Emotion Recognition using Speech Analysis for Mental Health Assessment by Parallelizing CNNs and Transformer-Encoders

Md. Touhidul Islam
ID: 23166028
Dept. of Computer Science and Engineering.
Dhaka, Bangladesh
md.touhidul.islam2@g.bracu.ac.bd

***Abstract*—**

***Index Terms*—Emotion recognition, speech analysis, machine learning, mental health assessment, Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), Convolutional Neural Networks (CNNs), Transfoer encoder, PyTorch.**

I. Introduction

The ability to identify emotions through speech analysis is a rapidly growing field of study with exciting potential for use in mental health evaluation. The capacity to precisely recognize and interpret emotions communicated through speech can offer insightful information about a person's emotional health, assisting in the early identification and treatment of mental health illnesses. Traditional approaches to emotion recognition have shown limitations in terms of accuracy and scalability. To improve the performance of emotion detection models for mental health evaluation, this research investigates the use of parallelizing CNNs and transformer encoders.

Understanding people's emotional states critically depends on the ability to recognize emotions like Joy, Anger, Sadness, Surprise, Fear, Disgust, and Neutral. Mental health experts can obtain better understanding of a person's emotional state and create focused interventions as a result of effectively identifying and classifying various emotions from speech signals. The complexity and diversity of speech patterns make it difficult to recognize emotions with high accuracy. To address these challenges, we are using model parallelism, a technique that uses two parallel Convolutional Neural Networks (CNN) in parallel with a Transformer Encoder network to classify audio data.

In this paper, we employ Convolutional Neural Networks (CNNs) as the primary methods for speech analysis and emotion recognition. These machine learning techniques have demonstrated effectiveness in capturing patterns and features from speech signals. We are building two parallel Convolutional Neural Networks (CNN) in parallel with a Transformer encoder network to classify audio data. We're working on the RAVDESS dataset to classify emotions from one of 8 classes.

II. Literature Review

Emotion recognition through speech analysis is a rapidly evolving field with various approaches and techniques being explored to improve accuracy and performance. In this literature review, we focus on three notable papers that have contributed to the advancement of emotion recognition in speech analysis.

MSCNN-SPU (Multi-Scale CNN with Statistical Pooling Units) is an important contribution in the field of emotion recognition through speech analysis. This model, introduced by Peng et al. [1], a multiscale CNN with statistical pooling units, for effective emotion recognition. This model learns speech and text modalities simultaneously, achieving a weighted accuracy of 79.5% and an unweighted accuracy of 80.4% on the widely used IEMOCAP dataset. The authors further propose MSCNN-SPU-ATT, an enhanced version with an attention module, which improves the performance to a weighted accuracy of 80.3% and an unweighted accuracy of 81.4%. The utilization of multiscale CNNs and attention mechanisms highlights their significance in capturing and leveraging temporal and contextual information for accurate emotion recognition in speech analysis.

In the field of emotion recognition through speech analysis, the use of mel-frequency cepstral coefficients (MFCCs) has emerged as a crucial component for capturing the spectral characteristics of speech signals. Arano et al. [2] emphasizes the importance of MFCCs in their approach to emotion recognition from speech. Arano et al. introduce a methodology that combines MFCCs and image features from pretrained convolutional neural networks (CNNs) for emotion classification from speech. By using a hybrid feature set combined with a support vector machine (SVM), an accuracy of 71% is achieved. Additionally, an MFCC-LSTM model is introduced, which slightly outperforms the hybrid feature set with an accuracy of 73%. The experiments are conducted on the RAVDESS dataset, a comprehensive audio-visual database of emotional speech and song. This work demonstrates the effectiveness of leveraging both acoustic and visual features to improve emotion recognition from speech.

A hybrid model combining a Long Short-Term Memory (LSTM) network and Transformer Encoder is proposed by Andayani et al. [3], for speech emotion recognition. The model employs Mel Frequency Cepstral Coefficients (MFCC) to extract speech features, which are then fed into the hybrid LSTM-Transformer classifier. The proposed model achieves high accuracy ranging from 72.49% to 85.55% on multiple datasets, including RAVDESS, Emo-DB, and Language-Independent datasets. This work showcases the effectiveness of combining LSTM and Transformer architectures for accurate emotion recognition from speech.

