Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

AUDIO BASED EMOTION CLASSIFICATION USING ENSEMBLE CLASSIFIER

Abdul Mudassar[1], Sana Ul Haq[2], Muhammad Imran Majid[3,*], Imtiaz Rasool[4], Muhammad Kamran[5]
[1] *Department of Computer Science, Edwardes College Peshawar, Pakistan*
[2, 4, 5] *Department of Electronics, University of Peshawar, Pakistan*
[3] *University of Warwick, UK*

[*]Corresponding author: *m.majid.1@warwick.ac.uk*

Abstract: **Recognition of emotion from speech is a challenging area of research. It is hard for the single classifier to provide better classification accuracy. For this reason, in recent years researchers have focused on ensemble classifier techniques to combine the best results of individual classifiers in an effective way to achieve overall higher classification performance. This paper presents a novel approach to combining classifiers outputs for audio emotion recognition. In this approach, the best results obtained for different emotion classes from various classifiers are combined to create a combined confusion matrix. It is because some classifiers with overall lower performance have better accuracy for a specific class as compared to others with overall higher accuracy. The performance of this approach was analyzed using three emotional speech databases in different languages, i.e., Berlin emotional speech database (EMO-DB), Italian emotional speech database (EMOVO-DB), and Surrey audio-visual expressed emotion database (SAVEE-DB). The openSMILE toolkit was used to extract a total of 8543 audio features. These features include pitch, energy, intensity, jitter, shimmer, formants, zero crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCCs), Mel-frequency bands (MFBs), line spectral pairs (LSPs) and spectral features. These features were normalized using the min-max normalization technique, while correlation-based feature selection (CFS) with a best-first search approach was used for feature reduction. The classification was performed using five different base classifiers, i.e., support vector machine (SVM), multi-layer perceptron (MLP), instance-based learner (IBK), adaptive boosting (AdaBoost), and Random Forest. The experimental results showed better performance for the proposed technique as compared to other state-of-the-art methods. The classification accuracies obtained for the seven emotion classes were 91.8%, 83.7%, and 80.5% for the EMO-DB, EMOVO-DB, and SAVEE-DB, respectively.**
Keywords: **Emotion Recognition, Audio Features, Feature Reduction, Feature Selection, Classifier Ensemble**

I. INTRODUCTION

Speech is an important modality of communication among humans and between humans and machines. Different human-computer interaction (HCI) applications are developed to enable humans to interact with machines using speech and other modalities. This interaction will be more effective and natural if machines can recognize human emotions and respond accordingly. The recognition of the human's emotional state has several applications including education, gaming, security, healthcare, and car driver's safety. Zhu and Luo [1] recognized emotions in e-learning using speech to cope with the deficiencies in e-learning systems. Intelligent pet machines with emotion recognition capability have been introduced in the market to engage home-alone elders, kids, and disabled people [2]. They communicate with people in a natural way to overcome their sorrows and record their emotions. Human speech contains both linguistic and paralinguistic contents. The paralinguistic content includes prosody and spectral features. Researchers have used prosody and spectral features for speaker recognition, speaker verification, and recognition of emotions [3]-[5]. The classification task in machine learning is normally performed using a single classifier, hierarchical classifier, or classifier ensemble approach. Araño et al. [6] utilized a hybrid set of features for classifying emotions from speech consisting of MFCCs and image features extracted from spectrograms. The[1] MFCCs features along with the long short-term memory (LSTM) network performed better as compared to the SVM classifier. Mannepalli

