EmoTech: A Multi-modal Speech Emotion Recognition Using Multi-source Low-level Information with Hybrid Recurrent Network

Shamin Bin Habib Avro, Taieba Taher, Nursadul Mamun
Robust Speech Processing Laboratory (RSPL)
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology, Chittagong
ovrohabib@gmail.com, taieba.athay@cuet.ac.bd, nursad.mamun@cuet.ac.bd

***Abstract*—Emotion recognition is a critical task in human-computer interaction, enabling more intuitive and responsive systems. This study presents a multimodal emotion recognition system that combines low-level information from audio and text, leveraging both Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory Networks (BiLSTMs). The proposed system consists of two parallel networks: an Audio Block and a Text Block. Mel Frequency Cepstral Coefficients (MFCCs) are extracted and processed by a BiLSTM network and a 2D convolutional network to capture low-level intrinsic and extrinsic features from speech. Simultaneously, a combined BiLSTM-CNN network extracts the low-level sequential nature of text from word embeddings corresponding to the available audio. This low-level information from both speech and text is then concatenated and processed by several fully connected layers to classify the speech emotion. Experimental results demonstrate that the proposed EmoTech accurately recognizes emotions from combined audio and text inputs, achieving an overall accuracy of 84%. This solution outperforms previously proposed approaches for the same dataset and modalities.**

***Index Terms*—Speech Emotion Recognition, Multimodal, BiLSTM, CNN, Text, MFCC**

I. Introduction

Emotion recognition from speech has gained significant attention due to its wide range of applications in human-computer interaction, mental health monitoring, and customer service [7], [8], [13]. The ability to accurately recognize emotions in speech can enhance the effectiveness of automated systems, making interactions more natural and responsive to users' emotional states. Traditional methods for emotion recognition have primarily focused on using either audio or text features independently. However, leveraging both modalities simultaneously can potentially improve the performance of Speech Emotion Recognition (SER) systems by capturing complementary information from both audio and text data.

Several studies have been proposed in recent years to incorporate audio and text to improve emotion recognition [3], [9]–[11]. For instance, Lee et al. proposed an SER system using text and audio features with a logical "OR" function at the decision level to combine acoustic and linguistic information [6]. Later, Jin et al. suggested combining auditory and linguistic information and training them with an SVM classifier to identify emotion categories [5]. Recent advancements in deep learning have also been utilized for emotion classification. Griol et al. trained three datasets using machine learning classifiers to categorize user emotions based on spoken utterances [4]. Yoon et al. employed RNN networks for both audio and text, while Atmaja et al. proposed LSTM and dense network architectures for emotion classification [1], [12].

This study proposes EmoTech, a multi-modal architecture for speech emotion recognition that combines both audio and text features at a low-feature level. In the proposed model, separate processing blocks are employed for audio and text inputs, which are subsequently concatenated and fed into a classification block to predict the emotion of the speech. The key contributions of this study include:

1) A multimodal architecture that effectively integrates audio and text modalities for emotion recognition.
2) The use of BiLSTM and CNN layers in both audio and text blocks to capture temporal dependencies and local features.
3) A comprehensive evaluation of the proposed model on standard speech emotion datasets, demonstrating its effectiveness compared to existing methods.

The paper is organized as follows: Section 2 describes the individual parameters of the proposed network used in this research. Section 3 presents the experimental results, followed by the conclusion in Section 4.

II. Methodology

This section presents the overall architecture of the proposed EmoTech for SER. It also describes the hyperparameter tuning of the network, along with the training and testing procedures, and the relevant dataset used for this study.

A. Dataset

This study utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [2] dataset to train and evaluate the proposed model. The IEMOCAP dataset, a widely used benchmark in speech emotion recognition research, comprises scripted and spontaneous acts of ten emotions. As shown in

Fig. 1, the dataset is unbalanced across these ten emotion categories, with some emotions having significantly fewer samples. To address this imbalance, we focused on the five dominant emotion categories: anger, sadness, happiness, neutrality, and excitement.

