Speech Emotion Recognition Using Deep Learning Models

Janani Ravikumar
Department Of Applied Science
San Jose State University
San Jose, USA
janani.ravikumar@sjsu.edu

Nikitha Bramadi
Department Of Applied Science
San Jose State University
San Jose, USA
nikitha.bramadi@sjsu.edu

Revathi Boopathi
Department Of Applied Science
San Jose State University
San Jose, USA
revathi.boopathi@sjsu.edu

Neetu Rasinger Babu
Department Of Applied Science
San Jose State University
San Jose, USA
neeturasinger.babu@sjsu.edu

***Abstract*–Emotion recognition is becoming integral to interaction technology with increasing human and computer interactions. Improving Speech Emotion Recognition (SER) from human speech can lead to reliable outcomes in human-computer interaction. Opportunistic prediction of mental health disorders has become essential due to the rise in mental health problems during the COVID-19 pandemic. A reliable SER establishes a platform to act toward helping humanity with mental health problems in real-time modes of interaction. While deep learning strategies have shown promise in the execution of SER compared to machine learning techniques, an extended deeper dive into using newer features is mostly limited to audio-related features. To address this limitation, an innovative approach is implemented using a well-established deep feature extraction technique, presenting the raw data through statistical features and making the recognition customized to gender through a gender feature. The proposed architecture with the new approach is SER-FusionNet, representing the fusion of announced feature perceptions. SER-FusionNet consists of two crucial deep learning modules: the 1-D Convolutional Neural Network used in deep feature extraction and the Multi-Layer Perceptron as a classification module. A test accuracy of 88% is achieved with SER-FusionNet on the augmented Ravdess dataset. A robust and stiff evaluation avoids any scope of overfitting and underfitting, for which custom cross-validation is implemented with each fold including unique actor representations and each fold having an equal number of actors. With the Ravdess dataset, a cross-validation accuracy of 0.45±0.05, is achieved. To increase the diversity in the dataset in terms of sentences spoken, voices involved, and the number of audios recorded, a merged dataset called SER Superset is created. The creation involves four datasets: Ravdess , CREMA-D, SAVEE, and TESS. The SER Superset generates a cross-validation accuracy of 0.43 ± 0.02, proving consistency over diverse data.**

Keywords- **Speech Emotion Recognition (SER), SER-FusionNet, Conv1D, MLP**

I. Introduction

Speech Emotion Recognition involves identifying and categorizing people's emotional states based on their voice signals utilizing computer approaches and machine learning algorithms. Pitch, intensity, tempo, spectral features, and phonetic content are auditory cues in human speech that can indicate the speaker's emotional state. By analyzing these acoustic aspects, SER systems attempt to recognize and classify emotions expressed in speech, such as happy, sad, neutral, fear, surprise, disgust, and more. Over traditional machine learning and rule-based approaches, deep learning's capacity to automatically create hierarchical representations from audio data, efficient feature learning, temporal dependency modeling, and high-dimensional data processing significantly made advances in speech-emotion recognition.

Existing research in SER encounters stagnation in recognizing an emotion, always only taking an evident traditional approach of extracting prominent acoustic features like mfcc, melspectrogram, zero crossing rate, or root mean square to train the model. Stepping back and looking into wider opportunities to provide new features to the model training to gain meaningful results is the aim of this project. The challenge is addressed using three perspectives in the case of SER model features. One is gender which inherently considers different vocal tones. Knowing which gender's speech is being evaluated could substantially impact the model's prediction outcome. A second perspective is to retain the raw data [8][9][10][11]. When audio is converted to an array of numbers, these numbers indicate the audio amplitude at every point sampling has been done. But the raw data has the drawback of having high dimensionality [8][9][10][11]. To reduce the dimension of the array generated and include the basic information, summarizing this data is the solution taken. An appropriate way to summarize the data is by extracting the various statistical features. A third reliable perspective being added here is to apply deep feature extraction. Deep learning methods are dependable for extracting high-level information from the input provided. Among multiple traditional acoustic filters applied whose result acts as features, instead, we want to reduce these features to have only the most important ones. A most suitable deep model would enable this through training and generating the weights acting as features. By combining these three perspectives, a final feature set is developed. Existing research in deep learning approaches to SER have used merged features. However we additionally experiment to check if the considered features make a significant impact to increase the model robustness, and increase the generalizability.

Apart from the unique feature selection, the proposed approach introduces a new cross-validation method as contrast to existing research where folds are created based on the actors. Creating SER Superset presents a unique experimental setup with wider train and test datasets variation. Towards modeling, having a two-step approach with a deep feature extraction model and a classification model makes it a unique procedure from existing methods in the literature.

