms constructed to learn from a set of classifiers responses and classify new input, from an appropriate fusion (Vasuki and Aravindan, 2012). In various classification systems, fusion is done at data level, feature level, response level and decision level. In fusion of features, researchers have worked with fusion of acoustic feature with linguistic feature (Polzhehl *et al*., 2011) and time domain features with frequency domain features (Rao and Koolagudi, 2013) and so on.

In this study, the system identifies the best feature to classify every individual emotion (class specific feature). We have proposed an adaptive ensemble classifier, which ensemble the weighted response of base classifiers built on best features. The weights of the classifiers are fixed on proportional $f_1$ basis. A rejection framework is designed to reject noisy input utterances. The overall accuracy obtained from our system is better than the accuracy obtained from classifier built on any other combination of given feature set.

**Features:** We have used time and frequency domain features and prosodic features to build our system.

**Mel frequency cepstral coefficients-MFCC:** MFCC is a representation of the short term power spectrum of a sound, based on linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The frame size is set to 25 msec at a rate of 10 msec. A Hamming function is used to window the frames and a pre-emphasis with k = 0.97 is applied. The (12+1) MFCC are computed from 26 Mel-bands computed from the FFT power spectrum. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. Figure 1 describes in detail about the MFCC feature extraction process.

**Perceptual linear coefficients-PLP:** Figure 2 illustrates the steps involved in the feature PLP-extraction.PLP analysis more consistent with auditory spectrum and PLP is efficient and low dimensional characteristics of speech.

**Prosodic features:** The prosodic features include the fundamental frequency (F0), the voicing probability and the loudness contours. Pitch may be estimated in three methods like autocorrelation, cepstral and SIFT (Simplified Inverse Filtering and Tracking) method. In this study, pitch was calculated based on the autocorrelation method and cepstrum based method. The collection 'prosodyAcf' uses autocorrelation and cepstrum based method to extract the fundamental frequency. The file collection prosodyShs extract the

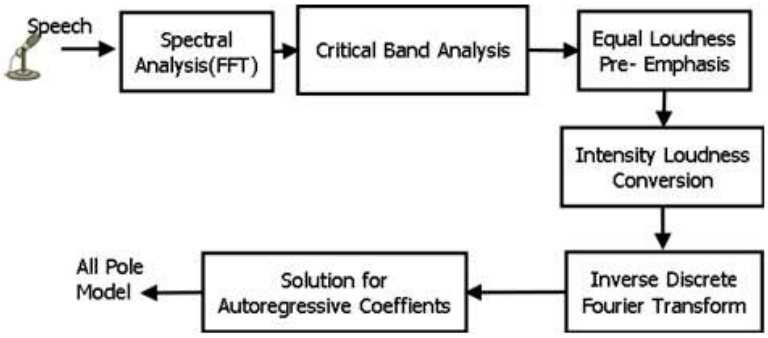


Fig. 2: PLP extraction feature

fundamental frequency via the Sub-Harmonic Sampling algorithm (SHS).

**Short term energy:** Short term energy is defined as the summation of the square of amplitude of the signal varies in time. So based on emotions the energy gets varied. This energy extraction is done by segmenting speech signal into frames and the energy for each frame has been calculated by the formula of:

$$E = \sum_{n=-\infty}^{\infty} s^2(n) \quad (1)$$

**Audio speculation:** It maps the power spectrum to an auditory frequency axis, by combining FFT bins into equally-spaced intervals on the bark axis (or one approximation of it).

**SVM classifier:** In SER, Support Vector Machines (SVMs), are supervised learning models with associated learning algorithms, used to analyze and recognize emotion patterns (Corninna and Vapnik, 1995). Given a set of training examples in each category of emotion, an SVM training algorithm builds a model that assigns new examples into one category of emotions (Giannakopoulos *et al*., 2009; Crammer and Singer, 2001).

**Cross validation:** The training and testing should happen in unbiased manner and the influence of number of training examples and the local optima in the behavior of classifiers have to be reduced. For this purpose the entire training set is classified into development set and test set. The development set is cross folded into many sets, in such a way that the training set includes all variations in recording environment like, time of recording, gender and speaker variability.

