Creating an Artificial Neural Network Model to Analyse Emotions from an Audio Database.

School of Graduate Studies: Applied Modelling and Quantitative Methods, MSc in Big Data Analytics Trent University
Peterborough, Ontario, Canada

AMOD-5610H-A-S01-2023GS-BIG DATA MAJOR RESEARCH PAPER
Professor Robert Sturgeon & Professor James Parker

Harsimran Kaur - 0735180
Hindu Raj - 0741561
Sharath Rupesh - 0744758

ACKNOWLEDGEMENT

We would like to express our gratitude to Professors James Parker and Rob Sturgeon for all of their help and guidance with this project.

Abstract

Emotional prediction by speech database analysis is an emerging research field aimed at detecting and predicting emotional states based on speech data. Emotions play an important role in human communication, and understanding and predicting emotions has numerous applications in various fields such as mental health, human-computer interaction, and entertainment. This article presents an innovative approach to sentiment prediction through speech database analysis. The proposed methodology consists of several steps. First, a comprehensive audio database is created containing a large number of audio samples for various emotional states. Features such as pitch, intensity, rhythm, and spectral characteristics are extracted from the audio data using advanced signal processing techniques. These features capture important acoustic signals associated with emotions.The extracted audio features and corresponding sentiment labels are then used to train machine learning algorithms such as support vector machines, neural networks, and deep learning models. Models can learn underlying patterns and relationships between acoustic features and emotional states to make accurate predictions .Extensive experiments are performed on audio databases to evaluate the effectiveness of the proposed approach. This result indicates the possibility of emotion prediction by speech database analysis. The proposed approach achieves expected accuracy in predicting emotional states from speech data. Furthermore, this experiment revealed the importance of specific acoustic features for distinguishing emotions.

Table of Contents

Introduction 5

Objective 5

Solution approach 6

Why is the study being done 6

Background 6

Literature Review 8

Dataset Description 10

Exploratory Data Analysis 11

Methodology 16

About Multi Layer Perceptron (MLP) Model 19

Other Alternatives to the model Used 20

Why was MLP chosen? 21

Pros of using MLP 21

Cons of using MLP 22

Results 23

Discussion and Conclusion 24

References 26

List of Figures and Tables

Figure 1. The waveplot 11

Figure 2. The Spectogram 11

Figure 3. Spectogram Color Shift 12

Figure 4. The Mel-Spectogram 12

Figure 5. Features Extracted from a Spectogram After Fourier Transformation 13

Figure 6. Mel Frequency Cepstral Coefficients 13

Figure 7. Features Extracted from a MelFrequency Cepstral Coefficients After Fourier Transformation 14

Figure 8. The Chroma Spectogram 14

Figure 9. Features Extracted from a Chroma Spectogram After Fourier Transformation 15

Figure 10. The zero crossing 15

Figure 11. Accuracy Validation Graph 18

Figure 12. Accuracy after training the model 23

Figure 13. Predicted Emotion 24

Table 1. Emotion Labels 16

INTRODUCTION

Understanding and predicting human emotions is an interesting and complex area of research with wide-ranging implications for fields as diverse as psychology, neuroscience, human-computer interaction, and entertainment. Emotions play an important role in human communication, decision-making, and overall well-being. The ability to accurately predict and interpret emotions will greatly improve our understanding of human behavior and enable the development of more empathic and responsive systems.

Traditionally, emotion detection and prediction relied on various modalities such as facial expressions, body language, and physiological cues. However, one modality that has grown in importance in recent years is audio data. Speech signals contain valuable information about emotional states such as intonation, speech patterns, and intonation. Analyzing and interpreting this information from speech databases provides valuable insight into human emotions.

The advent of big data, advances in signal processing techniques and machine learning have opened up new possibilities for analyzing speech databases to predict emotions. By extracting meaningful acoustic features and training machine learning models on labeled speech data, researchers can develop systems that can predict emotional states based solely on speech input.

The purpose of this article is to explore the field of emotion prediction using analysis of speech databases. We address the methodologies required to build comprehensive speech databases, extract relevant acoustic features, and train machine learning models to predict emotions. Furthermore, we discuss the potential applications of emotion prediction from speech data in various fields.

The two goals of this study are to highlight the possible applications of accurate emotional prediction in practice and to test how well audio database analysis predicts emotions. This research intends to contribute to the expanding body of knowledge in the field and stimulate further improvements by illuminating the advantages and disadvantages of this method.

The following section reviews the methodology, model and results related to sentiment prediction from speech database analysis. By exploring the possibilities of this approach, we hope to pave the way for future developments and applications that harness the power of speech data to understand and predict human emotions.

Objective:

Emotions have a critical role in human mental health. It is a way of expressing one's viewpoint or emotional condition to other people. Through their sensory systems, humans are able to detect one another's emotional condition. However, it is difficult to accomplish the

same with a computer. Speech emotion recognition seeks to gain the depth beneath material, which is difficult to obtain even though computers can quickly interpret content-based information.
Speech Emotion Recognition (SER) is the extraction of a speaker's emotional state from a speech stream. Any intelligent system with limited processing power can be taught to recognize or synthesize a small number of universal emotions. These include neutral, composed, fear, rage, joy, and sadness, among others. Speech emotion detection is increasingly being employed in industries like healthcare, entertainment, and education.
With the aid of the Keras and TensorFlow frameworks, we will create a model in this project that can identify emotions from audio samples. A MLP from the Sklearn package will also be used to create a model.

Solution approach:

You must first gather and prepare an audio database of emotional states. After that, you divided the data into sets for training, validation, and testing. Next, based on the type of data, you create the architecture of your neural network, such as a feedforward neural network or recurrent neural network. The model is then trained using the training data, with the model's parameters adjusted to reduce the discrepancy between the anticipated and real emotion labels. Using the testing set after the model has been trained, you may gauge its effectiveness by looking at measures like accuracy, precision, recall, and F1 score. To enhance outcomes, you can adjust the model's architecture and hyperparameters as necessary. Then, taking into account computing needs and keeping up with the most recent developments in audio and emotion analysis research, you deploy the model for emotion analysis on fresh audio samples.

Why is the study being done?

The study of building an ANN model to extract emotions from audio databases has important ramifications for our comprehension of emotions, enhancing human-computer interaction, assisting psychological research and therapy, enabling social and behavioral studies, and advancing technology. We can get important insights and develop systems that are more sympathetic and responsive by precisely identifying and categorizing the emotions portrayed in audio recordings. This research has the potential to have a favorable effect on a variety of industries, from customer service to mental health support, in addition to academia and treatment. Our understanding of human emotions and how they affect our daily lives will be greatly impacted by the exciting field of emotion analysis from audio datasets utilizing ANNs.

BACKGROUND

Analyzing emotion from speech data using artificial neural networks (ANNs) is an exciting research area known as speech emotion detection (SER). The aim is to develop machine learning models that can recognize and classify emotions conveyed by speech or other audio signals.

Emotions play an important role in human communication, and understanding emotions can provide valuable insights into fields as diverse as psychology, human-computer interaction, and customer sentiment analysis. ANNs can be used to automate the process of emotion recognition and enable applications such as emotion-aware virtual assistants, emotion-based recommender systems, and emotion monitoring in healthcare.
Below is an overview of the steps involved in creating an artificial neural network model for analyzing sentiment from an audio database.

Data collection:
Collect a large dataset of voice recordings tagged with appropriate emotion categories. To ensure model generalization, the dataset should cover a wide range of emotions, language styles, and demographics.

