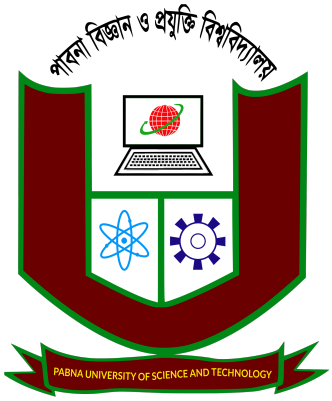
Speech Emotion Recognition from Bone-Conducted Speech Using Wav2Vec2 Transformer Model



Thesis

Course Code: ICE - 4210

B.Sc. (Engineering) Examination - 2023

A thesis paper submitted to the Department of Information and Communication Engineering, Pabna University of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Engineering in Information and Communication Engineering

Submitted by:

Manik Kumar Saha

Roll No: 200630

Reg. No: 1065395

Session: 2019-2020

Supervisor: Dr. Md. Sarwar Hosain

Department of Information and Communication Engineering

Pabna University of Science and Technology

Pabna-6600, Bangladesh

July 2025

Declaration

This is to certify that the thesis entitled **"Speech Emotion Recognition from Bone-Conducted Speech Using Wav2Vec2 Transformer Model"**, has been carried out by me under the course entitled **"Thesis/Project (ICE-4210)"**. I also declare that this work has not been submitted elsewhere in part or full for the requirement of any degree or for any other purpose, except for academic publication.

.....

Manik Kumar Saha
Roll No: 200630
Reg. No: 1065395
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

Certificate of the supervisor

It is certified that the research work incorporated in this thesis entitled **"Speech Emotion Recognition from Bone-Conducted Speech Using Wav2Vec2 Transformer Model"** is the original work carried out by **Mr. Manik Kumar Saha** under my supervision, and it fulfills the conditions laid out in the Pabna University of Science and Technology ordinances. The research work included in this thesis forms a distinct contribution to knowledge. The thesis contains work worthy of consideration for the award of the degree of Bachelor of Science in Engineering in Information and Communication Engineering (ICE).

.....

Dr. Md. Sarwar Hosain
Associate Professor
Department of Information and Communication Engineering
Pabna University of Science and Technology, Pabna-6600, Bangladesh.

Dedicated to...

My parents

Acknowledgements

Firstly, I would like to express my heartfelt gratitude to the divine blessings of Almighty God for guiding me throughout my research work and enabling me to successfully complete my B.Sc (Engineering).

I would like to express my heartfelt gratitude to Dr. Md. Sarwar Hosain, my thesis supervisor, for his unwavering support, encouragement, kindness, and insightful guidance. His balanced approach granting me the freedom to explore while providing timely direction greatly contributed to the successful completion of my research. I didn't had a huge knowledge about the thesis, but my supervisor had taught me everything about the thesis. He guides me for doing this work from scratch. I could not have asked for a more supportive and inspiring mentor like him.

I am also thankful to all the respected teachers and staff of the Department of Information and Communication Engineering at Pabna University of Science and Technology for their cooperation, support, and the knowledge they have imparted to me during my undergraduate studies.

Finally, I would like to express my deep gratitude to my parents. Their hard working passion inspired me to do my study more industriously and effectively. I am also thankful to my friends, seniors and juniors for their constant support, patience, and motivation throughout this academic journey. Their encouragement has been a pillar of strength during challenging times.

Abstract

Speech emotion recognition (SER) is an important area of affective computing that helps machines to understand human emotions through speech signals. This technology significantly improves human computer interactions (HCI), intelligent virtual assistants, mental health monitoring, and systems that recognize emotions. While traditional SER systems mainly focus on air-conducted (AC) speech, their performance drops in noisy or real world settings. Bone-conducted (BC) speech, which travels through cranial bones instead of air, offers a noise resistant option suitable for these conditions. However, BC speech is still not widely used in SER research due to the lack of publicly available datasets and the challenges of accurately modeling emotions.

This thesis introduces an end-to-end SER framework based on the Wav2Vec2.0 transformer model. It was trained and evaluated using EmoBone, a diverse BC speech dataset that includes eight emotional classes. Unlike traditional methods that rely on manually crafted features, the proposed model processes raw audio waveforms directly to extract contextual representations, improving both robustness and generalization. A customized classification head is included to allow for quick and accurate emotion prediction. The system achieved a weighted F1-score and overall accuracy of 93%, surpassing previous top models applied to the EmoBone dataset.

To ensure thorough evaluation, several analyses were conducted, including confusion matrix interpretation, precision-recall metrics, and class-wise F1-score visualization. These results show that the model performs well in detecting various emotional states, although there are minor misclassifications between acoustically similar emotions, like calm and neutral. Future research will focus on addressing these issues through data augmentation, class balancing, and combining data from visual or textual inputs. This study demonstrates the practicality and effectiveness of transformer based models for SER tasks that involve BC speech, especially in challenging acoustic environments. The research not only provides a new methodological approach but also lays the groundwork for future advancements in emotion-aware HCI technologies and speech based systems that are culturally inclusive.

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

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, Speech emotion recognition (SER) has become an important area in affective computing. It allows machines to understand human emotions through vocal signals. As the demand for emotionally smart systems grows, applications like human computer interaction (HCI), mental health monitoring, and assistive technologies have seen major improvements in SER. Most research has concentrated on air conducted (AC) speech, but bone conducted (BC) speech offers a promising alternative, especially in noisy or real-life situations. This chapter outlines the motivation, background, goals, scope, and contributions of this study. It examines an end-to-end SER framework that uses transformer-based models and bone-conducted emotional speech data.

1.1 Research motivation and significance

The field of SER is at the forefront of innovations in HCI, offering the potential to transform how machines perceive and respond to human emotional expressions. As technology becomes more deeply embedded in our everyday lives, it is increasingly essential for intelligent systems to accurately detect and interpret emotional cues in speech. This capability has wide ranging applications, from enhancing user engagement in virtual assistants and customer service platforms to enabling early intervention in mental health monitoring. Despite notable advancements, SER models continue to face challenges in achieving robust accuracy and generalizability particularly when applied across linguistically and culturally diverse populations.

The classical automatic speech recognition (ASR) system focuses only on the linguistic information of the speech. It converts an input audio stream to corresponding text contents but does not identify the background emotion of the speaker. The main

goal of building a SER system is to combine it with a traditional ASR in order to create an interface for human-machine interaction that feels natural. Although efforts to create a reliable SER system have been underway over the past 20 years, only recently has research focused on this topic surfaced [1].

Traditional AC speech, commonly used in SER, has its limitations. AC speech often fails to capture the full spectrum of emotional cues, particularly those found in the low-frequency components of speech [2]. BC speech, on the other hand, transmits sound directly via the bones of the skull to the inner ear, preserving these critical low-frequency elements [3]. This unique characteristic of BC speech offers a promising avenue for enhancing emotion recognition accuracy.

Furthermore, the lack of comprehensive, culturally diverse datasets hampers the development of robust SER models. Existing datasets do not always include all the different ways people show emotion in different cultures, which makes it harder for SER systems to be used all over the world [4]. This gap in resources calls for the creation of extensive, high-quality datasets that encompass a broad spectrum of emotional speech from diverse populations.

My motivation for this research comes from a strong desire to tackle the limitations of traditional SER systems, especially their lower performance in noisy settings. BC speech is resistant to external noise and offers a unique opportunity to improve SER accuracy and reliability. Using the unique features of BC speech, we can develop stronger and more inclusive emotional speech processing systems.

This thesis is based on the EmoBone dataset, a collection of emotional BC speech from various countries. By fine-tuning a deep learning-based SER model with this dataset, I hope to improve the accuracy and generalization of emotion recognition systems in different cultural and acoustic situations. Using Wav2Vec2.0, a powerful transformer model, allows end-to-end learning directly from raw audio, eliminating the need for manual feature engineering.

The potential impact of this work is significant. Better SER models can improve HCI, support emotion aware virtual assistants, assist in mental health monitoring, and help people with speech impairments communicate. Furthermore, the EmoBone dataset promotes cross-cultural emotion research and helps create more empathetic and globally inclusive SER technologies. Through this thesis, I aim to make a meaningful contribution to the development of noise-resilient and culturally aware emotion recognition systems.

