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

Md. Sarwar Hosain, Md. Rifat Hossen, Md. Uzzal Mia,
Yosuke Sugiura and Tetsuya Shimamura

Presented by

Md. Rifat Hossen
Pabna University of Science and Technology

Presentation Outline

- Motivation
- Research Question
- Objectives
- Introduction
- Dataset Preparation
- Dataset Evaluation
- Reliability of Evaluation
- Methodology
- Results
- Discussion
- Conclusions

Motivation

- Address **speech emotion recognition** issues in deep neural networks including degradation and information loss

- Utilize **bi-directional LSTM (BiLSTM)** to capture the temporal dynamics of emotional expression

- Provide a **novel bone-conducted** speech emotion identification system

Research Question

✓ **How can a BiLSTM network enhance its performance in emotion recognition using bone-conducted speech?**

Objective

- To develop and optimize a BiLSTM based model
- To evaluate and compare the performance of the model
- To attain **state-of-the-art EmoBone dataset** performance

Introduction

❖Speech:

- Speech is a primary mode of human communication
- It conveys not just information but also emotions
- Accurately recognizing emotions in speech is crucial for various applications

Figure 1: Speech delivery

Figure 2: Speech recognition

❖Speech Recognition:

- Computers can recognize our words and turn them into text.
- Speech recognition analyzes features like pitch, sound waves, and pronunciation
- Voice assistants (e.g., Siri, Alexa, Google Assistant)

Introduction(Cont.)

Figure 3: Speech emotion recognition

❖Speech Emotion Recognition (SER):

- SER aims to automatically recognize emotions from spoken language.
- SER analyzes features like tone, pitch, and energy to identify emotions
- Emotion recognition is important for understanding human behavior
- It can be used to improve social interactions, human-computer interaction (HCI), and affective computing
- SER has applications in diverse areas, such as call centers, in-vehicle services, and medical services etc.

Introduction(Cont.)

