Enhancing Bangla Speech Emotion Recognition

by

Masiat Mohammad Momshad
21241017
Jonathon Lenny Baroi
24341252
Raisa Tahasen
24141097
Tasnia Hossain
23341097

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Students Full Name & Signature:

Masiat Mohammad Momshad
21241017

Jonathon Lenny Baroi
24341252

Raisa Tahasen
24141097

Tasnia Hossain
23341097

Approval

The thesis/project titled Enhancing Bangla Speech Emotion Recognition submitted by

1. Masiat Mohammad Momshad(21241017)
2. Jonathon Lenny Baroi(24341252)
3. Raisa Tahasen(24141097)
4. Tasnia Hossain(23341097)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 24, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Jannatun Noor
Assistant Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethical Statement

In this research, audio data was collected from publicly available videos on YouTube to develop a diverse dataset for speech emotion recognition in the Bangla language. The videos used to create the dataset are openly accessible to the public and were used strictly for research and academic purposes, without any intent of commercial exploitation. According to the copyright and intellectual property laws of Bangladesh, the use of publicly available online content for non-commercial, academic, and research purposes is permissible[3]. In addition to Bangla-language content from Bangladesh, this study also utilized publicly available Bangla-language videos originating from India, under similar legal and ethical considerations[1]. Appropriate measures were taken to ensure that no personal, sensitive, or confidential information was used in this study. Furthermore, no attempt was made to identify or harm any individual or entity featured in the collected content. The purpose of utilizing this data is solely to advance the field of speech emotion recognition in the Bangla language and contribute to the development of language-specific machine learning models. Proper consideration was given to uphold ethical research standards, including respect for content ownership, privacy, and cultural sensitivity.

Abstract

Speech Emotion Recognition (SER) has grown to be a crucial part of human-computer interaction and other advanced speech recognition systems. Although Bangla is one of the most widely spoken languages in the world, there has been a lack of extensive research on Bangla speech emotion recognition. Speech, a complex analogue signal that varies over time, presents unique challenges and opportunities for researchers implementing machine learning techniques in audio signal classification. In this research we will introduce a new benchmark Bangla Speech Emotion recognition dataset. We will also develop a benchmark method to effectively comprehend emotion recognition from Bangla speech. The method will include complex deep-learning techniques to classify emotion from the audio speech. The research topic proposes many benefits such as the potential for developing more emotionally aware machine interactions in diverse and low-resource linguistic contexts. The broader implications encompass the advancement of human-computer interaction technologies, rendering them more inclusive and proficient in comprehending a broader spectrum of emotional expressions, including mixed emotions.

Keywords: Speech Emotion Recognition; Bangla SER; Deep Learning; Machine Learning

Acknowledgement

We would like to extend our heartfelt gratitude to our supervisor, Dr. Jannatun Noor, for her exceptional guidance throughout our thesis journey and for her invaluable contributions that greatly enriched our research.

We are also deeply thankful to our parents for their unwavering support, encouragement, and belief in us, without which this achievement would not have been possible.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several abbreviation that will be later used within the body of the document

AdamW Adam with Weight Decay

AE Autoencoder

BiLSTM Bidirectional Long Short-Term Memory

CASIA Chinese Academy of Sciences Emotional Speech Database

CNN Convolutional Neural Network

CREMA-D Crowd-sourced Emotional Multimodal Actors Dataset

CRNN Convolutional Recurrent Neural Network

CV-Transformer Convolutional-Transformer Hybrid Model

DBN Deep Belief Network

DCNN Deep Convolutional Neural Network

DCT Discrete Cosine Transform

ECOC Error Correcting Output Codes

EESDB Chinese Elderly Emotion Database

EMO-DB Berlin Database of Emotional Speech

Farsdat Farsi Speech Database

FBANK Filter Bank

FP Fourier Parameter

GC Gamma Classifier

GELU Gaussian Error Linear Unit

GMM Gaussian Mixture Model

GRU Gated Recurrent Unit

HMM Hidden Markov Model

HNR Harmonic to Noise Ratio

HTK Hidden Markov Model Toolkit

IEMOCAP Interactive Emotional Dyadic Motion Capture Database

KBES KUET Bangla Emotional Speech Dataset

KNN K-Nearest Neighbors

LFPC Log Frequency Power Coefficients

LOSO Leave-One-Speaker-Out

LSSED Large-Scale Speech Emotion Dataset

LSTM Long Short-Term Memory

MFCC Mel-Frequency Cepstral Coefficients

MLP Multi-Layer Perceptron

MSP-PODCAST MSP Podcast Dataset

NCA Neighborhood Component Analysis

PLP Perceptual Linear Prediction

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song

RCA Relevant Component Analysis

ReLU Rectified Linear Unit

RF Random Forest

RMSprop Root Mean Square Propagation

RNCA Regularized Neighborhood Component Analysis

RNN Recurrent Neural Network

SER Speech Emotion Recognition

SSA Self Speaker Attention

STFT Short-Time Fourier Transform

SUBESCO Bangla Speech Emotion Dataset

SVM Support Vector Machine

TDF Time-Distributed Flatten Layer

TEO Teager Energy Operator

UAR Unweighted Average Recall

ViT Vision Transformer

WA Weighted Accuracy

ZCR Zero Crossing Rate

Chapter 1

Introduction

1.1 Background

Presently technology is advancing faster than ever, and the realm of human- computer interaction has seen some incredible progress as well that enables users to communicate effectively with computers improving their usability. To enhance this goal further, an area of effective computing called speech emo- tion recognition seeks to train computers to detect peoples emotional states via speech which have potential applications in several different domains, such as virtual assistants, psychological evaluations, advertising, and consumer data mining, among others. While notable achievement has been made in emotion recognition from speech for several languages, there remains a notable gap in re- search focusing on underrepresented languages, such as Bangla. Despite its vast speaker population, research on emotion recognition in the Bangla language is limited, posing challenges to the development of emotion-aware systems tai- lored to Bangla speakers. This paper aims at advancing our understanding of emotion processing in the Bangla language, to contribute to the development of culturally sensitive and inclusive technologies that cater to diverse linguistic communities. We aim to produce a comprehensive and balanced data set in the Bangla language that will be a paramount addition to the existing resources in Bangla. Alongside this, we seek to investigate the acoustic features that will assist with both the expression and interpretation of emotions in Bangla speech. By leveraging deep learning techniques and data-driven approaches, we seek to develop a robust emotion recognition system capable of accurately identifying emotional states from Bangla speech signals.

1.2 Problem Statement

Emotion recognition contributes to various areas, such as human-computer interactions, monitoring mental health, market research, lie detection and many more. Therefore, the more we understand human emotions from speech, we can enhance the performance of these applications, to make them a more integral part of our lives. However, emotion can be a very subjective aspect and may be interpreted differently by individuals. This variability is further multiplied by cultural nuances of how we express ourselves. We also lack comprehensive datasets which adequately represent all the aspects of human emotions. Furthermore, emotions may be ambiguous due to cases of multiple emotions being expressed simultaneously. In recent work, SER

has employed deep learning techniques such as CNNs, RNNs and attention mechanisms. However, some problems still remain, such as the generalization of training for different languages, speakers and cultural contexts. Also, these models require an adequate amount of data, which is rarely available, especially in under-resourced languages.

1.3 Motivation

The motivation behind this research arises from the pressing need to address several critical gaps in the field of Speech Emotion Recognition (SER), particularly for the Bangla language. Although Bangla is the seventh most spoken language globally, it lacks comprehensive and diverse emotional speech datasets, with existing resources like BanglaSER and SUBESCO being limited in size, linguistic richness, and realism-often relying on acted emotions and scripted content. The creation of reliable, broadly applicable SER models is hampered by this paucity. Furthermore, the majority of SER research to date has been on Western languages, leaving Bangla underrepresented, because emotions are culturally and linguistically complex. In order to improve the authenticity and application of SER systems, our work gathers emotionally charged speech from authentic Bangla contexts, including discussion shows, news, and sermons. Additionally, using cutting-edge deep learning methods like Wav2Vec2 and ExHuBERT has the potential to greatly improve performance; however, these models need high-quality, domain-specific data in order to be fine-tuned effectively. This research aims to advance the field and contribute to more inclusive and culturally aware language technologies with broad applications in education, mental health, human-computer interaction, and more by creating a novel Bangla SER dataset and benchmarking it using transformer-based models as well as handcrafted feature-based CNNs.

1.4 Research objectives

This paper aims to advance Speech Emotion Recognition in the Bangla Language. Firstly, a new Bangla dataset from real-life talk shows will be created to provide a more authentic and diverse representation of natural human emotions as expressed in spontaneous speech. Next, various deep learning models will be trained using the new dataset. Pre-trained models will also be fine-tuned on the dataset to adapt them to the nuances of Bangla SER. Lastly, the performance will be evaluated against existing Bangla SER systems using standard metrics, thereby demonstrating the impact of real-life data on model accuracy and robustness.
Hence, the main objectives include:

1. Introduce a new dataset in Bangla
2. Train different models on the dataset
3. Fine-tune different models on the dataset
4. Evaluate the performances of the models

Chapter 2

Related Work

There have been various Speech Emotion Recognition datasets and models built over the years in English and other languages as well. In this section, we will critically examine some of these approaches.

2.1 English Datasets

Akinpelu and Viriri[36] has used the TESS dataset, which comprises seven emotions: angry, happy, fear, pleasant, disgust, sad, surprise and neutral, with more than 2800 speech samples. The mel-spectrogram from the audio signal was the feature of choice. This is a visual representation of sound signals and shows how the frequency changes over time. It uses the Mel scale on the y-axis, which imitates how humans hear variations in pitch. This image was fed into a DCNN model, followed by an attention layer and RNCA (regularized neighborhood component analysis). Classification was however done with primitive models: SVM, MLP and RF. The attention layer has a significant impact on training time and concentrates on features with more emotional data. The RNCA simplifies the problem using dimension reduction, while also preserving the local structure of the data. Two DCNNs were used and compared, the VGG-16 and VGG-19, local features were extracted using these pre-trained models. The feature selection process has significantly improved accuracy by 3.7%, showing the importance of feature selection. The attention-based DCNN+RNCA+RF model was finally proposed with a promising accuracy of 97.9%. The feature extraction techniques used in this paper are our primary inspiration.

Kim and Lee[32] introduce an innovative method that integrates vision transformers (ViT) with knowledge transfer to enhance speech emotion recognition (SER). Its uniqueness comes from utilizing vertically segmented spectrogram patches to more effectively capture temporal frequency correlations compared to traditional square patches, along with the application of image coordinate encoding to improve positional awareness. A teacher-student network setup is utilized, where the teacher network, equipped with a convolutional stem, learns spatial features and shares this knowledge with the student network, thereby decreasing computational complexity. The model demonstrates impressive accuracies: 99.47% on SAVEE, 99.76% on EmoDB, and 95.24% on CREMA-D, attributed to the effective combination of CNN-based feature extraction and transformer-based temporal modeling. Strengths encompass innovative patch design, efficient knowledge transfer, and improved posi-

tional encoding, whereas weaknesses pertain to possible difficulties in generalizing to datasets with higher variability, like IEMOCAP, and sensitivity to hyperparameter tuning. The method achieves a strong balance between accuracy and efficiency by utilizing transformers alongside CNN features and incorporating knowledge transfer, allowing it to surpass numerous current SER techniques.

Wani et al.[27] and El Ayadi et al.[8] present a detailed analysis of SER systems. Authors of[27] notes that deep learning approaches are the turning point in SER, despite the fact that much work has been done using traditional techniques. It also emphasises that various challenges remain to be addressed. Finally, it advises that the systems need more reliable algorithms to enhance performance so that accuracy rates increase, as well as a focus on identifying an acceptable set of features and competent classification techniques to further boost HCI. El Ayadi et al.[8] provides a broader perspective for pattern recognition researchers who do not have a strong foundation in speech analysis. It surveys three key components of speech emotion recognition: design criteria for emotional speech corpora, the effect of speech features on classification performance and methods. Further issues are drawn from concluding the study and the use of more robust techniques is proposed here too.

One of our primary focuses, the self-attention mechanism, was employed by Tarantino et al.[18], as it sees every frame of the utterance simultaneously, which means it can remember the long-term correlations in emotions, unlike RNNs which have a decaying memory. In addition, they used a global window (with 50% overlap) due to the fact that it can unveil relationships between data points which would otherwise be impossible. Indeed the results show that a larger overlapping results in a smaller loss. For classification, they have tried both regression and classification, as regression shows the degree of multiple emotions and their percentages present in the audio, unlike classification which simply outputs a single label. This aligns with the fact that emotions are multi-layered in terms of the emotions present in a speech window. While annotators often disagreed with each other on the labels, the results show that the accuracy increased as the degree of agreement between them increased. Another one of our inspirations, the RNCA was compared with NCA and RCA by[7], and RNCA overcomes the problem of overfitting which is faced by NCA and RCA in the case of high-dimensional data. Regularization prevents overfitting by imposing constraints on the model parameters. This can be achieved by applying a Gaussian prior.

Trinh Van et al.[35] made use of the IEMOCAP corpus, but only the speech audio signals, of approximately twelve hours of data. It includes the emotions of anger, happiness, sadness, and neutrality. Data augmentation such as adding noise and formant shifting helped to increase the size of the dataset four times. The feature set was divided into two groups: first denoting the characteristics of the speech signal spectrum (spectral flatness, mel-spectral coefficients, spectral bandwidth, spectral contrast, spectral centroid, pitch and roll-off frequency) and the second describing the intensity (FRMS). Three Deep Neural Networks were tested and compared, namely the CNN, GRU and CRNN. The GRU, a variant of the LSTM, was used.. It is worth noting that batch normalization has helped to stabilize the distribution of activation values throughout the training, thereby substantially reducing training

time. Similarly, ELU has contributed to speeding up learning times and improved classification accuracy. As for the model performances, for the mel-spectral coefficients, CRNN demonstrated the best performance while GRU exhibited the worst, however, the difference was only about 2%. However, for mel-spectral coefficients combined with the rest of the features, the GRU performed the best with CRNN the worst, with similar differences. It was also found that the spectral centroid followed by spectral roll-off and spectral flatness had the best influence on high accuracy.

