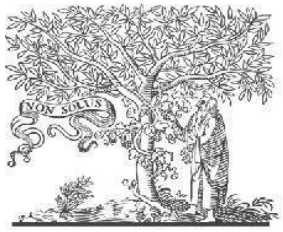
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 8th Aug 2024. Link

https://ijiemr.org/downloads.php?vol=Volume-13& issue=issue08

DOI: 10.48047/IJIEMR/V13/ISSUE 08/6

Title SPEECH EMOTION RECOGNITION FOR AUTISM SPECTRUM DISORDER

Volume 13, ISSUE 08, Pages: 41 - 49

Paper Authors
Dr. G.V. Ramesh Babu , C. Mounika

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

SPEECH EMOTION RECOGNITION FOR AUTISM SPECTRUM DISORDER

Dr. G.V. Ramesh Babu
Associate Professor, Department of Computer Science, Sri Venkateswara University, Tirupati
gvrameshbabu74@gmail.com
C. Mounika
Master of Computer Applications, Sri Venkateswara University, Tirupati.
mounika311019@gmail.com

Abstract
Our project demonstrates how an emotion can be inferred from the audio file in which the speaker has spoken. The primary goal of this project development is to assist children with Autism Spectrum Disorder (ASD) who cannot identify emotion from speech and may benefit from this project's ability to do o. We are concentrating mostly on this model's fundamental problem of high variance (Overfitting), which may be caused by the lack of audio recordings utilized to train the model. Ryerson's Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS) dataset were excellent training datasets for this model to overcome the overfitting issue. Noise removal will be performed as part of the pre-processing step using Python algorithms. Mel-frequency Cepstrum Coefficient (MFCC) simulates the audio features mechanism. Numerous applications support human-computer interactions, but in this case, we're introducing deep learning neural networks (CNN) and Multilayer Perceptron (MLP) to recognize and categorize the precise output. Our final model will classify 8 different emotions (anger, calm, disgust, fearful, happy, neutral, sad, and surprised) with better accuracy.
Keywords—Autism Spectrum Disorder (ASD), Mel frequency cepstral coefficients (MFCC), CNN, MLP, RAVDESS, TESS.

Introduction

The most efficient means of expressing one's opinions is through communication. Depending on their state of mind, humans communicate their sentiments through a variety of emotions, including anger, sadness, happiness, fear, and enthusiasm. Autism Spectrum Disorder is a condition that affects the functionality of the brain. No age group of people is exempt from this neurological disorder. ASD sufferers struggle to interact and socialize with others. Their sense of perception is distinctive. Since there is no precise medical test to identify ASD in humans, diagnosis is challenging. Doctors consider the child's behavior and development behavior to determine a diagnosis. In this paper, we intend to develop a model of speech emotion recognition by using machine learning and a deep learning approach that will make it easier for people with ASD to interact socially. Deep learning and machine learning employ statistical models to make predictions.

In the development of every machine learning model, the initial and most important step is data pre-processing. Data pre-processing is a procedure that takes raw data and converts it into a format that computers and machine learning software can understand and evaluate. Raw data often contains some noise. In this context, raw data is the audio files. The presence of Noise greatly affects the performance of the

model. So, in our proposed model we used a noisereduce algorithm in a pre-processing step. We found some limitations in the existing models. one major is the overfitting problem which we overcome in our proposed model by using multiple corpse datasets like RAVDESS and TESS. From large feature sets, we extract required acoustic features to train the model with the help of Mel frequency cepstral coefficients (MFCC). MFCCs reduce the quantity of data in a frame of speech's Fourier transform to a condensed set of values. A wide variety of classification techniques are present in the machine and deep learning methods to predict the emotion precisely and test the model's performance. The proposed model uses CNN and MLP classifiers as its learning methods.

Related Work

Md. Shah Fahad et al [1] described the study by examining speech emotion recognition in an outdoor setting. This study focuses on Speech Emotion Recognition (SER)i in a natural setting, including its benefits and drawbacks. They also talked about how to improve speech signals for real-world speech signals using techniques like spectrum subtraction, wiener filtering, and MMSE. It has been found that the spectral subtraction and MMSE approaches for emotion recognition are resistant to the airport and chatter noises. Deep learning-based difficulties in a natural setting are limited, and their solutions are not yet fully realized. Tzirakis et al [2] proposed an End-to-End multimodal Emotion Recognition Using Deep Neural Networks was proposed by Robust features must be retrieved from the proposed model's utilization of auditory and visual modalities to capture the emotional content of different speaking styles. They use convolutional neural networks to extract features from speech for this purpose.

