IEEE 10th International Conference on Electrical Engineering and Informatics (ICEEI 2025)

HOSTED BY: Universiti Kebangsaan Malaysia and Institut Teknologi Bandung
13-15 November 2025

Paper ID: 70

Paper Title : Enhancing Robustness and Accuracy of Bone-Conducted Speech Emotion Recognition via Transformer Models

Md. Rifat Hossen
Department of Information and Communication Engineering,
Pabna University of Science and Technology
Pabna, Bangladesh

OUTLINE OF PRESENTATION

- ❑ Motivation
- ❑ Research Question
- ❑ Objectives
- ❑ Introduction
- ❑ Methodology
- ❑ Result and Discussion
- ❑ Conclusion
- ❑ Reference

Motivation

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

❖ Utilize **Transformer Model** to capture the temporal dynamics of emotional expression

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

Research Question

✓ **How can a Transformer Model enhance its performance in emotion recognition using bone-conducted speech for Malaysian speakers?**

Objective

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

Introduction

❖Speech:

- oSpeech is a primary mode of human communication
- oIt conveys not just information but also emotions
- oAccurately recognizing emotions in speech is crucial for various applications

Figure 1: Speech delivery

Figure 2: Speech recognition

❖Speech Recognition:

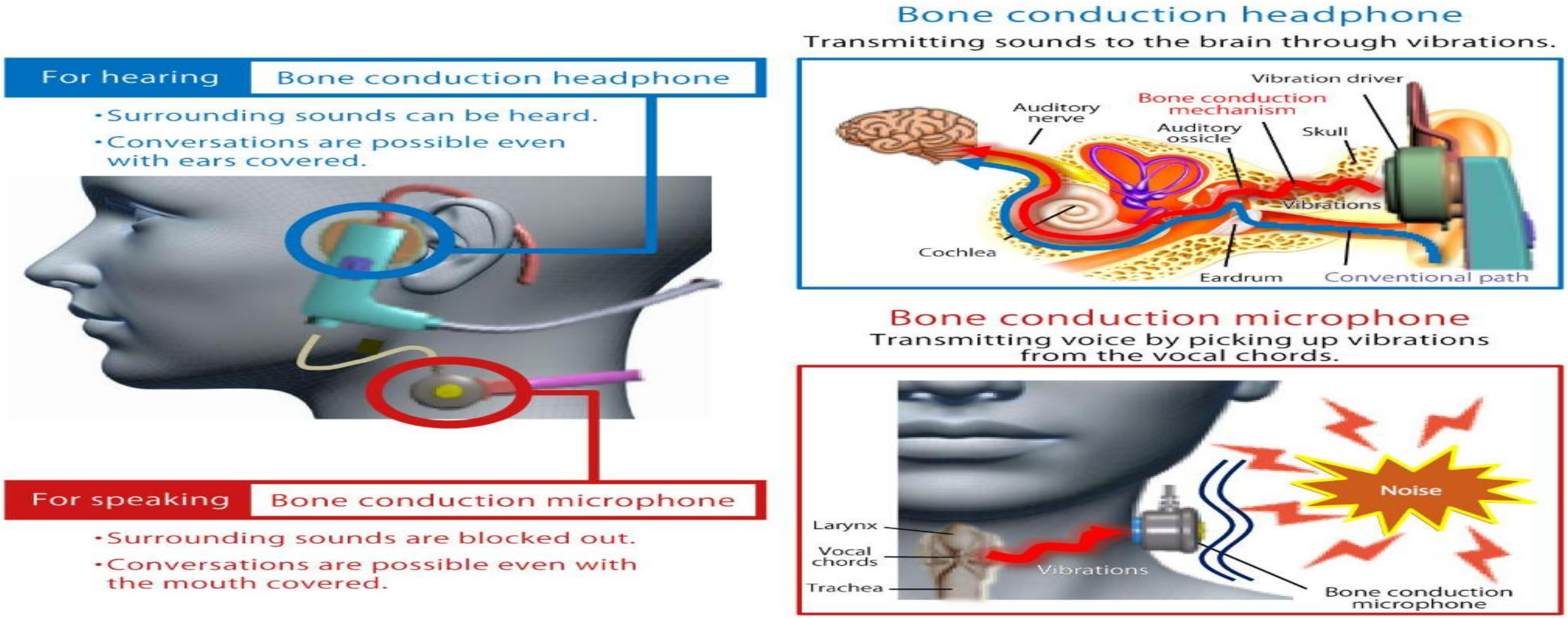
- oComputers can recognize our words and turn them into text.
- oSpeech recognition analyzes features like pitch, sound waves, and pronunciation
- o Voice assistants (e.g., Siri, Alexa, Google Assistant)

Introduction(Cont.)

