NCSP'25 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2025
February 27 - March 2, 2025, Pulau Pinang, Malaysia

Exploring the EmoBone Dataset with Bi-Directional LSTM for Emotion Recognition via Bone Conducted Speech

Md. Sarwar Hosain[1], Md. Rifat Hossen[1], Md. Uzzal Mia[1],
Yosuke Sugiura[2], and Tetsuya Shimamura[2]

[1]Pabna University of Science and Technology
Pabna, Bangladesh
E-mail: sarwar.ice@pust.ac.bd

[2]Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama 338-8570, Japan

Abstract

This research explores the application of deep learning techniques for emotion recognition using bone-conducted (BC) speech, utilizing the EmoBone dataset—the first dataset specifically created for this purpose. The study aims to investigate the effectiveness of deep learning approaches in recognizing emotions from BC speech while addressing challenges such as degradation and information loss in neural networks. By implementing advanced models, the research compares the accuracy of emotion recognition with other relevant techniques, showcasing the advantages of BC speech in terms of noise resilience and privacy. The findings contribute to the development of more accurate and reliable emotion recognition systems, creating the way for innovative applications in fields such as healthcare, human-computer interaction, and secure communication systems.

1. Introduction

In recent years, the development of human-computer interaction (HCI) systems has increasingly focused on making interactions more natural and responsive. A key aspect of achieving this goal is the ability to accurately recognize human emotions from speech. By identifying human emotions from voice signals, speech emotion recognition (SER) finds applications in diverse areas such as robotics, mobile services, contact centers, gaming, and psychological assessments [1]. A lot of research has been done on air-conducted (AC) speech in the SER system, but bone-conducted (BC) speech has its benefits, especially in noisy locations and when privacy is important. BC speech propagates auditory signals directly to the cochlea using bone conduction of the skull. Despite its potential, research on emotion recognition from BC speech remains largely unexplored [4]. The research aims to explore the potential of deep learning in recognizing emotions from BC speech and address the unique challenges associated with this task. Additionally, it seeks to compare the accuracy of emotion recognition achieved by deep learning models with other relevant approaches. Ultimately, this research aspires to contribute to the development of more accurate and reliable emotion recognition systems.

The noise-robust feature of BC speech, which captures the body's vibrations and makes it less susceptible to outside noise, is what drives its application in SER. This property of BC speech not only enhances the low-frequency components but also improves the accuracy of speech analysis [2]. In [2], although SER takes AC speech as input, there have been promising results in converting AC speech to equivalent BC speech using an infinite impulse response (IIR) filter. The expectation is that BC speech provides more valuable information for SER due to its enhanced low-frequency components, which are essential for recognizing emotions. They got 72.50% accuracy for recognizing BC speech emotion using convolutional neural networks (CNNs) with synthetic data. The authors of [3] have also emphasized the improved accuracy rates for speaker recognition when AC and BC speech are used together under noiseless conditions. Their findings indicate that the error rates for BC and AC speech are comparable, which suggests that BC speech can serve as a compensatory feature, potentially leading to better accuracy in SER models.

The rest of the paper is structured as follows: Section 2 provides a detailed description of the study protocol and evaluation methodology. The comparative findings and various statistical analyses are presented in Section 3. This section also divides into the corresponding discussion and limitations. The final conclusions are drawn in Section 4.

2. Methodology and structure

This section will describe the data collection meth-

ods, data analysis procedures, and data validations employed in this study. The LSTM is a variant of the RNN that consists of memory blocks connected and capable of maintaining temporal states through self-connections. It has three gate units: input, output, and forget gates. LSTM solves long-term dependence issues in RNNs and implements refined internal processing units for better storage and updating context information. It also overcomes gradient vanishing or explosion problems in standard RNNs. The BiLSTM model utilizes contextual information in both the forward and backward directions, enhancing its robustness by detecting hidden emotions through directional analysis [5]

