Journal of Advanced College of Engineering and Management, Vol. 10, 2025, Advanced College of Engineering and Management

CNN-TRANSFORMER BASED SPEECH EMOTION DETECTION

Rojina Baral [1], Sanjivan Satyal[2], Anisha Pokhrel[3]

[1,2] Department of Electronics and Computer Engineering Pulchowk Engineering Campus, IOE,
[3]Department of Electronics and Computer Engineering Western Region Campus, IOE
Email: rojinabaral28@gmail.com[1], sanjiwan.satyal@pcampus.edu.np[2], anishapokhrel01@gmail.com[3]

ABSTRACT

In this study, a parallel network technique trained on the Ryerson Audio-Visual Dataset of Speech and Song (RAVDESS) was used to perform an autonomous speech emotion recognition (SER) challenge to categorize four distinct emotions. To capture both spatial and temporal data, the architecture comprised attention-based networks with CNN-based networks that ran in tandem. Additive White Gaussian Noise (AWGN) was used as augmentation techniques for multiple folds to improve the model's generalization. The model's input was MFCC, which was created from the raw audio data. The MFCC were represented as images, with the height and breadth corresponding to the time and frequency dimensions of the MFCC, in order to take use of the proven effectiveness of CNNs in image classification. Transformer Encoder layer, an attention-based model, was used to capture temporal characteristics. The projects' findings demonstrated that the Parallel CNN-Transformer network's accuracy as 88.16% for 1-fold augmentation, 92.11% for 2-fold augmentation and 86.84% of accuracy for 3-fold augmentation.

Keywords: Speech Emotion Recognition, Convolution Neural Network, Mel Frequency Cepstral Coefficient, Additive White Gaussian Noise

1. Introduction

Emotion recognition is a broad field of study, one could say. In human communication, the way emotions are expressed is crucial when it comes to communicating the information that has to get over to the other person. Emotions in humans can be expressed in a wide variety of ways. It may involve body language, facial expressions, eye contact, laughter, and tone of voice [1]. As a whole we can say that it can be accomplished through nonverbal and verbal communication. Speech comes under the verbal communication. Although the human brain can recognize emotions from other people's speech, automatic emotion detection is a challenging task. But by mimicking human intelligence in robots to comprehend and behave like humans, the phrase ”artificial intelligence,” or ”AI,” makes that difficult task much easier. Speech recognition technologies are transforming and greatly impacting the current state of human-computer interaction. [2]. Due to differences in speech patterns, accents, and background noise, emotion identification in speech is complex. Consequently, the creation of an algorithm capable of efficiently and automatically analyzing speech emotions is required.

The system's capabilities should include processing audio input, extracting pertinent information including pitch, energy, and spectral qualities, and accurately classifying the speaker's emotional state. Robust and reliable emotion detection in real-world applications requires overcoming significant challenges related to data scarcity, variability, and computational efficiency.

Convolutional neural networks (CNNs) have shown to be an essential technique for extracting local spatial properties from audio. Conversely, CNN is a typical feed-forward deep network. The algorithm's accuracy is limited because it cannot recognize contextual timing information due to its

monotonously connected network structure and one-way information flow. The Recurrent Neural Network (RNN) operates based on the idea of loop formation, whereby the output of one layer is saved and fed back to the subsequent layer. The nature of RNN makes it possible to establish the global dependency of sequence data. However, as the length of the data sequence increases, the network is likely to have gradient vanishing issues. Long-term memorization is introduced via the long Nevertheless, compared to LSTM, the Transformers that employ the self-attention mechanism are better equipped to capture long-term dependencies.

The paper introduced a novel approach to speech emotion recognition (SER) that combined a Convolutional Neural Network (CNN) with Transformer architecture, as well as various data preprocessing techniques and data augmentation technique was examined, specifically tailored to improve the classification of emotional states from voice data. Our solution used both spatial and temporal information to solve inadequacies in classic SER models, which frequently struggled with contextual timing and long-term interdependence. We achieved considerable gains in accuracy and robustness using our hybrid design, suggesting a significant step forward over prior methodologies. Our results showed that this revolutionary architecture surpassed previous models, obtaining an accuracy of 92.11% with 2-fold augmentation on the RAVDESS dataset and thereby establishing a new standard in the field of automatic speech emotion recognition.

