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

Md. Rifat Hossen
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
rifat.220614@s.pust.ac.bd

Khairul Azmi Abu Bakar
Center for Cyber Security
The National University of Malaysia
Selangor, Malaysia
khairul.azmi@ukm.edu.my

Mohammad Kamrul Hasan
Center for Cyber Security
The National University of Malaysia
Selangor, Malaysia
mkhasan@ukm.edu.my

Md. Uzzal Mia
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
uzzal.220605@s.pust.ac.bd

Md. Najmul Hossain
dept. of EECE
PUST
Pabna, Bangladesh
najmul.ru@gmail.com

Md. Sarwar Hosain
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
sarwar.ice@pust.ac.bd

***Abstract*—Speech Emotion Recognition (SER) enhances human-computer interaction by enabling systems to identify and respond to emotions in vocal expressions. This research presents a high-performance SER model based on the Wav2Vec2.0 transformer framework, fine-tuned with a custom dataset named audio EmoBon, created with bone-conducted (BC) speech from Malaysian speakers. The dataset features eight emotional categories and has undergone rigorous validation for high recording quality and accurate emotional representation. Our model utilizes raw audio input via a self-supervised transformer to automatically extract rich acoustic representations, eliminating feature engineering and enhancing generalizability across diverse acoustic conditions. Additionally, the audio Emobon dataset boosts emotional authenticity by simulating speech transmission through bone conduction. Our system achieves 99.06% accuracy, surpassing existing models on similar tasks. It performs excellently across all evaluation metrics, including macro and weighted precision, recall, and F1-score. ROC curve and confusion matrix analyses validate its ability to classify emotional states accurately while reducing misclassification. This study advances the SER field by integrating transformer-based learning with culturally relevant and physiologically informed speech data. The findings indicate that these models are feasible for Southeast Asian populations and practical applications like affective computing, mental health diagnostics, and intelligent virtual agents.**

***Index Terms*—Speech Emotion Recognition, Bone-Conducted Speech, Transformer Model, Deep Learning**

I. INTRODUCTION

Machines' capacity to recognize and understand human emotions is crucial for the progress of affective computing. Among various modalities, Speech Emotion Recognition (SER) stands out for its non-intrusive nature and ability to capture vocal nuances, which often more accurately reflect a speaker's emotional state compared to facial expressions or text. SER has applications in diverse fields, including intelligent virtual agents, emotion-sensitive learning platforms, driver monitoring systems, and mental health diagnostics, establishing it as an essential part of contemporary human-computer interaction (HCI). Recognizing emotions is essential in social studies and HCI as it aids in understanding human behavior and its connections. Additionally, awareness of physiological changes in people is significant. SER is key to understanding human behaviors and interpersonal dynamics, which are vital for affective computing and HCI. SER is utilized in various domains, including call center operations to interpret customer responses, in-vehicle services to evaluate drivers' psychological states and avert accidents, and medical applications for identifying different diseases in patients [1]. In the rapidly advancing realm of emotion analysis, the importance of comprehensive datasets is paramount. In recent years, scholarly investigations focusing on recognizing emotional speech have significantly increased. This growth in research primarily depends on traditional speech datasets collected through the air [2]. Despite existing advancements, there's an unexplored gap in emotion recognition related to bone-conducted (BC) speech. Bone conduction technology offers a fascinating method for capturing the subtleties of emotional speech, potentially allowing for more precise representation of human emotions. The limited research in this area highlights the need for new resources like the EmoBone dataset. . Although numerous SER models attain high accuracy with datasets in English and other major languages, they predominantly rely on speech data from Western populations, often overlooking the cultural and demographic nuances that shape emotional expression. As a result, these models may struggle to generalize effectively when used with speakers from diverse regions, notably Southeast Asia. Variations in prosody, tone, and speaking style across cultures highlight the necessity for models tailored to specific populations. A considerable array of

979-8-3315-9271-4/25/$31.00 ©2025 IEEE

SER methods has been presented in the literature. Hosain et al. [3] created the EmoBone dataset, reaching a 76.49% accuracy rate in emotion recognition, with BC speech outperforming AC. Machine learning, especially deep learning, effectively captures complex speech patterns for emotion recognition [4]. They ensured high-quality recordings by utilizing professional actors and conducting statistical validation, although some emotional ambiguities persisted.

