A Privacy Preserving Speech Emotion Recognition using Federated Learning

Md. Reasad Zaman Chowdhury
ID: 23166016
BRAC University

Alvin Rahul Hore
ID: 23166005
BRAC University

Rabea Akhter
ID: 23366029
BRAC University

Mashfurah Afiat
ID: 23366039
BRAC University

Alex Sarker
ID: 23373008
BRAC University

***Abstract*—**

***Index Terms*—**

I. Introduction

Speech emotion recognition (SER) is a challenging task due to the variability of human speech and the lack of large, labeled datasets. It has a wide range of applications, such as healthcare, customer service, and security. Traditional SER approaches typically require centralized training, which can raise privacy concerns. In this paper, we propose a privacy-preserving SER approach using federated learning. Federated learning is a machine learning paradigm that allows multiple devices to train a shared model without sharing their data. This makes it a promising approach for SER, as it can help to protect the privacy of users' speech data. For example, the user's data could be used to track their emotional state, or it could be used to build a profile of their personality. Another concern is that the user's data could be de-anonymized. This means that the user's identity could be revealed from their data. Thus, if the user's data is combined with other data, such as their location data, it could be possible to identify the user.

Our proposed approach consists of two phases: a training phase and a testing phase. In the training phase, each device trains a local model on its own data. The local models are then aggregated to form a global model. In the testing phase, each device uses the global model to classify its own speech data. We evaluated our approach on the IEMOCAP and CREMA-D datasets. The results showed that our approach can achieve comparable performance to traditional SER approaches, while providing better privacy protection.

One of the main concerns is that SER can be used to track people's emotional states without their knowledge or consent. This could be used to discriminate against people based on their emotions, or to manipulate their emotions. Another concern is that SER can be used to extract sensitive information from people's speech, such as their political beliefs or their health status. This information could then be used to harm people or to violate their privacy. For these reasons, it is important to consider privacy when developing SER systems. There are a number of ways to protect privacy in SER, such as using anonymized data, federated learning, and differential privacy. Data privacy, Model privacy, Inference attacks are some of the privacy concerns that arise in SER.

Our proposed approach for privacy-preserving speech emotion recognition (SER) using federated learning has several benefits. Federated learning allows us to train a shared model without sharing the users' speech data. This means that the users' privacy is protected, as their data is not exposed to the server or to other users. Federated learning is a scalable approach, as it can be used to train models on a large number of devices. This makes it a promising approach for SER, as it can be used to collect data from a large number of users. Our approach can achieve comparable performance to traditional SER approaches. This means that we can still achieve good accuracy while protecting the users' privacy. Our proposed approach for privacy-preserving SER using federated learning has several limitations too. However, we believe that these limitations are outweighed by the benefits of the approach. We believe that our approach has the potential to make SER more privacy-preserving and accessible to users. There are a number of directions for future work on privacy-preserving speech emotion recognition (SER) using federated learning and that can be done by Improving accuracy, Reducing communication and computational overhead, Addressing security concerns and using other machine learning paradigms. We believe that this is a promising area of research, and we believe that future research will make it possible to develop more accurate, efficient, and secure SER systems that protect the privacy of users.

II. Literature Review

Traditional approaches to emotion recognition have focused on either audio or visual modalities, but recent research has shown that multimodal approaches can achieve better performance. This is because audio and visual modalities contain complementary information about emotion. For example, the audio modality can provide information about the pitch, loudness, and timbre of a person's voice, which can be indicative of their emotional state. The visual modality can provide information about a person's facial expressions, which can also be indicative of their emotional state.

The paper ”A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition” proposes a joint cross-attention model for audio-visual fusion in dimensional emotion recognition. The model first learns individual feature representations for each modality. Then, it uses a joint cross-attention mechanism to learn how to attend to both modalities simultaneously. This allows the model to learn how to fuse the complementary information from both modalities in order to make more accurate predictions of emotion.