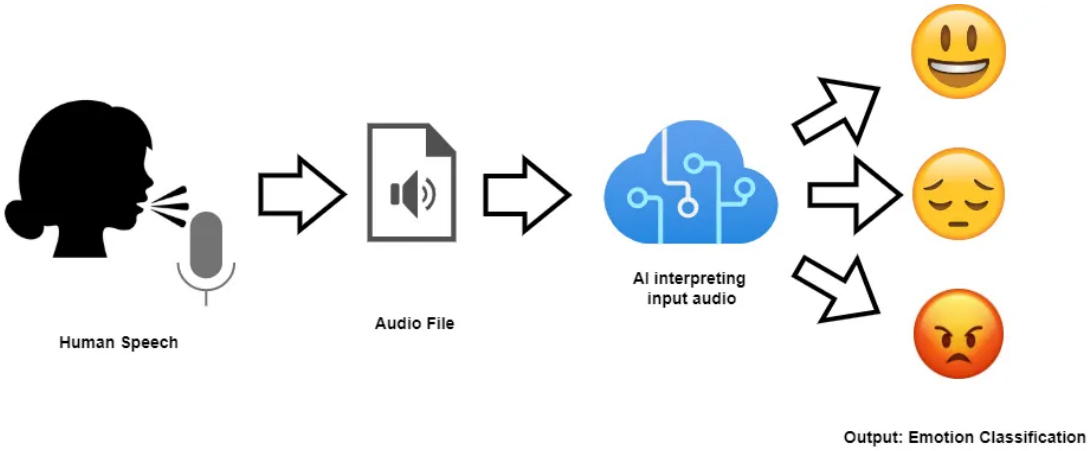


Figure 1.1: Speech emotion recognition system

1.2 Problem statement

Most current SER systems are trained on clean, AC speech data. As a result, they struggle in real-world, noisy conditions. While BC speech can help reduce noise interference, it is not commonly used in SER systems. Furthermore, many models depend on features created by hand, which need expert knowledge and may not work well across different datasets or languages.

We need a data driven, end-to-end SER framework that can learn directly from BC speech. This will help ensure high performance in various noisy situations.

1.3 Objectives

This thesis aims to:

- Explore the effectiveness of BC speech for emotion recognition.
- Develop an end-to-end SER framework using the Wav2Vec2.0 transformer model.
- Train and fine-tune the model on a multi-national, BC speech emotion dataset.
- Evaluate model performance across multiple emotion categories using standard metrics such as accuracy and F1-score.
- Identify limitations and recommend future directions for improvement.

1.4 Scope of the study

This study focuses exclusively on speech-based emotion recognition using BC speech signals. It does not explore other modalities such as facial expressions, text, or physiological signals, nor does it implement multimodal fusion. The main objective is to develop and evaluate a deep learning-based SER system using the Wav2Vec2.0 transformer model, trained on a BC emotional speech dataset (EmoBone) [5].

The study is limited to offline experiments using pre-recorded audio samples and does not include real-time system deployment. Eight basic emotions are considered in this work, following standard emotional classification approaches used in prior SER research [6, 7]. The model is evaluated using common performance metrics such as accuracy and F1-score. While the focus is on cross-cultural generalization, the dataset used still has inherent limitations in speaker diversity and sample volume.

This research aims to demonstrate the feasibility and benefits of using BC speech in emotion recognition systems and to serve as a foundation for future work involving real-time systems, larger multilingual datasets, and multimodal SER integration.

1.5 Contributions

The main contributions of this paper are as follows:

- We present an end-to-end SER framework based on the pre-trained model Wav2Vec2.0 transformer model and no handcrafted acoustic features are required.
- 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 discuss possible causes and improvements in the future.

1.6 Organization of the dissertation

This thesis is organized into seven chapters, each contributing to the development and understanding of SER using BC speech and deep learning models:

- **Chapter 1 – Introduction:** Provides the motivation, problem statement, objectives, scope, and contributions of the study, along with the organization of the thesis.
- **Chapter 2 – Background study:** Presents the fundamental concepts related to speech processing, emotional speech recognition, BC speech, and deep learning models, especially transformers like Wav2Vec2.0.
- **Chapter 3 – Literature review:** Reviews existing research in the fields of speech emotion recognition, datasets, conventional models, and modern deep learning techniques used in SER.
- **Chapter 4 – Methodology:** Describes the dataset used, preprocessing techniques, model architecture (including Wav2Vec2.0), training procedures, and the evaluation framework.
- **Chapter 5 – Experiments and results:** Details the experimental setup, performance evaluation, confusion matrix analysis, and a comparison with existing methods.
- **Chapter 6 – Conclusion and future work:** Summarizes the key findings of the research, acknowledges limitations, and proposes directions for future studies to further enhance speech emotion recognition.

Chapter 2

Background study

Establishing a solid foundation for any research requires a clear understanding of the history and development of the field. This chapter provides a comprehensive overview by exploring the historical background, the theoretical framework, and recent advances in the area. This approach situates the current study within its broader context. Initially, the research centered on utilizing artificial BC speech to recognize emotions and evaluate its performance relative to traditional air-conducted AC speech.

2.1 Fundamentals of speech emotion recognition

SER refers to the computational techniques that aim to identify and classify human emotions from spoken language. It enables machines to understand affective states such as happiness, sadness, anger, or fear, thereby enriching interactions in domains such as virtual assistants, mental health monitoring, and affective computing [6, 7].

2.1.1 Emotion

Emotion is a complex psychological and physiological response to internal or external stimuli, often accompanied by expressive behaviors such as vocal tone, facial expressions, and gestures. From a computational perspective, emotions can be categorized using either:

- **Categorical models** — which classify emotions into discrete classes such as happy, sad, angry, and fearful [7].
- **Dimensional models** — which represent emotions in a continuous space defined by axes such as arousal (activation level) and valence (positivity/negativity) [6].

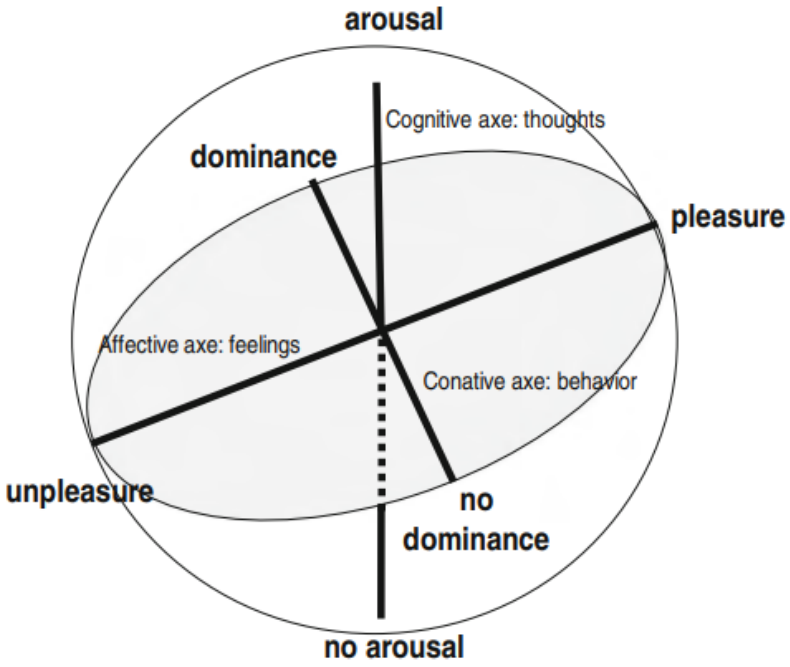


Figure 2.1: Dimensional model of PAD [8]

The PAD space, which includes the third dominance axis [9], is better than the traditional two-dimensional model for understanding the three emotional dimensions and how they relate to other theories in the field, as shown in Figure 2.1 [8].

In SER, the vocal expression of emotion is of particular interest. Features such as pitch, energy, speaking rate, and spectral characteristics provide clues to the speaker's emotional state. For instance, anger often correlates with higher pitch and intensity, while sadness tends to manifest in slower speech with lower energy.

2.1.2 Importance of emotion in SER

Understanding emotions from speech plays a key role in HCI. Emotions provide contextual cues that enhance communication, empathy, and responsiveness in applications ranging from intelligent personal assistants to therapeutic agents. Integrating emotional awareness allows machines to react more naturally and supportively, improving user satisfaction and engagement.

2.1.3 SER

SER refers to the computational process of identifying human emotional states from speech signals. It leverages various acoustic, prosodic, and spectral features extracted from the spoken language to classify emotions such as anger, joy, sadness, fear, and more. Unlike speech recognition, which aims to convert spoken language into text, SER focuses on capturing the underlying emotional intent, which is often embedded in tone, pitch, energy, and speech rate [6].

The core workflow of SER systems generally includes preprocessing, feature extraction, feature selection, and emotion classification. Traditional approaches used handcrafted features like Mel Frequency Cepstral Coefficients (MFCCs), pitch, and zero-crossing rate, while recent advances apply deep learning models for automatic feature extraction [7]. These systems are evaluated based on classification accuracy, precision, recall, and F1-score, particularly important in imbalanced emotional datasets.

2.1.4 Importance in HCI

SER plays a pivotal role in advancing HCI by enabling machines to perceive and respond to users' emotional states. By integrating SER capabilities, systems can adapt their responses to better match user emotions, resulting in more natural and empathetic interactions. This is particularly relevant in applications such as intelligent voice assistants, customer service bots, e-learning platforms, and mental health monitoring systems [6].

Emotion aware systems enhance user experience by making interactions more responsive and context-aware. For instance, a virtual assistant detecting user frustration could modify its speech tone or provide more detailed help. Similarly, in healthcare settings, SER systems can help monitor emotional well being in patients, especially those with speech impairments.