Both audio recordings and text transcriptions from the dataset were used. To increase the number of samples in the less dominant classes and achieve a more balanced dataset, various data augmentation techniques were applied. For audio data, techniques such as time stretching, pitch shifting, noise addition, time shifting, and volume adjustment were used. For text data, augmentation techniques included synonym replacement, random insertion, random deletion, and random swapping. The statistics of the total 5,633 data samples after augmentation are represented in Fig. 2.

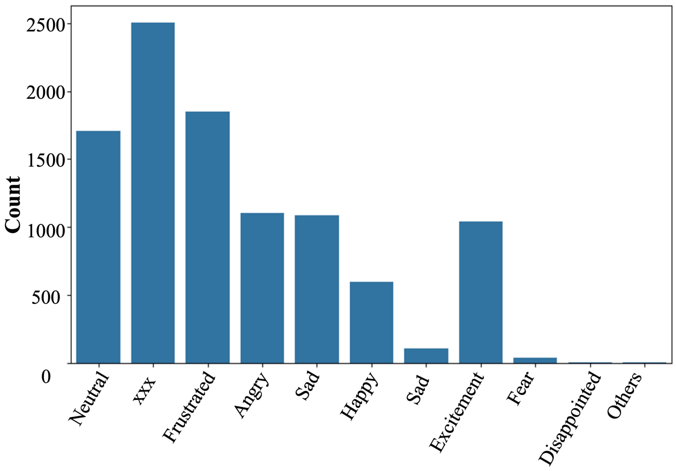

Fig. 1. The statistics of ten classes in IEMOCAP Dataset before augmentation

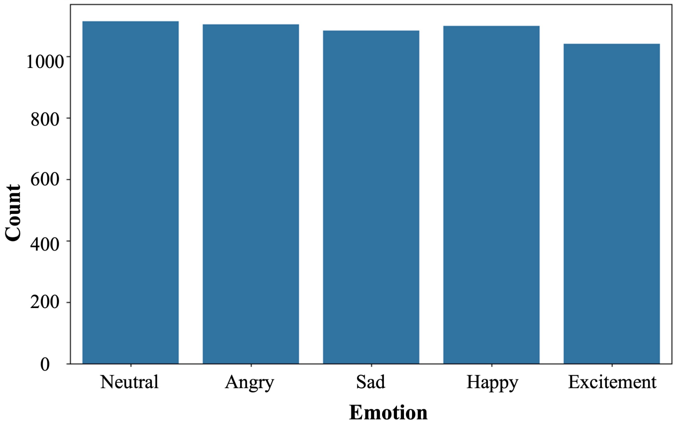

Fig. 2. The statistics of five dominated classes in IEMOCAP Dataset after augmentation

B. Network Architecture

Figure 3 illustrates the basic block diagram of the proposed EmoTech, designed for SER. The network comprises three main blocks: 1) Audio Block, 2) Text Block, and 3) Classification Block. The input audio data is first transformed into Mel Frequency Cepstral Coefficients (MFCCs) and processed through a recurrent network to capture temporal dependencies and a 2D convolutional network to extract spatial features. In parallel, the text data is converted into embeddings and processed through a recurrent network with global max pooling to capture sequential information. Additionally, the embeddings pass through a 1D convolutional network with global max pooling to convert them into low-dimensional feature maps. The outputs from the audio and text blocks are then concatenated, passed through several dense layers with decreasing units, and finally classified into one of the emotion categories: anger, sad, happy, excited, or neutral.

1) Audio Block: The detailed block diagram of the audio block is presented in Fig. 4. In EmoTech, MFCCs are utilized as the input for the audio block. Initially, sentence-level utterances are resampled to a sampling rate of 16 kHz, and silenced parts are removed using a threshold value of 20 dB. The number of cepstral coefficients used is 13, with the FFT window length set to 2048 (12.8 ms) and the hop length to 512. The 'Hann' window function is applied, and zero padding ensures that all MFCCs have a consistent shape. Consequently, the final shape of the MFCCs for each utterance is (740, 13), where 740 represents the number of time steps and 13 represents the number of cepstral coefficients.