These papers collectively demonstrate the importance of leveraging advanced techniques such as multiscale CNNs, attention mechanisms, hybrid feature sets, and fusion of acoustic and visual information to improve emotion recognition in speech analysis. Furthermore, the utilization of diverse datasets, such as IEMOCAP, RAVDESS, Emo-DB, and Language-Independent datasets, contributes to the validation and evaluation of the proposed models. Building upon these prior studies, our research aims to enhance the field by incorporating distributed model parallelism using TensorFlow or PyTorch. This approach offers the potential to further improve the accuracy and scalability of emotion recognition models for mental health assessment.

III. Datasets

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) datasets will be used to train and test the proposed models in this study on "Emotion Recognition Using Speech Analysis for Mental Health Assessment Through Model Parallelism." These datasets offer a wealth of information for researching emotional speech analysis and include a variety of professional actors' expressive emotional ranges.

A. RAVDESS

The RAVDESS dataset is a well-known and often used audio-visual collection that contains a broad range of emotional expressions. 24 professional actors—12 men and 12 women—perform 1400 audio recordings in all, portraying various emotions in a staged setting. vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. These are just a few of the numerous metadata included in the RAVDESS dataset that can be used to further analyze and comprehend emotional speech. The audio files are recorded in 16-bit mono at a 48 kHz sample rate.This makes it beneficial for a variety of audio analysis applications. The recordings in RAVDESS are of great quality and capture intonation, rhythm, and timbre as well as other elements of emotional speech. With the help of this dataset, researchers can investigate the auditory characteristics and patterns connected to various emotions, facilitating the creation and assessment of reliable emotion detection models.

The emotion recognition field has made substantial use of the RAVDESS dataset, giving researchers standardized and verified resources for creating and comparing emotion recognition algorithms. This dataset have aided a number of studies on the analysis of emotional speech, advancing the diagnosis and treatment of mental illness.

For the purpose of this research, the RAVDESS will be applied for preprocessing steps to ensure consistency and compatibility with the proposed model parallelism approach. Relevant acoustic properties, such as mel-frequency cepstral coefficients (MFCCs), will be retrieved from the audio recordings once they have been segmented. To allow supervised learning for emotion identification models, the emotional categories in the datasets will be accurately labeled.

IV. Proposed Methodology

The research proposal includes numerous essential steps, including preprocessing, feature extraction, classification, and decision making. We are implementing two parallel Convolutional Neural Networks (CNN) in parallel with a Transformer encoder network to classify audio data. The main data sources for training and testing are the RAVDESS and TESS databases.

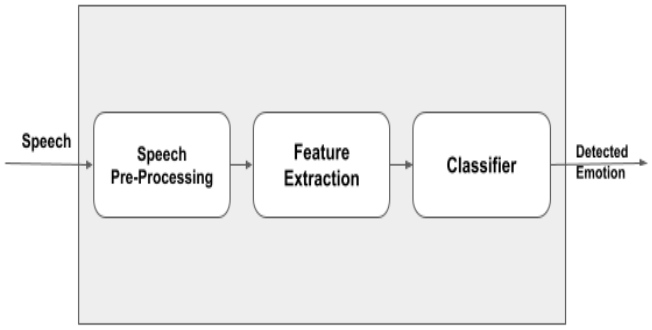

Fig. 1. Speech Emotion Recognition Step.

A. Preprocessing

The preprocessing stage involves several essential steps to prepare the audio data for subsequent analysis. These procedures include windowing, which divides the audio into smaller frames, silent removal, which gets rid of non-speech parts, background noise removal, which improves the quality of the speech signal, and normalizing, which maintains constant amplitude levels throughout several recordings. These preprocessing methods seek to raise the standard and dependability of the features that are extracted. The main important parts for preprocessing are data loading and labelling, converting audio to webforms, dataset splitting, augmenting the data with Additive White Gaussian Noise (AWGN) etc.

1) Data loading and labeling: In the beginning, we are loading audio data from the RAVDESS dataset and extract emotion labels. Additional labels like intensity and gender are

also obtained. Each audio file is loaded with 3 seconds of the file and we are cutting off the first 0.5s of silence. After that the waveform is extracted from the audio files. We are splitting training, validation, and test sets from the dataset. Separate the emotions one at a time, then shuffle the indices to establish balance. We are using 80% data for trainning, 10% for validation, and 10% for testing. This process ensures that all sets represent emotions in a diverse and equitable way.