**Ensemble techniques:** In Machine learning, ensemble is done to bring different factors together in input or output level of classifier. In input level different categories of features are mixed together to train and test the classifier, thus the response of the classifier will depend various factors. In classifier fusion, the classifiers results are involved in fusion (Fig. 3). Appropriate decision logic is designed for the classification of ensemble responses. In some research, meta classifier a (classifier trained based on the response of level one classifiers) is devised to take decision. According to literature survey, ensemble works better in SER applications (Rao and Koolagudi, 2013; Tariq *et al*., 2011).

**Fusion:** AdaBoost, short for "Adaptive Boosting", is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire. It is the learning algorithm to improve the performance of the existing classifiers. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier (Kittler *et al*., 1998).

AdaBoost is adaptive to change the change the weak learners misclassification. But the system has a disadvantage that AdaBoost is sensitive to noisy data and outliers. In some problems, however, it can be less susceptible to the over fitting problem. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing (i.e., their error rate is smaller than 0.5 for binary classification), the final model can be proven to converge to a strong learner. While every algorithm configured with different parameter set may suitable to some problem, adaboosting providing best out-of-box, which consolidates betterment of every classifier:

$$h_f(x) = \sum_{t=1}^{N} h_t(x) \quad (2)$$

The hypothesis of an ensemble classifier is aggregated hypothesis of all the classifier involved in ensemble.

## MATERIALS AND METHODS

In this study, we have built a well performing emotion recognition system. We have developed

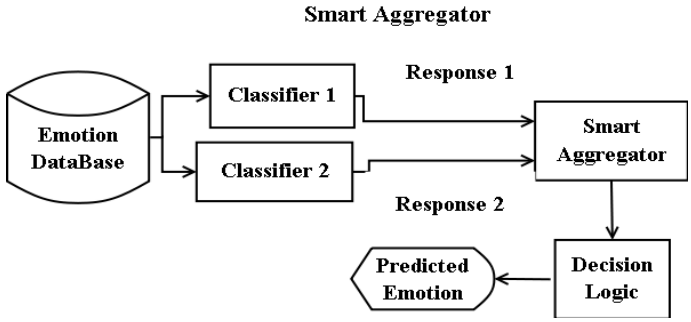


Fig. 3: Classifier fusion-simple aggregation

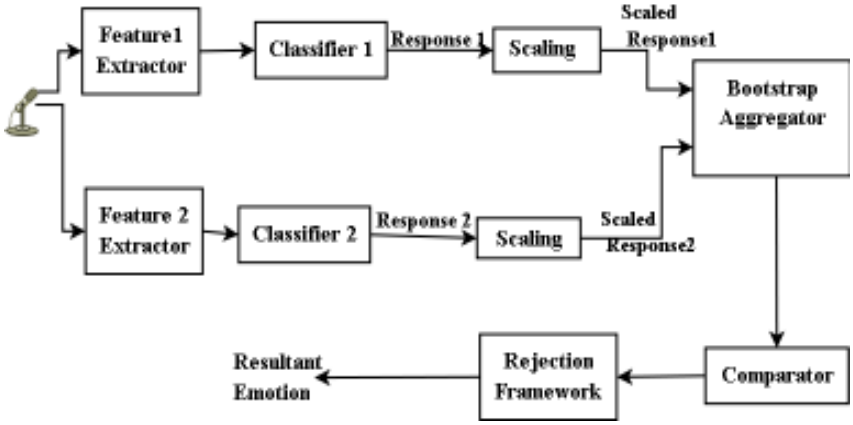


Fig. 4: System architecture

different learners using different features. Some learner may be best among available features on classifying a specific emotion. We have identified the best classification feature of an emotional class. Only features which identify at least an emotional class at the best recognition rate are selected for ensemble. Thus the best classifier is identified for every emotional class. The ensemble result is combination of responses of all such classifiers. Thus the collective response of ensemble classifier represents representations from dominating feature of every individual emotional class. A rejection framework has been developed to filter out noisy data.

