*Computer Systems Science & Engineering*
DOI: 10.32604/csse.2023.024441
***Article***

Tech Science Press

# Design of Hierarchical Classifier to Improve Speech Emotion Recognition

**P. Vasuki***

Department of IT, Sri Sivasubramaniya Nadar College of Engineering, Chennai, 603110, India
*Corresponding Author: P. Vasuki. Email: vasukip@ssn.edu.in
Received: 17 October 2021; Accepted: 13 December 2021

**Abstract:** Automatic Speech Emotion Recognition (SER) is used to recognize emotion from speech automatically. Speech Emotion recognition is working well in a laboratory environment but real-time emotion recognition has been influenced by the variations in gender, age, the cultural and acoustical background of the speaker. The acoustical resemblance between emotional expressions further increases the complexity of recognition. Many recent research works are concentrated to address these effects individually. Instead of addressing every influencing attribute individually, we would like to design a system, which reduces the effect that arises on any factor. We propose a two-level Hierarchical classifier named Interpreter of responses (IR). The first level of IR has been realized using Support Vector Machine (SVM) and Gaussian Mixer Model (GMM) classifiers. In the second level of IR, a discriminative SVM classifier has been trained and tested with meta information of first-level classifiers along with the input acoustical feature vector which is used in primary classifiers. To train the system with a corpus of versatile nature, an integrated emotion corpus has been composed using emotion samples of 5 speech corpora, namely; EMO-DB, IITKGP-SESC, SAVEE Corpus, Spanish emotion corpus, CMU's Woogle corpus. The hierarchical classifier has been trained and tested using MFCC and Low-Level Descriptors (LLD). The empirical analysis shows that the proposed classifier outperforms the traditional classifiers. The proposed ensemble design is very generic and can be adapted even when the number and nature of features change. The first-level classifiers GMM or SVM may be replaced with any other learning algorithm.

**Keywords:** Speech emotion recognition; hierarchical classifier design; ensemble; emotion speech corpora

## 1 Introduction and Motivation

Human-Computer-Interaction (HCI) aims to enhance the sophisticated usage of any system. If a system can to understand human emotions, the response could be much better [1]. Emotion Recognition has many practical applications. For instance, SER is used in an intelligent assistant to a learning system [2] or in understanding the usability experience of a toolkit/games from the user's conversation [3]. The accuracy of Speech Recognition also increases, when the word, decoded with the emotional content and context [4]. Speech Emotion Recognition (SER) research has been evolving for more than two decades [5]. The research on improving

the performance of SER has evolved in various directions as data collection, segmentation of data, feature extraction, classifier design, an ensemble of features and decisions in different aspects.

Early emotion recognition systems were based on temporal features of spectral features [6–8]. There are many other new acoustical features like low-level descriptive features (LLD) [9], and high-level prosody features like energy and pitch contours [10], linguistic & paralinguistic features were involved in emotion model building [11,12]. Later on, statistical variations like mean, median, high value and low value are added to the basic features of speech [13].

Various machine learning algorithms were involved in learning emotions from speech, such as K-Nearest Neighborhood Algorithm (KNN), Support Vector Classification algorithm (SVM) [14], Hidden Markov Model (HMM) [15] and, Ada Boosted Decision trees [16]. Some researchers applied the hierarchical emotion recognition technique, by using different ensemble techniques to improve the performance [17–19]. The system has been trained with low-level descriptive features like pitch, intensity and formant features with their variances. Some researchers tried to enhance the real-time emotion recognition with multiple corpora. Some of these systems were built to recognize arousal and valance factors across the corpus [20]. When emotion recognition is carried out in a mixed corpus, there is a possibility of data drift or bias on the dataset. Some researchers address this issue with domain-adaptation methods. An acoustic code vector enriched adverbially with a universal context database has been considered to represent emotion information in the specific dataset for the prediction of the context database [21,22]. After the revamp of Deep Neural Network, many researchers used various Deep Learning architectures like Convolution Neural Network, Recurrent Neural Network, LSTM etc. for SER [23–25]. In the deep learning approach the low-level input, features are given as an input and use Convolution Neural Network [26], Auto Encoder kind of techniques are to map the salient input features to recognize the emotion and these features are used to train CNN [27]. Also, Deep Nets lean a new feature vector from raw input speech for 1D architecture or the spectrogram image [28,29], log-Mel Spectrogram for 2D Network. To improve the accuracy some researchers were training Deep Networks to recognize emotions with acoustical cues and acoustical features separately [30]. Different networks are trained for learning local information from self-attention module and global cues from the social-attention module [31] and both the information are combined on the recognition of emotion.

Systems trained with one corpus recorded in a particular environment may not work well while testing with the corpus recorded in other environments. In research, many efforts are made to design a generic method, so that, systems evolve as speaker-independent and corpus independent with techniques like speaker normalization and corpus normalization [32,33] techniques. Schuler along with Zhang, Weninger and Wollmer have worked on cross-corpus emotion recognition using unsupervised learning [34]. Our research aim is to develop a robust SER system that performs well, despite the variations due to acoustical & cultural environment variations, length of utterance, language used, and speaker-related variations. To achieve this, we propose a two-level ensemble classification system.

## 2 Objectives of this Research

Our research aim is to develop a robust SER system that performs well, despite the variations due to acoustical & cultural environment variations, length of utterance, language used, and speaker-related variations. The variation due to these effects have been handled separately in the literature [35,36]. We have proposed a generic system to handle these changes. To achieve this, we propose a two-level ensemble classification system. The objectives of our research are to

- Design a Speech Emotion Recognition system that handles variations like Recording environment, Cultural environment, Gender of speaker and Age of speaker etc.
- Distinguish an emotion from acoustically resembling other emotions.

- Formation of integrated corpus which is versatile by combining different corpora of different cultural backgrounds, recording condition.

## 3 Hierarchical Classification

The primary issue of generic SER is variations of the recording environment, age & gender of the speaker and the acoustical resemblance of one emotion with others. To address this issue, we have planned to design an ensemble classifier.