Mountzouris et al.[42] uses an attention mechanism (CNN-ATN) together with a convolutional neural network (CNN) to solve the issues and challenges with speech emotion recognition (SER). The researchers trained and evaluated their models using the Ryerson Audio-Visual Database (RAVDESS) and the Surrey Audio-Visual Expressed Emotion (SAVEE) databases. Initially, the voice characteristics of the speakers were obtained by extracting Mel Frequency Cepstral Coefficients (MFCCs) using the Librosa library. This library ensured that the data was consistently processed and gave a comprehensive set of features. Moreover, these coefficients employ the Mel scale, which is derived from the way humans perceive various signal frequencies. Finally, when an attention mechanism was included in the fundamental CNN, their model could focus only on the auditory input portions that had greater emotional meaning., while disregarding portions where the speaker remains silent or utters non-emotionally charged words. These contributed to enhancing the capacity to forecast emotions. The findings demonstrated that the CNN-ATN surpassed their previous efforts against the most complex methods such as, an LSTM network, a LSTM network with an attention mechanism (LSTM-ATN), a CNN and a DBN. For the RAVDESS database, the CNN-ATN attained an accuracy of 77% and for the SAVEE database of 74%. The research establishes a basis for future investigations in SER. Nevertheless, the precision is insufficient. The growing intricacy of implementing attention-enhanced CNNs may provide difficulties, especially in contexts with limited resources. On the other hand Abdulmohsin et al.[22] presents a new statistical method for SER emphasizing the extraction of features that utilize various degrees of standard deviation (SD) from the mean of each feature. The aim is to obtain a more nuanced variance distribution of features, rather than relying on the typical 2 SD method often used in earlier studies. The method includes computing features like Zero Crossing(ZC), Mel-Frequency Cepstral Coefficients (MFCC), harmonic ratio, and additional metrics across various SD values, spanning from 0.25 to 4 SD on both sides of the mean. This helped them reach impressive accuracies of 86.1% for RAVEDESS and 91.7% for SAVEE dataset.

Moine et al.[26] introduces an innovative approach that incorporates speaker identity information into emotion classification, by introducing Self Speaker Attention (SSA) to prioritize the components of an utterance that are related to emotions. Moreover, they suggested a regularization method to mitigate the issue of class collapse in the classification layer. Lastly, it contributed to substantial enhancements in SER performance, specifically on the IEMOCAP dataset. For the datasets the authors used, Att-HACK and IEMOCAP, Att-Hack consists of 25 speakers who interpret 100 utterances in four social attitudes, with 3-5 iterations of each mood for a total of approximately 30 hours of speech. This dataset focuses on expressive emotions, such as seductive, friendly, dominant, distant. The dataset size is very

small for a deep learning model to be trained properly, similar to the Berlin dataset mentioned in[43]. In the case of IEMOCAP, which has 302 videos altogether across the collection, is made up of 151 recorded dialogue videos with two speakers per session. Similar to the other dataset, the number of data points is low, even though the dataset is unique in the sense that it consists of videos of speech. For preprocessing the authors used 3D log mel spectrograms which had an extra axis, the first derivative of the audio frequencies. Utilizing this more comprehensive representation may improve the model's capacity to extract distinguishing characteristics for speech emotion recognition in comparison to solely using 2D spectrograms, which have been used in most of the other papers. For classification[26] used ACRNN and ACRNN-r models. While both showed below 60% accuracy, their proposed model SSA-CRNN-r showed a higher accuracy of 77%. The use of SSA helped to focus on the important features of the audio and improved training of the model. Another novelty of[26] is the training regularization method, which helped to avoid class collapse. Without regularization the model focused on 1 or 2 of the 4 classes in the dataset causing few of the classes to be recognized better than the others. The regularization in the training phase along with the use of SSA was vital to get improved results. However, the accuracy could have been improved further using datasets with more utterances and more classes for the model to train on. Moine et al.[26] also lacks comparison with the state-of-the-art models such as MLP, SVM, etc. This comparison, along with the performance metrics could have given a better evaluation of the models.

Fayek et al.[13] evaluates deep learning approaches for speech emotion recognition. The paper reviews different state-of-the-art pre-processing steps along with Leave-One-Speaker-Out (LOSO) cross-validation. This was used to assess how well the model can generalize to new, unseen speakers. Moreover, by leaving one speaker out at a time LOSO helps to reduce bias. For classification, the authors used CNNs and LSTM-RNNs and compared them. The authors trained the models on utterance-based features and on frame-based features separately to compare them. Frame-based features include MFCC, LPC, pitch, energy, etc while utterance-level features include duration of speech, voice quality, spectral features etc. The utterance-based feature training averaged an accuracy of 58% roughly where the frame-based averaged 60%. This shows how important frame-based features are in training the deep learning models. However, in practice both types of features are combined and then worked on, like all the papers we mentioned did.

Aouani and Ben Ayed[19] presents a two-stage emotion recognition system that utilizes speech signals along with SVM classifiers. The system initially extracts a 42-dimensional feature set, which includes MFCCs, Zero Crossing Rate (ZCR), Teager Energy Operator (TEO), and Harmonic to Noise Ratio (HNR). To enhance classification, the study employs auto-encoders (AE) for feature dimension reduction and evaluates the outcomes using both basic AE and stacked AE. The performance is assessed through the RML database, which includes six emotions in various languages. Experiments indicate that the RBF kernel achieves the highest accuracy, reaching a recognition rate of 74.07% when utilizing the stacked AE. The research emphasizes that utilizing auto-encoder-based feature selection is more effective than relying solely on the raw 42-dimensional features, especially for emotions such as

"Angry" and "Happy." Nonetheless, the system shows weak performance on "Surprise," with accuracy consistently staying low throughout the experiments. The application of AE greatly improves the performance of SVM by preserving crucial information within a concise feature set. Future work involves testing the model on larger datasets and incorporating audiovisual features to enhance accuracy.

Issa et al.[20] presents a one-dimensional CNN-based model for Speech Emotion Recognition (SER) that incorporates five audio features: MFCC, Chromagram, Mel-spectrogram, Spectral Contrast, and Tonnetz. This blend of features boosts input representation and enhances classification. The model was trained with the RAVDESS and IEMOCAP datasets, achieving an accuracy of 71.61% on RAVDESS and 64.3% on IEMOCAP. The CNN architecture includes convolutional layers, batch normalization, dropout, and ReLU activation, with modifications implemented to improve the performance for each dataset. Methods like time-shifting and incorporating noise improve the model's resilience. While the model demonstrates improved performance compared to earlier techniques on RAVDESS it encounters difficulties with IEMOCAP due to the complexity of the dataset. The paper emphasizes the model's straightforwardness, effectiveness, and adaptability as key advantages over more complex multi-modal systems. Similarly Mustaqeem and Kwon[21] ran their models on IEMOCAP and RAVDESS, but their methology was different. This paper presents a novel SER framework that combines Clustering-Based Key Segment Selection with Deep BiLSTM networks, aiming to enhance both accuracy and computational efficiency. Radial-Based Function Networks (RBFN) are used to pinpoint significant segments from speech utterances, which are converted into spectrograms via STFT for further analysis. ResNet101 extracts sophisticated features, which are normalized before being fed into BiLSTM layers to capture temporal dependencies. A Softmax classifier predicts various emotion categories. The model showed an accuracy of 77.02% on RAVDESS and 72.25% on IEMOCAP, underscoring its strong generalizability. Improving performance that is not tied to specific speakers can be accomplished by normalizing data, and choosing important segments aids in reducing computational complexity.

Deep learning models require a training set with adequate diversity and large size to perform better according to Alomar et al.[38]. Data augmentation is a technique used to increase the number of data points and diversity as data scarcity is a challenge in the real world scenario. Data augmentation is also used to mitigate overfitting problems of deep learning models. There are several data augmentation methods suitable for datasets of SER systems, one of which is cycle-generative adversarial networks (cycle-GAN). The authors of[33] have used cycle-GAN to augment the dataset EMO-DB and used two classifiers SVM and DNN. They obtained an accuracy of 83.33% which was preferable than the baseline methods. Other methods include randomly increasing or decreasing the speed of the audio and randomly varying the volume as done by Braunschweiler et al.[24]. This technique was applied on the IEMOCAP dataset and accuracy improved with data augmentation.

Fayek et al.[13] look into the application of end-to-end deep learning to SER and critically evaluate how each of these architectures may be used in this purpose. Experiments show how feed-forward and recurrent neural network designs, and

their modifications, can be used for paralinguistic speech detection, notably emotion recognition. Convolutional neural networks (ConvNets) outperformed other architectures in terms of discriminative performance. The suggested SER system, which uses minimal speech processing and end-to-end deep learning in a frame-based formulation, produces a high accuracy of 64.78% for the frame-based approach on the IEMOCAP database for speaker-independent SER. The suggested SER system can be combined with automatic speech recognition, creating a unified model for speech processing by combining knowledge of the linguistic and paralinguistic components of speech. On the other hand, Han et al.[10] introduce a novel approach to SER by leveraging Deep Neural Networks (DNNs) for feature extraction and Extreme Learning Machines (ELMs) for emotion classification. This approach results in a significant improvement of 20% with a weighted accuracy of approximately 54.3% on the IEMOCAP database compared to existing state-of-the-art methods.

To address the scarcity of a diverse and large dataset for SER deep learning models, Crowd-sourced Emotional Multi-modal Actors Dataset (CREMA- D) was built consisting of 91 actors from different age groups and ethnicity groups by Cao et al.[9]. The dataset consists of three modalities, audio-visual, audio only and video only. There are 7442 clips for 6 emotions. These clips contain 12 different sentences in English acted by the actors with 3 different intensities: low, medium, and high, for the first sentence. The clips were then rated by crowd-sourcing for validation. While CREMA-D is a simulated dataset, Lotfian and Busso[17] created a more naturalistic dataset named MSP-PODCAST. The audio clips of MSP-PODCAST were extracted from YouTube podcasts in the English language.Techniques like speaker diarization automated some processes during the creation of the dataset. The goal of the authors was to create a dataset with natural interactions and diverse speakers for improving generalization of emotion classifiers. To solve overfitting problems of SER models and adapt real-world scenarios, authors of[25] built LSSED consisting of 820 speakers and 147,025 sentences in the English language. LSSED contains 11 emotion labels with fear and surprise classes each making up only 1% of the data.

2.2 Non-English Datasets

Zhao and Shu[43] explores the use of CNNs combined with gamma classifier (GC) - based error correcting output codes (ECOC) for analyzing emotions in speech. This study leverages CNNs to automatically feature extraction from raw audio signals, capitalizing on their ability to handle complex data patterns and capture spatial hierarchies. The Berlin dataset comprises 535 words articulated by 10 actors with the purpose of expressing one of the emotions: anger, boredom, contempt, anxiety/fear, happiness, sorrow, or neutral. The shEMO dataset consists of 3000 semi-natural utterances, which is comparable to 3 hours and 25 minutes of speech data taken from internet radio programmes. The datasets have numerous issues. Firstly, they have a very limited number of utterances which is not enough to effectively train a model causing overfitting. Secondly, in the Berlin dataset, the actors were asked to express certain emotions using specific words, which is less of a natural speech and more of a forced one. The authors of[43] used mono channeling, resampling and normalization for the preprocessing stage, and for the feature description spectro-temporal modulation and entropy, features were used before using CNN to extract

the important features. Additionally, it introduces a novel gamma classifier-based ECOC approach to enhance the classification accuracy of speech emotions by transforming multiclass classification problems into multiple binary classification problems. The authors validate their proposed method through extensive experiments, demonstrating that the combination of CNNs and gamma classifier-based ECOC significantly improve the accuracy of speech emotion recognition such as 93.3% and 85.78% on different datasets, compared to traditional methods. However, according to our analysis, the use of datasets with such low data points is also responsible for achieving high accuracy, as the model could have memorized the training data and learned patterns. Even though the GC and ECOC can provide high accuracy, they also require high computational power which can be challenging. Zhao and Shu[43] highlights the capacity of deep learning approaches to analyze the emotional characteristics of speech. Furthermore, conducting research on integrating the ECOC model with other established classifiers to enhance the precision of an emotion identification system can be a potential area of investigation.

Mannepalli et al.[14] proposed a model called the Adaptive Fractional Deep Belief Network (AFDBN) to overcome the challenges of SER in a real-world scenario. In this paper, the authors extracted four signal features which are tonal power ratio, spectral flux, MFCC, and pitch chroma, and then based on the feature vectors, they combined the fractional theory with Deep Belief Network (DBN) to form their proposed model AFDBN which finds out the weight vectors optimally. Two datasets, the Telugu database and the Berlin Database of Emotional Speech are used. In the training phase, the input signal extracts four spectral features and is then processed by DBN and Fractional Calculus. The classifier that has been proposed is then used to find weights iteratively. During the testing phase, weights that were learned are utilized to recognize the speakers emotion. The outcomes were compared with those of other systems which include DBN and FDBN and the model proposed achieved a higher accuracy of 97.74% for the Telugu database and 99.17% for the Berlin database. This novel AFDBN algorithm contributed to high accuracy in speaker-independent emotion recognition.

Zhang and Jia[29] and Wang et al.[12] conducted experiments using the EMODB and the CASIA, a Chinese database. Nevertheless, Zhang and Jia[29] introduced a technique that involves merging the Mel-Spectrogram and CapsNet. It is capable of acquiring both the characteristics of frequency domain signals and the characteristics of time domain signals. Subsequently, a CNN is employed to extract additional depth details from the Mel-Spectrogram. Ultimately, the CapsNet and dynamic routing method are employed to discern the profound characteristics acquired from the Mel-Spectrogram in order to enhance the accuracy of classification and recognition. Alternatively, Wang et al.[12] proposed a novel Fourier parameter model that incorporates the perceptual content of voice quality, as well as the first- and second-order differences, for the purpose of recognizing speaker-independent speech emotion. The findings presented by Zhang and Jia[29] demonstrate that the depth features derived from the Mel-Spectrogram are more suited for classifying speech emotions. The utilization of Mel-Spectrogram and CapsNet can enhance the recognition outcome by 2.0% and 9.8% respectively in the EMO-DB and CASIA database, as compared to conventional classification approaches. The findings by

Wang et al.[12] demonstrate that the suggested Fourier parameter (FP) characteristics are efficient in discerning different emotional states in voice signals. They enhance the accuracy of recognizing emotions compared to approaches that utilize MFCC characteristics by 16.2, 6.8, and 16.6 points on EMODB, CASIA, and the Chinese elderly emotion database (EESDB). Specifically, the integration of FP with MFCC leads to an additional enhancement in recognition rates of 17.5, 10, and 10.5 points on the relevant databases.

