

End to End Speech Emotion Recognition

Srivatsa V¹, Vivek B V², Dr. Sumana M³ and Vashista A N⁴

¹⁻⁴ M S Ramaiah Institute of Technology, Information Science and Engineering, Bangalore, India
Email: 1MS20IS116@GMAIL.COM, SUMANA.M@MSRIT.EDU, 1MS20IS135@GMAIL.COM ,
1MS20IS131@GMAIL.COM

Abstract— The proposed model combines machine learning, artificial neural networks (ANN), and natural language processing (NLP) techniques for a robust analysis of emotions in speech. The process begins by extracting relevant features from the audio signal, such as Mel Frequency Cepstral Coefficients (MFCC), which capture the spectral characteristics of the speech signal. These features are used to train a feedforward artificial neural network (ANN) to classify the emotions in the speech, with the model optimized using techniques like dropout and adaptive learning rates to prevent overfitting and improve generalization performance. In addition to the audio features, the model incorporates natural language processing (NLP) techniques to analyse the content of the speech. It uses speech-to-text conversion to obtain the transcript of the speech, and then applies sentiment interpretation using TextBlob to calculate the sentiment polarity of the text. This information is combined with the emotion predicted by the ANN model to create a more effective emotion categorization, taking into account both the acoustic and linguistic aspects of the speech.

Index Terms— Machine learning, artificial intelligence, natural language processing.

I. INTRODUCTION

Human emotions play a significant role in communication, as they convey essential information that goes beyond the mere content of spoken words. With the rapid advancements in artificial intelligence and natural language processing, there is an increasing need for systems capable of recognizing and interpreting emotions in speech. Speech Emotion Recognition (SER) is a growing interdisciplinary field that aims to develop algorithms and techniques to automatically identify and classify human emotions from spoken language. The potential applications of SER systems are vast, spanning various domains such as virtual assistants, customer service, mental health monitoring, and affective computing.

In the proposed work, an end-to-end speech emotion recognition system that harnesses the power of machine learning, artificial neural networks (ANN), and natural language processing (NLP) to provide a comprehensive and robust analysis of emotions in speech. The primary objective of the proposed work is to design and implement a system that effectively combines acoustic and linguistic features to improve the accuracy and reliability of emotion classification in speech. For achieving this goal, the model begins by extracting relevant acoustic features from the audio signals, such as Mel-Frequency Cepstral Coefficients (MFCC), which effectively capture the spectral characteristics of speech. These features serve as the input for a feedforward artificial neural network (ANN) trained to classify emotions in speech. To enhance the model's performance, employment of optimization techniques such as dropout and adaptive learning rates to mitigate overfitting and promote generalization. In addition of acoustic features, our SER system incorporates Natural Language Processing (NLP) techniques to

interpret the textual content of speech. It utilizes speech-to-text conversion to obtain the transcripts of spoken words, followed by sentiment analysis using TextBlob to compute sentiment polarity scores. This textual information is integrated with the ANN model's predictions to derive more effective emotion classification that is considered for both acoustic and linguistic aspects of speech. The end-to-end speech emotion recognition system is thoroughly evaluated on well-known benchmark dataset, the RAVDESS dataset, to assess its performance in classifying emotions accurately. The incorporation of NLP techniques for sentiment analysis demonstrates a notable improvement in the overall achievement of the system, underscoring the benefits of combining audio and text-based features for robust and accurate speech emotion recognition.

II. OBJECTIVES

- The ANN model is trained from the labelled data using the RAVDESS audio dataset.
- Emotion of the unlabelled data is predicted using extracted features from MFCC.
- The speech in audio file is converted to text and the polarity of the text calculated
- The combined effective emotion of the entire audio file is estimated and displayed

A. Dataset Description:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a popular dataset for speech emotion identification. It contains audio and video recordings of actors performing scripted speech and songs with distinct emotions.

