EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems Research Article EAI.EU

Emotional Inference from Speech Signals Informed by Multiple Stream DNNs Based Non-Local Attention Mechanism

Manh-Hung Ha[1,*], Duc-Chinh Nguyen[1], Long Quang Chan[1], Oscal T.C. Chen[1,2]

[1] Faculty of Applied Sciences, International School, Vietnam National University, Hanoi 100000, Vietnam.
[2] Department of Electrical Engineering, National Chung Cheng University, Chiayi, 62102, Taiwan.

Abstract

It is difficult to determine whether a person is depressed due to the symptoms of depression not being apparent. However, the voice can be one of the ways in which we can acknowledge signs of depression. Understanding human emotions in natural language plays a crucial role for intelligent and sophisticated applications. This study proposes deep learning architecture to recognize the emotions of the speaker via audio signals, which can help diagnose patients who are depressed or prone to depression, so that treatment and prevention can be started as soon as possible. Specifically, Mel-frequency cepstral coefficients (MFCC) and Short Time Fourier Transform (STFT) are adopted to extract features from the audio signal. The multiple streams of the proposed DNNs model, including CNN-LSTM based on an attention mechanism, are discussed within this research. Leveraging a pretrained model, the proposed experimental results yield an accuracy rate of 93.2% on the EmoDB dataset. Further optimization remains a potential avenue for future development. It is hoped that this research will contribute to potential application in the fields of medical treatment and personal well-being.

Keywords: Convolution Neural Network, LSTM, Attention mechanism, Emotion, Classification.

Received on 20 02 2024, accepted on 01 07 2024, published on 02 08 2024

doi: 10.4108/eetinis.v11i4.4734

1. Introduction

In human communication, emotions play a pivotal role conveying information and establishing a foundation for understanding the speaker's mood, emotions, and responses. The capability to identify and analyse emotions in natural language has emerged as a significant tool, particularly within the realm of mental health treatment, offering advantages in recognizing the emotional states of patients. This facet aids in cost savings associated with human resources and ensures accuracy unaffected by subjective factors. The discernment and examination of emotional expressions in natural language holds considerable promise for applications in mental health diagnostics, providing a nuanced and objective insight into the emotional well-being of individuals. The intrinsic value of this analytical process lies not only in its potential cost-effectiveness but also in its capacity to yield reliable and unbiased assessments, thus contributing to advancements in the field of healthcare.

This research endeavours to develop a novel methodology for analysing and recognizing emotions based on audio data. The primary aim of the study is to classify and predict emotional states, with a specific focus on detecting negative emotions, particularly depressive states. Depression has emerged as a prevalent psychological issue in modern society, evident not only through psychological expressions but also through physical health symptoms [1][2]. According to statistics, 97% of suicide cases are linked to mental health conditions during the period of illness, with the highest incidence observed in cases of depression. Therefore, the prevention and treatment

*Corresponding author. Email: hunghm@vnu.edu.vn

of depression play a pivotal role in enhancing the quality of life and mental health within the community.

To assess the state of an individual's depression, information can be gathered from various perspectives, including changes in physical health and daily habits. However, analysing information from linguistic expressions and speech patterns is considered a more convenient method, as speech serves as a rich source of signals reflecting mood and emotions. Advanced tools and algorithms in this field have the capability to analyse and predict speech data, contributing to a more detailed and reliable assessment of an individual's emotional state.

In today's digital age, the ability to understand and analyse human emotions through language has become a crucial application, especially across various domains. To meet the increasing demand for this understanding, emotion recognition models based on language input, known as Natural Language Processing (NLP), have become a focal point of research. The integration of deep learning methods and natural language processing techniques is driving significant advancements in the development and application of models designed for emotion recognition in the realm of natural language

Motivation and Objective:

With the advancement of science and technology in today's society, human life is becoming increasingly hectic and complex. The modern living environment, characterized by a fast-paced and expansive lifestyle, coupled with pressures from various aspects, has laid the foundation for the development of a range of health-related issues, particularly those concerning mental health.

Continual pressure from the environment can lead to anxiety, depression, or mood instability, potentially generating negative and prolonged impacts on both mental and physical health. Mental illnesses, especially depression, have become prominent and concerning issues in today's society, with the potential to cause profound effects on daily life and the overall community healthcare system.

In Europe, the consumption of antidepressants is rapidly increasing in developed countries, according to data from the OECD reported on Euro News [3]. During the period from 2000 to 2020, the average consumption of antidepressants per 1000 people per day in some European countries has doubled, tripled, or even quintupled. The most significant growth occurred in countries such as Czechia, Estonia, and Slovakia, where the increase was approximately fivefold, particularly notable in Czechia with a 577% rise. Other countries also experienced high growth rates ranging from 100% to 300%. There are relatively few countries with lower growth rates, and no country in the statistics recorded a decrease in the antidepressant usage rate.

Developed countries in Europe are grappling with various pressures from daily life, including work-related stress and concerns about external issues. Advances in medical science have also facilitated the early detection and treatment of depression, leading to an increasing trend in the use of antidepressants among individuals seeking to mitigate potential negative consequences.

The concept of emotion recognition for detecting depression is implemented through the application of deep learning methods, combined with the use of accompanying patient-worn devices to assess speech characteristics relevant to the depressive state. This technology utilizes speech as a means to predict emotions and is rapidly advancing.

In this research, we aim to apply deep learning to detect emotional features in the speaker for assessing the potential signs of depression or underlying depressive tendencies. Simultaneously, this study focuses on evaluating emotional expressions during speech to identify indications of depression in their verbal communication.

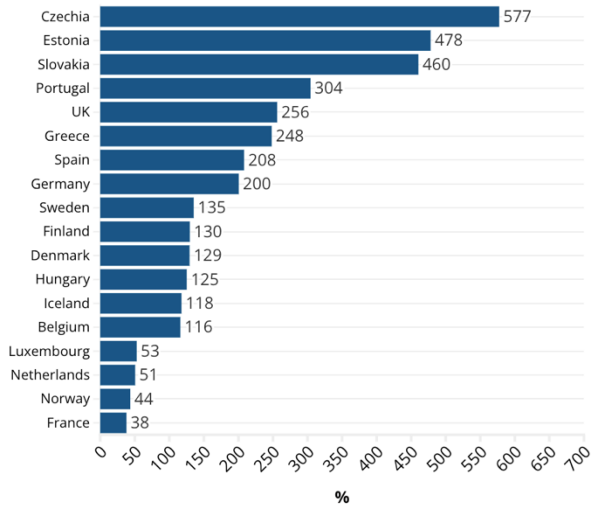

Figure 1: Increase in consumption of antidepressant drugs in the last 20 years (Percentage of change between 2000 and 2020) [3]

In this study, we conducted experiments and developed artificial neural network models to predict and analyse emotions based on labelled audio data. This process not only focused on testing various methods and algorithms but also emphasized the application of processing techniques such as Mel-frequency cepstral coefficients (MFCC) and Short Time Fourier Transform (STFT) for audio data processing and feature extraction. Simultaneously, we explored and implemented neural network architectures like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), along with attention mechanisms to construct models aimed at improving accuracy in emotion recognition and classification. Through these methods and tools, the research aims to contribute to enhancing the performance of the emotion analysis and recognition process from audio data.