**Materials:**

**Data:** We have used EMO-DB-a Berlin emotional speech corpus for this experiment.

**EMO-DB emotional speech corpus from Berlin:** An European emotional database EMO-DB, recorded in Berlin region in German language, is one of the most commonly used emotional corpus for emotion recognition (Zixing *et al*., 2011; Shami and Verhelst, 2007; Bitouk *et al*., 2010; Schuller *et al*., 2005; Kamaruddin *et al*., 2012; Giannakopoulos *et al*., 2009). Burkhardt *et al*. (2005, 2009) recorded EMO-DB with the help of ten male and female professional actors, who were asked to utter a sentence with the predefined emotions. The number of speech files in each category varies from 50 to 100 and in overall, there are 840 utterances are available. The database consists of 7 different emotions namely anger, boredom, disgust, happiness, sadness, fear and neutral. 20 observers were involved in the perception test and 67% accuracy is obtained from the test.

**Data preparation:** The entire database is partitioned into two as training (development) set and testing set. 75% of the data are used in development and rest is used for training. Ten fold cross validation is done on the development set to ensure that the system behaves in unbiased manner towards the selection of training and testing set. The training parameters are set based on the performance of classifiers on development set.

**Methodology:** Figure 4 shows the detail description of the architecture of the proposed methodology. The System collects responses from different classifiers and aggregates the responses using weighted average method. The resultant response is evaluated by rejection framework to find out whether the input utterance has mixture of emotions or single emotion. If the reject frame work accepts the input, the final response is forwarded to the decision logic to decode the dominant emotion present in input utterance.

**Training:** During training the speech input utterances of development set are cross folded and fed to the

system. There are number classifiers equivalent to number of features are built, each learner is built with different feature. Every emotional class is modeled in all learners and tested. The performance on development set of the each classifier is measured and recorded. Based on $f_1$ measure, the best learner of every emotional class is chosen. During testing, the responses of best learners are fusion to generate combined response. Thus the weight of each learner involved in ensemble is calculated based on normalized average $f_1$ measure of best learners and recorded.

**Ensemble:** The simple fusion simply aggregates the response of the classifiers. The adaptive aggregator aggregates the response of the classifiers with the weight factor. The weight for every classifier is initially fixed based on their $f_1$ measure. Using the development set of data the weight vector (Weights of all classifiers) is updated.
Weight of $i^{th}$ classifier:

$$\omega_i = \frac{f_{1i}}{\sum_{i=1}^{N} f_{1i}} \quad (3)$$

Now the new weights are to be normalized using the formula:

$$\omega_i = \frac{\omega_i}{\sum_{i=1}^{N} \omega_i} \quad (4)$$

The final weight vector is stored in a file.

**Testing:** During testing, the features identified as the best during training are extracted from the test speech utterance and fed to the appropriate classifier. The responses of each SVM classifier are scaled down to zero to one. The weights of the classifiers are read from the training output file. The scaled down weighted response is the result of the resultant classifier:

$$h_f(t) = \sum_{t=1}^{n} \omega_i * h_i(x) \quad (5)$$

N = The total number of base classifiers
$\omega_i$ = Weight of $i^{th}$ classifier
$h_i(x)$ = Response of the $i^{th}$ classifiers

These weights will be used in testing set to identify emotion of test utterance.

The response of the resultant classifier is given to decision logic. The decision logic identifies the emotional class:

$$e = \max_{e_i \, \varepsilon \, E} e_i \quad (6)$$

The emotion which has highest result will be labeled as the emotional class of input utterance.

**Rejection framework:** Generally the system performance is affected by the presence of outlier; our system identifies outlier using a rejection algorithm and rejects it from classification.

The system calculates the fusion response of all classifiers. A minimal threshold has been fixed as minimum distance of top two emotional classes of the combined response. Outlier may not have clear distinctions among resembling emotions. To identify outlier based on distance among top two emotions; difference between first highest response and the second highest response is calculated. The optimal threshold value has been fixed based on development set. During testing, the differences between top two emotional class is calculated and if the difference is less than the threshold value, it is identified that the input utterance is outlier and can't distinguish its identity as a single class and this outlier data may be rejected from classification. And the performance of the system is evaluated on the rest of the utterances.

## EXPERIMENTAL RESULTS

Experiments have been conducted to evaluate the performance of our ensemble classifier on Speech Emotion Recognition.
Objectives of the experiments:

- To ascertain that performance of our ensemble classifier is better compared to traditional classifier
- To ascertain that our rejection framework rejects outliers
- To ascertain that the performance of our ensemble classifier is better than the classifiers reported in literature

The features are extracted using open smile feature extraction tool. To implement SVM for training and testing emotional utterance we used the tool SVM-MULTICLASS.SVM multiclass uses an algorithm is based on Structural SVMs and it is an instance of SVMstruct (Joachims, 1999) (Hofmann). For linear kernels, SVM multiclass V2.20 is very fast and runtime scales linearly with the number of training example.