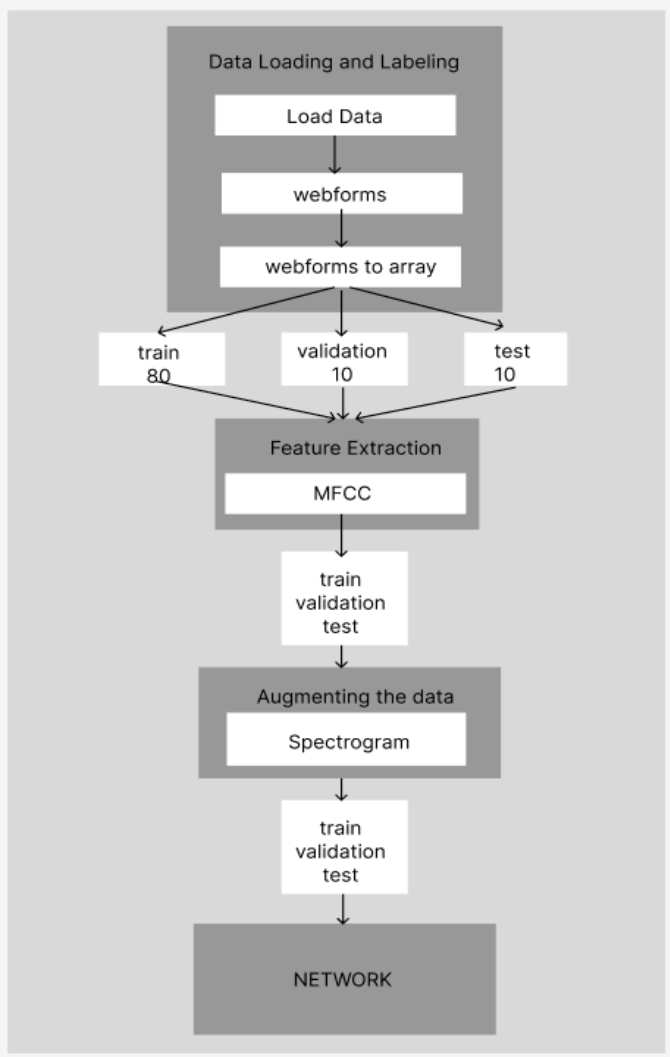

Fig. 2. Preprocessing Steps.

2) Feature Extraction: In the feature extraction phase, features are extracted by using Mel-Frequency Cepstral Coefficients (MFCCs) from the audio waveforms. This process creates a set of 40 coefficients for each waveform, capturing important spectral characteristics.

3) Augmenting the data: Given the small size of our dataset, the risk of overfitting is high, especially when dealing with complex deep neural network models. To address this challenge, we will employ data augmentation techniques. While generating additional real samples is challenging, we can effectively augment the dataset by introducing white noise into the audio signals. This serves a dual purpose: it helps mask the impact of random noise present in the training data and generates pseudo-new training examples. This augmentation strategy helps counterbalance the influence of intrinsic noise in the dataset. Our chosen approach is Additive White Gaussian Noise (AWGN) augmentation. It's "additive" because we add it to the original audio signal, "Gaussian" because the noise vector is sampled from a normal distribution with a mean of zero (zero-mean), and "white" because the noise, after a specific transformation, uniformly increases the power of the audio signal across the frequency spectrum.

B. Classification

Our neural network architecture designed to predict emotions from spectrogram-like input data. CNN and transformer blocks are used in parallel in this architecture, taking advantage of each one's unique characteristics to analyze the incoming data thoroughly. Using both spatial and temporal aspects, this parallel structure is effectively constructed.

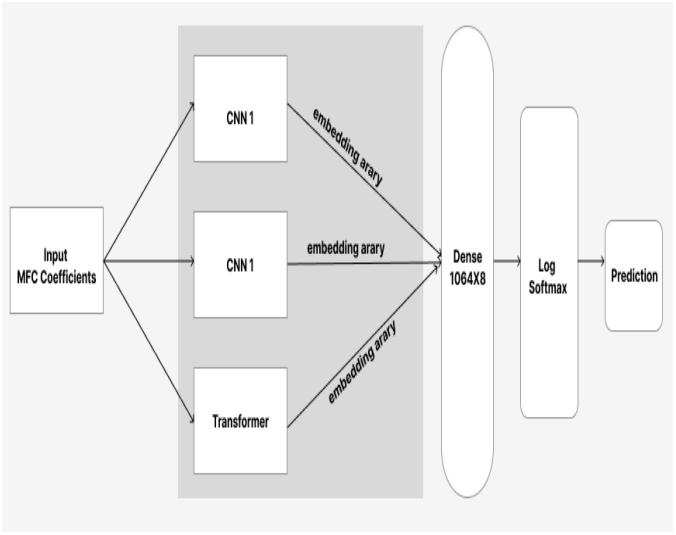

Fig. 3. Speech Emotion Recognition Step.

