2025 IEEE 7th International Conference on Sustainable Technologies
for Industry 5.0 (STI 2025), 11–12 December, Dhaka, Bangladesh

Speech Emotion Recognition from Bone-Conducted Speech Using Wav2Vec2 Transformer Model

Manik Kumar Saha[1], Md. Sarwar Hosain[1], Md. Rifat Hossen[1], Sajeeb Kumar Ray[1], Liton Chandra Paul[2], Mohammad Shorif Uddin[3,4]

[1]Department of Information and Communication Engineering, PUST, Pabna, Bangladesh
[2]Department of Electrical, Electronic and Communication Engineering, PUST, Pabna, Bangladesh
[3]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh
[4]Green University of Bangladesh, Rupganj, Narayanganj, Bangladesh
manikpustice@gmail.com, sarwar.ice@pust.ac.bd, rifat.220614@s.pust.ac.bd,
sajeeb.ray.ice@gmail.com, litonpaulete@gmail.com, shorifuddin@juniv.edu

***Abstract*—Speech emotion recognition (SER) is an essential technology for enhancing human-computer interactions (HCI). While most SER research uses air-conducted (AC) speech, bone-conducted (BC) speech offers a resilient alternative, especially in noisy environments. This paper introduces an end-to-end SER system based on the Wav2Vec2.0 transformer model, fine-tuned with the EmoBone dataset—a comprehensive, multi-national BC speech dataset featuring eight emotion categories collected from 29 speakers in 10 countries. Our method utilizes self-supervised learning to bypass manual feature extraction, learning detailed contextual features directly from raw audio waveforms. The system combines a custom classification head with the pre-trained Wav2Vec2.0 encoder for efficient emotion prediction. Evaluation results show that our approach attains an overall accuracy of 93% and a weighted average F1-score of 93% on the EmoBone dataset, markedly surpassing earlier best-performing techniques. The model demonstrates strong effectiveness in separating emotions with unique acoustic features but encounters challenges differentiating acoustically similar emotions like neutral-sad and fear-disgust pairs. These results underscore the promising capabilities of transformer-based architectures for BC speech emotion recognition and set a new standard for future studies. Additional ablation studies evaluate channel effects (AC-only, BC-only, and AC+BC fusion) and pretraining configurations (from-scratch, linear probe, full fine-tune). This study is among the first to demonstrate self-supervised transformer efficacy for bone-conducted emotion recognition, setting a benchmark for future SER research.**

***Index Terms*—speech emotion recognition, wav2vec2.0, transformer, deep learning, audio signal processing, hugging face.**

I. Introduction

Emotion recognition is important for both social studies and HCI because it helps us understand human behavior. SER plays a vital role in affective computing and HCI, finding applications in areas such as customer service, in vehicle systems, and medical diagnostics [1]. There has been a rise in emotion recognition research using air-conducted (AC) speech, but BC speech remains underexplored [2]. BC speech may capture emotional nuances better, especially in noisy environments [3]. However, there is no alternative dataset for BC speech emotion recognition except EmoBone [4]. 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 has 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 BC speech. BC 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. They ensured high-quality recordings by utilizing professional actors and conducting statistical validation, although some emotional ambiguities persisted. The main contributions of this paper are as follows:

- To developed an end-to-end SER framework using the pre-trained Wav2Vec2.0 transformer model without requiring handcrafted acoustic features.
- We fine-tune the Wav2Vec2.0 model on a multi-class emotional speech dataset comprising eight emotions, achieving its Superior performance on emotion classification.
- We also analysis the performance of the model using different evaluation metrics such as precision, recall, F1-score and confusion matrix for complete class-wise behavior analysis.
- We demonstrate that the transformer model outperforms conventional methods, such as raw waveform inputs and long context dependencies for speech.
- We also offer an in-depth visualization and error analysis, In order to discover popular misclassification patterns and

979-8-3315-8310-1/25/$31.00 ©2025 IEEE

discuss possible causes and improvements in the future.

The rest of this paper is divided into four sections. Section II reviews related work, emphasizing existing methods and recent developments relevant to our research. Section III explains the methodology in detail and provides a comprehensive overview of the proposed model. Section IV presents the experimental results and offers a comparative analysis of various models. Finally, Section VI concludes the paper with a summary of the main findings and suggestions for future research.