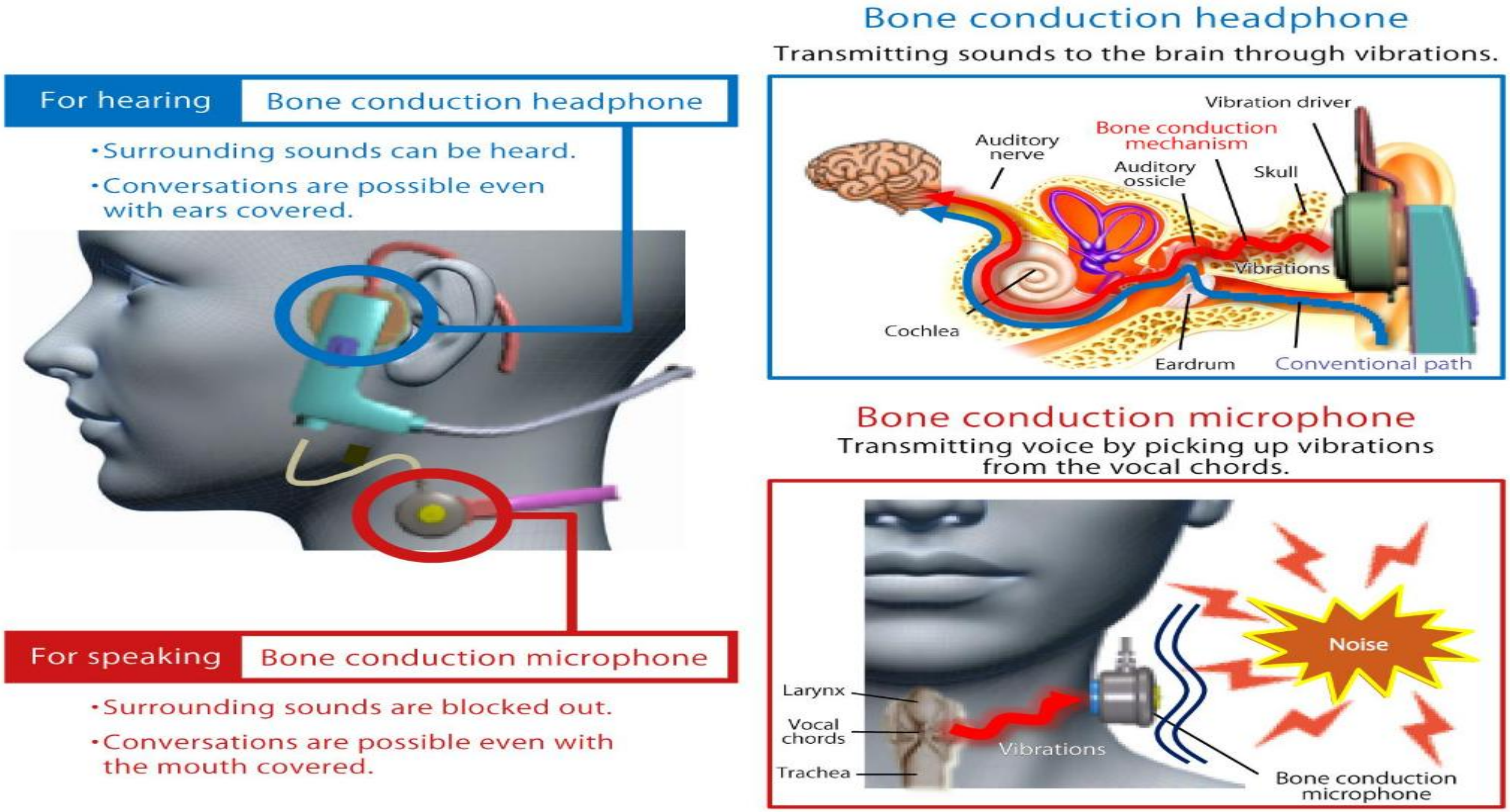


Figure 4: Bone conduction technology

Introduction(Cont.)

Probable application areas of BC speech emotion recognition system

❖ **Human Computer Interaction :**
- Smartphones and wearables devices
- Virtual assistants and chatbots
- Biometric authentication

❖ **Healthcare and Mental Health**:
- Mental health monitoring
- Telehealth and remote monitoring
- Speech therapy and language learning

❖ **Education and Learning**:
- Personalized learning environments
- Educational games and applications
- Sports training and performance analysis

❖ **Security and Law Enforcement**:
- Stress detection in high-pressure situations
- Passenger screening at airports or borders
- Lie detection and deception analysis

❖ **Customer Service and Marketing**:
- Empathy detection in call centers
- Targeted marketing and advertising

Dataset Preparation

Figure 5: Emotion categories

Emotion Categories

- ✓ Happy
- ✓ Angry
- ✓ Sad
- ✓ Calm
- ✓ Disgust
- ✓ Neutral
- ✓ Fear
- ✓ Surprise

Dataset Preparation(Cont.)

Text	Speaker	Training	Recording	Editing	Validation
We chose ten English sentences for the speakers to record	A multi-phase selection process identified 28 speakers from 10 different countries	Reviewed examples of audio and video clips and practiced the assigned sentences	The recording process was spanned multiple sessions, and experts carefully reviewed the recording and discarded any unusable audio clips	All clips were stored in (.wav) format, averaging 4.5 sec and renamed according to the filename convention	A group of 80 raters participated in the validation process and the overall accuracy rate achieved was more than 76%

Figure 6: Flowchart for dataset preparation

Table 1: Sentences used for dataset

No	Sentences
1	We have to cancel our plans for tonight.
2	Argentina won the FIFA World Cup in Qatar.
3	Life is too short to waste time on regrets.
4	It is very cold outside today in Saitama.
5	Do not go outside at night.
6	Students are gossiping in the class.
7	Never underestimate the power of a positive attitude.
8	He loves his family very much.
9	The cat chases the mouse around the house.
10	They are planning to go to Bangladesh.

Dataset Preparation(Cont.)

Table 2: Dataset summary

Parameters	Types/Value
Year of production	2023
Language	English
Dataset type	Acted
File type	Audio only
Sampling rate	48KHz
Speakers number	28
Emotions	7
Sentences	10
Number of audio clips	15680
Average duration	4.5 sec
Software	Ocenaudio
Dataset duration	70560 sec=19 hours 36 min
Validators	80
Recognition rate	76.49%

Table 3: Speaker number and language status by country

Country	Speaker	English language status	Age groups
Japan	3	Officially recognized	30-40
China	2	Officially recognized	25-30
Bangladesh	13	Officially recognized	30-42
Myanmar	3	Officially recognized	25-35
Sri Lanka	2	Officially recognized	30-35
Nigeria	1	Official	30-35
Nepal	1	Officially recognized	30-35
Malaysia	1	Officially recognized	25-30
Afghanistan	1	Officially recognized	25-30
Pakistan	1	Official	30-35

Dataset Evaluation

- It examined how well-untrained listeners identified emotional content in speech
- Resource limits, student workload, and finding female raters were challenges
- Collaboration with a Bangladeshi university provided access to a larger pool of raters
- 80 raters, 40 males and 40 females, assessed two sets of recordings
- The evaluation focused solely on the acoustic information
- 40 audio sets were carefully picked out, each consisting of 392 files
- Participants chose one emotion to match an audio sample