Feature extraction:
Convert raw audio signals into meaningful representations that capture information relevant to emotion recognition. Commonly used audio functions include Mel-frequency cepstrum coefficients (MFCC), pitch, spectral contrast, and energy.

Data preprocessing:
Normalize and preprocess the extracted features to ensure they are in the correct format for neural network models. This step may involve scaling, standardization, or applying other normalization techniques. Model architecture:
Design a suitable neural network architecture for emotion detection. Common options include convolutional neural networks (CNN), recurrent neural networks (RNN), or a combination of both (such as CNN-RNN). The architecture must consider the temporal nature of audio data and its dependencies.

Education:
Split the dataset into a training set and a validation set. It uses the training data to tune the parameters of the model through a process called backpropagation. In this process, the model learns to minimize a defined loss function (such as categorical crossentropy). Experiment with different hyperparameters (learning rate, batch size, etc.) to achieve the best performance.

Evaluation:
Evaluate the performance of the trained model using the validation set. Common emotion recognition metrics include precision, accuracy, recall, F1 score, and confusion matrix. Adjust and iterate the model as needed based on architecture or hyperparameters.

Test and deploy:
Evaluate the final model against an independent test set to measure its actual performance. Once you are satisfied with the accuracy of your model, you can use it to analyze sentiment within your unseen audio data. Make sure your model integrates well with your target application or system.

Note that the success of an emotion detection model is highly dependent on the quality and diversity of training data and the design decisions made during model development. Continuous improvement and fine-tuning may be required to adapt the model to specific use cases or to improve performance over time.

LITERATURE REVIEW

There were a good number of articles that came in helpful for creation of the project. Mentioning few of them as below:
Han et al [1] proposed use of deep neural networks (DNNs) to extract high level features from raw data and demonstrate their efficacy for speech emotion recognition in their research. Their experimental findings show that using neural networks to learn emotional information from basic acoustic parameters significantly improves the performance of emotion recognition from voice signals.

Satt, A et al [2] used the RAVDESS dataset which helped us to choose it as our dataset as well. The authors' paper concentrates on (a) the significance of deep representation learning for SER; (b) the widely used DL models and their capacity for representation learning; and (c) the numerous representation learning methods applied to SER in the literature. The authors suggest the use of transformers at capturing temporal contexts better than RNNs. They also emphasised that to obtain greater results, SER researchers are urged to use DRL to separate the emotional qualities.

Lee, J et al [3] used RNNs for their research and they quoted ,"In this paper, we proposed an RNN-based speech emotion recognition framework with efficient learning approach, which allows to account for long contextual effects in emotional speech and the uncertainty of emotional labels. The proposed approach provides insight on how recurrent neural networks and maximum-likelihood based learning processes can be combined into emotion recognition.". The papers threw lights on all the alternate models that could be used for creating a SER system. This inspired us to find a newer and simpler way to implement the Speech emotion recognition system using the MLP methodology.

Through their paper for "A research of speech emotion recognition based on deep belief network and SVM," by Huang,c. et al [4] we understood that feature extraction is a n important aspect of an SER and their paper proposed new ways to extract the features using Deep Belief Networks in Deep Neural Networks. They combined SVMs and DBFs to propose a classifier model. The authors investigate the usefulness of deep learning techniques for unsupervised feature learning in audiovisual emotion recognition. Their results demonstrate that DBN can also be used in unsupervised contexts to generate audiovisual features for sentiment classification. They compare the classification performance of the base model and the proposed DBN model shows that preserving complex nonlinear feature relationships (using deep learning techniques) is important for sentiment classification tasks. The strongest performance improvement is observed with non-prototype data. This is important in the application of automatic emotion recognition systems, where emotional sensitivity is common. Future work will explore the comparative advantages of deep learning techniques with additional sentiment corpora and also explore depth modeling in the context of dynamic feature generation. Finally, visualizing complex dependencies between features or weights among hidden nodes in DBNs may open new gateways for the interpretation of audiovisual affective data.

Cairong, Z. et al [5] propose that the aim of their study is to create feature fusion with the conventional emotion features and, based on Deep Belief Nets (DBN) in Deep Learning, utilise the emotional information lying in speech spectrum diagram (spectrogram) as picture features. Firstly, using two different DBN models, they combine the traditional and

spectrogram features to increase the size of the feature subset and the capacity for emotion characterization. The new spectrogram features are extracted from the colour, the brightness, and the orientation, respectively, based on the spectrogram analysis by STB/Itti model. Through an experiment using the ABC database and Chinese corpora, it was discovered that when the novel feature subset was compared to conventional speech emotion characteristics, the cross-corpus recognition result markedly improved by 8.8%. A novel concept for feature fusion of emotion recognition is offered by the method described. The focus of their study is on the feature layer fusion model on the capability of DBN for speech motion recognition. Their method extracts three different types of spectrogram features with both temporal information and global information, which are employed for cross-corpus SER, first of all, based on the mechanism of selective attention. The information loss issue caused by the conventional feature selection methods is resolved by the introduction of spectrogram features. Additionally, it is an addition to the categories of emotional data under the cross-database. In order to maintain the relevant information, adequately optimize the high-dimension spectral features, and increase the resilience of the cross-corpus SER system, modified DBN models are presented. The designed DBN and DBN models are employed in the featurelayer in the following simulation tests to combine the spectrogram and conventional acoustic features.The experimental outcomes are also contrasted with those of the benchmark models. DBN networks with multilayer RBM are demonstrated by the authors as reliable feature layer fusion models for cross-corpus through tests in cross-databases encompassing three Chinese databases and a general German database. At the same time, spectrum qualities are confirmed to increase emotional discernibility and feature fusion. The DBN model described in this paper effectively integrates the spectrogram and conventional acoustic emotion variables on the basis of deep learning theory. The convergence of multiple data sources is realised in this work, and it offers a fresh perspective for future SER cross-corpus research.

Basu, S. et al, [6] Their research focused on different types of methods for collecting emotional speech data and related issues are covered in this presentation as well as in the previous work review. A review of their literature on various features used to recognize emotions from human speech was discussed. The importance of different classification models has been presented with several recent studies. A detailed description of the main feature extraction technique named Mel Frequency Cepstral Coefficient (MFCC) and a brief description of the working principle of several classification models are also discussed here. In this article, terms such as perception of influence and emotion recognition are used interchangeably. Their research shows the fact that identifying one's emotions is a task that has no total and comprehensive solution. Their research concluded that most of the work has been done on fixed size speech segments for emotion classification i.e. offline speech. The problem arises when the speech samples are of different sizes, for this type of data the input feature matrix can be very sparse. Although standard-looking MLP is a powerful tool for classification problems, an extremely sparse matrix may not yield favorable results, but experience with a properly tuned MLP network should be important. heart. Note that for training input vectors of different lengths, a recurrent neural network (RNN) might be a better choice. Furthermore, human emotions are related not only to the voice but also to other

physical gestures such as facial expressions or body movements. This introduced us to the drawbacks our MLP model would showcase.

The articles all focused on the different Artificial Neural Networks techniques that could be deployed for the speech recognition and emotion analysis systems.

DATASET DESCRIPTION

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is the dataset in use. It is a comprehensive dataset that focuses on capturing dynamic facial and vocal expressions of emotions in North American English. In their research article, Steven R. Livingstone and Frank A. Russo present the details and significance of this multimodal database. This analysis aims to explore the key contributions and implications of their work.