### *3.1 Suitability of GMM & SVM Classifiers for Ensemble*

The ensemble classifier may perform better when the first level classifiers are diverse. The diverse nature of generative classifier GMM and discriminative classifier SVM are analyzed. A good classification system produces the right response for acoustically strong utterances. The system may generate response for weak utterance, based on the classification algorithm. A good classification system produces the right response for the utterances which are acoustically strong. The system may generate a response for weak utterance, based on the classification algorithm. SER system using GMM & SVM is built with EMO-DB corpus and MFCC_D_A feature vector. During testing, for a given test utterance, The GMM classifier generates maximum likelihood values and SVM generates a confidence score for every emotion class. These functional values generated by both the classifiers for emotions Anger, Fear, Happiness and Neutral samples are presented in Figs. 1–4 respectively. Both SVM and GMM are classifying most of the samples correctly. As There is acoustical resemblance among some of the emotions, which is not well discriminated in the utterances, any one of the classifiers may wrongly interpret it. When wrongly classified samples are not coinciding, then such classifiers will be suitable for the ensemble. The maximum likelihood values produced by GMM and SVM classifiers for four emotions; anger, fear, happiness and neutral for anger samples. The graph has been plotted against sample number and Maximum Likelihood Value. The response of SVM is presented in blue color whereas GMM response in red color and plotted against the sample number. Different style of line style has been used to connect the responses generated for the various classes. The response has been compared with ground truth and found different samples were false negative for the different classifier. During False-negative cases, the behavior of both the classifiers is different, whereas for most true positive cases both the classifiers are coinciding with each other.

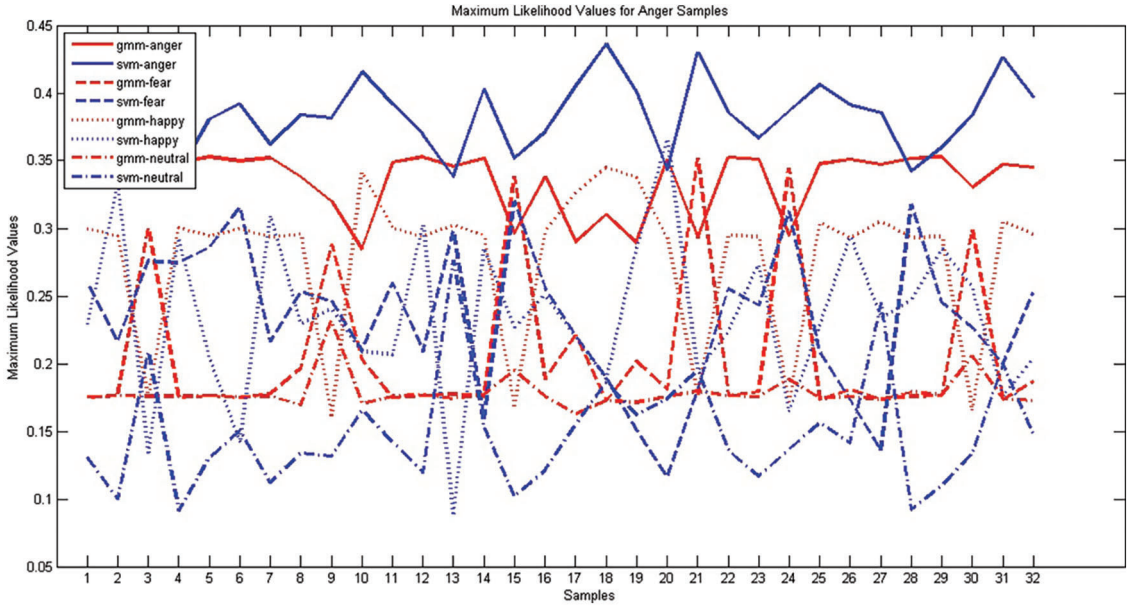


**Figure 1:** Comparison of GMM, SVM classifiers using samples of the emotion 'Anger' EMO-DB corpus

In Fig. 1, the samples numbered 9, 15, 21, 24 GMM result in false negatives. These samples were not recognized as anger and but give the second highest likelihood value for anger, whereas the SVM classifier correctly recognizes the emotion of these samples. Sample 9 was interpreted as happy and 15, 21 and 24 were interpreted as fear by GMM. The sample numbered 20, was wrongly classified as happy by SVM, whereas GMM correctly recognized it. For instance, for the 20'th sample, with reference to SVM's response, anger's likelihood value is only 0.02 less than happy. GMM produces the highest likelihood for anger and the second-highest value for happiness with the difference of 0.05, which shows that anger closely resembles happiness in the response.

Fig. 2 shows a comparison of maximum likelihood values generated for different emotions by The GMM and SVM classifiers. Samples 2, 9, 11 and 16 of the Fear class have been wrongly classified by SVM but correctly classified by GMM. Samples 3, 6, 7, 8, 13 are wrongly classified by GMM but rightly recognized by SVM. Samples 10, 12 are labeled as neutral by both the classifiers. Sample 16 has been identified as happy by SVM and angry by GMM. Fig. 3 shows the response of the classifier on Happy Samples. Samples numbered 4, 12, 15, 801 are the wrong coincidences of GMM and SVM classifiers. SVM fails to recognize samples numbered 2, 3, 8, 10, 13 but GMM succeeds, whereas samples numbered 5, 6, 9 are classified wrongly by GMM and rightly by SVM classifiers. The maximum likelihood values of GMM & SVM Classifiers for four emotions obtained when tested with Neutral samples are presented in Fig. 4. In The case of Neutral, GMM recognized all samples correctly, SVM also correctly recognized all samples except sample 7. The overall performances of both the classifiers were good. Due to environmental factors or acoustical resemblance, some samples were wrongly classified by one classifier, while another classifier recognizes some other sets of samples wrong. However, in most of the false-negative cases, the classifier produces at least, second-highest likelihood values for actual emotion. Thus, when these interpretations are used as additional evidence to train the second level classifier, the performance of ensemble learners will be higher.