Reliability of Evaluation

- The evaluation process was conducted in a controlled classroom environment
- User login and activation-deactivation sessions were implemented for data security
- Raters were required to register and verify their details
- After registration, raters could access designated audio files
- Audio tracks were shuffled before playback to prevent predictability
- The submit button remained inactive until the user had fully experienced the audio
- Raters were allowed a 15-minute break after each 45-minute session
- The administrator prevented raters from retaking the experiment

Methodology

- **Data Collection and Preprocessing:**
  - Utilized the EmoBone dataset, processed using **Python's torchaudio library**, resampled, and transformed into feature representations using MFCC
- **Model Architecture Design:**
  - **BiLSTM model:** develop with bidirectional LSTM layers and a fully connected layer, and train it for emotion classification using PyTorch with the Adam optimizer over 100 epochs
- **Training Process:**
  - trained using the same dataset, minimizing cross-entropy loss, and adjusting the learning rate to prevent overfitting and improve generalization performance
- **Evaluation Measures:**
  - Assess model performance with accuracy, confusion matrices, and classification reports.

Results

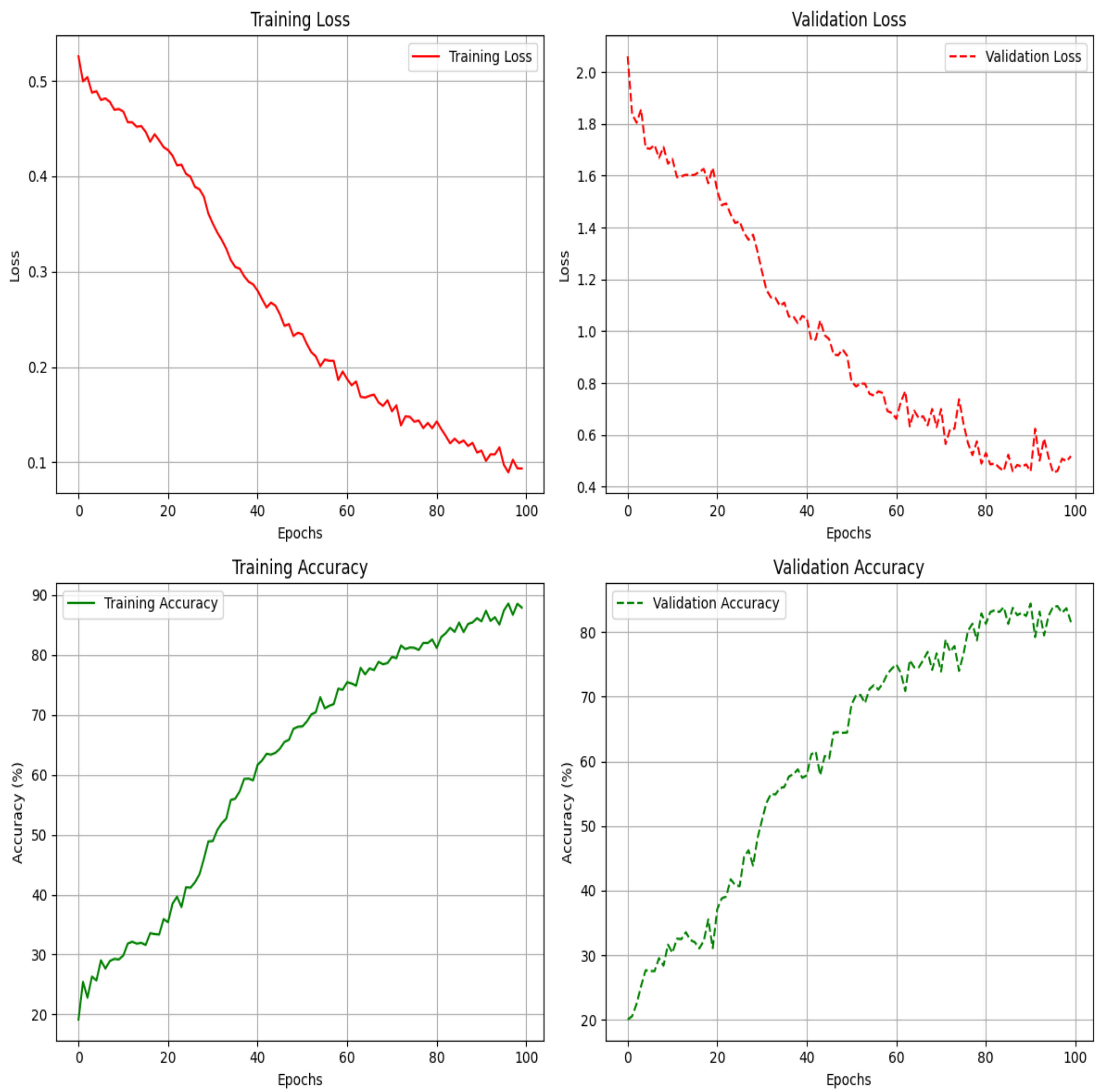


Figure 7: Training and Validation Loss and Accuracy over 100 epochs

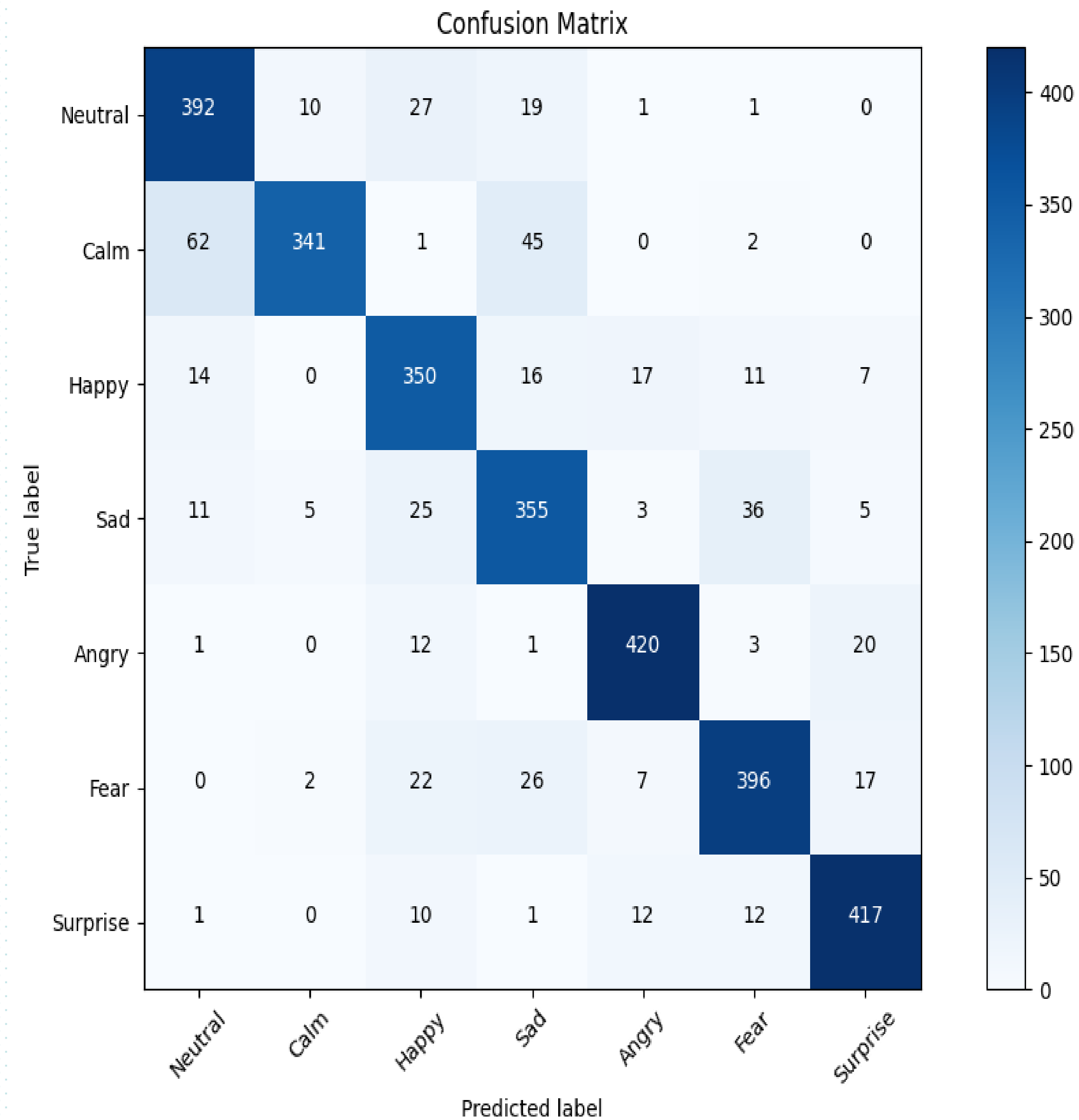
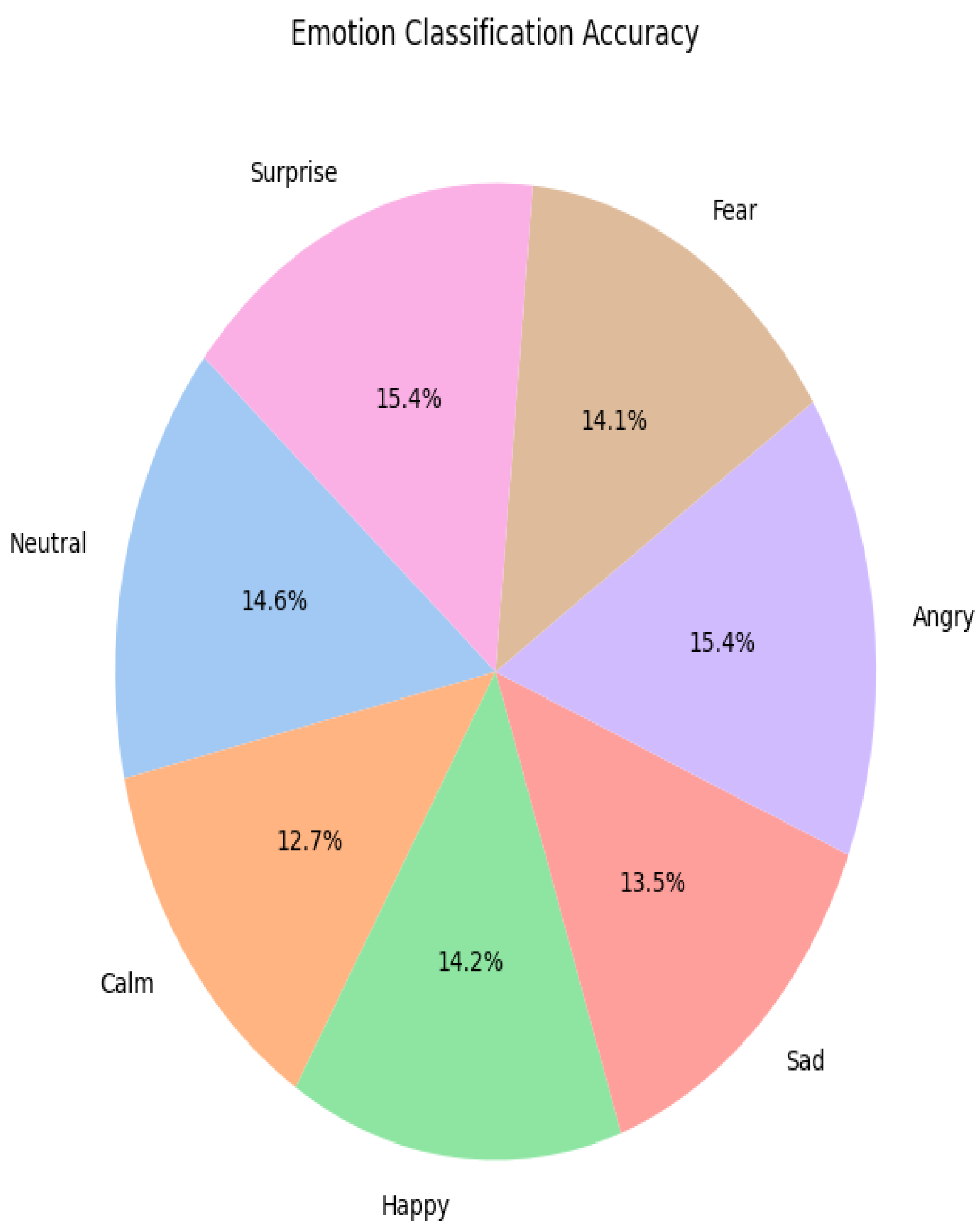