Rezwan Matin et al [3] suggested a spoken emotion recognition model to help children with Autism Spectrum Disorder in recognizing emotion based on the Support Vector Machine. The model was trained using data from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) corpus. To train the model, the first 26 Mel frequency Cepstral Coefficients (MFCCs) and the zero-crossing rate (ZCR) are retrieved. A test accuracy of 77% was provided by the final SVM model. The issue of large variance is one of the model's main flaws. It might be because there weren't enough audio recordings to train the model. Minajul Haque et al [4] suggested a model for removing voice background noise using various linear filtering approaches. To eliminate background noise from the speech, adaptive &Kalman filters that have been constructed and tested over voice signals were presented in the proposed model. It has been found that the Kalman filter outperforms other adaptive filters like the LMS and NLMS for AWGN (Gaussian) noise. The above methodology's drawback is that performance changes when noise becomes practical.

Siddique Latif et al [5] put forth a multi-task, semi-supervised, adversarial autoencoding algorithm for speech emotion recognition. Recognition of emotions and speech Adversarial Autoencoding with Multi-Task, Semi-Supervised With the aid of 10-fold cross-validation, compare the performance of the technique with other frameworks and improve the auxiliary tasks by adding more data, which mostly contributes to improving the primary job. For leave-one-speaker-out validation on the IEMOCAP and MSPIMPROV datasets, combining auxiliary tasks is preferable to using them separately since it increases the precision of the main job. However, this strategy offers the most accuracy for the primary task. when both category and dimensional emotion representations are used along with the auxiliary tasks.

Shiqing Zhang et al [6] suggested a method using Deep Convolutional Neural Networks (DCNN) and SVM classifiers for voice emotion recognition. This method uses Discriminant Temporal Pyramid Matching and deep convolutional neural networks. Effective characteristics from speech might be extracted using DCNN models trained on extensive ImageNet data and input obtained from Mel-spectrogram segments. As a result, training DCNN with less annotated speech data is made simpler. We encounter several

issues with this method since it is unable to handle continuous, multidimensional emotion detection and cannot identify spontaneous emotions.

Nasir Saleem et al [7] proposed On Learning Spectral Masking for Single Channel Speech Enhancement Using Feedforward and Recurrent Neural Networks. In the proposed approach, to clear the time-domain speech from background noise, which is further used in the single-channel speech enhancement approach, they used RNNs and DNNs followed by acoustic features. They found that the evaluated values of speech quality and intelligibility in terms of overly noisy conditions (SDR, PESQ, STOI, and ESTOI) are improved by 2.75 dB, 8.67%, and 4.76%, respectively. In the TIMIT dataset's environment when compared to baseline methods, this approach achieves less complexity in computation.

Nikolaos Vryzas et al [8] proposed continuous speech emotion recognition with convolutional neural networks. In the proposed system, they use the Acted Emotional Speech Dynamic Database (AESDD), which takes audio streams as input followed by frame-level Mel-scale bands on continuous speech. Data augmentation is also applied to generalize and improve the designed network robustness. The overall performance accuracy of this approach is 6.4% higher than that of other ML models like SVM, so it is very effective. But the problem with using this approach is that the classification accuracy is sometimes altered due to data augmentation.

Table 1: Existing Techniques Analysis

S.No	Author	Technique	Strengths	Limitations Or Future Work
1	Md. Shah Fahad	Conventional model, DNN	The weighted sparse representation model used in this study, which is based on the maximum likelihood estimate, increased the classifier's accuracy.	It may perform better when distinct subsets of features are chosen for each node's clean and noisy environments.
2	Tzirakis	Long short-term memory model (LSTM) and Convolutional neural network (CNN).	Both in terms of valence and arousal, the proposed method performs significantly better than the traditional methods.	Future studies will involve expanding the method's modalities to enhance performance.
3	Rezwan Matin	Support Vector Machine (SVM)	The performance is improved by applying the native sampling technique.	The low amount of audio files was the cause of the high variance issue.
4	Minajul Haque	adaptive filtering, kalman filter	NLMS converges more quickly than LMS.	The performance changes in the presence of practical noise.
5	Siddique Latif	Adversarial autoencoders (AAE)	Utilizing the unlabelled data for auxiliary tasks improves accuracy while overcoming the	In the future, this architecture will incorporate reinforcement learning for better results.

