Emotion Recognition using Speech Analysis for Mental Health Assessment Through Distributed Model Parallelism

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), 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 distributed model parallelism using PyTorch.

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 distributed model parallelism, a technique that distributes the computational workload by partitioning the emotion recognition model and utilizing parallel processing, this approach aims to improve the efficiency and scalability of the emotion recognition process. The distributed model parallelism approach will be implemented and tested on a single computer or machine.

In this paper, we employ Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and 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. Additionally, by using libraries and frameworks that provide automatic parallelization capabilities, such as PyTorch we can define and train models using high-level APIs while automatically handling the distribution of computations across available resources on a single machine. Furthermore, Model parallelism has a number of benefits over traditional approaches. Model parallelism eliminates the need for a distributed cluster of workstations by enabling the application of distributed computing capabilities on a single machine. The solution is therefore simpler to implement and uses fewer resources and sophisticated processes. Distributed model parallelism facilitates faster inference and training by enhancing the scalability and effectiveness of emotion recognition models.

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) and TESS (Toronto Emotional Speech Set) 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 Distributed 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 7356 audio recordings in all, portraying various emotions in a staged setting. Each actor does two trials for each of the seven basic emotions: neutral, calm, happy, sad, angry, afraid, and disgusted. Age, race, and gender 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 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.

B. TESS

Another useful tool for study on emotion recognition is the TESS dataset, sometimes referred to as the Toronto Emotional Speech Set. It features two female performers expressing their emotions in their genuine way. TESS comprises of 2800 audio recordings in total, 200 samples for each of the 7 emotional categories of surprise, neutral, fear, anger, disgust, and happiness. Researchers can examine the audio signals and patterns connected to distinct emotional states because to the dataset's rich collection of emotional states. To ensure controlled and reliable emotional reactions, the recordings in TESS have been carefully created to elicit particular emotions.. With the use of this dataset, it is possible to examine the performance of emotional speech and evaluate how well emotion detection models work when used with various emotional categories.

The emotion recognition field has made substantial use of the RAVDESS and TESS datasets, giving researchers standardized and verified resources for creating and comparing emotion recognition algorithms. These datasets 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 and TESS datasets will be applied for preprocessing steps to ensure consistency and compatibility with the proposed distributed 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. Convolutional neural networks (CNNs), Gaussian mixture models (GMMs) and Support vector machines (SVMs) are all required for the implementation and evaluation of the suggested approach. The main data sources for training and testing are the RAVDESS and TESS databases.

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.

B. Feature Extraction

In the feature extraction phase, various acoustic features are extracted from the preprocessed audio files. These characteristics may include pitch, loudness, intensity, rhythm, and other pertinent traits that capture significant patterns and cues associated with emotional expression in speech. The requirements for the research as well as the qualities of the dataset will determine which specific features are chosen. For next classification algorithms, the retrieved features are used as input.

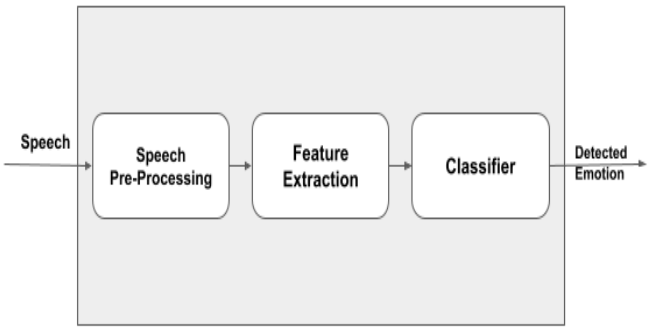

Fig. 1. Speech Emotion Recognition Step.

C. Classification

To begin the classification phase of the emotion recognition task, an initialization classifier is set up. The initial classifier may then be trained using a portion of the available data, and its performance may then be assessed. Therefore, cutting-edge methods like neural networks, including SVM, GMM, and CNN architectures, are used to improve the reliability and accuracy of the emotion identification models. These classifiers develop their knowledge from the extracted features and base their predictions on discovered patterns and connections.

D. Decision

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 surprise, fear, disgust, and neutral. Other emotional states represented by the output include joy, rage, sadness, and neutrality. Based on the greatest confidence or likelihood score the classifier has given, a final determination is made regarding the emotion conveyed in the speech.

With the use of distributed model parallelism approaches, the project intends to efficiently identify and categorize emotions from speech signals. Utilizing SVMs, GMMs, and CNNs allows for the investigation of complex machine learning methods as well as the employment of parallel computing resources for enhanced performance and scalability. The RAVDESS and TESS datasets, which offer a wide spectrum of emotional expressions for in-depth analysis and testing, will be used to validate and assess the study methods. The experimental findings and performance assessment of the suggested strategy on these datasets support the

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.