2. Related Work

Traditional machine learning methods have been employed for the identification of emotional patterns in speech, with the most widely used approach being the application of classifiers. The fundamental concept involves establishing a classification model based on a large amount of known data, accurately assigning data to the appropriate categories. With a sufficiently large dataset, these models can achieve accuracy when applied to new data [4][5]. Common classification models include Support Vector Machine (SVM), Gaussian Mixture Models (GMM), Hidden Markov Model (HMM), and various other models. Minji Gil and colleagues conducted research on predicting the risk of depression in students using machine learning algorithms such as sparse logistic regression (SLR), support vector machine (SVM), and random forest (RF) [6]. However, these methods often require complex preprocessing and meticulous feature selection, which can increase complexity and processing time.

When employing these methods for identification, the process is typically divided into several fundamental steps: pre-processing of the input signals, feature extraction, and classification. The pre-processing stage may address issues with the input data, such as denoising the signal, segmenting or merging audio segments of varying lengths to a uniform length, or selecting feature parameters through feature selection. After pre-processing, the data is fed into the model to extract features before moving on to the classification step for identification.

The pre-processing steps are crucial for traditional models because audio data can exhibit considerable noise and variability. In reality, each of us speaks with different tones, speech habits, and intonations. Moreover, recorded speech may differ from real-life speech. Therefore, normalization is an essential step to standardize the audio, creating favorable conditions for feature extraction in subsequent steps. Cleaning methods are commonly applied to minimize noise in audio data, using filtering techniques such as high-pass or low-pass filters, referred to as Noise Reduction. Additionally, Silence Removal is an important step to eliminate non-informative silence from the audio signal. Many research groups aiming to improve model accuracy have employed Volume Normalization to adjust the volume of all audio segments to the same level, avoiding unnecessary discrepancies. Trimming and Alignment helps to cut audio segments to the same length and align them in time to ensure consistency. Resampling ensures that all audio segments have the same sampling rate, synchronizing the data. Feature extraction involves various techniques to derive important features from audio data. Numerous methods are employed in audio processing, such as Energy, Zero-Crossing Rate, and MFCC. One notable technique in audio processing is MFCC. Kunxia Wang et al. [7] conducted research on the impact of these methods on emotion recognition tasks. Typically, this process generates 12 coefficients. By incorporating the energy of the audio signal, we can have 13 coefficients. MFCC may become unstable in the presence of noise, however, so they are often normalized to minimize noise. The effectiveness of this process has been studied [8].

Temporal Features include Zero-Crossing Rate (ZCR), which measures the rate at which the audio signal crosses the zero axis, and Energy, which measures the energy of the audio signal in small time frames. Statistical Features such as mean, variance, skewness, and kurtosis provide statistical information about the audio signal that can be leveraged to enhance performance.

Following that, the feature selection process is conducted to reduce the number of features used by identifying the best feature combinations from the initial parameters. This aims to reduce computational complexity, accelerate execution speed, and enhance recognition performance. Feature selection methods include: statistical (such as Chi-Square test, ANOVA, correlation coefficient), filter methods (such as Variance Threshold, Mutual Information, ReliefF), wrapper methods (such as Forward Selection, Backward Elimination, Recursive Feature Elimination), and embedded methods (such as L1 Regularization, decision trees, Elastic Net). Dimensionality reduction techniques like PCA and LDA are also used to optimize data. However, feature selection can lead to the loss of important information if not performed carefully.

With the continuous expansion of deep learning applications, speech recognition has achieved promising results. Numerous studies have integrated deep learning methods for emotion recognition in speech. A neural network combined with CNN and LSTM achieved a resolution rate of 68% for DAIC-WOZ [9]. In 2017, A. M. Badshah and colleagues utilized a three-layer CNN architecture with a fully connected layer for emotion recognition in speech on the EmoDB database, achieving a recognition rate of 84.3% [10]. In the same year, Haytham and the research group [11] employed a CNN-RNN structure on the IEMOCAP database and achieved a recognition rate of 64.78%. While deep learning methods have improved recognition performance, they often require significant computational resources and long training times.
In 2018, S. Tripathi [12] applied a three-layer LSTM architecture to classify emotions on the IEMOCAP dataset, achieving a recognition rate of 71.04%. Subsequently, in 2019, J. Zhao and colleagues [13] used a CNN-LSTM architecture. This marked the first time CNN was used to extract features from speech signals, followed by the application of LSTM to analyze the temporal relationships of these features. They achieved a recognition rate of 92.9% in EmoDB and IEMOCAP classification. Details regarding the datasets will be discussed and explained in the following chapter. Table 1 below provides a detailed comparison.

Table 1. Comparison of related documents

Ref.	# emotion	Data	model	Acc
[9]	2	DAIC-WOZ	CNN-LSTM	68%
[10]	7*	EmoDB	CNN	84.3%
[11]	5**	IEMOCAP	CNN-RNN	64.78%
[12]	4***	IEMOCAP	3 layers LSTM	71.04%
[13]	7*/6****	EmoDB/IEMOCAP	CNN-LSTM	92.9%

anger, boredom, disgust, fear, joy, sadness, neutral
*** angry, happy, neutral, sad, silence*
****anger, happiness, sadness, neutral*
*****anger, excited, frustrated, happiness, neutral, sadness*

Emotional audio data is inherently complex due to diversity in emotional expression across languages, regions, genders, and ages. These variations, along with different data collection methods such as real-life scenarios, interviews, films or television shows, and professional acting, add to the complexity. Throughout our research, we have identified several commonly used datasets, including DAIC-WOZ, EmoDB, and IEMOCAP.

The DAIC-WOZ database contains clinical interviews designed to assist in diagnosing psychological conditions such as anxiety, depression, and post-traumatic stress disorder. The data includes audio and video recordings, along with responses from extensive questionnaires. Interviews are conducted by a remotely controlled virtual interviewer, facilitating natural and authentic data collection. The data has been transcribed and annotated for various speech and non-speech characteristics, with durations ranging from 7 to 33 minutes (average of 16 minutes). However, the naturalness and diversity of the data can introduce noise, complicating processing and analysis.