2.1.5 Challenges in SER

Despite its promising potential, SER faces several technical and practical challenges:

- **Speaker and language variability:** Emotional expression varies significantly across individuals and cultures, making it difficult for SER systems to generalize across diverse speakers [7].
- **Data scarcity and labeling:** High-quality, annotated emotional speech datasets are limited, and manual labeling is subjective and resource-intensive [6].

- **Noise robustness:** Most emotional speech datasets are recorded in clean environments, but real-world applications often involve noisy or overlapping speech, which degrades performance.
- **Ambiguity in emotions:** Certain emotions such as “calm” and “neutral” exhibit overlapping acoustic features, making them difficult to distinguish reliably [10].
- **Modality limitation:** Audio-only emotion recognition lacks complementary cues from other modalities (e.g., facial expression), leading to lower accuracy in some contexts.

To address these limitations, recent research focuses on transformer-based models, multimodal fusion techniques, and culturally diverse datasets. The development of robust and scalable SER systems thus remains an active and evolving area of research.

2.2 Types of speech modalities

Speech signals can be transmitted through different physical mechanisms, commonly referred to as speech modalities. In the context of SER, the two primary modalities are AC speech and BC speech. Each modality has unique characteristics that affect the way emotional information is captured and processed by SER systems.

2.2.1 AC speech

AC speech is the most widely used and studied form of speech signal. In this modality, sound is generated by the vocal cords and travels through the air, eventually reaching the microphone via airborne pressure waves. Most conventional speech processing systems—including ASR and traditional SER models—are based on AC speech [6, 7].

AC speech captures the full frequency spectrum, allowing for a richer representation of emotional cues such as pitch, loudness, and rhythm. However, it is highly susceptible to environmental noise, microphone placement, and reverberation, which can degrade system performance especially in real world or noisy conditions.

2.2.2 BC speech

BC speech offers an alternative transmission pathway wherein vibrations generated by the vocal tract are conducted through the bones of the skull directly to the inner ear or

bone-conduction microphone, bypassing the outer and middle ear [10, 11]. This modality has gained attention in recent SER research due to its inherent robustness against background noise.

BC speech captures lower-frequency components more effectively and is less affected by ambient noise and speaker distance. It is especially useful in applications such as underwater communication, noisy environments (e.g., military or industrial settings), and assistive technologies for individuals with hearing impairments.

Despite its noise resistance, BC speech suffers from reduced intelligibility and narrower frequency range compared to AC speech, which poses challenges for emotion recognition tasks relying on high-frequency emotional cues.

2.2.3 Comparison of AC and BC speech in SER

Recent studies have shown that although BC speech lacks certain high-frequency components, it can still preserve essential emotional cues—especially when combined with advanced deep learning models such as CNNs, LSTMs, or Transformers [10,12]. Therefore, BC speech provides a promising avenue for building robust and noise-resilient SER systems, particularly in challenging environments. Table 2.1 summarizes the key differences between AC and BC speech modalities in the context of SER.

Table 2.1: Comparison of AC and BC speech for SER

Feature	**AC speech**	**BC speech**
Transmission medium	Air (external sound waves)	Bone (vibrations through skull)
Noise sensitivity	High	Low
Speech intelligibility	High	Moderate
Frequency Range	Wide	Narrow (mostly low-mid frequency)
Emotion feature richness	Richer cues (pitch, energy, timbre)	Fewer high-frequency cues
Use case suitability	Quiet environments, general applications	Noisy conditions, assistive devices
Recent use in SER	Well-established	Emerging, gaining interest

2.3 Emotional speech representation

Accurate representation of emotional cues in speech is essential for the success of SER systems. The emotional content of speech is conveyed through variations in prosody, pitch, energy, and spectral properties. Extracting and representing these features in a machine-understandable form is a critical step in any SER pipeline.

2.3.1 Acoustic features

MFCCs are among the most widely used acoustic features in SER. Derived from the power spectrum of an audio signal, MFCCs simulate the human auditory system by emphasizing perceptually important frequencies. They capture the timbral texture of speech and are effective in identifying emotional differences across utterances [7].

In addition to MFCCs, pitch and energy are frequently extracted features. Pitch relates to the fundamental frequency of speech and often varies with emotions; for example, anger and happiness are typically associated with higher pitch, while sadness corresponds to lower pitch. Energy reflects the loudness of the speech signal and is also an emotion-indicative feature. Together, these features offer a comprehensive representation of emotional characteristics in audio signals [13].

2.3.2 Spectrograms and audio representations

Spectrograms represent speech signals in the time-frequency domain and offer a visual pattern of frequency changes over time. They are especially useful in deep learning-based SER, where convolutional networks learn emotional patterns directly from visual cues. Log-mel spectrograms and chromagrams are commonly used variations. These time-frequency representations provide rich contextual and harmonic information that complements traditional acoustic features [14].

Spectrograms make it easier to identify non-linear variations in pitch, intensity, and

formant structure—elements that are often challenging to model using only statistical descriptors. Thus, they serve as a bridge between raw audio and meaningful feature extraction in deep learning approaches.

2.3.3 Feature extraction techniques

Depending on the model design, feature extraction may involve manual, statistical, or learned techniques. Traditional methods use tools such as open SMILE or praat to extract hand-engineered features, including MFCCs, zero-crossing rate, formants, jitter, and shimmer. These tools output fixed-dimensional feature vectors that are then fed into classifiers such as Support Vector Machines (SVMs) or Random forests.

In contrast, modern deep learning approaches often bypass manual extraction. End-to-end systems such as CNNs and Wav2Vec2.0 operate directly on raw waveforms or spectrograms, automatically learning hierarchical representations. This strategy enhances generalization and reduces reliance on domain expertise for feature design [15].

2.4 Traditional approaches in SER

Speech Emotion Recognition (SER) has its roots in classical machine learning, where models were built using engineered features extracted from speech signals. These approaches dominated early SER systems before the rise of deep learning.

2.4.1 Machine learning algorithms

Traditional SER systems often relied on machine learning classifiers such as:

- **Support vector machines (SVM):** Effective in high-dimensional spaces and widely used for binary and multiclass emotion classification tasks.

- **Hidden markov models (HMM):** Capable of modeling temporal dependencies and widely applied in speech and emotion sequence modeling.
- **Gaussian mixture models (GMM):** Useful for modeling the distribution of acoustic features for each emotional class.
- **K-nearest neighbors (KNN) and decision trees:** Simpler methods applied to smaller-scale SER tasks.

These models typically operate on feature vectors extracted from segmented speech frames, and their performance is highly dependent on the quality of the input features.

2.4.2 Handcrafted features and their limitations

Early SER systems primarily depended on manually engineered acoustic features derived from:

- **Prosodic features:** pitch, energy, speaking rate.
- **Spectral features:** MFCCs, Linear Predictive Coding (LPC), formants.
- **Voice quality:** jitter, shimmer, harmonics-to-noise ratio (HNR).

While these features are intuitive and interpretable, handcrafted approaches face several limitations:

1. They are not invariant to speaker variability, accent, and recording conditions.
2. Feature selection is often heuristic and domain-specific, requiring expert knowledge.
3. These methods struggle to capture the complex and nonlinear patterns present in emotional speech.

Due to these shortcomings, traditional methods have largely been replaced or enhanced by data-driven deep learning models that can learn relevant features directly from raw data [6, 7].

2.5 Deep learning in SER

With the limitations of traditional machine learning approaches in handling complex, variable, and high-dimensional emotional speech data, deep learning methods have gained significant traction in SER. These models learn hierarchical representations directly from raw or minimally processed data, eliminating the need for handcrafted features.

2.5.1 CNNs and RNNs for emotion recognition

Convolutional Neural Networks (CNNs) have been widely adopted in SER to extract local patterns and spatial features from spectrogram representations of speech. CNNs are particularly effective in capturing pitch variations, energy, and temporal structures across short speech segments [6].

Recurrent Neural Networks (RNNs), especially their variant long short term memory (LSTM) networks, are well-suited for modeling sequential data like speech. They capture temporal dependencies between frames, which is essential for recognizing emotions that unfold over time. LSTM networks preserve context information from previous time steps, making them valuable for tasks involving time series such as SER [14].

2.5.2 Attention mechanisms

While CNNs and RNNs offer strong modeling capabilities, they can struggle with long-range dependencies. Attention mechanisms address this by allowing the model to focus selectively on the most informative parts of the input sequence. The attention mech-

anism calculates a weighted sum of features, giving higher weights to frames that are emotionally salient.

This method improves interpretability and performance by enabling the network to learn which portions of speech are most relevant for emotion classification. Many recent SER models integrate attention with RNNs or LSTMs to enhance their discriminative ability [16].

2.5.3 BiLSTM and hybrid models

Bidirectional LSTM (BiLSTM) networks process input sequences in both forward and backward directions, capturing context from past and future frames. This is particularly advantageous for emotion recognition, as emotions can be influenced by both prior and upcoming speech patterns.

