Speech Emotion Recognition using Deep Learning

Aashvik Chennupati
Computer Science
Arizona State University
Tempe, AZ, US
achennu1@asu.edu

Ravi Tej Chaparala
Data Science, Analytics and Engineering
Arizona State University
Tempe, AZ, US
rchapara@asu.edu

Jaya Shankar Maddipoti
Data Science, Analytics & Eng
Arizona State University
Tempe, AZ, US
jmaddipo@asu.com

Hemanth Pasula
Computer Science
Arizona State University
Tempe, AZ, US
hpasula@asu.edu

Dinesh Koushik Deshapathi
Computer Science
Arizona State University
Tempe, AZ, US
ddeshapa@asu.edu

1. Abstract

This progress report details the ongoing efforts of Group-26 in the project titled "Speech Emotion Recognition using Deep Learning." The project's objective is to develop a deep learning model that can accurately recognize emotions from speech audio recordings. This endeavor is motivated by the increasing importance of Speech Emotion Recognition (SER) in applications like virtual assistants, mental health monitoring, and customer service analysis. Our research focuses on curating a diverse dataset, designing a specialized Convolutional Neural Network (CNN) architecture, and implementing the model using TensorFlow and Keras. We aim to surpass existing SER approaches by exploring novel techniques, including transfer learning. In addition to evaluating the model's performance with standard metrics, we plan to conduct a subjective evaluation by human evaluators. The report also outlines the project management team's roles and responsibilities and provides insights into an extension of the project, which involves Scam Detection through Voice Analysis.

2. Introduction

Speech Emotion Recognition (SER) plays a pivotal role in the realm of human-computer interaction, enabling applications like virtual assistants, mental health monitoring, and customer service analysis to better understand and respond to human emotions. Group-26 has embarked on a project aimed at advancing SER through deep learning methods, with a focus on recognizing a wide range of emotions, including happiness, sadness, anger, and more, from speech audio recordings.

In this project, we tackle the real-world challenge of speech emotion recognition, aiming to introduce a novel formulation of the problem and dataset. While the problem of recognizing emotions from speech has been widely studied, our approach incorporates innovative techniques, including a specialized Convolutional Neural Network (CNN) architecture and the integration of diverse datasets such as RAVDESS, CREMA-D, TESS, and SAVEE. By combining these unique elements, we present a distinctive formulation of the problem, enhancing the accuracy and robustness of our model. Additionally, our project delves into an innovative extension, addressing the problem of scam detection through voice analysis, further expanding the scope of our research. Through these novel contributions, we strive to advance the field of speech emotion recognition and provide valuable insights for real-world applications.

The project seeks to address the fundamental question: Can a deep learning model, trained on a diverse dataset, accurately recognize emotions from speech audio recordings? The team's hypothesis is that their model will achieve state-of-the-art performance in terms of accuracy and F1 score, surpassing existing approaches. To achieve this, the project leverages a comprehensive dataset and employs advanced deep learning techniques.

In addition to SER, the team extends its research to Scam Detection through Voice Analysis. This extension involves identifying fraudulent callers who attempt to scam innocent individuals. By classifying the authenticity of a caller's voice, this system will provide a valuable tool for fraud prevention and user protection.

The project report will detail the team's approach, data collection and preprocessing steps, the use of Convolutional Neural Networks (CNNs) optimized for Mel-frequency cepstral coefficients (MFCCs), TensorFlow and Keras for model implementation, and methods for enhancing model performance. Furthermore, the team will compare their model with existing state-of-the-art SER models using various metrics, both objective and subjective, to assess its real-world applicability.

With a well-structured management plan, the team is composed of members with distinct roles, ensuring efficient coordination and collaboration in achieving project objectives. As the project progresses, the team will strive to make significant contributions to the field of speech emotion recognition and further extend their work to combat phone scams through voice analysis.

3. Related Work

In the realm of Speech Emotion Recognition (SER), the combination of neural networks and spectrograms has significantly improved accuracy. Papers like [1], [5], and [6] highlight the

importance of deep learning techniques, particularly RNNs, LSTM networks, and multi-headed attention mechanisms, for capturing emotional nuances in speech. Spectrograms provide vital spectral and temporal information.

Addressing underrepresented languages, [1] introduced an Algerian dialect dataset, achieving a remarkable 93.34% classification accuracy with an LSTM-CNN model. [2] showcased the effectiveness of combining CNN and LSTM for SER. Multiple studies emphasize feature extraction techniques like MFCC and spectrograms.