EmoDB takes a different approach by collecting emotions from actors in a controlled environment. This German database contains approximately 500 utterances from ten different actors, portraying six basic emotions and a neutral state. The data was recorded in an anechoic chamber at the Technical University of Berlin. Although this data lacks the naturalness of DAIC-WOZ, it has the advantage of minimizing noise and easily controlling variables.

The IEMOCAP database is a multimodal, multi-speaker acted database, collected at the SAIL lab at USC. It contains approximately 12 hours of audiovisual data, including video, speech, facial motion capture, and text transcripts. Dyadic sessions, where actors perform improvised or scripted scenarios specifically selected to elicit emotional expressions, are included. The IEMOCAP data is annotated by multiple individuals with categorical labels. With its multimodal data and large size, IEMOCAP offers a rich resource for studying and developing emotion recognition models, although its size and diversity can complicate processing and analysis.

These databases, each with unique characteristics and diverse collection methods, provide an essential foundation for researching and developing speech emotion recognition models.

3. Proposed DNN for Emotional Classification

As shown in Fig. 2, the proposed DNN consisting of two 2D CNNs, 1D CNN + LSTM and Non-Local Attention generation layers. The utilization of these three processing streams aims to optimize the extraction of features from audio data, integrating information from both raw signals and spectral representations to enhance the accuracy of emotion recognition. First of all, audio signal is decomposed into multiple segments each of which has an interval of few seconds that are sufficient to contain an emotion. Owing to the fixed dimension of the input neural layer in the proposed DNN, the frame size in a audio segment may need to be converted. One using 1D data and the other using 2D data. Both models were provided with 8-second audio inputs sampled at 16 kHz. The 1D processing stream captures information directly from the raw audio signal, enabling the model to learn frequency and temporal characteristics in detail. This stream leverages raw data without complex preprocessing steps, minimizing information loss and maintaining the integrity of the original signal. Furthermore, the 1D CNN can detect short-term frequency features in the audio data, while the LSTM aids in retaining and processing long-term dependencies, facilitating emotion recognition.

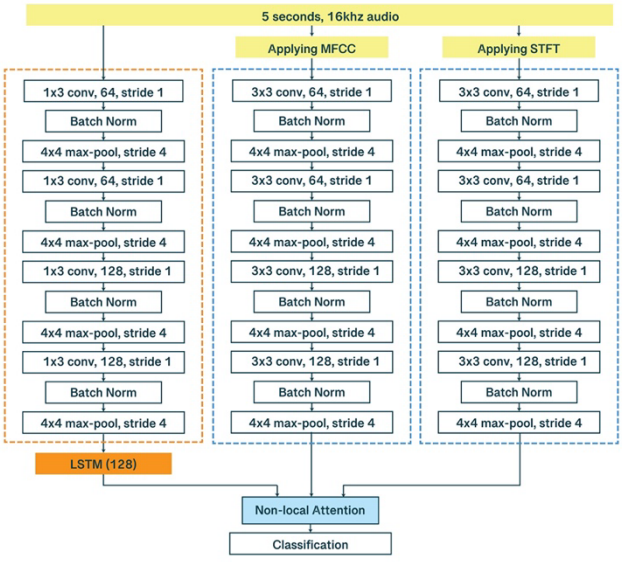

Figure 2. Overall proposed DNN architecture

In contrast, for the 2D data model, instead of using direct audio signals, the authors applied the Mel-frequency cepstral coefficients (MFCC) encoding method and applying the STFT processing technique for second 2D CNN. MFCC is a widely-used technique in speech recognition that helps the model focus on the most crucial features of the audio signal and filter out unnecessary noise. STFT, by applying Fourier transforms to short time frames of the signal, provides information on the frequency variations over time, allowing the model to recognize short-term changes and dynamic features in the audio.By combining both MFCC and STFT methods, the model can leverage the advantages of both techniques, including extracting significant frequency features and tracking frequency changes over time. This enhances the model's emotion recognition capability by offering a more comprehensive view of the audio signal characteristics.

Figure 2 illustrates the overall architecture of the model when using this input data. The input data consists of 5-second audio samples processed concurrently in three streams as described.

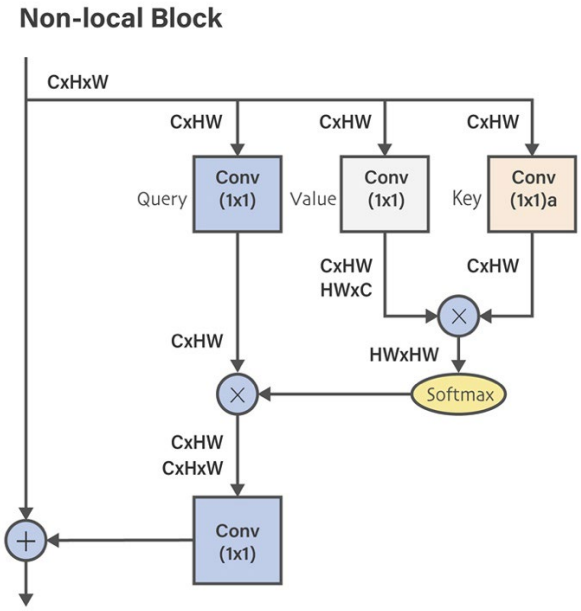

Figure 3: Modified Non-Local Attention Mechanism block [14]

In the 1D CNN stream, a sequence of 4 convolutional blocks is connected consecutively, each block comprising a 1D convolutional layer combined with batch normalization and pooling. The input data for this stream are 8-second audio segments sampled at 16 kHz, converted into one-dimensional vectors with a length of 80,000 audio points. These vectors contain complete frequency and temporal information of the raw audio signal, allowing the model to learn complex features without extensive preprocessing. As mentioned, processing the raw audio data directly enables capturing short-term frequency characteristics, and we use an LSTM layer at the end to retain and process long-term dependencies, facilitating emotion recognition.

The two 2D processing streams have identical structures except for the different audio preprocessing techniques. As discussed, MFCC and STFT were used to preprocess the audio. The audio signal, After applying MFCC, is transformed into a set of Mel-frequency cepstral coefficients, representing important frequency features of the audio signal in a 2D matrix. This matrix is then used as input for the first 2D CNN. The transformed data has dimensions of 128x157, with 128 MFCC coefficients representing frequency features and 157 time frames representing temporal features of the audio segment.

The audio signal, after applying STFT, is converted into a spectrogram, displaying frequency content over time. This spectrogram is a 2D matrix, showing the frequency variations over time, and is used as input for the second 2D CNN. The output after STFT processing also has dimensions of 128x157, with 128 representing frequency bands and 157 representing time frames.