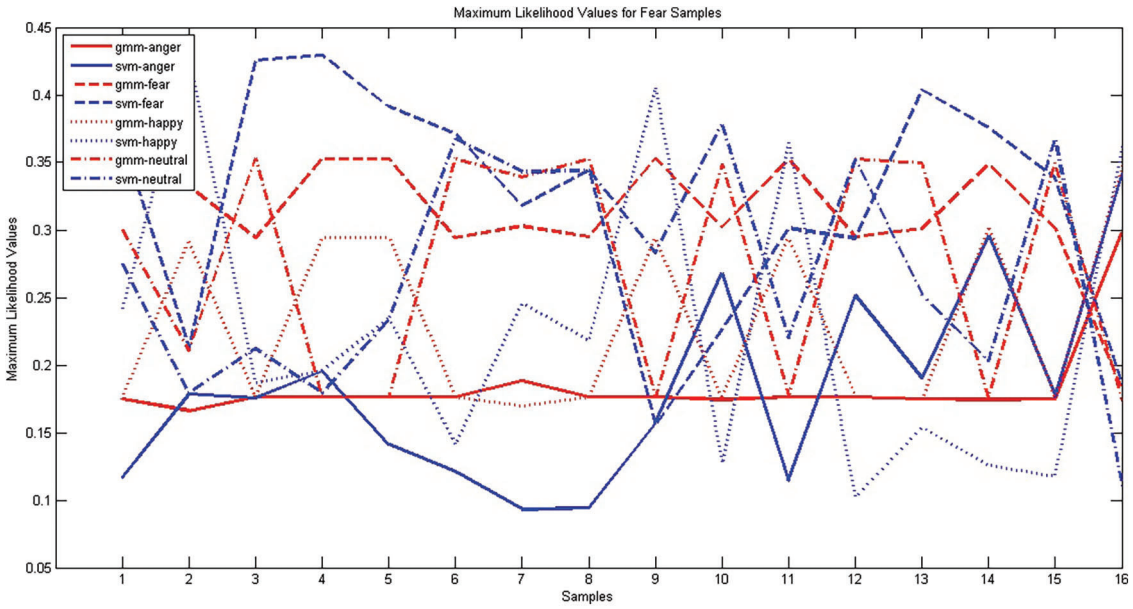


**Figure 2:** Comparison of GMM, SVM classifiers using samples of the emotion 'Fear' EMO-DB corpus

While noticing the responses of the wrongly classified utterances of SVM and GMM Classifiers, one classifier produces positive peaks while the other one either gives flat or negative peaks. Thus when these

two classifiers responses are directly added it may produce a neutral effect. Instead, if this pattern is given for a third classifier, it may learn to interpret the right response. An ideal emotion recognition system must recognize emotions present in input utterances irrespective of these variations. The literature shows that the ensemble of data, features and responses of preliminary classifiers enhances the performance of SER. The meta-learning approaches used in classifier fusion have been used to model emotions from a high-level abstraction of utterances. As meta-learner combines the knowledge from independent learning, it is expected to be a higher level learned model to explain emotion classes well. Thus GMM & SVM Classifiers are found to be suitable for designing ensemble classifiers.