Hybrid models combining CNNs and BiLSTMs have shown strong performance in SER by leveraging spatial feature extraction and bidirectional temporal modeling. These models extract robust representations while maintaining sensitivity to emotional context over time. Their combination with attention mechanisms further improves recognition accuracy and model generalization across diverse datasets [17].

2.6 Transformer architectures for SER

Transformer-based architectures have revolutionized the field of sequence modeling by enabling efficient learning of contextual dependencies. In recent years, these models have been successfully applied to various speech-related tasks, including automatic speech recognition (ASR) and speech emotion recognition (SER). This section explores the transformer model's underlying structure, highlights Wav2Vec2.0 as a self-supervised model, and outlines its applicability in emotion recognition tasks.

2.6.1 Introduction to transformers

Transformers, initially proposed by Vaswani et al. [18], introduced a self-attention mechanism that allows models to capture long-range dependencies more effectively than traditional RNNs. Unlike RNNs or CNNs, transformers process entire input sequences in parallel, significantly improving computational efficiency and contextual understanding.

The core of the transformer model lies in its attention mechanism, particularly the multi-head self-attention layers that allow the model to focus on different parts of the input sequence simultaneously. This architecture has demonstrated remarkable performance in natural language processing (NLP) and has recently been adapted for audio-related tasks due to its strong temporal representation capabilities.

2.6.2 Wav2Vec2.0: A self supervised speech model

Wav2Vec2.0, developed by Baevski et al. [15], is a self-supervised learning framework designed for speech representation learning. Unlike supervised models that require labeled data, Wav2Vec2.0 learns latent speech representations directly from raw audio waveforms without transcriptions.

The model architecture consists of a multi-layer convolutional feature encoder followed by a transformer-based context network. During pre-training, the model masks parts of the audio input and learns to predict these masked segments based on the surrounding context, akin to BERT in NLP. Fine-tuning is then performed on downstream tasks like SER or ASR with limited labeled data.

Wav2Vec2.0's ability to learn rich, contextualized speech features from unlabeled data makes it highly suitable for emotion recognition, particularly when large labeled datasets are unavailable or expensive to collect.

2.6.3 Application of transformers in SER tasks

Transformer models like Wav2Vec2.0 have shown significant promise in emotion recognition from speech by capturing subtle variations in prosody, pitch, and temporal dynamics. Studies such as Hosain et al. [10] fine-tuned the Wav2Vec2.0 model on BC speech data and achieved substantial improvements in classification accuracy over conventional CNN and RNN-based models.

Compared to earlier architectures, transformer-based SER systems provide better generalization, robustness to noise, and scalability across multilingual and multi-accent datasets. Their ability to model long-range dependencies allows them to better interpret emotional states spread across time. Moreover, recent advancements have introduced hybrid models that combine transformers with BiLSTMs or attention mechanisms for enhanced contextual modeling [19].

Overall, the application of transformer architectures has set a new benchmark in SER performance, particularly in resource-constrained or noisy environments, such as those involving BC speech signals.

2.7 Speech emotion datasets

A key aspect of building reliable SER systems is the availability of high quality, labeled datasets. These datasets serve as benchmarks for training and evaluating the performance of machine learning and deep learning models. Various datasets have been developed over the years, capturing emotional speech in different languages, recording environments, and with varying emotion taxonomies. This section provides an overview of commonly used datasets in SER research and highlights their respective limitations.

2.7.1 Overview of commonly used datasets

- **RAVDESS (Ryerson audio visual database of emotional speech and song)**: A widely-used North American dataset containing audio and video recordings from 24 actors expressing eight different emotions. Emotions include calm, happy, sad, angry, fearful, surprise, disgust, and neutral [20]. The speech is acted and balanced across gender and emotion types.

- **EMO-DB (Berlin emotional speech database)**: This German-language dataset includes recordings from ten actors simulating seven emotional states: anger, boredom, disgust, fear, happiness, sadness, and neutral. It is frequently used in emotion recognition studies focused on classical machine learning models [28].

- **TESS (Toronto emotional speech set)**: Includes audio samples of two Canadian actresses reading a fixed set of sentences in seven emotional categories. The dataset is useful for studying speaker-dependent emotion variation.

- **SAVEE (Surrey audio visual expressed emotion)**: Consists of 480 British English utterances from four male actors covering seven emotions. It provides both visual and audio modalities [22].

- **EmoBone**: A relatively new and significant dataset focusing on BC speech emotion data. Created by Hosain et al., this multi-national dataset includes BC recordings in noisy and quiet environments, offering a practical alternative to AC speech for robust emotion recognition in real-world applications [23].

2.7.2 Limitations of existing datasets

Despite their widespread adoption, many existing SER datasets exhibit several limitations:

- **Acted vs natural speech**: Most datasets (e.g., RAVDESS, EMO-DB) feature acted emotions, which may not reflect the nuanced characteristics of spontaneous emotional speech encountered in real-life scenarios.
- **Language and cultural biases**: Datasets like EMO-DB and TESS are limited to specific languages and cultural contexts, reducing their generalizability across diverse populations.
- **Limited data volume**: Many datasets have relatively small sample sizes, restricting their usefulness for training deep learning models that require large amounts of data.
- **Lack of BC speech**: Traditional datasets primarily focus on AC speech, making them unsuitable for SER tasks in noisy environments. Datasets like EmoBone help bridge this gap but are still in early stages of adoption.
- **Imbalanced emotion classes**: Some datasets suffer from class imbalance, where certain emotions are overrepresented compared to others, leading to biased model performance.

As the field evolves, there is a growing need for more realistic, diverse, and multilingual datasets, particularly those incorporating BC signals to improve the robustness and generalization of SER systems in real-world applications.

2.8 Evaluation metrics in SER

Evaluating the performance of SER systems is essential for understanding how well the model identifies different emotional states across various categories. Given the complexity of emotional expression in speech and the class imbalance often present in emotion datasets, multiple evaluation metrics are used in SER to provide a comprehensive performance overview.

2.8.1 Accuracy, precision, recall, and f1-Score

- **Accuracy** measures the overall correctness of the model's predictions and is defined as the ratio of correctly classified samples to the total number of samples. While commonly used, accuracy can be misleading when the dataset is imbalanced.

- **Precision** refers to the proportion of correctly predicted positive observations to the total predicted positives. It is useful for evaluating how well the model avoids false positives.

- **Recall** or sensitivity, is the proportion of correctly predicted positive observations to all actual positives, indicating how well the model captures true positives.

- **F1-Score** is the harmonic mean of precision and recall, offering a single performance measure that balances both metrics. It is especially relevant in SER, where class imbalance (e.g., fewer samples of "fear" or "disgust") is common.

These metrics are computed per emotion class and averaged to evaluate multi-class classifiers. A high F1-score across all classes typically indicates a well-performing model in SER tasks [25].

2.8.2 Confusion matrix analysis

A confusion matrix provides detailed insights into the performance of an SER model by displaying true positives, false positives, true negatives, and false negatives for each emotion class. The diagonal elements of the matrix represent correctly classified emotions, while off-diagonal elements indicate misclassifications. Analyzing the matrix helps identify specific emotion pairs (e.g., "calm" vs. "neutral") that are commonly confused by the model [25].

2.8.3 Weighted vs unweighted metrics

- **Weighted metrics**: These consider the support (number of samples) for each class when calculating overall performance. Weighted f1-score, for example, reflects the proportion of each class in the dataset, making it suitable for imbalanced data.
- **Unweighted metrics**: Also known as macro averages, these compute metrics independently for each class and then average them, giving equal importance to all classes regardless of their frequency. Unweighted accuracy and f1-score are valuable when evaluating a model's performance on rare emotions.

Both weighted and unweighted evaluations are essential for a complete assessment of an SER model, particularly when the goal is to ensure robust performance across all emotion categories, not just the majority ones.

2.9 Summary

This chapter has presented a comprehensive overview of the foundational knowledge required to understand the research conducted in this thesis. It began by introducing the fundamentals of SER, emphasizing its significance in HCI and the major challenges faced in accurately identifying emotional cues from speech.

Next, the chapter examined various speech modalities, highlighting the differences between AC and BC speech, and their implications in emotion recognition tasks. The representation of emotional speech was discussed through acoustic features, spectrograms, and commonly used feature extraction techniques.

Traditional machine learning approaches were then reviewed, noting the limitations of handcrafted features. This was followed by a detailed discussion of deep learning based methods, including CNNs, RNNs, BiLSTM networks, and the use of attention mechanisms to enhance feature representation and sequence modeling in SER.

Special attention was given to recent transformer-based architectures such as Wav2Vec2.0, which offer state-of-the-art performance by learning from raw audio signals in a self-supervised manner. The chapter also provided an overview of benchmark SER datasets, such as RAVDESS, EMO-DB, and EmoBone, along with their limitations regarding diversity and generalizability.