2. Related Work

Previously, identifying emotions in speech was done using a deep learning architecture. They used the IEMOCAP datasets to train their model. It used spectrograms and a combination of convolutional and recurrent neural networks to detect both short-term and long-term properties in speech input. The paper [4] examined data augmentation techniques such as vocal track length perturbation, layer-wise optimizer adjustment, and batch normalization of recurrent layers, resulting in 64.5% weighted accuracy and 61.7% un-weighted accuracy on four emotions (happy, sad, angry, and neutral). The model's combination of convolutional and recurrent neural networks enabled it to effectively capture subtle patterns in the speech stream, such as pitch, tone, and rhythm, all of which are important for emotion recognition. Furthermore, the use of batch normalization in recurrent layers helped to alleviate vanishing gradient issues, stabilize training, and improve convergence.

Another paper [5] produced a novel deep dual recurrent encoder model for speech emotion recognition. The model used both audio and text data to improve emotion detection. Unlike typical models, which solely focus on audio features, their model employed dual RNNs to handle both audio signals and text sequences, resulting in a more extensive speech analysis. The model took advantage of the complementary nature of audio and written cues, allowing for a more nuanced comprehension of emotional context. By combining the two modalities, it was able to record small differences in speech patterns and word selections, improving the accuracy of emotion classification. Experiments on the IEMOCAP dataset indicated that the model exceeded past state-of-the-art techniques, reaching classification accuracy of 68.8% to 71.8% for the emotions angry, pleased, sad, and neutral.

The paper [6] proposed a model using the CNN (Convolutional Neural Network) along with LSTM (Long Short-Term Memory) to increase the accuracy of the current available models. The model used was a Time Distributed Convolutional Neural Network. The main concept behind the related network was to build a method across the log-mel spectrogram. Every window in that mechanism as defined had access to a convolutional neural network composed of four local feature learning blocks, and whatever yield from those convolutional networks was then catered in a repeated neural network made up of two LSTM (Long Short Term Memory) cells to learn the long-term contextual dependencies. Finally, emotions were predicted from recorded speech using a complete linked layer

and Softmax activation. Their proposed model outperformed previous state-of-the-art methods by 72% to 75% when applied to the RAVDESS dataset.

This article [1] discusses research on speech emotion recognition using deep neural networks like CNN. For the study, the researcher employed the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus, which included four emotions: anger, happiness, sorrow, and neutral. Mel spectral coefficients, as well as other spectrum and intensity factors, were used for recognition. The data augmentation was performed by altering the voice and adding white noise.

This paper [7] used parallel-based networks with CNN and attention-based models to categorize eight emotions from the Ryeson Audio-Visual Dataset of Speech and Song (RAVDESS). To improve generalization, raw audio data was transformed into Mel-Spectrograms and augmented with several approaches such as Additive White Gaussian Noise (AWGN), SpecAugment, Room Impulse Response (RIR), and Tanh Distortion. CNNs extract spatial characteristics from spectrograms, while Transformer and BLSTM-Attention modules capture temporal features. The suggested CNN-Transformer and CNN-BLSTM-Attention architectures obtained high accuracy of 89.33% and 85.67%, respectively, on a 10% hold-out test set. They outperformed solo and hybrid models while using fewer training parameters. This demonstrates the effectiveness and efficiency of parallel CNN-Attention networks in SER tasks.

Another paper [2] recently suggested a model for automatic speech emotion recognition utilizing a hybrid deep neural network, CNN-BiLSTM. They mostly employed prosodic and spectral features in their work. These acoustic properties, which can distinguish between low-level and high-level features in the ASER, were used to classify emotions. In that scenario, the CNN with convolutional layers, a single batch normalization (BN) layer, a pooling layer, and a fully connected layer was designed for feature extraction, whereas the Bi-LSTM network was designed specifically for long-term sequence dependencies. They compared their results in terms of accuracy and speed to prior work on the RAVDESS and IEMOCAP datasets. The CNN-BiLSTM model revealed its capacity to accurately capture both spatial and temporal relationships in speech data, making it ideal for emotion recognition tasks. The model provided a comprehensive analysis of prosodic and spectral aspects by integrating CNNs for feature extraction and Bi-LSTMs for sequential information processing. The comparison results revealed that their hybrid approach beat traditional models not just in terms of accuracy but also in computational efficiency, demonstrating its durability across both the RAVDESS and IEMOCAP datasets.