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

et al. [7] introduced multiples support vector neural network classifier for speech emotion recognition. The proposed model performed better as compared to the adaptive fractional deep belief network (AFDBN), fractional deep belief network (FDBN), and deep belief network (DBN). Alluhaidan et al. [8] combined the MFCCs and time-domain features (MFCCT) to achieve better classification accuracy. The convolutional neural network was used for classification, which performed better as compared to other machine learning classifiers. ER [9] combined the acoustic and deep features for speech emotion recognition. The SVM classifier was used for the classification. The proposed technique achieved classification accuracies of 79.4%, 90.2%, and 85.4% for the RAVDESS, EMO-DB, and IEMOCAP datasets, respectively. Liu et al. [10] proposed a brain-emotional learning model for speech emotion recognition. The weights of the model were updated using a genetic algorithm. The MFCC related features were used for classification. The proposed method obtained average classification accuracies of 90.3% on CASIA, 76.4% on SAVEE, and 71.1% on FAU Aibo datasets for the speaker-dependent scenario. To make a final decision about the classification of an instance, the ensemble technique aggregates the outputs of base classifiers. The combining classifiers technique has been observed to perform better as compared to a single classifier model. This technique has been used in different areas of machine learning including biometrics [11], streaming data [12], concept drift, and incremental learning [13]. Three reasons are suggested by Dietterich [14] for using an ensemble classifier that might provide improved performance as compared to a single classifier: statistical, computational, and representational. Numerous theoretical analyses [15]-17], experimental comparisons [18,19], and reviews [20,21] propose how to combine classifiers to achieve better results. Mohan et al. [22] combined the 2D convolutional neural network (2D-CNN) and extreme grading boosting (XG-Boost) to achieve 96.5% classification accuracy for 16 emotion classes of the RAVDESS dataset. The MFCC features were used for the classification. The propose ensemble model performed better as compared to Random Forest and CNN-LSTM. Bhanusree et al. [23] proposed a model that used a time-distributed attention-layered CNN for feature extraction and a Random Forest for classification. The proposed model achieved classification accuracies of 92.2% and 90.3% on the RAVDESS and IEMOCAP datasets, respectively. Badr et al. [24] proposed a hybrid neural network model using the Convolutional and LSTM (ConvLSTM) networks. An average classification accuracy of 91.0% was obtained on the RAVDESS dataset using MFCC features. Novais et al. [25] used Random Forest, AdaBoost, Neural Network, and their ensemble using majority vote for audio emotion classification. The classification accuracy of 75.6% was achieved on the RAVDESS dataset using Random Forest and 86.4% on a group of datasets consisting of RAVDESS, SAVEE, and TESS using Neural Network. The overall performance of the ensemble method using a majority vote was lower as compared to individual classifiers. Chalapathi et al. [26] proposed ensemble learning using high-dimensional acoustic features for audio emotion recognition. The AdaBoost classifier was used for classification. An average accuracy of 94.8% was obtained for seven emotions of the RAVDESS dataset. Speech emotion recognition is a challenging field of speech processing. It is difficult for an individual classifier to provide higher classification performance. For this reason, in recent years researchers have focused on ensemble classification techniques. In ensemble learning, the best results of individual classifiers are combined in an effective way to achieve higher classification accuracy. In this research, a novel classifier ensemble technique is proposed to classify six basic emotions plus neutral using speech. This technique is called as combined confusion matrix, which combines the best results obtained for various emotions using different base classifiers. The combined confusion matrix approach is based on the fact that some classifiers with overall lower performance have better accuracy for a specific class as compared to a classifier with overall higher accuracy. Thus, by combining the best results obtained for each emotion class from different base classifiers, an overall improved classification performance is achieved. The rest of the paper is organized as follows: emotional speech databases, methodology, results and discussion, and conclusion.

II. EMOTIONAL SPEECH DATABASES

The speech databases used in this research are the EMO-DB [27], EMOVO-DB [28], and SAVEE-DB [29]. To authenticate the generalization of the proposed approach, three databases in different languages were chosen. The EMO-DB is in German language, EMOVO-DB is in Italian language and SAVEE-DB is in English language. All the databases were recorded in six basic emotions plus a neutral state. In addition, these databases have been widely used in emotion recognition research.