Akinpelu and Viriri[30] introduce a SER model that employs deep transfer learning using VGGNet for feature extraction. The system uses mel-spectrograms as input and applies NCA to detect relevant emotional features, effectively reducing dimensionality and misclassification. Two classifiers, specifically MLP and SVM, are used to evaluate performance. The study reaches an impressive accuracy of 94.3% on EMO-DB and 97.2% on TESS, surpassing multiple earlier models. Stopping 19 pretrained layers is a notable advantage, as it significantly reduces computational expenses and improves efficiency. Utilizing various datasets, such as TESS, EMO-DB, and a combined set, improves the generalizability of the findings. However, a major drawback is the reliance on acted datasets, which might not truly reflect real-world speech. Using pretrained convolutional neural networks from non-speech tasks, like ImageNet, might limit the precision of feature extraction for recognizing emotions in speech. The performance of feature selection may also be influenced by how sensitive NCA is to noisy data. The paper makes a valuable contribution by optimizing feature selection and attaining high accuracy; however, additional testing on real-world datasets is necessary to confirm its robustness. Similarly, Khalil et al.[16] explain the usage of deep neural networks in SER. They examine various neural network architectures, including CNN, RNN, DNN, DBN, and Autoencoders, and analyze their effectiveness in different emotion recognition tasks. The study examines findings from various well-known speech emotion datasets, such as Emo-DB and RECOLA, where CNN and RNN-based models have attained the highest recognition accuracies. A combination of 1D and 2D CNN-LSTM networks achieved a 92.9% accuracy on the Emo-DB dataset, demonstrating the effectiveness of deep learning techniques in speech emotion recognition tasks.

Liu and Yuan[41] approach speech emotion recognition using two types of features: paralinguistic and spectral. They employed the Berlin dataset. To begin with, the input samples go through pre-processing techniques such as removing silent segments, by utilizing the sample rate to specify useful and useless segments. For feature extraction, MFCC and openSMILE were used for spectral and paralinguistic features respectively. As for openSMILE, it extracted features based on its standard emobase dataset with 998 acoustic features. A feature selection and enhancement procedure was performed by LLCF (for paralinguistic features) and by UDSF (for spectral features). Finally, for classification, MLP and SVM were used to test the performance, for both paralinguistic and spectral features, as well as with and without data augmentation. The results indicate that the MLP performs best using paralinguistic features and data augmentation, while the SVM achieves better results with spectral features.

Zehra et al.[28] created a majority voting ensemble learning method for precise

emotion detection within a corpus, as well as across different corpora. This project aims to address emotions across different languages and empower robots to operate on a worldwide scale. This study utilizes four corpora (EMO-DB, URDU, SAVEE and EMOVO) which encompass a range of languages (Urdu, German, Italian and English). The researchers chose to carry out studies using Urdu as the foundational language for different situations, comparing it to the three other languages. The study used the binary valence (positive and negative) methodology to establish consistent class labels across all datasets for researchers. To feed the classifier, prosodic and spectral data are combined from raw audio recordings. The ensemble learning technique of majority voting is employed to effectively train the model for the correct classification of emotions into their respective categories. It has been noted that several classifiers exhibited varying performance across different languages, prompting the question of which classifier performs optimally across all languages. All languages produced comparable results when the ensemble learning strategy was used with three of the most widely used machine learning algorithms and a majority voting technique. This approach offers the advantage of amalgamating the effects of all classifiers instead of selecting a single classifier and compromising the accuracy of specific language corpora. When testing within the different corpora, the experiments showed an increase in accuracy of 8% for the German corpus, 13% for the Urdu corpus, 5% for the English corpus, and 11% for the Italian corpus. In cross-corpus trials, training on Urdu data and testing on German data produced a 2% improvement; training on Urdu data and testing on Italian data produced a 15% improvement. A 7% increase in accuracy was obtained by training on German data and then testing on Urdu data. In a similar vein, accuracy increased by 3% when testing on Urdu data and training on Italian data and by 5% when testing on Urdu data and training on English data. However, obtaining corpora for many languages from their native habitat would provide a challenge, as there is a limited availability of such resources. Therefore, the task at hand is to choose algorithms that consistently perform well across all languages, in both natural and recorded situations.

Zhu et al.[15] present a new classification technique that integrates DBN and SVM to address the difficulty of accurately recognizing emotions from Chinese speech. This approach is necessary since the intricacies of the Chinese language make it challenging to achieve high accuracy using only one of these methods. Experiments that vary depending on gender are carried out using the emotive speech database from the Chinese Academy of Sciences, which may be found at http://www.datatang.com/data/39277. Initially, DBN is implemented to extract deep speech features from unprocessed vocal samples. Nevertheless, the uppermost layer is substituted by SVM. SVM is trained utilizing deep speech features, which are feature vectors. In other words, the output of the DBN serves as the input for the SVM, and they are combined to form a complete classifier. The suggested classifier combines the advantageous features of both SVM and DBN, and can reach optimal results without requiring extensive parameter adjustment. To accelerate the training process and reduce the Mean Square Error (MSE) between the output and label, a conjugate gradient approach is incorporated. This method calculates the ideal solution of the weight matrix. The proposed classification approach attained a mean accuracy of 95.8%, surpassing that of either solely SVM or DBN. The study showed that low-level

speech features could effectively capture high-level emotional states, indicating that deep features can be used to reflect these states. This strategy is advantageous as it reduces the need for manual optimization of neural networks, resulting in time and effort savings. The combination strategy of using SVM and DBN proved highly user-friendly. Furthermore, the findings indicate that the proposed approach is exceptionally efficient when applied to training databases with restricted dimensions. Similar to this research, Huang et al.[11] also used an SVM classifier along with CNN to classify emotions. The researchers trained their model on two different datasets, SAVEE and EMO-DB. They reached a very high accuracy of 89.7% on the SAVEE and 93.7% on the EMO-DB dataset.

Gharavian and Ahadi[6] used the Farsdat speech corpus. This is in the Farsi language, with 1800 utterances, which were selected for emotion recognition. This is important as speech can vary in many ways in terms of culture, and the nuances of how emotion is portrayed can vary greatly among various languages. Thus it is crucial to make sure that models are able to adapt to such variations. As for the emotions, grief, anger and neutral were chosen as the only emotional states, and were tested in two groups: with and without the neutral state. Using HTK, HMMs were created to label the speech data, which was then used to extract formants using linear prediction spectra, and the pitch as PDA. Important conclusions showed significant differences in vowel durations, formant frequencies and pitch, among the emotions. It is important to note that cepstral mean subtraction has led to an improvement in the overall results. Formants and their slopes, along with pitch frequency, showed potential for improving recognition performance. GMM-based emotion recognition and maximum likelihood approach were both tested. While grief and anger were easily distinguishable from each other, it is not the case for grief and neutral, due to grief having different levels of energy. Anger has a strong effect on energy, thus it is easily recognisable. Upon comparison, the Gaussian Mixture Model (GMM) performed significantly better than the maximum likelihood approach, which used a single Gaussian distribution. However, anger still remains the easiest to recognize with a recognition rate of 98.7%, and grief and neutral having 87.3% and 80.72% respectively, mostly due to their similarities.

Schuller et al.[5] and Nwe et al.[4] use hidden Markov models (HMMs) using specifically created datasets to classify emotions into discrete categories. However, Schuller et al.[5] employs a speech corpus in German and English with seven distinct emotion labels. Then, two continuous hidden Markov models are utilized. The first method uses Gaussian mixture models to classify a global statistics framework of an utterance based on derived aspects of the speech signal's raw pitch and energy contours. A second method increases temporal complexity by utilizing continuous hidden Markov models with several states and low-level instantaneous features instead of global statistics. The results exceeded an 86% recognition rate against a 79.8% human judgment rate. In contrast, Nwe et al.[4] employs an emotion database in Burmese and Mandarin, which includes six archetypal emotions. The voice signals are represented using short-time log frequency power coefficients (LFPC) and classified using a discrete hidden Markov model. The results demonstrate that the suggested approach achieves an average accuracy of 78% and a maximum accuracy of 96% in the classification of six emotions.

2.3 Bangla Datasets

In spite of being the seventh most spoken language globally, Bangla stays classified as a low-resource language for SER tasks due to the insufficient availability of Bangla data on which SER models can be built. Das et al.[31] present the creation of the BanglaSER dataset, which aims to address this gap by providing a collection of Bangla speech-audio recordings that have been acted, specifically for SER. For BanglaSER there were 34 participants between the ages of 19 and 34. There were an equal number of males and females, that is 17 males and 17 females who were non-professional actors. The total dataset includes five classes of emotions, namely angry, happy, neutral, sad, and surprise, and 1467 recordings. The recordings were made by repeating 3 statements in Bangla, three times. While four emotions were done by 34 speakers, the neutral class was done by 27 speakers, leading to an imbalanced dataset. Besides repeating only 3 statements throughout the dataset creates a homogeneous dataset. For recording audio, the authors used their smartphones, laptops, and microphones. Each recording lasted between 3 to 4 seconds and background noise was removed via Audacity Software. To validate the dataset, they asked fifteen native Bangla speakers who did not participate in producing the dataset to recognize the intended emotion. The confusion matrix for BanglaSER of actual versus perceived emotion yields an accuracy of 80.5% where the highest scoring class was the angry emotion with 84.5%. The surprise class was often confused with the happy class and scored the least at 77%. This suggests that the dataset creation should be done alongside proper validation from multiple annotators. The sad emotion was sometimes confused with the neutral class as well. The authors said they eliminated the disgust emotion from the dataset as it often overlaps with the angry emotion making it difficult for recognition. Even though BanglaSER is a popular dataset for SER tasks in Bangla, it lacks diversity and robust validation methods which requires further research in the area.

There have also been attempts to create datasets in languages other than English. Among them, Aziz et al.[39] is notable due to its high accuracy on Bangla datasets. The researchers used SUBESCO and BanglaSER datasets and got average accuracies of 90% and 78% respectively. This is mainly due to the data augmentation techniques they applied. Each class of the dataset contains 306 samples except for the neutral emotion, with 243 recordings. The SUBESCO dataset is not the best dataset to use due to the fact that how it was created. This is an simulated dataset where the authors created an artificial situation so that vocalists spoke a certain speech in a certain emotional class, for which the audio seemed to be more forced and less natural. On the other hand BanglaSER, inspired by the RAVEDESS dataset, itself is an imbalanced dataset. It has more happy and neutral emotion classes than surprise and sad classes. This caused the models to train properly. In the pre-processing stage they used MFCC for feature extraction and Discrete Cosine Transform (DCT) for compressing the signals. It functions as a measurement for the perception of pitch, measuring pitches that are regarded by listeners as being equally distant from each other. Moreover, the authors used data augmentation methods to improve the accuracy of the model. Each sample in this experiment underwent six separate augmentation processes to improve performance by producing synthetic data. It provided assistance by protecting models from the negative effects

of background noise. Time stretching on the other hand is particularly valuable in activities such as adjusting the speed of audio and enhancing time-domain data, where precise management of timing is crucial. Despite all these this caused data integrity issues and overfitting issues if the data became too extreme and unrealistic. For the classification the authors used KNN, RNN, CNN and Random Forest models. Among these models, the fine tuned CNN achieved superior performance because of its suitability for processing time-varying sequential input, such as voice. Additionally, the convolutional layers effectively capture local patterns throughout the time dimension. Aziz et al.[39] lastly suggests exploring the application of alternative Deep Learning methodologies, such as Transformer-based models. Moreover, expanding the suggested approach to include the identification of emotions from several medium, such as facial expressions and text, could enhance the precision of emotion recognition in practical scenarios.

Ali Shuvo and Khan[37] build on the standard SER models aiming at increasing their overall precision through the incorporation of both local and global qualities of human speech, which have not been done in other models. The authors propose a hybrid Convolutional-Transformer (CV-Transformer) model that does this work. The convolutional layers in neural networks excel at extracting local features only, whereas transformers perform better at capturing global characteristics from speech. However, both of these are required for successful SER and thus a hybrid model is proposed. The MFCC was then extracted from these files along with a feature matrix. The authors tested these data with four different deep-learning techniques. The first one is a Fully Connected Network (FCN) with ReLU activations. The final layer in the model consisted of 5 neurons corresponding to the target emotion classes. The second model is the CNN-based approach, which adopts a VGG-like network design. The third approach is the Transformer-based approach in which the MFCC sequence is passed into a transformer encoder. The final and their proposed model consisted of a hybrid CV-Transformer Model. Here the local features were first extracted using the CNN layers. The high-level features selected were subsequently passed through a transformer encoder to obtain global features. The CNN and transformer features were then put together and passed into the final linear classifier layer. The results revealed that the hybrid CV-Transformer model significantly outperformed the other models, achieving a weighted accuracy of 92%. This hybrid model is particularly useful when the target emotion is present only in a few frames of the audio data and the remaining majority of frames are neutral. In their model, they seek to mitigate the risk of losing such important features. The models lowest performance was in distinguishing between the surprise and happy classes, where the surprise class scored 86.05% due to frequent misclassification as happy emotion. The authors did a good job of taking advantage of both convolutional and transformer architectures, which produce a comprehensive feature set required for robust Speech Emotion Recognition and, as a result, sets a new standard in the field through superior accuracy.

Billah et al.[40] produced a novel dataset which they named as the KUET Bangla Emotional Speech (KBES) dataset that addresses the need for a more diverse Bangla resource for SER tasks. This dataset comprises of 900 audio signals collected from 35 professional actors, 15 males, and 20 females, from various age groups in Bangladesh.

It has been collected from numerous Bangla Telefilms, Dramas, TV Series, and Web Series that were publicly available. The authors of this paper categorized their dataset into five emotion classes, which are, happy, sad, neutral, disgust, and angry. Except for the neutral class, the four emotions were classified into High and Low-Intensity levels. This produced a Bangla dataset containing nine classes for emotion classification which includes neutral, sad (High), sad (Low), happy (High), happy (Low), angry (High), angry (Low), disgust (High), and disgust (Low). Unlike datasets such as SUBESCO and BanglaSER, which feature repeating dialogues from a small number of actors, the KBES dataset includes dialogues that are unique across a larger pool of actors, increasing its realism and diversity. The dataset is symmetrical because it has 100 samples of speech audio in each class that is evenly split between male and female actors. The authors collected the emotional speech dialogue audio from sources that were publicly available on social networking sites like YouTube and Facebook and were extracted using VideoProc Converter software. Each clip was then converted to a standard audio format and trimmed to a duration of 3 seconds using the Any Video Converter software. Initially, they collected over 5000 video clips and analyzed them, however, trimmed only 3000 emotional dialogues for further categorization. Finally, they selected 900 clips to be included in the dataset which were chosen by the three authors who used the validation method to confirm the correctness of emotion and intensity labeling. However, relying on a small number of validators and the categorization process being conducted by the same individuals who developed the dataset leads to potential biases and subjectivity that ruin the reliability of the emotion and intensity labels. A more robust approach should include a more diverse and larger group of validators to ensure broader consensus and mitigate individual biases, thus enhancing the overall validity and reliability of the dataset's labels.