To address these deficiencies, our project focuses on expanding datasets, refining deep learning models, and exploring transfer learning. We also consider attention mechanisms as in [5] and [8]. Our comprehensive approach encompasses recognizing complex emotions, diverse language datasets, various deep learning architectures, noise reduction, multimodal data integration, and real-world applications, aiming to advance the field of Speech Emotion Recognition.

4. Dataset Description

We are using four datasets containing short voice messages (<3s) with english phrases voiced by professional actors. The datasets used are Ravee, Crema, Savee, and Tess.

These datasets contain approximately seven main emotions: Happy, Fear, Angry, Disgust, Surprised, Sad, or Neutral.

4.1. Data Preparation:

The data preparation process involved extracting emotion labels and file paths from multiple datasets: Ravdess, Crema, Tess, and Savee. In the Ravdess dataset, emotions were categorized as neutral, calm, happy, sad, angry, fear, disgust, and surprise, while the Crema dataset included sad, angry, disgust, fear, happy, neutral, and unknown emotions. The Tess dataset recorded surprise and other emotions. The Savee dataset used specific prefixes in file names to indicate emotions such as 'a' for anger, 'd' for disgust, 'f' for fear, 'h' for happiness, 'n' for neutral, 'sa' for sadness, and 'su' for surprise. This data extraction enabled subsequent analysis and modeling for emotion recognition.
The experiment mapped these emotions to their corresponding categories.

4.2. Data Augmentation:

The experiment employed data augmentation techniques, including noise injection, stretching, shifting, and pitching, to increase the variety of training data.
These techniques can create variations of audio samples to improve model robustness.

4.3. Data Processing:

The extracted features were processed, and the data was standardized to prepare it for model training. Standardization helps ensure that features have similar scales.

We will split the dataset into a training set (80%), a validation set (10%), and a test set (10%). Preprocessing steps will involve extracting Mel-Frequency Cepstral Coefficients (MFCCs) from audio recordings and normalizing the data.

Number of Instances	2076(Ravdess), 7442(Crema), 480(Savee), 2800(Tess)
Number of Features	Typically ranges from 300 to 400
Class Distribution (Number of Instances in Each Class)	Crema(anger-1026, disgust-981, fear-1017, happiness-1069, neutral-2349, sadness-1000), Ravdess(anger-252, calm-252, disgust-252, fear-252, happiness-252, sadness-252), Savee(anger-60, disgust-60, fear-60, happiness-60, neutral-120, sadness-60), Tess(anger-390, disgust-400, fear-400, happiness-390, neutral-600, sadness-620)
Dataset Splits	Training: (80%) Validation: (10%) Testing: (10%)

Fig. 1. shows the total count of each emotion across all the datasets(Ravdess, Crema, Savee and Tess) combined together.

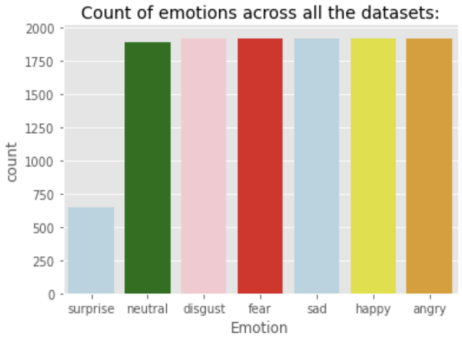

Figure 1: Total count of all emotions

5. Approach

5.1. Pre-Processing:

To prepare audio data for analysis and comparison, we need to take several essential steps. First, we should normalize the audio data by dividing it by its maximum amplitude, ensuring consistent volume levels across all clips. Next, we should identify and remove any silence at the beginning and end of the clips by trimming them appropriately. This step helps to focus on meaningful audio content. Additionally, it's important to resample all clips to a uniform sampling rate, simplifying the comparison of extracted audio features. Lastly, we should extract specific audio characteristics, such as zero-crossing rate, energy, RMS, energy entropy, spectral contrast, spectral flatness, spectral rolloff, chroma

features, short-time Fourier transform, mel spectrogram, and MFCCs, which are valuable for subsequent analysis.

Furthermore, after extracting these audio features, it's crucial to standardize them. This involves the process of subtracting the mean and dividing by the standard deviation for each feature. Standardization enhances the comparability of features across different audio clips and significantly improves the performance of machine learning models. These steps collectively form a comprehensive procedure for audio data preprocessing, ensuring that the data is consistent, informative, and ready for advanced analysis and modeling. Furthermore, we have decided to make use of zero-crossing rate, RMS and MFCC as feature extraction methods in an experimental manner.