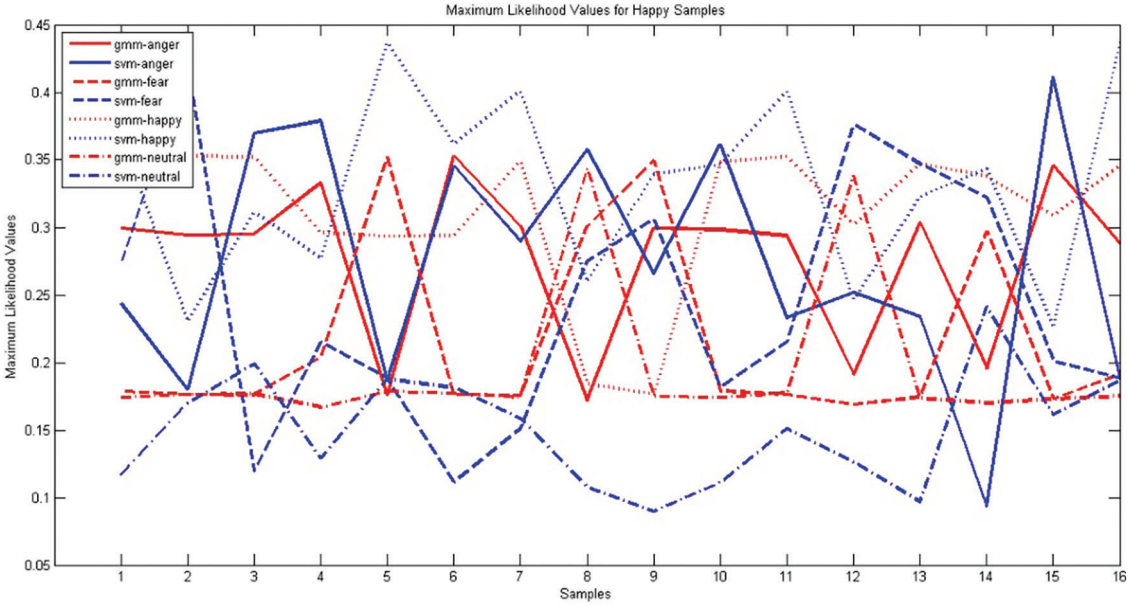


**Figure 3:** Comparison of GMM, SVM classifiers using samples of the emotion 'happy' EMO-DB corpus

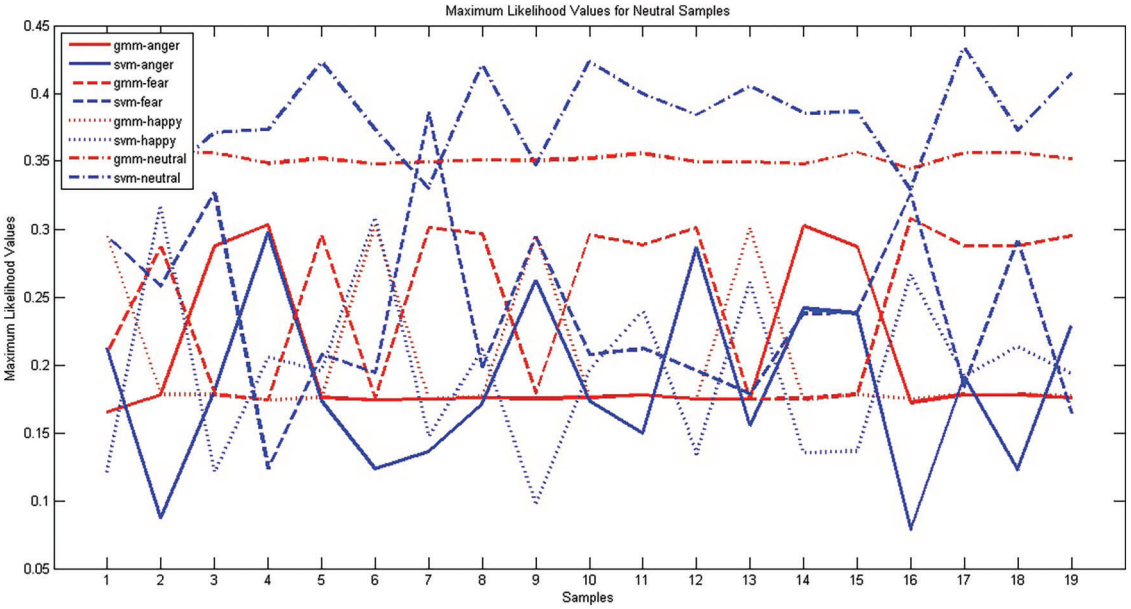


**Figure 4:** Comparison of GMM, SVM classifiers using samples of the emotion 'neutral' EMO-DB corpus

### *3.2 Hierarchical Classification Using Interpreter of Responses (IR)*

In this method, the SER system has been designed as an ensemble classifier for emotion recognition from speech. From the above analysis, we found GMM & SVM are suitable for the ensemble. Thus in the first level classifications, we used GMM & SVM and trained them with a standard feature to classify emotions. In the second level, we used SVM, as it is a good discriminating classifier, which classifies emotions, based on interpretation of first level classifiers. SVM has been trained with the responses of first level classifiers on given input along with the input features to decode emotion.

The functionality of the system is described in Fig. 5. Both of the first level classifiers are trained with the four emotion classes. The GMM classifier generates a Model for every emotion class.

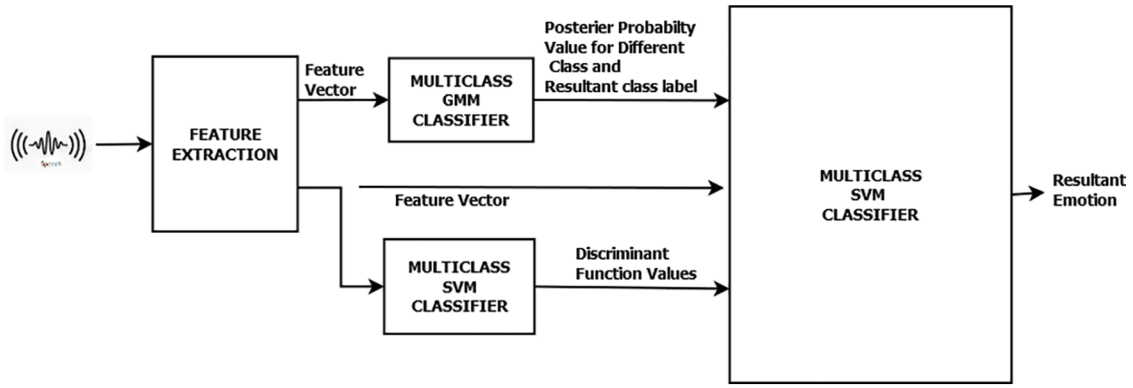


**Figure 5:** Framework of the proposed classifier

Eq. (1) represents the parametric GMM,

$$N(o,\ \mu,\ \sum) = \frac{1}{(2\pi)^{D/2}|\varepsilon^{-1}|^{1/2}} e^{-1/2(o-\mu)^T} \sum^{-1}(o-\mu) \tag{1}$$

where'o'-observation vector

μ-mean vector

∑-covariance matrix

When tested with test utterances, the GMM classifier generates the probability of the utterance belonging to the model of a given category. The probability distribution of the test sample ' x ' over 'N' components is calculated using Eq. (2).

$$P(x) = \sum_{n=1}^{N} \alpha_n N(x/\mu_n,\ \varepsilon) \tag{2}$$

where N-number of Gaussian mixtures.

$\alpha_n$ is the prior probability and (mixture weight), $0<\alpha_n<1$.

Thus GMM provides a probability value of test utterance belonging to different classes.

Similarly, SVM Classifier is used to model emotions. The linear SVM has many advantages-It is steady, doesn't over-fit and has therefore been used in our experiments. SVM takes an input feature vector and generates models for every emotion class. Using these emotion models, SVM classifies input utterances and also produces a discriminant function value, which comprises a score value for every emotion class.

The mapping function is f(x, w) = argmax(y), where w is the linear function of x and y. x is the corresponding input feature vectors, and y is the emotion class label. Finally, the test sample takes up the emotion label of the class, which produces maximum functional value among all the emotion classes. The radial basis function is given by Eq. (3).

$$f(x) = \beta_0 \sum_{i \in S}^{N} \alpha_n K(x,\ x_i) \tag{3}$$

x is the feature vector extracted from the test utterance.

$x_i$ is the feature vector extracted from the i'th sample in set S belongs to an emotion class.

Where $\alpha_n$ is the learning parameter.

$$k(x, xi) = \exp\left(-\frac{\|x - xi\|^2}{2\varepsilon}\right) \tag{4}$$

Whereas $K(x, x_i)$ is the Radial basis Kernel given by Eq. (4).