The initial model architecture developed in this study was a BiLSTM network. This model was built to take advantage of the sequential structure of voice data by capturing both forward and backward temporal dependencies. The architecture consists of several layers that hierarchically process the input data. The input layer receives the MFCC feature vectors, which serve as the model's raw data input. This layer is followed by a stack of two bidirectional LSTM layers, each with 128 hidden units. These LSTM layers are responsible for processing the input sequences and generating a sequence of hidden states that capture the temporal dependencies within the data. The LSTM layers in the model can include both past and future context when processing each time step due to their bidirectional nature, enhancing its ability to capture relevant The final output of the LSTM layer is sent into a fully linked layer, which transforms the LSTM's output into the seven emotion classes. This mapping is achieved through a linear transformation, followed by a Softmax activation function. The Softmax function converts the raw scores into probabilities, providing a measure of the model's confidence in each emotion class. The choice of a Softmax activation function is motivated by its ability to normalize the output probabilities, ensuring that they sum to one. This property is essential for multi-class classification tasks, as it allows the model to express uncertainty in its predictions. The model was implemented using the PyTorch framework, which offers a versatile and effective framework for constructing and instructing deep learning models. To optimize the parameters of the model, we employed the Adam optimizer, a modified version of stochastic gradient descent that adjusts the learning rate for each parameter using estimates of its first and second moments. The learning rate was set to 0.001, a commonly used value that balances the need for quick convergence with the risk of overshooting the optimal solution. The model performed 100 epochs of training, utilizing a batch size of 8, to achieve sufficient exposure to the training data while also retaining computational efficiency.

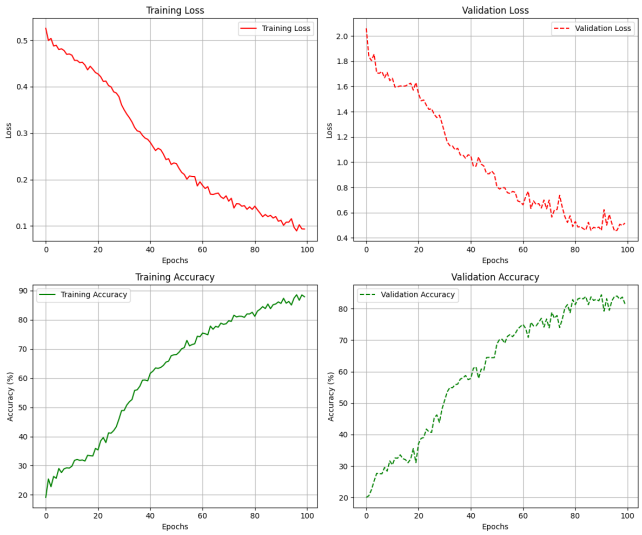

Figure 1: Training and Validation Loss and Accuracy over 100 epochs.

The BiLSTM model was trained using the EmoBone dataset. In order to train the model, we minimized the cross-entropy loss, which is the difference between the actual emotion labels and the predicted probabilities. To avoid overfitting, we checked the accuracy of the model on a validation set while it was being trained. We used an early stopping mechanism to cease training after three consecutive epochs if the validation loss failed to improve. This method is useful for preventing the model from becoming "overfit" to its training data, thereby improving its generalization performance. If the validation loss did not improve, a learning rate scheduler was employed to lower the learning rate, so assisting the model in achieving a more optimal minimum. The performance of the model was evaluated using classification accuracy on a held-out test set. The BiLSTM model achieved an accuracy of 85%, indicating its ability to capture temporal dependencies in the speech data. To further analyze the models' performance, we computed confusion matrices and classification reports for both models. The confusion matrix for the BiLSTM model showed a significant lower number in misclassification. The classification report highlighted the improvement in precision, recall, and F1-score for most of the emotion classes, indicating a more balanced and robust performance.

3. Experimental results

The EmoBone dataset was utilized to train and evaluate the BiLSTM models. The dataset's distribution across the seven emotion classes (neutral, calm, happy, sad, angry, fear, and surprise) is visualized in Figure