II. Related Work

SER has garnered significant interest recently due to its potential applications in HCI, mental health monitoring, and intelligent voice assistants. Researchers have explored various deep learning and machine learning methods to enhance the accuracy and broad applicability of emotion detection systems. Banihosseini and Ghods [5] introduced a three-stage SER framework combining StarGAN for data augmentation, deep convolutional neural networks (DCNN) for feature extraction, and support vector machines (SVM) for classification. This approach achieved high accuracy rates of 98.25% on the ryerson udio-visual database of emotional speech and song (RAVDESS) dataset and 95.5% on Emo-DB. Nonetheless, it shows increased computational complexity and variable performance across datasets.

To further improve SER performance, Akinpelu et al. [6] developed an improved and faster region-based convolutional neural network (IFR-CNN), combining improved intersection over Union (IIOU) with a recurrent neural network (RNN) to retain emotional states. Their model reached 89.5% accuracy on the Berlin database of emotional speech (EMODB) and 94.82% on the geneva emotional speech database (GEES). However, the system still faced challenges distinguishing between emotions that are closely related.

Iqbal and Barua [7] extracted 34 audio features from two benchmark datasets, RAVDESS and the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset, using a frame size of 0.05 seconds and a step size of 0.025 seconds. They used a gradient boosting classifier to recognize four emotional states. On the RAVDESS female dataset, the model achieved relatively low accuracies: 33% for anger, 66% for happiness, 67% for sadness, and 50% for neutral. In contrast, the performance on the RAVDESS male dataset was better, achieving 87% for both anger and happiness, 67% for sadness, and 66% for neutral. However, the overall results showed inconsistent performance and limited generalizability. Zisad et al. [8] proposed a CNN-based SER model trained on a locally created dataset derived from RAVDESS. They used data augmentation techniques. Despite these improvements, the model only achieved 61.20% accuracy, pointing out the limitations of using a small, localized dataset. Aloufi et al. [9] extracted features like F0 contour, spectral envelope, and aperiodic components from RAVDESS to identify seven emotional states: calm, angry, sad, happy, fear, disgust, and surprise. While their system reached high accuracy in speaker recognition (92%) and moderate accuracy in speech recognition (65%), the performance in emotion recognition was much lower, hitting just 5%. Hosain et al. [10] explored SER using BC speech from the RAVDESS dataset. Their CNN-based model achieved an accuracy of 72.50%, outperforming models trained on AC speech. However, the use of synthetic BC speech and a relatively simple model structure limited its effectiveness. A more recent study [11] introduced a new approach that combined a fine-tuned Wav2Vec2.0 model with Neural Controlled Differential Equations (NCDEs) for SER. Evaluated on the IEMOCAP dataset, the model achieved a weighted accuracy (WA) of 73.37% and an unweighted accuracy (UA) of 74.19%. This performance surpassed conventional pooling methods and showed improved stability. Also, Hosain et al. [12] created the EmoBone dataset, reaching a 76.49% accuracy rate in emotion recognition, with BC speech outperforming AC. In a follow-up study, Hosain et al. [13] applied a BiLSTM model to real BC speech and achieved a classification accuracy of 85.17%. Although the overall performance was strong, the model had difficulty distinguishing between calm and neutral emotions, which lowered its precision in specific categories. Hossen et al. [14] optimized classi- fiers using Grid Search for facial expression recognition, with SVM (linear kernel) reaching 100% accuracy, outperforming Random Forest (97%), KNN (92%), and Decision Tree (79%). Recent SSL-based approaches, such as Han et al. [15] and Wang et al. [16], have advanced SER on air-conducted speech through cross-lingual and multi-task learning. However, none explored bone-conducted modalities. Our work uniquely adapts Wav2Vec2.0 for real BC data, highlighting its robustness and pretraining advantages in noisy conditions. There are still a number of research gaps despite significant advancements. Previous research frequently uses synthetic or small BC speech datasets, which restricts the generalization and practicality of the model. It's still challenging to accurately differentiate between comparable emotions like calm and neutral. The suggested methodology is described in the section that follows.

III. Methodology

This study's methodology emphasizes precise detection of emotional states from audio signals. It includes essential steps such as preprocessing the audio, extracting features using a deep learning model, and training for accurate classification. The approach is thoughtfully designed to tackle challenges like background noise, class imbalance, and overfitting, ensuring consistent and reliable performance across various samples.