Figure 8: Confusion matrix using BLSTM

Results (Cont.)



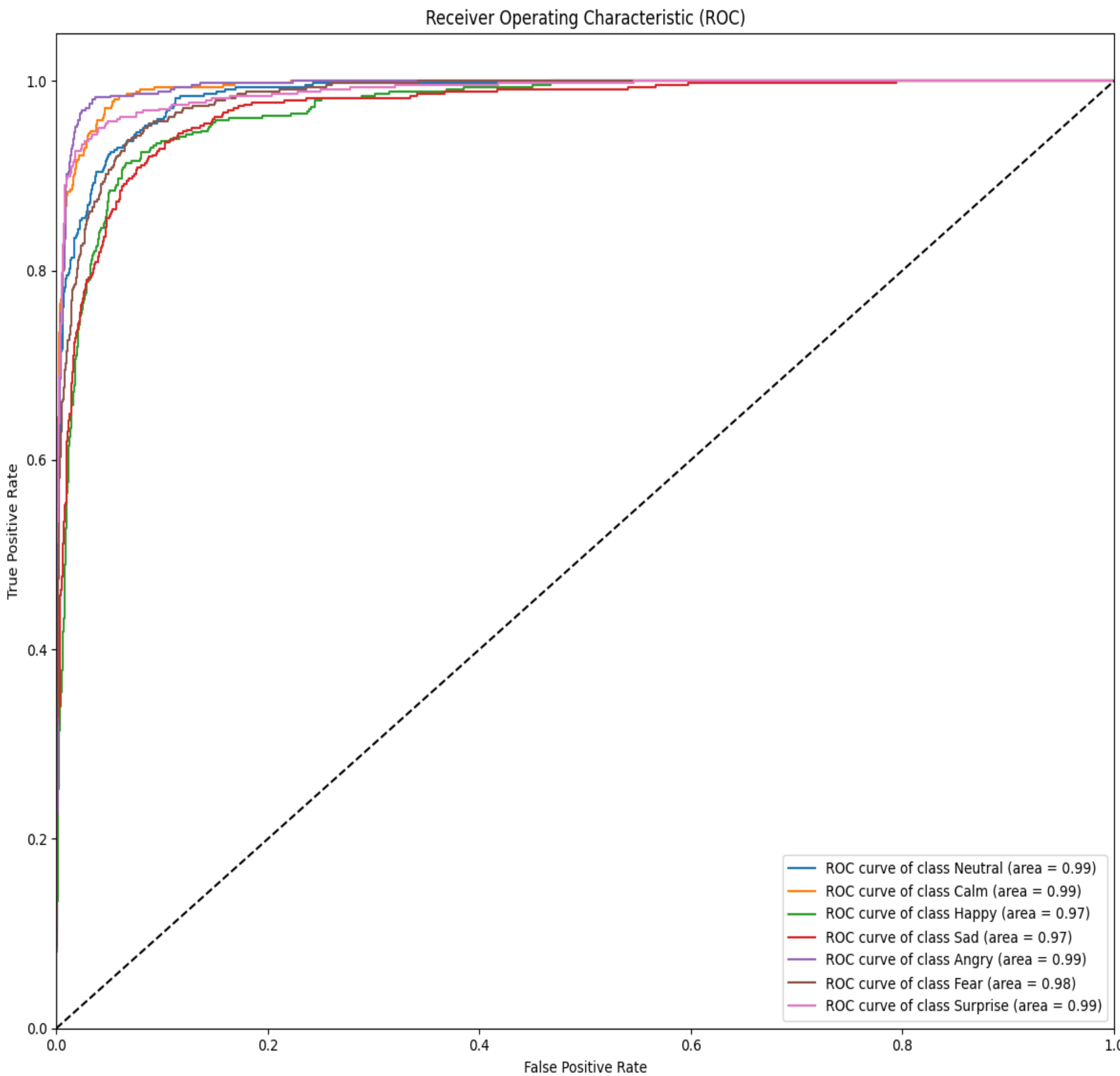


Figure 9: Emotion wise pie chart

Figure 10: Receiver operating curve

Results(Cont.)