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

A. EMO-DB:

The EMO-DB is an acted German emotional speech corpus recorded by 10 actors (5 male and 5 female) of diverse ages. It consists of 10 utterances used in everyday oral communication. It was recorded in six basic emotions (anger, disgust, fear, happy, sad, and surprise) plus neutral in an anechoic chamber with high-fidelity recording equipment. The total number of speech samples in this database is 535. To evaluate and guarantee the quality and naturalness of the corpus, a human perception test was carried out by 20 subjects. The samples with more than 80% recognition rate and more than 60% naturalness rate as judged by the subjects were labeled for voice-quality, phonatory and articulatory settings, and articulatory features. The average human accuracy for the database is 86.1% for seven emotions.

B. EMOVO-D

The EMOVO-DB is an emotional speech corpus recorded in the Italian language. The database was recorded by 6 actors (3 male and 3 female) with professional backgrounds having ages in the range of 23 to 30 years. Each actor uttered 14 sentences per emotion in six basic emotions [5] plus neutral. To evaluate the corpus a subjective test was performed by 12 subjects. The average human classification accuracy is 80% for seven emotion classes.

C. SAVEE-DB

To develop an automatic emotion recognition system, the SAVEE-DB was recorded in the English language. Four male actors of diverse origins and ages participated in the recording. The database was recorded in six basic emotions (anger, disgust, fear, happy, sad, and surprise) plus a neutral state. For recording, phonetically balanced sentences (15 per emotion) were selected from the TIMIT dataset. Sophisticated recording equipment was used to record the data in a visual media lab. Ten subjects evaluated the quality of the data. For audio data, the average human accuracy is 66.5% for seven emotions.

III. METHODOLOGY

The process of emotion classification using speech consists of feature extraction, feature normalization, feature selection, and classification, as shown in Figure 1.

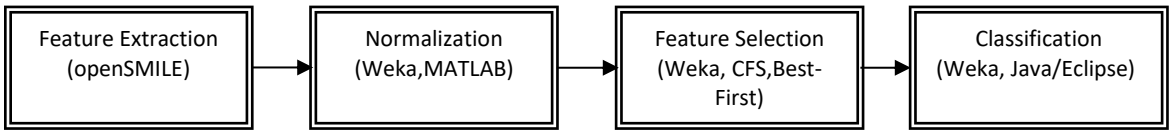

Figure 1: Block diagram of emotion classification methodology

A. Feature Extraction:

Audio features are extracted using different tools including PRAAT [30], Speech Filling System (SFS) [31], Hidden Markov toolkit (HTK) [32], and openSMILE toolkit [33]. In this research, the openSMILE toolkit was used for the extraction of audio features. A total of 73 low-level descriptors (LLDs) were extracted. These LLDs included MFCCs, MFBs, LSPs, ZCR, intensity, energy, formants, pitch, jitter, and spectral features. Delta and delta-delta coefficients were computed for these LLDs. Thirty-nine functions were applied to the LLDs and their delta and delta-delta coefficients. These functions included extremes, regression, moments, percentile crossings, peaks, and means. The total number of extracted features for each instance was 8543. The details of extracted features are given in Table 1.

B. Feature Normalization:

Normalization is the process of scaling feature space to a common range. The extracted features need to be normalized before the application of feature selection and classification algorithms. Different techniques have been used for data normalization. The common feature normalization techniques are z-score [34] and min-max [35]. In this research, both techniques were analyzed, and the min-max technique was observed to be better than the z-score method.

Z-Score Normalization: Z-score normalizes the data to zero mean and unit variance. It is defined by the following relation

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Where μ is the mean and σ is the standard deviation of feature x.