The research question is to analyze if the fusion of three types of features provides a competitive speech emotion recognition approach. To answer this, two suitable models for deep feature extraction and classification over experiments with potential well-versed deep learning models. Additionally, a real-time testing framework to showcase the performance of the proposed approach is implemented.

Making machines appear and behave more human-likely has proven to require the addition of emotions. This research contributes to developing effective, real-time techniques for identifying emotions in various human-machine interface users, including call center agents, customers, pilots, and drivers. This is especially in those with speech impediments in several industries, including education, criminal investigation, customer service chatbots, psychiatry, and healthcare.

II. Related Work

Emotion reputation is explored in diverse modalities, including facial expressions and textual content evaluation. However, tackling the challenge of figuring out emotions from audio files gives enormous demanding situations [1]. Emotions, including happiness, sadness, anger, surprise, disgust, and fear, were considered among other emotions. A critical factor in the SERs ability to recognize emotions effectively in speech is the integration of elements like MFCC, pitch, zero-crossing rate, and energy. These operations provide priceless data that classifiers may utilize to discern distinct emotional states [2] accurately. The number of zero-crossings in a signal segment measures how smooth that specific signal is [3]. Some of the existing works leveraged the statistical representation of audio features into account. Features such as average amplitude, voice volume, duration of speech, etc., were among them. [4]. Various machine learning classification algorithms aid in classifying speech emotions. They involve statistical or domain-based techniques to extract the most relevant features [5]. Since pitch is a physiological trait, it can be used to identify a speaker's gender. Deep learning based approaches to solving the problem of recognizing emotions remain state-of-the-art. CNN has been observed to outperform BiLSTM in speech emotion recognition. This can be attributed to the fact that CNN emphasizes local features and strongly correlates with critical aspects of speech, such as pitch, maximum, and minimum values of speech features [6]. Cross-validation techniques have consistently proved the robustness of machine learning and deep learning models. The number of folds in the cross-validation depends on the sample size [7].

III. Methodology

A. Datasets

Four different datasets are considered to fulfill the data requirement, the primary dataset being Ravdess [8]. The dataset consists of 12 male and 12 female actors, with 1,440 audio recordings equally divided among both genders. Two sentences are used to generate all the audio recordings. The dataset covers eight target emotional labels: calm, neutral, happy, surprised, sad, disgusted, fear, and angry. A sample visual of an emotion can be seen in Fig. 1 [8].

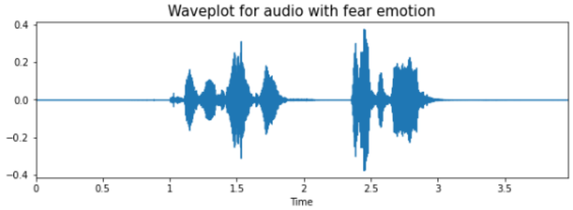


Figure 1: Sample visual plot of fear emotion

More sentences and audio recordings are brought together to increase the generalizability of the model being developed. Apart from Ravdess, Crema-d, Savee, and Tess are combined to create an SER Superset [9] [10] [11]. The superset created consists of a 12,162 file count, 121 actors with 57 female and 64 male actors. In this dataset, 229 sentences generate all the audio files. The disadvantage of combining different datasets gives rise to imbalanced audio recording for each emotion. Based on Fig. 2, surprise and calm have low file counts and will be eliminated to model on the SER Superset data.

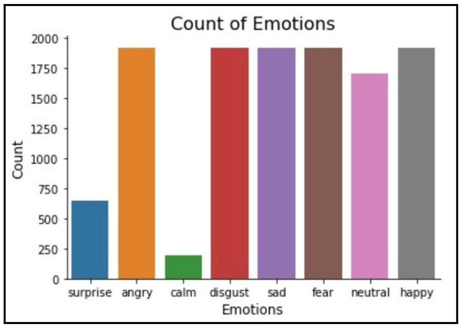


Figure 2: Class Imbalance present in SER Superset

Each dataset specified is collected from Kaggle, where each dataset is published separately [12] [13] [14] [15].

B. Preprocessing

Data is prepared by extracting information like gender indicated from the filename or a mapping sheet provided. While reading the audio file, a sampling rate of 22,050 Hz is used to avoid the curse of dimensionality. Another notable alteration to the audio signals is sniping out the starting and ending empty audio signals based on the general observation.