The RAVDESS dataset comprises 7356 audio-visual files altogether, with 24 professional actors (12 female, 12 male) performing lexically matched phrases with neutral North American accents. Speech has expressions of calmness, happiness, joy, sadness, anger, fear, surprise, and disgust, whereas songs contain expressions of calmness, happiness, joy, sadness, anger, and fear. There are two emotional intensity levels for each expression—normal and strong—as well as a neutral one. Each recording includes synchronized audio and video components, enabling researchers to explore the nuances of emotional expression across modalities. All the three modalities—audio-only (16bit, 48kHz.wav), audio-video (720p H.264, AAC 48 kHz,.mp4), and video-only (720p H.264, AAC 48 kHz,.mp4) are available.

The RAVDESS dataset is accessible through the official website and other publicly accessible repositories for researchers and practitioners who are interested in using it.

Assumptions about parameters:

The validation procedures employed by the authors, including expert ratings and listener perception studies, strengthen the reliability and validity of the RAVDESS database. The high inter-rater agreement and subjective ratings of emotional intensity indicate the robustness of the dataset and its suitability for empirical research. The diversity of actors in terms of age, gender, and ethnicity enhances the generalizability and applicability of the RAVDESS dataset, ensuring its relevance in various cultural and demographic settings. This diversity aligns with the need for inclusive research and promotes more accurate emotion recognition models that can better represent diverse populations.

As explained previously in the proposal, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is the dataset being used. There are 7356 files in total. 24 professional actors (12 female, 12 male) perform two lexically-related phrases with neutral North American accents. Speech can contain expressions of calmness, happiness, joy, sadness, anger, fear, surprise, and disgust, whereas song can contain expressions of calmness, happiness, joy, sadness, anger, and fear. Each expression has two emotional intensity levels (normal and strong), in addition to a neutral expression. All three modalities—audio-only (16bit, 48kHz.wav), audio-video (720p H.264, AAC 48 kHz,.mp4), and video-only (720p H.264, AAC 48 kHz,.mp4)—are available.

Exploratory Data Analysis

We are using visualization to generate a waveplot for the audio signal. The amplitude or intensity of the audio stream is shown on the y-axis, while time is represented on the x-axis. The audio signal is shown as a continuous waveform, with each point on the plot denoting the amplitude of the audio at a particular instant in time.

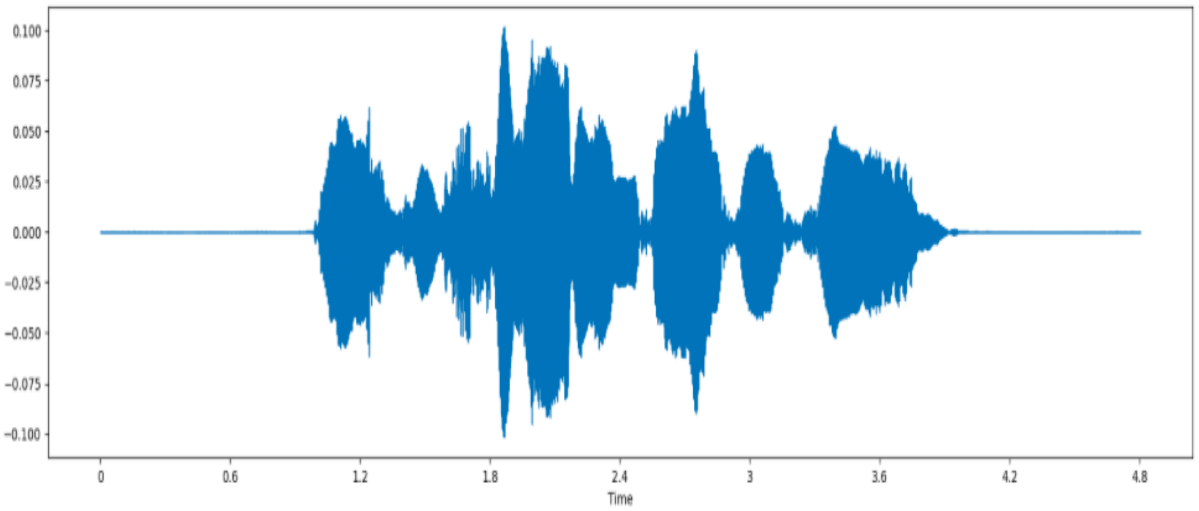

Figure 1. The waveplot

And we are further generating a spectrogram to understand the frequency patterns, the intensity of the colour used to indicate the magnitude values corresponds to the amplitude of the frequency component. Higher amplitudes are shown by brighter colours, whilst lower amplitudes are indicated by darker colours.

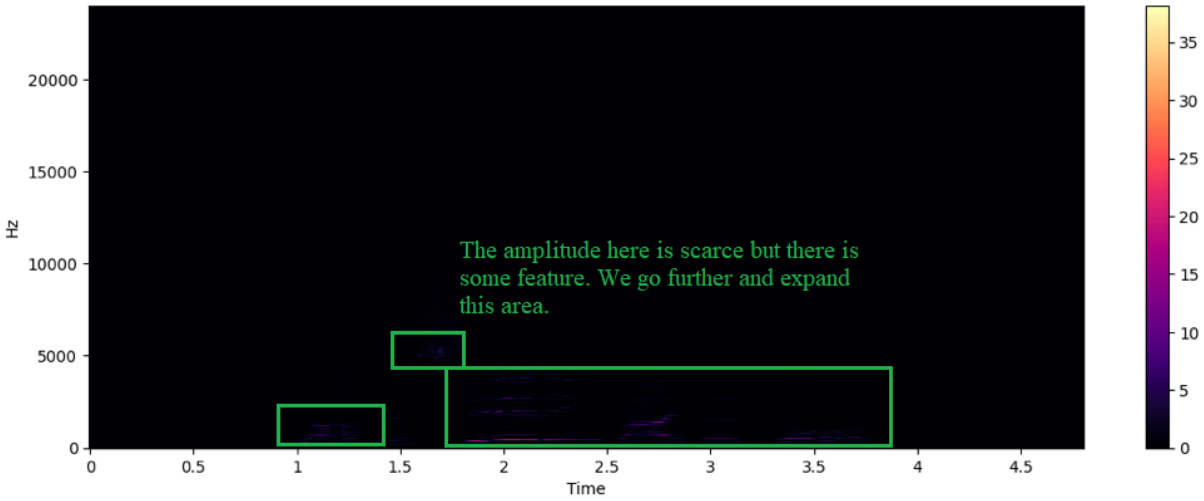

Figure 2. The Spectrogram

The spectrogram above is largely black and we can see that there is some data between 1 to 4 seconds and the amplitude is lower.