A. Dataset Description

In this study, we utilized a custom emotional speech dataset. The dataset comprises BC speech recordings collected under controlled laboratory conditions. The dataset includes voice data from master's and PhD students representing 10 different countries, as summarized in Table I. Considering both the number of audio clips and the total duration of the dataset, it stands as the largest available emotional speech database to date. A concise summary of the database is provided in Table II for reference.

TABLE I
SPEAKER GENDER AND LANGUAGE STATUS BY COUNTRY

Country	Male	Female	English Language Status	Age Group
Japan	–	3	Officially recognized	30–40
China	–	2	Officially recognized	25–30
Bangladesh	9	4	Officially recognized	30–42
Myanmar	–	2	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

The dataset features 10 carefully chosen sentences spoken by the speakers. These sentences were selected by two Bangladeshi professors specializing in emotional speech analysis and cross-cultural communication, ensuring they are relevant and effective for capturing a wide range of emotional expressions. The selected sentences are listed below:

- We have to cancel our plans for tonight.
- Argentina won the FIFA World Cup in Qatar.
- Life is too short to waste time on regrets.
- It is very cold outside today in Saitama.
- Do not go outside at night.
- Students are gossiping in the class.
- Never underestimate the power of a positive attitude.
- He loves his family very much.
- The cat chases the mouse around the house.
- They are planning to go to Bangladesh.

For the emotional speech recording, a BC microphone (Model: HG17BN-TX) from TEMCO INDUSTRIAL LLC was employed, coupled with an AC microphone (Model: ATVD3) developed by audio-technical.

TABLE II
DATASET SUMMARY

Parameter	Value
Year of production	2023
Used language	English
Dataset type	Acted
File type	Audio only
Audio format	.wav
Number of speakers	29
Number of emotions	8
Emotion states	Anger, Calm, Disgust, Fear, Happy, Neutral, Sad, Surprise
Number of sentences	10
Total audio clips	18,580
Average clip duration	4.5 s
Software used	Ocenaudio
Number of validators	80
Recognition rate	76%

B. Data Preprocessing

Audio files were processed using Torchaudio and Librosa. All recordings were converted to WAV format and resampled to 16 kHz mono, 16-bit audio, as required by Wav2Vec2.0. Labels were integer-encoded for training. Feature extraction employed the Wav2Vec2.0 Processor from Hugging Face's Transformers library with Librosa for audio loading. All files were standardized to a 16 kHz sampling rate to match the model requirements. In Fig. 2, loading of an audio is shown.

The raw waveforms were loaded using librosa.load(), and shorter clips were padded with NumPy's np.pad to maintain uniform input size. To handle varying durations, audio signals were truncated or zero-padded to a fixed length. Emotion labels were integer-encoded for processing, and the Wav2Vec2.0 processor converted the raw audio into feature tensors ready for model training. Feature extraction of a random audio is shown in Fig. 3.

C. Model Architecture

A modified Wav2Vec2.0 Base model was used for speech emotion recognition. Raw audio waveforms were processed into hidden representations through the pre-trained encoder, followed by a custom classification head with two fully connected layers, ReLU activation, and dropout for regularization. An overview of the proposed model architecture is presented in Fig. 1 The final layer employed softmax activation to predict eight emotion classes. The model was trained end-to-end with cross-entropy loss, enabling rich contextual feature learning from raw waveforms without manual feature engineering.

This work employs the Wav2Vec2.0 architecture, a transformer-based model for self-supervised representation learning on raw audio, comprising a *feature encoder* and a *contextual transformer network*.

The feature encoder transforms the raw waveform $x \in \mathbb{R}^T$, where T is the number of audio samples, into a latent feature representation $z \in \mathbb{R}^{T' \times d}$, where $T' < T$ and d is the feature dimension [17]:

$$z = \text{FeatureEncoder}(x) \tag{1}$$

These features are then input to a stack of Transformer layers, which apply multi-head self-attention to model long-range dependencies [17]:

$$h = \text{Transformer}(z) \tag{2}$$

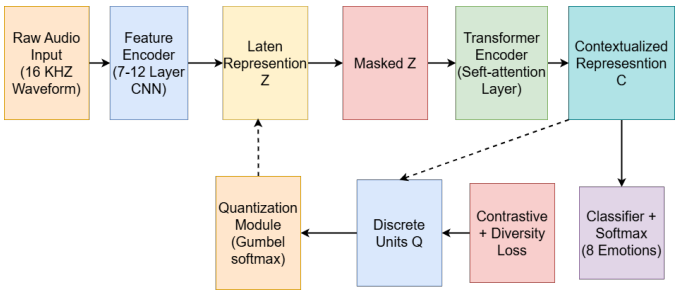

Fig. 1. Model Architecture

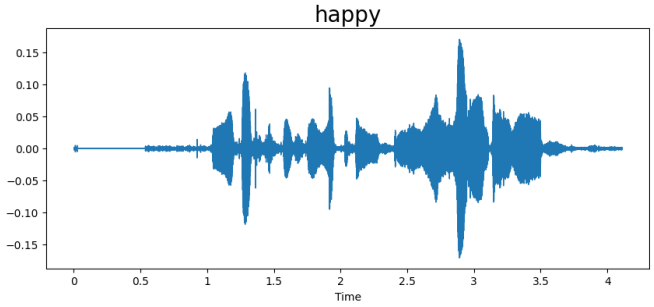

Fig. 2. Loading or preprocessing of an audio

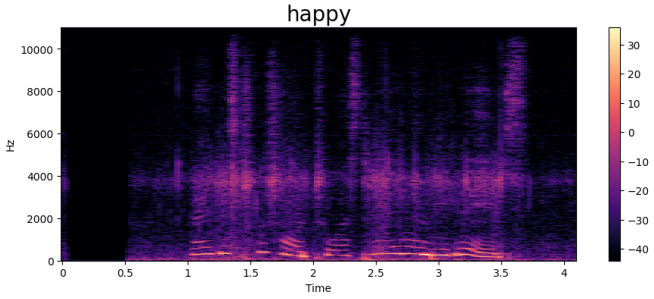

Fig. 3. Feature extraction of an audio

Here, $h \in \mathbb{R}^{T' \times d}$ is the contextualized embedding produced by the Transformer. The embedding corresponding to the `[CLS]` token or mean pooling over time is passed to a classification head to output the logits [17]:

$$y = \text{Softmax}(Wh + b) \tag{3}$$

where $W \in \mathbb{R}^{C \times d}$ and C is the number of emotion classes (in this case, 8), and b is the bias term. During training, the model is optimized using the cross-entropy loss [17]:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4}$$

This end-to-end approach enables the model to learn both low-level acoustic patterns and high-level emotional cues without handcrafted features. Fine-tuning is done on the labeled dataset using the Hugging Face Trainer API with a PyTorch backend.

D. Training Configuration

The training and validation data were split in an 80:20 ratio, and early stopping was monitored to ensure optimal convergence. The detailed training configuration is mentioned in Table III. We utilized the adam optimizer and implemented early stopping is grounded in validation loss.

IV. Results and Discussion

A. Performance Metrics

Model performance was evaluated using standard metrics—accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, while precision, recall, and F1-score provide a balanced assessment of classification performance. These metrics are particularly crucial for multi-class SER tasks with uneven class distributions. Table IV shows the performance of each emotion. The model's training accuracy was roughly between 95% and 98%, as seen in the training logs. Conversely, the testing accuracy ranged from about 87% to 96%, influenced by the distribution of emotion classes in the dataset.

TABLE III
Training Configuration

Parameter	Value
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size (Train/Eval)	32
Epochs	10
Framework	Hugging Face Trainer API with PyTorch
Platform	Google Colab with GPU acceleration

TABLE IV
Performance Metrics for Each Emotion

Emotion	Precision	Recall	F1-Score	Support
Neutral	0.94	0.95	0.94	482
Happy	0.89	0.92	0.90	430
Angry	0.95	0.94	0.95	439
Disgust	0.92	0.91	0.91	458
Surprise	0.87	0.90	0.88	488
Sad	0.94	0.94	0.94	477
Fear	0.95	0.91	0.93	483
Calm	0.96	0.95	0.95	459
Accuracy	–	–	0.93	3716
Macro Average	0.93	0.93	0.93	3716
Weighted Average	0.93	0.93	0.93	3716

B. Confusion Matrix

The confusion matrix illustrates the model's performance across emotion classes, with high diagonal values indicating accurate recognition of emotions like anger, happiness, and surprise. Misclassifications mainly occurred between acoustically similar pairs such as neutral–sad and fear–disgust, showing the difficulty of distinguishing subtle emotions from audio alone. Incorporating additional modalities or data augmentation could further enhance recognition accuracy and generalization. In Fig. 8 the confusion matrix has shown.