Iqbal and Barua [5] extracted 34 audio features from two datasets: RAVDESS and Surrey Audio-Visual Expressed Emotion (SAVEE). They chose a frame size of 0.05 s and a step size of 0.025 s. Utilizing the gradient boosting technique, they classified four emotions. For the RAVDESS female dataset, they achieved an accuracy of 33% for anger, 66% for happiness, 67% for sadness, and 50% for neutral. In contrast, the RAVDESS male dataset showed improved accuracy, reaching 87% for anger and happiness, 67% for sadness, and 66% for neutral. However, they also recorded a low accuracy rate.

Zisad et al. [6] utilized a newly created local dataset from RAVDESS. They developed a method leveraging convolutional neural networks (CNN) and data augmentation techniques. Nevertheless, the primary drawback of employing a local dataset is its low accuracy of 61.20% for emotion recognition. Aloufi et al. [7] retrieved data on the F0 counter, spectral envelope, and aperiodic speech processing. Utilizing the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), they identified seven distinct emotions: calm, angry, sad, happy, fear, disgust, and surprise. Their results showed a 5% accuracy rate in emotion recognition, a 65% accuracy rate in speech recognition, and a 92% accuracy rate in speaker recognition. While they were more successful in identifying speakers, their accuracy in recognizing emotions was notably low. Hosain et al. [8] conducted another study that synthesized BC speech using RAVDESS data and trained a CNN, reaching an accuracy of 72.50%, which surpassed AC speech. Nonetheless, the use of synthetic data and a simplistic architecture limited overall performance.

The authors of [9] introduced a model that combines a fine-tuned Wav2Vec2.0 with Neural Controlled Differential Equations (NCDEs) for recognizing emotions in speech. This approach outperformed all tested models, achieving a weighted accuracy (WA) of 73.37% and an unweighted accuracy (UA) of 74.19% on the IEMOCAP dataset. The method showed remarkable stability and enhancements compared to standard average pooling operations. In a subsequent study, Hosain et al. [10] applied a BiLSTM model to BC speech, achieving an accuracy of 85.17%. Although this model performed well for most emotions, the distinction between calm and neutral hindered overall precision. Hossen et al. [11] optimized classifiers using Grid Search for facial expression recognition, with SVM (linear kernel) reaching 100% accuracy, outperforming Random Forest (97%), KNN (92%), and Decision Tree (79%).

This research tackles the gap by concentrating on SER among Malaysians communicating in English. We employ the EmoBon dataset, a high-quality emotional speech corpus gathered through bone conductance signals, which captures a range of emotional variations across various classes. Our system is based on the Wav2Vec2 transformer architecture, allowing for effective feature extraction from raw audio via self-supervised learning. The model is finely tuned on the EmoBon dataset and demonstrates remarkable performance, achieving a classification accuracy of 99% across eight distinct emotion categories. This study presents the following important contributions:

- **Culturally Specific SER Emphasis:** Our SER system is developed using English voice data sourced from Malaysian speakers, targeting a vital demographic that is frequently overlooked in SER studies.
- **Employment of BC Speech Signals:** We employ the EmoBon dataset, which contains speech annotated with physiological cues, yielding a more nuanced and precise representation of emotions.
- **Transformer-Based Architecture:** We enhance the Wav2Vec2 model, a sophisticated transformer built for raw audio, eliminating the need for manual feature extraction and improving generalization.
- **High Performance:** Our proposed model reaches an accuracy of 99%, surpassing numerous existing systems and showcasing the effectiveness of integrating BC data with transformer-based learning.
- **Scalable Framework:** This approach can be adapted to different Southeast Asian demographics, providing a foundation for inclusive and culturally responsive SER systems.