Min-Max Normalization: This technique normalizes the data in a range $[v_{min}\ \ v_{max}]$ using the following relation

$$\bar{x} = \left[\frac{x - x_{min}}{x_{max} - x_{min}} \times (v_{max} - v_{min})\right] + v_{min} \quad (2)$$

All features were normalized to the range $[-1\ \ 1]$. For this range, the Equation 2 becomes

$$\bar{x} = \left[\frac{x - x_{min}}{x_{max} - x_{min}} \times 2\right] - 1 \quad (3)$$

Table 1: Audio features extracted using the openSMILE toolkit.

Feature Group	No. of Features	Delta Coefficients	Delta-Delta Coefficients	Functions	Total
MFCCs (0-12)	13	13	13	39	1521
MFBs (0-25)	26	26	26	39	3042
LSPs (0-7)	8	8	8	39	936
ZCR	1	1	1	39	117
Pitch	3	3	3	39	353
Intensity	2	2	2	39	234
Energy	1	1	1	39	117
Formants	4	4	4	39	468
Pitch Jitter	3	3	3	39	351
Spectral	12	12	12	39	1404
	73	73	73	39, 2 extra for the F_0 final	8543
(73 + 73 + 73) × 39 + 2 = 8543					

C. Feature Selection

The CFS with the best first search method was used for feature selection. The best first technique searches the space of attribute subsets by greedy hill climbing. The number of selected features for the EMO-DB was 266, for the SAVEE-DB was 123, and for the EMOVO-DB was 121. The selected features for the three databases are given in Table 2.

D. Classification

To perform emotion classification, five different base classifiers were used including SVM [36], MLP [37], IBK [38], AdaBoost [39], and Random Forest [40]. These classifiers were chosen based on the different approaches they used

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

for classification. SVM transforms the data from low to high dimensional space where it can be easily separated. It is faster and performs better with less training data and high dimensional space. MLP is a feed-forward artificial neural network that consists of multiple fully connected layers. It requires a large number of training data to provide better classification results. IBK is the non-parametric k-nearest neighbor classifier, which utilizes vicinity to estimate the group of a data point. AdaBoost builds a robust classifier by uniting numerous weak classifiers to obtain higher accuracy. Any classifier with adjustable weights can be used as a base classifier. A Random Forest fits numerous decision trees on several sub-samples of the dataset. The classification accuracy is improved using averaging.

Table 2: Selected features for the EMO-DB, EMOVO-DB, and SAVEE-DB using CFS with the best first search method.

Feature Group	**Number of Selected Features**		
	EMO-DB	**EMOVO-DB**	**SAVEE-DB**
MFCC	47	35	18
MFB	95	30	35
LSP Freq	18	7	15
Spectral	33	17	0
F Bands	19	0	17
Intensity	3	3	0
Energy	1	0	4
Pitch	24	13	27
Jitter	3	1	1
Shimmer	6	6	0
Voicing	4	2	2
Formant	6	7	5
ZCR	7	0	0
Total	**266**	**121**	**123**

The experiments were performed for seven emotion classes with 10-fold cross-validation. Table 3 shows the class-wise and overall classification performance of base classifiers for the EMO-DB. The SVM resulted in the highest accuracy of 90.8%, while in terms of individual classes it provided better results for happy, neutral, and sad emotions. The overall accuracy of IBK was 84.7% which was the lowest, however, it resulted in much better classification accuracy for the disgust emotion as compared to other classifiers. The MLP classifier provided better results for the fear, sad, and surprise classes, while achieving an overall accuracy of 90.1%. Random Forest achieved better performance for the anger and sad classes, although its overall performance was lower than the SVM and MLP classifiers. The AdaBoost classifier did not performed the best for any individual emotion, although its overall performance was better than IBK. For the sad emotion, three classifiers provided the equal best result, i.e., Random Forest, MLP, and SVM, but we chose SVM because of its overall best performance.

The analysis of these results shows that the overall best performance does not guarantee the best results for all individual emotion classes. A classifier with an overall lower accuracy can provide better results for individual classes. These results guided us to introduce a novel approach, the combined confusion matrix.