❖Speech Emotion Recognition (SER):

- SER aims to recognize emotions from spoken language automatically.
- 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.)



For hearing
Bone conduction headphone
•Surrounding sounds can be heard.
•Conversations are possible even with ears covered.
For speaking
Bone conduction microphone
•Surrounding sounds are blocked out.
•Conversations are possible even with the mouth covered.
Bone conduction headphone
Transmitting sounds to the brain through vibrations.
Vibration driver
Auditory nerve
Bone conduction mechanism
Auditory ossicle
Skull
Vibrations
Cochlea
Eardrum
Conventional path
Bone conduction microphone
Transmitting voice by picking up vibrations from the vocal chords.
Noise
Larynx
Vocal chords
Trachea
Vibrations
Bone conduction microphone

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

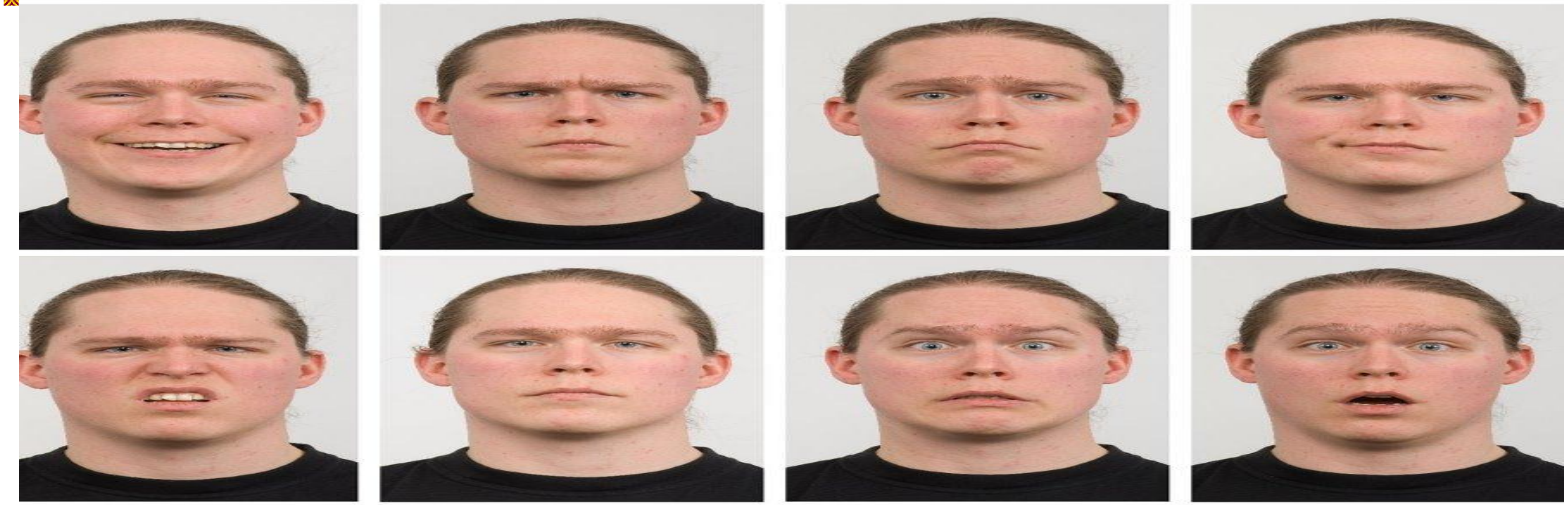
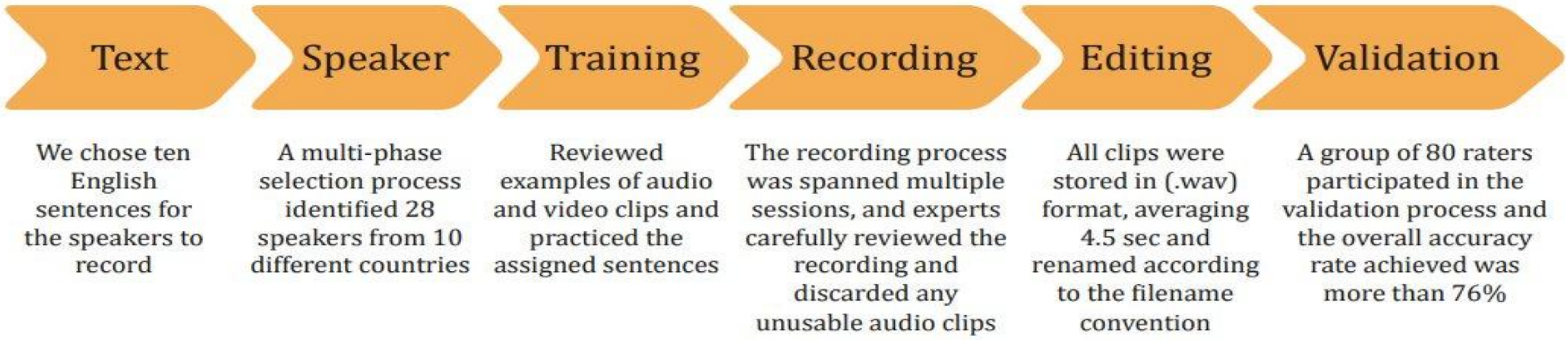