Opensmile is a freeware open source toolkit developed by Eyben *et al*. (2013). Many features can be extracted from speech starts from basic feature set too many statistical variations of feature sets like mean, max and median of different parameters etc. The results are obtained either in ARFF format or CSV format.

**Features:** We have tested our system with various features which includes variations of MFCC and PLP along with energy and prosodic features. The features used in our system are listed below.

**MFCC:**
**MFCC_12_0_D_ A:** It consists of MFCC coefficients 13 MFCC coefficients derived from 26 band filters, 13

delta coefficients and 13 acceleration coefficients are added to it.

**MFCC12_E_D_A:** It is same as MFCC12_0_D_A, log energy is added instead of zero coefficients.

**MFCC12_0_D_A_Z:** This configuration is the same as MFCC12_0_D_A, except that the features are meaning normalized with respect to the full input sequence.

**Perceptual linear coefficients-PLP:** PLP Cepstral Coefficients (PLP-CC) with many variations are extracted are listed.

**PLP_0_D_A:** It consists of PLP prediction, delta and acceleration coefficients.

**PLP_E_D_A:** This is same as PLP_0_D_A the same as PLP_0_D_A, except that the log energy is appended to the PLP 1-5 instead of the 0-th PLP.

**PLP_0_D_A_Z:** This configuration is the same as PLP_0_D_A, except that the features are mean normalized with respect to the full input sequence (usually a turn or sub-turn segment).

**PLP_E_D_A_Z:** This configuration is the same as PLP_E_D_A, except that the features are mean normalized with respect to the full input sequence.

**Prosodic features:** These files extract the fundamental frequency (F0), the voicing probability and the loudness contours. The file prosody Acf uses the `cPitchACF' component to extract the fundamental frequency via an autocorrelation and cepstrum based method. The file prosodyShs uses the `cPitchShs' component to extract the fundamental frequency via the Sub-Harmonic Sampling algorithm (SHS).

**Evaluation parameters:** We have used the evaluation parameters Accuracy, precision, recall and $f_1$ measure. The parameters are calculated using the formulae:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$f_1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

TP : True Positive
TN: True Negative
FP : False Positive
FN: False Negative

**Results:** We have presented the result obtained in the experiments using various classifier and ensemble classifier.

Table 1 shows $f_1$ value obtained on development set from SVM classifier trained with different features. MFCC12_0_D_A, MFCC12_E_D_A and PLP_O_D_A are found to be the best features at different emotional classes. From the result, the best feature of an emotional classifier is identified and tabulated in Table 2.

The average value of evaluation parameters of all emotional classes obtained by classifier of trained with every features individually is presented in Table 3. The

Table 1: Performance of classifiers built on individual feature

| Feature | Anger | Boredom | Disgust | Fear | Happy | Neutral | Fear |
|---|---|---|---|---|---|---|---|
| Audspec | 0.510 | 0.311 | 0.521 | 0.143 | 0.152 | 0.431 | 0.494 |
| Energy | 0.398 | 0.240 | 0.101 | 0.382 | 0.064 | 0.418 | 0.425 |
| MFCC12_0_D_A | 0.833 | 0.455 | **0.791** | 0.462 | **0.596** | 0.509 | 0.733 |
| MFCC_12_0_D_A_Z | 0.593 | 0.361 | 0.593 | 0.298 | 0.368 | 0.359 | 0.358 |
| MFCC12_E_D_A | **0.858** | 0.331 | 0.692 | 0.384 | 0.493 | 0.469 | 0.718 |
| MFCC12_E_D_A_Z | 0.626 | 0.435 | 0.593 | 0.384 | 0.340 | 0.304 | 0.588 |
| PLP_0_D_A | 0.707 | **0.591** | 0.583 | **0.532** | 0.107 | **0.614** | **0.807** |
| PLP_0_D_A_Z | 0.562 | 0.423 | 0.496 | 0.318 | 0.052 | 0.442 | 0.456 |
| PLP_E_D_A | 0.690 | 0.419 | 0.521 | 0.431 | 0.041 | 0.569 | 0.743 |
| PLP_E_D_A_Z | 0.594 | 0.175 | 0.473 | 0.396 | 0.021 | 0.442 | 0.561 |
| ProsodyAcf | 0.523 | 0.389 | 0.329 | 0.148 | 0.409 | 0.391 | 0.538 |
| ProsodyShs | 0.387 | 0.367 | 0.380 | 0.209 | 0.352 | 0.272 | 0.619 |