Here is a complete description of the RAVDESS dataset:

B. Audio-Visual Samples:

24 professional actors (12 male, 12 female) participated in the recording sessions. Each actor performed two scripted statements (a neutral statement and a statement with emotional content) with a total of 8 emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Actors also performed the exact emotional expressions with a strong intensity level. Speech and song modalities are available, with a total of 7356 recordings. All audio files are in WAV format, with a 48 kHz sampling rate and 24-bit depth. Video files are in MP4 format, with a 30-fps frame rate and 720p resolution. The duration of each audio sample varies between 3 and 5 seconds.

C. File Naming Convention:

The file names follow a specific pattern: Modality-Emotion-Intensity-Actor-Statement-Take

Modality: 01 for speech, 02 for song

Emotion: 001 = neutral, 002 = calm, 003 = happy, 004 = sad, 005 = angry, 006 = fearful, 007 = disgust, 008 = surprised

Intensity: 01 = normal, 02 = strong

Actor: 01 to 24, with even numbers representing male actors and odd numbers representing female actors

Statement: 01 = statement 1, 02 = statement 2

Take: 01 or 02 (each statement was performed twice)

III. METHODOLOGY

A. Data Collection and Preprocessing:

Collect a large dataset of speech samples with labelled emotions (RAVDESS, or TESS datasets). Convert the audio files to a suitable format (e.g., WAV) and Perform voice activity detection (VAD) to remove non-speech segments and reduce computational complexity.

B. Feature Extraction:

Extract acoustic features from the speech signals using open-source libraries like librosa. Common features include MFCCs, chroma features, spectral contrast, and pitch. For NLP-based analysis, use an ASR system like Google Speech-to-Text to convert the speech signals into text transcripts.

C. Text Preprocessing and Sentiment Analysis:

Use an NLP library like NLTK to perform sentiment analysis on the text data and compute polarity scores.

D. Model Development:

For the acoustic model, use a deep learning architecture like feed forward neural networks. Concatenate the extracted acoustic features and NLP-based polarity scores as input for a new model, which can be another deep learning architecture.

E. Training and Validation:

Split the dataset into training and testing sets (our case 80% and 20%).

Train the model on the training set using a suitable optimizer (e.g., Adam).

F. Evaluation and Optimization:

Evaluate the trained models on test set using relevant metrics such as accuracy, F1-score, and confusion matrix.

IV. LITERATURE SURVEY

In [1], The proposed work aims to classify speech into four emotions: sadness, anger, fear, and happiness. Speech samples from LDC and UGA databases were used, and important characteristics such as energy, pitch, MFCC and LPCC coefficients, and speaker rate were extracted. SVM was used as the classifier with two strategies: One against All (OAA) and Gender Dependent Classification. The study also conducted a comparative analysis between the two strategies and two algorithms, LPCC and MFCC. In [2], The paper "End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network" proposes a context-stacking dilated convolutional neural network (CS-DCNN) which is an end-to-end architecture that offers as a novel method for voice emotion recognition. Without requiring any additional intermediate feature extraction processes, the suggested model is able to learn the emotional elements straight from the voice signal. In [3], The paper "Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable- Level Prosodic Features" by Starlet Ben Alex, Leena Mary, and Ben P. Babu uses attention mechanisms, feature selection approaches, and prosodic features at the utterance and syllable levels to suggest a method for automatic speech emotion recognition.

In [4], it suggests a novel end-to-end method for speech emotion recognition. It is divided into two stages: a stage for feature extraction and another for classification. The acoustic-to-word ASR model is used in the feature extraction stage to convert the speech signal into a series of word-level embeddings. The convolutional neural network (CNN) is then utilized as input to extract high-level characteristics important for emotion recognition from the word-level embeddings. In [5] The research suggests a speech emotion identification system that takes into account both verbal and nonverbal speech sounds. The system is built on a deep neural network that accepts facial expressions and speech signals as inputs. Facial expressions are extracted using a deep convolutional neural network. In [6], A self-attention mechanism and multitask learning are used to increase the accuracy and resilience of the model in this upgraded end-to-end speech emotion identification system. The system is made up of a feature extraction module that extracts spectral and prosodic features from speech signals, a self-attention module that determines the significance of each feature for emotion recognition, and a multitask learning module that simultaneously learns to predict speaker identities and emotion labels.