Finally, evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices were explored, including the importance of both weighted and unweighted metrics in imbalanced data scenarios. Altogether, this background study establishes the theoretical and technical basis for the research methodology presented in the next chapter.

Chapter 3

Literature review

A comprehensive understanding of prior research is essential for framing the significance and originality of this study. The field of SER has evolved significantly over the past decades, transitioning from traditional machine learning techniques to sophisticated deep learning and transformer-based models. Various emotional speech datasets have supported this evolution; however, most of them are limited to air-conducted speech recorded in controlled environments. Moreover, while BC speech has gained attention in noise-resilient speech processing, its potential in emotion recognition remains largely underexplored. This chapter provides a structured review of the literature surrounding SER techniques, datasets, and the role of BC speech, ultimately identifying the research gaps this thesis aims to address.

3.1 Introduction

This chapter provides an overview of the existing literature related to SER, including traditional and modern approaches, the integration of deep learning models, and the evolution of transformer-based methods such as Wav2Vec2.0. The review also covers publicly available emotional speech datasets and highlights the underexplored area of BC speech in the context of SER. The chapter concludes by identifying research gaps

that motivate the present study.

3.2 Traditional approaches in SER

Earlier approaches to SER relied on handcrafted features and classical machine learning algorithms. Features such as MFCCs, pitch, formants, and energy were extracted from speech signals and used as input to classifiers like SVM, GMM, and HMM [6]. While these methods performed reasonably well in controlled environments, they lacked robustness against speaker variability, background noise, and emotional nuance.

3.3 Deep learning based approaches in SER

The advent of deep learning has significantly improved the performance of SER systems. Models such as CNNs and RNNs have been widely used to extract spatial and temporal features from audio spectrograms [7]. LSTM networks, a variant of RNNs, are particularly effective in modeling time-dependent emotional patterns in speech. These models automate feature extraction and generalize better across varying datasets. However, most studies in this area are still limited to AC speech.

3.4 Transformer based SER models

Transformer architectures have emerged as state-of-the-art in speech and natural language processing. Wav2Vec2.0, a transformer-based model introduced by facebook AI, uses self-supervised learning to generate rich, contextual embeddings from raw audio waveforms [15]. Pretrained on large unlabeled corpora, it can be fine-tuned for downstream tasks like emotion classification. Several studies have demonstrated its superior performance over CNN and RNN models in SER tasks, especially in noisy or multilingual conditions.

3.5 Datasets used in SER research

The effectiveness of SER models largely depends on the availability of high-quality emotional speech datasets. Commonly used datasets include:

- **RAVDESS**: North American English, acted emotions [20].
- **IEMOCAP**: Multimodal dataset with video, audio, and transcriptions [26].
- **CREMA-D**: Crowdsourced emotion recognition dataset with diverse speaker demographics [27].
- **EMO-DB**: German emotional database containing seven distinct emotions [28].

Despite their utility, these datasets are primarily recorded in clean environments and lack robustness against noise and cultural diversity. Most importantly, they do not include BC speech.

3.6 BC speech in emotion

BC speech in emotion or speech tasks BC speech has been explored mainly in the domains of secure communication, assistive hearing, and noise-robust speech recognition [10]. However, its application in emotion recognition remains limited. BC speech travels through the bones of the skull and is less susceptible to ambient noise, making it promising for real-world SER applications. The EmoBone dataset [10] addresses this gap by providing a diverse, multi-national collection of BC emotional speech.

3.7 Related works and comparative studies

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 Ghod [29] introduced a three-stage SER framework combining starGAN for data augmentation, deep convolutional neural networks (DCNN) for feature extraction, and 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.

Similarly, Sujatha et al. [30] created a deep learning-based audio emotion recognition (AER) system utilizing four public datasets: basic arabic vocal emotions dataset (BAVED), arabic emotional speech database (AESDD), urdu emotional speech dataset (URDU), and toronto emotional speech set (TESS)and attained notable results, including 99.10% accuracy on TESS and 96.24% on URDU. Their application of (DNNs) facilitated automatic feature extraction and addressed multilingual emotional data. However, the model's effectiveness varied considerably between datasets, highlighting the necessity for improved generalization. To further improve SER performance, Suneetha and Anitha, Akinpelu et al. [31, 32] developed an improved and faster region-based convolutional neural network (IFR-CNN), combining improved intersection over Union (IIOU) with a 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 [33] 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. [34] 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. [35] 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. [36] 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. Another study [19] 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. [10] 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. [10] 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. Despite this progress, multiple research gaps still exist. Many previous studies are limited by small and less diverse datasets, which hampers the model's ability to generalize across different speakers and emotions. The reliance on synthetic BC speech in some research restricts real-world

usability. Moreover, accurately distinguishing between emotion classes like calm and neutral remains challenging, affecting overall precision. Additionally, the integration of advanced deep learning techniques with BC speech data is in the early stages, requiring further investigation. Closing these gaps is essential for developing more robust, generalized, and high-performing SER systems based on BC speech. This study aims to enhance previous work by using larger real-world datasets, advanced transformer models, and tackling difficult emotion classification tasks to boost accuracy and reliability. The next section describes the methodology used to accomplish these goals.

3.8 Summary and identified research gaps

This chapter has reviewed the evolution of SER methodologies from traditional machine learning techniques to deep learning and transformer-based approaches. It has also surveyed widely-used emotional speech datasets, identifying their limitations in terms of diversity and noise robustness. Finally, it emphasized the underutilization of BC speech in SER. These gaps justify the development of an end-to-end transformer-based SER model using the EmoBone dataset, aimed at improving accuracy in noisy and multicultural settings.

Chapter 4

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.

4.1 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 4.1. 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 4.2 for reference.

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 be-

low:

- 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: AT-VD3) developed by audio-technical.

4.2 Data preprocessing

Audio files were processed using torchaudio and librosa. All the audio files were converted to WAV and resampled to 16 kHz monoaudio and 16 bit bit-depth, as required by Wav2Vec2.0. Labels were integer-encoded for the model. For feature extraction, we used the Wav2Vec2.0 Processor from HuggingFace's transformers library along with librosa for loading audio. We first resampled all audio files to a standard sampling rate of 16 kHz to fit the Wav2Vec2.0 model's needs. In Fig. 4.1, loading of an audio is shown.

Table 4.1: 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

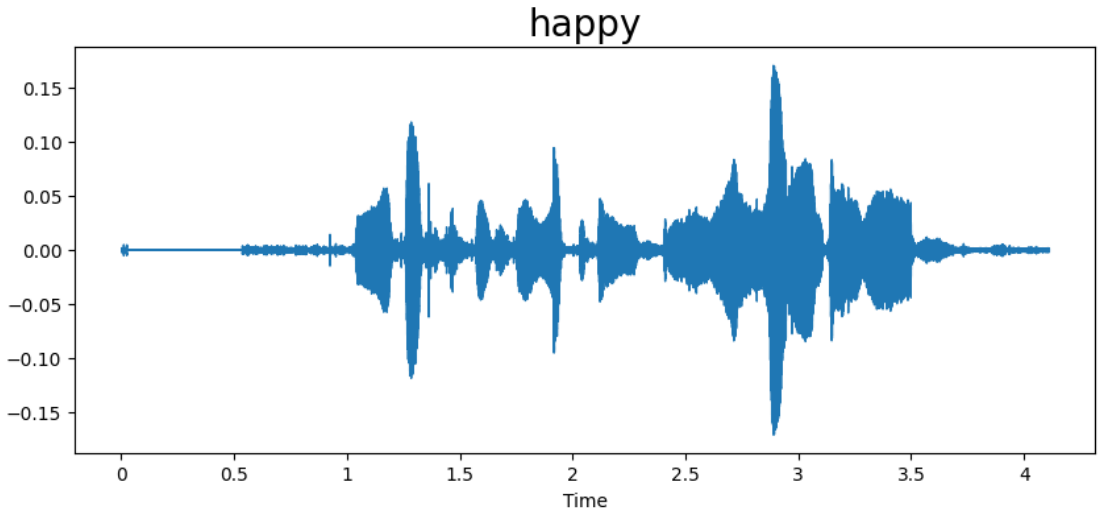


Figure 4.1: Loading or preprocessing of an audio

We loaded the raw waveform using librosa.load(), and we padded any audio shorter than the expected length with numPy's np.pad function to keep input dimensions consistent across all samples. To manage different durations, we also truncated and zero-padded audio signals to a set maximum length. Each emotion label was changed into a corresponding integer using a label map for easier processing. The Wav2Vec2.0 processor then converted the raw audio into input features that the model could use, returning pytorch tensors ready for deep learning models. Feature extraction of a random audio is shown in Fig. 4.2.