Table 2: Class specific feature

| Emotion | Best feature |
|---|---|
| Anger | MFCC12_E_D_A |
| Boredom | PLP_0_D_A |
| Disgust | MFCC12_0_D_A |
| Fear | PLP_0_D_A |
| Happiness | MFCC12_0_D_A |
| Neutral | PLP_0_D_A |
| Sadness | PLP_0_D_A |

Table 3: Result on test set tested with classifiers built on different features

| Feature | Accuracy (%) | Precision (%) | Recall (%) | $f_1$ (%) |
|---|---|---|---|---|
| Audspec | 40.63 | 40.86 | 37.57 | 38.31 |
| Energy | 45.83 | 51.43 | 44.00 | 43.64 |
| MFCC_12_0_D_A_Z | 50.00 | 53.00 | 49.57 | 50.13 |
| MFCC12_E_D_A_Z | 50.00 | 49.57 | 48.43 | 48.23 |
| PLP_0_D_A_Z | 44.79 | 42.57 | 41.71 | 41.46 |
| PLP_E_D_A | 56.25 | 54.43 | 49.57 | 49.19 |
| PLP_E_D_A_Z | 40.63 | 36.71 | 36.00 | 35.06 |
| ProsodyAcf | 34.38 | 32.43 | 30.43 | 30.31 |
| ProsodyShs | 35.42 | 34.86 | 34.71 | 34.19 |

Table 4: Test set output of classifiers built on best features and their ensemble classifier

| Feature | Accuracy (%) | Precision (%) | Recall (%) | $f_1$ (%) |
|---|---|---|---|---|
| MFCC12_0_D_A | 67.71 | 70.71 | 64.14 | 65.27 |
| MFCC12_E_D_A | 68.75 | 71.14 | 65.29 | 66.53 |
| PLP_0_D_A | 62.50 | 64.86 | 55.43 | 54.83 |
| Adaptive | 70.83 | 71.00 | 64.43 | 65.07 |
| After rejection of outliers | 74.70 | 77.40 | 68.79 | 72.84 |

Table 5: Confusion matrix (feature: PLP_0_D_A)

| Emotion | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| Anger | 24 | 0 | 0 | 0 | 0 | 0 | 0 |
| Boredom | 2 | 9 | 0 | 0 | 0 | 2 | 2 |
| Disgust | 1 | 1 | 2 | 1 | 1 | 0 | 2 |
| Fear | 2 | 0 | 0 | 8 | 1 | 0 | 1 |
| Happy | 10 | 0 | 0 | 0 | 1 | 1 | 0 |
| Neutral | 0 | 5 | 0 | 1 | 0 | 8 | 0 |
| Sad | 1 | 1 | 0 | 1 | 0 | 0 | 8 |

Accuracy: 62.50

Table 6: Confusion matrix (feature: MFCC_12_0_D_A)

| Emotion | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| Anger | 20 | 0 | 0 | 1 | 2 | 0 | 1 |
| Boredom | 0 | 12 | 0 | 0 | 0 | 1 | 2 |
| Disgust | 0 | 1 | 4 | 1 | 0 | 1 | 1 |
| Fear | 3 | 0 | 0 | 7 | 1 | 1 | 0 |
| Happy | 7 | 0 | 0 | 0 | 4 | 1 | 0 |
| Neutral | 1 | 4 | 0 | 0 | 0 | 9 | 0 |
| Sad | 0 | 1 | 0 | 1 | 0 | 0 | 9 |

Accuracy: 67.71

Table 7: Confusion matrix (feature: MFCC_12_E_D_A)

| Emotion | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| Anger | 20 | 0 | 0 | 2 | 2 | 0 | 0 |
| Boredom | 0 | 12 | 0 | 1 | 0 | 1 | 1 |
| Disgust | 0 | 1 | 4 | 1 | 0 | 1 | 1 |
| Fear | 2 | 0 | 0 | 7 | 2 | 1 | 0 |
| Happy | 7 | 0 | 0 | 0 | 5 | 0 | 0 |
| Neutral | 0 | 4 | 0 | 1 | 0 | 9 | 0 |
| Sad | 0 | 1 | 0 | 1 | 0 | 0 | 9 |

Accuracy: 68.75

results shows that the class specific features' average performance of all emotional classes is also good.