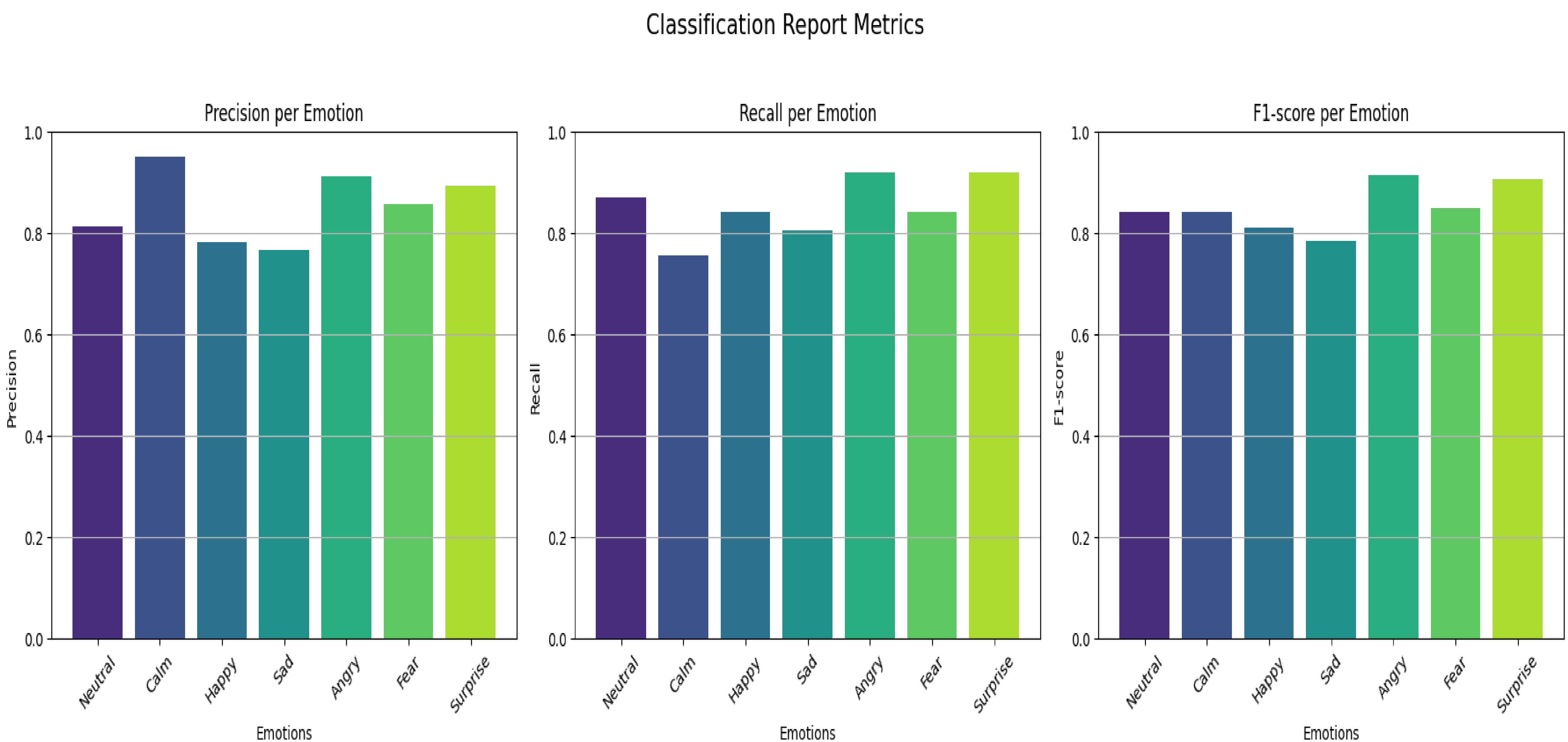


Figure 11: Classification accuracy

Discussion

Table 4: Comparison of accuracy across different studies

Work	Dataset	Accuracy
Proposed model (BiLSTM)	EmoBone	85.17%
Zhao et al. (2019) [1]	IEMOCAP	69%
Mustaqeem and Kwon (2020) [2]	IEMOCAP	81.75%
Mustaqeem and Kwon (2020) [2]	RAVDESS	79.5%
Chen et al. (2018) [3]	Emo-DB	82.82%
Chen et al. (2018) [3]	IEMOCAP	64.74%
Etienne et al. (2018) [4]	IEMOCAP	64.5%
Zhao et al. (2018) [5]	IEMOCAP	68%
Satt et al. (2017) [6]	IEMOCAP	66%
Badshah et al. (2017) [7]	Emo-DB	56%
Hosain et al. (2023) [8]	Synthetic BC speech data	72.50%

Discussion(Cont.)

- **Balanced dataset impact**:
  - ✓ Distribution of seven emotion classes enables unbiased model training and provides a robust foundation for performance evaluation
- **BiLSTM model performance**:
  - ✓ Effectively learns with steady loss reduction and accuracy gains, but struggles to distinguish between similar emotions like calm and neutral
- **Confusion matrix analysis**:
  - ✓ Show that the prominent diagonal elements and improved performance across all emotion classes
- **Overall Accuracy Enhancement:**
  - ✓ Achieves a notable accuracy of 85.17%, surpassing the baseline of 72.50%, highlighting the BLSTM effectiveness in improving emotion recognition from bone-conducted speech

Conclusion

- **Achievements**: **State-of-the-art accuracy of 85.17%** using the **EmoBone dataset**.
- Demonstrated the effectiveness of **BiLSTM techniques**.
- **Significance**: Improved emotion classification for **BC speech**.
- Addressed key challenges such as **information loss and degradation**.
- **Future Scope**: Incorporate transfer learning and multi-modal fusion for further performance improvement.

References

[1] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller,“Attention-enhanced connectionist temporal classification for discrete speech emotion recognition,” in Proc. Interspeech, pp. 206–210, 2019.

[2] N. Mustaqeem and S. Kwon, “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition,” Sensors, vol. 20, no. 1, p. 183, Dec. 2019, doi: 10.3390/s20010183

[3] M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recur rentneural networks with attention model for speech emotion recogni tion,”IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018

[4] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch,“CNN+LSTM architecture for speech emotion recognition with data aug-mentation,” 2018, arXiv:1802.05630.