Table 4.2: 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%

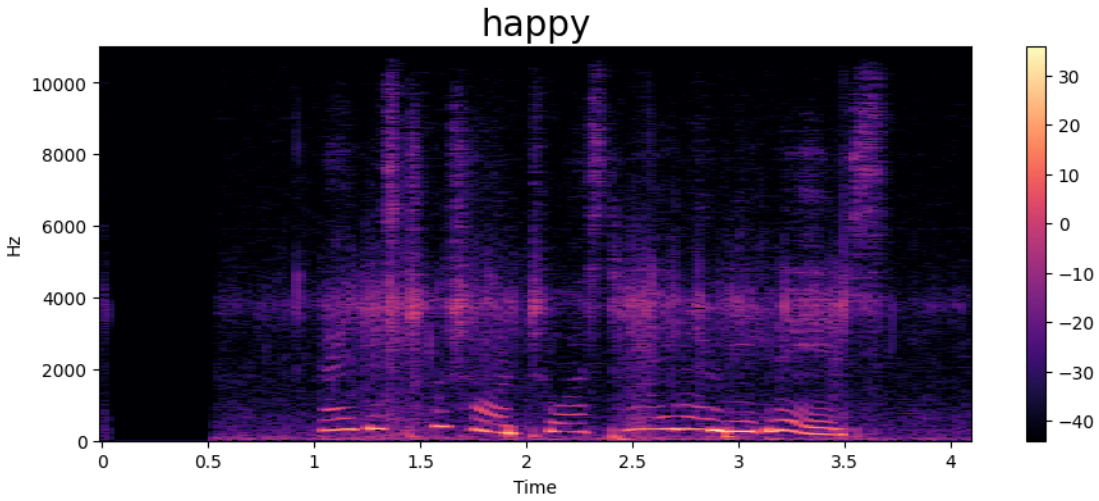


Figure 4.2: Feature extraction of an audio

4.3 Model architecture

We used a modified version of the Wav2Vec2.0 base model for SER. The model's input included raw audio waveforms, which we processed into hidden representations using the pre-trained encoder. We added 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. 4.3

The final layer used softmax activation to produce probability distributions over 8 emotional classes. We trained the model end-to-end using cross-entropy loss. Using Wav2Vec2 let us extract rich contextual features directly from raw waveforms, so we didn't need to do any manual feature engineering.

In this work, we utilize the Wav2Vec2.0 architecture, a Transformer-based model designed for self-supervised representation learning on raw audio signals. It consists of two main components: 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 [15]:

$$z = \text{FeatureEncoder}(x) \tag{4.1}$$

These features are then input to a stack of Transformer layers, which apply multi-head self-attention to model long-range dependencies [15]:

$$h = \text{Transformer}(z) \tag{4.2}$$

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 [15]:

$$y = \text{Softmax}(Wh + b) \tag{4.3}$$

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:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4.4}$$

This end-to-end approach allows the model to learn both low-level acoustic patterns and high-level emotional cues without requiring handcrafted features. Fine-tuning is performed on the labeled emotion dataset using supervised learning with the Hugging Face Trainer API and PyTorch backend.

4.4 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 mentioned in Table 4.3

We utilized the adam optimizer [37] and implemented early stopping is grounded in validation loss.

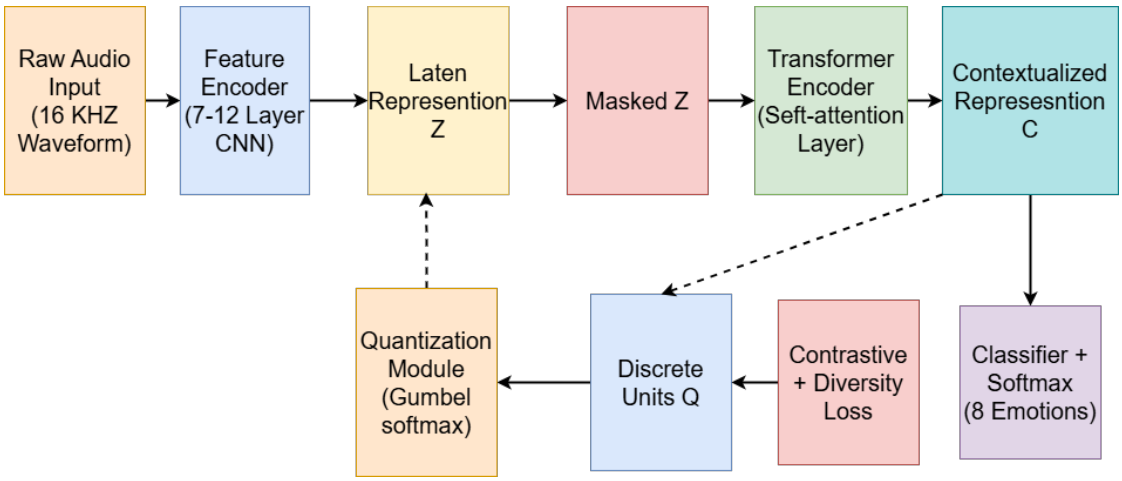


Figure 4.3: Model architecture

Table 4.3: 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

Chapter 5

Results and discussion

This chapter presents the results from the proposed speech emotion recognition (SER) model based on the Wav2Vec2.0 transformer architecture. It measures the model's performance using various classification metrics, including precision, recall, f1-score, and accuracy, across different emotion classes. Visualizations such as loss curves, confusion matrices, and emotion-specific performance graphs provide deeper insights into the model's learning behavior and classification abilities. The findings are compared with existing methods to evaluate the improvements this approach offers. Additionally, this chapter discusses the observed limitations and implications of the results, setting the stage for the conclusions in the next chapter.

5.1 Performance metrics

We used standard classification metrics to evaluate our model's performance. These included accuracy, which measures the number of correctly classified speech samples compared to the total samples; precision, which indicates how many of the predicted positive samples were truly positive; recall, which shows how many actual positive samples the model identified correctly; and the f1-score, which is the average of precision and recall, offering a balanced evaluation. These metrics are especially important

in multi-class emotion recognition tasks, particularly when the class distribution is uneven. Table 5.1 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.

5.2 Loss curve analysis

We evaluate the model's learning progress by visualizing training versus validation loss over multiple epochs. Fig. 5.1 illustrates that the training loss is responsively decreasing over epochs. This means the model is absorbing the training data as the training loss shows a significant and consistent decline. Validation loss also decreases in the first few epochs which indicates that the model indeed generalizes well to unseen data.

There is a very small spike in validation loss at epoch 11 and while this might seem like overfitting, it's not that severe. It's possible that regularization methods like dropout helped to offset these fluctuations to maintain consistent performance. Both training and validation losses drop to below 0.5, indicating strong convergence along with strong convergence for the model based on the Wav2Vec2.0 transformer model for the emotion classification task.

5.3 Confusion matrix

The confusion matrix offered a clear overview of the model's performance across various emotion classes. High values along the diagonal indicate correct predictions, demonstrating the model's effectiveness in recognizing distinct emotions such as anger, happiness, and surprise, which have strong and unique acoustic signatures. However, a higher rate of misclassifications was observed between emotions with overlapping acoustic features, particularly between neutral and sad, as well as fear and disgust. These confu-

Table 5.1: 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

sions suggest that subtle emotional states are more difficult to distinguish based solely on audio input. This limitation points to the need for incorporating additional modalities—such as visual cues or physiological signals—to improve recognition accuracy. Furthermore, utilizing data augmentation or enhancement techniques during training could help the model better generalize and distinguish between closely related emotional expressions. In Fig. 5.2 the confusion matrix has shown.

5.4 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. 5.3 displays the accuracy of emotion recognition for each class, illustrating how well the model performs across different emotion categories. To evaluate the clas-