Table 4 presents the performance of SVM classifier built on individual feature and ensemble of best features (PLP_0_D_A, MFCC12_0_D_A and MFCC12_E_D_A) and also shows outlier removed result. As we have considered $f_1$ measure, the feature selection will be generic. The best accuracy obtained from individual acoustical feature is 68.75 with MFCC12_E_D_A which is lesser than 74.70, the accuracy of Adaptive ensemble classifiers performance.

From the observation of Table 1 and 3, it is found that the spectral features are dominating than time domain and prosodic feature and when zero crossing rates are added with the features the performance drops down as the z coefficients and energy are similar behavior in resembling emotions and increases the ambiguity among emotions which are closer in acoustical space.

Table 5 to 7 represent confusion matrix of the classifier built with features PLP_0_D_A, MFCC12_0_D_A and MFCC12_E_D_A respectively.

Table 8 and 9 shows the confusion matrix produced by adaptive fusion and outlier rejected classifiers response, respectively.

Figure 5 compares the classifiers accuracy trained with independent feature to the ensemble classifier output with adaptive fusion and outlier removed state.

Table 8: Confusion matrix (adaptive fusion)

| Emotion | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| Anger | 23 | 0 | 0 | 0 | 1 | 0 | 0 |
| Boredom | 0 | 12 | 0 | 0 | 0 | 2 | 1 |
| Disgust | 0 | 1 | 4 | 1 | 0 | 1 | 1 |
| Fear | 2 | 0 | 0 | 7 | 2 | 1 | 0 |
| Happy | 8 | 0 | 0 | 0 | 3 | 1 | 0 |
| Neutral | 0 | 4 | 0 | 0 | 0 | 10 | 0 |
| Sad | 0 | 1 | 0 | 1 | 0 | 0 | 9 |

Accuracy: 70.83

Table 9: Confusion matrix (after rejection framework)

| Emotion | A | B | D | F | H | N | S |
|---|---|---|---|---|---|---|---|
| Anger | 22 | 0 | 0 | 0 | 1 | 0 | 0 |
| Boredom | 0 | 12 | 0 | 0 | 0 | 0 | 0 |
| Disgust | 0 | 1 | 4 | 1 | 0 | 1 | 1 |
| Fear | 2 | 0 | 0 | 7 | 1 | 0 | 0 |
| Happy | 7 | 0 | 0 | 0 | 2 | 0 | 0 |
| Neutral | 0 | 4 | 0 | 0 | 0 | 7 | 0 |
| Sad | 0 | 1 | 0 | 1 | 0 | 0 | 8 |

Accuracy: 74.70

Table 10: Comparison of our system with existing works

| Article | Author | Recognition rate (%) |
|---|---|---|
| Acoustic emotion recognition: a benchmark comparison of performances | Schellur *et al*. (2009) | 73.20 |
| Emotion recognition in speech using MFCC and wavelet features | Krishna *et al*. (2013) | 51.00 |
| Two stage emotion recognition based on speaking rate | Koolagudi and Krothapalli (2011) | 63.61 for seven emotions |
| Proposed system | | 74.70 |

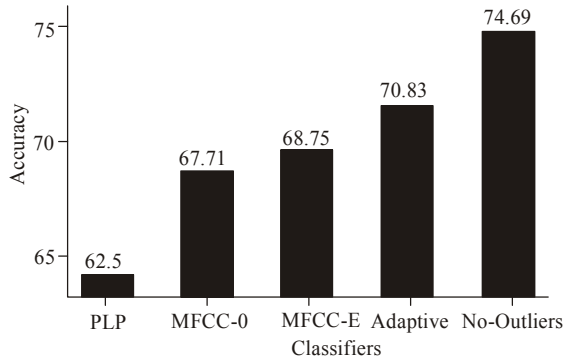


Fig. 5: Comparison of accuracy of classifiers

Table 10 presents the comparison of our system with some of the existing system based on recognition rate. The table illustrates, our classifier design performs better than other system reported in literature.