Sultana et al.[34] have put forth a deep neural network model which is the combination of a DCNN and BLSTM complete with a time-distributed flatten layer (TDF). Here, two datasets have been employed for two different languages - SUBESCO for Bangla and RAVDESS for English. The Log-Mel spectrograms are then passed into a DCNN as inputs. Next, it undergoes a TDF layer to assess temporal correlations. Finally, the output is passed onto a BLSTM layer, the function of which is to record sequential information. The SUBESCO dataset was implemented with the help of 7000 stimuli that encapsulate seven separate emotions; on the other hand, 2068 files encapsulating five emotions were used for the RAVDESS dataset. After the model was executed, it was found to yield a weighted accuracy (WA) of 82.7% for RAVDESS, and 86.9% for SUBESCO. This considerably surpasses the performances of models such as 4CNN+TDF+LSTM and 7CNN+TDF+BLSTM. The authors, however, cannot claim this comparison between the models to be absolutely fair. Firstly, the RAVDESS dataset is not gender-balanced; second, the number of data points and the emotions being evaluated are not equal across both datasets. For this paper, the experimental setups included cross-lingual training, which is the training of models on one dataset and subsequently testing it on another; multilingual training, which is the utilization of combined data from both datasets; and transfer learning, which is the adaptation of pre-trained models that have frozen datasets to new datasets. However the inspections showed that the cross-corpus tests did not produce accurate results, owing to the datasets being of different languages.

Although the use of transfer learning improved the models performance, the outputs were still not up to the mark. Models that used multilingual datasets also did not achieve very high accuracy. These findings shed light on the challenges of cross-lingual adaptation and underline the need for more diverse datasets that will enhance the model's robustness and generalizability.

2.4 Comparative analysis

Tables 2.1, 2.2, and 2.3 provides a comparison of various SER methods, focusing on the key strengths, weaknesses, and factors that contribute to their performance. Almost all the papers mentioned here used MFCC and log mel-spectrogram for feature selection. The actual difference in accuracies was mainly due to different model architectures used in different papers. Models using ViT-based architectures by Kim and Lee[32] achieved the highest accuracy of 99.76% on Emo-DB by leveraging temporal-frequency correlation and positional encoding through knowledge transfer, which identifies emotional cues effectively and simultaneously reduces complexity. Hybrid CNN-Transformer models by Akinpelu and Viriri[30] and Shuvo and Khan[37] performed well on structured datasets like BanglaSER 92.3% but performed poorly on more complex or improvisational data like IEMOCAP. Traditional MLP, SVM, and CNN models by Akinpelu and Viriri[30], Nwe et al.[4] and El Ayadi et al.[8] showed excellent performance, with accuracies up to 97.2% on TESS, due to the use of transfer learning and stacked features.However, these models struggle with variability in cross-corpus scenarios or real-world data, such as IEMOCAP. Ensemble learning and attention-based models by Zehra et al.[28] and Mountzouris et al.[42] improved cross-lingual recognition but had declining performance on English datasets due to emotional variability. Advanced architectures like Teacher-Student networks by Sultana et al.[34] effectively reduce complexity through knowledge distillation, enhancing temporal and spatial feature learning. In contrast, simpler models like SVM with Autoencoders by Aouani and Ayed[19] and Fourier parameter-based models by Wang et al.[12] achieve moderate accuracies as they rely on a lesser number of features. Overall, high accuracies are achieved by models that combine deep learning, feature extraction, and transfer learning techniques, particularly on smaller, acted datasets like Emo-DB and SAVEE, while performance tends to decline on datasets with spontaneous emotions like IEMOCAP.

Table 2.1: Comparative Analysis (Part 1 of 3)

Ref	Classifier Types	Feature Types	Recognition Accuracy	Dataset	Methods
[30]	MLP, SVM	Mel-spectrogram, NCA-selected features	94.3% (EMO-DB), 97.2% (TESS)	EMO-DB, TESS, Combined	Deep Transfer Learning, VGGNet for feature extraction
[32]	Vision Transformer (ViT)	Log mel-spectrogram	99.76% (EmoDB), 99.47% (SAVEE), 95.24% (CREMA-D)	SAVEE, EmoDB, CREMA-D	Deep CNN, TDF, BLSTM, Transfer Learning
[34]	4CNN + TDF + BLSTM	Log mel-spectrogram	86.9% (SUBESCO), 82.7% (RAVDESS)	SUBESCO, RAVDESS	Teacher-Student Knowledge Transfer, Temporal-Frequency Correlation, ViT with Positional Encoding
[11]	Semi-supervised CNN + SVM	Mel-spectrogram	89.7% (SAVEE), 93.7% (EMO-DB)	SAVEE, Emo-DB	Semi-supervised CNN, SVM for final classification
[20]	1D CNN	MFCC, Chromagram, Mel-spectrogram, Spectral Contrast, Tonnetz	71.61% (RAVDESS), 86.1% (EMO-DB), 64.3% (IEMOCAP)	RAVDESS, EMO-DB, IEMOCAP	Incremental 1D CNN, feature stacking, data augmentation
[37]	Hybrid CNN-Transformer	MFCC	92.3%	BanglaSER	CNN backbone, Transformer encoder, Linear classifier
[21]	CNN + BiLSTM	Spectrogram, CNN-learned features	72.25% (IEMOCAP), 85.57% (EMO-DB), 77.02% (RAVDESS)	IEMOCAP, EMO-DB, RAVDESS	RBF key segment selection, CNN, BiLSTM for temporal modeling

Table 2.2: Comparative Analysis (Part 2 of 3)

Ref	Classifier Types	Feature Types	Recognition Accuracy	Dataset	Methods
[22]	DNN	Zero Crossing, MFCC, Harmonic Ratio, Statistical Features	86.1% (RAVDESS), 96.3% (EMO-DB), 91.7% (SAVEE)	RAVDESS, EMO-DB, SAVEE	Feature selection using F-score, NN classifier
[16]	CNN, RNN, LSTM, DNN, Autoencoder	MFCC, PLP, FBANK, Log-mel Spectrogram	86.99% (IEMOCAP), 92.9% (EMO-DB)	IEMOCAP, Emo-DB, RECOLA, AVEC	Deep models: CNN, RNN, Autoencoder
[19]	SVM	39 MFCC, ZCR, TEO, HNR	72.83% (Basic AE), 74.07% (Stacked AE)	RML Database	SVM with AE for feature selection
[39]	CNN	MFCC	90% (SUBESCO), 78% (BanglaSER)	SUBESCO, BanglaSER	CNN with max-pooling, hierarchical feature extraction
[42]	CNN with Attention Mechanism (CNN-ATN)	MFCC	74% (SAVEE), 77% (RAVDESS)	SAVEE, RAVDESS	CNN with attention mechanism to focus on emotionally intense segments
[28]	Ensemble Learning (Majority Voting)	Spectral and prosodic features	Within-corpus: 96.75% (URDU), 89.75% (EMO-DB), 69.31% (SAVEE), 87.14% (EMOVO)	SAVEE, URDU, EMO-DB, EMOVO	Majority voting ensemble of classifiers for cross-lingual emotion recognition
[15]	SVM + DBN	Deep speech features (DBN), feature vectors	95.8%	Chinese Academy of Sciences Emotional Speech Database	Deep belief network (DBN) extracts deep features, SVM for classification

Table 2.3: Comparative Analysis (Part 3 of 3)

Ref	Classifier Types	Feature Types	Recognition Accuracy	Dataset	Methods
[12]	Fourier Parameter (FP) Model	MFCC + Fourier Parameters	87.5%	EMO-DB, CASIA, EESDB	Fourier parameter model based on voice quality, combined with MFCC for improved recognition

Chapter 3

Methodology and Model Development

Our research is divided into two parts: collecting and building the Bangla Speech Emotion Recognition dataset as well as explanations from human annotators and developing a model based on state-of-the-art pre-trained models to predict and classify emotions from the input audio speech.

3.1 Dataset Creation and Preparation

3.1.1 Dataset Description

There are only a limited number of datasets available for Bangla Speech Emotion Recognition. BanglaSER is a dataset among them. However, it contains only 1,467 recordings across five emotional categories. This is not enough to train robust machine-learning models. It also focuses on a few fixed sentences, which reduces linguistic diversity. Moreover, the speech-audio recordings are collected from non-professional actors, reducing the authenticity of the emotions. All these could have an impact on how well models trained on this data perform in real-world scenarios.

While the largest dataset for Bangla ser system is SUBESCO, containing 7000 utterances, the audios are acted and not natural. The dataset also lacks diversity in terms of speakers and uttered sentences as there are only 20 speakers and 10 sentences throughout the full dataset. This can give rise to overfitting problems as realistic scenarios contains diverse speakers and varied sentences.

Our dataset aims to capture a wide range of genuine human emotions, providing a thorough representation of natural affect. To accomplish this, we gathered audio clips from publicly accessible videos on YouTube, such as speeches, talk shows, news segments, and waaz. These sources offer natural emotional expressions in different contexts, which enhances the dataset's authenticity. Each clip is meticulously labeled according to standard emotional categories, ensuring both consistency and reliability for SER tasks. This method takes advantage of the depth of spontaneous emotions found in real-world interactions, making the dataset a valuable resource for training effective machine learning models. Initially we annotated the emotions, but later they were validated by our group members to strengthen the authenticity.

All of our group members are native Bangla speakers.

Paralinguistic Feature-Based Approach The concentration of our research was solely on paralinguistic emotion detection, i.e. the non-verbal elements of speech that convey emotional information beyond literal word meaning. This focuses on expression from the voice and its tone.

3.1.2 Dataset Creation Methodology

This dataset was collaboratively constructed by all four members of our research team. As native Bangla speakers with expertise in Bangla speech and emotion analysis, we ensured the dataset was meticulously curated. This approach min- imized errors during construction that might arise from a lack of familiarity with emotion recognition nuances. To capture authentic emotions, we sourced speech samples from real-life contexts such as live talk shows, waaz (Islamic sermons), and news broadcasts. These platforms were selected because they feature unscripted, natural conversations rather than fabricated content. The initial dataset comprised of 7,000 utterances, covering seven emotions: happi- ness, sadness, surprise, anger, neutrality, fear, and disgust.

Here is a description of the methodology we used to create the dataset:

1. Audio Collection: We collected the links for videos, such as Waaz, talk shows, news, and Bangla Natoks, that are publicly available on YouTube.

2. Clip Extraction: We used a custom-made interface built for clipping and extracting audio segments from the selected YouTube video to ease our workflow and automate the task. We uploaded the selected video on the interface and extracted the desired video segments that contain one of the seven emotions we are working with. Then we labelled the audio segment with the emotion name and the gender by selecting the correct option from the interface. Following the above method, we extracted 7000 audio clips, 1000 for each emotion, utilizing various videos from YouTube.

3. Standardization of Clip Length: Ensure that each audio clip is exactly 3 seconds long for uniformity.

4. Background Music Removal: Seperation of vocal from non-vocal elements using machine learning models. We used Ultimate Vocal Removal to remove the music, but we did not remove any background noise from the audio files. This made our dataset more realistic.

5. Emotion-Based Labeling: Assign labels to each clip according to predefined emotion categories (Happiness, Sadness, Surprise, Neutral, Fear, Disgust, Anger) following a consistent naming convention.

6. Validation Process: To ensure high-quality labeled data for our audio classification task, we employed Cleanlabs Datalab[46] to automatically detect potential label errors in our dataset. Upon identification of these noisy labels, we re-visited these datapoints and performed relaballing using majority

voting, as well as removed ambiguous ones, reaching our final dataset size of 5250. This methodology has been used earlier by authors in[47]

The dataset creation process is shown in Figure 3.1.

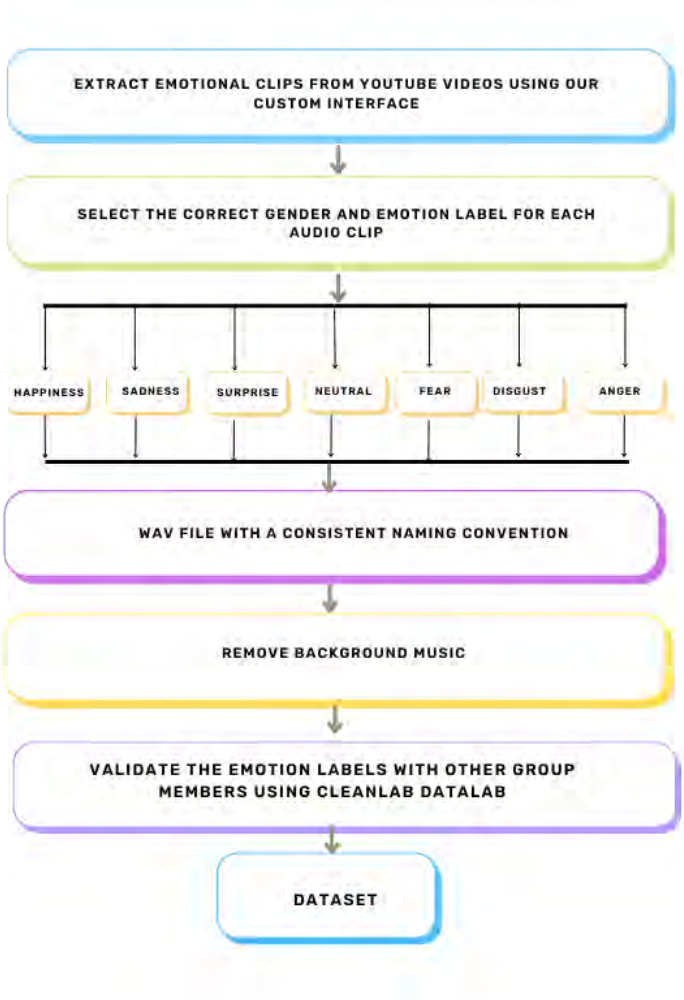


Figure 3.1: Overview of dataset creation process

3.1.3 Indeterminancy

This study acknowledges several inherent limitations based on the inherent nature of human emotions as well as modern research methods. First, emotions are not mutually exclusive categories as they are typically represented in classification models. As a matter of fact, people rarely feel purely discrete and isolated emotions like “pure anger” or “pure sadness.” For example, fear involves elements of surprise and disgust, and bittersweet joy includes elements of happiness and sadness such as the feeling of graduating and saying goodbye to friends. The complex nature of emotions makes precise categorization essentially difficult and perhaps reductionist.

Second, emotions are dynamic phenomena rather than static ones. They possess the ability to change rapidly over short time periods, for example, in the space of seconds of spoken communication or social interaction. Someone can start a sentence in frustration and finish it with sarcasm or humor; likewise, in longer speech

or exchanges, emotional states can change relentlessly. This temporal instability means that static labels might poorly capture the richness of human emotional expression, so time-dependent modeling strategies become necessary to do justice to such subtleties.