The MFCCs with a shape of (740, 13) are then passed as input to the BiLSTM network. The BiLSTM network comprises two layers, each with 64 hidden units and employing the 'tanh' activation function. The output of the BiLSTM network results in a shape of (128,).

Simultaneously, the same MFCCs are used as input for the Conv2D network. This network consists of three Conv2D layers, each followed by batch normalization and a MaxPooling layer. The convolution layers have 32, 64, and 128 kernels, respectively, with each kernel having a shape of (3, 3). Each MaxPooling layer utilizes a window shape of (2, 2).

Following the flatten operation in the Conv2D network, a tensor with a shape of (11776,) is generated and used as input for the dense network within the audio block. This dense network's output has a shape of (128,). The dense network comprises two layers, with the first dense layer having 512 hidden units and the second dense layer having 128 hidden units. A dropout layer with a dropout rate of 0.2 is placed between the two dense layers.

Finally, the outputs of the BiLSTM network (with an output shape of (128,)) and the dense network (also with an output shape of (128,)) are concatenated. This concatenation results in the final output of the audio block having a shape of (256,).

2) Text Block: As shown in Fig. 5, the corresponding transcriptions of the audio signals are utilized as the text input in this study. Each text sentence is tokenized and zero-padded to ensure uniform length. These tokenized sentences are then processed through an embedding layer. The maximum length of a sentence is set to 98 words, and the vocabulary consists of 2843 words. The integer-encoded sentences, having a shape of