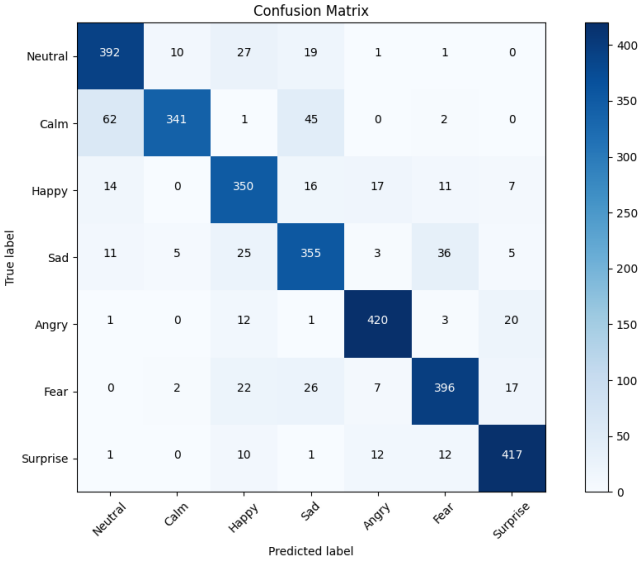

Figure 2: Confusion matrix using BLSTM

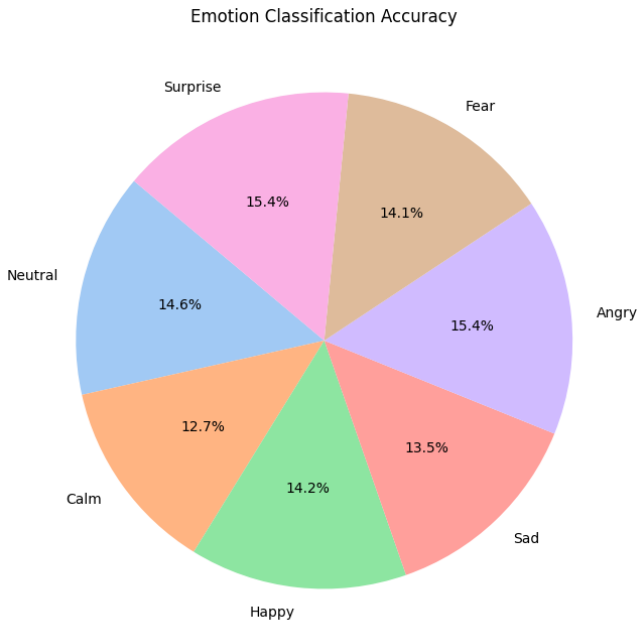

Figure 3: Emotion wise pie chart

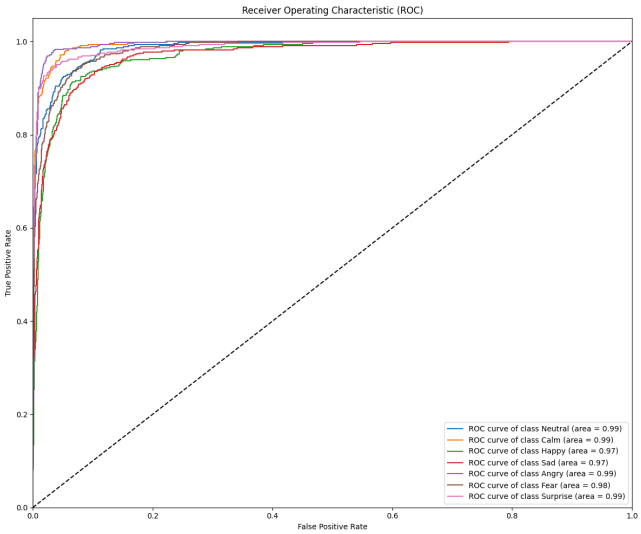

Figure 4: Receiver operating curve

5.2. The figure illustrates a relatively balanced distribution, with each emotion class represented by a substantial number of samples. This balance is crucial for preventing the model from being biased towards any particular emotion class during training.

The performance of the BiLSTM model is depicted in Figure 1. The figure presents the training and validation loss and accuracy curves over 100 epochs. The loss curves demonstrate a steady decrease in both training and validation loss, indicating that the model is learning to fit the training data effectively. The accuracy curves show a corresponding increase in both training and validation accuracy, further confirming the model's learning progress. The intersection of the training and validation curves implies that the model is not overfitting, which is a positive sign for its generalization capabilities. The confusion matrix for the BiLSTM model is presented in Figure 2. The matrix provides a thorough analysis of the approach estimates, highlighting its advantages and disadvantages in categorizing various emotions. The diagonal elements of the matrix represent the number of correct predictions for each emotion class, while the off-diagonal elements indicate misclassifications. The matrix shows that the model performs well for most emotion classes, with high accuracy for emotions like angry and surprise. However, there is some confusion between similar emotions like calm and neutral, suggesting room for improvement.