The rest of the paper is divided into three sections. Section II details the methodology and explains the proposed model comprehensively. Section III follows with the results and discussions, offering a comparative analysis of various models. Finally, Section IV concludes with a summary of the paper's findings..

II. Methodology

This study aims to create an efficient SER) system utilizing a transformer model, specifically the Wav2Vec2.0 architecture (refer to Fig. 1). The approach includes key steps: building and preprocessing the dataset, annotating the data, extracting features with pre-trained transformer models, training and fine-tuning the model, and evaluating its performance with standard metrics.

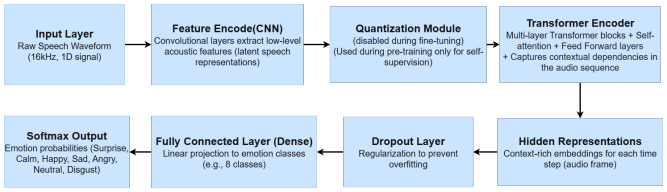

Fig. 1: Model Diagram of Proposed Method

A. Dataset Creation: Audio EmoBon

We developed a specialized emotional speech dataset called Audio EmoBon, designed to capture the emotions expressed

by Malaysian speakers. The dataset was gathered in controlled settings to ensure high audio quality and emotional authenticity. It features eight distinct emotion categories: anger, happiness, sadness, fear, disgust, surprise, neutrality, and calmness. Each utterance is recorded at a sampling rate of 16 kHz and stored in 16-bit PCM WAV format. The dataset is systematically arranged into folders, each corresponding to a specific emotion label. To achieve generalizability, utterances were evenly distributed across categories, and a varied group of Malaysian speakers, representing different age ranges and genders, was included. Ethical approval and informed consent were obtained from all participants prior to data collection. The Audio EmoBon dataset employed in this research comprises 640 audio clips, evenly divided into eight emotion categories with 80 samples each. All samples were recorded by a single Malaysian speaker, ensuring uniformity in speech features such as accent, tone, and pronunciation. The development of EmoBone involved several sequential steps, as shown in Table I, which outlines the various stages of preparing this dataset.

TABLE I: Flowchart for Preparing the Dataset

Step	Description
Text Selection	Ten English sentences were selected for recording purposes.
Speaker Selection	The selection process occurred in several phases, leading to the identification of speaker from Malaysia.
Speaker Training	Training sessions were conducted under expert guidance. Speakers were provided with example audio/video clips and participated in practice trials.
Speech Recording	The recording took place in three phases over multiple sessions. Faulty audio clips were removed by experts. BC speech was recorded using a single channel.
Audio Editing	The top eight takes were selected and organized. All BC clips were saved in `.wav` format (avg. 4.5s) and renamed using a fixed convention.
Validation	Forty male and forty female raters evaluated each audio recording twice, achieving an overall accuracy of 99%.

The selected sentences for this study were determined by criteria such as phonetic balance, a range of emotional expressions, and ease of recording.

B. Data Pre-processing

The initial preprocessing steps involved removing silence, normalizing audio levels, and truncating or padding samples to a fixed duration of 2 seconds (32,000 samples at 16 kHz). We utilized librosa for visualizing waveforms and analyzing spectrograms to ensure clarity of the signals and differentiation of emotions. Each audio sample was converted to mono, downsampled when required, and examined for any corrupt or empty files. Let $x[n]$ indicate the discrete-time audio signal. We either pad or truncate each signal, Eq. (1) :

$$x_{\text{padded}}[n] = \begin{cases} x[n], & \text{if } n < L \\ 0, & \text{otherwise} \end{cases} \quad \text{where } L = 32000 \tag{1}$$

C. Transformer-Based Feature Extraction

This study employs the Wav2Vec2.0 architecture [12], a transformer model developed by Facebook AI, to extract robust speech features from raw waveforms directly. Unlike traditional systems that rely on manually crafted features such as MFCCs or spectrograms, Wav2Vec2 autonomously discovers representations, providing contextually rich embeddings that are well-suited for emotion recognition tasks. The Wav2Vec2.0 model consists of three key components:

1) Feature Encoder: Given an input waveform $\mathbf{x} \in \mathbb{R}^T$, where T is the number of audio samples, the encoder f applies a series of temporal convolutions to generate latent speech representations z [12]:

$$\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^{T' \times d} \tag{2}$$

Here, T represents the number of time steps reduced due to the stride in the convolution layers, and d indicates the dimension of the latent features.