Other two researchers conducted research on Persian voice recognition. The authors presented two models, one based on spectrograms and the other on audio, fine-tuned with the shEMO dataset. These models greatly improved the accuracy of prior systems, boosting it from around 65 to 80 percent on the specified dataset. The same models were then fine-tuned twice to see how multilingualism affected the fine-tuning process. They were fine-tuned first using the English IEMOCAP dataset, followed by the Persian shEMO dataset. The study demonstrated the effectiveness of using multilingual datasets to improve emotion recognition in non-English languages. By fine-tuning the models first on English data, which has more diverse emotional expressions, and then refining them with Persian data, the system was able to better capture the nuances of Persian speech. This cross-linguistic fine-tuning strategy not only enhanced model accuracy, but also revealed the feasibility of utilizing multilingual training strategies to improve performance in under-resourced languages such as Persian. This results in an enhanced accuracy of 82% for the Persian emotion identification system [8].

The study [9] developed a system that uses convolutional neural networks and transformers to model speech sequences and use attention mechanism to increase features in time, space, and channel, based on which their model was designed. A set of preprocessing processes were carried out on the input

speech stream. Specifically, varied durations of speech sequences were uniformly processed into 1.8s, with longer sequences split into sub segments and shorter sequences processed via loop filling, before MFCC properties of speech were extracted as input to the model. The local features of the speech were first retrieved by a CNN block, which used irregular-sized time-frequency domain convolution to obtain the speech's time and frequency domain features. The features were then improved with a T-Sa attention mechanism block, which includes a bilstm attention module that models the features in temporal order, followed by a spatial-channel attention mechanism that focuses on spatial and channel information. Finally, the global and local information of speech were learned interactively, allowing the model to learn at various scales. Evaluations of the IEMOCAP and Emo-DB datasets revealed significant performance gains over existing approaches.

The methodology given in the study [10] attempts to combine Deep and Machine learning methodologies. The proposed SER system is separated into two main parts: deep learning, which extracts features from the Wav2vec2 and HuBERT models, and machine learning, which uses the linear SVM classifier to classify emotions. The study examined the impact of embedded features in Wav2vec2 and HuBERT models on SER. Two types each module were tested: Wav2vec2 basic, Wav2vec2 big, HuBERT large, and HuBERT X-large. In addition, a linear Support Vector Machine (SVM) as a downstream model to recognize emotions was adopted. The proposed approach relying on the combination of HuBERT X-large features with the SVM model led to the highest recognition rate of 82.6% on the RAVDESS database.

In the paper [11] the preprocessing signal for speech emotion recognition was introduced. The discrimination between speech and music files was performed depend on a comparative between more than one statistical indicator such as mean, standard deviation, energy and silence interval. Preprocessing, which includes silence removal, pre-emphasis, normalizing, and windowing, is crucial for obtaining a pure signal for subsequent feature extraction. This work uses preprocessing to create a clean signal for feature extraction. After the features of this signal were extracted then it was used to distinguish the emotion.

2.1 Comparison of Related Work

Table 1: Comparison of previous work based on accuracy.

Paper	Datasets	Accuracy
CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation [4].	IEMOCAP	WA=64.5% UA=61.7%
Multimodal Speech Emotion Recognition Using Audio and Text [5].	IEMOCAP	71.8%
Speech Emotion Detection Using State of the Art CNN and LSTM.[6]	RAVDESS	75%
Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation.[7]	RAVDESS	89.33%
Unveiling embedded features in Wav2vec2 and HuBERT models for Speech Emotion Recognition. [10]	RAVDESS	82.6% while using HuBERT X-large, 52.4% while using Wav2Vec2 large.
Automatic Speech Emotion Recognition Using Hybrid Deep Learning Techniques.[2]	RAVDESS, IEMOCAP	82.07% on IEMOCAP, 80.25% on RAVDES

3. Material and Methods

Fig 3.1 illustrates the method implemented for speech emotion recognition. The process started with data collection, data-preprocessing to the feature extraction. The time-frequency acoustic features i.e MFCC as the feature was extracted. The equivalent cnn and transformer receives each time-frequency characteristics determined by the audio preprocessing task each determining the spectral and temporal dependencies respectively. These output was then passed to linear followed by loss entropy block and softmax activation function which classify the speech based on the emotion.