PJETS
E-ISSN 2224-2333
P-ISSN 2222-9930

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020

Table 2: Overall and class-based classification accuracies (%) of base classifiers on the EMO-DB.

Classifier	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall Accuracy
AdaBoost	94.4	75.6	83.5	76.3	90.5	94.2	94.5	87.0
Random Forest	99.0	80.4	83.2	63.1	90.3	98.0	91.0	86.4
IBK	94.2	87.0	86.3	48.0	86.2	97.0	94.3	84.7
MLP	93.5	84.4	88.0	77.7	94.4	98.0	95.0	90.1
SVM	95.6	84.6	87.0	80.5	96.0	98.0	94.0	90.8

Combined Confusion Matrix: The results obtained for the base classifiers in Table 3 demonstrate that some classifiers provided better classification results for one or more individual classes irrespective of their overall lower performance. Based on these facts, we combined the best classification results for each emotion class irrespective of base classifiers to obtain a combined confusion matrix. The combined confusion matrices for the EMO-DB, EMOVO-DB, and SAVEE-DB are given in Tables 4, 5, and 6, respectively. The results of the base classifiers were combined at the decision level. The classification performance of base classifiers on training data was used to find out which classifier provides overall better results for which specific emotion. Once these classifiers were identified, in the next step their results were combined based on recognized emotions. A classifier that performs better for a specific emotion was given priority over others that provide lower performance for that specific emotion. In this way, the emotion-specific best classifier decisions were combined to achieve an overall better performance. This approach has resulted in improved classification results because the best emotion-specific results were combined.

Table 3: Combined confusion matrix for the EMO-DB.

	Recognized Emotion						
Acted Emotion	**anger**	**disgust**	**fear**	**happy**	**neutral**	**sad**	**boredom**
anger	**126**	**0**	**0**	**1**	**0**	**0**	**0**
disgust	**1**	**40**	**1**	**0**	**2**	**1**	**1**
fear	**2**	**2**	**60**	**3**	**2**	**0**	**0**
happy	**8**	**2**	**3**	**56**	**2**	**0**	**0**
neutral	**0**	**0**	**1**	**1**	**76**	**1**	**0**
sad	**0**	**0**	**0**	**0**	**61**	**1**	**0**
boredom	**0**	**0**	**1**	**0**	**1**	**2**	**77**
Overall classification accuracy = 91.8%							

Table 4: Combined confusion matrix for the EMOVO-DB.

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

	Recognized Emotion						
Acted Emotion	**anger**	**disgust**	**fear**	**happy**	**neutral**	**sad**	**surprise**
anger	72	2	0	8	2	0	0
disgust	2	72	2	1	3	3	1
fear	3	6	64	2	0	2	7
happy	5	5	1	59	1	0	13
neutral	0	2	1	1	80	0	0
sad	0	1	3	0	2	78	0
surprise	0	1	5	10	0	1	67
Overall classification accuracy = 83.7%							