Regarding structure, the processing blocks of the two 2D streams are similar, each comprising 4 convolutional blocks, each starting with a 2D convolutional layer combined with batch normalization and a pooling layer. The number of channels as the data flows through each layer gradually increases from 1 to 64, ending with 128 layers.

With numerous modifications from the input data format to the architecture of the model, the accuracy did not significantly improve, and in some architectures, the accuracy even decreased compared to the original one. Instead of adjusting or removing components of the original model, we considered enhancing certain elements to improve the model's accuracy. One of the techniques we explored for this purpose is the Non-Local attention mechanism as shown in Figure 3.

The non-local Attention mechanism is integrated to enhance the model's ability to recognize and analyze critical features of the audio signal. Unlike traditional convolutional layers, which focus only on local regions of the data, the non-local block can consider and integrate information from the entire temporal and spatial sequence of the input data. This enables the model to recognize long-term relationships and broader context in the data, which conventional CNN layers might miss. The use of non-local blocks aims to integrate and optimize information from the three processing streams (1D CNN and 2D CNN). This helps the model not only focus on short-term and long-term frequency features but also enhances the ability to distinguish important features from different representations of the audio data.

The desired outcome of using non-local blocks is to improve emotion recognition accuracy by emphasizing important data regions and minimizing the impact of noise. The non-local block operates by generating an attention weight matrix from the input matrices of the three processing streams. This matrix is then multiplied with the original data matrix to create new data matrices, highlighting important regions.

The output matrix from the input data goes through an additional layer of either a convolutional or fully connected layer to create an attention weight matrix. This matrix is then multiplied with the original data matrix to generate a set of new data matrices, where more important regions are highlighted through the applied attention mechanism.

3.1 Dataset and Pre-processing

The data used for this study is selected from the EmoDB [15] database of Germany's Berlin Institute of Technology. Although the data volume in EmoDB is smaller compared to some other databases, it is sufficient to achieve the research objectives. Each data segment is a short sentence, which is convenient for emotion determination. Compared to other databases, EmoDB may be considered more suitable for this research. The main reason is the moderate length of the audio, where each sentence represents a single emotion, eliminating the need for cutting or creating excessively long data segments. Applying certain methods to other databases may lead to inaccurate classification results. This database is vital for emotion recognition from speech, containing 535 audio samples recorded by 10 participants (5 males and 5 females) shown in Table 2. It includes 7 different emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral. The diversity in gender and emotion types makes Berlin EmoDB increasingly popular, providing a foundation for in-depth analysis and research on emotions expressed through speech.

Furthermore, a strategy is employed to prepare the audio data in EmoDB, following the method cited from [12], which involves standardizing the length of the audio to 8 seconds. This is achieved by adjusting the length of audio below 8 seconds to 8 seconds through zero-padding and handling the remainder. In EmoDB, the length of the audio typically ranges from 1 to 8 seconds, with the longest breakpoint being 8.9 seconds. This technique also allows controlling the audio length within 8 seconds by removing the head and tail portions.

It is noted that using zero-padding can lead to filling too many 0 values, resulting in the loss of useful information and blurring the data. Therefore, the mirror padding method has been applied to prepare the audio data to 8 seconds, as illustrated in Figure 4 below.

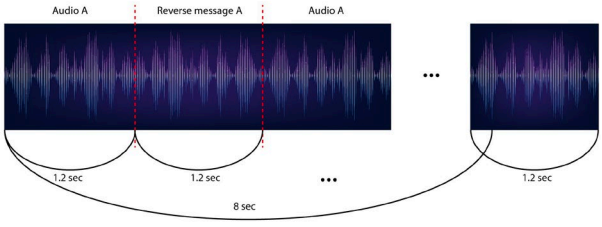

Figure 4. Example of the mirror padding scheme for audio segment

To enhance the accuracy of recognition, we implemented cross-validation by removing one individual from the dataset. This allowed us to estimate the model's performance by testing with data from this person and the remaining data from others. Given the uneven distribution of data for each emotion, we standardized the length of each audio segment and ensured a consistent number of training samples. The table below illustrates the data segmentation methods and details of the data volume.

For a specific approach, we took the resting data of the first person as an example. In this case, the test data was identified from this person, and the remaining training and testing data were divided for days 2 to 10. We extracted individual data, shuffled, and standardized data for each emotion. Initially, we randomly selected 15 samples for verification and used the remaining part to create the training set.

A challenge arose when the number of training samples for some emotion types was too low. This was due to using the same data to create additional training data, leading to a shortage of training data and a decrease in the recognition rate. To address this issue, we applied data augmentation. For example, to augment the data, we removed some samples of the angry emotion from the first person. At this point, out of the required 500 samples, data would be generated from the original data. After concatenating to reach a length of 8 seconds using the mirror padding method as previously described, we generated a total of 97 data. Next, each of the 97 data was shifted to the left by 0.1 seconds, creating 97 new data, and this process was repeated until we collected 500 samples. This helped ensure the diversity of the training data to improve emotion recognition performance.

Table 2. EmoDB dataset feature

	samples	rate	max length	participant	emotions
EmoDB	535	16khz	8.7s	10 people	7

The input data size under consideration is 8 seconds, with a sampling frequency of 16kHz. Each one-dimensional audio data can be represented by a data array of 128,000 points. For two-dimensional data, we used the librosa library to create an MFCC parameter matrix for audio data. Simultaneously, 8 seconds of audio data could produce an MFCC parameter matrix of size 128 * 251. The sampling parameters were set to 2048 points for an audio frame and 512 points for a hop size.

During the testing process, using the first individual as the test data for a small-scale evaluation could lead to inaccuracies in assessing the model's performance. The results would either be entirely correct (100%) or incorrect (0%), resulting in an imprecise evaluation. Therefore, in subsequent experiments, we randomly selected a subset of individuals from the total of 10 to represent the test data. This

approach aimed to create a more diverse testing environment, providing a better reflection of the model's performance.

In the original document, the authors only mentioned the output of each layer in the form of computational formulas. When we conducted practical research, we computed this information into specific numerical values, which may lead to differences from the original structure in the reference literature. Using MFCC with 128 parameter groups, each audio frame's size being 2048 points, and a hop size of 512, we obtained specific parameters.

Figure 5: STFT scheme for Audio as spectrogram

The reference literature utilized two similar models, one using 1D data and the other using 2D data. Both models were provided with 8-second audio inputs sampled at 16 kHz. For the 1D data model, with the specified audio parameters, the input size of the model is a one-dimensional vector with a length of 128,000 audio points. In contrast, for the 2D data model, instead of using direct audio signals, the authors applied the Mel-frequency cepstral coefficients (MFCC) encoding method.