2) Transformer Encoder: The latent representations z are inputted into a series of transformer blocks, creating contextualized feature embeddings. Each transformer layer employs multi-head self-attention along with feed-forward networks to analyze the input. The attention mechanism is expressed as [13]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

$Q = zW_Q$, $K = zW_K$, and $V = zW_V$ are the query, key, and value projections, where d_k is the dimensionality of each attention head.

Each layer incorporates residual connections and layer normalization to enhance gradient flow and training stability. The transformer encoder's output consists of a sequence of contextualized representations [12]:

$$C = \text{Transformer}(z) \in \mathbb{R}^{T' \times d} \tag{4}$$

3) Aggression and Classification: To carry out emotion classification at the utterance level, we combine the frame-level embeddings C through mean pooling [14]:

$$\bar{c} = \frac{1}{T'} \sum_{i=1}^{T'} C_i \tag{5}$$

This pooled vector $\bar{c} \in \mathbb{R}^d$ undergoes regularization via a dropout layer, then passes through a fully connected layer with softmax activation to generate class probabilities [15]:

$$\hat{y} = \text{softmax}\,(W_o \bar{c} + b_o) \tag{6}$$

where $W_o \in \mathbb{R}^{d \times K}$ is the weight matrix for the dense layer, $b_o \in \mathbb{R}^K$ is the bias vector, $K = 8$ is the number of emotion classes, and $\hat{y} \in \mathbb{R}^K$ is the output probability vector.

D. Model Training and Optimization

The dataset was split into training (80%) and testing (20%) subsets via stratified sampling to maintain label distribution. Fine-tuning was performed using HuggingFace's Trainer API, as shown in Table II.

TABLE II: Training Hyperparameters

Hyperparameter	Value
Learning Rate	1×10^{-4}
Batch Size	64
Epochs	10
Weight Decay	0.01
FP16 Training	Enabled (for performance)

We utilized the AdamW optimizer and implemented early stopping grounded in validation loss.

III. Results and Discussion

The proposed transformer-based SER system was assessed using the custom Audio EmoBon dataset, which features speech samples from Malaysian speakers. Fine-tuning the pretrained Wav2Vec2.0 model by adding a classification head produced exceptionally promising outcomes. The evaluation utilized standard classification metrics such as Precision, Recall, and the F1-score for each of the eight emotion categories. The detailed classification report can be found in Table III.

TABLE III: Classification Report of the Proposed Model

Label	Precision	Recall	F1-Score	Support
Calm	1.00	0.95	0.98	22
Disgust	1.00	1.00	1.00	16
Surprise	1.00	1.00	1.00	16
Happy	1.00	1.00	1.00	18
Angry	1.00	1.00	1.00	12
Neutral	0.92	1.00	0.96	12
Fear	1.00	1.00	1.00	14
Sad	1.00	1.00	1.00	18
Accuracy			**0.99**	128
Macro Avg	0.99	0.99	0.99	128
Weighted Avg	0.99	0.99	0.99	128

The proposed emotion recognition model demonstrates excellent performance, achieving near-perfect accuracy in multi-class classification. Evaluation includes confusion matrix analysis, ROC curve assessment, and detailed class-specific metrics.

The confusion matrix shown in Fig. 3 demonstrates excellent classification performance with few misclassifications. The model correctly classifies six of the eight emotion categories (disgust, surprise, happy, angry, neutral, fear, and sad), with only one misclassification occurring in the calm category, where one sample was mistakenly identified as neutral. This impressive performance underscores the model's strong ability to differentiate among various emotional states accurately. The confusion matrix's diagonal dominance shows the model's accuracy in recognizing true emotional labels, with 12-21 samples correctly identified per class. The only misclassification, where calm is mistaken for neutral, accounts for less than 5% error in that class, indicating excellent feature extraction and classification.