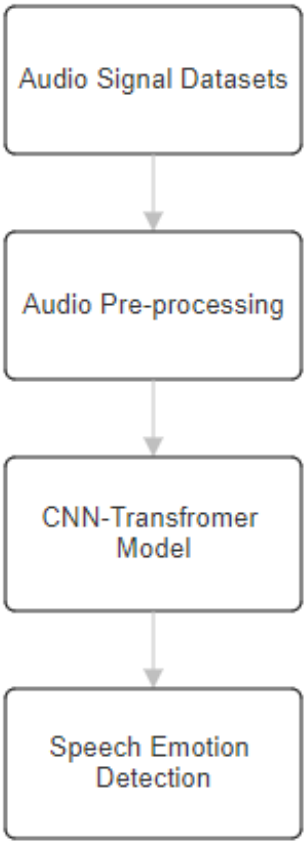

Fig 3.1: Flow diagram for recognition of speech emotion.

3.1 Audio Datasets

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was employed to train and evaluate the effectiveness of the implemented model. The RAVDESS dataset is a comprehensive resource that includes audio-visual recordings, audio-only recordings, and video-only recordings, making it versatile for a wide range of emotion recognition tasks. This dataset was meticulously curated with recordings from 24 professional actors, consisting of 12 males and 12 females, ensuring a balanced representation of gender. The actors perform a diverse set of emotions, providing a rich and varied dataset ideal for developing robust emotion recognition models. RAVDESS includes recordings of eight distinct emotional expressions: calm, neutral, happy, sad, angry, fearful, surprise, and disgust. Each of these emotions is expressed by the actors at different levels of intensity, adding depth and complexity to the dataset. However, for the purposes of this project, a subset of these emotions—namely, happy, sad, angry, fearful, and surprised was focused.

This selection was made to streamline the classification process and target emotions that are more commonly analyzed in emotion recognition studies.

3.2 Audio Pre-processing

Before proceeding with the analysis, several essential pre-processing steps were taken to prepare the RAVDESS dataset for effective training and evaluation. Among 1440 audio data of RAVDESS datasets, the 950 data were taken of the classes, happy, sad, angry, fearful and surprised. To simplify the dataset, the 'happy' and 'surprised' emotion classes were merged into a single 'happy' category.
Due to the merging of two classes of data into single it introduced an imbalance in the dataset, with some classes being underrepresented compared to others. To address this imbalance, oversampling technique was applied, ensuring that each class was adequately represented in the dataset. This oversampling brought the number of 950 data to 1520. This step was crucial for preventing the model from being biased toward the more dominant classes, thus improving its ability to recognize emotions across all categories.
Next, the dataset was divided into three distinct subsets: training, validation, and testing, following an 80:10:10 ratio. This split allowed for a comprehensive evaluation of the model's performance at different stages of development. The training set was utilized for model learning, the validation set for adjusting and preventing overfitting, and the testing set for evaluating overall performance.

After splitting the data, each splitted data was then converted to MFCC. The extraction of speech features is a significant and difficult task in speech emotion detection. The accuracy of the final emotion recognition algorithm and the efficiency of future model training are directly impacted by the feature extraction process. Speech characteristics are classifiable. Characteristics based on deep learning and acoustics, where the former can be roughly divided into phonetic features, spectral-based correlation features, and rhythmic features. The aspects of the signal in the frequency domain, are reflected by the spectral-based correlation features. Linear spectrum and inverse spectrum, linear prediction coefficients (LPC), log frequency power coefficients (LFPC), and so on are based on the spectral correlation features; the inverse spectrum comprises Mel-Frequency Cepstrum Coefficients (MFCC), linear prediction cepstrum coefficients (LPCC), and so on. MFCC is considered a low-level feature in speech recognition, based on human knowledge and is widely employed. In this model too, the MFCC was extracted. To extract this MFCC the Python library librosa was used with sample rate of 48000, an FFT window length of 1024, hamming window of length 512, and 128 mel bins as its parameters.