Figure 5: Emotion categories

Emotion Categories

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

Dataset Preparation(Cont.)



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

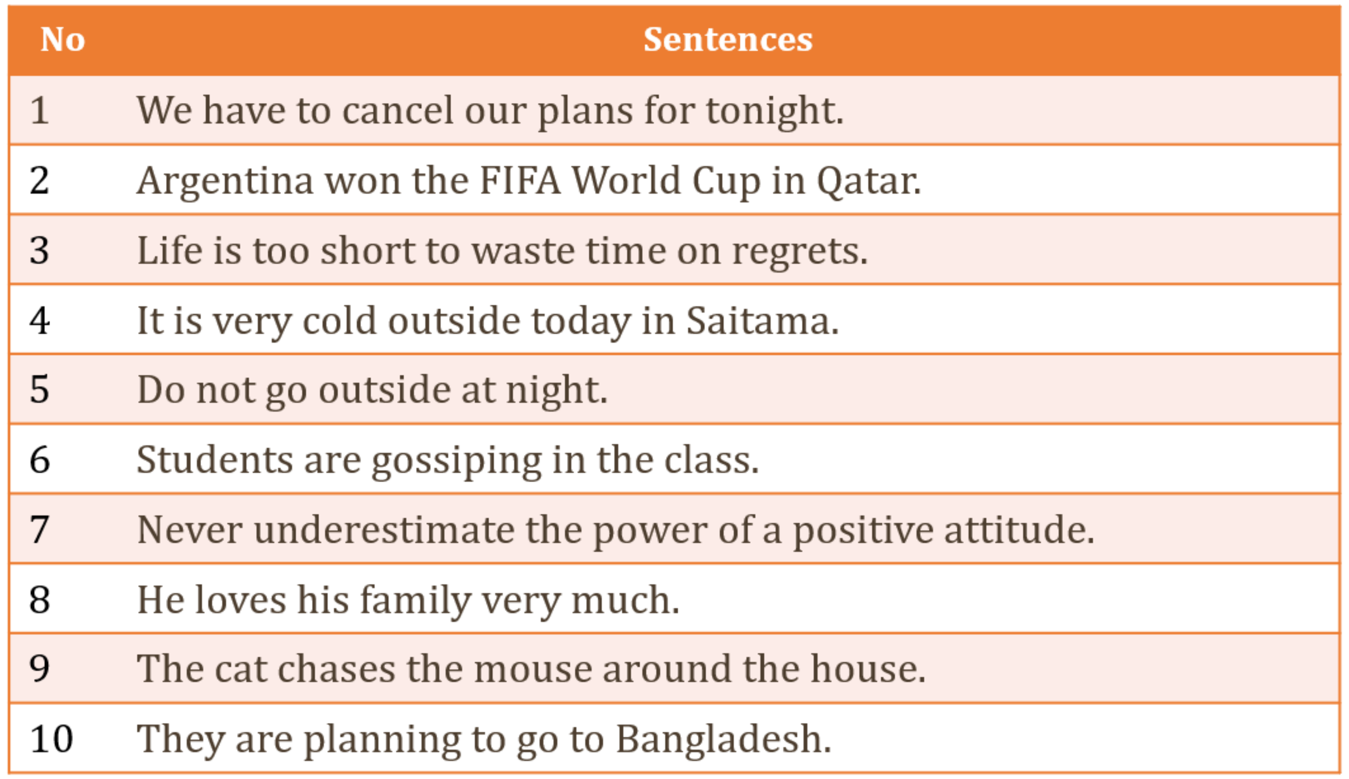
The recording process was spanned multiple sessions, and experts carefully reviewed the recording and discarded any unusable audio clips