In [7] gives a thorough overview of deep learning methods used in speech emotion recognition, including convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), and gated recurrent units (GRU). The attention mechanisms that let the models concentrate on the most crucial elements of the input signal are also covered by the authors. In [8], The paper "Speech emotion recognition using recurrent neural networks with directional self-attention" presents a novel method for speech emotion recognition utilizing a directed self-attention mechanism built inside a recurrent neural network (RNN). The directional self-attention method enables the model to concentrate on the speech signal's most important components and ignore its less important components. In [9], in this paper, initially it is trained and tested neural network architecture for emotion recognition from speech on the widely used corpus available to the community, IEMOCAP, in order to validate its performance. We then used the same architecture to the CEMO corpus, which contains 440 dialogs (2 hours, 16 minutes) from 485 speakers in real life.

In [10], In this study, the autoencoders are used to extract features from LibriSpeech for model pre-training, and then carried out experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets for classification. In [11], A technique termed "Head Fusion" is used, which is based on the multi-head attention mechanism to enhance the SER's accuracy. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set is used to develop an attention-based convolutional neural network (ACNN) model and run tests. To 76.18% (weighted accuracy, WA) and 76.36% (unweighted accuracy, UA) respectively, the accuracy has increased. In [12], The authors present a novel structure that extracts various sound features such as Mel-frequency cepstral coefficients, chromogram, Mel-scale spectrogram, Tonettes representation, and spectral contrast from audio files. These features are then fed into a one-dimensional Convolutional Neural Network which is designed to recognize emotions. The network is evaluated using samples from three datasets, namely RAVDESS, EMO-DB, and IEMOCAP.

In [13], This research encompasses an extensive review of deep learning methodologies for Speech Emotion Recognition (SER), in addition to an examination of traditional machine learning techniques within the same domain. The study goes further to draw a comparison across various neural network strategies in practical SER scenarios. The aim is to furnish a comprehensive overview of the state-of-the-art techniques for discrete speech emotion recognition. In [14], A cutting-edge framework for Speech Emotion Recognition (SER) is put forth by the authors, which involves selecting key sequence segments based on similarity measurements using radial basis function networks (RBFN) within clusters. These selected sequences are transformed into spectrograms through the Short-Time Fourier Transform (STFT) algorithm. Subsequently, a CNN model processes these spectrograms to extract distinct and significant features from the speech spectrogram. In [15], The study introduces the concept of using ladder networks for emotion recognition, which integrates an unsupervised auxiliary task along with the main task. The principal task involves regression analysis to predict emotional attributes, while the secondary task focuses on reconstructing intermediate feature representations via a denoising autoencoder. This approach allows for semi-supervised learning by employing large volumes of unlabeled data. The study demonstrates that this methodology significantly outperforms fully supervised single-task learning and multi-task learning benchmarks in SER.