By applying the STFT processing technique, we could represent the data as a spectrogram, showing the frequency content over time by applying the Fourier transform to short time frames (see Figure 5) of the audio signal. The frequency spectrum obtained from STFT was used as input for the neural network. Figure 21 illustrates the overall architecture of the model when using this input data.

3.2 Proposed Multiple DNNs

In Figure 1, the architecture is comprised of two main Convolutional Neural Network (1D CNN and 2D CNN, as shown in more detail in Figure 6) blocks and a Long Short-Term Memory (LSTM) block. The working principle of this architecture involves using CNN to extract information from the audio signals and generate sequential parameters over time. Subsequently, LSTM is applied to leverage its ability to analyse time-dependent data, using the sequential parameters created by CNN. Finally, this parameter set, containing temporal information, is passed through fully connected layers for classifying the data. In the reference literature, the authors implemented two similar structures. One structure uses direct one-dimensional audio data, while the other structure utilizes the encoded results from Mel-frequency cepstral coefficients (MFCC) and SIFT to create two-dimensional data. The 1D CNN and 2D CNN parameter are illustrated in Table 3 and Table 4.

Figure 6 provides a detailed breakdown of the CNN blocks used in our architecture. The 1D CNN block is designed with multiple layers to effectively capture and process temporal features from one-dimensional audio data. The first layer is a 1x3 convolutional layer with 64 filters and a stride of 1. This choice of kernel size allows the model to capture local patterns within the audio signal, while the number of filters ensures that a diverse set of features is learned. The inclusion of batch normalization after the convolutional layers helps in stabilizing the learning process and accelerating convergence. Following this, a 4x4 max-pooling layer with a stride of 4 is used to down-sample the data, reducing the computational load and focusing on the most salient features.

This configuration is repeated with an increased number of filters - 128 in the second set of layers - to allow the model to capture more complex features as the depth increases. The same pattern of convolution, batch normalization, and max-pooling is applied, which ensures that the model is capable of learning hierarchical feature representations from the audio data.

Similarly, the 2D CNN block is structured to process two-dimensional data, such as those derived from Mel-frequency cepstral coefficients (MFCC). The first layer in the 2D CNN block is a 3x3 convolutional layer with 64 filters and a stride of 1, followed by batch normalization and a 4x4 max-pooling layer with stride 4. The choice of a 3x3 kernel size is standard in many deep learning applications because it is effective in capturing spatial hierarchies and details within the data. The subsequent layers follow the same pattern but increase the number of filters to 128, allowing for more complex and abstract feature extraction.

The meticulous selection of parameters, including the number of layers, the size of filters, and the use of batch normalization and max-pooling, reflects the ambition to create a robust and efficient model. The multiple layers of convolutions ensure that both local and global patterns are captured, while the progressive increase in filter numbers allows the model to learn increasingly abstract features as the data flows through the network. These design choices are aimed at maximizing the model's ability to generalize from training data to unseen samples, ultimately leading to more accurate and reliable classifications.

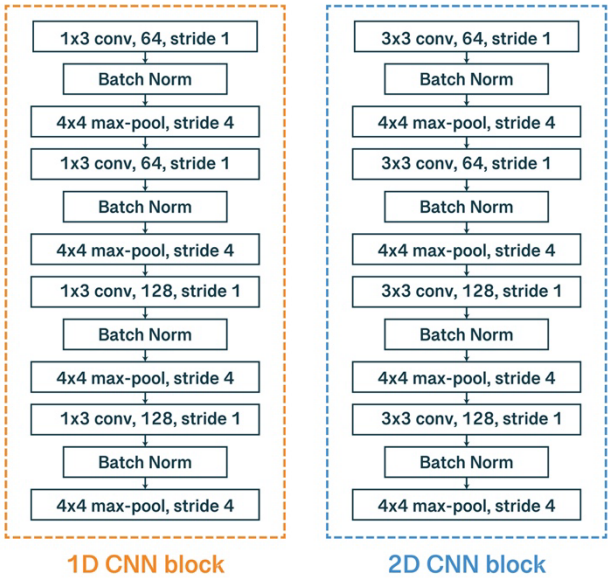

Figure 6: 1D CNN and 2D CNN Block

A feasible variant of these models is to switch from using a 16kHz frequency to an 8kHz frequency. The main idea behind this adjustment is the recognition that the frequency of human speech primarily focuses below the 4kHz threshold, so data with frequencies above 4kHz is relatively scarce. We performed high-frequency removal, retaining only data with an 8kHz frequency, and applied it to the adjusted models with an attention mechanism.

Table 3: Parameter for 1D CNN

1D CNN Network architecture parameters						
Layer	input dimensions	Output dimensions	Filter	Kernel	Stride	Activation function
1 Conv	128000	128000*64	64	3	1	
Batch normalization						ReLU
1 pooling	128000*64	32000*64		4	4	
2 Conv	32000*64	32000*64	64	3	1	
Batch normalization						ReLU
2 pooling	32000*64	8000*64		4	4	
3 Conv	8000*64	8000*128	128	3	1	
Batch normalization						ReLU
3 pooling	8000*64	2000*128		4	4	
4 Conv	2000*128	2000*128	128	3	1	
Batch normalization						ReLU
4 pooling	2000*128	500*128		4	4	
Lstm	500*128	1*256	256			Tanh
Flatten	1*256	256				
fc	256	7	7			softmax

Table 4: Parameter for 2D CNN

2DCNN Network architecture parameters						
Layer	input dimensions	Output dimensions	Filter	Kernel	Stride	Activation function
1 Conv	128*251	128*251*64	64	3*3	1*1	
Batch normalization						ReLU
1 pooling	128*251*64	64*125*64		2*2	2*2	
2 Conv	64*125*64	64*125*64	64	3*3	1*1	
Batch normalization						ReLU
2 pooling	128*251*64	16*31*64		4*4	4*4	
3 Conv	16*31*64	16*31*128	128	3*3	1*1	
Batch normalization						ReLU
3 pooling	16*31*128	4*7*128		4*4	4*4	
4 Conv	4*7*128	4*7*128	128	3*3	1*1	
Batch normalization						ReLU
4 pooling	4*7*128	1*1*128		4	4	
reshape	1*1*128	1*128				
Lstm	1*128	1*256	256			Tanh
Flatten	1*256	256				
fc	256	7	7			softmax

Diference type for comparation with proposed DNN: To understand the positive impact of the proposed model, we conduct a comparison with four models as list in Figure 7, Figure 8, Figure 9, and Figure 10.