## CONCLUSION

We have implemented an adaptive sensor fusion technique for improving emotion recognition from speech. In this study a formula for calculating weight of classifiers involved in ensemble based on $f_1$ measure is also proposed. We have derived twelve different features from speech and built a classifier for each feature. The performance of classification based on different features was analyzed and best feature which produces better $f_1$ for a particular emotional class has been identified. We ensemble response of all individual classifiers responses trained with class specific features. Thus we have representation for all emotional classes. Some of the utterances may be noisy and couldn't express a specific emotion. We have devised a rejection framework, which rejects such confusing and noisy utterances as junk emotion. In EMO-DB three features MFCC12_0_D_A, MFCC12_E_D_A and PLP_0_D_A are identified as best features for recognition of different emotions. The responses of classifiers trained with these three emotions are combined. We have also proposed a rejection framework which rejection rejects some of the input utterances as outliers and the recognition accuracy obtained after rejection of outliers is 74.70%. Accuracy of emotion recognition is affected due to the acoustically resembling emotions. Happy is misclassified as anger and neutral is misclassified as boredom. The ambiguity of these resembling emotions have to be resolved using any other feature or any other classification system. The system developed is a generic one, can be tested with addition of any other acoustical feature for any corpus.

## ACKNOWLEDGMENT

We would like to thank the management of SSN College of Engineering for funding the High Performance Computing Lab, where this research was carried out. We also thank Dr. Chandrabose Aravindan and Dr. T. Nagarajan, Professors of SSN College of Engineering, for their valuable technical suggestions and motivation. We would like to thank the resource people of emotional corpus EMO-DB for providing their corpus for our research.

## REFERENCES


Bitouk, D., R. Verma and A. Nenkova, 2010. Class-level spectral features for emotion recognition. Speech Commun., 52(7-8): 613-625.

Böck, R., D. Hübner and A. Wendemuth, 2010. Determining optimal signal features and parameters for HMM-based emotion classification. Proceeding of the 15th IEEE Mediterranean Electrotechnical Conference (MELECON).

Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, 2005. A database of german emotional speech. Proceeding of the Interspeech. Lissabon, Portugal, pp: 1517-1520.

Burkhardt, F., M. van Ballegooy, K.P. Engelbrecht, T. Polzehl and J. Stegmann, 2009. Emotion detection in dialog systems: Applications, strategies and challenges. Proceeding of the Affective Computing and Intelligent Interaction and Workshops.

Chen, L., X. Mao, Y. Xue and L.L. Cheng, 2012. Speech emotion recognition: Features and classification models. Digit. Signal Process., 22(6): 1154-1160.

Corninna, C. and V. Vapnik, 1995. Support-vector networks. J. Mach. Learn., 20(3): 273-297.

Crammer, K. and Y. Singer, 2001.On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res., 2: 265-292.

Dellaert, F., T. Polzin and A. Waibel, 1996. Recognizing emotion in speech. Proceeding of the 4th International Conference on Spoken Language (ICSLP'96). Philadelphia, PA, 3: 1970-1973.

Dileep, A.D. and C. Chandra Sekhar, 2014. Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines. Speech Commun., 57: 126-143.

Eyben, F., F. Weninger, F. Gross and B. Schuller, 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. Proceeding of the 21st ACM International Conference on Multimedia (MM, 2013). Barcelona, Spain, pp: 835-838.

Fulmare, N.S., P. Chakrabarti and D. Yadav, 2013. Understanding and estimation of emotional expression using acoustic analysis of natural speech. Int. J. Nat. Lang. Comput., 2(4).

Giannakopoulos, T., P. Aggelos and T. Sergios, 2009. A dimensional approaches of emotion recognition from movies. Proceeding of the IEEE International Conference on Acoustic Speech Signal Processing.

Joachims, T., 1999. Making Large-Scale SVM Learning Practical. In: Schölkopf, B., C.J.C. Burges and A.J. Smola (Eds.), Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, MA.

Kamaruddin, N., A. Wahab and C. Quek, 2012. Cultural dependency analysis for understanding speech emotion. Expert Syst. Appl., 39(5): 5115-5133.

Kittler, J., M. Hatef, R.P.W. Duin and J. Matas, 1998. On combining classifier. IEEE T. Pattern Anal., 20(3).