Thirdly, emotions are inherently subjective and context-dependent. The same facial expressions or vocal modulations can have different meanings in different cultural, social, or personal contexts. For example, a high vocal pitch can mean anger in one cultural context, but excitement in another. This contextuality implies that models learned in one context (e.g., a dataset, culture, or language) can have limited generalizability to other contexts, as we have seen in our cross-dataset experiments.

Finally, the question of labeling clarity poses a basic problem in the context of emotion recognition research. Human raters tend to show inconsistency in the emotions recognized from specific samples, and inter-rater agreement measures tend to fall below optimal thresholds, reflecting this inherent uncertainty. As such, labels used in emotion data sets need to be seen as approximations and not as absolute values, which impacts training and evaluation approaches.

3.1.4 Dataset Validation

Label Error Detection with Cleanlabs Datalab: Cleanlabs Datalab was used to audit the dataset for label issues. We initialized Datalab with our dataset and labels. We used 5-fold cross-validation to generate out-of-sample predicted probabilities for each audio clip, ensuring that predictions were not influenced by the same data points during training. This step also provided an unbiased estimate of model performance. The out-of-sample predicted probabilities from cross-validation were passed to `find_issues()`, which computed a label quality score for each example (lower scores indicate higher likelihood of label errors). It also produced a boolean flag indicating whether each example was likely mislabeled. The results were summarized in a report, highlighting the most probable label errors.

We manually inspected the audio clips flagged by Cleanlab to confirm mislabeling. Many identified issues were confirmed as true label errors (e.g., a clip labeled "surprise" actually contained "disgust"). Some clips were ambiguous, meaning multiple authors struggled with classification. A few clips were of poor quality and were labelled as faulty.

A subset of speech samples that had previously been determined to include noisy or unreliable labels were used to assess the inter-rater agreement. Then, four different annotators relabeled these samples. We calculated **Fleiss' Kappa** to measure the consistency of this relabeling process, and the result was a score of **0.8096**. The interpretation scale developed by Landis and Koch[2], which is commonly used, states that this score represents "Almost Perfect Agreement" between annotators. This high degree of consistency indicates that the ambiguity in the original labels was effectively resolved by the relabeling effort, producing a more dependable and trustworthy annotation set for further analysis or model training.
Fleiss' Kappa (κ), calculation formula:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{3.1}$$

where $\bar{P}$ is the mean observed agreement and $\bar{P}_e$ is the mean expected agreement by chance.

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \right] \tag{3.2}$$

$$\bar{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \tag{3.3}$$

Correcting these label errors, and removing ambiguous ones improved model accuracy by **8%** on the wav2vec2 model, demonstrating the importance of label quality. This approach provides an efficient, automated way to audit audio datasets without manually reviewing every sample.

3.1.5 Dataset Description

The finalized dataset consists of 5,250 audio recordings in WAV format, each with a fixed duration of three seconds, amounting to a total size of 2.3 GB. All samples are in the Bangla language and encompass seven distinct emotional categories, representing both spontaneous (natural) and simulated (acted) emotional expressions. The audio files adhere to the following naming convention: *Emotion_FileNo_Gender.wav*. The data sources comprise publicly available videos from YouTube, and the dataset includes recordings from both male and female speakers. The table below shows a comparison between our dataset and existing Bangla SER Datasets:

Name of the Dataset	**Language**	**Number of data points**	**Number of unique Sentences**	**Number of emotions**	**Acted or natural**	**Validated by**
Our Dataset	Bangla	5250	5250	7	Mixed	Cleanlab and authors
BanglaSER	Bangla	1467	3	5	Acted	Student/Faculty
SUBESCO	Bangla	7000	10	7	Acted	50 students from SUST
KBES	Bangla	900	Almost 900	5	Acted	Authors

Table 3.1: Comparison of Bangla SER datasets with our dataset.

3.2 Data Processing and Model Pipeline

3.2.1 Overview

The audio was passed through the following preprocessing steps:

1. Data Augmentation: Increasing dataset size and variability by introducing noise, pitch and time shifts. This helped mitigate class imbalance and improve model generalization.

2. Data Scaling: Normalization of data ensures that input features fall within a consistent range, which is critical for stable and efficient training of neural networks. We normalized the audio signal amplitudes to a standard range like [0, 1] to reduce the influence of loudness variations and to ensure convergence during training. For image-based models, such as CNNs, pixel values of mel spectrograms were scaled between 0 and 1.

3. Audio Resampling: Each audio file is loaded and resampled to 16 kHz using librosa, ensuring consistency in sampling rate across the dataset. Stereo files are also converted to mono automatically.

4. Padding or Truncation: All audio signals are adjusted to a fixed length of 32,000 samples. Longer files are truncated, and shorter ones are zero-padded to match the input size required by the model.

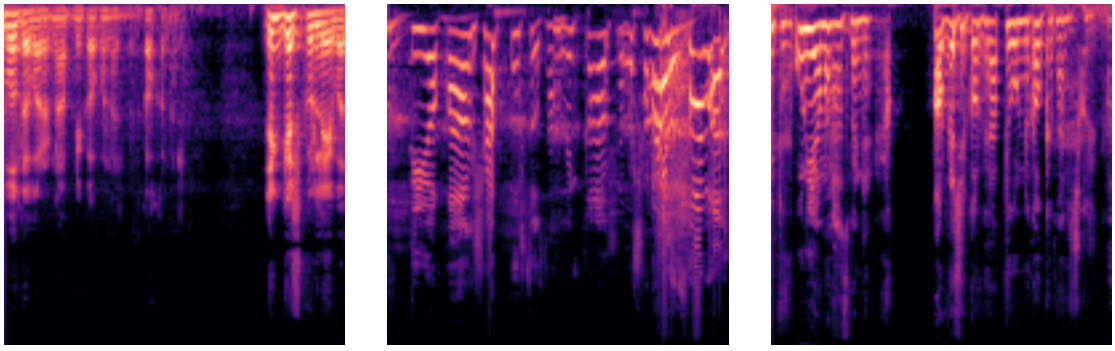
5. Tokenization: In the Wav2Vec2 Base and ExhuBERT models , tokenization was used to transform raw audio waveforms into input tensors usin Wav2Vec2Processor. It normalizes and segments the waveform into fixed-size input representations suitable for the transformer encoder.

6. Mel Spectrogram Generation: A time-frequency representation of audio, in an image format, with the frequency axis scaled according to the mel scale, which is similar to the sensitivity of the human auditory system. This representation was used as input only for the CNN model, transforming audio classification into an image classification task.

7. Label Encoding: Emotion labels are converted into integer classes using a label map. This numeric encoding is necessary for classification by the models.

3.2.2 Feature Extraction

The study employed two complementing feature extraction techniques to capture both learnt and handmade representations of emotion in Bangla speech. The pre-trained facebook/wav2vec2-base model, which employs self-supervised learning to directly extract contextualized features from raw audio waveforms, was used for the transformer-based model. The related Wav2Vec2Processor took care of padding, tokenization, and audio normalization. To anticipate emotions, these high-level features were sent straight to a softmax classification head. In parallel, we used the Librosa package to extract a vast collection of manually created audio features for the CNN-based model. Mel-frequency cepstral coefficients, mel spectrogram, chroma STFT, zero-crossing rate, and root mean square energy were calculated for each 3-second audio clip with a 0.6-second offset. Each audio sample was represented by a single feature vector created by concatenating these features. A comparison of deep contextual representations and conventional signal processing features for speech emotion recognition was made possible by this hybrid technique.



(a) Angry (b) Disgust (c) Fear

Figure 3.2: Mel-spectrograms of some emotions.

3.2.3 Data splitting

To provide unbiased and confidence-based evaluation over models and datasets, all the data was divided into train and test sets. For the CNN model, we employed an 80:20 train-test division using Scikit-learn's train_test_split functionality. Data was shuffled first to eliminate any sequence bias, and then reshaped to the CNN input size by adding an extra dimension to the feature vectors along which they were expanded. For the ExHuBERT and transformer-based Wav2Vec2 models, we employed the HuggingFace DatasetDict format, and the same 80:20 ratio was utilized to divide the training and test sets. We kept the class distributions in both sets well-balanced so that the low-frequency emotion classes were equally well-covered, allowing for better model generalization across emotional classes.

3.2.4 Training Procedure

Every model was trained using architecture-specific steps. The CNN model was trained with the Keras Sequential API, categorical crossentropy loss, and AdamW optimizer and a learning rate of 0.0001. The model was trained for 50 epochs with

a batch size of 64 and validation accuracy was checked after each epoch. Wav2Vec2 and ExHuBERT's training pipeline were constructed with HuggingFace's Trainer class. A few of the most significant training parameters were: learning rate 2e-5, training for 10 epochs, batch size 16, weight decay 0.01. Early stopping with patience of 3 epochs was used in order to avoid overfitting, and custom callbacks were used in order to log train accuracy. The models were trained end-to-end on waveform inputs tokenized by their respective processors, allowing for end-to-end training and emotion classification from raw audio.

3.2.5 Model fine-tuning

Hyperparameter tuning was carried out to fine-tune performance by adjusting the hyperparameters. In the CNN model, various combinations of convolution filters, dropout rates, and dense layer sizes were tried manually in a grid search-like process. The final model utilized a deep Conv1D layer stack in conjunction with several dense layers employing dropout regularization (0.3) with its hyperparameters tuned according to validation accuracy and loss trends. For the transformer models, fine-tuning included modifying the learning rate (ranges tried: 1e-5 to 3e-5), training epochs (515), and regularization hyperparameters like weight decay and warm-up steps. While no formal GridSearchCV was used explicitly, a set of controlled training experiments were made to arrive at the best configurations. Checkpoints with best validation accuracy were chosen for final evaluation.

3.2.6 Models

Wav2Vec2-Base Architecture

Overview Wav2Vec2 is a self-supervised speech representation model developed by Facebook AI that learns directly from raw audio waveforms. It operates in two stages: pretraining and fine-tuning. In the pretraining phase, it masks portions of the latent speech audio representation and learns to predict the proper context through a contrastive loss. The model architecture consists of a convolutional feature encoder, a Transformer-based context network, and a quantization module. This design enables Wav2Vec2 to generate robust and generalized speech representations, which require minimal labeled data for downstream fine-tuning tasks such as speech emotion recognition.

Feature Encoder The feature encoder is a 7-layer 1D convolutional network that transforms raw 16kHz waveforms into lower-resolution latent feature vectors. Each layer uses temporal convolutions with specific kernel sizes and strides to reduce the sequence length while preserving essential acoustic details. Typically, the encoder outputs one vector for every 20 ms of audio. Layer normalization and GELU activation functions are applied after each convolution to stabilize training and introduce non-linearity into the feature extraction process.

Context Network The context network is a Transformer encoder that processes the latent features to produce contextualized speech representations. It comprises 12 to 24 Transformer blocks, each consisting of multi-head self-attention and feedforward layers. Wav2Vec2 employs convolutional relative positional encodings instead

of fixed positional embeddings, allowing the model to attend over the full sequence. This capability enables the extraction of global speech patterns-such as intonation, rhythm, and phrasing-which are critical for emotion recognition tasks.

Pretraining with Contrastive Learning During pretraining, Wav2Vec2 randomly masks parts of the encoder output and replaces them with learned vectors. It then uses a contrastive loss to predict the correct quantized latent representation from among a set of distractors. Quantization is performed using Gumbel-softmax over multiple precomputed codebooks, which transforms continuous features into discrete codes. The loss function encourages high similarity between actual masked targets and model predictions, while discouraging similarity with negative samples. Additionally, a diversity loss term is included to ensure effective utilization of the full codebook space.

Fine-Tuning for Downstream Tasks After pretraining, the model is fine-tuned on labeled data for the target task. For emotion recognition, a classification head is added on top of the final Transformer layer, typically incorporating global average pooling over time. The entire model or selected components is then trained with a supervised loss function, such as cross-entropy, to classify speech features into emotion categories. Fine-tuning enables the model to adapt its generalized speech representations to emotion-specific features like pitch variation, energy, and prosody. This makes Wav2Vec2 particularly well-suited for Bangla speech emotion recognition with limited labeled data.

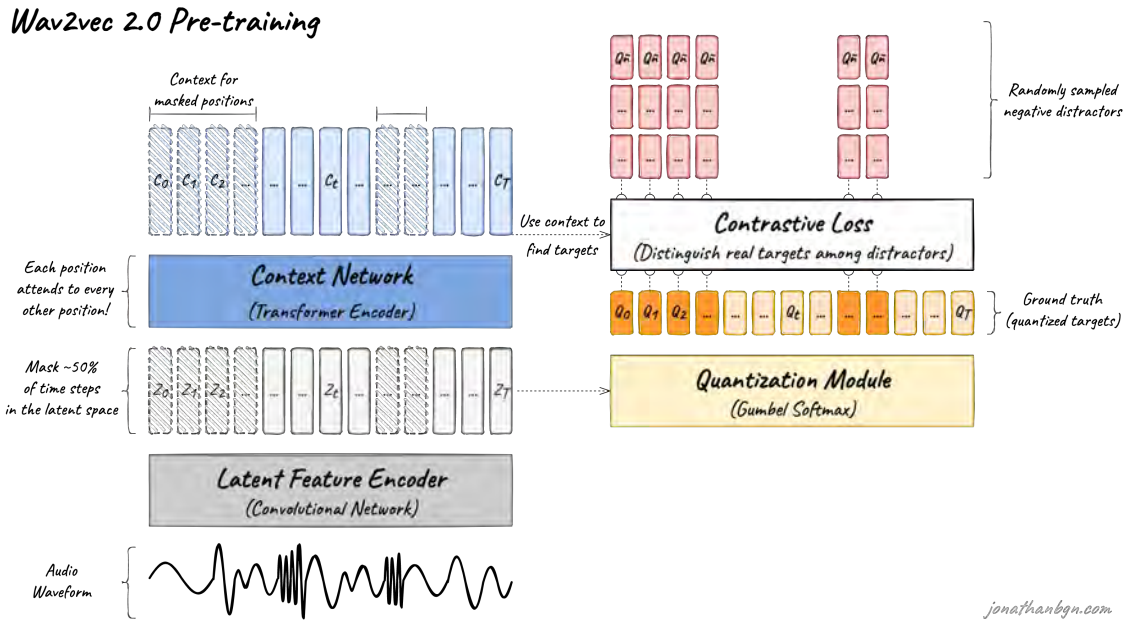


Figure 3.3: Wav2Vec2 model architecture (Source:[23]).