The level 2 classifier SVM has been trained with the feature vector formed using three components; namely discriminant function values of SVM, acoustical likelihood functional value of GMM, and Acoustical feature vectors of input speech. Interpreter of Responses (IR) has been trained with these input feature vectors. On training, support vectors are formed and the system predicts the test sample based on the distance with these vectors. IR maps the input speech to an emotion label which has a higher support value for the input sample among all emotional categories.

## 4 Empirical Environments

This section deal with the corpus, features and experimental details like the ratio of training and testing files and the environment.

### *4.1 Speech Corpora*

The extensive empirical analysis was carried out with the various emotional speech corpora namely; EMO-DB, SAVEE Database, Spanish Emotional Speech Corpus, IITKGP-SESC, and CMU's Woogles corpus evaluate the performance of the classifiers. The details of those corpora are provided in Tab. 1.

**Table 1:** Database description

| Corpus | Source | Number and list of emotions | No. of samples (For 4 emotions) | # Speakers |
|---|---|---|---|---|
| EMO-DB | [36] | 7 anger, boredom, fear, happy, sadness and neutral | 341 | 5 male, 5 female |
| IITKGP-SESC | [37] | 8 anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise | 240 for a speaker | 5 male, 5 female |
| SAVEE | [38] | 7 anger, disgust, fear, happy, sad, surprise and neutral | 299 | 4 male |
| CMU's Woogles | [39] | 5 happy, sad, anger, fear and neutral | 918 | 7 female |
| Spanish emotional corpus | [40] | 5 anger, fear, neutral and happy | 3850 | a male & a female |

### *4.2 Integrated Corpus*

A strong system to predict emotion class from speech may be built using a corpus with wide variations. Researchers are working with merging two different corpora and allowing two classes of emotions for the same emotional category (example berlin_Anger and Danish_Anger). Enormous variations are present in an integrated corpus, thus acts a good base for real-time application development. To achieve such corpus, we integrated the above said 5 corpora into a single one by combining utterances from alike emotion categories of different corpora into a unique category. For example, afraid of Woogles corpus and fear of other corpora are combined as a unique fear collection. Thus the proposed SER system developed this integrated corpus of versatile is expected to be more robust than the system built on the single corpus.

### *4.3 Acoustical Features*

The experiments were carried out to build and evaluate conventional systems and proposed systems using standard feature MFCC_D_A and the paralinguistic feature set EC-2010 provided for Emotion Challenge conducted in Inter speech conferences in the year 2010.

The acoustical feature MFCC is found to be a simple, reasonably good, and dominant feature for analyzing emotions in speech. The empirical analysis carried out shows that, 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. We considered 39 dimensions of MFCC vectors with 13 MFCC coefficients along with 13 delta coefficients and 13 acceleration coefficients.

The feature set EC 2010 contains Low-Level Descriptors (LLD) like Intensity, Loudness, 12 MFCC, Pitch ($F_0$), The feature set also includes delta coefficients computed from LLD and the functional values like Min/Max, arithmetic mean. Finally, there are 1582 features formed from these LLD features.

### *4.4 Experimental Environments*

The cross-validated data partition ensures the stability of the system. The samples to form the training set for every emotion class was formed by picking samples at arbitrary style. Also, it is ensured that the set covers utterances from all the speakers. With each fold of corpus, experiments were carried out with 4 emotion categories which are common across all corpora, GMM, SVM and Fused classifier IR separately. Further tests were carried out by having 75% of training files and 25% testing files. The reported results were obtained from the 75:25 strategic partition. The evaluation parameters recall, precision, accuracy and $F_1$measure are used for the performance comparison analysis.

Experiments are carried out with the following objectives:

- To ascertain that the proposed hierarchical classification system achieves better performance than traditional monolithic systems in a specific corpus environment.
- To ascertain that the proposed hierarchical classifier system performs better than conventional monolithic classifiers in an integrated corpus environment.
- To ascertain that the proposed system outperforms existing systems available in the literature.

#### *4.4.1 Corpus Specific Emotion Recognition*

The corpus-specific emotion recognition system is built using a specific emotion corpus. These systems were built individually with the above said five corpora. Tab. 2 shows the recall, accuracy, precision & $F_1$ measure produced by conventional classifiers and IR classifier when tested in a specific corpus environment using feature vector MFCC_D_A.

**Table 2:** Performance of GMM, SVM, IR classifiers when employed on different corpora & feature vector MFCC_D_A

| Recall | | | | | | |
|---|---|---|---|---|---|---|
| Classifier/DB | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 72.66 | 76.67 | 67.47 | 60.81 | 77.74 | 71.07 |
| SVM | 74.22 | 91.67 | 68.28 | 65.04 | 78.25 | 75.49 |
| IR | 81.5 | 91.67 | 91.67 | 87.88 | 85.69 | 87.68 |
| **Gain of IR** | **9.81** | **0** | **34.26** | **35.11** | **9.5** | **17.73** |
| Accuracy | | | | | | |
| Classifier/DB | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 77.11 | 76.67 | 72.97 | 53.95 | 78.77 | 71.89 |
| SVM | 79.52 | 91.67 | 72.97 | 67.98 | 82.62 | 78.95 |
| IR | 84.34 | 91.67 | 93.24 | 86.4 | 88.14 | 86.76 |
| **Gain of IR** | **6.06** | **0** | **27.78** | **27.1** | **6.68** | **13.52** |
| Precision | | | | | | |
| Classifier/DB | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 77.45 | 77.55 | 67.7 | 55.49 | 76.96 | 71.03 |
| SVM | 76.84 | 91.67 | 67.74 | 72.1 | 80.26 | 77.72 |
| IR | 85.37 | 0 | 91.67 | 85.95 | 86.81 | 86.58 |
| **Gain of IR** | **10.22** | **0** | **35.84** | **19.2** | **8.17** | **14.69** |
| $F_1$ measure | | | | | | |
| Classifier/DB | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 74.98 | 77.11 | 67.59 | 58.03 | 77.35 | 71.01 |
| SVM | 75.5 | 91.65 | 68.01 | 68.39 | 79.24 | 76.56 |
| IR | 83.39 | 91.67 | 91.84 | 86.9 | 86.25 | 86.12 |
| **Gain of IR** | **10.44** | **0.02** | **35.05** | **27.06** | **8.84** | **16.28** |

Concerning recall value, the hierarchical classifier is always a boon to the base level classifier. In the case of SAVEE and CMU's Woggles corpus, the yield is too high at the maximum of 35.11 and an average gain of 17.73%.