			problem of the restricted availability of emotion datasets.	
6	Shiqing Zhang	Deep Convolutional Neural Networks (DCNN) and SVM classifier.	The suggested method makes it simpler for DCNN to train with a small amount of annotated voice data.	This method cannot handle continuous dimensional emotion detection and cannot identify unprovoked emotions.
7	Nasir Saleem	recurrent neural network (RNN) and Deep neural network (DNN).	Compared to the standard DNN/RNN algorithms, this model was able to attain less computational complexity and quick convergence.	The suggested approach is limited to a sole task and ineffective for other voice recognition tasks.
8	Nikolaos Vryzas	Convolutional Neural Networks (CNNs)	The robustness and generality of the database are enhanced through data augmentation.	We can investigate speaker-dependent and speaker-independent techniques in the future.

Proposed Methodology

Fig. 1 shows the block diagram of the proposed model. The selection of the databases is a crucial step in speech emotion recognition because the model's effectiveness depends on how natural the database is. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto Emotional Speech Set (TESS) English language databases are used in this work. We pre-process the audio recordings present in those two datasets to remove any noise before using them. To train the model, the acoustic features are retrived using MFCC. These features are provided to classifiers to help them recognize emotions and to evaluate their accuracy. MLP and CNN are the classifiers utilized in this work.

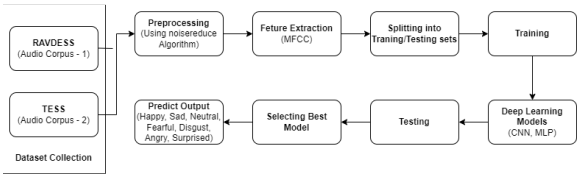


Fig. 1: System Architecture

A.Datasets Description

1. RAVDESS

It is a multimodal database including recordings in North American English, both audio and video. Researchers from Ryerson University's SMART Lab developed the corpus. Anyone can access it because it is a public database with a Creative Commons Attribution license. 24 actors who are professional (12 men and 12 women) totaled took part in the recording. RAVDESS has 7,356 files totaling 24.8 GB. There are 1,440 files total in the audio-only recordings, 60 trials for each of the 24 actors (215 MB of data). The audio was captured in WAV format with a bit depth of 16 bits and a sampling rate of 48,000 Hz. The speech recordings showed eight different emotions: happy, sad, neutral, surprised, calm, angry, fearful, and disgust.

2. TESS

Two actresses (26 and 64 years old) recited a set of 200 target words in the carrier phrase "Say the word _," and recordings of the set evoking each of the eight emotions were made (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

There are a total of 2800 data points (audio files).

B. Pre-processing

Before collecting the signal's features from the voice samples, a procedure called pre-processing is used. Speech samples include undesirable information that was present during the recording of the speech, such as noise and some environmental variations. Noise is an erroneous signal that can occur during audio signal
transmission or recording. The noise removal method should be used as a first step before various advanced sound processing activities. The model's accuracy will be impacted by the ambient noise. As a result, as part of the emotion recognition model's pre-processing, we eliminate background noise. Therefore, we employ the noisereduce algorithm.

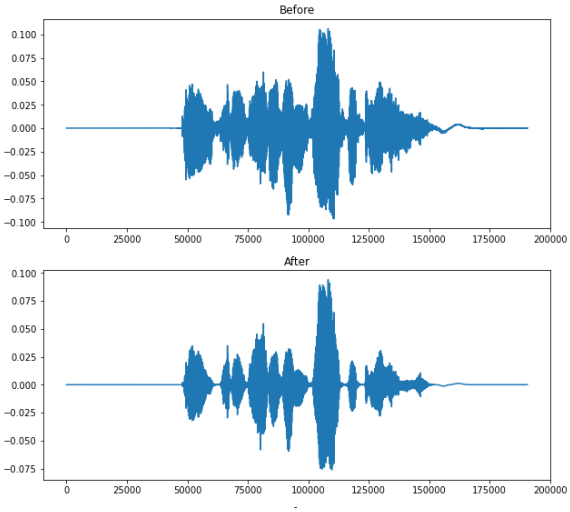


Fig. 2: Applying pre-processing on an audio clip

Python's noisereduce algorithm for reducing noise in time-domain inputs, including voice, bioacoustics, and physiological signals It makes use of a technique known as "spectral gating," which is a kind of Noise Gate. It computes the spectrogram of the noise signal and calculates the noise threshold for each frequency band. The frequency-varying threshold is used to generate a mask, which gates noise below it. The method requires two inputs: a noise clip with the clip's prototype noise and a signal clip with the intended target signal and noise. Over the audio clip with the noise, a spectrogram is computed. Statistics are computed over the noise spectrogram (in frequency) Using the noise statistics, a threshold is determined. The signal spectrogram is compared to the threshold to determine the mask. Using a filter, the mask is rounded off in both frequency and time. The mask is inverted and applied to the signal's spectrogram. Fig. 2 depicts the audio clip's graph displaying the results of the pre-processing before and after.