To improve accuracy, we use a pretrained model strategy. This involves initializing the weights of a pre-trained model through transfer learning, using large-scale data from image classification [16] and speech recognition [17]. Despite speech and image tasks being different fields, they share the same network input after preprocessing. Our CNN model uses the spectrum map as input, which is treated as an image. Experiments confirmed the effectiveness of transfer learning. Initially, the model uses weight parameters from a natural scene image database (ImageNet) with 1,000 classes, then fine-tunes the weights using our speech emotion database.

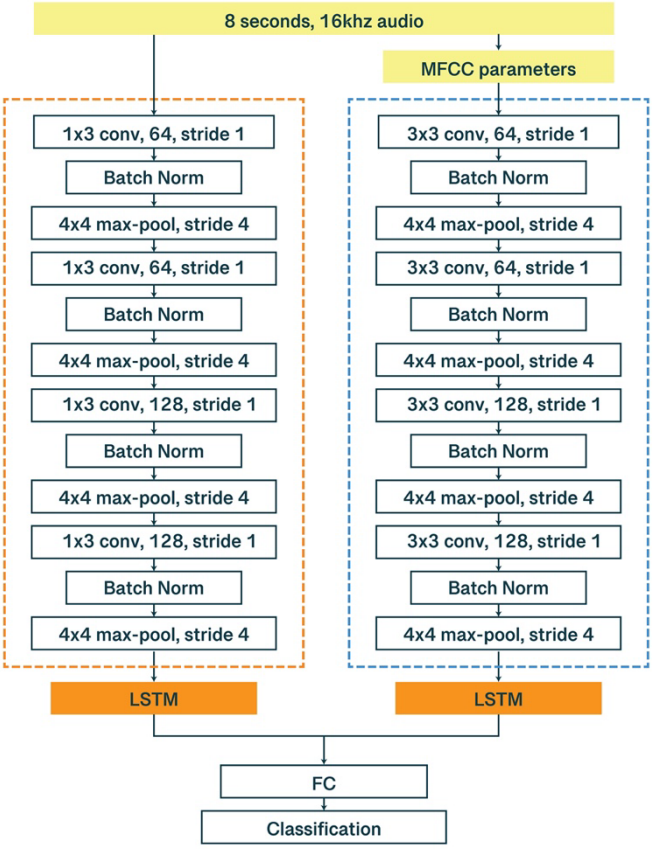

Figure 7: Two stream CNN-LSTM as Type 1

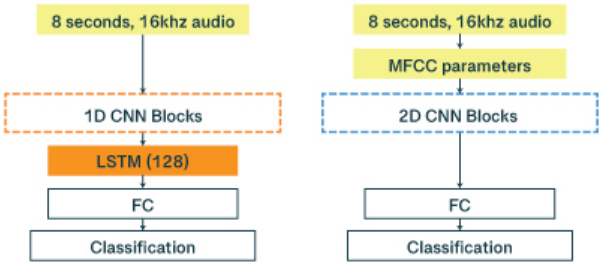

Figure 8: 1D CNN- LSTM as Type 2(a) and 2D CNN as Type 2 (b)

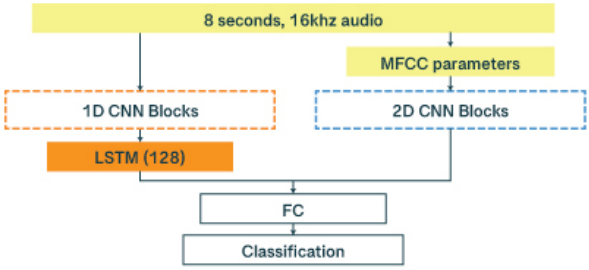

Figure 9: Two stream 1DCNN-LSTM and 2D CNN as Type 3

4. Experiment Results and Discussion

4.1 Dataset and Parameter Setups for Experiments

In this study, we utilized the EmoDB dataset to evaluate the models. The dataset comprises 535 audio recordings collected from 10 participants (5 males and 5 females), each exhibiting 7 different emotions. To ensure fairness in testing, we rigorously adhered to uniform data splitting for all tests. The data was divided into two sets with an 8:2 ratio, ensuring even distribution across the emotional classes.

All experiments were conducted on Google Colab servers, leveraging their robust computational resources for training and evaluating the proposed models. For a detailed insight into the server's computational capacity, the CPU specifications are as follows:

- Processor: Intel(R) Xeon(R) CPU @ 2.20GHz
- CPU Family: 6
- Model: 79
- Cache Size: 56320 KB
- Cores: 1
- Siblings: 2

The server configuration facilitated multitasking capabilities and efficient memory utilization, which are essential for the successful training and evaluation of complex deep learning models.

After training the proposed model on the training dataset, we conducted testing on a separate test dataset and computed

performance metrics to assess its classification capability. Accuracy measures the percentage of correctly classified samples, calculated as the ratio of the sum of true positive (TP) and true negative (TN) samples to the total number of samples:

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

4.1 Evaluation Impact of the Difference Model

We began evaluation of the efficiency model architecture as proposed. The goal was to analyse and suggest changes to improve the model's accuracy as shown in Table 5.

1D + 2D combines two architectures as Type 1: During the experimentation process, we conducted testing on three different model architecture. We built the 1D CNN and 2D CNN models with four main blocks, each comprising one convolutional layer, one batch normalization, and one max-pooling layer. For the 1D CNN processing block, the number of data channels is expanded from 1 to 64, ending at 128. This expansion similarly occurs sequentially in the blocks of the 2D CNN stream. Detailed parameters of the two models are presented in Tables 2 and 3. However, due to the unclear points in the reference literature, we encountered differences between these architecture versions. This led to experimental results that did not align with what was initially reported in the reference document. Specifically, the accuracy rates of the 1D CNN and 2D CNN architectures were 77.17% and 77.99%, respectively. When we experimented with combining the two architectures in a parallel model, as illustrated in Figure 7 below, the recognition rate reached 79.51%. This result indicates that combining the architectures in a parallel model led to a significant improvement in recognition performance compared to using each architecture independently.

This can be explained by the fact that when using 1D CNN, the model can focus on the short-term frequency characteristics of the audio signal, capturing important information related to amplitude and frequency at a specific time. This is particularly useful in recognizing audio features such as rhythm and timbre.

Meanwhile, 2D CNN has the ability to process the spatiotemporal aspects of the audio signal, representing frequency information over time. This allows the model to recognize variations in the audio signal over time, suitable for elements such as fluctuations and spectral content.