The accuracy obtained by Hierarchical classifier IR in comparison with GMM/SVM classifier is better than the best. Depending upon the variation in the expression, the yield of IR varies. For instance, in the Spanish corpus, the number of speakers is two, IR gets little improvement than the performance of the traditional classifier. The IIT corpus was recorded in a noise-free environment, and the corpus available with us consists of only single speaker recordings.

Thus we have obtained good performance metrics from all the classifiers compared to other corpora, despite less number of training samples involved in this corpus (Total samples 60 per class, which is lesser than other corpora). IR retains the best accuracy.

Even if one of the classifiers is performing well, the hierarchical classifier gets the influence of a better recognizing classifier. In the case of the SAVEE database, the average precision obtained by GMM and SVM are equal, but the percentage of recognition is not evenly obtained from 4 emotional classes in these classifiers. The accuracy obtained for Fear is found to be superior at the GMM classifier, where the SVM classifier's accuracy toward happiness is better than GMM. Thus, the hierarchical classifier IR is functioning well in these situations to remove the ambiguity with a high relative yield.

Similar to recall, the gain of accuracy is high in SAVEE, and Woogles corpus reaches a maximum gain of 27.78% for SAVEE, and the overall average gain from all 5 corpora is 13.52%.

Precision identifies the truth of the prediction, the percentage of the correctness if the response of the classifier. In the SAVEE database, the precision obtained by IR is better than the other classifiers by 35.84%, which is the maximum gain among all corpora. On the average of all corpora, IR gains 14.69% compared to traditional classifiers.

A good classifier should not only, recognize, its class, but also reject the negative samples. The $F_1$ measure measures the harmonic mean between accuracy and precision. $F_1$ measure shows how well does the system reject negative example and how well does it identify a positive example. Hierarchical Classifier-IR is performing better than conventional classifiers in this $F_1$ measure too and provides an average improvement of 16.28%.

The results show that the system works well when built using standard feature MFCC_D_A and the stability of the system has been validated further using feature set EC-2010. Various performance metrics of the systems namely accuracy, recall, F1 score are presented in Tab. 3. The results show that the hierarchical classifier IR is functioning better than conventional classifiers concerning all evaluation parameters in a specific-corpus environment.

**Table 3:** Comparison of classifiers using various parameters with different corpora using feature set **EC2010**

| Accuracy | | | | | | |
|---|---|---|---|---|---|---|
| Classifier | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 78.31 | 93.33 | 78.38 | 62.28 | 79.9 | 78.44 |
| SVM | 83.13 | 93.33 | 83.78 | 67.54 | 87.92 | 83.14 |
| IR | 86.75 | 93.33 | 93.33 | 86.84 | 88.13 | 89.68 |
| **Gain of IR** | **4.35** | **0** | **11.4** | **28.57** | **0.24** | **8.91** |
| Precision | | | | | | |
| Classifier | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 77.3 | 93.32 | 79.97 | 62.68 | 77.11 | 78.08 |
| SVM | 84.93 | 93.32 | 81.93 | 67.67 | 86.8 | 82.93 |
| IR | 87.51 | 93.32 | 92.59 | 86.26 | 87.22 | 89.38 |
| **Gain of IR** | **3.03** | **0** | **13.01** | **27.48** | **0.48** | **8.8** |
| Recall | | | | | | |
| Classifier | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 74.18 | 93.33 | 73.33 | 65.18 | 77.89 | 76.78 |
| SVM | 79.4 | 93.33 | 80.8 | 64.72 | 85.54 | 80.76 |

(Continued)

**Table 3 (continued)**

| Recall | | | | | | |
|---|---|---|---|---|---|---|
| Classifier | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| IR | 83.86 | 93.33 | 91.88 | 88.27 | 85.82 | 88.63 |
| **Gain of IR** | **5.61** | **0** | **13.7** | **35.43** | **0.33** | **11.01** |
| $F_1$ measure | | | | | | |
| Classifier | EMO-DB | IIT | SAVEE | Woo | Spanish | Average |
| GMM | 75.71 | 93.33 | 76.51 | 63.91 | 77.5 | 77.39 |
| SVM | 82.07 | 93.33 | 81.36 | 66.16 | 86.16 | 81.82 |
| IR | 85.65 | 93.33 | 92.23 | 87.25 | 86.51 | 88.99 |
| **Gain of IR** | **4.35** | **0** | **13.36** | **31.88** | **0.4** | **10** |

*4.4.2 Emotion Recognition in Integrated Corpus*

The Tab. 4 shows the consolidated performance of the various classifiers when trained and tested on an integrated corpus. The performance of IR is better than the primary level classifier used in system building in both feature sets MFCC_D_A as well as in the EC-2010 Feature set. The EC-2010, recognize emotion better than MFCC_D_A.

**Table 4:** Performance of classifiers on integrated corpus

| Performance on classifiers employed with MFCC_D_A feature set | | | |
|---|---|---|---|
| Parameters | GMM | SVM | IR |
| Accuracy | 64.51 | 69.35 | 74.18 |
| Precision | 62.64 | 66.88 | 72.54 |
| Recall | 62.31 | 66.37 | 71.97 |
| $F_1$ | 62.48 | 66.62 | 72.25 |
| Performance on classifiers employed with 'EC-2010' feature set | | | |
| Classifier/parameter | GMM | SVM | IR |
| Accuracy | 64.98 | 75.37 | 76.23 |
| Precision | 63.75 | 74.11 | 74.97 |
| Recall | 62.9 | 73.24 | 73.93 |
| $F_1$ | 63.32 | 73.67 | 74.45 |

Results of our systems are compared with published results of existing systems on emotion recognition when tested with appropriate corpora and presented in Tabs. 4 and 5. From the Tables of the result, it is observed that our system is performing significantly better compared to existing systems in a specific corpus environment.