C. Feature Extraction

The model's accuracy and effectiveness are determined by the feature extraction of the speech signal. To examine the signal, the acoustic properties of the speech signal, such as Mel Frequency Cepstral Coefficients, are retrieved. The most significant and effective method utilized in speech-related applications is called MFCC. The human vocal tract's structure filters the sound that is produced and determines what sound emerges. The speech signal is modeled by MFCC as the short-term power spectrum of sound. The speech signal's frequency is measured in mels. In the low-frequency area, MFCC provides improved frequency resolution. The seven steps of the MFCC are depicted in Fig. 3.

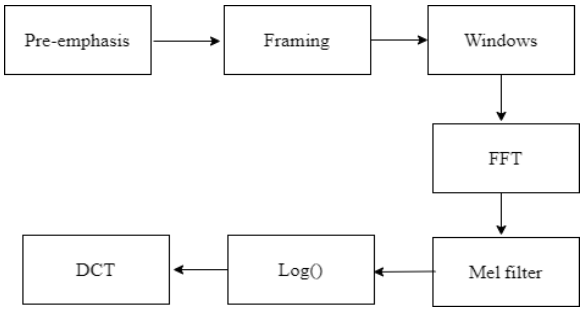


Fig. 3: Steps in Mel Frequency Cepstral Coefficients

In comparison to the high-frequency range, the voice signal is more energetic at low frequencies. Pre-emphasis is used to increase the amount of energy in the high-frequency region. An astable signal is speech. So, the signal is fragmented into tiny

timed frames. The signal's component frames are each subjected to windowing which decreases the spectral distortion and frame discontinuities at the beginning and conclusion of the signal. The windowing signal is subjected to FFT in order to transform it into the frequency domain and obtain the signal spectrum. For frequencies under 1000 Hz, the Mel-frequency scale has linear spacing, whereas, for frequencies above, it has logarithmic spacing. The speech signal's real frequency (Hz) is converted to the mel frequency using the mel function. The logarithm is used to calculate the values of the amplitude. Using a Mel-frequency scale as an expression, the real logarithm of the short-term energy spectrum underwent a cosine transform to produce MFCC.

For the MFCC extraction, we employ the Librosa audio processing library.

D. Classification

The primary goal of machine learning methodologies is to create an appropriate classification framework that identifies the class to which an observation belongs. For many applications involving emotional intelligence, the capacity to categorize emotions precisely is particularly significant. Here, CNN and MLP classifiers are proposed classification models for recognizing emotions.

MLP

A feed-forward artificial neural network called a Multi-Layer Perceptron (MLP) has layers that are all fully coupled to one another. Multi-layer perceptrons may comprise one or more hidden layers along with one input layer and one output layer. There will be a lot of hidden layers, and the number of layers can be altered depending on the situation. The features that are collected from the audio file will be input into the input layer. An activation function is used by the hidden layer to react to the input data and process the data. The logistic activation function is the one that is employed. The information that the network has learned is output by the output layer. Using the computation carried out by the hidden layer, this layer categorizes the input and output of the expected emotion.

CNN

A Convolutional Neural Network (CNN) is an artificial neural network for deep learning that is fed forward. Convolutional layers, pooling layers, and fully connected layers make up its fundamental structure. Filters are convolved or moved across the input in a convolutional layer. In order to lower the resolution of the feature map produced by the convolutional layer, pooling layers are introduced. Additionally, CNNs are capable of having multiple fully linked layers, each of which has a direct connection to every neuron in the layer below it. There can be more than two layers in a convolutional neural network, and each layer can be trained to recognize various features. Each training set of data is subjected to filters, and the output of each layer serves as the input for the subsequent layer.