6. Methods:

Description: This section outlines the methods used in the experiment for speech emotion recognition (SER) using a 1-dimensional Convolutional Neural Network (ConvNN).

6.1. Model Architecture:

6.1.1. CNN:

Our model architecture is designed specifically for speech emotion recognition and consists of multiple 1D convolutional layers followed by batch normalization, max-pooling layers, and fully connected layers. This design enables the model to learn hierarchical features from the audio data. The key architectural elements are as follows:

Convolutional Layers (conv1d, conv1d_1, conv1d_2, conv1d_3, conv1d_4):

- These layers are like filters that scan the input to find important patterns or features.
- They start with larger filters and gradually reduce the size of the features they look for.

Batch Normalization (batch_normalization, batch_normalization_1, batch_normalization_2, batch_normalization_3, batch_normalization_4, batch_normalization_5):

These layers help keep the network stable during training by adjusting the data.

Max Pooling Layers (max_pooling1d, max_pooling1d_1, max_pooling1d_2, max_pooling1d_3, max_pooling1d_4):

- These layers shrink the data, keeping the most important information while reducing the size.

Flatten Layer (flatten):

- This layer takes the shrunken data and makes it flat, turning it into a simple list of numbers.

Dense Layers (dense, dense_1):

- These layers are like thinking layers. They use the flattened data to make final predictions.
- The last dense layer has 7 neurons, making predictions for 7 different things.

In total, your model has over 7 million parameters that it learns from data during training.

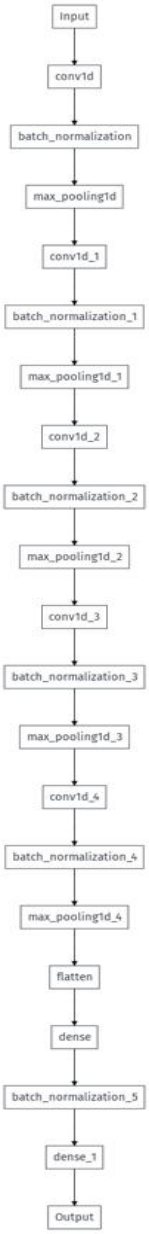

Figure 2: CNN Architecture

In conclusion, our project employs a combination of data preprocessing, specialized CNN architecture design, custom evaluation metrics, and advanced techniques to achieve the goal of speech emotion recognition. This comprehensive approach is expected to surpass existing SER approaches, making it suitable for various applications, including virtual assistants, mental health monitoring, and customer service analysis.

6.1.2. RNN Model:

Our model architecture for speech emotion recognition is tailored to leverage the temporal dependencies and sequential nature of

audio data. It predominantly relies on Recurrent Neural Networks (RNNs) and includes the following components:

Recurrent Layers: The core of our model consists of recurrent layers, specifically Long Short-Term Memory (LSTM) layers. These layers are designed to capture temporal features in the audio data, making them well-suited for understanding how emotions evolve over time.

Batch Normalization: To stabilize and expedite training, we apply batch normalization within the recurrent layers. This ensures that the model can learn emotional patterns efficiently and adapt to different emotional expressions in the audio.

Fully Connected Layers: Following the recurrent layers, our architecture incorporates fully connected layers. These layers are essential for mapping the learned temporal features to specific emotional categories. They enable the model to make emotion predictions based on the extracted audio features.

RNN is a robust and effective tool for the task of SER, which involves identifying and classifying emotions from different forms of speech. This model can be used to train on a diverse dataset comprising a large number of audio files, including the RAVDESS, CREMA-D, TESS, and SAVEE datasets. These audio files can encompass a wide range of emotions, including neutral, happy, sad, angry, fear, and disgust.

The data preprocessing phase is meticulous, involving audio trimming to remove silence and standardization to ensure a consistent length of samples for each audio clip. Emotion labels are encoded into numerical values, facilitating the training process.

The core of the model architecture consists of Long Short-Term Memory (LSTM) layers. The LSTM layers, with 64 units each, are well-suited for capturing temporal dependencies in the audio data, making them particularly effective for SER tasks. The model is concluded with a dense output layer with six units, corresponding to the six distinct emotion classes. During training, the categorical cross-entropy loss function and the RMSProp optimizer are employed to optimize the model's performance.