When combining both architectures in a parallel model, the model can learn both types of features simultaneously. This leads to a more robust model capable of recognizing and identifying more complex classification patterns than using each architecture independently. Through combination, the model can leverage the distinct advantages of each architecture to improve overall recognition performance, especially when the audio signal features depend on both spatiotemporal and frequency domains.

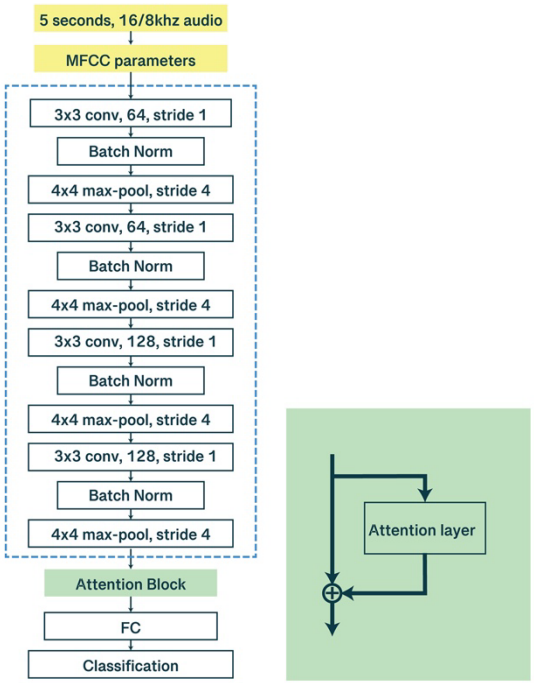

Figure 10: One stream followed by Attention Mechanism as Type 4

In the unique directional neural network architecture, the initial output of the LSTM block was set to 256. However, since the output size of the fourth CNN layer is 500*128, to match the input size, we adjusted the output of the LSTM block from 256 to 128. This helps balance the sizes between the inputs and outputs of the layers in the model.

Modified 1D and 2D CNN architecture (Type 2) shown in Figure 8 for the bidirectional model, the output size of the fourth CNN layer is only 11128, indicating a single time step in the data. Due to the concentration of data into a single time step, using an LSTM block to learn is no longer necessary, as the temporal relationships have already been captured in the preceding layers. With the neural network architecture described earlier, the experimental results indicate that the adjustments did not lead to a significant improvement in accuracy. The accuracy of the 1D and 2D models remained at 78.92% and 79.97%, respectively.

Modified 1d combined with 2d + Spectrum when changed to input as Type 3 (Figure 9): After conducting experiments, we observed that the changes did not lead to a significant improvement in recognition performance. Therefore, we delved deeper by examining the characteristics of the dataset. Statistical analysis of the data revealed that out of a total of 535 samples, only 25 samples had a duration exceeding 5 seconds, and only 1 sample had a duration exceeding 8 seconds. This indicates that the majority of the data in the dataset has a duration of less than 5 seconds. To meet the requirement for an 8-second duration, pre-processing the data by adding padding to create samples of the desired length would be reduced if we chose to fix the audio samples at 5 seconds instead of 8 seconds. We applied this modification to both types of models, including the original model and the adjusted models. After experimenting with both types of architectures, we obtained results showing that reducing the input size from 8 seconds to 5 seconds improved performance for the majority of cases. Type 3 reached 88.4% at 5 seconds and employing the first subject as the test data in the leave-one approach.

The proposed DNN shown in Figure 2, we considered for pre-processing the input data was using Short-Time Fourier Transform (STFT) instead of employing the Mel-frequency cepstral coefficients (MFCC) technique. The reason for this choice is that STFT focuses on representing audio signals in the frequency-time domain. The testing results of the proposed model applied to 5-second data for each type of model, including the 1D CNN model and two 2D CNN model base on Non-Local Attention, achieved accuracies of 90.1%. While approach pretrained model reached 93.2% as well.

Table 5: Comparison proposed DNN and Its variants

Input audio length	8 seconds	5 seconds	5 seconds ***	5 seconds ****
Data volume	3500 subjects for training and 105 subjects for verification Test set is not fixed (49 samples from the initial subjects for leave-one)			
1D CNN architecture	68.49%	77.17%*		
2D CNN architecture	76.96%	74.99%*		
1D + 2D combines two architectures as **Type 1**	77.81%	79.51%*		
Modified 1D CNN architecture as **Type 2 (a)**	69.82%*	78.92%*		
Modified 2D CNN architecture as **Type 2(b)**	78.14%*	79.43%*	79.97%*	78.03%
Modified 2D architecture **Type 2(b)** + reduced from 16k to 8k input		68.18%*	82.51%*	76.96%
Modified 1d combined with 2d + Spectrum when changed to input as **Type 3**	82.25%*	83.10%*	88.41%*	85.14%
Modified 2D architecture + attention mechanism as **Type 4**		79.43%*	83.55%*	80.51%
Modified 2D architecture + attention mechanism as **Type 4** + input reduction		79.97%*	85.55%*	78.51%
Combining input time spectrum with modified 2d architecture	69.27%*	74.18%*		
1d CNN architecture adds attention mechanism	19.29%*	57.06%*		
Proposed DNN without Pretrained	83.0%*	82.7%*	**90.1%***	87.3%
Pretrained + **Proposed** DNN	87.0%*	88.5%*	**93.2%***	91.7%

*(Without *) is the complete identification result obtained using the leave-one method.*

** Experimental results were derived by employing the first subject as the test data in the leave-one approach.*

*** Referencing the paper's experimental outcomes; however, due to the unavailability of data segmentation details, only the reference is provided.*

**** Data padding correction was performed for mirror padding.*

***** Besides using mirror padding, data augmentation also employed time difference amplification to enhance data differences.*

The impact of model design on experimental results is significant. The careful selection and combination of model architectures, such as integrating 1D and two 2D CNNs, enabled the model to capture both short-term frequency characteristics and long-term spatiotemporal features of the audio signals. This dual approach ensured a more comprehensive feature extraction process, leading to higher accuracy rates. Adjusting the input sizes from 8 seconds to 5 seconds also contributed to performance improvement by better aligning with the characteristics of the dataset. Moreover, using STFT for pre-processing allowed the model to effectively represent audio signals in the frequency-time domain, further enhancing recognition accuracy. These thoughtful design choices, including the configuration of CNN layers, filter sizes, and the use of advanced pre-processing techniques, collectively improved the model's ability to process and classify audio data accurately, demonstrating the importance of a well-designed architecture in achieving optimal experimental results.

Table 6: Lists the comparison of average accuracies from the proposed DNN and conventional ones. The structure of combinatiion CNN-LSTM reveal good accuracy of 92.8 and 92.9 while our proposed DNN is higher in magin of 0.03%. The proposed DNN adopted pretrained model achieves the second best performance which is still higher than that from the CNN-LSTM [13] on EmoDB dataset.