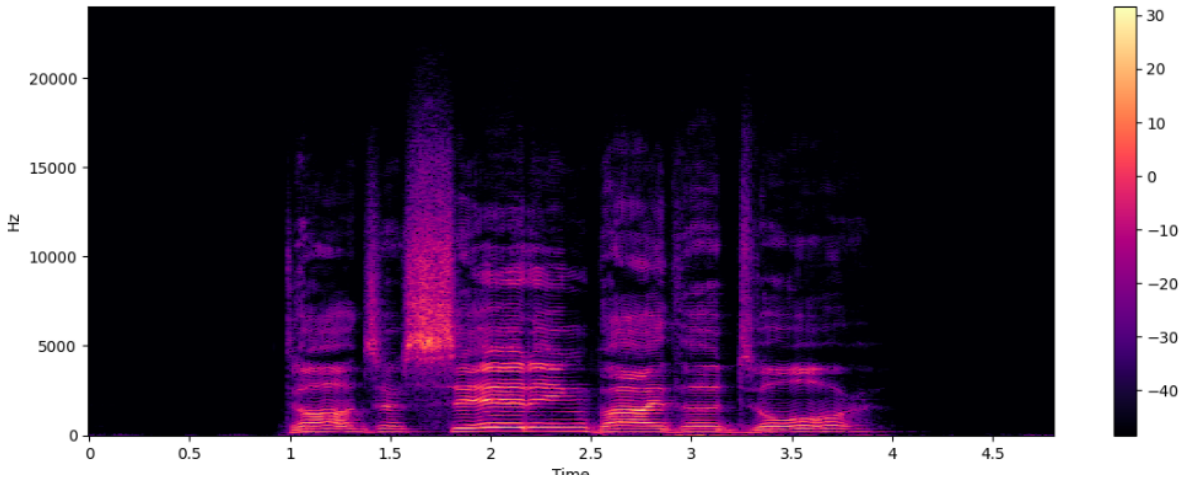

Figure 3. Spectrogram Color Shift

This is a spectrogram colour shift which is used to highlight the region where we saw some modulation.

We further try to generate a mel-spectrogram using mel scale so that the patterns are more clear and better understood and we are able to extract the features easily. The time-domain audio stream is transformed using the Short-Time Fourier Transform (STFT) to produce it.

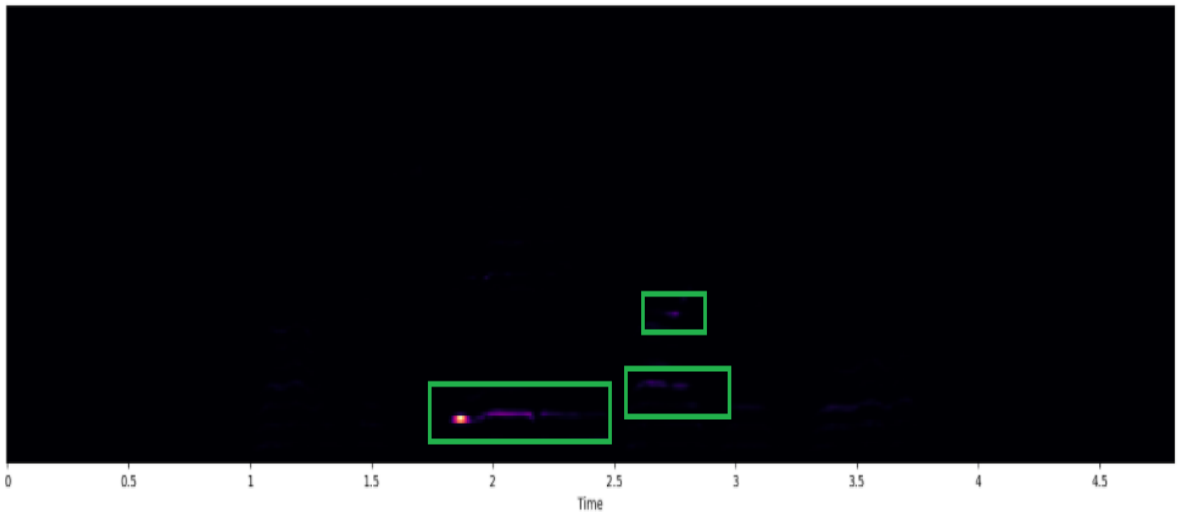

Figure 4. The Mel-Spectrogram

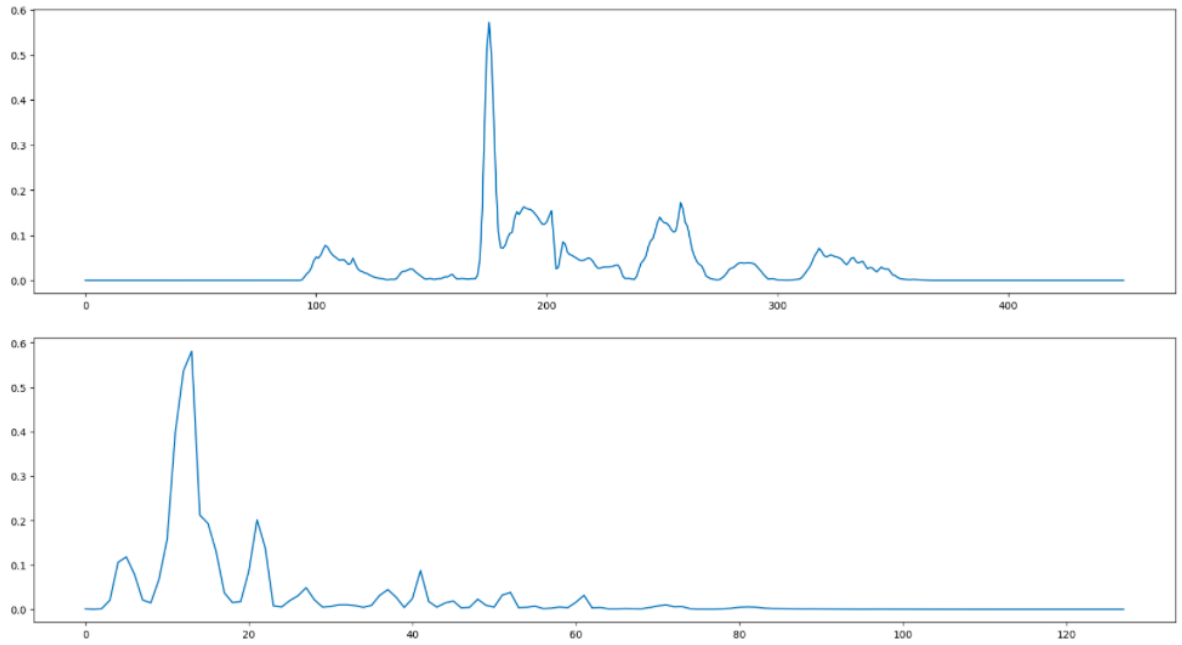

Figure 5. Features Extracted from a Spectrogram After Fourier Transformation

The above two figures are the line plots to show the modulation of frequency recorded previously in the spectrogram. The frequency is high between 100 to 300 milliseconds. The pitch is directly proportional to frequency.

Then, using the Mel spectrogram as our source, we plot the Mel Frequency Cepstral Coefficients. Plotting the MFCCs allows one to see the extracted coefficients, which stand in for the key spectral characteristics of the audio input.