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:**
  - ➢ **Transformer model:** a **speech emotion recognition pipeline** where raw audio is processed through CNN-based feature extraction, a Transformer encoder for contextual embeddings, and a dense layer with softmax to classify emotions.
- ❑ **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

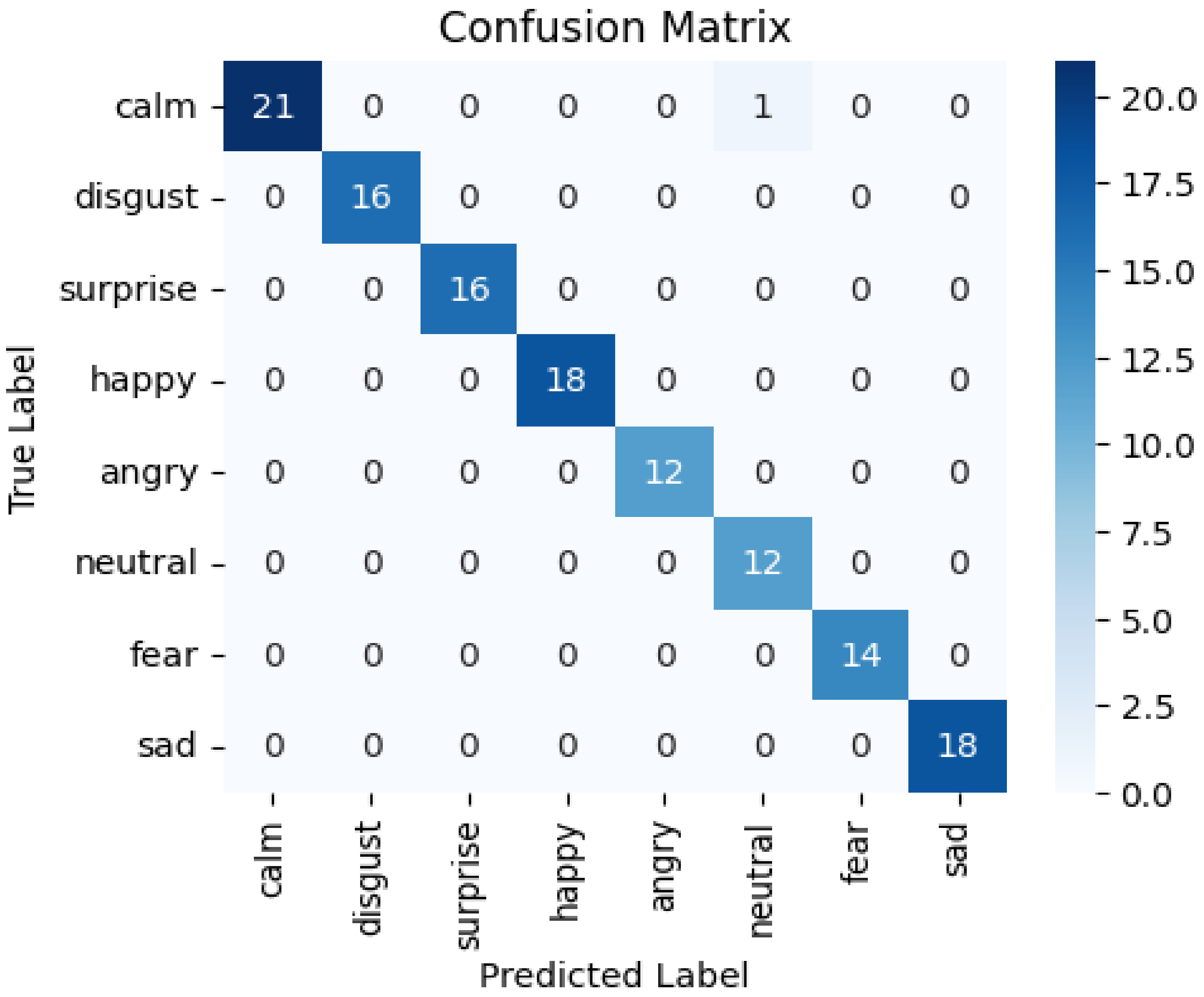


Figure 8: Confusion matrix using Transformer

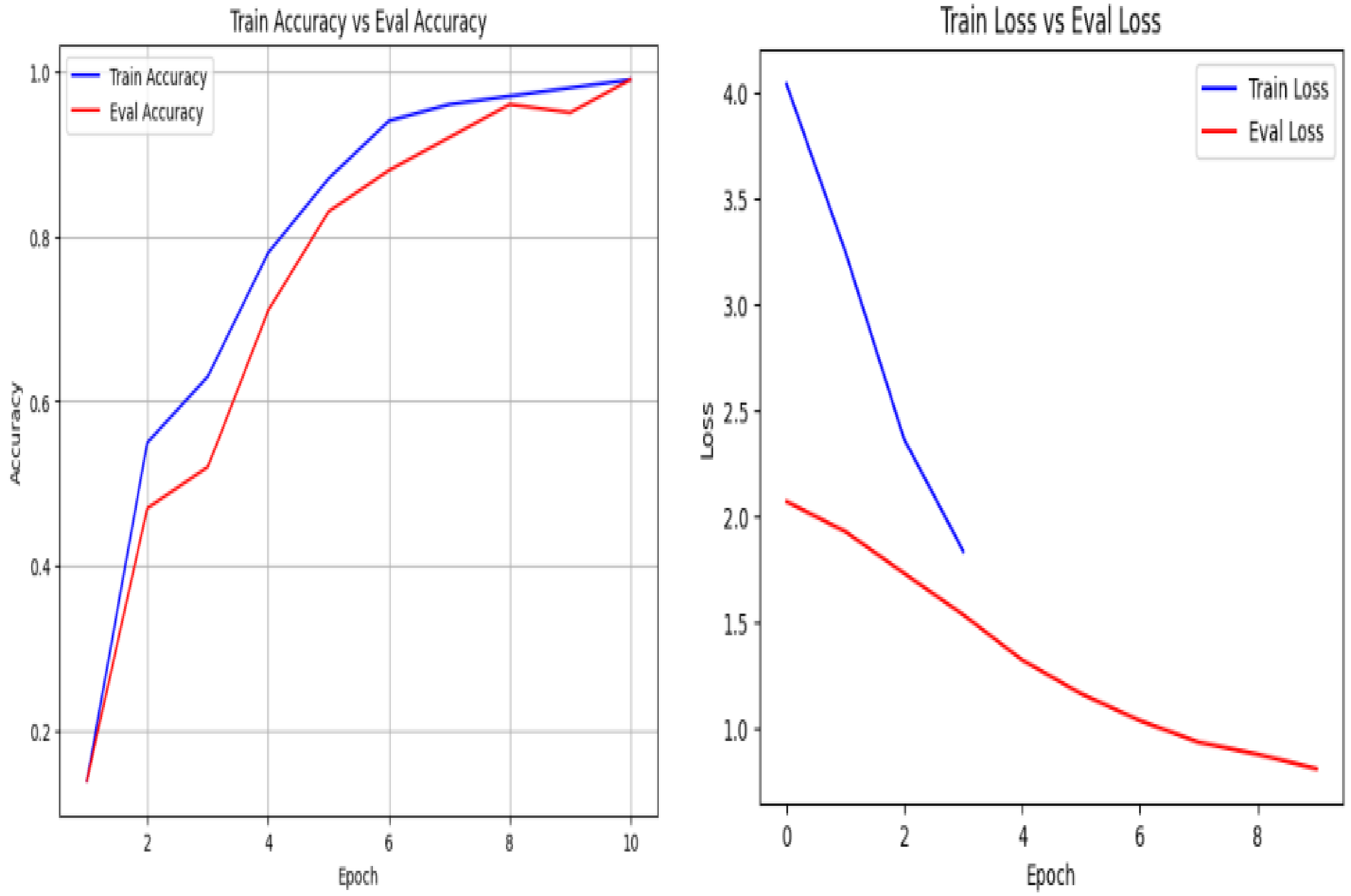


Figure 7: Training and Validation Loss and Accuracy

Results

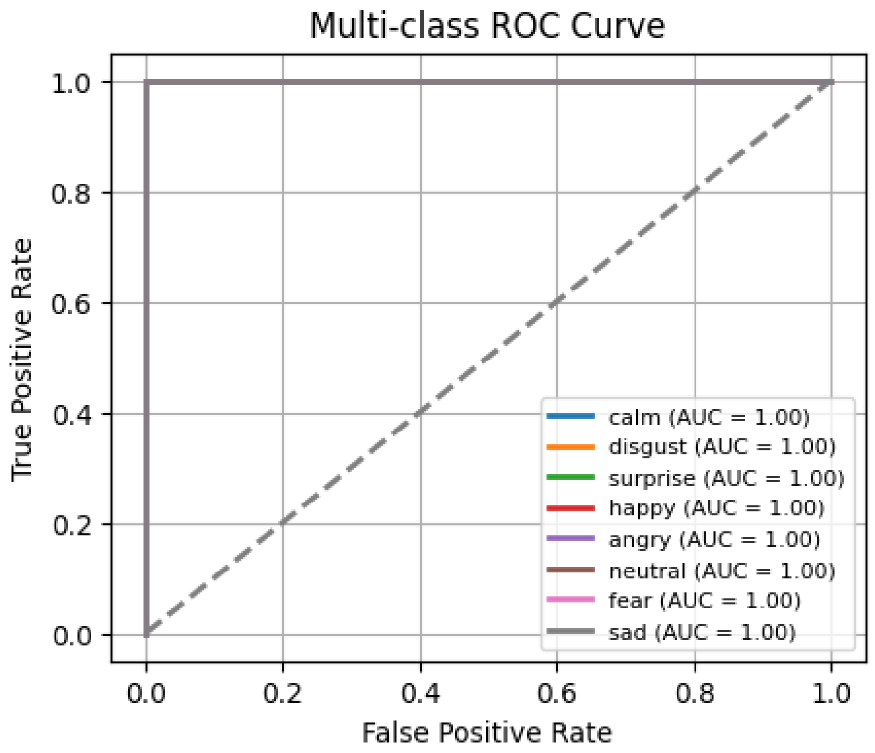


Figure 10: Receiver operating curve

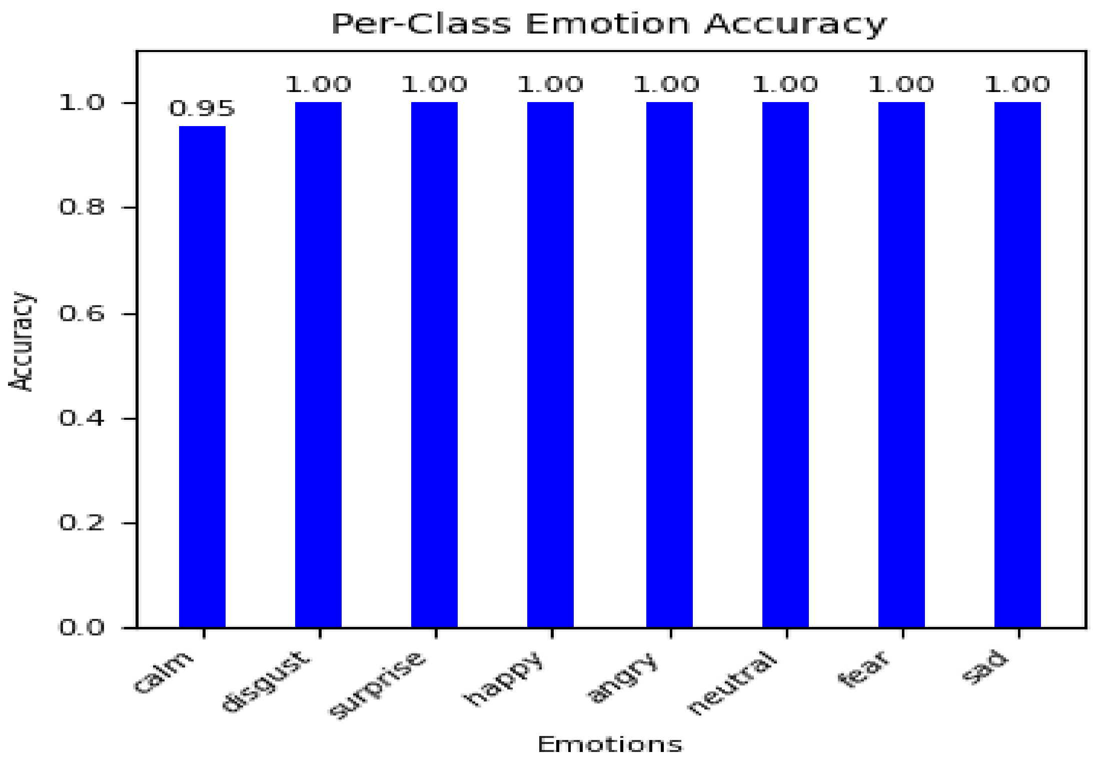


Figure 10: per Class Emotion Accuracy

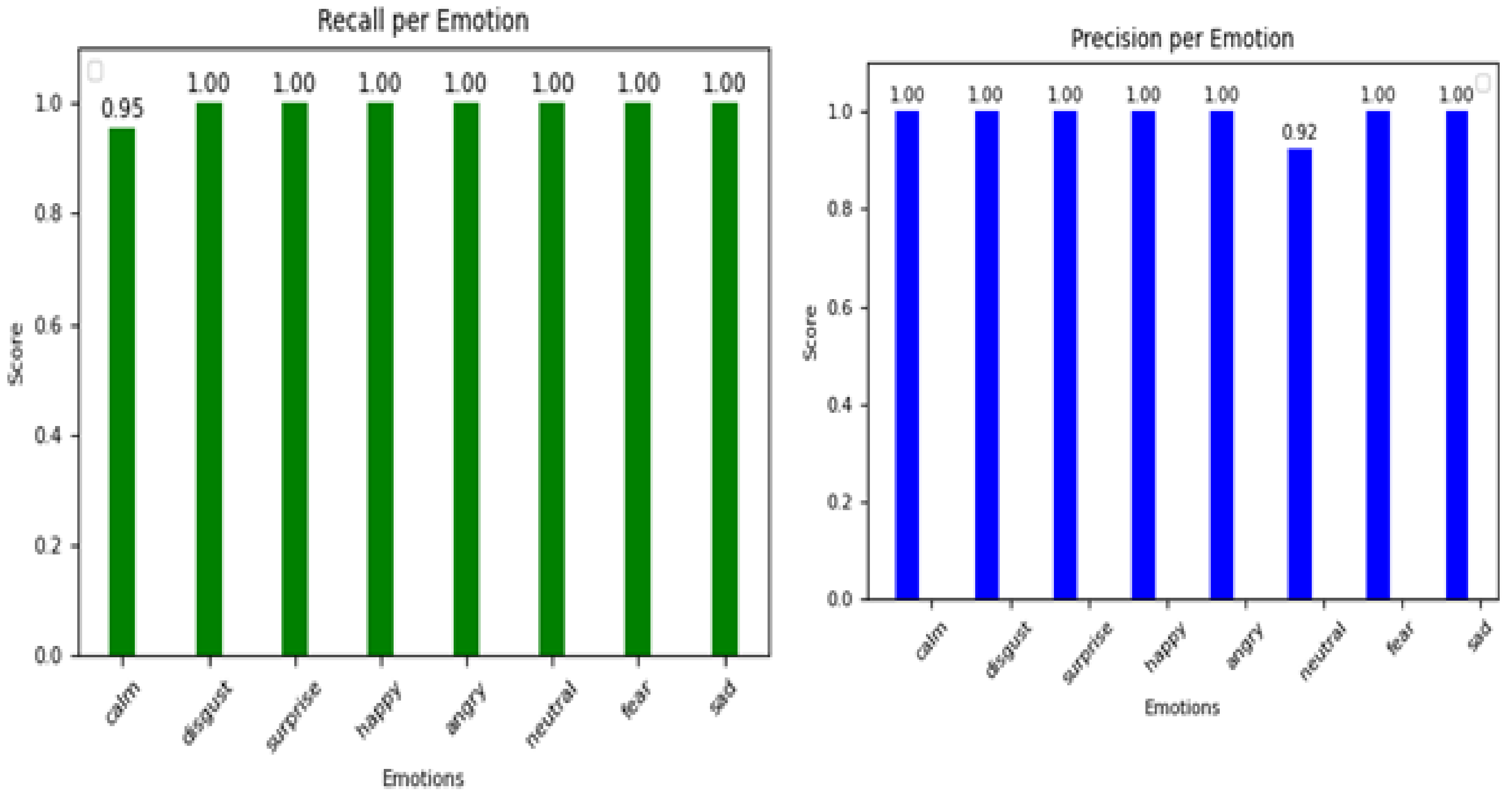


Figure 10: per Class Emotion Accuracy F1 score and Precision

Discussion

TABLE IV: Comparison of Speech Emotion Recognition