The accuracy of each emotion class for the BiLSTM model is visualized in Figure 3 using a pie chart. The chart provides a clear visual representation of the model's performance across different emotions. The chart shows that the model achieves high accuracy for

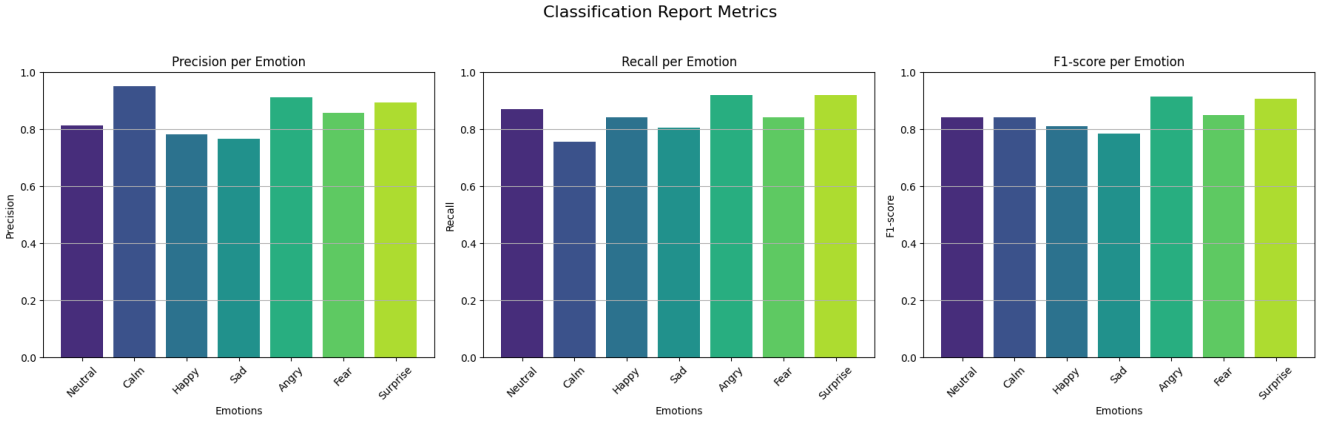

Figure 5: Classification accuracy diagram spanning both columns.

most emotions, with accuracy exceeding 80% for several classes. However, the accuracy for Calm is relatively lower, indicating a potential area for improvement. Figure 4 presents the ROC curves for the BiLSTM model. ROC curves graphically represent the relationship between the true positive rate and the false positive rate across various classification levels. The curves illustrate the ability of the model to differentiate across several emotion categories. The area under the curve (AUC) for each emotion class is also provided, serving as a quantitative measure of the model's performance. The AUC values are generally high, indicating good discriminative power for most emotions.

The overall classification accuracy for the BiLSTM model without attention is presented in Figure 5. The figure shows that the model achieves an accuracy of 85.17%, which is a respectable performance considering the complexity of the task. This result serves as a baseline for comparison with the Bi-LSTM model with attention.

4. Conclusions

This study highlights the feasibility of using BiLSTM networks for emotion recognition via bone-conducted speech, providing a baseline performance for the EmoBone dataset. While the results are promising, the findings suggest that further enhancements, such as incorporating advanced preprocessing techniques or combining BiLSTM with other neural network architectures, could improve performance. This work lays the groundwork for future research in BC speech emotion recognition and underscores the significance of exploring novel datasets like EmoBone to advance human-computer interaction. Future research could explore the integration of other advanced techniques, such as transfer learning and multi-modal fusion, to further enhance the performance and robustness of speech emotion recognition systems.

References

[1] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," Information Fusion, vol. 102, p. 102019, Feb. 2024.

[2] M. S. Hosain, Y. Sugiura, N. Yasui, and T. Shimamura, "Deep-learning-based speech emotion recognition using synthetic bone-conducted speech", Journal of Signal Processing, vol. 27, no. 6, pp. 151-163, 2023.

[3] S. Tsuge, N. D. Koizumi, M. Fukumi, and S. Kuroiwa, "Speaker verification method using bone-conduction and air-conduction speech," Dec. 2009.

[4] M. S. Hosain, Y. Sugiura, M. S. Rahman and T. Shimamura: EmoBone: A Multinational Audio Dataset of Emotional Bone Conducted Speech, IEEJ Trans. Electrical and Electronic Engineering, Publisher: IEEJ and Wiley, vol. 19, no. 9, 2024.

[5] S. Sultana, M. Z. Iqbal, M. R. Selim, Md. M. Rashid, and M. S. Rahman, Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks, IEEE Access, vol. 10, pp. 564–578, Jan. 2022.