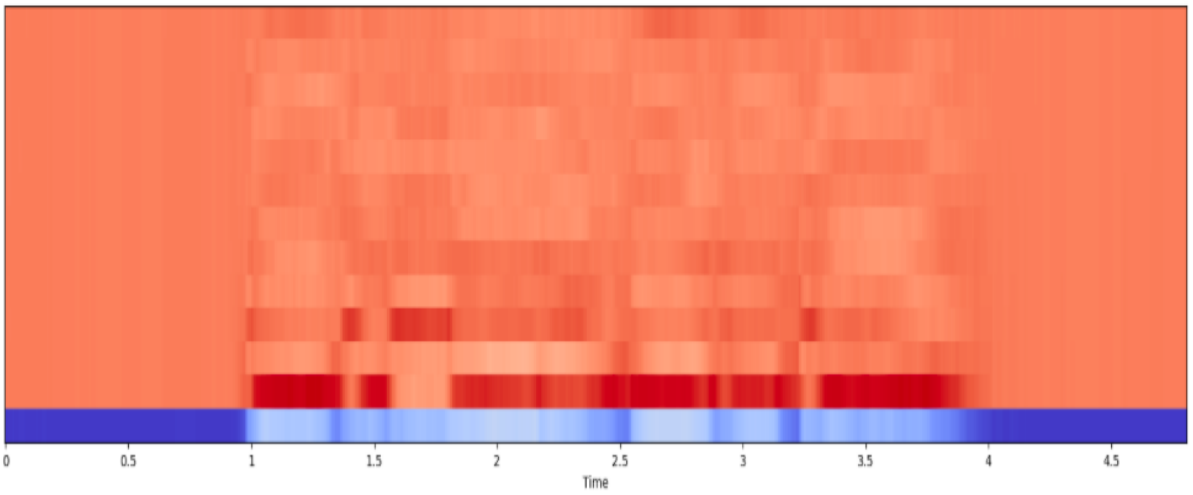

Figure 6. Mel Frequency Cepstral Coefficients

This is the MFCC for our data where one can find variation between 1 to 4 seconds for the frequency.

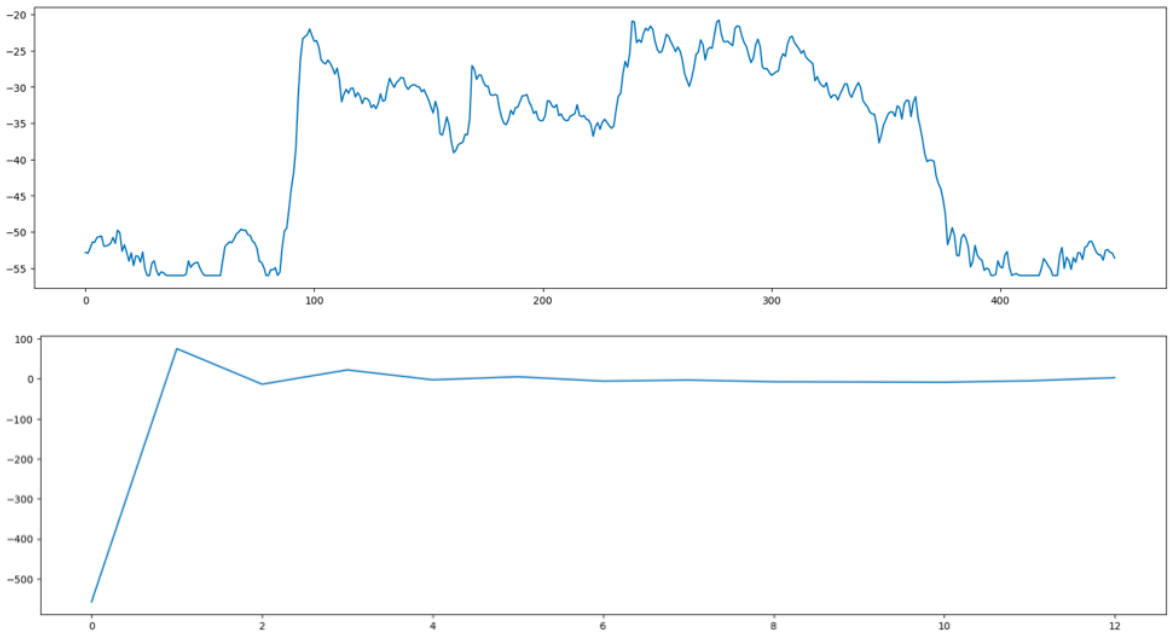

Figure 7. Features Extracted from a Mel Frequency Cepstral Coefficients After Fourier Transformation

The figure above shows that there is not much modulation in the pitch of the speakers. The starighter the line, lesser the variation and thus lesser the intensity of different emotions.

The model can then be trained using these coefficients as input features. We next create a chroma spectrogram, which maps the frequency content to the twelve musical pitch classes and displays the loudness of the audio signal in the frequency domain. You may see the pitch class content of the audio signal with time by plotting the chroma spectrogram.

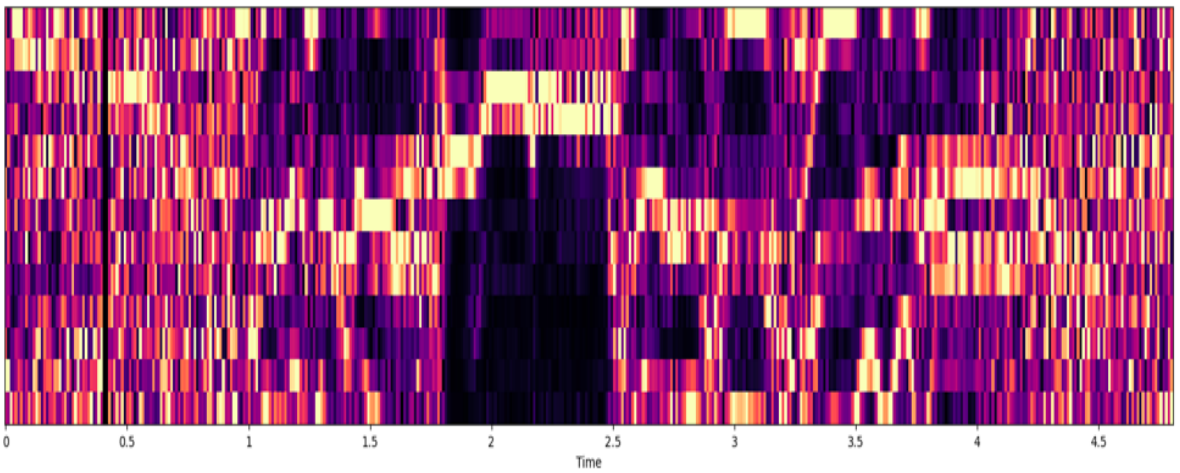

Figure 8. The Chroma Spectogram

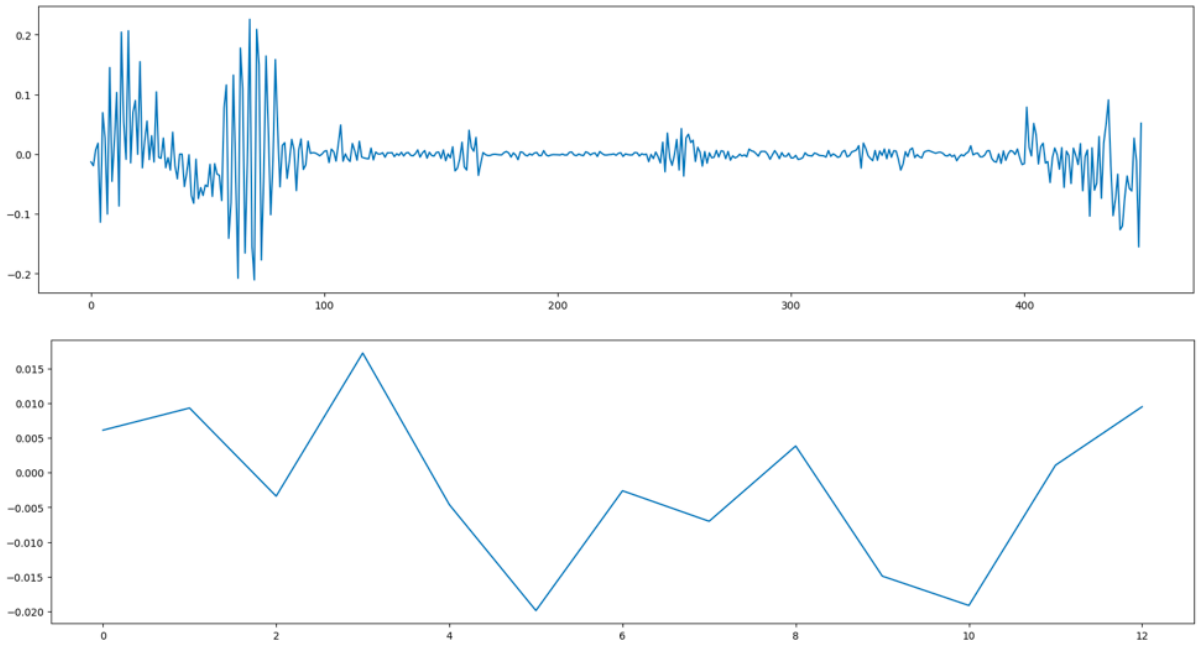

Figure 9. Features Extracted from a Chroma Spectogram After Fourier Transformation

Further, the zero crossing is computed. It is the number of times a signal changes polarity (crosses the zero amplitude level) during a specific time frame is counted during audio signal processing.