V. SYSTEM ARCHITECTURE:

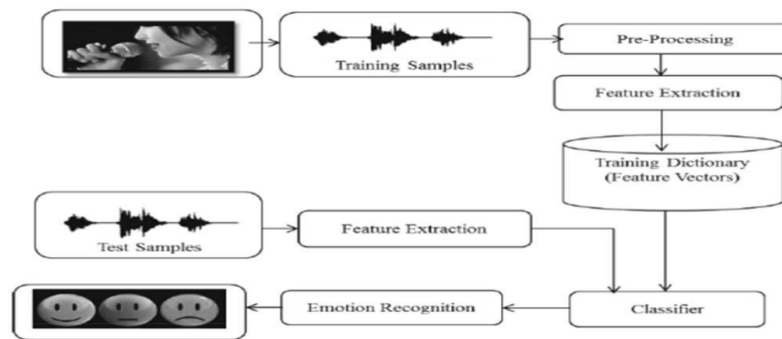


Figure 1: Architecture Diagram of the Model

A. Data Pre-processing

- **Audio Pre-processing:** Load the dataset audio files (e.g., RAVDESS dataset in WAV format). Perform noise reduction or filtering, if necessary.
- **Feature Extraction:** Extract all pertinent features from the audio samples, such as Mel-frequency cepstral coefficients, chroma features, spectral contrast, etc.

B. Data Splitting

Divide the dataset into training, validation, and testing sets.

C. Model Development

Design an appropriate ANN architecture, such as a combination of convolutional layers, recurrent layers (LSTM or GRU), and dense layers, or use a pre-trained model with transfer learning.

D. Model Training

Train the ANN model using the training set and adjust hyperparameters based on the validation set performance.

E. Model Evaluation

Test the model's performance using the testing set, and analyse metrics like accuracy, F1-score, and confusion matrix.

F. NLP for Context:

Speech-to-Text: Convert the audio samples into text transcripts using a speech-to-text service or library.

Sentiment Analysis: Perform sentiment analysis on the transcripts using NLP techniques, such as TextBlob, BERT.

G. Use Case Diagram

Figure 2 shows a use case diagram of the proposed work.

a. Actors:

- *User*: A person or another system that interacts with the speech emotion recognition system to analyse emotions in audio data.
- *System*: The speech emotion recognition system.

b. Use Cases:

- *Input Audio File*: The user input audio data, including noise.
- *Load Audio*: The system loads the audio from the user and filters out the noise during preprocessing.
- *Feature Extraction Using MFCC*: The system extracts the 39 main attributes of the audio using MFCC (13 of energy features, 13 delta features, 13 double delta features).
- *Polarity determination using NLP*: The system converts audio samples to text using a speech-to-text API or a pre-trained model and calculates the polarity.
- *Predict Effective Emotion*: The system performs sentiment analysis on the transcriptions using NLP techniques to get context and the ANN model determines the emotion by the training and the combined effective emotion is calculated by the system. Display The Predicted Emotion: The system displays the predicted effective emotion to the user.

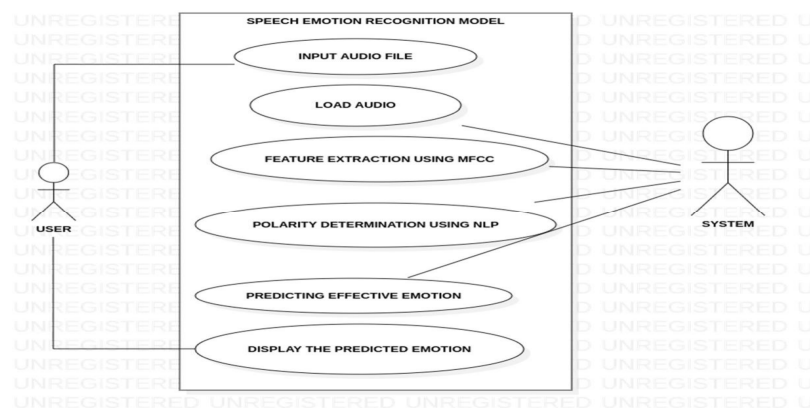


Figure 2: User System Interaction Diagram

VI. PROPOSED SOLUTION

The solution for the proposed model consists of several stages such as preprocessing, feature extraction, feature fusion, and classification. Here's an outline of the model:

A. Preprocessing

Load the speech data and convert it into a suitable format. Perform noise reduction and normalization to improve the grade of the audio signal. Apply windowing and frame segmentation to divide the audio signal into overlapping frames.