From these observations, it is found that the hierarchical hybrid ensemble classifier outperforms in every case, concerning the evaluation parameter recall, precision, accuracy and$F_1$ measure. The results show that on average, in the corpus specific environment, our hierarchical classifier performance is improved by 13.23% for SVM and 17.57% concerning the GMM classifier as shown in Tab. 3.

**Table 5:** Comparison of the proposed system with existing systems in mixed corpus environment

| Article | Corpora | Methodology | Emotional class description | Accuracy |
|---|---|---|---|---|
| [18] | EMO-DB, NTU-Asian dataset, NTU-American dataset | Multilevel approach using ANFIS, GenSoFNN and MLP | 4 classes; anger, fear, happy, neutral. | 59.5% was obtained with a mixture of 3 corpora & 64.6% is for 2 corpora mixture with strata of test: train 1:1 |
| [31] | BabyEars & Kismet, Berlin-Danish emotional corpus | Segment based approach employed in integrated corpus | Berlin_neutral, happy, anger and sadness, Danish_neutral, happy, anger and sadness | 72.2% using SVM classifiers |
| [33] | BabyEars & Kismet | SBA in IC | 6 Classes approval, attention prohibition_Kismet, and approval, attention, prohibition_Babyears | 74.70% |
| [41] | Spanish, English, German and Arabic | Subjective analysis | 6 emotional classes | 42% with monolingual speakers |
| [42] | IEMOCAP, FAU AIBO | MFCC, SDC/ DNN | 7 | 51.5 |
| [43] | Berlin, Beumo, DES, RUSL | Speaker, power AND L2 normalization | 8 | 62.2 |
| **Proposed system** | **Integrated corpus** | **IR** | **4 emotions** | **74.18% strata of 1:4** |

### *4.5 Comparison with Existing Systems*

This section provides a comparative analysis of the proposed system with the state of art systems. The comparison is based on the factors; conditions used for selecting a strategic partition of training files and testing files, and the methodology used for building the system. Emotion varies across various parameters, to reduce the effect on variations the new classifier has been designed. The system has been compared with various existing systems in literature. Though many systems have been developed in specific-corpus environments, the emotion recognition in wild or in real-time work is comparatively lesser. To achieve that, we have taken the mixed corpus environment to test the emotion. Here the system

has been trained and tested with a corpus of different acoustical, cultural and linguistic environments. The proposed system is providing better accuracy compared to the other system present in the literature.

## 5 Conclusion

The background is defined as the recording environment, gender, age language of the speaker and the text utterance. Since, the background, has a strong influence on emotion recognition, the accuracy reduces when training and testing corpora are of different backgrounds. Thus we formed an integrated corpus by merging various corpora, reordered from different backgrounds. SER systems modeled with mixed corpora is more versatile in recognizing emotion from different backgrounds.

An ensemble technique has been proposed to address this issue. The proposed system is a two-level classification system. At the primary level of the classifier is built with SVM and GMM classifiers. In the secondary level, the intermediate responses of primary classifiers are handled using an ensemble classifier-Interpreter of Responses (IR)-a multi-class SVM. As the intermediate responses of primary classifiers are used to train the second level classifier. Consequently, the response of the final classifier is expected to be better even if there is a less negative correlation on direct outputs of primary classifiers of the test utterances.

EMO-DB, IITKGP-SESC, Spanish Emotional Corpus (ELRA-S0329), SAVEE, and CMU'S Woggles corpus were used for the extensive empirical analysis in a specific corpus and integrated corpus environments. It is observed from the experiment that there is an improvement in performance regarding the accuracy, precision, recall, and F1 measure. The average of specific corpus environments of 5 above said corpora.

GMM and SVM classifiers recognize emotions with the accuracy of 71.89% and 78.95% respectively, whereas IR approaches 86.76%. IR method also produces an improvement of 14.22% in F1 measure over conventional approaches. Furthermore, the performance of the proposed system is better than the published results of SERs in multilingual and mixed corpus environments too.

The proposed approach produces an accuracy of 74.18% in an integrated corpus developed from 5 corpora, while 74.7% was reported recognition accuracy with a mixture of two corpora, and 18% accuracy obtained when emotion was recognized against two corpora of two different regional languages. To conclude, the proposed classifier is stable & generic despite changes in the recording environment.

The proposed work may be tested with some more classifiers as primary classifiers. The work may be extended by synthesizing augmented data for the corpora which are weak in the number of samples. The generic nature of the proposed classifier has to be enriched by increasing various corpora, so that the system may be suitable for recognizing spontaneous real-time emotion accurately.

**Acknowledgement:** We would like to thank the management of SSN College of Engineering. We also thank Dr Chandrabose Aravindan Professor & Head of IT, SSN College of Engineering and Dr T. Nagarajan, Professor & Head of CSE, SNU, for the motivation and encouragement provided. We would like to thank the resource people of emotional corpora EMO-DB, IITKGP-SESC, SAVEE, Spanish Emotion Corpus, SAVEE and Woogles Corpus for providing their corpus for our research.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References


[1] M. Anjum, "Emotion recognition from speech for an interactive robot agent," in *IEEE/SICE Int. Symp. on System Integration (SII)*, Paris, France, pp. 363–368, 2019.

[2] A. Arguel, L. Lockyer, G. Kennedy, J. M. Lodge and M. Pachman, "Seeking optimal confusion: A review on epistemic emotion management in interactive digital learning environments," *Interactive Learning Environments*, vol. 27, no. 2, pp. 200–210, 2019.