One of the earliest studies in this area was conducted by Picard et al. in 1997. This study used a multimodal approach to recognize emotions in speech and facial expressions. The study found that the multimodal approach was able to achieve better performance than either modality alone.

More recent studies have further improved the performance of multimodal emotion recognition. For example, Zeng et al. proposed a multimodal deep learning model that was able to recognize emotions in speech and facial expressions with high accuracy

The joint cross-attention mechanism works by first computing attention weights for each modality. The attention weights for a modality indicate how much attention the model should pay to that modality. The attention weights are then used to fuse the feature representations from the two modalities. The fused feature representations are then used to make a prediction of the person's emotional state.

The authors evaluated their model on the AffWild2 dataset, which is a large-scale dataset of in-the-wild facial and vocal expressions. The results showed that the proposed model outperformed state-of-the-art methods on the AffWild2 dataset.

The paper ”A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition” makes a number of contributions to the field of emotion recognition. First, the paper proposes a novel joint cross-attention mechanism for audio-visual fusion. Second, the paper provides a comprehensive evaluation of the proposed model on a challenging dataset. Third, the paper shows that the proposed model can outperform state-of-the-art methods on the AffWild2 dataset.

The authors of the paper discuss several directions for future work. One direction is to investigate the use of the proposed model for other tasks, such as facial expression recognition and affect detection. Another direction is to explore the use of different attention mechanisms for audio-visual fusion. Finally, the authors plan to investigate the use of the proposed model in real-world applications.

Emotion recognition from speech is a challenging task due to the variability of human speech and the lack of large, labeled datasets. Recent advances in deep learning have led to significant improvements in the performance of emotion recognition systems. However, there is still room for improvement, especially in terms of the robustness of these systems to noise and other factors that can degrade the quality of the speech signal.

A number of different approaches have been proposed for emotion recognition from speech. Early approaches typically used handcrafted features, such as pitch, energy, and spectral features. However, these approaches were often limited in their ability to capture the complex patterns of acoustic cues that are associated with different emotions. In recent years, there has been a growing trend towards using deep learning for emotion recognition from speech. Deep learning models have been shown to be able to learn more complex patterns of acoustic cues than handcrafted features. Additionally, deep learning models can be trained on large datasets, which can help to improve their performance. Some of the most common deep learning models that have been used for emotion recognition from speech include LSTMs (long short-term memory networks), CNNs (convolutional neural networks), and Transformers. LSTMs are well-suited for this task because they can learn long-term dependencies in the speech signal. CNNs are good at extracting local features from the speech signal. Transformers are able to attend to important features in the speech signal, which can help to improve the performance of emotion recognition systems.

The proposed model in the paper is a hybrid LSTM-Transformer model. The model combines the strengths of LSTMs and Transformers to learn long-term dependencies and attend to important features in the speech signal. The model is evaluated on three datasets: RAVDESS, Emo-DB, and language-independent. The results show that the proposed model achieves state-of-the-art performance on all three datasets.

The paper makes several contributions. First, the authors propose a novel hybrid model that combines the strengths of LSTMs and Transformers. Second, the authors evaluate the proposed model on three different datasets, showing that it achieves state-of-the-art performance. Third, the authors discuss the challenges of emotion recognition from speech audio files and how their proposed model addresses these challenges.

The paper is a well-written and informative contribution to the field of emotion recognition from speech audio files. The proposed model achieves state-of-the-art performance, and it is evaluated on three different datasets. The paper also discusses the challenges of this task and how the proposed model addresses these challenges.

The authors of the paper suggest several directions for future work. These include investigating the use of other deep learning models for emotion recognition from speech audio files, using the proposed model for other tasks, and addressing the limitations of the proposed model, such as its sensitivity to the quality of the speech signal. Overall, the paper is a valuable contribution to the field of emotion recognition from speech audio files. The proposed model is a promising approach for this task, and it is likely to be further improved in future work.