Feature Extraction: Three types of features are estimated in the procedure and will be utilized over different stages of the architecture proposed. The feature extraction methods are as follows:

a) *Acoustic Features:* Acoustic features are critical descriptors of audio material. They extract helpful information from audio signals and make various audio-related jobs easier. Some essential functions of the Librosa audio library are zero crossing rate, chroma stft, mel frequency cepstral coefficient, root mean square, and mel spectrogram to generate features later fed to the deep feature extraction model.

b) *Statistical Features:* It acts as a numerical measure to represent raw data in this case. Six different features are extracted using mean, variance, kurtosis, skewness, mean of spectral centroids, and mean of spectral bandwidth.

c) *Gender Feature:* A gender feature based on whether the actor is male or female is selected and indicated by 0 and 1, respectively.

C. Data Augmentation

Numerous changes and alterations are made to the current data samples for data augmentation. To increase the generalization and robustness of the model, data augmentation aims to produce fresh training samples that are comparable to the original data but show minor changes. The techniques used are pitch shift, noise induction, stretch, and loss-pass filter. Utilizing the method, the dataset is increased to double the original audio count.

D. Preparation

The data preparation is done one step ahead of the model training, validation, and testing. The data is split into 80:10:10 for train, test and validation respectively. While splitting, stratified process is ensured for no leakage between datasets.The evaluation metrics from each training stage, validation, and test are compared to evaluate for overfitting or underfitting.

As a part of preparation special attention is required to be put into the procedure to split the data during cross-validation on the Ravdess dataset involves dividing the dataset into six-folds, where each fold solely consists of the information from a set of four actors. A similar cross-validation preparation is conducted on the SER Superset data with 11 folds, each with 11 actors.

E. Proposed Approach

SER FusionNet encompasses a phased approach. Firstly, it captures the acoustic features from audio and passes it through a feature extractor unit consisting of two stacked Conv1D layers. This network is known to extract the local time series pattern efficiently. This finally reduces the high dimensional acoustic features of the audio to 32 features.

Furthermore, the extracted features are combined with gender and statistical features of the audio to get 41 fused features. Finally, the features are fed to a classifier unit which consists of a Multi-Layer Perceptron (MLP) that distinguishes various emotions. MLP shows improved performance as compared to experimented LSTM classifiers. A detailed SER FusionNet architecture is shown in Fig. 3.

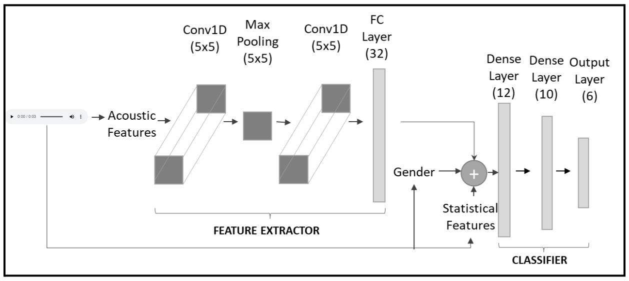


Figure 3: SER-FusionNet Architecture

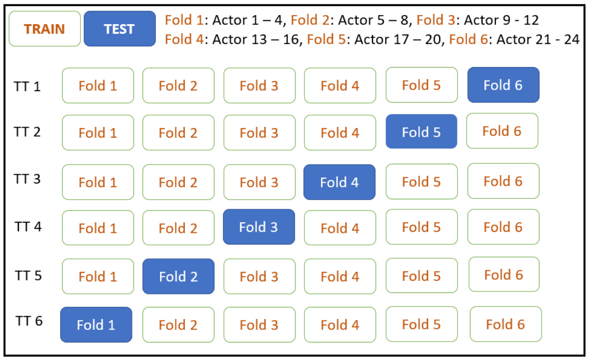
F. *Evaluation Metric*

As a primary metric of evaluation, accuracy is used. Apart from this, recall, precision, and f1-score are also measures in consideration. A confusion matrix is plotted to see the values. Fig. 4 illustrates the confusion matrix [17]. The loss values are evaluated during training to establish an early-stop approach.

		Gold standard outcome		
		Positive	Negative	total
Test outcome	Positive	True positive (TP)	False positive (FP)	Bias (P')
	negative	False negative (FN)	True negative (TN)	Negative bias (N')
total		Prevalence (P)	Negative prevalence (N)	Sample size

Figure 4: Confusion matrix

Cross-validation as an evaluation procedure offers a method for analyzing and contrasting models, choosing the best model, fine-tuning hyperparameters, and determining the model's generalizability. It ensures the model works well with anonymous data and prevents frequent problems like overfitting. Six iterations of the model train and test are performed for the Ravdess dataset and 11 iterations for SER Superset dataset. In the end, an average of the accuracy is considered the final evaluation outcome. Fig. 5 showcases how Ravdess data is trained in iterations.