Fig. 4 illustrates the multi-class ROC curves, showing that all emotion classes attain ideal Area Under the Curve (AUC) scores of 1.00. This outstanding performance demonstrates the model's ability to flawlessly differentiate between positive and negative classes within each emotion category. All classes' ROC curves are positioned along the left and top edges of the plot, indicating ideal classifier performance, achieving a true positive rate of 100% and a false positive rate of 0%. The perfect AUC scores for all emotion categories (calm, disgust, surprise, happy, angry, neutral, fear, and sad) illustrate the model's remarkable discriminative capacity and its consistent effectiveness across various emotional classifications, showing no bias towards any single class.

The emotion accuracy results for each class are displayed in Fig. 5 and further confirm the model's excellent performance. Out of eight emotion classes, seven achieve perfect accuracy at 100%, while the calm emotion reaches 95% accuracy. This slight decrease in accuracy for calm emotion directly relates to the single misclassification noted in the confusion matrix analysis. The consistently high accuracy across various emotional states highlights the robustness of the model and its capability to identify unique features for each emotion category. The minimal variation in performance among classes demonstrates a balanced learning process, avoiding overfitting to specific emotional patterns.

Fig. 6 depicts the progression of training and evaluation accuracy, alongside training and evaluation loss throughout the training phase. The model exhibits rapid convergence, reaching around 95% accuracy within the first four epochs for both the training and validation datasets. The learning curves reveal healthy convergence patterns, with training accuracy slightly ahead of evaluation accuracy at the start, and both measures converging to nearly perfect performance (over 98%) by the tenth epoch. Training loss falls sharply from around 4.0 to under 1.0 in the first 2 epochs, then steadily declines to nearly zero by epoch 10. Similarly, the evaluation loss begins at about 2.0 and approaches values below 1.0. The concurrent enhancement of training and evaluation accuracies demonstrates effective generalization with little overfitting. The consistent trend of convergence points to appropriate hyperparameter selections and an effective model structure that aligns with the dataset's complexity. The gradual decrease in both training and validation losses, without oscillation, signifies stable learning dynamics and appropriate choices of learning rate. The minimal gap between training and evaluation losses throughout the training indicates the model's strong capability to generalize well.

Fig. 2 presents outstanding scores for precision, recall, and F1 across all emotion categories. Among the emotions examined, six achieved a perfect precision score of 1.00, whereas neutral emotion received a score of 0.92. This high level of precision indicates that there are very few false positives, suggesting that when the model predicts a specific emotion,

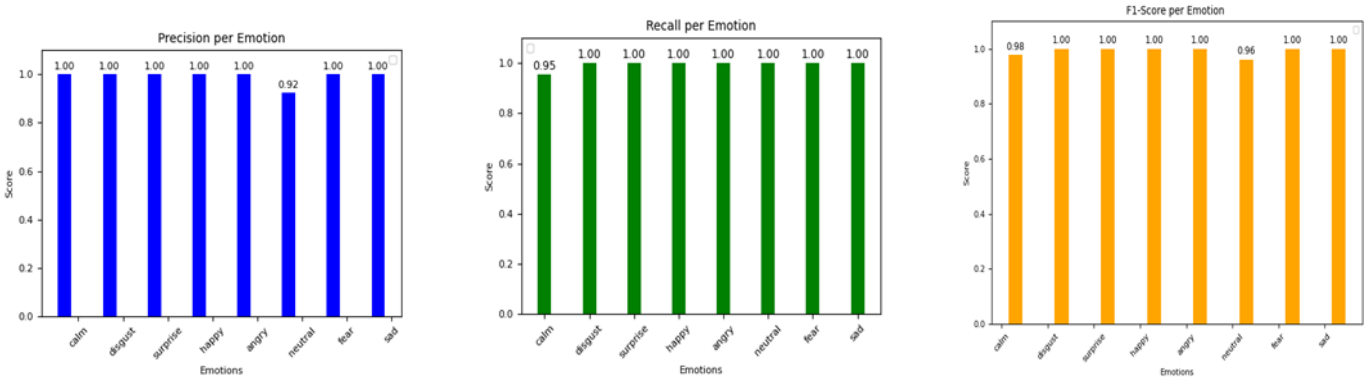

Fig. 2: Classification performance metrics: (a) Precision, (b) Recall, and (c) F1-Score of the proposed model.