IV. RESULTS AND DISCUSSION

The classification results of individual base classifiers for the EMO-DB, EMOVO-DB, and SAVEE-DB are given in Table 7. The best classification accuracies of 90.8%, 82.7%, and 77.1% were obtained for the EMO-DB, EMOVO-DB, and SAVEE-DB, respectively. The classification accuracies for classifier ensemble methods using different combination rules such as sum, maximum, minimum, majority vote, product, and combined confusion matrix for the three datasets are given in Table 8. The best results obtained for the individual base classifiers, classifier combination rules, and combined confusion matrix are summarized in Table 9. In general, it was observed that the different combination rules, i.e., sum, max, min, majority vote, and product, did not provide any significant improvement in the classification accuracy in comparison to the best results obtained with the individual base classifiers. The sum rule performed better as compared to other combination rules. The results demonstrate that the combined confusion matrix approach outperformed both the base classifiers and classifier combination rules. The comparison of the proposed combined confusion matrix approach with the existing state-of-the-art techniques and human is given in Table 10. Schuller et al. [41] achieved an average accuracy of 87.5% for seven emotions on EMO-DB using the best 75 features using the SVM classifier. Tawari and Trivedi [42] reported 83.7% classification accuracy for seven emotions of the EMO-DB using the SVM Classifier. Hassan and Damper [43] proposed a 3DEC hierarchical model to achieve 89.0% accuracy for seven emotion classes on EMO-DB. Yüncü et al. [44] achieved an average accuracy of 82.9% for seven emotions on the EMO-DB, while 73.8% for six emotions on the SAVEE-DB. Mao et al. [45] reported 85.2% and 73.6% accuracies for the EMO-DB and SAVEE-DB, respectively, using a convolutional neural network. Er [9] used the acoustic and deep features with the SVM classifier and achieved classification accuracies of 90.2% on EMO-DB. Liu et al. [10] proposed a brain-emotional learning model with MFCC features for emotion classification. The proposed method obtained an average classification accuracy of 76.4% on the SAVEE database for the speaker-dependent scenario. The human classification accuracies for EMO-DB, SAVEE-DB, and EMOVO-DB are 86.1%, 66.5%, and 80.0%, respectively. The comparisons of these results indicate that the proposed combined confusion matrix technique outperformed the state-of-the-art methods by exploiting the best class-wise performance of base classifiers.

E-ISSN 2224-2333
P-ISSN 2222-9930

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020

Table 5: Combined confusion matrix for the SAVEE-DB.

	Recognized Emotion						
Acted Emotion	**anger**	**disgust**	**fear**	**happy**	**neutral**	**sad**	**surprise**
anger	49	1	2	7	0	0	4
disgust	2	41	2	1	11	2	1
fear	2	2	39	5	1	2	9
happy	7	2	1	46	0	0	4
neutral	0	1	0	0	118	1	0
sad	0	4	0	0	8	48	0
surprise	1	0	7	4	0	0	48
Overall classification accuracy = 80.5%							

Table 7: The classification accuracy (%) of base classifiers for the EMO-DB, EMOVO-DB, and SAVEE-DB.

Base Classifier	**Database**		
	EMO-DB	**EMOVO-DB**	**SAVEE-DB**
AdaBoost	87.0	82.7	76.7
Random Forest	86.4	78.9	71.7
IBK	84.7	71.7	63.3
MLP	90.1	78.0	77.1
SVM	90.8	78.7	72.9

Table 8: The comparison between classification accuracies (%) of different classifier ensemble methods.

Database	**Classifier Ensemble Method**					
	Sum	**Max**	**Min**	**Majority Vote**	**Product**	**Combined Confusion Matrix**
EMO-DB	91.0	88.2	89.1	82.3	84.4	91.8
EMOVO-DB	81.3	81.0	80.5	80.0	79.3	83.7
SAVEE-DB	77.4	75.1	76.2	76.1	72.6	80.5

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020
E-ISSN 2224-2333
P-ISSN 2222-9930

Table 9: The best classification accuracy (%) obtained for base classifiers and classifier ensemble methods.

Database	Base classifier	Classifier Ensemble Method	
		Sum	Combined confusion matrix
EMO-DB	90.8	91.0	91.8
EMOVO-DB	82.7	81.3	83.7
SAVEE-DB	77.1	77.4	80.5

Table 10: The comparison between the classification accuracy (%) of state-of-the-art techniques, human, and the proposed combined confusion matrix approach.