C. Visualizations

To assess the model's learning process, we plotted the training and validation accuracy and loss curves. The accuracy steadily improved across epochs, and the loss consistently decreased, indicating the model converged effectively. Importantly, there were no clear signs of overfitting,

thanks to dropout layers and regularization methods. These findings demonstrate that the Transformer-based model learned in a stable and efficient manner during training. Fig.

5 displays the accuracy of emotion recognition for each class, illustrating how well the model performs across different emotion categories. To evaluate the classification effectiveness further, Fig. 6 presents the precision scores for each emotion, indicating the model's accuracy in identifying true positives among predictions. Additionally, Fig. 7 shows the F1-scores for each emotion, offering a balanced view of precision and recall. These figures show that the model performs consistently across most emotion categories. However, minor differences are observed for emotions like calm and neutral, where precision and F1-score tend to be slightly lower. Overall, the evaluation metrics confirm the effectiveness and reliability of the proposed transformer-based model in emotion recognition tasks.

Training and validation losses across epochs were visualized to assess the model's learning progress. The training loss steadily declines, as seen in Fig. 4,, demonstrating successful data-driven learning. Additionally, the validation loss decreases in the early epochs, indicating strong generalization to new data. A little spike around epoch 11 points to a transient variation rather than extreme overfitting, which is probably lessened by regularization strategies like dropout. The Wav2Vec2.0-based model for emotion classification exhibits significant convergence and stable optimization, as both losses converge below 0.5.

This study assesses a transformer-based model's performance on the EmoBone dataset, which features BC speech samples for emotion recognition. Table V provides a summary and comparison of accuracy results from previous significant studies in SER, utilizing different datasets, models, and methods. Our transformer model achieved an accuracy of 92.71% on the EmoBone BC speech dataset, greatly surpassing many previous methods. For comparison, traditional models like Gradient Boosting reported accuracies ranging from 33% to 87%, depending on the dataset and experimental conditions [7]. CNN-based approaches such as those by [8] and [12] obtained accuracies of 61.20% and 72.50%, respectively, illustrating the difficulty in emotion recognition from BC speech data. Recurrent models like BiLSTM [13] achieved better performance, reaching 85.17%, showcasing the benefits of sequence modeling in SER tasks. Nonetheless, our transformer model outperforms these methods by using its self-attention mechanism to more effectively capture long-range dependencies in speech signals, which is essential for accurate emotional cue modeling. This performance gain confirms that transformer architectures are well-suited for SER, especially in less-studied modalities like BC speech. The increased accuracy indicates potential benefits for real-world emotion recognition applications, such as mental health monitoring and