Study	Dataset	Model	Accuracy
[3]	EmoBone	Not Specified	76.49%
[4]	RAVDESS, SAVEE	Gradient Boosting	33%,87%
[5]	Local RVDESS	CNN	61.20%
[7]	RAVDESS	CNN	72.50%
[9]	BC speech	BiLSTM	85.17%
Our	EmoBone	Transformer	99.00%

Discussion

- **Balanced dataset impact**:
  - ✓ Distribution of seven emotion classes enables unbiased model training and provides a robust foundation for performance evaluation
- **Transformer 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 99.00%, highlighting the Transformer’s effectiveness in improving emotion recognition from bone-conducted speech

Discussion

- ❑ **Achievements**: **State-of-the-art accuracy of 99%** using the **EmoBone dataset**.
- ❑ Demonstrated the effectiveness of **Transformer 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] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, “Speech emotion recognition using machine learning—a systematic review,” Intelligent systems with applications, vol. 20, p. 200266, 2023.

[2] B. W. Schuller, “Speech emotion recognition,” Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[3] M. S. Hosain, Y. Sugiura, M. S. Rahman, and T. Shimamura, “Emobone: A multinational audio dataset of emotional bone conducted speech,” IEEJ Transactions on Electrical and Electronic Engineering, vol. 19, no. 9, pp. 1492–1506, 2024.

[4] M. R. Hossen, E. Hossain, J. Al-Faruk, J. Sultana, M. B. Islam, and M. S. Hosain, “Tversky loss mechanisms: A resunet approach to improving brain tumor segmentation,” in 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), 2025, pp. 1–6.

[5] A. Iqbal and K. Barua, “A real-time emotion recognition from speech using gradient boosting,” in 2019 international conference on electrical, computer and communication engineering (ECCE). IEEE, 2019, pp. 1–5.