From this paper “Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions” proposes a speech emotion recognition method that combines speech features and speech transcriptions (text) to improve emotion detection. The authors experimented with several Deep Neural Network (DNN) architectures, which take in different combinations of speech features and text as inputs. The proposed network architectures achieve higher accuracies when compared to state-of-the-art methods on a benchmark dataset. The combined MFCC-Text Convolutional Neural Network (CNN) model proved to be the most accurate in recog-

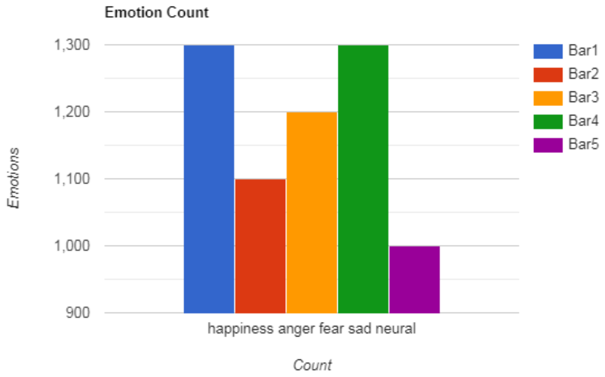

Fig. 1. Enter Caption

nizing emotions in IEMOCAP data.

The authors suggest that speech features such as Spectrogram and Mel-frequency Cepstral Coefficients (MFCC) help retain emotion-related low-level characteristics in speech whereas text helps capture semantic meaning, both of which help in different aspects of emotion detection. This approach has the potential to improve the accuracy of speech-based emotion recognition systems.

III. Dataset

We are using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Crowded Source Emotional Multimodal Actors Dataset(CREMA-D) · **IEMOCAP** dataset is collected by following theoretical theory to simulate natural interaction between the actors. Here we are using only five emotional categories :happy, angry, sad, fear, and neural to compare the performance of our model with other research using the same categories. In this session IEMOCAP dataset has five sessions, from every session there are two speakers , one male and another female. We need to process using 10 unique speakers for consistent comparison with other works. Then we collected all data with happiness, anger, sadness, fear, neural (1636, 608, 1103, 1084, 1708) with 6139 utterances. Utterances spoken with fixed script are spontaneously spoken. In this work we focus on the session following form [1][2] paper for our experiment. · **CREMA-D** is an emotional multimodal actor data set. Actors spoke from a selection of sentences.The sentences were presented using one of different emotions. For our research paper we are using 5 data sets(happiness, anger, fear, sad, neural).We are using bar graphs to show our result. A speech emotion detection classifier on CREMA dataset. This classifier attempts to recognize human emotions.

Figure 1: Emotions Count using Dataset(CREMA-D)

Figure 2: Wave plot for particular wave emotion sad

As we have done this experiment with those dataset and our work is continuing with this experiment.

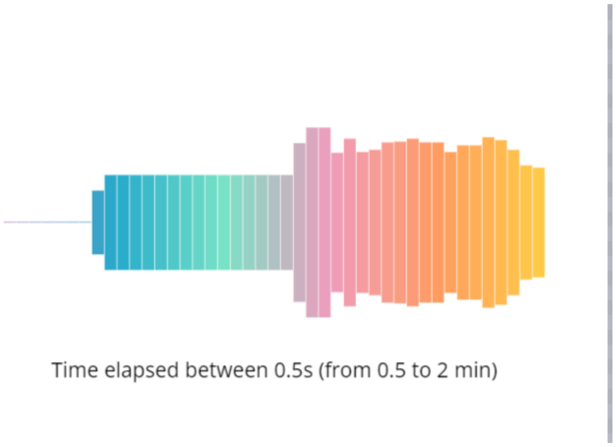

Fig. 2. Enter Caption

IV. Methodology

In this paper, we propose a privacy-preserved federated learning approach for audio emotion recognition using Mel-frequency cepstral coefficients (MFCC) feature. Different machine learning and deep learning-based architecture are used for the clients and server side in our federated learning-based approach. The methodology encompasses data preprocessing, client-side model training and server-side model aggregation and performance evaluation. Fig 3 shows the workflow of the proposed methodology.