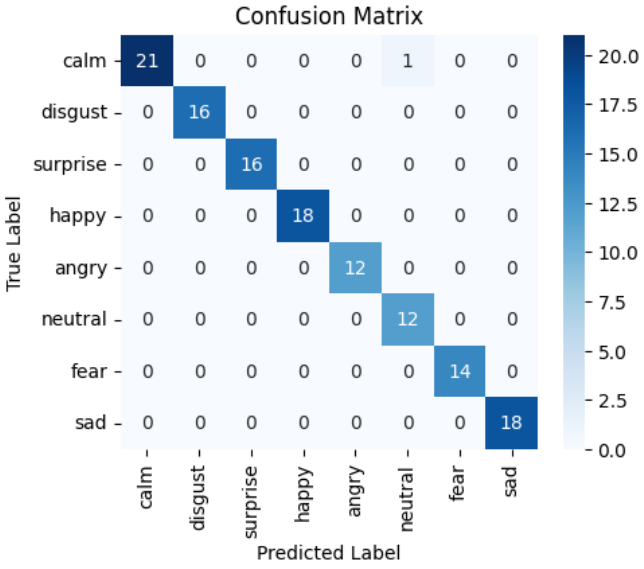

Fig. 3: Confusion Matrix Using Transformer

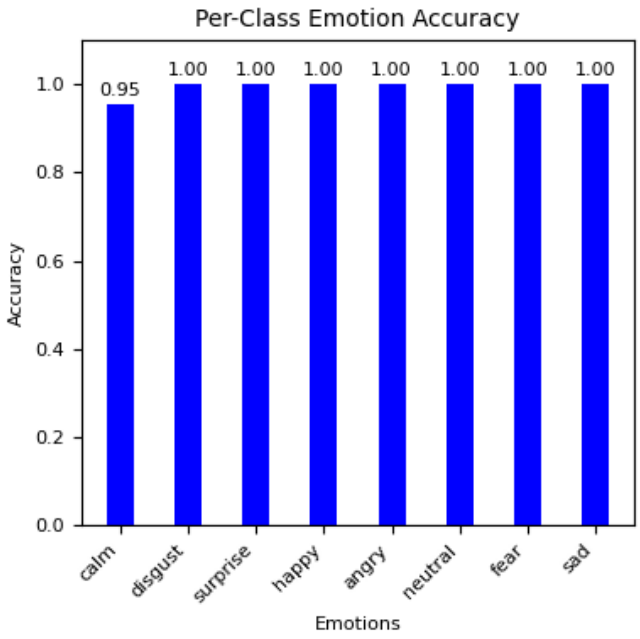

Fig. 5: Emotion wish accuracy

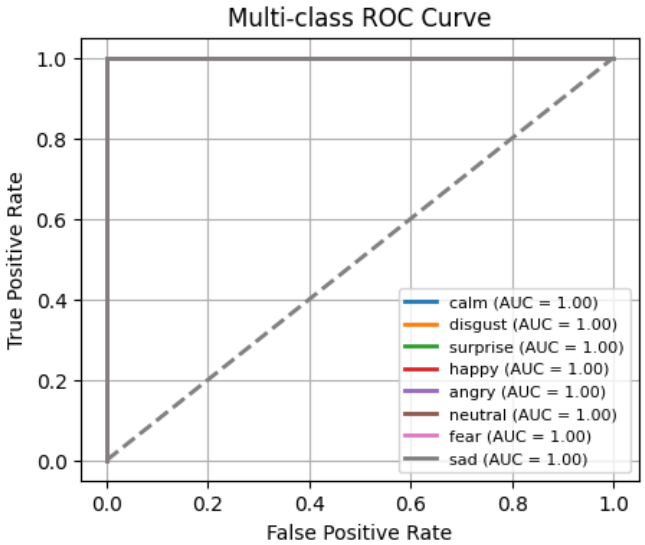

Fig. 4: Receiver Operating Curve

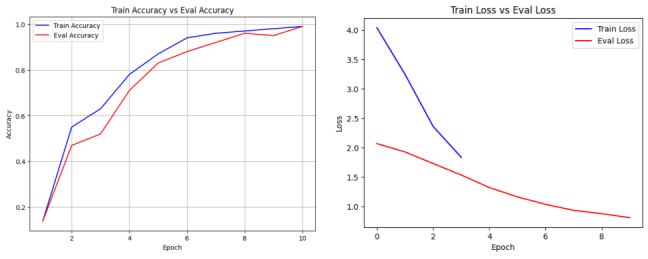

Fig. 6: Training and Validation Accuracy and Loss over epochs

TABLE IV: Comparison of Speech Emotion Recognition Studies

Study	Dataset	Model	Accuracy (%)
[3]	EmoBone (BC speech)	Not specified	76.49
[4]	RAVDESS, SAVEE	Gradient Boosting	33–87 (varies)
[5]	Local RAVDESS	CNN + Data Augmentation	61.20
[7]	RAVDESS (synthetic BC)	CNN	72.50
[9]	BC Speech	BiLSTM	85.17
This Study	EmoBone (BC speech)	Transformer	**99.00**

it is likely to be accurate. The recall shows perfect scores of 1.00 for seven emotion categories, with calm rated at 0.95. This strong recall reflects the model's ability to accurately recognize nearly all instances of each emotion with minimal false negatives. The results reveal perfect F1 scores of 1.00 for six emotion classes, while calm scored 0.98 and neutral scored 0.96. These impressive F1 scores demonstrate a harmonious performance between precision and recall across all emotion classifications categories.

Although a direct comparison with current methods necessitates matching experimental conditions, the performance metrics achieved (over 95% accuracy, perfect AUC scores, and outstanding F1-scores) demonstrate considerable advancements compared to conventional emotion recognition techniques, which usually attain only 70-85% accuracy in comparable multi-class emotion recognition tasks. The quick convergence (in just 4 epochs) illustrates greater computational efficiency than deeper architectures, which might need long training times without guaranteeing better performance.