[6] S. N. Zisad, M. S. Hossain, and K. Andersson, “Speech emotion recognition in neurological disorders using convolutional neural network,” in International conference on brain informatics. Springer, 2020, pp. 287–296.

[7] R. Aloufi, H. Haddadi, and D. Boyle, “Emotionless: Privacy-preserving speech analysis for voice assistants,” arXiv preprint arXiv:1908.03632, 2019.

[8] M. S. Hosain, Y. Sugiura, N. Yasui, and T. Shimamura, “Deep-learningbased speech emotion recognition using synthetic bone-conducted speech,” Journal of Signal Processing, vol. 27, no. 6, pp. 151–163, 2023.

[9] N. Wang and D. Yang, “Speech emotion recognition using fine-tuned wav2vec2. 0 and neural controlled differential equations classifier,” PloS one, vol. 20, no. 2, p. e0318297, 2025.

[10] M. S. Hosain, M. R. Hossen, M. U. Mia, Y. Sugiura, and T. Shimamura, “Exploring the emobone dataset with bi-directional LSTM for emotion recognition via bone conducted speech,” in 2025 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), 2025, pp. 97–100.

[11] M. R. Hossen, M. U. Mia, R. Islam, M. S. Hosain, M. K. Hasan, and T. Shimamura, “Facial expression recognition: A machine learning approach with svm, random forest, knn, and decision tree using grid search method,” in 2025 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), 2025, pp. 421–424.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in neural information processing systems, vol. 33, pp. 12 449– 12 460, 2020.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[14] D. Jurafsky and J. H. Martin, Speech and Language Processing, 2025, draft of January 12, 2025. [Online]. Available: https://web.stanford.edu/ jurafsky/slp3/

[15] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org.

Thank You for Your Kind Attention …