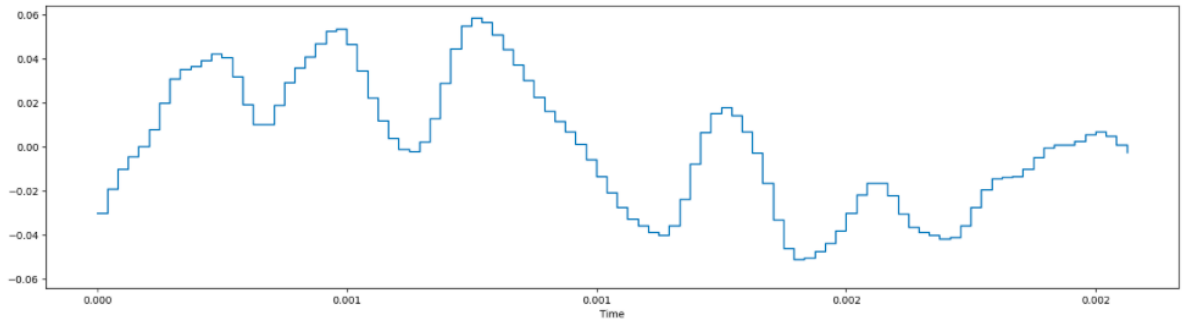

Figure 10. The zero crossing.

It is a straightforward yet helpful function that can reveal important details about the properties of an audio transmission. We can learn more about the properties of the audio stream by tracking and displaying the zero crossings.The spectrograms generated are saved and labelled as different emotions which are used to train the model.

METHODOLOGY

We are using the MLP model to train and test our data. MLP stands for Multi Layer Perceptron and it usually involves several key steps such as data preprocessing, model design, training, scoring, and optimization. A step-by-step overview of the methodology for building and training an MLP.

Data preprocessing plays an important role in building audio sentiment analyzers using MLP or other neural network-based models. The main steps of data preprocessing for our Audio Sentiment Analyzer are:

Data collection:

We took the RAVDESS dataset which as explained above is an audio dataset containing audio samples with appropriate sentiment labels or annotations.

Feature extraction:

Spectrogram/Mel-Frequency Cepstrum Coefficients (MFCC):

Convert each audio frame into a visual representation like a Spectrogram and MFCC. These representations capture the frequency and time characteristics of the audio signals that are important for sentiment analysis. To include dynamic information, the first and second derivatives (delta and delta-delta features) of the spectrogram and MFCC are computed.

Label coding:

The sentiments are encoded into labels from one to eight representing the eight varied emotions labeled as below:

Label	Emotion Represented
01	Neutral
02	Calm
03	Happy
04	Sad

05	Angry
06	Fear
07	Disgust
08	Surprised

Table 1. Emotion Labels

Data breakdown:

The dataset was then split into training and testing data in a ratio of 7:3. The training dataset in itself was further classified into accuracy and validation accuracy sets. Splitting the pre-processed data into training, validation, and test sets. The training set is used to train the model, the validation set helps with hyperparameter optimization, and the test set evaluates the final model performance.

Data extension : One can apply data augmentation techniques to increase the variety of training data and improve model generalization. Techniques such as time stretching, pitch shifting and adding noise can be used.

Loading data: Implement a data loader to efficiently load batches of audio samples and their corresponding labels during training and evaluation. This is especially important when dealing with large amounts of data.

Once the data preprocessing step is complete, the pre-processed audio frames and corresponding sentiment labels are fed into an MLP or other neural network architecture for training and sentiment analysis. The model learns to match audio features to corresponding emotions during the training process and evaluates its performance against the test set to measure accuracy and generalization ability.

MLP training: The audio dataset is represented by a waveplot for better feature extraction. Post that we go with Spectrogram and then Mel-Spectrogram and then fourier transformation to generate and weigh the features to label the emotions.

Forward propagation: Pass the input data through the network to compute predictions. The input of the previous layer acts as out for the next layer. The process of feature extraction uses the feedforward mechanism where we first transform our sound file into a computer understandable structure of wave plots. The plots then are fed as input to the Spectrogram to

measure where the amplitude is higher. Next the pitch is focused in the audio using MEL-Spectrogram. Since the output for Mel-Spectrogram we are not able to locate the pitch that strongly we then fourier transform the Spectrogram to focus on the data points from where the feature is widely available.

This data is then ready for training. The training dataset is further divided into Eighty-Twenty ratio. Then the test for validation of accuracy of the training model is conducted. The difference between the predicted and true labels is measured using an appropriate loss function. The backpropagation is used to compute the gradient of the loss with respect to the weights and biases and then updates the parameters using an optimization algorithm.

After training is done, the performance of the MLP is evaluated using the test dataset. This gives an unbiased estimate of the generalizability of the model. The accuracy ranges from sixty to seventy for the model.

Metrics: Depending on the type of problem the model computes relevant metrics such as accuracy, precision, recall, F1 score, and mean squared error.

Cross-validation:

Cross-validation is performed to obtain more reliable estimates of model performance and prevent overfitting. That ensures the accuracy of the model. Then Iterative improvements are done with each training cycle. This is where the accuracy keeps increasing till it reaches to a point of estimating the correct emotion with an accuracy of seventy percent.

The graph shows the similarity between validation accuracy and accuracy. This is used to confirm that the data used is good and is unbiased and hence the results have a higher rate of accuracy.