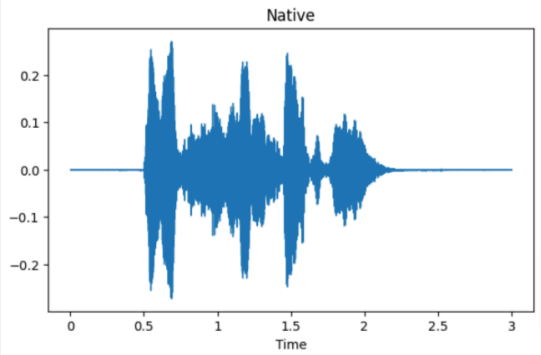

Fig 3.2: Original Speech Waveform.

The robustness of the model was enhanced further by performing data augmentation on each of these subsets. Specifically, Additive White Gaussian Noise was employed as the augmentation technique

for multifold of N. AWGN uses a Gaussian noise vector from a normal distribution with a zero mean time average, applied equally over the frequency distribution. To implement noise addition, two signals were added and the resultant signal is scaled to adjust the signal-to-noise ratio (SNR). The SNR was randomized and consistently picked in the decibel scale, which suits a logarithmic scale rather than linear, comparable to human hearing.
In this study the size of N made equals to 1, 2 and 3 increasing the datasets by two times, three times and four times of original datasets respectively. The AWGN augmentation utilized a minimum and maximum sound-to-noise ratio of 15 and 30, respectively. By introducing noise into the audio samples, this augmentation method helped the model learn to generalize better to new, unseen data, thereby reducing the risk of overfitting and improving its overall performance across diverse audio conditions.

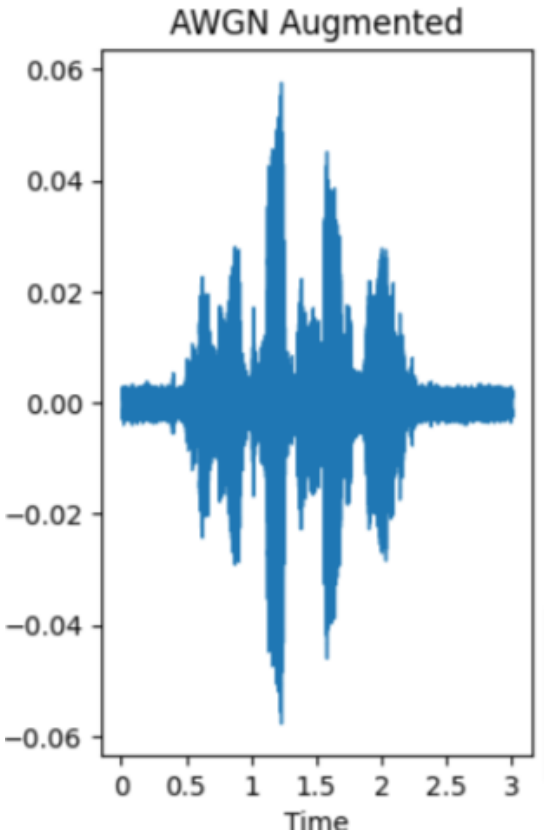

Fig 3.3: Augmented Speech Waveform

3.3 CNN-Transformer Model

In the Parallel CNN- Transformer model, two 3-layered CNN blocks were implemented parallel with a Transformer Encoder layer. The input feature takes the shape of (N, 1, 40, 282).The architecture's CNN-Transformer module was made to process the input of mel-frequency cepstral coefficients (MFCC), collecting intricate details that represent the underlying frequency speech patterns.