In order to check and identify features that specifically reflect the input item, the complexity of the filters increases with each additional layer. As a result, the input for the subsequent layer is the output of each layer or the partially recognized data after each layer. The CNN detects emotions at the final layer. These layers carry out operations on the data in order to discover characteristics unique to the data. Convolution, activation, and pooling are three of the most used layers.

Through a series of convolutional filters, which sequentially activate different aspects of the input data, convolution processes the signals. By preserving positive values and projecting negative values to zero, activation enables quicker and more efficient training because only the active features are carried over to the following layer. By conducting nonlinear down sampling on the output, pooling reduces the number of parameters the network needs to learn. Each layer learns to recognize various traits as these procedures are repeated over several layers.

Experimental Results

In our work, 25% of the dataset is reserved for testing, with the remaining 75% being

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

used for training. Fig. 4 and Fig. 5 illustrate the performance of the classifiers MLP and CNN, which are compared afterward.

MLP

Its activation function is the tanh function. Backpropagation is a supervised learning method that is used by MLP during training. It is a linear function that translates the weighted inputs to the output of each neuron in a multilayer perceptron if all of the neurons have a linear activation function. 200 epochs of the model were run. With a weight decay of 0.0001, the starting learning rate was set to 0.001. Overall accuracy is 81% as a result.

	precision	recall	f1-score	support
angry	0.78	0.92	0.85	153
calm	0.63	0.94	0.76	69
disgust	0.99	0.71	0.83	111
fearful	0.74	0.79	0.77	164
happy	0.88	0.78	0.83	153
neutral	0.95	0.86	0.91	118
sad	0.86	0.65	0.74	162
surprised	0.77	0.93	0.84	121
accuracy			0.81	1051
macro avg	0.83	0.82	0.81	1051
weighted avg	0.83	0.81	0.81	1051

Fig. 4: The MLP model's test set results for each class

CNN

Fig. 5 shows how the CNN classifier is structured for classification. It has three layers and Relu serves as the model's activation function. A 2 × 2 pooling size maxpool layer is placed after each CNN layer. The Flatten layer, which makes the feature map compatible with the succeeding connected layers, is placed after the final CNN layer and is followed by a dropout layer with a dropout set to 0.1. Training optimization makes use of the Adam optimizer. Training is conducted at a learning rate of 0.00001. 200 epochs were run by the model. As a backend, TensorFlow and Keras are employed. The eight emotions are represented here as (neutral = 0; calm = 1; happy = 2; sad = 3; angry = 4; fearful = 5; disgust = 6; surprised = 7). It generates an accuracy of 83%.

Layer (type)	Output Shape	Param #
conv1d_54 (Conv1D)	(None, 40, 64)	384
activation_72 (Activation)	(None, 40, 64)	0
dropout_54 (Dropout)	(None, 40, 64)	0
max_pooling1d_36 (MaxPooling1D)	(None, 10, 64)	0
conv1d_55 (Conv1D)	(None, 10, 128)	41088
activation_73 (Activation)	(None, 10, 128)	0
dropout_55 (Dropout)	(None, 10, 128)	0
max_pooling1d_37 (MaxPooling1D)	(None, 2, 128)	0
conv1d_56 (Conv1D)	(None, 2, 256)	164096
activation_74 (Activation)	(None, 2, 256)	0
dropout_56 (Dropout)	(None, 2, 256)	0
flatten_18 (Flatten)	(None, 512)	0
dense_18 (Dense)	(None, 8)	4104
activation_75 (Activation)	(None, 8)	0

Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0

Fig. 5: The mechanism of the CNN classifier is described in detail.

	precision	recall	f1-score	support
0	0.95	0.86	0.90	140
1	0.69	0.89	0.78	109
2	0.84	0.81	0.82	205
3	0.89	0.75	0.82	211
4	0.86	0.90	0.88	192
5	0.81	0.76	0.78	172
6	0.77	0.88	0.82	140
7	0.82	0.85	0.84	144
accuracy			0.83	1313
macro avg	0.83	0.84	0.83	1313
weighted avg	0.84	0.83	0.83	1313

Fig. 6: The CNN model's test set results for each class

Table 2: MLP and CNN models' F1-scores for each class were compared.

CLASS	MLP	CNN
NEUTRAL	91	90
CALM	76	78
HAPPY	83	82
SAD	74	82
ANGRY	85	88

FEARFUL	77	78
DISGUST	83	82
SURPRISED	84	84

By examining the results from above Table 2, as a whole we can conclude that CNN is better than MLP at identifying emotions in speech samples but MLP is better at classifying neutral, happy, and disgust classes. The performance of the CNN for classification is significantly better. Fig. 7 and Fig. 8 depict the CNN model's reliability. In Fig. 7, we can see how, up to the 200th epoch, the value of loss tends to go down on both the training set and the test set. From the 75th epoch onward, the reduction in the test set is less obvious but still observable. According to Fig. 8, the average accuracy value across all classes rises as the number of epochs rises.