B. Feature Extraction

Acoustic Features: Extract relevant acoustic characters, such as Mel-frequency cepstral coefficients (MFCCs), pitch, and energy, which are important emotion recognition.

Linguistic Features: Convert audio data to its textual form using speech recognition techniques (e.g., Google Speech-to-Text API). Then, apply NLP techniques like sentiment analysis (e.g., using TextBlob) to obtain sentiment polarity scores and other linguistic features.

C. Feature Fusion

Combine the extracted acoustic and linguistic features into a single feature vector, capturing complementary information from both sources.

D. Classification

Train an ANN model (e.g., a feed-forward neural network) using the fused feature vectors and corresponding emotion labels. Optimize the model by fine-tuning hyperparameters, such as the count of hidden layers, neurons, and learning rate.

E. Evaluation

Evaluate the model's performance on a test dataset by calculating accuracy, precision and other metrics. Analyze the results and compare them with some other proposed solutions to assess the efficiency of the model.

Post-processing (optional): Implement a decision-making mechanism that considers both the ANN's output and the sentiment polarity to determine the final emotion label, potentially improving the overall performance.

The proposed solution aims to realize the requirements associated with recognition of emotions in audio data by integrating acoustic and linguistic features and leveraging the power of ANN for classification. By incorporating NLP techniques and feature fusion, the model has the calibre to achieve high accuracy and robustness in recognizing emotions across various speakers and real-world conditions. The proposed is an integrated model for automatic identification and classification of human emotions from vocal signals. It employs a multi-stage process that includes preprocessing of audio signals, extraction of both acoustic and linguistic features, fusion of these features into a single vector, and classification using an optimized Artificial Neural Network (ANN) model. The system also evaluates its performance using various metrics and optionally implements a decision-making mechanism that considers both the ANN's output and sentiment polarity to determine the final emotion label. By integrating machine learning, ANN, and natural language processing techniques, the system aims to achieve high accuracy and robustness in recognizing emotions across various speakers and real-world conditions.

VI. PSEUDO-CODE AND RESULTS

A. Algorithm

Load Audio Data for Emotion Classification

Input: data_path (Directory containing the audio files)

Output: data (A list of tuples, each containing file path and associated emotion label)

1. Define a dictionary 'emotion_labels' with:
 - '01' -> 'neutral' - '02' -> 'calm' - '03' -> 'happy' - '04' -> 'sad' - '05' -> 'angry' - '06' -> 'fearful' - '07' -> 'disgust' - '08' -> 'surprised'
2. Initialize an empty list 'data'
3. For each subdirectory and file in 'data_path':
 - a. If the file has an extension ".wav":
 - i. Compute the full 'file_path'
 - ii. Extract the 'file_name' without extension
 - iii. Determine 'emotion' by splitting 'file_name' on "-" and using the third segment to lookup 'emotion_labels'
 - iv. Add the tuple (file_path, emotion) to the 'data' list
4. Return the 'data' list

B. Designing and Training a Neural Network Model for Emotion Classification

1. Initialize a sequential neural network model.
2. Add to the model:
 - a. A dense layer with: 256 neurons, 'ReLU' activation, Input shape equal to the number of features in the training data
 - b. A dropout layer with a rate of 0.5 (50% dropout)
 - c. A dense layer with: 128 neurons, 'ReLU' activation
 - d. A dropout layer with a rate of 0.5 (50% dropout)
 - e. A dense layer with: 64 neurons, 'ReLU' activation
 - f. A dropout layer with a rate of 0.5 (50% dropout)
 - g. An output dense layer with: 8 neurons (corresponding to the number of classes/emotions), 'Softmax' activation (for multiclass classification)
3. Compile the model with:
 - a. Loss function: 'categorical_crossentropy' (suitable for multiclass classification)
 - b. Optimizer: 'Adam' with a learning rate of 0.001
 - c. Metrics: 'accuracy'
4. Train the model using:
 - a. Training data: X_train, y_train
 - b. Batch size: 32
 - c. Number of epochs: 100
 - d. Validation split: 0.1 (10% of the data used for validation)