[3] J. Jia and X. Dong, "User experience classification based on emotional satisfaction mechanism," in *Int. Conf. on Human-Computer Interaction*, Orlando, FL, USA, pp. 450–459, 2019.

[4] M. Bojanić and V. Delić, "Automatic emotion recognition in speech: Possibilities and significance," *Electronics*, vol. 13, no. 2, pp. 35–40, 2009.

[5] I. R. Murray and J. L. Arnott, "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*, vol. 93, no. 2, pp. 1097–1108, 1993.

[6] R. Böck, D. Hübner and A. Wendemuth, "Determining optimal signal features and parameters for hmm-based emotion classification," in *Melecon 2010-2010 15th IEEE Mediterranean Electrotechnical Conf.*, Valletta, Malta, pp. 1586–1590, 2010.

[7] S. Yun and C. D. Yoo, "Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, pp. 4169–4172, 2009.

[8] K. K. Kishore and P. K. Satish, "Emotion recognition in speech using MFCC and wavelet features," in *3rd IEEE Int. Advance Computing Conf. (IACC)*, Ghaziabad, India, pp. 842–847, 2013.

[9] A. Milton and S. T. Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," *Computer Speech & Language*, vol. 28, no. 3, pp. 727–742, 2014.

[10] K. S. Rao and S. G. Koolagudi, "Robust emotion recognition using combination of spectral and prosodic features," *Springer Briefs in Electrical and Computer Engineering*, pp. 71–84, 2013, ISBN: 9781461463603, 1461463602.

[11] D. Bitouk, R. Verma and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[12] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[13] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang *et al.,* "Speaker independent speech emotion recognition by ensemble classification," in *IEEE Int. Conf. on Multimedia and Expo*, Amsterdam, Netherlands, pp. 864–867, 2005.

[14] M. Shami and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication*, vol. 49, no. 3, pp. 201–212, 2007.

[15] T. L. Nwe, S. W. Foo and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[16] C. C. Lee, E. Mower, C. Busso, S. Lee and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9–10, pp. 1162–1171, 2011.

[17] P. Vasuki and C. Aravindan, "Improving emotion recognition from speech using sensor fusion techniques," in *TENCON IEEE Region 10 Conf.*, Cebu, Philippines, pp. 1–6, 2012.

[18] N. Kamaruddin, A. Wahab and C. Quek, "Cultural dependency analysis for understanding speech emotion," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5115–5133, 2012.

[19] Mustaqeem and S. Kwon, "1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features," *Computers Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[20] A. Meftah, Y. Seddiq, Y. Alotaibi and S. Selouani, "Cross-corpus Arabic and English emotion recognition," in *IEEE Int. Symp. on Signal Processing and Information Technology (ISSPIT)*, Bilbao, Spain, pp. 377–381, 2017.

[21] C. M. Chang and C. C. Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," in *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, pp. 5820–5824, 2017.

[22] C. M. Chang and C. C. Lee, "Adversarially-enriched acoustic code vector learned from out-of-context affective corpus for robust emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 7395–7399, 2019.

[23] J. Kim, K. P. Truong, G. Englebienne and V. Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition," in *Seventh Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, TX, USA, pp. 383–388, 2017.

[24] Y. Zhang, Y. Liu, F. Weninger and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, pp. 4990–4994, 2017.

[25] T. Polzehl, A. Schmitt, F. Metze and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Communication*, vol. 53, no. 9-10, pp. 1198–1209, 2011.

[26] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[27] Z. Huang, M. Dong, Q. Mao and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. of the 22nd ACM Int. Conf. on Multimedia*, Orlando, Florida, USA, pp. 801–804, 2014.

[28] N. Anand and P. Verma, S, "Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data," *Technical Report*, pp. 1–6, 2015.

[29] M. Mustaqeem and S. Kwon, "Speech emotion recognition based on deep networks: A review," in *Proc. of the Korea Information Processing Society Conf.*, Korea, pp. 331–334, 2021.

[30] M. Mustaqeem and S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 5116–5135, 2021.

[31] M. Mustaqeem and S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, pp. 1–11, 2021.

[32] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," *Speaker Classification II*, pp. 43–56, 2007. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-74122-0_5.

[33] M. Bhaykar, J. Yadav and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM," in *National Conf. on Communications (NCC)*, New Delhi, India, pp. 1–5, 2013.

[34] Z. Zhang, F. Weninger, M. Wöllmer and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, Waikoloa, HI, USA, pp. 523–528, 2011.

[35] P. Vasuki, "Speech emotion recognition using adaptive ensemble of class specific classifiers," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 9, no. 12, pp. 1105–1114.

[36] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, "A database of German emotional speech," in *INTERSPEECH 2005-Eurospeech, 9th European Conf. on Speech Communication and Technology*, Lisbon, Portugal, pp. 1517–1520, 2005.

[37] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti and K. S. Rao, "IITKGP-SESC: Speech database for emotion analysis," in *Int. Conf. on Contemporary Computing*, Noida, India, pp. 485–492, 2009.

[38] S. Haq, P. J. Jackson and J. Edge, "Speaker-dependent audio-visual emotion recognition," in *Int. Conf. on Audio-Visual Speech Processing*, Norwich, UK, pp. 53–58, 2009.

[39] F. Dellaert, T. Polzin and A. Waibel, "Recognizing emotion in speech," in *Proc. of Fourth Int. Conf. on Spoken Language Processing, ICSLP'96*, Philadelphia, PA, USA, pp. 1970–1973, 1996.

[40] J. M. Montero, J. M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera *et al..,* "Emotional speech synthesis: From speech database to TTS," in *Fifth Int. Conf. on Spoken Language Processing*, Sydney, Australia, pp. 923–926, 1998.

[41] M. D. Pell, L. Monetta, S. Paulmann and S. A. Kotz, "Recognizing emotions in a foreign language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, 2009.

[42] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub *et al.,* "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, no. 1, pp. 22–35, 2019.

[43] H. Kaya, and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, no. 1, pp. 1028–1034, 2018.