In terms of quantitative results, this model can be used to demonstrate impressive accuracy. This high accuracy is indicative of the model's ability to accurately classify emotions in speech, making it a promising tool for real-world applications.
By utilizing recurrent layers in our architecture and complementing them with emotion-specific evaluation metrics, we aim to build a robust and accurate model for speech emotion recognition. This approach ensures that the model excels not only in achieving high accuracy but also in recognizing and classifying emotions effectively, making it well-suited for applications in speech and affective computing.

6.1.3. Other models:

Limited Feature Extraction: Traditional non-CNN models, such as logistic regression and random forests, rely heavily on handcrafted feature engineering. For speech emotion recognition, extracting relevant features from raw audio data is a challenging and time-consuming task. CNNs, on the other hand, can automatically learn and extract relevant features directly from the spectrograms or raw audio, eliminating the need for manual feature engineering.

Spatial and Temporal Patterns: Emotion recognition in speech often involves capturing both spatial and temporal patterns in audio data. CNNs excel at this task by applying convolutional filters across different regions of the input data, capturing spatial features effectively. Additionally, they can handle the sequential nature of audio data by learning temporal patterns, which is a challenging feat for non-CNN models.

Complexity and Depth: CNNs can handle the inherent complexity and depth of audio data. They can effectively model intricate relationships between audio features and emotional expressions, which might be beyond the capability of traditional models. Speech emotion recognition is a nuanced task, and CNNs can capture the subtle variations and dependencies in the data.

Scale and Generalization: As the volume of speech data grows, CNNs can scale efficiently and generalize well. Their ability to adapt to different datasets and handle variations in speech characteristics is a key advantage. Non-CNN models might struggle to generalize to diverse datasets without extensive customization.

State-of-the-Art Performance: In recent years, CNN-based models have achieved state-of-the-art performance in various audio-related tasks, including speech emotion recognition. Their ability to capture both local and global patterns in audio data has led to impressive results in accurately recognizing emotions from speech recordings.

Adaptation to New Data: CNN models are often more adaptable to new and unseen data. They can continue to perform well even when faced with variations in speech characteristics, accents, or noise. Non-CNN models might require re-engineering and retraining when presented with different data distributions.

7. Experiments:

In pursuit of maximizing the predictive accuracy and test performance of our convolutional neural network (CNN) model, we embarked on a comprehensive series of experiments. Our goal is to explore various hyperparameters and architectural choices. We investigated the number of filters, kernel sizes, and strides in convolutional layers, allowing us to comprehend their impact on feature extraction and scale. Furthermore, the effects of batch normalization were assessed by considering its placement before or after activation functions. We delved into MaxPooling parameters, examining pool size, strides, and padding to regulate feature downsampling. The architecture's capacity was investigated

through variations in the number of units in dense layers and activation functions. Learning rate, optimizers, and loss functions were scrutinized, enabling us to tailor them to our specific task. Additionally, batch size and the number of training epochs were fine-tuned to strike a balance between convergence and overfitting. Regularization techniques, including dropout and L1/L2 regularization, were employed to enhance generalization. Data augmentation techniques were tested to increase the diversity of the training dataset. Early stopping and learning rate schedulers were adopted to expedite training and boost convergence. Different CNN architectures and neural network types were explored to gauge their suitability for the task. Hyperparameter optimization techniques were applied systematically, utilizing grid search and random search to identify optimal configurations. Each experiment was meticulously evaluated on validation and test datasets, culminating in the selection of the most effective hyperparameters for our specific problem, ensuring the model's optimal performance. We still are working on different models with different values of parameters.

8. Algorithm Choice and Implementation:

It is not finalized yet, at the moment, we have good accuracy with CNN model. We will update the finalized algorithm in the final report of the project.

9. Results and Evaluations:

9.1.1. CNN : Achieved an accuracy of approximately 95.91% while working on your Convolutional Neural Network (CNN) model. We are still in the process of improving it to achieve even higher accuracy.