TRAIN **TEST** Fold 1: Actor 1 – 4, Fold 2: Actor 5 – 8, Fold 3: Actor 9 - 12
Fold 4: Actor 13 – 16, Fold 5: Actor 17 – 20, Fold 6: Actor 21 - 24

TT 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	**Fold 6**
TT 2	Fold 1	Fold 2	Fold 3	Fold 4	**Fold 5**	Fold 6
TT 3	Fold 1	Fold 2	Fold 3	**Fold 4**	Fold 5	Fold 6
TT 4	Fold 1	Fold 2	**Fold 3**	Fold 4	Fold 5	Fold 6
TT 5	Fold 1	**Fold 2**	Fold 3	Fold 4	Fold 5	Fold 6
TT 6	**Fold 1**	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6

Figure 5: Cross-validation splits for Ravdess data

IV. Experimental Setup

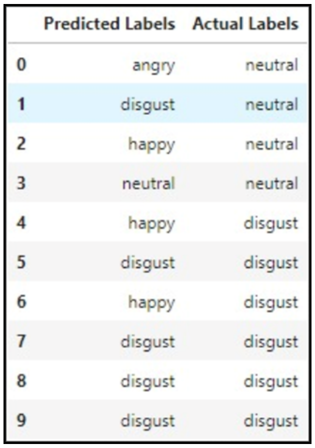
Random seed was set during the experiments to ensure reproducibility. The hyperparameters are tuned by carefully comparing the evaluation results based on cross-validation and choosing the ones that showed better evaluation metrics. The feature extractor and the classifier model were run for 25 epochs to show a stable performance on the validation set. Reduced Plateau learning rate is set with a minimum learning rate of 10^{-7}. ReLu activation is used throughout the neural network architecture because of its simplicity. Adam optimizer is leveraged because of its greater convergence speed. A batch size of 64 is chosen for efficient GPU utilization.

Similar model and hyperparameter set-up is established to execute four experiments. SER FusionNet is implemented with Ravdess with and without cross-validation. At the same time SER Superset as well is implemented with and without cross-validation.

V. Results and Analysis

A. *Results*

Finally, the predicted emotions are compared with actual ground truth to perform visual inspection of the model's effectiveness. Figure 6 shows the predicted labels and actual labels.



	Predicted Labels	Actual Labels
0	angry	neutral
1	disgust	neutral
2	happy	neutral
3	neutral	neutral
4	happy	disgust
5	disgust	disgust
6	happy	disgust
7	disgust	disgust
8	disgust	disgust
9	disgust	disgust

Figure 6: Actual and predicted labels

In the given experimental setup, multiple observations were made with respect to the SER-FusionNet model. Table 1 represents the comparison of validation of test accuracies of Ravdess and SER Superset data. Observations on Ravdess and SER Superset yield test accuracy of 88.07% and 68.04%, respectively.

Table 1: Experimental Results with validation and test accuracies

Data Set	Train Data	Test Data	Validation Accuracy	Test Accuracy
RAVDESS	5-24 Actors	1-4 Actors	72.63 %	88.07 %
SER Superset	5-121 Actors	1-4 Actors	61.12 %	68.04 %

For the data distributed into folds for the cross-validation procedure, we divide the train and validation data, as shown in Table 2. The maximum validation accuracy on Ravdess is greater than that of SER Superset.

Table 2: Experimental Results with a cross-validation accuracy

Data Set	Train Data	Validation Data	Validation Accuracy
RAVDESS	5 * (4 Actor fold)	1 * (4 Actor fold)	45 ± 5 %
SER Superset	10 * (11 Actor fold)	1 * (11 Actor fold)	43 ± 2 %

Table 3 compares the proposed Fusion features extraction and deep features for the Ravdess and SER Superset data. The results of the fusion feature have greater test accuracies establishing that fusion features perform better than just deep features here extracted using Conv1D.

Table 3: Experimental Results to compare with fusion features and just deep features

Dataset	Features	Validation Accuracy	Test Accuracy
RAVDESS	Fusion features	72.63 %	88.07 %
	Deep features	71.69 %	86.65 %
SER Superset	Fusion features	61.12 %	68.04 %
	Deep features	61.20 %	67.19 %

To compare the results from the SER Superset data with SER-FusionNet to existing research conducted using models like Random Forest and a deep learning 1-D CNN model, the SER-FusionNet has demonstrated greater accuracy when compared to the models presented by other research, as shown in the comparison table 4 [16].