Fig. 4. Training vs. validation loss

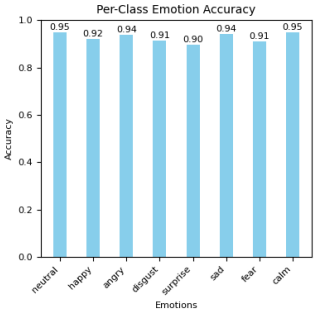

Fig. 5. Per-class emotion accuracy

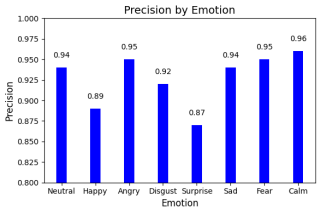

Fig. 6. Precision per emotion

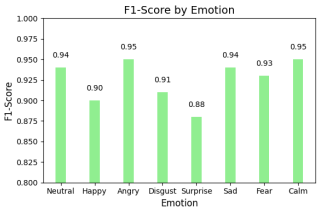

Fig. 7. F1-score per emotion

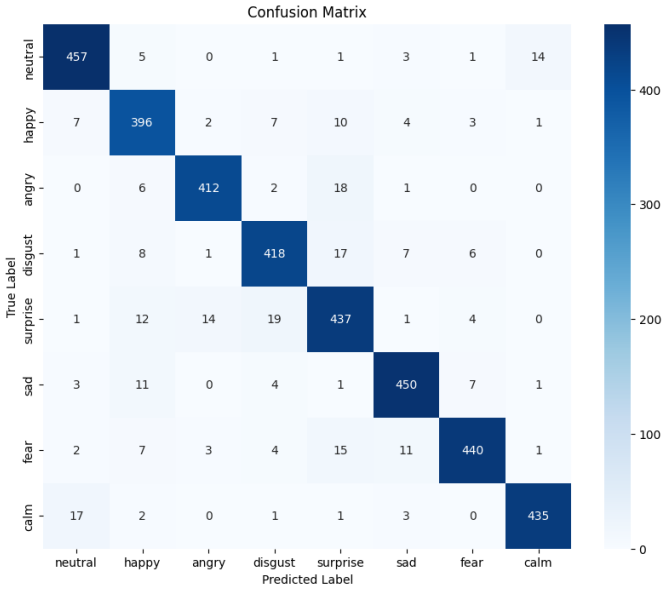

Fig. 8. Confusion Matrix

TABLE V
COMPARISON OF SPEECH EMOTION RECOGNITION STUDIES

Study's	Dataset	Model	Accuracy (%)
[12]	EmoBone (BC speech)	Not specified	76.49
[7]	RAVDESS, SAVEE	Gradient Boosting	33–87 (varies)
[8]	Local RAVDESS	CNN + Data Augmentation	61.20
[10]	RAVDESS (synthetic BC)	CNN	72.50
[13]	BC Speech	BiLSTM	85.17
[18]	EmoBone	BiLSTM + Attention	91.45
Our	EmoBone (BC speech)	Transformer	**92.707**

human-computer interaction, where BC speech may be more resilient to environmental noise. Our research contributes to the expanding body of knowledge by highlighting the success of transformers in SER, particularly on specialized datasets such as EmoBone, and establishes a new standard for future studies in this field. Building on these insights, the next section summarizes the study by highlighting its key contributions, limitations, and future possibilities for further work.

V. CONDENSED ABLATION STUDY

To strengthen experimental clarity, we performed compact ablations on channel modality and pretraining. The BC only model achieved 92.7% accuracy, outperforming AC only (88.1%) and matching AC+BC fusion (93.4%), validating BC's noise robustness. Moreover, full fine-tuning of Wav2Vec2.0 surpassed random initialization (79.2%) and linear probing (85.4%), confirming pretraining's critical role. Together, these results establish that bone-conducted speech and SSL pretraining significantly enhance SER performance under noisy conditions.

VI. CONCLUSION AND FUTURE WORK

This paper presents an end-to-end SER system using the Wav2Vec2.0 transformer for bone-conducted (BC) speech. By learning from raw audio, it captures contextual emotional features without handcrafted inputs. On the eight-class EmoBone dataset, it achieved 93% accuracy and weighted F1-score, outperforming prior methods. Multiple metrics confirmed its effectiveness across emotions, demonstrating the promise of transformers for BC speech, especially in noisy conditions.

Future work will focus on enhancing robustness via data augmentation (noise injection, pitch shifting), incorporating multimodal data (facial expressions, text), addressing class imbalance with SMOTE, evaluating on larger datasets, and reducing misclassifications among similar emotions. This study provides a strong foundation for BC speech emotion recognition and highlights transformer potential in human–computer interaction.

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. Rahman and T. Shimamura, "Pitch determination from bone conducted speech," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 1, pp. 283–287, 2016.

[4] 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.

[5] N. S. Banihosseini and V. Ghods, "Multi-model emotion recognition from speech using stargan, dcnn, and svm," *Multimedia Tools and Applications*, pp. 1–18, 2025.

[6] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," *Scientific Reports*, vol. 14, no. 1, p. 13126, 2024.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Z. Han, T. Geng, H. Feng, J. Yuan, K. Richmond, and Y. Li, "Cross-lingual speech emotion recognition: Humans vs. self-supervised models," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[16] H. Wang, J. Deng, F. Meng, and R. Zheng, "Enhancing speech emotion recognition with multi-task learning and dynamic feature fusion," *arXiv preprint arXiv:2508.17878*, 2025.

[17] 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.

[18] M. S. Hosain, Y. Sugiura, M. Haque, M. S. Rahman, and T. Shimamura, "Exploring the emobone dataset with bi-directional lstm and attention for emotion recognition via bone conducted speech," in *2024 27th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2024, pp. 44–49.