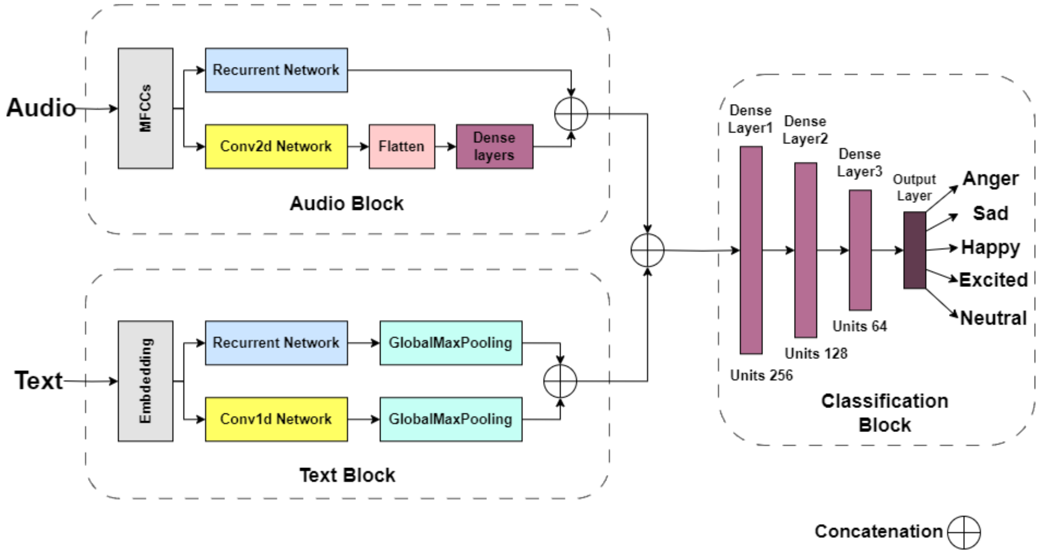

Fig. 3. Basic block diagram of the proposed EmoTech Architecture

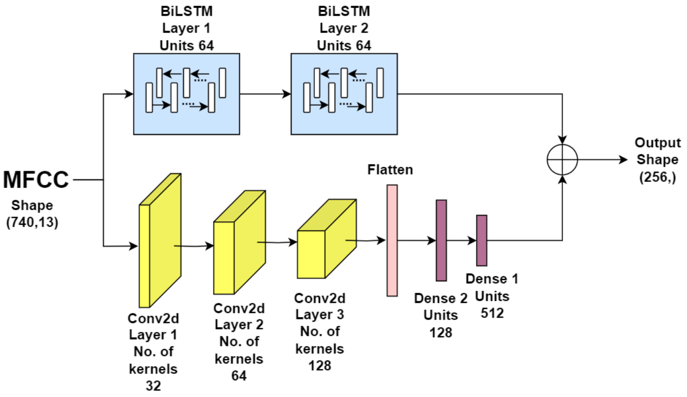

Fig. 4. A detail block diagram of audio signal processing in EmoTech

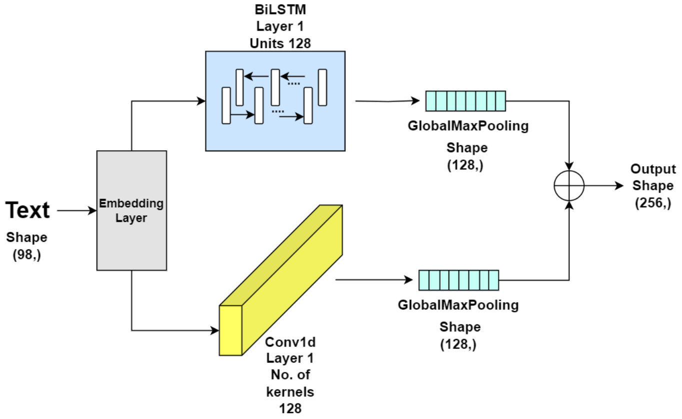

Fig. 5. A detail block diagram of text data processing in EmoTech

(98,), are used as input for the embedding layer. The number of embedding dimensions is 200, resulting in an output shape of (98, 200).

The output of the embedding layer is then used as input for two parallel networks: a BiLSTM network and a Conv1D network. The BiLSTM network is configured with 64 hidden units. The output of the BiLSTM layer is passed into a GlobalMaxPooling layer, resulting in an output shape of (128,). Simultaneously, the output of the embedding layer is used as input for the Conv1D network, which uses 128 kernels with a kernel size of 5, and applies the 'ReLU' activation function for each convolution layer. Following the Conv1D layer, a GlobalMaxPooling layer is utilized, also resulting in an output shape of (128,).

The final output of the text block is achieved by concatenating the outputs of both the Conv1D and BiLSTM networks, resulting in a shape of (256,).

3) Classification Block: This section of EmoTech combines the low-level features from the audio and text blocks, as shown in Fig. 6. The classification block consists of three dense layers with unit sizes of 256, 128, and 64, respectively. A dropout layer with a rate of 0.2 follows the first two dense layers to prevent overfitting. The final output layer is configured with 5 units, corresponding to the five emotion classes. Each of the dense layers is activated using the 'ReLU' function, except for the output layer, which uses the 'Softmax' activation function.

C. Hyperparameters and Model Training

Table 1 represents the different hyperparameters used to train the proposed EmoTech network. The proposed model is trained on 5,633 data samples using 5-fold cross-validation to ensure robust performance evaluation and prevent overfitting.

A total of 30 epochs is specified, with a batch size set to 32. The optimization process is carried out using the