Table 4: Comparison of existing literature with SER-FusionNet

Dataset	Model	Accuracy
SER Superset	Radom Forest	43.6 %
	1-D CNN	48.96 %
	SER-FusionNet	68.04 %

B. Discussion

The experiments proved that extracting native audio features using deep learning model doesn't perform well and shows very less accuracy of 13%. This triggered the need for the acoustic feature extraction technique which efficiently transforms the data in frequency domain.

Firstly, the SER-FusionNet model performs the best on the Ravdess dataset. This can be explained by the limited number of sentences and actors who spoke. Figure 7 shows the learning curve for Ravdess.

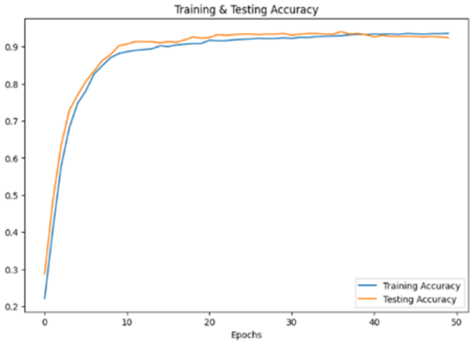


Figure 7: Learning curve for Ravdess

However, generalizability is ensured in the model results by adding CREMA, TESS, and SAVEE datasets with more actors and various sentences.

Based on the observed results in Table 3, the SER-FusionNet approach of having fusion features performs better than just using deep features. Despite not being a significant rise, even a slight improvement in unknown data explains the model's generalizability.

C. *Strengths and Weaknesses*

Robust performance evaluation provides a more reliable estimate of the model's performance by evaluating it on multiple subsets of the data. Combining datasets offers generalization, increased data diversity, an enlarged training set, mitigation of dataset bias, improved model robustness, enhanced model generalization, and addressing limited data challenges. These strengths improve performance and more reliable emotion recognition in real-world scenarios. The system does not explicitly identify which acoustic or statistical features are most relevant for recognizing emotions. Understanding the underlying mechanisms and cues that contribute to emotion recognition makes it challenging.

VI. Conclusion

The proposed approach SER-FusionNet puts forth a new concept of combining deep features, statistical features, and gender features to classify emotions. Apart from evaluating the model on an individual dataset, an ensemble dataset utilization combined with cross-validation has proven to provide a powerful model that considers a high variability in data and performs consistently. Alongside accuracy other metrics, such as precision, backs the satisfactory performance. This ensures that the proposed model is competitive in performance.

The research could further be extended to incorporate a fusion of facial expressions, and spectral image-based features, to improve its predictive power.

References

[1] Kumar, Yogesh, and Manish Mahajan. "Machine learning-based speech emotions recognition system." Int. J. Sci. Technol. Res 8, no. 7 (2019): 722-729.

[2] R.Kawade, R.Konade, P.Majukar, S.Patil. (2022). Speech Emotion Recognition Using 1D CNN LSTM Network on Indo-Aryan Database. *IEEE.*. In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT).

[3] Bisio, I., Delfino, A., Lavagetto, F., Marchese, M., & Sciarrone, A. (2013). Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications. IEEE Transactions on Emerging Topics in Computing, 1(2), 244–257.

[4] Fahmi, F., Jiwanggi, M. A., & Adriani, M. (2020). Speech-Emotion Detection in an Indonesian Movie. In Workshop Spoken Language Technologies for Under-resourced Languages (pp. 185–193).

[5] Krishna, K. V., Sainath, N., & Posonia, A. M. (2022). Speech Emotion Recognition using Machine Learning. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC).

[6] Zhang, L.-M.; Li, Y.; Zhang, Y.-T.; Ng, G.W.; Leau, Y.-B.; Yan, H.(2023). A Deep Learning Method Using Gender-Specific Features for Emotion Recognition. Sensors 2023, 23, 1355.

[7] Rumagit, R. Y., Alexander, G. L., & Saputra, I. (2021). Model Comparison in Speech Emotion Recognition for Indonesian Language. Procedia Computer Science, 179, 789–797.

[8] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

[9] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

[10] Jackson, P., & Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK.

[11] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS)-Younger talker_Neutral.

[12]https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[13]https://www.kaggle.com/datasets/ejlok1/cremad

[14]https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee

[15]https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

[16] Mittal, R., Vart, S., Shokeen, P., & Kumar, M. (2022, June). Speech Emotion Recognition. In 2022 2nd International Conference on Intelligent Technologies (CONIT) (pp. 1-6). IEEE.

[17] O'Reilly, C.,Nielsen, T. (2013b) .Revisiting the ROC curve for diagnostic applications with an unbalanced class distribution.8th International Workshop on Systems, Signal Processing and Their Applications (WoSSPA).