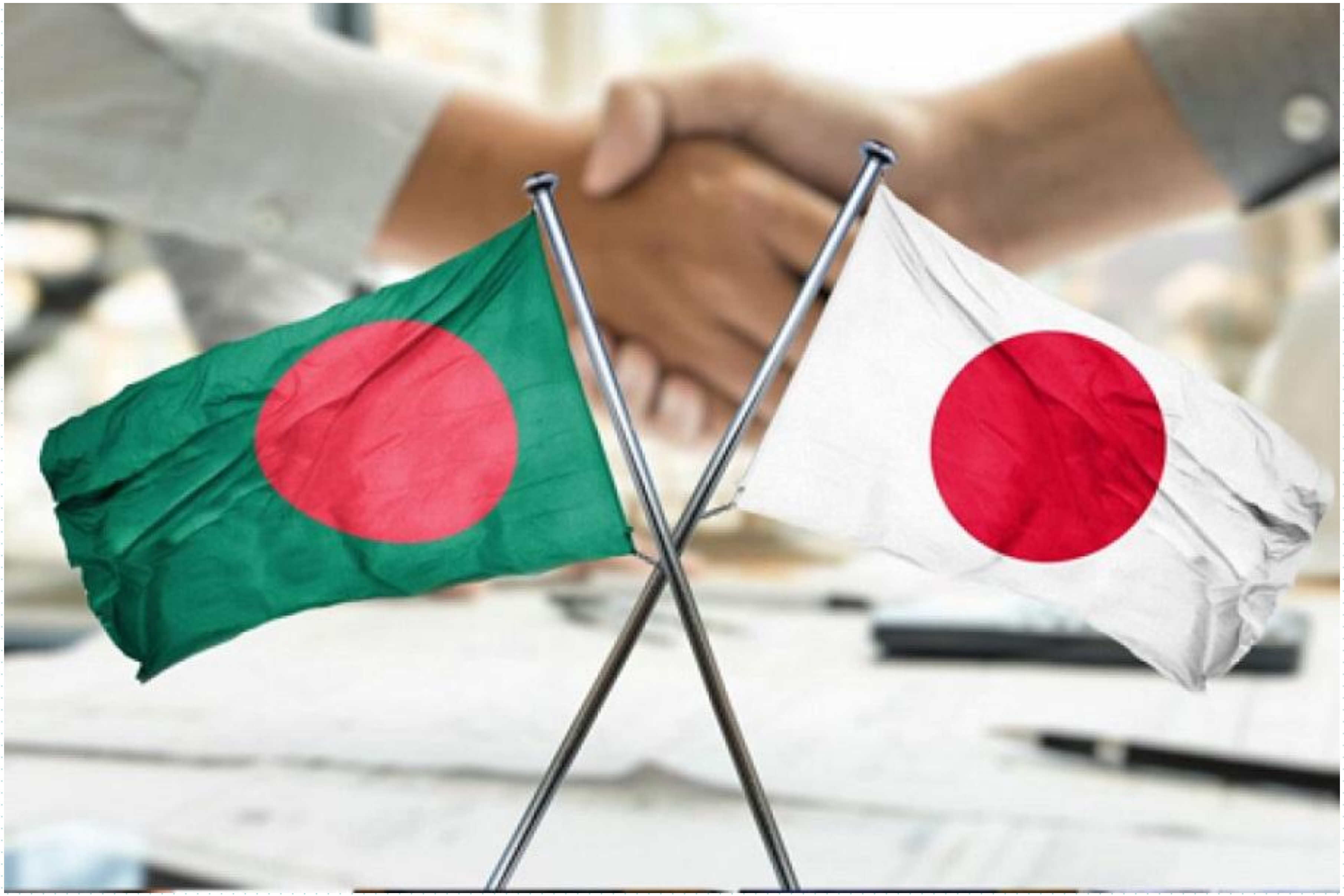
[5] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, “Deep spectrum feature representations for speech emotion recognition,” in Proc. JointWorkshop 4th Workshop Affect. Social Multimedia Comput. 1st Multi-Modal Affect. Comput. Large-Scale Multimedia Data, pp. 27–33, 2018.

[6] . Satt, S. Rozenberg, and R. Hoory, “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms,” Aug. 2017, doi: 10.21437/interspeech.2017-200.

[7] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech Emo tion Recognition from Spectrograms with Deep Convolutional Neural Network,” Feb. 2017, doi: 10.1109/platcon.2017.7883728.

[8] 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, Nov. 2023, doi: 10.2299/jsp.27.151.

Saitama University



Thank You for Your Kind Attention

Please Give Your Valuable Comments and Suggestions