This architecture uses transformer encoder. In the tersformer block the input data undergoes a max-pooling operation, facilitating its efficient processing by the subsequent transformer encoder layer. This layer, comprising four stacked units, capitalizes on multi-head self-attention mechanisms and feedforward networks. Simultaneously, two parallel convolutional blocks engage with the input. Each convolutional block consists of three sequential 2D convolutional layers, strengthened by batch normalization, ReLU activation, max-pooling, and dropout. This parallel arrangement enables the network to simultaneously capture fine-grained local patterns through convolutional operations and learn broader temporal dependencies using the transformer mechanism.

Proceeding further, the embeddings extracted from both convolutional blocks are flattened and merged with the transformer's output. This encapsulates the collective insights from

both spatial and temporal analysis. Subsequently, a linear layer generates logits, forming the basis for loss computation. A softmax layer creates probability distributions for emotion prediction

In a nutshell, this architecture skillfully makes use of parallelism to simultaneously capture intricate local details and overarching temporal context. By integrating convolutional and transformer blocks, the model acquires a holistic understanding of spectrogram-like data, ultimately achieving remarkable precision in emotion prediction.

C. Predictions

Once the classification models have been trained and evaluated, they are ready to make predictions on new, unseen data. The output of the models represents the predicted emotional states, which are connected to particular emotional categories like calm, happy, sad, angry, fearful, surprise, and disgust expressions. Based on the score the classifier has given, a final determination is made regarding the emotion conveyed in the speech.

V. Result Analysis

VI. Challenges

The development of the emotion detection model using deep learning techniques poses several challenges that merit attention. One major challenge is the availability of high-quality and diverse datasets for training and evaluation. While the current study utilized the RAVDESS dataset, finding more extensive and representative datasets encompassing a wide range of emotions, cultural backgrounds, and linguistic variations remains a significant challenge. Addressing this challenge is essential to ensure the model's generalization to real-world scenarios and across different populations.

Another critical challenge is the tuning of hyperparameters. The architecture involves a combination of different neural network components, each with its own set of hyperparameters. Finding the optimal values for parameters such as learning rates, dropout rates, batch sizes, and transformer layer configurations requires extensive experimentation. Hyperparameter tuning significantly impacts the model's performance, and striking the right balance between components is a non-trivial task.

VII. Future Work

There are several promising directions for future work in the domain of emotion detection from audio signals. One intriguing section is the incorporation of Recurrent Neural Networks (RNNs) into our current architecture. RNNs are adept at capturing temporal dependencies in sequential data, such as audio signals, which can provide a deeper understanding of how emotions evolve over time. This can potentially lead to enhanced emotion recognition accuracy by capturing nuanced emotional transitions that occur within audio clips.

Furthermore, alongside the sophisticated deep learning models, exploring traditional machine learning approaches like Support Vector Machines (SVMs) can offer valuable insights. By combining both deep learning and classical techniques, we can perform a comparative analysis to identify scenarios where one approach outperforms the other. This comparative assessment can provide a comprehensive perspective on the strengths and limitations of different methods in various emotion recognition scenarios.

VIII. Conclusion

In conclusion, the present study has successfully developed an emotion detection model using a two parallel CNNs in parallel with a Transformer encoder. This architecture has shown promising results in accurately recognizing emotions from audio data. However, there are exciting opportunities for future research to further enhance the model's performance. Incorporating Recurrent Neural Networks (RNNs) to capture temporal dependencies, exploring traditional machine learning approaches like Support Vector Machines (SVMs) for comparative analysis, and fine-tuning the existing architecture are potential avenues. Additionally, expanding the emotion categories and dataset can lead to a more comprehensive emotion recognition system. By advancing in these directions, we can continue to improve the accuracy and robustness of emotion detection from audio signals.

References

[1] Z. Peng, Y. Lu, S. Pan and Y. Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3020-3024, doi: 10.1109/ICASSP39728.2021.9414286.

[2] Arano, Keith and Gloor, Peter and Orsenigo, Carlotta and Vercellis, Carlo. (2021). When Old Meets New: Emotion Recognition from Speech Signals. Cognitive Computation. 13. 1-13. 10.1007/s12559-021-09865-2.

[3] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," in IEEE Access, vol. 10, pp. 36018-36027, 2022, doi: 10.1109/ACCESS.2022.3163856.