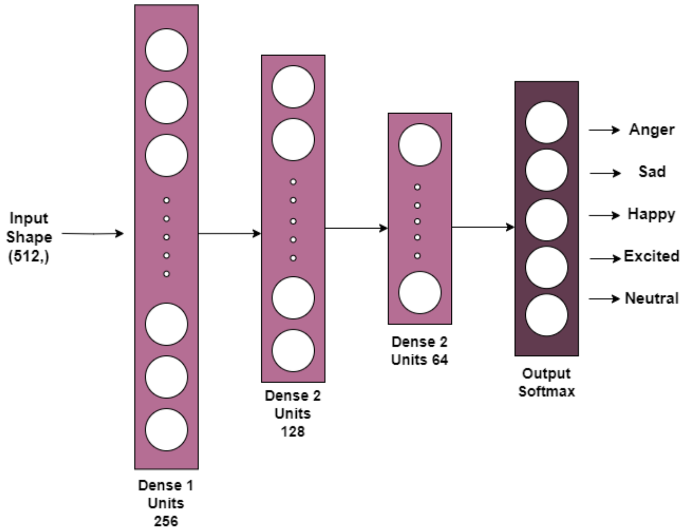

Fig. 6. A detail block diagram of classification processing in EmoTech

Adam optimizer. The model's loss is calculated through the Categorical Cross Entropy function. An initial learning rate of 0.001 is defined, while the minimum learning rate is adjusted to 0.000001 during training.

To further optimize training, EarlyStopping is employed to halt training when the validation loss ceases to improve, and ReduceLROnPlateau is used to reduce the learning rate when the validation performance plateaus, thereby enhancing the model's convergence efficiency.

The total number of parameters in the model is 7,295,821, indicating a complex architecture designed to capture intricate patterns in the data. The training process is conducted on Google Colab, utilizing a T4 GPU as the accelerator to expedite computations and leverage the parallel processing capabilities of the GPU. This setup allows for efficient handling of the model's extensive parameter space and accelerates the training process significantly.

III. Results and Discussions

This section presents the simulated results of the proposed EmoTech algorithm, evaluated through objective metrics including precision, recall, and F1-score. The performance of EmoTech is compared with three existing algorithms to provide a benchmark. Additionally, the impact of different emotions on the accuracy of the algorithm is analyzed and discussed.

A. Effect of data augmentation on EmoTech

Table I presents the impact of data augmentation on minority classes and feature modality in terms of classification accuracy before and after augmentation. The classification accuracy is shown for different modalities. Generally, the overall accuracy of the model is higher when EmoTech uses combined speech and text for classification rather than a single feature modality.

TABLE I
The impact of data augmentation on accuracy

Feature	Augmentation	Accuracy
Speech	No	0.7022
Speech	Yes	0.7184
Text	No	0.7133
Text	Yes	0.7423
Speech + Text	No	0.8105
Speech + Text	Yes	0.8352

TABLE II
Different classification metrics for individual classes

Class	Precision	Recall	F1-score	Accuracy
Anger	0.9641	0.9729	0.9685	0.9728
Excited	0.9083	0.9252	0.9167	0.9252
Happy	0.9224	0.9456	0.9339	0.9456
Neutral	0.8660	0.8153	0.8399	0.8153
Sad	0.9654	0.9696	0.9675	0.9695

When combining time-domain audio features and text-embedding features, EmoTech demonstrated the best performance, regardless of augmentation. Moreover, the classification accuracy is higher after augmentation than before augmentation, irrespective of the feature modality used. Therefore, the combination of augmented speech and text is utilized as features for the proposed network and for further evaluation.

B. Performance evaluation for different classes

Table II presents the accuracy of individual emotion classes for SER. The accuracy metrics are provided for five emotion classes: anger, sadness, happiness, neutrality (denoted as ”neu”), and excitement. These metrics include precision, recall, F1-score, and overall accuracy.

In general, the accuracy is notably high for the ”angry” and ”sad” emotions, demonstrating robust performance in both precision and recall. Conversely, the ”neu” (neutrality) emotion tends to have lower accuracy compared to the others. Specifically, the highest accuracy score observed is 97.28% for the ”angry” emotion.

The confusion matrix for individual emotions is shown in Fig. 7. The results indicate that the neutral emotion is frequently misclassified, particularly confused with excitement and happiness.