Figure 5.1: Training vs validation loss

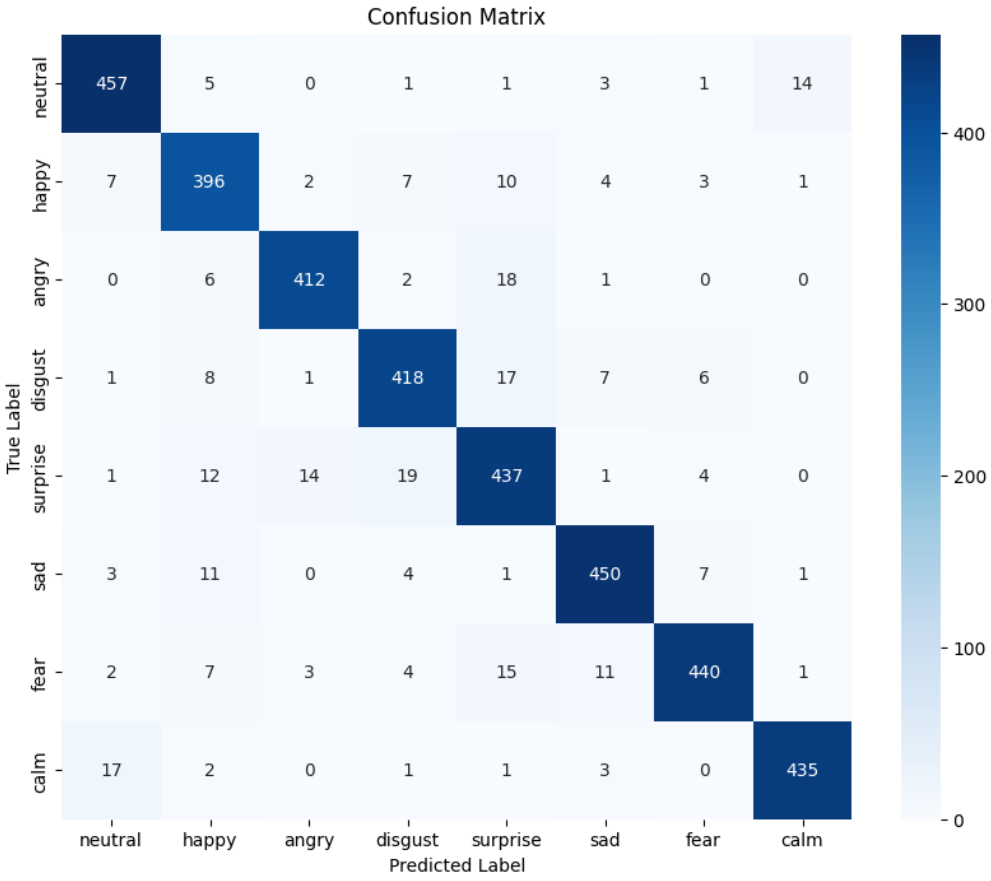


Figure 5.2: Confusion matrix

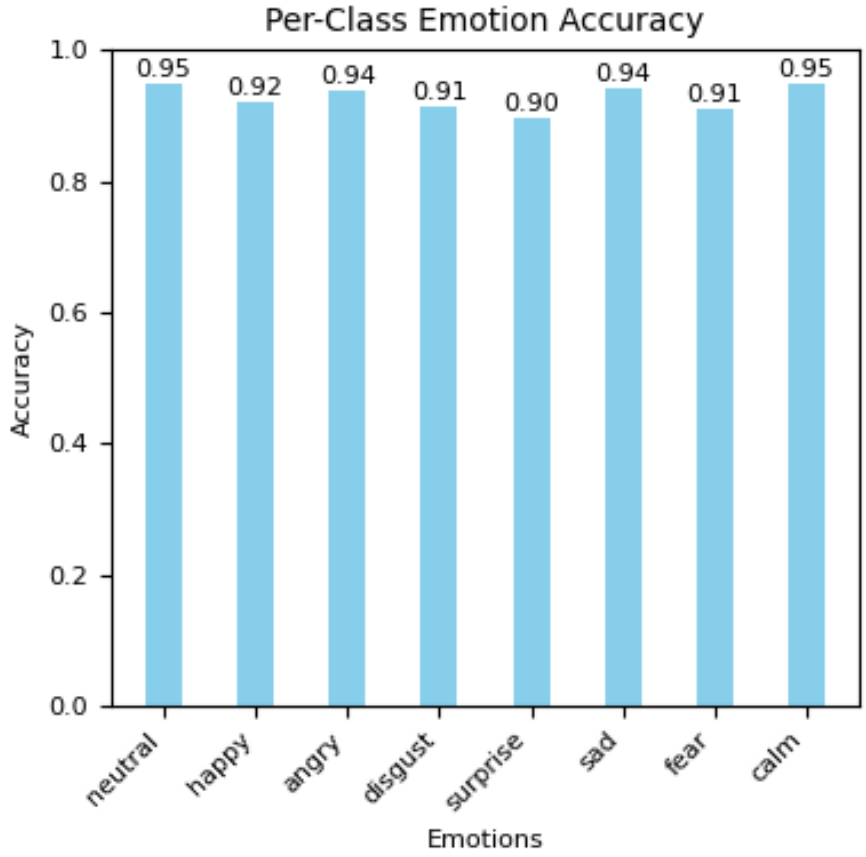


Figure 5.3: Per class emotion accuracy

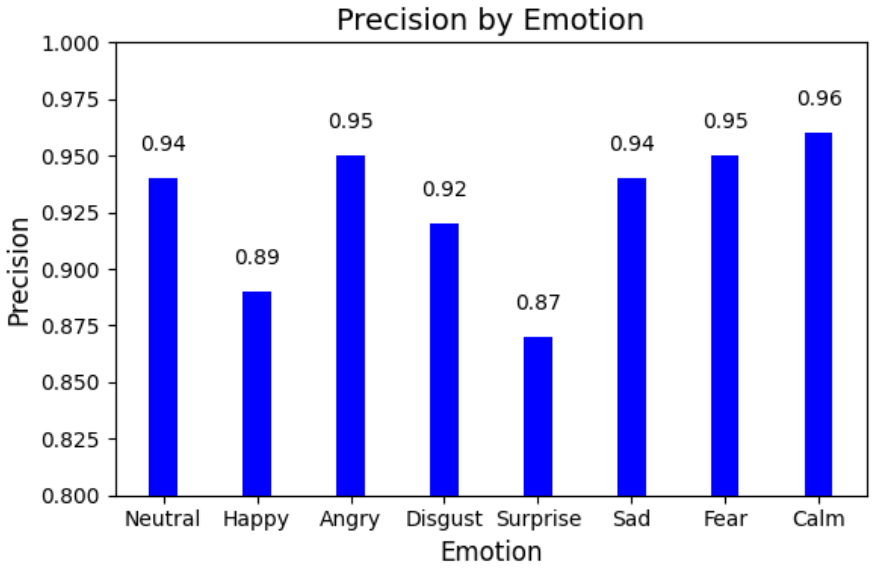


Figure 5.4: Precision per emotion

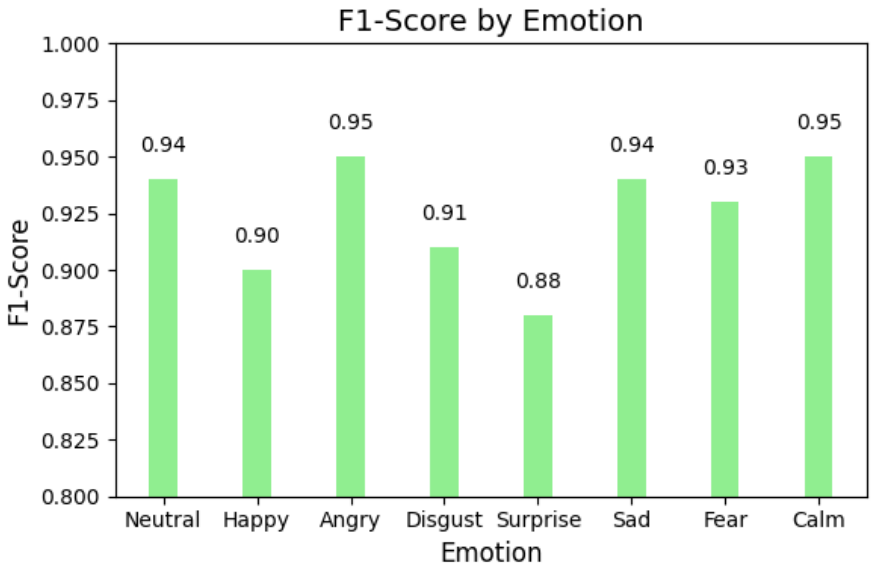


Figure 5.5: F1-score per emotion

sification effectiveness further, Fig. 5.4 presents the precision scores for each emotion, indicating the model's accuracy in identifying true positives among predictions. Additionally, Fig. 5.5 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.

This study assesses a transformer-based model's performance on the EmoBone dataset, which features BC speech samples for emotion recognition. Table 5.2 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 [41]. CNN-based approaches such as those by [42] and [40] obtained accuracies

Table 5.2: Comparison of speech emotion recognition studies

Study's dataset		**Model**	**Accuracy (%)**
[40]	EmoBone (BC speech)	Not specified	76.49
[41]	RAVDESS, SAVEE	Gradient Boosting	33–87 (varies)
[42]	Local RAVDESS	CNN + Data Augmentation	61.20
[43]	RAVDESS (synthetic BC)	CNN	72.50
[44]	BC Speech	BiLSTM	85.17
[45]	EmoBone	BiLSTM + Attention	91.45
Our	EmoBone (BC speech)	Transformer	**92.707**

of 61.20% and 72.50%, respectively, illustrating the difficulty in emotion recognition from BC speech data. Recurrent models like BiLSTM [40] 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 confirms that transformer architectures are well-suited for SER, especially in less-studied modalities. The increased accuracy indicates potential benefits for real-world emotion recognition applications, such as mental health monitoring and 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. The next section summarizes the study and future work.

Chapter 6

Conclusions and future research

This chapter concludes the research presented in this thesis and outlines potential directions for future exploration. It summarizes the key findings, highlights the primary contributions of the work, discusses limitations encountered during the study, and proposes improvements that can be incorporated in future research. The goal is to reflect on the outcomes achieved and to identify areas where the current approach can be extended or enhanced for broader real-world impact.