Kobayashi, V.B. and V.B. Calag, 2013. Detection of affective states from speech signals using ensembles of classiers. Proceeding of the IET Intelligent Signal Processing Conference.

Koolagudi, S.G. and R.S. Krothapalli, 2011. Two stage emotion recognition based on speaking rate. Int. J. Speech Technol., 14(1): 35-48.

Krishna Kishore, K.V. and P. Krishna Satish, 2013. Emotion recognition in speech using MFCC and wavelet features. Proceeding of the IEEE 3rd International Advance Computing Conference (IACC).

Maaoui, C. and A. Pruski, 2008. A comparative study of SVM kernel applied to emotion recognition from physiological signals. Proceeding of the 5th International Multi-Conference on Systems, Signals and Devices, pp: 1-6.

Mehmet, S.U., O. Kaya and A. Coskun, 2009. Emotion recognition using neural networks. Proceeding of the 10th WSEAS International Conference on Neural Networks, pp: 82-85.

Milton, A. and S. Tamil Selvi, 2013. Class specific multiple classifiers scheme to recognize emotions from speech signal. Comput. Speech Lang., 28(3): 727-742.

Neiberg, D., K. Elenius and K. Laskowski, 2006. Emotion recognition in spontaneous speech using GMMs. Proceeding of the 9th International Conference on Spoken Language Processing (INTERSPEECH).

Polzhehl, T., A. Schmitt, F. Metze and M. Wagner, 2011. Anger recognition in speech using acoustic and linguistic cues. Speech Commun., 53(9-10): 1198-1209.

Rao, K.S. and S.G. Koolagudi, 2013. Robust emotion recognition using combination of excitation source, spectral and prosodic features. In: Sreenivasa Rao, K. and S.G. Koolagudi (Eds.), Robust Emotion Recognition using Spectral and Prosodic Features. SpringerBriefs in Electrical and Computer Engineering, pp: 71-84.

Schuller, B., S. Reiter and G. Rigoll, 2006. Evolutionary feature generation in speech emotion recognition. Proceeding of the IEEE International Conference on Multimedia and Expo.

Schuller, B., S. Reiter, R. Muller, M. Al-Hames, M. Lang and G. Rigoll, 2005. Speaker independent speech emotion recognition by ensemble classification. Proceeding of the IEEE International Conference on Multimedia and Expo.

Schuller, B., B. Vlasenko, F. Eyben, G. Rigoll and A. Wendemuth, 2009. Acoustic emotion recognition: A benchmark comparison of performances. Proceeding of the IEEE Conference on Automatic Speech Recognition and Understanding.

Shami, M. and W. Verhelst, 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun., 49(3): 201-212.

Tariq, U., L. Kai-Hsiang, Z. Li, X. Zhou, Z. Wang, V. Le, T.S. Huang, X. Lv and T.X. Han, 2011. Emotion recognition from an ensemble of features. Proceeding of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops.

Tin, L.N., S.W. Foo and L.C. De Silva, 2003. Speech emotion recognition using hidden marcov models. Speech Commun., 2-3: 603-623.

Trabelsi, I., D. Ben Ayed and N. Ellouze, 2013. Improved frame level features and SVM supervectors approach for the recogniton of emotional states from speech: Application to categorical and dimensional states. Int. J. Image Graph. Signal Process., 5(9): 8-13.

Valery, A.P., 1999. Emotion in speech: Recognition and application to call centers. Proceeding of the Conference on Artificial Neural Networks in Engineering, pp: 7-10.

Vasuki, P. and C. Aravindan, 2012. Improving emotion recognition from speech using sensor fusion techniques. Proceedings of the the IEEE Region 10 Conference TENCON, 2012.

Wu, S., T.H. Falk and W.Y. Chan, 2011. Automatic speech emotion recognition using modulation spectral features. Speech Commun., 53(5): 768-785.

Yuen, C.T., W.S. San, J.H. Ho and M. Rizon, 2013. Effectiveness of statistical features for human emotions classification using EEG Biosensors. Res. J. Appl. Sci. Eng. Technol., 5(21): 5083-5089.

Zixing, Z., F. Weninger, M. Wollmer and B. Schuller, 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding. Waikoloa, HI, pp: 523-528.