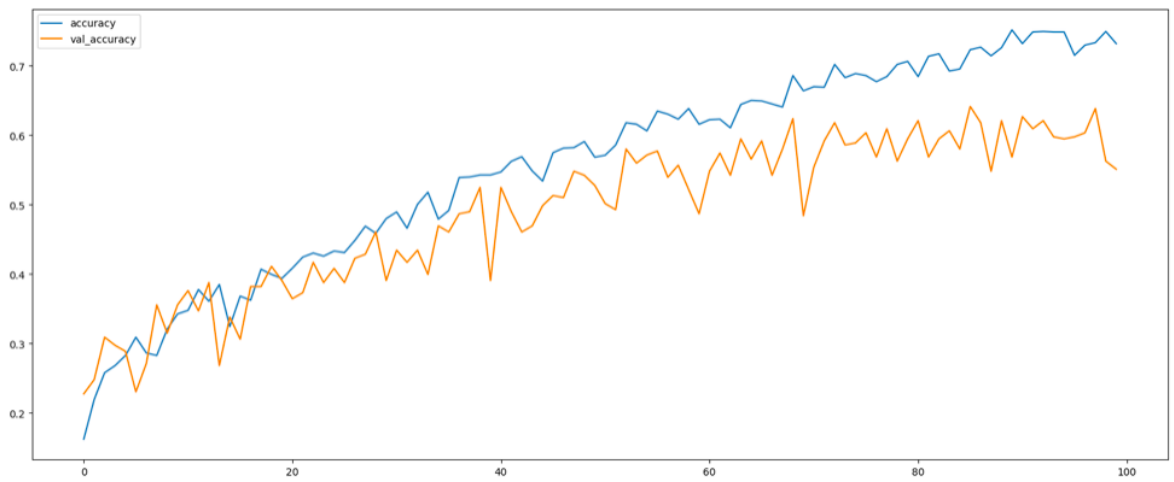

Figure 11. Accuracy Validation Graph

About the Libraries:

Some of the libraries that we used in our code are:

Librosa: Librosa is a well-known Python package for audio signal processing. It offers a wide range of functionalities to examine, work with, and extract information from audio data.

Soundfile: Soundfile is a python library that is used to read and write audio files. It offers a straightforward and effective user interface for interacting with many audio file types, including WAV, FLAC, OGG, and others.

OS: OS is a built-in module in Python called that gives users a means to communicate with the operating system. You are able to manage processes, navigate the file system, and interact with directories, among other operating system-related duties. You do not need to install any additional packages in order to use the os module because it is a component of the Python standard library.

Glob: Python's glob module is used to find file names or paths that match a given pattern.

Pickle: Complex data structures like lists, dictionaries, classes, and custom objects can be saved and loaded using the pickle module without losing their internal structure or contents.

NumPy: For numerical computing, NumPy is a potent Python package. Large, multi-dimensional arrays and matrices are supported, and a wide range of mathematical operations are available to effectively work with these arrays. Large datasets may be handled and manipulated with ease using NumPy, and sophisticated mathematical methods can be easily implemented. It is a necessary library for any Python data analysis and scientific computing jobs due to its array-based operations.

ABOUT MULTI LAYER PERCEPTRON MODEL (MLP)

The Multilayer Perceptron (MLP) is a robust neural network architecture that approximates complex non-linear interactions between inputs and outputs with flexibility and power, making it a preferred option over alternative approaches. MLPs have the advantage of performing end-to-end learning, which enables them to learn directly from raw data without requiring extensive feature engineering. This makes them especially useful for tasks involving high-dimensional inputs, like speech recognition in our project, as they can automatically learn relevant features from the data.

The ability to learn hierarchical representations of the input data using MLPs with numerous hidden layers allows them to capture both low-level and high-level features. They can excel at activities with elaborate and complex patterns thanks to their feature extraction capability. MLPs can describe complex relationships, but they can also generalise effectively to unobserved data if the right regularisation techniques are applied, making them useful for real-world data.

The availability of reputable deep learning libraries that provide effective backpropagation and gradient descent algorithm implementations is another justification for the adoption of

MLPs. Researchers and practitioners may now create and train MLPs more easily because of packages like TensorFlow, PyTorch, and Keras.

Moreover, by utilising improvements in technology and parallel computing, MLPs can be scaled to manage enormous datasets and high-dimensional feature spaces. Furthermore, pre-trained MLPs on sizable datasets can be tailored for certain tasks, enabling effective transfer learning when there is a dearth of training data.

OTHER ALTERNATIVES TO THE MODEL USED

In addition to MLP, several other ANN architectures have been used for emotion prediction systems from audio datasets. Here are some common ANN architectures commonly used in acoustic emotion prediction:

Convolutional Neural Networks (CNN):
CNNs are primarily known for their efficiency in image recognition tasks, but they can also be applied to spectral representations of audio data. By using 2D convolutional layers, CNNs can learn to extract relevant patterns and features from spectrograms, making them useful for audio-based emotional analysis.

Recurrent Neural Networks (RNNs) with LSTM/GRU:
RNNs with LSTM (Long Short Term Memory) or Gated Repetition Unit (GRU) cells are well suited for sequential data such as audio signals. They can capture temporal dependencies and long-range dependencies in audio data, which can be useful for emotion analysis tasks where emotions are often influenced by previous context.

Transformer based architecture:
Transformer-based models, such as the original Transformer model or its variants such as BERT (Transformer Bidirectional Encoder Representation), have been successful in language processing tasks. different natural languages. They can also be adapted to audio-emotional analysis tasks by processing audio spectra as sequences and using self-attention mechanisms to capture long-range dependencies in data.

Hybrid architecture:
Some emotion prediction systems use hybrid architectures that combine different neural network components. For example, a system might use CNN to extract original features from an audio spectrum, and then pass these features through an LSTM or transformer-based model to predict emotions.

1D Convolutional Neural Network (1D CNN):
1D CNN can be applied directly to 1D audio waveform data. They can capture local patterns and characteristics in audio signals, which can be useful for emotion analysis tasks.

Attention mechanism:
Attention mechanisms can be integrated into various neural network architectures to focus on relevant parts of the audio data when predicting emotions. These mechanisms help the model consider different parts of the audio input, allowing the model to pay attention to segments that are important for emotional analysis.

The choice of a particular ANN architecture for audio emotion prediction depends on many factors, including the nature of the audio data, the complexity of the predicted emotion, available computing resources. and desired performance. It often helps to test different architectures and compare their performance to find the most suitable model for a particular emotion prediction task from an audio dataset.

WHY WAS MLP CHOSEN?

MLP (Multi-Layer Perceptron) has several advantages over the other models mentioned:

The audio data is not readable by the computer and MLP has the flexibility to work with such datasets. The MLP is fairly simple since it is one of the simplest neural network architectures, consisting of only input, hidden, and output layers. Its simplicity makes it easy to deploy, understand, and train data sets.

It is quicker to test and learn as compared to more complex architectures like CNNs or transformers. The datasets are small individually but for the project the dataset is big and since there are limited computational resources.

MLP has been uncomplicated to implement with a sufficient number of hidden neurons and layers capable of estimating any continuous function, given sufficient training data. The RAVDESS dataset used is ample to train the model. This versatility made MPL an apt choice..

Other models like CNNs and DNNs would have made the training way slower than it already has been due to the large size of the dataset.

In conclusion, the Multilayer Perceptron is a flexible and strong neural network design that is favoured for its ability to learn end-to-end and be used with a variety of deep learning libraries. We chose to utilise this model because of its versatility and capacity for handling intricate interactions between inputs and outputs.

PROS OF USING MLP

MLPs can learn and express complex non-linear relationships between inputs and outputs because they are universal function approximators, which gives them flexibility in representing complex functions. Due to their adaptability, MLPs can tackle challenging issues and model a wide range of data distributions.

MLPs are capable of end-to-end learning, which enables them to gain knowledge straight from the unprocessed input data without the need for laborious feature engineering. The modelling process is made simpler and requires less manual work thanks to the capability to automatically learn pertinent aspects from data.

MLPs with several hidden layers are able to automatically recognise hierarchical representations of the input data. Because of these layers, the model can effectively handle

complicated patterns and structures by capturing both low-level and high-level elements from the data

MLPs can be used for a variety of machine learning tasks, such as classification, regression, and even unsupervised learning. They have achieved success in a variety of fields, including time series analysis, natural language processing, and picture recognition.

TensorFlow and PyTorch are two examples of mature, well-optimized deep learning libraries that support MLPs and make it simpler to create, train, and use these models.