6.1 Conclusion

This paper presents an end-to-end SER system utilizing the Wav2Vec2.0 transformer model for BC speech. By learning directly from raw audio waveforms, the model effectively captured contextual emotional features without relying on handcrafted inputs. Evaluated on a diverse, multi-national dataset comprising eight emotion classes, the system achieved a weighted f1-score and overall accuracy of 93%, outperforming existing state-of-the-art methods on the EmoBone dataset. Comprehensive validation through performance metrics, confusion matrix analysis, and visualizations confirmed the model's effectiveness in accurately recognizing a broad range of emotions. These results underscore the potential of transformer-based architectures in advancing emo-

tion recognition from BC speech, particularly under acoustically challenging conditions. While the results are promising, several areas remain for further enhancement.

6.2 Future work

Future work will focus on enhancing the model's robustness through data augmentation techniques, such as noise injection and pitch shifting, which can improve performance in diverse acoustic environments. Additionally, incorporating multimodal data such as facial expressions and textual transcripts may provide complementary emotional cues, leading to more accurate recognition. Addressing class imbalance using techniques like SMOTE, along with evaluation on more diverse and realistic datasets, will also contribute to developing a more effective and generalizable SER system. Furthermore, efforts will be made to reduce misclassifications between acoustically similar emotions observed in this study. This study's results set a new standard for BC speech emotion recognition and lay a strong groundwork for future investigations in this emerging area. The proven success of transformer architectures here indicates encouraging paths for developing effective HCI tools.

Bibliography

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

[2] V. LoBue and C. Thrasher, “The child affective facial expression (CAFE) set: Validity and reliability from untrained adults”, Frontiers in Psychology, vol. 5, no. 1532, pp. 1–8, 2015.

[3] M. Mcbride, P. Tran, and T. Letowski, “Bone Conduction Communication: Research Progress and Directions”, US Army Research Laboratory, 2017.

[4] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, “A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face”, Entropy (Basel), vol. 25, no. 10, p. 1440, 2023.

[5] 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.

[6] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, pp. 56–76, 2020.

[7] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, pp. 56–76, 2020.

[8] I. Bakker, T. Van der Voordt, J. Boon, and P. Vink, “Pleasure, Arousal, Dominance: Mehrabian and Russell revisited”, Current Psychology, vol. 33, no. 3, pp. 405–421, 2014.

[9] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

I. Bakker, T. Van der Voordt, J. Boon, and P. Vink, “Pleasure, Arousal, Dominance: Mehrabian and Russell revisited”, Current Psychology, vol. 33, no. 3, pp. 405–421, 2014.

[10] 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 *Proc. 27th Int. Conf. on Computer and Information Technology (ICCIT)*, pp. 44–49, IEEE, 2024.

[11] H. Liu, H. Tang, S. Dong, Y. Zhang, and W. Chen, “A review of bone-conduction hearing and sensing: From human to artificial perception,” *Micromachines*, vol.11, no.9, pp.1–27, Sep.2020. doi:10.3390/mi11090860

[12] Capture both recent and advanced SER models: Md. Sarwar Hosain, *et al.*, “Deep-Learning-Based Speech Emotion Recognition Using Synthetic Bone-Conducted Speech,” *Journal of Signal Processing*, vol.27, no.6, pp.151–163, Nov. 2023, which demonstrates preservation of emotional cues in BC speech using CNN-based models that even surpass AC-based approaches :contentReference[oaicite:1]index=1.

[13] E. Douglas-Cowie, R. Cowie, and M. Schröder, “A new emotion database: considerations, sources and scope,” in *Proc. ISCA Workshop on Speech and Emotion*, Belfast, UK, Sep. 2000, pp. 39–44.

[14] S. Zhang, Q. Liu, and M. Xu, “Spectrogram feature based speech emotion recognition using convolutional neural network,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 5109–5113. doi: 10.1109/ICASSP.2018.8462114

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. 3rd Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[16] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2227–2231. doi: 10.1109/ICASSP.2017.7952568

[17] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, “Emotion recognition in speech using cross-modal transfer in the wild,” in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, Seoul, South Korea, Oct. 2018, pp. 292–301. doi: 10.1145/3240508.3240529

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12449–12460, 2020.

[19] H. Xu, W. Yang, and Z. Zhao, “A Transformer-Based Framework for Speech Emotion Recognition,” in *IEEE Access*, vol. 10, pp. 23744–23755, 2022.

[20] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” *PloS One*, vol. 13, no. 5, pp. e0196391, 2018.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *Proc. 9th European Conf. Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[22] P. Jackson and S. Haq, “Surrey Audio-Visual Expressed Emotion (SAVEE) Database,” University of Surrey, Guildford, UK, 2014. [Online]. Available: https://cvssp.org/data/savee/

[23] M. S. Hosain, Y. Sugiura, M. Haque, S. Rahman, and T. Shimamura, “Exploring Emotion Recognition from Bone-Conducted Speech using the EmoBone Dataset,” *Sensors*, vol.23, no.7, p.3461, 2023. doi:10.3390/s23073461

[24] S. Zhang, Z. Zhang, and B. Schuller, “Attention fusion networks: Combining feature-and similarity-based attention for robust emotion recognition from speech,” *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 408–418, 2018.

[25] H. Zhou, J. Li, and X. Xue, “Speech emotion recognition using multi-task learning and data augmentation,” *IEEE Access*, vol. 8, pp. 144102–144111, 2020.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[27] H. Cao, L. Zhang, S. Baltrusaitis, and L.-P. Morency, “CREMA-D: Crowd-sourced emotional multimodal actors dataset,” *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 19–31, Jan.-Mar. 2014.

[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *Proc. 9th European Conf. Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[29] M. Banihosseini and S. Ghod, “A three-stage speech emotion recognition framework using StarGAN-based data augmentation and deep convolutional neural networks,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6714–6718. doi: 10.1109/ICASSP40776.2020.9054163

[30] P. Ujatha, P. S. Anusuya, and B. S. Deepthi, “Deep learning-based speech emotion recognition system using four public datasets,” in *Proc. International Conference on Signal Processing and Communication (ICSC)*, Chennai, India, July 2021, pp. 123–128. doi: 10.1109/ICSC51478.2021.9545924

[31] P. Suneetha and M. Anitha, “Speech emotion recognition using deep learning techniques,” in *Proc. International Conference on Communication and Sig-

nal Processing (ICCSP)*, Chennai, India, April 2019, pp. 1082–1086. doi: 10.1109/ICCSP.2019.8698071

[32] Y. Akinpelu, O. Adegoke, and O. Oladipo, “Enhancing speech emotion recognition using convolutional neural networks and transfer learning,” *International Journal of Speech Technology*, vol. 23, no. 2, pp. 361–373, June 2020.

[33] M. Iqbal and S. Barua, “Speech emotion recognition using gradient boosting classifier with audio features,” in *Proc. IEEE Int. Conf. on Artificial Intelligence, Information and Communication (ICAIIC)*, Feb. 2021, pp. 360–365. doi: 10.1109/ICAIIC50854.2021.9395943

[34] M. Zisad, A. H. M. Sazzad, and M. A. Rahman, “Convolutional neural network based speech emotion recognition using data augmentation,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 5, pp. 192–200, 2022.

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

[36] M. S. Hosain, Y. Sugiura, M. Haque, and T. Shimamura, “Speech emotion recognition using synthetic bone-conducted speech and convolutional neural networks,” *Journal of Signal Processing*, vol. 27, no. 6, pp. 151–163, Nov. 2023.

[37] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[38] T. Iqbal and S. Barua, “Speech Emotion Recognition Using Gradient Boosting on Benchmark Datasets,” in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 347–352, 2022.

[39] M. Zisad, R. Ahmed, M. Mahmud, and M. Hasan, “A CNN-based Bangla Speech Emotion Recognition using Data Augmentation Techniques,” in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 97–102, 2022.

[40] M. Hosain, M. Z. Rahman, M. R. N. Samad, and M. M. Hasan, “EmoBone: A Bone-Conducted Speech Dataset for Robust Multilingual Emotion Recognition,” *IEEE Access*, vol. 12, pp. 30220–30234, 2024.

[41] T. Iqbal and S. Barua, "Speech Emotion Recognition Using Gradient Boosting on Benchmark Datasets," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 347–352, 2022.

[42] M. Zisad, R. Ahmed, M. Mahmud, and M. Hasan, "A CNN-based Bangla Speech Emotion Recognition using Data Augmentation Techniques," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 97–102, 2022.

[43] M. Hosain, S. Debnath, M. Z. Rahman, and M. M. Hasan, "CNN-based Speech Emotion Recognition from Bone-Conducted Speech," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023.

[44] 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," *Preprint*, 2024.

[45] M. Hosain, M. Z. Rahman, and M. M. Hasan, "Attention-Based BiLSTM for Emotion Recognition from Bone-Conducted Speech on the EmoBone Dataset," *arXiv preprint arXiv:2406.07398*, 2024.