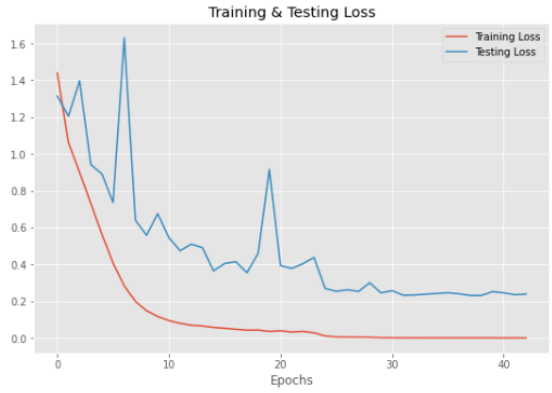

Figure 3: Training and Testing loss VS Epochs Graph

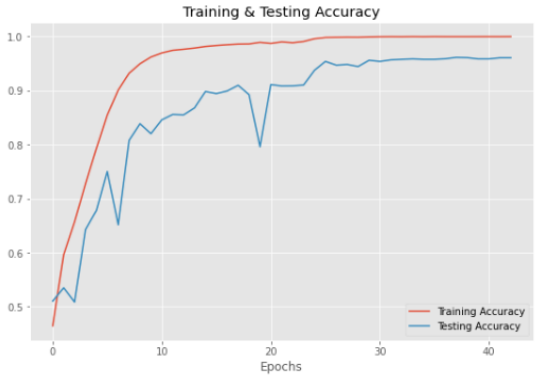

Figure 4: Training and Testing Accuracy VS Epochs Graph

Evaluation:

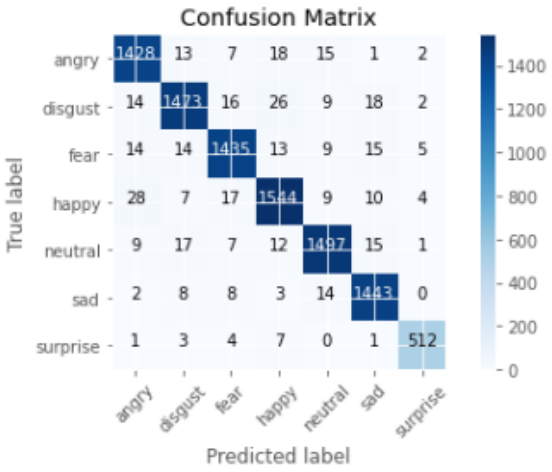

Figure 5: Confusion Matrix

Description of results for CNN:

- Class 0: There were 1,428 instances where the actual class was 0, and the model correctly predicted them as class 0. There were 13 instances of class 0 that the model incorrectly predicted as class 1, 7 as class 2, 18 as class 3, 15 as class 4, 1 as class 5, and 2 as class 6.
- Class 1: For class 1, there were 1,473 instances correctly predicted as class 1, but there were also misclassifications, such as 14 instances predicted as class 0, 16 as class 2, 26 as class 3, 9 as class 4, 18 as class 5, and 2 as class 6.
- Class 2: Similarly, for class 2, there were 1,435 instances correctly predicted as class 2, with misclassifications as well.
- Class 3: Class 3 had 1,544 correct predictions, along with some misclassifications.

- Class 4: Class 4 had 1,497 correct predictions.
- Class 5: Class 5 had 1,443 correct predictions.
- Class 6: Finally, for class 6, there were 512 correct predictions, and a few misclassifications.

Other metrics for CNN:
Still working on this.

9.1.2. RNN : Achieved an accuracy of approximately 93% while working on your Recurrent Neural Network (RNN) model. We are still working on identifying ideal parameters to improve the accuracy of the model and also compare these results with any other model used.