C. Evaluate and Predict Emotion from an Audio File

1. Use the trained model to predict emotions on the test data:
 - a. ``y_pred`` = Predicted values from ``X_test`` using the model
 - b. Extract the class with the highest prediction score for each sample in ``y_pred`` to get ``y_pred_classes``
 - c. Extract the actual class labels from ``y_test_encoded`` to get ``y_test_classes``
2. Evaluate the model's performance:
 - a. Calculate and print the accuracy between ``y_test_classes`` and ``y_pred_classes``
 - b. Compute and print the confusion matrix between ``y_test_classes`` and ``y_pred_classes``
3. Define a function "predict_emotion" to predict the emotion from an audio file:
 - Inputs: ``file_path`` (path to the audio file), ``model`` (trained neural network model), ``scaler`` (preprocessing scaler for features)
 - a. Extract MFCC features from the audio file at ``file_path`` using librosa with 40 coefficients
 - b. Scale the computed features using the provided ``scaler``
 - c. Predict the emotion using the ``model`` on the scaled features
 - d. Extract the emotion label corresponding to the highest prediction score using a predefined list of emotion labels
 - e. Return the predicted emotion label
4. Test the "predict_emotion" function using a sample audio file path
 - Print the predicted emotion for the provided file path

D. Deriving Effective Emotion from Speech and Sentiment Analysis

1. Define a function "speech_to_text" to convert speech in an audio file to text:
 - Input: ``file_path`` (path to the audio file)
 - a. Initialize a speech recognizer
 - b. Open the audio file at ``file_path`` and record its content
 - c. i. Convert the recorded audio data to text using Google's speech recognition API
 - ii. Return the recognized text
 - d. If speech is not recognized:
 - i. Return "Unrecognized speech"
 - e. If there's an API request error:
 - i. Return an error message
2. Define a function "sentiment_polarity" to compute the sentiment polarity of a text:
 - Input: ``text`` (the text to be analyzed)
 - a. Compute the sentiment of the text using TextBlob
 - b. Return the sentiment polarity value
3. Define a function "effective_emotion" to determine the effective emotion based on sentiment polarity:
 - Inputs: ``emotion`` (original emotion), ``polarity`` (sentiment polarity of the text)
 - a. If polarity is greater than 0.2:
 - Return "positive"
 - b. Else if polarity is less than -0.2:
 - Return "negative"
 - c. Else:
 - Return the original ``emotion``
4. Load audio data and associated emotions from a given directory
5. For each audio file and its associated emotion in the loaded data:
 - a. Convert the speech in the audio file to text
 - b. Compute the sentiment polarity of the converted text
 - c. Determine the effective emotion based on the original emotion and sentiment polarity
 - d. Print details: file path, original emotion, converted text, sentiment polarity, and effective emotion

```
Emotion: neutral
Text: Tonk first sitting by the door
Polarity: 0.25
Effective emotion: positive
```

Figure 3: Emotion from Neural Network and Sentiment Analysis

```

Emotion: happy
Text: casual talking by the door
Polarity: -0.5000000000000001
Effective emotion: negative

```

Figure 4: Emotion and Polarity

VIII. CONCLUSION AND FUTURE SCOPE

Conclusion

In conclusion, the system considered, that synergistically leverage machine learning, artificial neural networks, and natural language processing have achieved remarkable strides in the domains of human-computer interaction and affection sensitive computing. The seamless integration of these cutting-edge technologies has resulted in a more accurate and reliable identification of emotions from speech signals, paving the way for a myriad of practical applications. Such applications encompass mental health monitoring, where these systems can aid in detecting early signs of emotional distress, thus facilitating timely intervention and support. In the customer service industry, these systems can be employed to gauge customer sentiment and automatically route calls to appropriate agents or escalate issues, leading to improved customer satisfaction. Furthermore, the enhancement of human-computer interfaces through emotion-aware systems can foster more natural and empathetic interactions, contributing to a more immersive user experience.