Method	Database		
	EMO-DB	SAVEE-DB	EMOVO-DB
Schuller et al. [41]	87.5		
Tawari and Trivedi [42]	83.7		
Hassan and Damper [43]	89.0		
Yüncü et al. [44]	82.9	73.8	
Mao et al. [45]	85.2	73.6	
Er [9]	90.2		
Liu et al. [10]		76.4	
Human	86.1	66.5	80.0
Combined Confusion Matrix	91.8	80.5	83.7

V. CONCLUSION

The research domain of emotion recognition has attracted many researchers to improve human-computer interaction. In this research, emotional speech databases in three different languages were used for the speaker-independent emotion classification, i.e., EMO-DB, EMOVO-DB, and SAVEE-DB. In the first step, audio features were extracted using the openSMILE toolkit. The feature sets were then normalized to the [1, 1] range using the min-max technique. It was followed by feature selection using the Correlation-based feature selection with the best first search method. In the last step, classification was performed using individual base classifiers as well as different classifier ensemble techniques. Different classifier combination rules including sum, max, min, majority vote, and product were investigated. It was observed that the classifier combination rules did not provide any significant improvement in the classification accuracy in comparison to individual base classifiers. The proposed combined confusion matrix approach resulted in better performance as compared to individual base classifiers and classifier combination rules. The proposed technique resulted in classification accuracies of 91.8%, 83.7%, and 80.5% for seven emotions of the EMO-DB, EMOVO-DB, and SAVEE-DB, respectively.

PJETS
E-ISSN 2224-2333
P-ISSN 2222-9930

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020

REFERENCES

[1] Aiqin Zhu and Qi Luo, "Study on speech emotion recognition system in E-learning," in *International Conference on Human-Computer Interaction*, Beijing, 2007, pp. 544-552.
[2] Yongming Huang, Guobao Zhang, and Xiaoli Xu, "Speech emotion recognition research based on wavelet neural network for robot pet," in *International Conference on Intelligent Computing*, Ulsan, 2009, pp. 993--1000.
[3] Douglas A Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, vol. 741, pp. 659-663, 2009.
[4] Jayant M Naik, "Speaker verification: A tutorial," *IEEE communications magazine*, vol. 28, no. 1, pp. 42-48, 1990.
[5] Roddy Cowie and Randolph R Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1-2, pp. 5-32, 2003.
[6] Keith April Araño, Peter Gloor, Carlotta Orsenigo, and Carlo Vercellis, "When old meets new: emotion recognition from speech signals," *Cognitive Computation*, vol. 13, pp. 771-783, 2021.
[7] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman, "Emotion recognition in speech signals using optimization based multi-SVNN classifier," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 2, pp. 384-397, 2022.
[8] Ala Saleh Alluhaidan, Oumaima Saidani, Rashid Jahangir, Muhammad Asif Nauman, and Omnia Saidani Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Applied Sciences*, vol. 13, no. 8, p. 4750, 2023.
[9] Mehmet Bilal Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," *IEEE Access*, vol. 8, pp. 221640-221653, 2020.
[10] Zhen-Tao Liu et al., "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145-156, 2018.
[11] Salil Prabhakar and Anil K Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognition*, vol. 35, no. 4, pp. 861-874, 2002.
[12] Valerio Grossi and Franco Turini, "Stream mining: a novel architecture for ensemble-based classification," *Knowledge and information systems*, vol. 30, pp. 247-281, 2012.
[13] Ryan Elwell and Robi Polikar, "Incremental learning of concept drift in nonstationary environments," *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1517-1531, 2011.
[14] Thomas G Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1-15.
[15] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226-239, 1998.
[16] Giorgio Fumera and Fabio Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 6, pp. 942-956, 2005.
[17] Robi Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21-45, 2006.
[18] Kaushik Ghosh, Yew Seng Ng, and Rajagopalan Srinivasan, "Evaluation of decision fusion strategies for effective collaboration among heterogeneous fault diagnostic methods," *Computers & chemical engineering*, vol. 35, no. 2, pp. 342-355, 2011.
[19] Li Zhang and Wei-Da Zhou, "Sparse ensembles using weighted combination methods based on linear programming," *Pattern Recognition*, vol. 44, no. 1, pp. 97-106, 2011.
[20] Sergey Tulyakov, Stefan Jaeger, Venu Govindaraju, and David Doermann, "Review of classifier combination methods," *Machine learning in document analysis and recognition*, pp. 361-386, 2008.
[21] Zhi-Hua Zhou, *Ensemble methods: foundations and algorithms*.: CRC Press, 2012.
[22] Meera Mohan, P. Dhanalakshmi, and R. Satheesh Kumar , "Speech Emotion Classification Using Ensemble Models with MFCC," *Procedia Computer Science*, vol. 218, pp. 1857-1868, 2023.
[23] Yalamanchili Bhanusree, Samayamantula Srinivas Kumar, and Anne Koteswara Rao, "Time-Distributed Attention-Layered Convolution Neural Network with Ensemble Learning using Random Forest Classifier for Speech Emotion Recognition," *Journal of Information and Communication Technology*, vol. 22, no. 1, pp. 49-76, 2023.
[24] Youakim Badr, Partha Mukherjee, and Sindhu Thumati, "Speech Emotion Recognition using MFCC and Hybrid Neural Networks," in *International Joint Conference on Computational Intelligence*, 2021, pp. 366-373.
[25] Rui M.B. Novais, Pedro J.S. Cardoso, and João M.F. Rodrigues, "Emotion Classification from Speech by an Ensemble Strategy," in *International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, Lisbon, 2022, pp. 85-90.
[26] M. M. Venkata Chalapathi, M. Rudra Kumar, Neeraj Sharma, and S. Shitharth, "Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal," *Security and Communication Networks*, vol. 2022, pp. 1-10, 2022.
[27] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of German emotional speech," in *Interspeech*, Lisbon, 2005, pp. 1517-1520.
[28] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco, "EMOVO corpus: an Italian emotional speech database," in *International conference on language resources and evaluation*, 2014, pp. 3501-3504.