Ref.	model	Accuracy (%)
[10]	CNN	84.3%
[18]2021 a	CNN	77%
[19]2020 b	CNN and Bi-LSTM	94%
[20]2021 c	SVM	80.05
[21]2022 d	CNN-VGG16	92.8%
[22]2023 e	VQ-MAE-S-12 (Frame) + Query2Emo	90.2
[13]	CNN-LSTM	92.9%
Proposed DNN	Multiple CNN-LSTM based Non-Local Attention	90.1%
	Pretrained + Multiple CNN-LSTM based Non-Local Attention	93.2%

In addition, differences could also arise from data labelling. The reference document did not provide specific information about the number of participants in the training and testing data. Acknowledging that such variations could occur, after each improvement and correction step, we addressed errors in the padding section, causing the entire data to become sparse. Simultaneously, we adjusted the data augmentation approach. Instead of copying data with the same quantity, we transitioned to time warping, a method previously applied and shown positive results in augmenting training data.

Despite the trimming and adjustments made to better fit the model architectures, such as resizing and removing unnecessary layers, the recognition results did not show significant improvement. Through further investigation, we found that the input data itself.

In choosing a duration for the neural network, we adopted for 5 seconds based on dataset statistics, acknowledging that a simpler network structure improved recognition performance for the task. However, the ability to recognize based on the time domain was not effective when testing with 8-second input data. Ultimately, incorporating attention mechanisms into the model structure rendered testing with 8-second data entirely unviable, while maintaining good recognition rates with 5-second input data. Although it is not certain whether extending to 8 seconds is a good approach, it did completely eliminate the inherent characteristics of the dataset. Based on the test results above, if we focus on audio inputs and apply MFCC processing, the length of the audio is a factor that influences recognition results to some extent.

5. Conclusion

In this work, we have successfully developed a new multiple DNN based on Non-local Attention. Experimental results show that proposed DNN using pretrained achieves superior accuracies than its modified version. Although this research has not achieved an state of the art result, with the highest recognition rate reaching about 93.2%, it has incorporated and compared various architectures, laying the groundwork for future research to expand and compare with even more architectures. The ultimate goal is to find an increasingly better model as an optimal tool for emotion recognition tasks. From the perspective of this study, proposed DNN for emotion recognition from speech is a futher promising, especially from the research community, its application in recognizing emotional content in speech could open up new dimensions, including predicting emotional states and psychological states in the future. We expect to continue developing in this direction in the future.

In the future, we will focus on enhancing the training data by collecting and integrating larger and more diverse datasets to improve the model's generalization ability and accuracy. Additionally, we will explore new audio preprocessing methods and compare them with other advanced deep learning models, such as Transformer and RNN variants, to optimize the model architecture. Furthermore, research and optimization of the non-local attention mechanism will continue to improve the model's performance.

We anticipate that this model will not only be applicable to emotion recognition from speech but also open up other practical applications such as predicting psychological states and aiding psychological therapy. Recognizing more complex emotional states could become an important tool in

fields like mental health care and user behavior analysis. This model also has the potential for widespread application in analyzing emotions from speech in everyday communication scenarios, thereby contributing to the improvement of human-machine communication systems and enhancing user experience in intelligent interaction services.

References

[1] Depressive disorder (depression), https://www.who.int/newsroom/factsheets/detail/depression.

[2] Mental Health Conditions: Depression and Anxiety, https://www.cdc.gov/tobacco/campaign/tips/diseases/depression-anxiety.html

[3] Europe's mental health crisis: Which country uses the most antidepressants? https://www.euronews.com/next/2023/09/09/europes-mental-health-crisis-in-data-which-country-uses-the-most-antidepressants

[4] M. -H. Ha and O. T. -C. Chen, "Deep Neural Networks Using Residual Fast-Slow Refined Highway and Global Atomic Spatial Attention for Action Recognition and Detection," IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3134694.

[5] M Gil, SS Kim, EJ Min - Frontiers in Public Health, 2022, "Machine learning models for predicting risk of depression in Korean college students: Identifying family and individual factors"

[6] K. Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech Emotion Recognition Using Fourier Parameters," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015, doi: 10.1109/TAFFC.2015.2392101.

[7] K. Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech Emotion Recognition Using Fourier Parameters," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015, doi: 10.1109/TAFFC.2015.2392101.

[8] A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea (South), 2017, pp. 1-5, doi: 10.1109/PlatCon.2017.7883728.

[9] Ma, Xingchen & Yang, Hongyu & Chen, Qiang & Huang, di. (2016). DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. 35-42. 10.1145/2988257.2988267.

[10] A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea (South), 2017, pp. 1-5, doi: 10.1109/PlatCon.2017.7883728.

[11] Fayek, Haytham & Lech, Margaret & Cavedon, Lawrence. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks. 92. 10.1016/j.neunet.2017.02.013.

[12] Tripathi, Samarth & Beigi, Homayoon. (2018). Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning.

[13] Zhao, Jianfeng & Mao, Xia & Chen, Lijiang. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control. 47. 312-323. 10.1016/j.bspc.2018.08.035.

[14] Manh-Hung Ha and Osacl T C Chen "Non-local Spatiotemporal Correlation Attention for action recognition" in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2022

[15] Busso, C., Bulut, M., Lee, CC. et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resources & Evaluation 42, 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6

[16] Krizhevsky A., Sutskever I., Hinton G. E. (2012). "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 1097–1105. 10.1145/3065386

[17] Dahl G. E., Yu D., Deng L., Acero A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 20, 30–42. 10.1109/TASL.2011.2134090

[18] M. H. Pham, F. M. Noori and J. Torresen, "Emotion Recognition using Speech Data with Convolutional Neural Network," *2021 IEEE 2nd International Conference on Signal, Control and Communication (SCC)*, Tunis, Tunisia, 2021, pp. 182-187, doi: 10.1109/SCC53769.2021.9768372.

[19] A. Yadav and D. K. Vishwakarma, "A Multilingual Framework of CNN and Bi-LSTM for Emotion Classification," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225614.

[20] Haider, Fasih, and Saturnino Luz. "Affect Recognition Through Scalogram and Multi-Resolution Cochleagram Features." 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021. 2021.

[21] Rudd, David Hason, Huan Huo, and Guandong Xu. "Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2022.

[22] Sadok, Samir, Simon Leglaive, and Renaud Séguier. "A vector quantized masked autoencoder for speech emotion recognition." arXiv preprint arXiv:2304.11117 (2023).