The CNN structure consisted of two parallel convolutional blocks, each with a set of three 2D convolutional layers. These layers add convolutional filters to the MFCC input, allowing the network to recognize local properties. The first layer used a 1 × 3 × 3 filter to produce 16 channels. The

resultant feature map was batch normalized before being activated with ReLU. After activation, a max-pooling layer was used, followed by a drop-out layer with a probability of 0.3 for all following layers. The second layer widens the output feature map to 32 channels and max-pool kernel size was increased. Finally, the third convolutional block expanded the output feature map to a depth of 64 channels. From the MFCC input, the two convolutional blocks independently learn different sets of features. After that, the outputs of those blocks were flattened, which reduced the multidimensional data to a one-dimensional vector and made it simpler to integrate with the model's later layers.

To comprehend the timing and sequence of the audio signals, the Transformer Encoder was utilized in the architecture to capture the temporal dependencies included in the speech data. After passing via a max-pooling layer with kernel size of 1*4 and stride of 1*4, the MFCC features provide the input for the Transformer Encoder. In order to make the input data manageable for the Transformer to handle, max-pooling helps to downsample it while keeping the most important characteristics. Using the input sequence with 60 features per element, multi-head self-attention layer with 6 heads, drop-out with the probability of 0.4, the network took into consideration multiple previous time steps when predicting the next.

The temporal relationships learned by the Transformer Encoder and the features recovered by the CNN's parallel convolutional layers were combined. After that, a linear layer received the combined output, which then included the temporal dynamics from the Transformer as well as the detailed frequency-based features from the CNN. The combined features were converted by this layer into a set of logits, which were then utilized to forecast the speech signal's ultimate emotional state. By integrating both the temporal and frequency domain information, the model can made more accurate and context-aware predictions about the emotional content of the audio.

The Adam as optimization algorithm was used, the learning rate was set to 0.0001, weight decay to 1e-5 and batch size to 32.

4. Results and Discussion

Important insights into the Speech Emotion Recognition (SER) CNN-Transformer model's performance can be gained from its study. Confusion matrices, loss curves, accuracy, precision, recall and f1-score of the model offer a thorough understanding of its strengths and areas for improvement. As discussed in audio-preprocessing techniques the AWGN was applied for multiple folds i.e 1-fold, 2-folds and 3-folds.

Table 2: Accuracy, Precision, Recall and F1-score for N-fold augmentation

N-Fold	Accuracy	Precision	Recall	F1-Score
1-fold	88.16%	0.89	0.88	0.88
2-fold	92.11%	0.93	0.92	0.92
3-fold	86.84%	0.89	0.87	0.88