Custom CNN

Overview Originally created for image processing, Convolutional Neural Networks (CNNs) are deep learning models that are now commonly used for 1D data,

including speech, ECG, and other time-series signals. CNNs are ideal for extracting temporal dependencies from sequence data because they use local filters that automatically learn hierarchical feature representations. CNNs greatly increase performance and scalability by eliminating the need for manual feature engineering, in contrast to traditional models.

For our research, we designed a custom 1D CNN architecture tailored specifically for speech-based emotion recognition. The model consists of four convolutional layers with increasing filter sizes, allowing it to capture both low-level and high-level temporal features. Each convolutional layer is followed by batch normalization and max pooling to stabilize training and reduce spatial dimensions while retaining important features. This is then followed by fully connected layers with dropout to prevent overfitting. Our architecture was experimentally optimized to balance complexity and performance for multiclass emotion classification.

Custom CNN Architecture Four convolutional blocks, three dense layers, and an output layer make up the CNN model used in this investigation. A frame or feature from the preprocessed audio is represented by each time step in the network's 1D input of shape; timesteps 1.

Convolutional Blocks Each convolutional block consists of the following sequence:

- A Conv1D layer with ReLU activation
- A BatchNormalization layer
- A MaxPooling1D layer with a pool size of 1

The number of filters increases progressively to enable the network to learn both low-level and high-level representations.

Convolutional Layer The convolutional layer is designed to detect local temporal patterns by sliding filters over the input sequence. In 1D CNNs, each filter moves along the time axis and captures relevant features such as pitch changes or amplitude variations. The number of filters is increased in successive layers to capture more complex patterns. For instance: 128 filters in the first layer detect basic signal changes, 1024 filters in the final block capture more abstract features useful for classification.

Activation Function Each convolutional layer is followed by a non-linear activation function to introduce non-linearity into the model, enabling it to learn complex representations. The Rectified Linear Unit (ReLU) is used after each Conv1D layer. Defined as: The ReLU function $f(x) = \max(0, x)$ increases training efficiency, mitigates the vanishing gradient problem, and promotes sparsity in neuron activations. Its computational simplicity also makes it ideal for deep networks.

Batch Normalization Batch Normalization is applied after every convolutional layer. This technique normalizes the activations of the previous layer for each mini-batch, stabilizing and accelerating the training process. It also acts as a form of regularization by reducing internal covariate shift, thereby improving generalization.

Pooling Layer A MaxPooling1D layer with a pool size of one comes at the end of each convolutional block. Using a pool size of 1 here helps to limit feature redundancy while maintaining temporal resolution, even though pooling with a larger window is more common. In tasks involving emotion recognition, where exact temporal changes in the signal may be crucial, this option is especially helpful.

Dense (Fully Connected) Layers After the convolutional feature extractor, the output is flattened and passed through three fully connected (Dense) layers. These layers integrate the high-level features extracted from previous layers and perform final feature abstraction before classification.

Dropout Regularization Dropout is a regularization technique applied to the dense layers to prevent overfitting. It randomly sets a fraction of input units to zero during training, forcing the network to learn more robust and generalized features. In our model: A 30% dropout is applied after the first dense layer, A 50% dropout is applied after the second and third dense layers.

Output Layer The final dense layer contains 7 neurons with a softmax activation function, corresponding to the 7 emotion classes. The softmax function outputs a probability distribution over the classes:

$$P(y = j) = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}} \tag{3.4}$$

This allows the model to predict the most probable emotion class for each input sample.

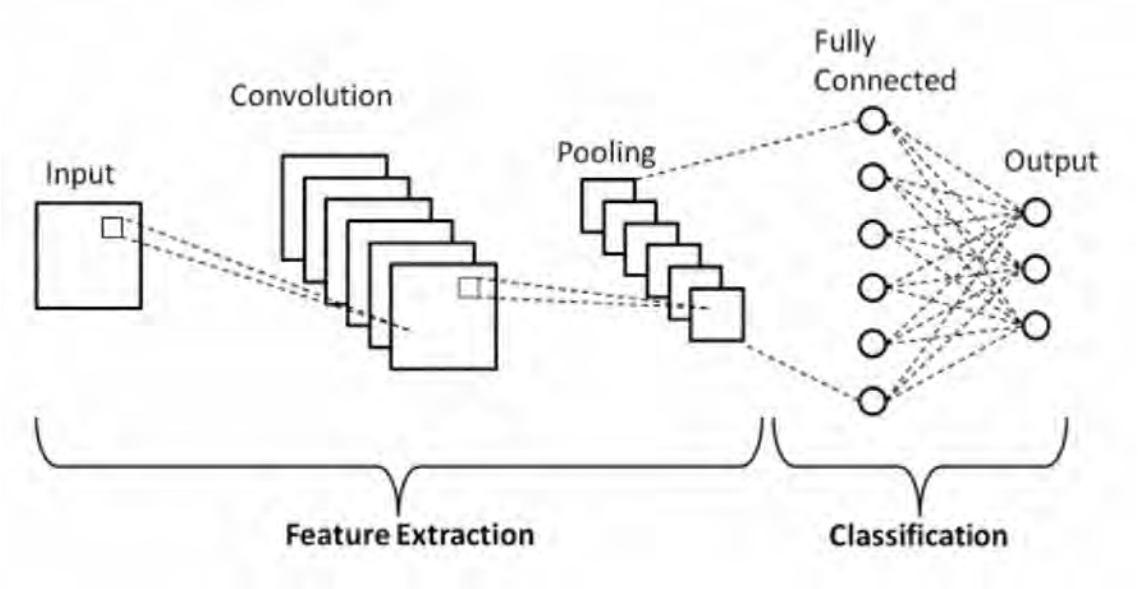


Figure 3.4: CNN model architecture

ExHuBERT

Overview The ExHuBERT model is an advanced model which builds on the foundation of the HuBERT model, which leverages both architectural innovations and large scale, multi-corpus fine-tuning. While large transformer-based models like HuBERT have made improvements through self-supervised pre-training, there remained a gap in developing promising performance across emotional speech corpora without corpus-specific strategies. The ExHuBERT which was pretrained on the EmoSet++ dataset, a multi-cultural corpus consisting of 37 existing emotion datasets. This was designed to maximize diversity in language, culture, and expression.[44]

Architecture: Backbone Block Extension It's novel architecture modification includes the Backbone Block Extension, where each encoder layer in the HuBERT model is duplicated, and the weights of the first copy are frozen. A zero-initialized linear layer and skip connections are added, which allowed the second copy of each layer preserve the original mode's learned representations, while adapting during fine-tuning.[44]

Training Methodology The model was fine-tuned using a round-robin approach, which ensured a balanced exposure to all datasets. Here, only the newly introduced layers are trainable, which assisted in stabilized training and enhances transferability[44].

Performance ExHuBERT demonstrated superior performance in comparison to the original HuBERT as well as other state-of-the-art SER models, on various unseen benchmark datasets, consistently achieving higher Unweighter Average Recall (UAR) scores, indicating strong generalization across domains.[44].

The block extension strategy is particularly effective in the cases where the domain shift between the souce and target datasets is small, but even with larger shifts, ExHuBERT ourperforms previous models. Notably, unfreezing the original HuBERT layers after block extension degrades performance, underscoring the importance of freezing pre-trained weights for transfer learning. In terms of convergence, ExHuBERT not only achieves higher accuracy but also converges faster during training on new datasets compared to models without any pre-training[44].

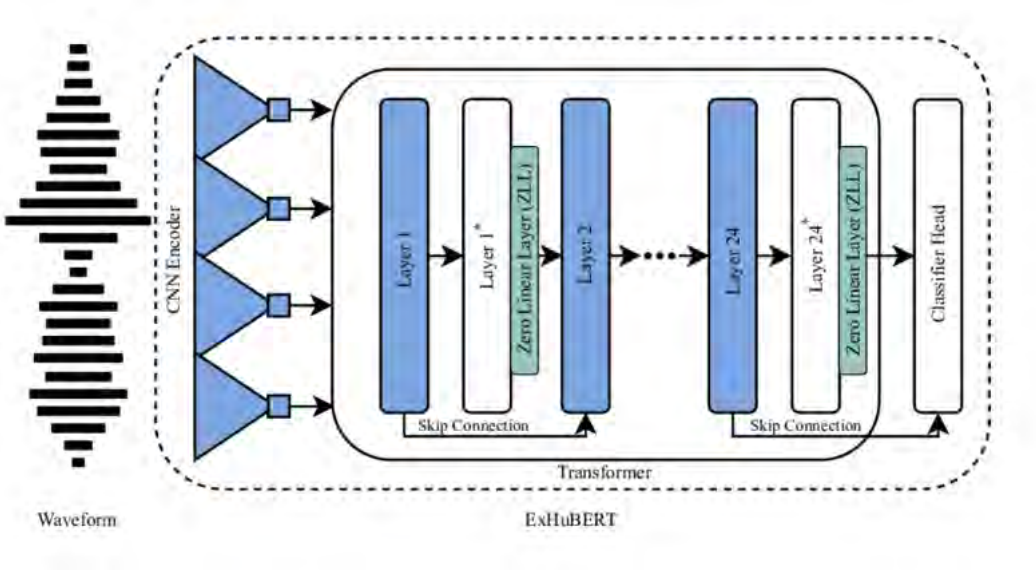


Figure 3.5: Overview of ExHuBERT's architecture (Source:[45]).

Chapter 4

Outcome and Discussion

4.1 Evaluation

We tested several different models on our dataset: CNN, wav2vec2 base and Ex-HuBERT.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

Model	**Dataset**	**Accuracy**	**Macro F1**	**Macro Precision**	**Macro Recall**
CNN	KBES	75%	74%	76%	74%
	BanglaSER	**88%**	**88%**	**88%**	**88%**
	SUBESCO	**88%**	**88%**	**88%**	**88%**
	Our Dataset	82%	82%	82%	82%
Ex-HuBERT	KBES	67%	68%	72%	67%
	BanglaSER	**88%**	**88%**	**89%**	**88%**
	SUBESCO	**93%**	**93%**	**93%**	**93%**
	Our Dataset	82%	82%	85%	82%
Wav2Vec2-Base	KBES	77%	78%	78%	78%
	BanglaSER	**91%**	**91%**	**91%**	**91%**
	SUBESCO	**97%**	**97%**	**97%**	**97%**
	Our Dataset	88%	88%	88%	88%

Table 4.1: A comparison of 3 models across various Bangla SER datasets.

On the BanglaSER dataset, Wav2Vec2 achieved 91% accuracy, followed by CNN and ExHuBERT both at 88%. For the KBES dataset, Wav2Vec2 was the best-performing model at 77%, with CNN at 75% and ExHuBERT at 67%. On the SUBESCO dataset, Wav2Vec2 showed exceptional performance at 97% accuracy, ExHuBERT at 93%, and CNN at 88%. For Our dataset, CNN performed at 82% accuracy, ExHuBERT also at 82%, while Wav2Vec2 was the best at 88%.

4.2 Hyperparameters

Hyperparameter	**Value**
Batch size	64
Optimizer	AdamW
Weight Decay	1×10^{-2}
Early Stopping	patience = 7
Learning Rate	1×10^{-4}
Epochs	50

Table 4.2: Hyperparameters used for training CNN

Hyperparameter	**Value**
Batch size	16
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	1×10^{-2}
Early Stopping	patience = 3
Epochs	20

Table 4.3: Hyperparameters used for training ExHuBERT

Hyperparameter	**Value**
Batch size	16
Optimizer	AdamW
Learning Rate	2×10^{-5}
Early Stopping	patience = 3
Weight Decay	1×10^{-2}
Epochs	20

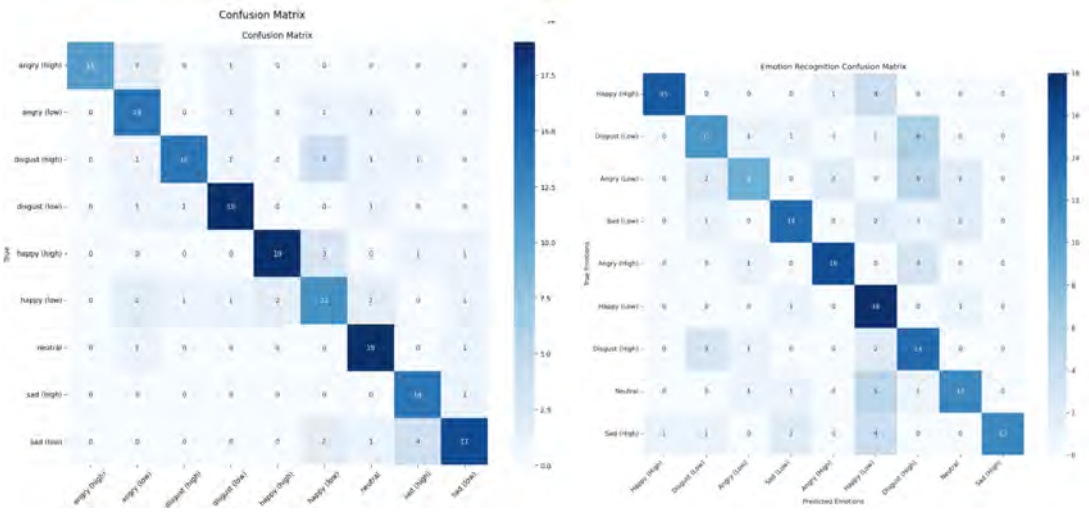
Table 4.4: Hyperparameters used for training Wav2Vec2-Base

The configurations of training for the Convolutional Neural Network (CNN), ExHuBERT, and Wav2Vec2 Base models are outlined in Tables 4.2, 4.3, and 4.4, respectively. Every model employed the AdamW optimizer with a weight decay of 1×10^{-2} for preventing overfitting through the promotion of reduced weight magnitudes. Each model's hyperparameters were chosen to accommodate the particular characteristics and computational needs of its respective architecture. The CNN, which was lighter and less memory-intensive, was trained using a larger batch size of 64 and a higher learning rate of 1×10^{-4}, allowing for faster convergence within 50 epochs, with an early stopping patience of 7 to prevent overtraining. In contrast, the ExHuBERT and Wav2Vec2 Base models, with their significantly larger depths and more extensive pretraining, were fine-tuned using a lower batch size of 16 and a smaller learning rate of 2×10^{-5} in an attempt to ensure stability in training and prevent catastrophic forgetting. These models were trained for up to 20 epochs, with

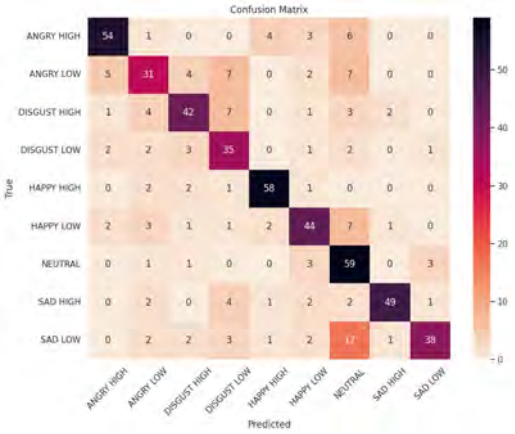
early stopping patience set at 3, given their inclination towards earlier convergence as a result of previous pretraining. In general, the hyperparameter selection process was carried out through empirical tuning, guided by the underlying architecture and optimization dynamics of each model.