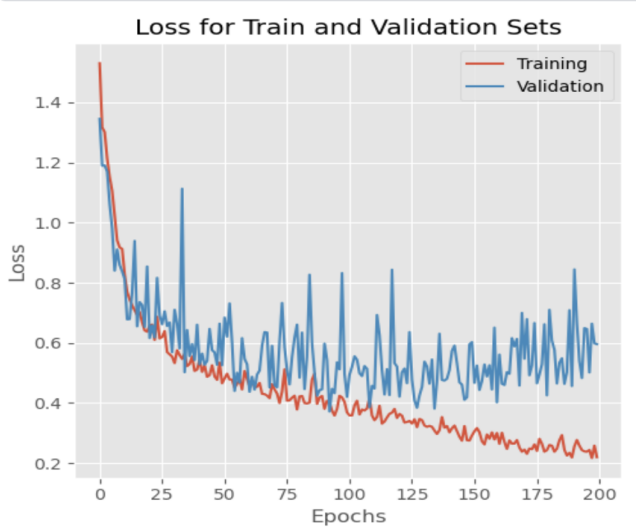

Figure 6: Training and validation loss VS Epochs Graph

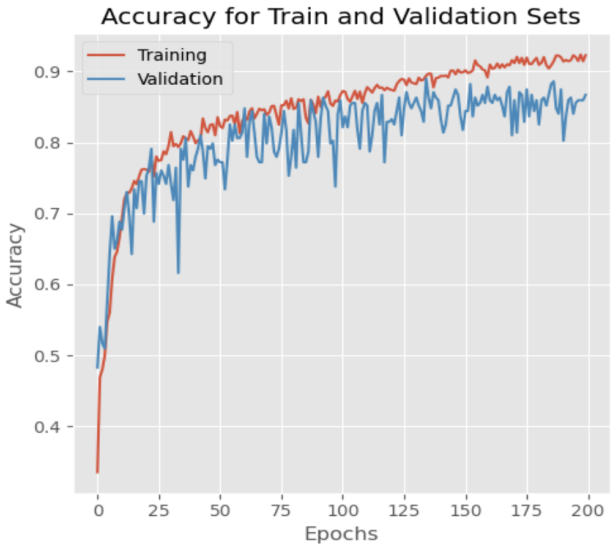

Figure 7: Train and validation Accuracy VS Epochs Graph

Evaluation:

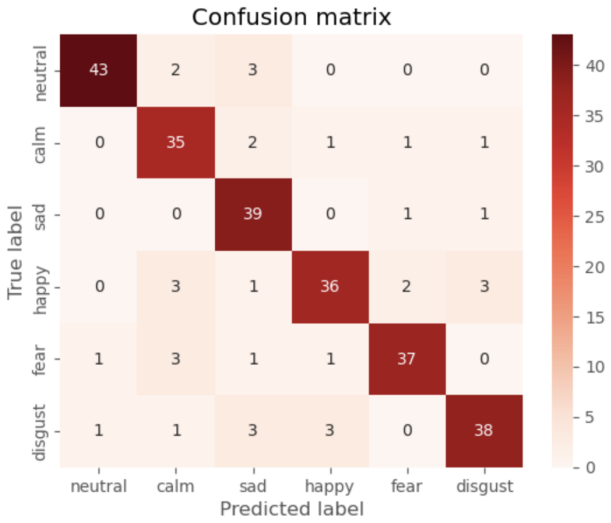

Figure 8: Confusion Matrix

Other metrics for RNN:
Still working on this.

Description of results for RNN:

- Class 0: There were 43 instances where the actual class was 0, and the model correctly predicted them as class 0. There were 2 instances of class 0 that the model incorrectly predicted as class 1, 3 as class 2.
- Class 1: For class 1, there were 35 instances correctly predicted as class 1, but there were also misclassifications, such as 2 instances predicted as class 2.
- Class 2: Similarly, for class 2, there were 39 instances correctly predicted as class 2, with misclassifications as well.
- Class 3: Class 3 had 36 correct predictions, along with some misclassifications.
- Class 4: Class 4 had 37 correct predictions.
- Class 5: Class 5 had 38 correct predictions.

10. REFERENCES

[1] R. Y. Cherif, A. Moussaoui, N. Frahta and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430224.

[2] H. Li, X. Zhang and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 795-799, doi: 10.1109/ICSIP52628.2021.9689043.

[3] Javier de Lope and Manuel Graña. 2023. An ongoing review of speech emotion recognition - ScienceDirect. ScienceDirect. Retrieved October 17, 2023.

[4] J. Bhanbhro, S. Talpur and A. A. Memon, "Speech Emotion Recognition Using Deep Learning Hybrid Models," 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC), Jamshoro, Sindh, Pakistan, 2022, pp. 1-5, doi: 10.1109/ICETECC56662.2022.10069212.

[5] X. Ying and Z. Yizhe, "Design of Speech Emotion Recognition Algorithm Based on Deep Learning," 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2021, pp. 734-737, doi: 10.1109/AUTEEE52864.2021.9668689.

[6] P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5089-5093, doi: 10.1109/ICASSP.2018.8462677.

[7] R. Ranjan, "Analysis of Speech Emotion Recognition and Detection using Deep Learning," 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2022, pp. 1-5, doi: 10.1109/IATMSI56455.2022.10119297.

[8] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1771-1775, doi: 10.23919/APSIPA.2018.8659587.

[9] M. Saloumi et al., "Speech Emotion Recognition Using One-Dimensional Convolutional Neural Networks," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 212-216, doi: 10.1109/TSP59544.2023.10197766.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327- 117345, 2019, doi: 10.1109/ACCESS.2019.2936124.