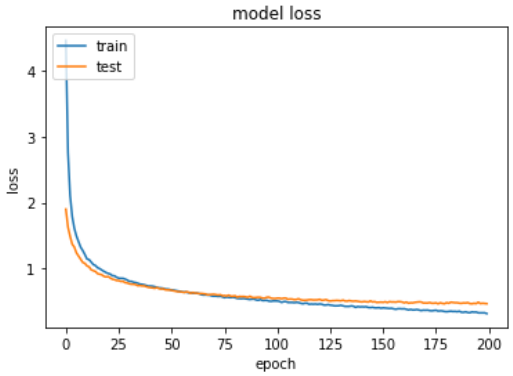


Fig. 7: A graph of our CNN's loss function over 200 epochs

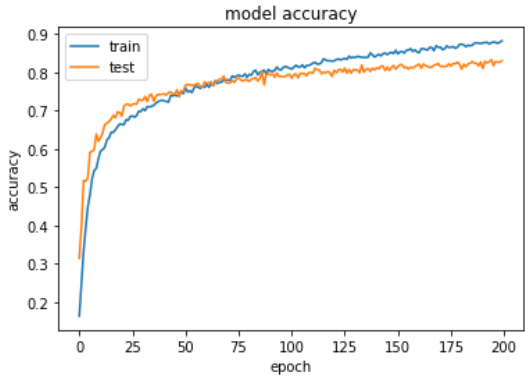


Fig. 8: A graph of our CNN's accuracy over 200 epochs

Conclusion

When we studied the baseline approaches, we found that background noise and overfitting are the main reasons why emotions are misclassified. So, we combined RAVDESS and TESS datasets as one to overcome the overfitting problem. In this work, we proposed a model based on deep neural networks for the classification of emotions using audio recordings from the combined dataset. The model has been trained to categorize eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). Using the noise reduce technique, we removed noise from the audio files used for training before extracting the MFCC features to achieve this result. We used the MLP and CNN Classifiers, both of which were trained on the same dataset and achieved an average F1 score of 0.81 and 0.83 across the 8 classes. The deep learning model we ultimately selected had an F1 score of 0.83 on the test set. The good results obtained suggest that such approaches based on deep learning methods are an excellent basis for speech emotion recognition.

References

[1] Md. Shah Fahad, Ashish Ranjan, Jainath Yadav, Akshay Deepak., A survey of speech emotion recognition in natural environment, Digital Signal Processing, Volume 110,2021,102951, ISSN 1051-2004, https://doi.org/10.1016/j.dsp.2020.102951.

[2] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal emotion recognition using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

[3] Rezwan Matin and Damian Valles, "A Speech Emotion Recognition solution-based on Support Vector Machine for children with Autism Spectrum Disorder to help identify Human Emotions", Intermountain Engineering, Technology and Computing (IETC), 2020, doi: 10.1109/IETC47856.2020.9249147

[4] Haque, M., & Bhattacharyya, K. (2018). Speech Background Noise removal using different Linear Filtering techniques. In

Lecture Notes in Electrical Engineering (pp. 297 – 307). Springer Singapore. https://doi.org/10.1007/978-981-10-8240-5_33.

[5] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps and B. W. Schuller, "Multi-Task Semi-Supervised Adversarial Autoencoding for speech emotion recognition," IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 992-1004, 1 April-June 2022, doi: 10.1109/TAFFC.2020.2983669.

[6] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576 -1590, June 2018, doi: 10.1109/TMM.2017.2766843.

[7] N. Saleem, M. I. Khattak, M. Al-Hasan and A. B. Qazi, "On Learning Spectral Masking for single-channel speech enhancement using Feedforward and Recurrent Neural Networks," in IEEE Access, vol. 8, pp. 160581-160595, 2020, doi: 10.1109/ACCESS.2020.3021061.

[8] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Continuous speech emotion recognition with Convolutional Neural Networks," J. Audio Eng. Soc., vol. 68, no. 1/2, pp. 14-24, 2020 January. doi: https://doi.org/10.17743/jaes.2019.0043