4.3 Results on Bangla SER Datasets

4.3.1 Results on KBES



(a) Wav2Vec2 Base (b) ExHuBERT

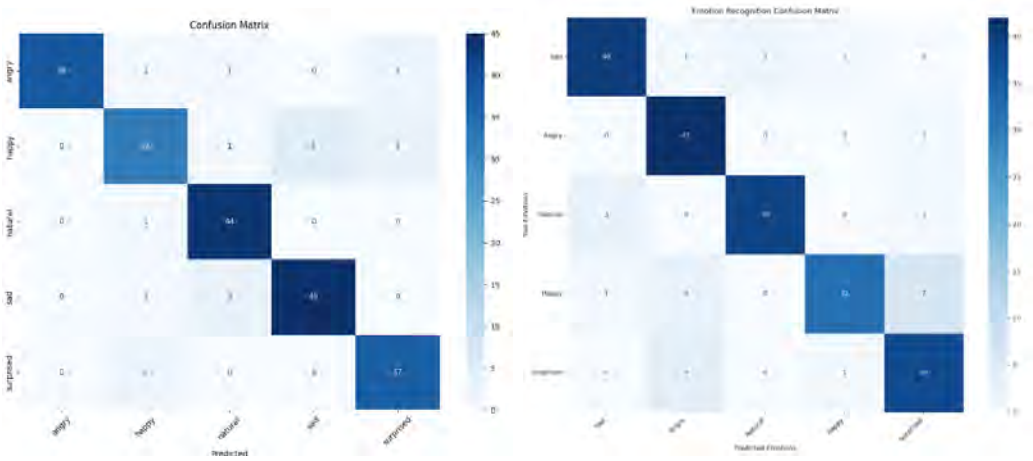


(c) CNN

Figure 4.1: Confusion matrices for each model on KBES

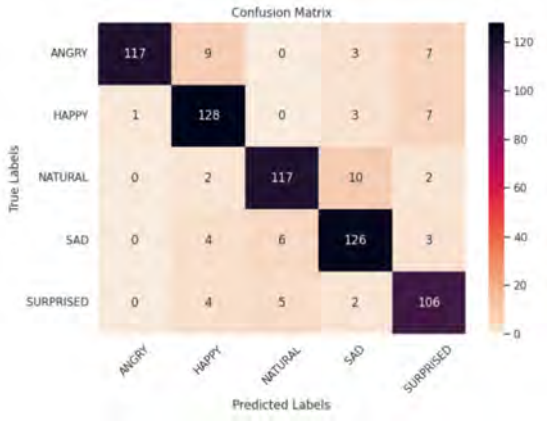
The KBES dataset proves tougher with its intensity-emotion variants. CNNs major confusions on KBES are calling Sad Low as Neutral, several emotions (Angry High/Low, Happy Low) as Neutral, and Disgust High/Angry Low as Disgust Low. ExHuBERT confuses intensity variations, calling Disgust Low and Angry Low as Disgust High, and Neutral/Sad High as Happy Low. Wav2Vec2 exhibits a distinctive trend of downgrading high-intensity emotions, calling both Disgust High and Happy High as Happy Low.

4.3.2 Results on BanglaSER



(a) Wav2Vec2 Base

(b) ExHuBERT

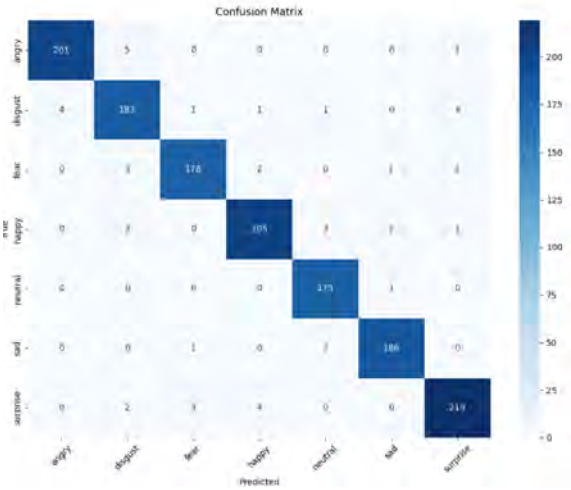


(c) CNN

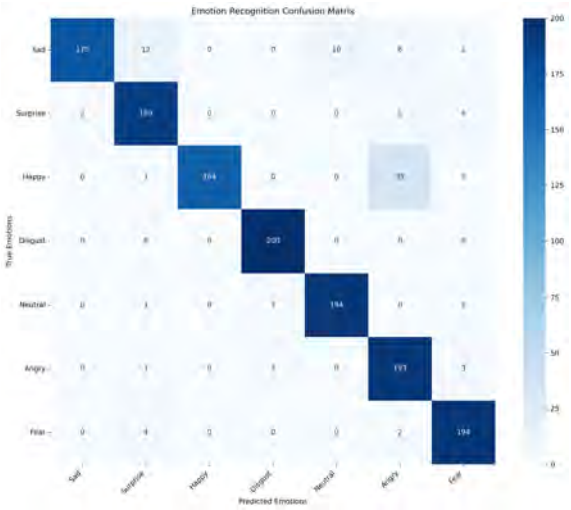
Figure 4.2: Confusion matrices for each model on BanglaSER

On the BanglaSER dataset, all three models register high diagonal domination, other than minor cases, for example when the CNN misclassified 'Natural' as 'Sad' and 'Angry' as 'Happy'.

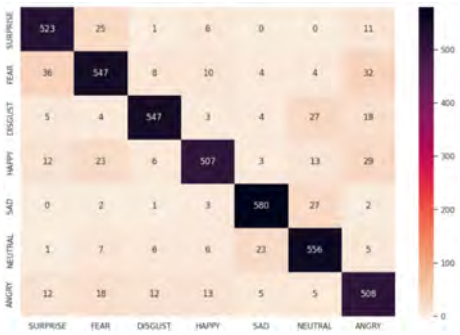
4.3.3 Results on SUBESCO



(a) Wav2Vec2 Base



(b) ExHuBERT

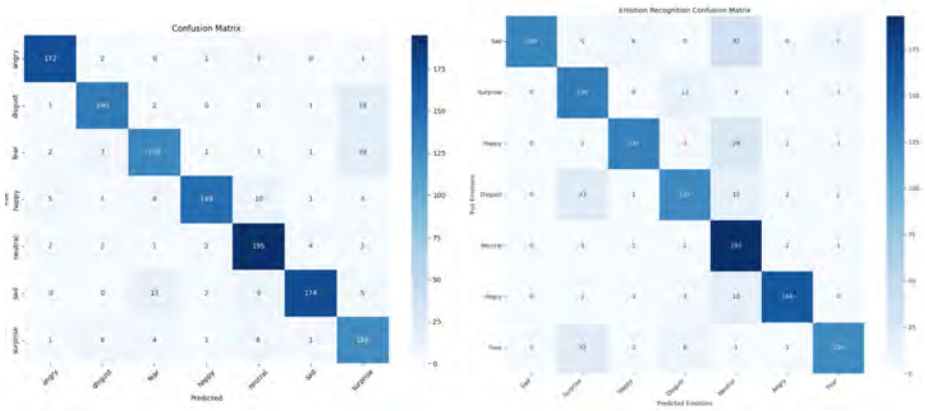


(c) CNN

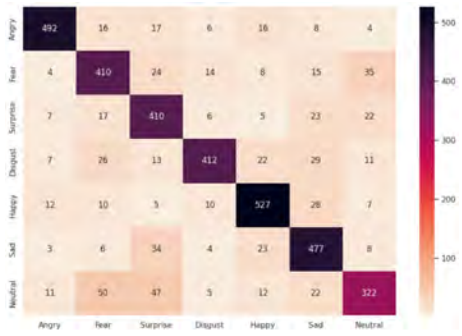
Figure 4.3: Confusion matrices for each model on SUBESCO

On the SUBESCO dataset, all three models register exceptionally high performance, all maintaining diagonal domination. Notably, in the CNN, Fear and Surprise is confused, along with Fear and Happy misclassified in Angry in more frequent cases. Disgust is sometimes misclassified as Neutral.

4.3.4 Results on our dataset



(a) Wav2Vec2 Base (b) ExHuBERT



(c) CNN

Figure 4.4: Confusion matrices for each model on Our Dataset

Our dataset exhibits variant confusion trends: CNN confuses Neutral with Fear/Surprise and Sad with Surprise; ExHuBERT over-predicts Neutral (misclassifying Sad, Happy, Disgust as Neutral) and confuses Fear/Disgust as Surprise; while Wav2Vec2 continues to exhibit confusion trends with Disgust/Fear as Surprise, Happy as Neutral, and Sad as Fear.

Table 4.5: Custom CNN Classification Report on our Dataset

	precision	**recall**	**f1-score**	**support**
Angry	0.92	0.88	0.90	559
Disgust	0.77	0.80	0.78	510
Fear	0.75	0.84	0.79	490
Happy	0.90	0.79	0.84	520
Neutral	0.86	0.88	0.87	599
Sad	0.79	0.86	0.82	555
Surprise	0.79	0.69	0.73	469
accuracy			0.82	3702
macro avg	0.82	0.82	0.82	3702
weighted avg	0.83	0.82	0.82	3702

Table 4.6: ExHuBERT Classification Report on our Dataset

	precision	**recall**	**f1-score**	**support**
Sad	0.9480	0.8770	0.9111	187
Surprise	0.8194	0.7471	0.7815	170
Happy	0.8873	0.7730	0.8262	163
Disgust	0.9306	0.7746	0.8454	173
Neutral	0.6498	0.9650	0.7767	200
Angry	0.9848	0.7027	0.8202	185
Fear	0.7120	0.8718	0.7839	156
accuracy			0.8185	1234
macro avg	0.8474	0.8159	0.8207	1234
weighted avg	0.8472	0.8185	0.8213	1234

Table 4.7: Wav2Vec2 Classification Report on our Dataset

	precision	**recall**	**f1-score**	**support**
Angry	0.9399	0.9609	0.9503	179
Disgust	0.8750	0.8589	0.8669	163
Fear	0.8442	0.8075	0.8254	161
Happy	0.9551	0.8613	0.9058	173
Neutral	0.8705	0.9375	0.9028	208
Sad	0.9560	0.8571	0.9039	203
Surprise	0.7200	0.8571	0.7826	147
accuracy			0.8801	1234
macro avg	0.8801	0.8772	0.8768	1234
weighted avg	0.8857	0.8801	0.8811	1234

A side-by-side comparison of the high-level summary figures (accuracy, macro average, and weighted average F1-score) easily places Wav2Vec2 as the top-performing model. It achieves an overall accuracy of 88.01% and a weighted average F1-score of 0.8811. It beats the Custom CNN (82% accuracy, 0.82 weighted F1-score) and ExHuBERT (81.85% accuracy, 0.8213 weighted F1-score).

Although the Custom CNN and ExHuBERT models are both similarly high performing overall, they are outperformed by Wav2Vec2, respectively. Wav2Vec2 achieves the highest overall accuracy rate while remaining well-balanced and stable among various emotion categories. It achieves an F1-score of above 0.90 for 'Angry', 'Happy', and 'Sad' predictions, indicating high recall and precision. The worst category is 'Surprise' with an F1-score of 0.7826; however, this performance is nevertheless remarkable and outperforms other models' worst categories. The Custom CNN is highly performing on 'Angry' (0.90 F1-score) and 'Happy' (0.84 F1-score), but its performance is very poor on 'Surprise,' which is way below at 0.73 F1-score. This indicates the model architecture has some trouble separating the features of surprise from other categories.

ExHuBERT shows a mixed but interesting characterization. The model is extremely accurate in its prediction of the emotion 'Angry' (0.9848) and 'Sad' (0.9480), pointing to a very high probability of being correct in predictions for these classes. The reverse is the case for its recall on these same emotions, especially 'Angry' (0.7027), meaning that a lot of instances of this emotion are not detected. 'Happy' is the least well-classified emotion, with an F1-score of 0.8262. This model seems to be more task-specialized, with a proficiency for spotting exact emotions with precision but at the expense of general detection rates.

Overall, the Wav2Vec2 model is the top performer for the specified emotion classification task based on data that has been analyzed. It achieves both high accuracy and a good, balanced F1-score across almost all emotional classes. The ExHuBERT model has promise for high-accuracy applications where it is imperative to maintain false positives at a low level for specific emotions such as 'Angry' or 'Sad,' despite this resulting in missing actual instances. The Custom CNN gives a good baseline performance but does not have the fine-tuned accuracy and balance that the Wav2Vec2 model has.

4.3.5 Cross-Dataset Evaluation Results

A cross-dataset study was conducted to see how generalizable our models were. In particular, our gathered dataset was used to test models trained on the SUBESCO dataset, and vice versa. We were able to assess the transferability of learned representations and patterns from one dataset to another using varying recording settings and speaker attributes.

Table 4.8: Training Wav2Vec2-Base on SUBESCO and testing on Our Dataset

Category	**precision**	**recall**	**f1-score**	**support**
angry	0.3211	0.4700	0.3815	932
disgust	0.3651	0.1847	0.2453	850
fear	0.2918	0.5772	0.3877	816
happy	0.3068	0.0889	0.1379	866
neutral	0.5921	0.2252	0.3263	999
sad	0.3505	0.5805	0.4371	925
surprise	0.3177	0.2433	0.2756	781
accuracy			0.3396	6169
macro avg	0.3636	0.3385	0.3131	6169
weighted avg	0.3692	0.3396	0.3153	6169

Table 4.9: Training Wav2Vec2-Base on Our Dataset and testing on SUBESCO

Category	**precision**	**recall**	**f1-score**	**support**
angry	0.9320	0.0960	0.1741	1000
disgust	0.1781	0.0390	0.0640	1000
fear	0.4412	0.2890	0.3492	1000
happy	0.3542	0.0340	0.0620	1000
neutral	0.5497	0.6640	0.6014	1000
sad	0.6120	0.6040	0.6080	1000
surprise	0.2334	0.8710	0.3681	1000
accuracy			0.3710	7000
macro avg	0.4715	0.3710	0.3181	7000
weighted avg	0.4715	0.3710	0.3181	7000

The accuracy of Wav2Vec2 fell to 37% (trained on our dataset, evaluated on SUBESCO) and 34% (trained on SUBESCO, evaluated on our dataset), indicating a fundamental restriction in the generalizability of emotion recognition.