PJETS
E-ISSN 2224-2333
P-ISSN 2222-9930

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 02, PP. 01-11, December, 2023
https://doi.org/10.22555/pjets.v11i2.1020

[29] Sanaul Haq and Philip JB Jackson, "Multimodal emotion recognition," in *Machine audition: principles, algorithms and systems*.: IGI Global, 2011, pp. 398-423.
[30] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341-345, 2001.
[31] M Huckvale. (2000) Speech filing system: Tools for speech research.
[32] Steve J Young and Sj Young, "The HTK hidden Markov model toolkit: Design and philosophy," University of Cambridge, Department of Engineering Cambridge, 1993.

[33] Florian Eyben, Martin Wöllmer, and Björn Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *International conference on affective computing and intelligent interaction and workshops*, Amsterdam, 2009, pp. 1-6.
[34] Anil Jain, Karthik Nandakumar, and Arun Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270-2285, 2005.
[35] Y Kumar Jain and Santosh Kumar Bhandare, "Min max normalization based data perturbation method for privacy protection," *International Journal of Computer & Communication Technology*, vol. 2, no. 8, pp. 45-50, 2011.
[36] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machine," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1-27, 2011.
[37] Matt W Gardner and SR Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627-2636, 1998.
[38] David W Aha, Dennis Kibler, and Marc K Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, pp. 37-66, 1991.
[39] Y Schapire Freund and E Schapire, "RE Experiments with a New Boosting Algorithm," in *International Conference on Machine Learning*, 1996, pp. 148-156.
[40] Leo Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.
[41] Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Interspeech*, Lisbon, 2005, pp. 805-808.
[42] Ashish Tawari and Mohan Manubhai Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502-509, 2010.
[43] Ali Hassan and Robert I Damper, "Classification of emotional speech using 3DEC hierarchical classifier," *Speech Communication*, vol. 54, no. 7, pp. 903-916, 2012.
[44] Enes Yüncü, Hüseyin Hacihabiboglu, and Cem Bozsahin, "Automatic speech emotion recognition using auditory models with binary decision tree and svm," in *International conference on pattern recognition*, 2014, pp. 773-778.
[45] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014.