Overall, the ongoing advancements in end-to-end speech emotion recognition systems have the potential to revolutionize various industries and lead to innovative solutions that address complex challenges. As these systems continue to evolve, they will create contemporary opportunities for fostering more effective communication and empathetic understanding between humans and machines, ultimately transforming the way users interact with technology.

Future Scope:

Multimodal emotion recognition: Integrating other modalities, such as facial expressions, body language, and physiological signals, could significantly improve the functioning of emotion recognition systems. Combining these modalities would provide a more comprehensive interpretation of human emotions and lead to more robust recognition systems.

Cross-cultural and cross-linguistic studies: Extensive research should be conducted to understand the variations in emotional expressions across various cultures and languages. This will enable the bloom of culturally-sensitive emotion recognition systems that can cater to a diverse global population.

Real-world applications: With the improvement of emotion recognition systems, more practical applications will emerge in diverse fields like health, education, marketing, entertainment, and social robotics. This will open up new avenues for R&D, as well as create openings for technology transfer and commercialization.

REFERENCES

- [1] Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, Rajesh Kumar Muthu (2020). Speech Emotion Recognition using Support Vector Machine. *Journal of Imaging*, 7(9), 165.
- [2] Duowei Tang, Peter Kuppens, Luc Geurts, and Toon van Waterschoot. (2021). End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network. *Intelligent Automation & Soft Computing*, 36(2).
- [3] Starlet Ben Alex, Leena Mary, Ben P. Babu (2020, May). Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. In 2018 IEEE Canadian conference on electrical & computer engineering (CCECE) (pp. 1-4). IEEE.
- [4] Han Feng, Sei Ueno, Tatsuya Kawahara, 2020, End-to-End Speech Emotion Recognition Combined with Acoustic- to-Word ASR-Model. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)* Volume 09, Issue 06 (June 2020).
- [5] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen (2019, June). Speech emotion recognition using deep neural networks considering verbal and non-verbal speech sounds. In 2019 IEEE international advance computing conference (IACC) (pp. 617-622). IEEE.
- [6] Yuanchao Li, Tianyu Zhao, Tatsuya Kawahara. (2019). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Real-Time Image Processing and Deep Learning 2020* (Vol. 11401, pp. 13-22). SPIE.

- [7] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. (2021). A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. In 2016 2nd international conference on next generation computing technologies (NGCT) (pp. 347-351). IEEE.
- [8] Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, Zhe Wang. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. *ICTACT Journal on soft computing*, 3(4), 563-575.
- [9] Théo Deschamps-Berger; Lori Lamel; Laurence Devillers. (2021) End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings, 9th International Conference on Affective Computing and Intelligent Interaction (ACII).
- [10] Yangwei Ying, Yuanwu Tu, and Hong Zhou, (2021) Unsupervised Feature Learning for Speech Emotion Recognition Based on Autoencoder, *Electronics* 2021, 10(17), 2086.
- [11] Mingke Xu, Fan Zhang, Samee U. Khan, And Wei Zhang, (2021) Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset, *IEEE Access* (Volume: 9).
- [12] Dias Issa, M. Fatih Demirci, Adnan Yazici, (2020) Speech emotion recognition with deep convolutional neural networks, *Biomedical Signal Processing and Control*, Volume 59, May 2020, 101894.
- [13] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby, (2021) Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models, *Sensors* 2021, 21(4), 1249.
- [14] Mustaqeem; Muhammad Sajjad; Soonil Kwon, (2020) Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, *IEEE Access* (Volume: 8).
- [15] Srinivas Parthasarathy; Carlos Busso, (2020) Semi-Supervised Speech Emotion Recognition With Ladder Networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (Volume: 28).