4.4 Discussion

4.4.1 Comparison with existing results

Our results demonstrate significant improvements over previous benchmarks established in the literature. On the BanglaSER dataset, our models achieved accuracies of 88-91%, substantially outperforming the previously reported accuracy of 80.5%. Similarly, on the SUBESCO dataset, our models reached 88-97% accuracy, surpassing the earlier benchmark of 90%. These improvements over state-of-the-art approaches like VGG16 and DCNN suggest that modern transformer-based architectures (ExHuBERT, Wav2Vec2) and well-designed CNN implementations provide superior feature extraction capabilities for emotion recognition tasks.

4.4.2 Unexpected Results and Their Impact

Our evaluation revealed a number of surprising findings that require thorough investigation. Most importantly, examination of varied datasets illustrated unforeseen shortcomings of generalization when models were trained on one dataset and later evaluated on another. In particular, training Wav2Vec2 on our self-created dataset and then evaluating it on SUBESCO saw the accuracy drop to a mere 37%. Similarly, training on SUBESCO and validating on our dataset saw only 34% accuracy. This severe deterioration in performance was unforeseen, especially given the strong individual performances of each dataset, and highlights a core limitation within the domain of emotion recognition: the lack of cross-dataset generalizability.

Another surprising and captivating finding was the significant improvement (an 8% accuracy increase) achieved with CleanLab to detect and fix incorrect labels in our own proprietary dataset. This observation is significant, as it suggests that even carefully curated emotion datasets contain high levels of labeling errors, which have

a significant impact on model performance. The success of the automated noise detection highlights that traditional human annotation processes, while necessary, are insufficient to guarantee the creation of high-quality emotion recognition datasets.

4.4.3 Future Research Directions

Our findings suggest a number of promising directions for future research that might overcome current limitations and further develop the field. First, future research should consider exploring multimodal approaches that pair spoken interaction with natural language processing of lexical features with facial expression-derived visual cues. The integration of these modalities would provide richer contextual information and aid in the disambiguation of emotional expressions that would be difficult to identify from audio only.

Additionally, research efforts should focus on developing culturally specific emotion recognition models and datasets. The results of our research show how important the cultural context is for emotion expression and recognition, and universal approaches will not be effective for global use. Future work should systematically examine how emotional expressions vary across cultures and develop solutions for the creation of culturally appropriate emotion recognition systems.

Another important direction of investigation regards the development of robust evaluation protocols that can enable cross-dataset comparisons and the corresponding methodologies related to transfer learning. The poor generalizability across datasets, as evident from our work, further emphasizes the need for models that can more effectively adapt to new domains and datasets. Investigating domain adaptation techniques, few-shot learning, and meta-learning strategies would greatly improve the real-world usability of emotion recognition systems.

4.4.4 Wider Consequences and Significance

This research marks significant progress in the field of speech emotion recognition, particularly in multilingual and multicultural environments. The main contribution involves the improvement of emotion recognition in Bengali speech by designing and testing models based on datasets that accurately reflect particular cultural settings. This research aims to solve a critical gap in emotion recognition research, which has largely been centered on Western languages and cultural contexts. By showing that emotional cues are inherently tied to cultural context and require specific learning methods, this research opens the door to designing more universal and inclusive emotion recognition systems.

The developments highlighted here have far-reaching implications for various fields, including human-computer interaction systems, mental well-being monitoring devices, and learning technologies for Bengali speakers. In enabling the development of strong and culturally adapted emotion recognition abilities, this study helps in the creation of more efficient and inclusive technological interventions that can better serve the needs of various global populations. Our methodological observations regarding data collection, cross-dataset evaluation, and noise mitigation also pro-

vide important suggestions for future research into emotion recognition in various languages and cultural settings.

4.5 Research Limitations

Although this research makes meaningful contributions to recognition of Bangla speech emotion, we acknowledge several limitations: There was a notable con- fusion among the emotional categories, fear, surprise, and disgust. The acoustic similarities and subtle distinctions between these emotions made them difficult to differentiate both during annotation and during model training, affecting the classification accuracy of these classes. Model trainings were carried out with limited GPU resources and time restrictions, which constrained the exploration of more complex models, larger datasets, and extended training durations that could potentially improve performance. Emotional annotation is inherently subjective, as perception of emotion can vary between individuals. Despite following consistent guidelines, some degree of human bias or inconsistency in labeling is unavoidable and may have influenced the quality of the dataset. The models demonstrated poor generalizability in cross-corpus evaluations, indicat- ing that the models performance significantly drops when tested on external datasets differing in recording conditions, speakers, or content. This highlights the need for further work to improve the robustness of models across diverse data sources.

Conclusion

This research presents a thorough study on improving speech emotion recognition in Bangla, a low-resource and culturally distinct language. By creating a diverse real-life dataset and evaluating three deep learning models CNN, ExHuBERT, and Wav2Vec2 across four Bangla SER datasets, the study shows that Wav2Vec2 consistently achieves the highest accuracy, reaching up to 97and surpasses previous benchmarks. However, a major challenge arose from poor cross-dataset generalizability, as performance dropped sharply when models were tested outside their training area. The successful use of CleanLab to find mislabeled data, resulting in an 8% accuracy improvement, further highlights the importance of data quality. These findings stress the need for culturally specific datasets, multimodal emotion recognition methods, and strong domain adaptation techniques. Overall, this work contributes significantly to creating more accurate, inclusive, and culturally aware SER systems. It has potential applications in virtual assistants, mental health monitoring, and educational technologies designed for Bangla-speaking users.

Bibliography

[1] Government of India, Ministry of Law and Justice, *Copyright rules, 1957 section 52*, Accessed: 2025-06-23, Office of the Copyright Administrator, 1957. [Online]. Available: https://copyright.gov.in/documents/copyrightrules1957.pdf.

[2] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. DOI: 10.1177/001316446002000104. [Online]. Available: https://www.jstor.org/stable/2529310.

[3] Government of the People's Republic of Bangladesh, *Copyright act, 2000 section 36: Use for the purpose of criticism, education, or research*, http://bdlaws.minlaw.gov.bd/act-846/section-30171.html, Accessed: 2025-06-23, 2000.

[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003, ISSN: 0167-6393. DOI: https://doi.org/10.1016/S0167-6393(03)00099-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639303000992.

[5] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 2, 2003, pp. II–1. DOI: 10.1109/ICASSP.2003.1202279.

[6] D. Gharavian and S. Ahadi, "Recognition of emotional speech and speech emotion in farsi," in *The Proceedings of international symposium on Chinese spoken language processing*, Citeseer, vol. 2, 2006, pp. 299–308.

[7] Z. Yang and J. Laaksonen, "Regularized neighborhood component analysis," in *Image Analysis: 15th Scandinavian Conference, SCIA 2007, Aalborg, Denmark, June 10-14, 2007 15*, Springer, 2007, pp. 253–262.

[8] 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, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2010.09.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320310004619.

[9] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014. DOI: 10.1109/TAFFC.2014.2336244.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Sep. 2014.

[11] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14, Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 801–804, ISBN: 9781450330633. DOI: 10.1145/2647868.2654984. [Online]. Available: https://doi.org/10.1145/2647868.2654984.

[12] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015. DOI: 10.1109/TAFFC.2015.2392101.

[13] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[14] K. Mannepalli, P. N. Sastry, and M. Suman, "A novel adaptive fractional deep belief networks for speaker emotion recognition," *Alexandria Engineering Journal*, vol. 56, no. 4, pp. 485–497, 2017, ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2016.09.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016816302484.

[15] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from chinese speech for smart affective services using a combination of svm and dbn," *Sensors*, vol. 17, no. 7, p. 1694, 2017. DOI: 10.3390/s17071694. [Online]. Available: https://doi.org/10.3390/s17071694.

[16] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019. DOI: 10.1109/ACCESS.2019.2936124.

[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019. DOI: 10.1109/TAFFC.2017.2736999.

[18] L. Tarantino, P. N. Garner, A. Lazaridis, *et al.*, "Self-attention for speech emotion recognition.," in *Interspeech*, 2019, pp. 2578–2582.

[19] H. Aouani and Y. Ben Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020. DOI: 10.1016/j.procs.2020.08.027. [Online]. Available: https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S1746809420300501.

[20] D. Issa, F. M. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101 894, 2020. DOI: 10.1016/j.bspc.2020.101894.

[21] M. S. Mustaqeem and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020. DOI: 10.1109/ACCESS.2020.2990405.

[22] H. A. Abdulmohsin, H. B. Abdul Wahab, and A. M. J. Abdul Hossen, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers & Electrical Engineering*, vol. 93, p. 107 172, 2021. DOI: 10.1016/j.compeleceng.2021.107172.

[23] J. Bgn, *An illustrated tour of wav2vec 2.0*, Online blog post, Figure 1 (Wav2Vec 2.0 architecture overview), Sep. 2021. [Online]. Available: https://jonathanbgn.com/2021/09/30/illustrated-wav2vec-2.html (visited on 06/23/2025).

[24] N. Braunschweiler, R. Doddipatla, S. Keizer, and S. Stoyanchev, "A study on cross-corpus speech emotion recognition and data augmentation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia, 2021, pp. 24–30. DOI: 10.1109/ASRU51503.2021.9687987.

[25] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: A large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 641–645. DOI: 10.1109/ICASSP39728.2021.9414542.

[26] C. L. Moine, N. Obin, and A. Roebel, "Speaker attentive speech emotion recognition," *arXiv preprint arXiv:2104.07288*, 2021.

[27] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021. DOI: 10.1109/ACCESS.2021.3068045.

[28] W. Zehra, A. Javed, Z. Jalil, *et al.*, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex Intell. Syst.*, vol. 7, pp. 1845–1854, 2021. DOI: 10.1007/s40747-020-00250-4. [Online]. Available: https://doi.org/10.1007/s40747-020-00250-4.

[29] W. Zhang and Y. Jia, "A study on speech emotion recognition model based on mel-spectrogram and capsnet," in *2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, 2021, pp. 231–235. DOI: 10.1109/IAECST54258.2021.9695802.

[30] S. Akinpelu and S. Viriri, "Robust feature selection-based speech emotion classification using deep transfer learning," *Applied Sciences*, vol. 12, no. 16, p. 8265, 2022. DOI: 10.3390/app12168265. [Online]. Available: https://www.mdpi.com/2076-3417/12/16/8265.

[31] R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "Banglaser: A speech emotion recognition dataset for the bangla language," *Data in Brief*, vol. 42, p. 108 091, 2022, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2022.108091. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235234092200302X.

[32] J.-Y. Kim and S.-H. Lee, "Accuracy enhancement method for speech emotion recognition from spectrogram using temporal frequency correlation and positional information learning through knowledge transfer," *IEEE Access*, 2022, Supported by the National Research Foundation of Korea (NRF) under Grant NRF-2022R1F1A1066371 and by the Regional Innovation Strategy (RIS).

[33] A. Shilandari, H. Marvi, H. Khosravi, *et al.*, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," *SIViP*, vol. 16, pp. 1955–1962, 2022. DOI: 10.1007/s11760-022-02156-9. [Online]. Available: https://doi.org/10.1007/s11760-022-02156-9.

[34] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," *IEEE Access*, vol. 10, pp. 564–578, 2022. DOI: 10.1109/ACCESS.2021.3136251.

[35] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," *Sensors*, vol. 22, no. 4, p. 1414, 2022.

[36] S. Akinpelu and S. Viriri, "Speech emotion classification using attention based network and regularized feature selection," *Scientific Reports*, vol. 13, no. 1, p. 11 990, 2023.

[37] S. H. Ali Shuvo and R. Khan, "Bangla speech-based emotion detection using a hybrid cnn-transformer approach," in *2023 8th International Conference on Communication, Image and Signal Processing (CCISP)*, 2023, pp. 163–167. DOI: 10.1109/CCISP59915.2023.10355685.

[38] K. Alomar, H. I. Aysel, and X. Cai, "Data augmentation in classification and segmentation: A survey and new strategies," *J. Imaging*, vol. 9, no. 2, p. 46, 2023. DOI: 10.3390/jimaging9020046. [Online]. Available: https://doi.org/10.3390/jimaging9020046.

[39] S. Aziz, N. H. Arif, S. Ahbab, S. Ahmed, T. Ahmed, and M. H. Kabir, "Improved speech emotion recognition in bengali language using deep learning," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2023, pp. 1–6.

[40] M. M. Billah, M. L. Sarker, and M. A. H. Akhand, "Kbes: A dataset for realistic bangla speech emotion recognition with intensity level," *Data in Brief*, vol. 51, p. 109 741, 2023, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2023.109741. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923008107.

[41] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 23, 2023.

[42] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech emotion recognition using convolutional neural networks with attention mechanism," *Electronics*, vol. 12, no. 20, p. 4376, 2023. DOI: 10.3390/electronics12204376. [Online]. Available: https://doi.org/10.3390/electronics12204376.

[43] Y. Zhao and X. Shu, "Speech emotion analysis using convolutional neural network (cnn) and gamma classifier-based error correcting output codes (ecoc)," *Scientific Reports*, vol. 13, no. 1, p. 20 398, 2023.

[44] S. Amiriparian, F. Packa, M. Gerczuk, and B. W. Schuller, "Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets," *Interspeech 2024*, pp. 2635–2639, Sep. 2024. DOI: 10.21437/interspeech.2024-280.

[45] S. Amiriparian, F. Packa, M. Gerczuk, and B. W. Schuller, *Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets*, ResearchGate Preprint, Figure retrieved from https://www.researchgate.net/figure/Outline-of-the-proposed-ExHuBERT-architecture-including-skip-connections-and-zero-linear_fig1_381484831, 2024.

[46] C. Inc., *Detecting issues in an audio dataset with datalab*, https://docs.cleanlab.ai/stable/tutorials/datalab/audio.html, Accessed: 2025-06-18, 2025.

[47] T. Momenpour and A. Abu Mallouh, "Optimizing cnn-based diagnosis of knee osteoarthritis: Enhancing model accuracy with cleanlab relabeling," *Diagnostics*, vol. 15, no. 11, 2025, ISSN: 2075-4418. DOI: 10.3390/diagnostics15111332. [Online]. Available: https://www.mdpi.com/2075-4418/15/11/1332.