C. Comparison with existing networks

To analyze the performance of the proposed network over existing methods, three different models are evaluated with the IEMOCAP dataset. The results are evaluated in terms of

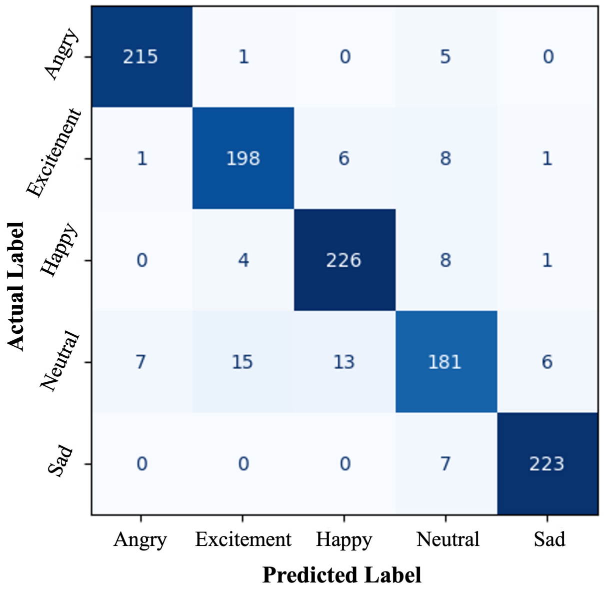

Fig. 7. Confusion Matrix in classifying five emotion using EmoTech architecture.

TABLE III
COMPARISON OF DIFFERENT ALGORITHMS IN SER

Model	Feature	Accuracy(%)
Yoon [12]	Speech+Text	71.80
Yenigalla [11]	Speech+Phoneme	73.90
Atmaja [1]	Speech+Text	75.40
EmoTech	**Speech+Text**	**83.52**

overall accuracy and presented in Table III. In general, the accuracy is high when a hybrid model is used to capture the detailed information from speech and text. The result shows that the score is high when emotions are classified using the proposed EmoTech network.

IV. Conclusion

This study introduces a multi-modal architecture for SER that effectively integrates audio and text features to enhance classification performance. By incorporating BiLSTM and CNN layers within dedicated audio and text blocks, the model successfully captures both temporal dependencies and local features, thereby improving its ability to detect nuanced emotional expressions. Evaluation on the IEMOCAP dataset, a widely accepted benchmark in emotion recognition, illustrates that the proposed model surpasses traditional single-modal approaches and achieves competitive accuracy comparable to state-of-the-art methods. Future research directions include extending the modalities to include audio, text, and video inputs, thereby broadening the scope of emotion recognition applications.

References

[1] Bagus Tris Atmaja, Kiyoaki Shirai, and Masato Akagi. Speech emotion recognition using speech feature and word embedding. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 519–523. IEEE, 2019.

[2] 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*, 42:335–359, 2008.

[3] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Deep neural networks for emotion recognition combining audio and transcripts. In *Interspeech*, pages 247–251, 2018.

[4] David Griol, José Manuel Molina, and Zoraida Callejas. Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. *Neurocomputing*, 326:132–140, 2019.

[5] Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. Speech emotion recognition with acoustic and lexical features. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4749–4753. IEEE, 2015.

[6] Chul Min Lee, Shrikanth S Narayanan, and Roberto Pieraccini. Combining acoustic and language information for emotion recognition. In *INTERSPEECH*, pages 873–876. Citeseer, 2002.

[7] Samaneh Madanian, David Parry, Olayinka Adeleye, Christian Poellabauer, Farhaan Mirza, Shilpa Mathew, and Sandy Schneider. Automatic speech emotion recognition using machine learning: digital transformation of mental health. In *Proceedings of the Annual Pacific Asia Conference on Information Systems (PACIS)*, 2022.

[8] Valery Petrushin. Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering*, volume 710, page 22, 1999.

[9] Seme Sarker, Khadija Akter, and Nursadul Mamun. A text independent speech emotion recognition based on convolutional neural network. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–4. IEEE, 2023.

[10] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.

[11] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*, volume 2018, pages 3688–3692, 2018.

[12] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE, 2018.

[13] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.