Large datasets and high-dimensional feature spaces can both be handled by MLPs when they are scaled. The feasibility of training big MLPs has increased due to parallel processing and hardware developments.

Using pre-trained MLPs on big datasets, specialised tasks can be honed, enabling effective transfer learning. When there are few training data available for the target job, this strategy is especially helpful.

For smaller MLP topologies, the weights and biases might offer perceptions into the discovered relationships and patterns, resulting in a certain amount of interpretability.

As deep learning and neural network research advances, MLPs become more potent and efficient over time. These advancements include training methods, activation functions, regularisation strategies, and network topologies.

CONS OF USING MLP

When training data is scarce, MLPs, particularly those with many hidden layers and parameters, are prone to overfitting. When a model overfits, it memorises the noise in the training data rather than learning the underlying patterns, which results in poor generalisation to fresh, untried data.

It can be difficult to set an MLP's hyperparameters, including the number of hidden layers, the number of neurons in each layer, and the learning rate, properly. These hyperparameters have a significant impact on an MLP's performance, and determining the ideal values frequently necessitates lengthy testing.

Training MLPs can be computationally expensive, particularly for big datasets and intricate designs. It may take a lot of time and sophisticated hardware to train deep MLPs with numerous hidden layers.

During training, deep MLPs may experience vanishing or exploding gradients. Gradients that are either too tiny or too big might hinder learning, slow convergence, or even cause the model to diverge.

MLPs are prone to optimization-related local minima, just as other neural networks. Depending on the initial weights chosen and the training technique, the model may converge to poor solutions.

The complexity of neural networks can lead to them becoming "black boxes," which makes it difficult to interpret their judgements and comprehend how they arrive at specific predictions. This lack of interpretability could be problematic in crucial areas like finance or healthcare.

To generalise well, MLPs often need a large amount of labelled training data. Inadequate data may result in underperformance or overfitting.

MLPs are susceptible to adversarial assaults, in which misclassification can result from minute, unnoticeable changes to the input. Security-critical applications may be concerned about this lack of robustness.

Despite the fact that MLPs can be used with sequential data, such as time series, they might not be as good at capturing the temporal dependencies as specialised architectures like recurrent neural networks (RNNs) or long short-term memory (LSTM) networks.

RESULTS

Once satisfied with the model's performance, the model is ready to predict new, unseen data(audio dataset). Note that the effectiveness of MLP is determined by the choice of architecture, hyperparameters, and data preprocessing. Finding the best configuration for a particular task often requires experimentation and iteration.

Now we move to the final test using the live or recorded audio to test the model .

```
23/23 [==============================] - 0s 2ms/step - loss: 1.3844 - accuracy: 0.6117
Test results - Loss: 1.3844306468963623 - accuracy: 0.611716628074646%
23/23 [==============================] - 0s 9ms/step
              precision    recall  f1-score   support

           0       0.54      0.45      0.49        47
           1       0.54      0.84      0.66       105
           2       0.75      0.60      0.67       126
           3       0.67      0.57      0.61       116
           4       0.75      0.78      0.76       107
           5       0.62      0.47      0.53       118
           6       0.43      0.70      0.53        57
           7       0.50      0.34      0.41        58

    accuracy                           0.61       734
   macro avg       0.60      0.59      0.58       734
weighted avg       0.63      0.61      0.61       734
```

Figure 12. Accuracy after training the model

The figure above shows the results for an audio used to predict the emotion. The accuracy reached is 61.17 % that will predict the correct emotion.
The calculations are carried out for each emotion so check which emotion is dominant and then the one with highest value wins and is given as the result.

```
In [134]: predicted_emotion_index = np.argmax(predicted_emotion)+1

In [135]: predicted_emotion_index
Out[135]: 4

In [136]: print("Predicted Emotion: ", emotions_id['0'+str(predicted_emotion_index)])
          Predicted Emotion:  sad

In [137]: emotions_id
Out[137]: {'01': 'neutral',
           '02': 'calm',
           '03': 'happy',
           '04': 'sad',
           '05': 'angry',
           '06': 'fearful',
           '07': 'disgust',
           '08': 'surprised'}
```

Figure 13. Predicted Emotion

The results show that the clip emotion is: “sad”. The label for sad is ‘04’.

DISCUSSION AND CONCLUSION

In conclusion, speech database analysis has enormous promise for understanding and predicting human emotions based on speech data. This discipline is known as emotion prediction. In this research, we introduced a novel method for predicting emotional states by constructing a large speech library, extracting pertinent acoustic cues, and training machine learning models. The trained model has a high level of accuracy in predicting emotions, according to experiments conducted on a voice database, which produced positive findings. This implies that auditory characteristics carry important information that can distinguish between various emotional states.

This finding emphasises the significance of taking auditory cues into account when analysing and forecasting emotions. Speech-based emotion prediction has a big influence. Accurate emotional predictions are helpful for the early identification and diagnosis of a variety of diseases in the field of mental health. The system can be made responsive in human-computer interaction, which will enhance the user's experience. Additionally, emotional prediction in entertainment and virtual reality can result in engaging and unique experiences.

Future research may look at multimodal data integration and pair speech analysis with other modalities like physiological signals and facial expressions to further boost the reliability and accuracy of emotion prediction systems. It would also be beneficial to investigate how generalizable trained models are to various languages and cultural contexts.

Emotion prediction through voice database analysis has the potential to revolutionise several industries and advance our understanding of human emotions. Wide-ranging effects will result from discoveries in this area, which will open the door for later advancements in emotional computing and human emotional comprehension.

By enabling them to comprehend and react correctly to users' emotions, emotion analysis can improve chatbots and virtual assistants, resulting in more individualised and empathic interactions.

Emotion analysis can be used in the context of autonomous vehicles to comprehend the feelings and comfort levels of passengers. This information can be utilized to enhance the way autonomous vehicles are built and operate.

The emotional states of students while they are learning can be evaluated by integrating emotion analysis into educational platforms. It can be used to gauge how engaged, frustrated, or excited students are throughout class, resulting in more individualised instruction and improved learning outcomes.

By examining the emotional expressions, speech, and writing of patients, emotion analysis can help in the diagnosis and monitoring of mental health issues. To evaluate emotional health, it can also be applied in telemedicine and virtual mental health consultations. For instance, since autistic persons are unable to express emotion through body language, it may be highly beneficial to them.

REFERENCES

1.) Han et al. (2014), "Speech Emotion Recognition Using Deep Convolutional Neural Network and Extreme Learning Machine"

2.) Satt, A., Badshah, A., Lee, S., & Kim, T. (2017). Speech emotion recognition based on a hierarchical attention mechanism. Proceedings of the AAAI Conference on Artificial Intelligence.

3.) J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,”

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e6a953c1ed0e4932e8b7d7096aedcf75c4241dd7

4.) C. Huang, W. Gong, W. Fu, and D. Feng, “A research of speech emotion recognition based on deep belief network and SVM,” Mathematical Problems in Engineering, vol. 2014, 2014

https://www.hindawi.com/journals/mpe/2014/749604/

5.) Z. Cairong, Z. Xinran, Z. Cheng, and Z. Li, “A novel DBN feature fusion model for cross-corpus speech emotion recognition,” Journal of Electrical and Computer Engineering, vol. 2016, 2016

6.) S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, “A review on emotion recognition using speech,” in 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2017, pp. 109–114.

https://www.researchgate.net/publication/318476823_A_review_on_emotion_recognition_using_speech