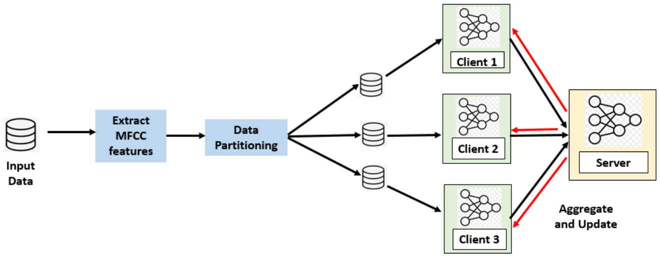

Fig. 3. Overall Workflow of Proposed Methodology

A. Data Preprocessing

Raw audio speeches are often unsuitable for direct use in audio recognition tasks due to their inherent redundancy and noise. The presence of redundant information and noise can significantly hinder the accuracy of recognition systems. To overcome this challenge, we leveraged the mel-frequency cepstral coefficients (MFCC) features for this task. The process of MFCC calculation underwent seveal stages. First, the signals were normalized and subjected to a denoising algorithm. Frame segmentation was performed with a frame size of 20 milliseconds and a step size of 20 milliseconds. MFCC features were extracted through a series of steps including pre-emphasis, windowing, FFT, Mel-filter bank, logarithmic compression, and DCT. A subset of 13 discriminatory MFCC coefficients was selected. The preprocessed data, along with emotion labels, were split into client-specific subsets for federated learning, ensuring balanced emotion distribution.

B. Client-side Model Training

On each client device, three different deep learning models, namely Deep Neural Network (DNN), Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) were trained using the MFCC data for audio emotion recognition.

1) Deep Neural Network: The DNN architecture consists of multiple hidden layers of interconnected neurons that learn complex representations from the input data. During training, the model learns to map the input features to the corresponding labels by minimizing a loss function through backpropagation. The model parameters are updated using optimization algorithms to improve classification accuracy. It leverages the power of deep learning to capture intricate patterns and relationships within the data.

2) Convolutional Neural Network: The CNN architecture is characterized by convolutional layers, pooling layers, and fully connected layers. The convolutional layers employ filters to extract local features from the input, capturing important spectral and temporal patterns. The pooling layers downsample the extracted features, enhancing the model's ability to generalize. Finally, the fully connected layers combine the learned features for classification. The hierarchical structure of CNN makes it particularly suitable for extracting spatial patterns from data.

3) Long Short Term Memory (LSTM): LSTM architecture is specifically designed to capture temporal dependencies in sequential data. It consists of memory cells, input gates, and output gates that retain information over time and selectively update or retrieve information as needed. The MFCC features are treated as sequential inputs, allowing the LSTM model to model the context and temporal dynamics in the audio data. LSTM's ability to capture long-term dependencies and handle variable-length sequences makes it effective in many tasks.

C. Server-side Model Aggregation:

The server-side model aggregation phase combined the client-trained models into a global model using federated averaging. Synchronized parameters were shared between the server and clients, ensuring consistency. Through iterative communication rounds, the global model benefited from the diverse expertise of the client models, improving audio emotion recognition performance while preserving data privacy and security. The server-side model aggregation phase enhanced the global model's accuracy and effectiveness.

D. Performance Evaluation

To assess the effectiveness of the audio emotion recognition system, several evaluation metrics such as accuracy, precision, recall and F1 were employed on the global model. By evaluating the trained models using these metrics, a comprehensive understanding of our proposed method is achieved.

V. Result Analysis

VI. Conclusion

References

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018, pp. 1771–1775.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.

[4] Tripathi, S., Kumar, A., Ramesh, A., Singh, C., Yenigalla, P. (2019). Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. arXiv preprint arXiv:1906.05681.