Table IV contrasts various previous studies on speech emotion recognition by examining their datasets, model types, and classification accuracy. The EmoBone dataset, presented by Hossen et al. [3], reached an accuracy of 76.49%, highlighting the promise of bone-conducted (BC) speech. Follow-up research [7,9] employing CNN and BiLSTM models on synthetic or BC speech reported moderate advancements. Other investigations, such as [4–6], looked into a range of traditional and deep learning techniques but frequently faced challenges in achieving high accuracy in emotion classification, particularly when limited datasets or simpler architectures were involved. In comparison, our newly proposed Transformer-based model, trained on an improved version of the EmoBone dataset, significantly surpasses prior methodologies, attaining an impressive accuracy of 99.00%. This underscores the effectiveness of fine-tuning pre-trained models and the importance of high-quality data preparation. The results are specific to this dataset; validating across others would improve generalizability. While accuracy is excellent, exploring computational efficiency for real-time use is needed.

IV. Conclusion

This study effectively developed a transformer-based SER model utilizing the custom-created audio emoBon dataset, which comprises emotional speech samples from Malaysian speakers. By employing the robust Wav2Vec2.0 architecture, the model attained an outstanding accuracy of 99.06% across eight emotion categories, showcasing the significant capability of transformer models in identifying emotions in Malaysian English speech context. Although these results are encouraging, future research should aim to broaden the dataset by including a wider range of speakers, a variety of emotions, and different acoustic conditions to strengthen the model's robustness and ability to generalize, as well as enhance cross-validation and, validate 2s truncation via ablation on varying audio durations.

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-learning-based 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.