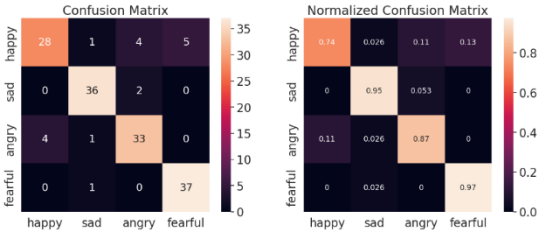

Fig 4: Confusion Matrix for 1-Fold Augmentation

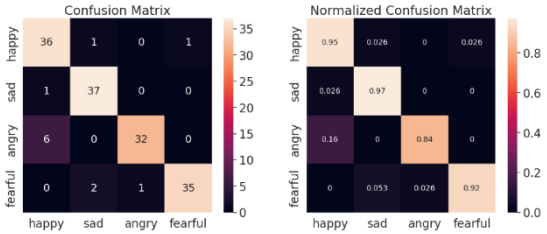

Fig 5: Confusion Matrix for 2-fold Augmentation

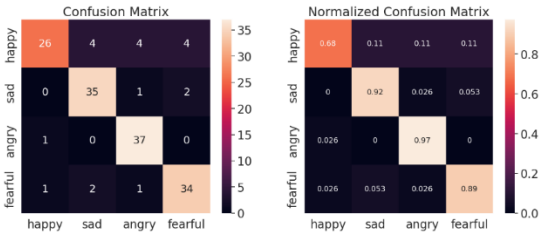

Fig 6: Confusion Matrix for 3-fold Augmentation

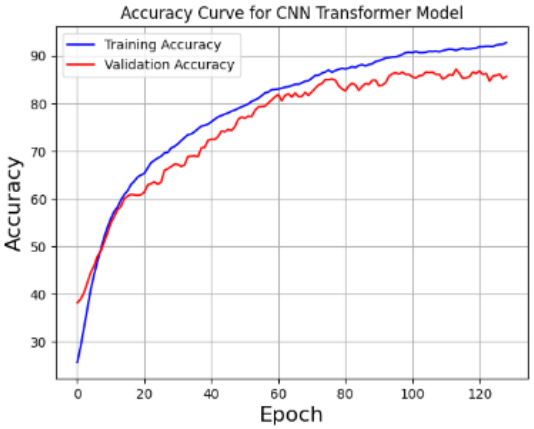

Fig 7: Accuracy Curve for 1-fold Augmentation

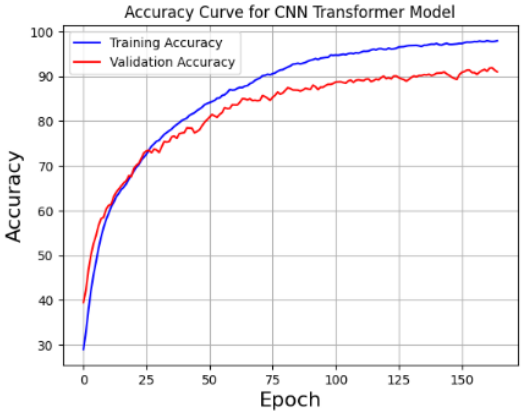

Fig 8: Accuracy Curve for 2-fold Augmentation

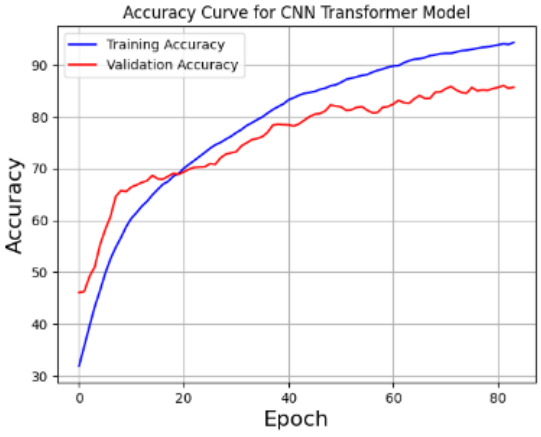

Fig 9: Accuracy Curve for 3-fold Augmentation

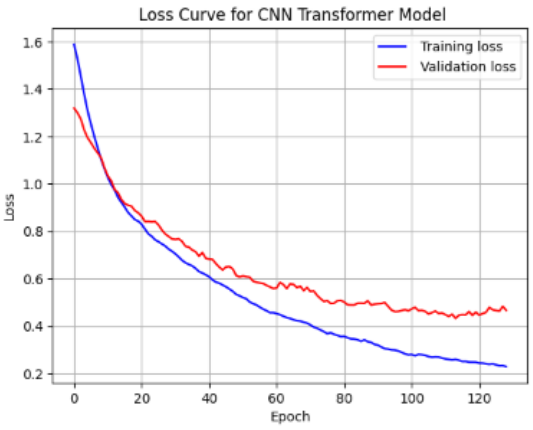

Fig 10: Loss Curve for 1-fold Augmentation

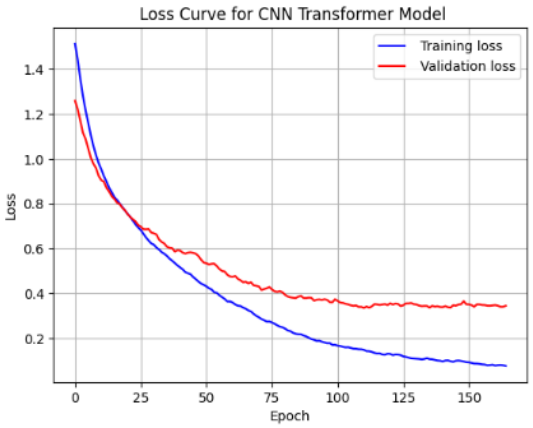

Fig 11: Loss Curve for 2-fold Augmentation

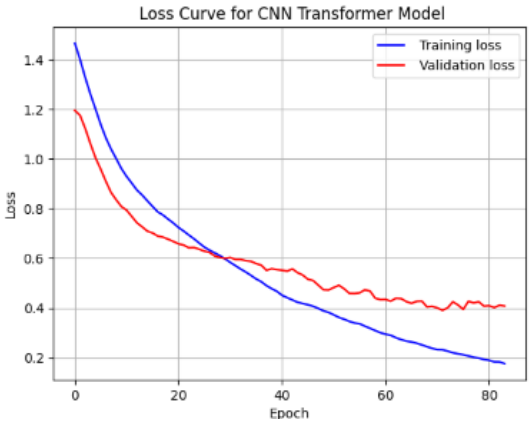

Fig 12: Loss Curve for 3-fold Augmentation

5. Conclusions and Future work

In this study, a complex hybrid CNN-Transformer model with multi-fold augmentation was built for speech emotion recognition, with a focus on classifying four important emotions: happy, sad, angry, and fearful. Among the evaluated N-fold augmentation, 2-fold augmentation outperformed the others obtaining the better accuracy, precision, recall and F1-score than other two, 1-fold and 3-fold augmentation.

The project findings show that the CNN-Transformer technique is promising for voice emotion identification tasks. However, future work might concentrate on strengthening the model by including more diverse datasets, boosting real-time performance, and addressing issues such as speaker variability and cross-dataset generalization. This study lays the groundwork for future research into emotion-aware systems that could be used in real-world applications such as emotional intelligence in virtual assistants, mental health monitoring, and human computer interaction systems.

Acknowledgments:

I would like to express my deepest gratitude to individuals and institutions who have contributed in the completion of the research.

First of all I would like to thank the M.Sc. coordinator, Assistant Professor Bibha Sthapit, of M.Sc. in Computer System and Knowledge Engineering, for her invaluable guidance and encouragement throughout the research process. Her insights and support have been crucial in refining our ideas and crafting a compelling report.

I would like to express my sincere gratitude to the Department of Electronics and Computer Engineering, IOE, Pulchowk Engineering Campus for their continuous support on this reasearch.

References

1. Loan Trinh Van, Thuy Dao Thi Le, Thanh Le Xuan, and Eric Castelli. Emotional speech recognition using deep neural networks. Sensors, 22(4), 2022.
2. Bilal Hikmat Rasheed, D. Yuvaraj, Saif Saad Alnuaimi, and S. Shanmuga Priya. Automatic speech emotion recognition using hybrid deep learning techniques. International Journal of Intelligent Systems and Applications in Engineering, 12(15s):87–96, Feb. 2024.Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).

3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.
4. Andrei Petrovskii Laurence Devillers Benoit Schmauch Caroline Etienne, Guillaume Fidanza. Cnn+lstm architecture for speech emotion recognition with data augmentation. Workshop on Speech, Music and Mind 2018, 2018.
5. Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. CoRR, abs/1810.04635, 2018.
6. Dipesh Sachdev Yash Jajoo Anchit Banga, Bhavik Baheti. Speech emotion detection using state of the art cnn and lstm. International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), 2021.
7. John Lorenzo Bautista, Yun Kyung Lee, and Hyun Soon Shin. Speech emotion recognition based on parallel cnn-attention networks with multi-fold data augmentation. Electronics, 11(23), 2022.
8. Minoo Shayaninasab and Bagher Babaali. Persian speech emotion recognition by finetuning transformers. 2024.
9. Xiaoyu Tang, Yixin Lin, Ting Dang, Yuanfang Zhang, and Jintao Cheng. Speech emotion recognition via cnn-transforemr and multidimensional attention mechanism. 2024.
10. Adil CHAKHTOUNA, Sara SEKKATE, and ADIB Abdellah. Unveiling embedded features in wav2vec2 and hubert msodels for speech emotion recognition. Procedia Computer Science, 232:2560–2569, 2024.
11. Bashar M. Nema and Ahmed Amer Abdul-kareem. Preprocessing signal for speech emotion recognition. Al-Mustansiriyah Journal of Science, 2018.