Speech Emotion Recognition

by 181579 Karim Hatem Hamed Abdelsalam

Submission date: 17-Feb-2022 08:28PM (UTC+0200)

Submission ID: 1764780725

File name: 167914_181579_Karim_Hatem_Hamed_Abdelsalam_Speech_Emotion_Recognition_231070_730923568.docx (1.98M)

Word count: 6094

Character count: 33507

OCTOBER UNIVERSITY FOR MODERN SCIENCES AND ART

Faculty of Computer Science

Graduation Project

Speech Emotion Recognition

Supervisor: Dr. Islam El-Shaarawy

Name: Karim Hatem Hamed

ID: 181579

Table of Contents

Abstract II
Chapter 1: Introduction 1
1.1 Introduction 2
1.2 Problem statement 3
1.3 Objective 3
1.4 Motivation 3
1.5 Thesis layout 3
Chapter 2: Background and Literature Review 5
2.1 Background 6
2.1.1 Machine Learning 6
2.1.2 Neural Network and Deep Learning 7
2.1.3 Long Short-term memory 7
2.2 Previous Work 7
2.2.1 Research 1 7
Chapter 3: Material and Methods 9
3.1 Materials 10
3.1.1 Data 10
3.1.2 Tools 11
3.1.3 Environment 11
3.2 Methods 11
3.2.1 System architecture Overview 11
Chapter 4: System Implementation 13
4.1 System Development 14
4.2 System Structure 15
4.2.1 System Overview 15
4.2.2 Class Diagram or Tensor Board 16
4.3 System Running 19
4.3.1 Component A 19
Chapter 5: Results and Evaluation 21
5.1 Testing Methodology 22
5.2 Results 22

5.2.1 Best Results Cases 23
5.2.2 Limitations 24
5.3 Evaluation 24
5.3.1 Accuracy Evaluation 24
5.3.2 Time Performance 26
Chapter 6: Conclusion and Future Work 27
6.1 Conclusion 28
6.2 Problem Issues 28
6.2.1 Technical issues: 28
6.2.2 Scientific issues: 29
6.3 Future Work 29
References 30

Table of Figures

Figure 1: Emotion detection result on sample images ……3
Figure 2: Annotation interface of objects in the image ……...6
Figure 3: Wave Dataset samples……9
Figure 4: Chromogram Dataset samples 10
Figure 4: LSTM system architecture overview.....12
Figure 6: AudioSigment Audio Sample……14
Figure 7: Normalized Audio Sample……14
Figure 8: Trimmed Audio Sample……15
Figure 9: Padded Audio Sample……15
Figure 10: Reduced silence Audio Sample……...16
Figure 11: Classification System architecture overview 17
Figure 12: Graph Example 20
Figure 13: System Running 20
Figure 14: Class Diagram Example 22
Figure 15: Random Results from the Best DNN Classification Model 23
Figure 16: Graphs of the best DNN model loss and Accuracy 27
Figure 17: Graph of the CNN results of different Activation Functions 28

Table of Tables

Table 1: First Summary of the DNN Classification Models with their Accuracy 24

Abstract

Theoretical definitions, categorizations, and modalities of emotion manifestation are all discussed. To carry out this research, a SER system built on various classifiers and feature extraction algorithms has been developed. The voice signals are analyzed for Mel-frequency cestrum coefficients (MFCC) and modulation spectral (MS) characteristics, which are then used to train several classifiers. In order to find the most relevant feature subset, feature selection (FS) was used. For the emotion categorization challenge, a variety of machine learning methods were used. Seven emotions are initially classified using a recurrent neural network (RNN) classifier. Their results are then compared to approaches such as multivariate linear regression (MLR) and support vector machines (SVM), which are commonly employed in the field of emotion recognition for spoken audio sources. The experimental data collection is made up of datasets from Berlin and Spain. When speaker normalization (SN) and feature selection are applied to the features in the Berlin database, all classifiers obtain an accuracy of 83 percent. The RNN classifier without SN and with FS achieves the highest accuracy (94%) in the Spanish database

Chapter 1: Introduction

1.1 Introduction

Speech is the most natural way for humans to communicate. Extending the same communication medium to computer applications is therefore just common sense. SER systems are composed of methodologies that process and classify speech signals to identify the emotions contained within them. extract emotional features from speech signals by computer and analyses the features and emotional change with contrast and contrast. According to the final law, speech and emotion are subject to scrutiny, and emotional states are judged under the final law. Speech emotion recognition was a technology that extracted emotional features from speech signals using a computer and then contrasted and analyzed the characteristic parameters and emotional change obtained. Finally, the law of speech and emotion was established, and speech emotional states were adjudicated in accordance with the law. Speech emotion detection is currently a growing cross-field of artificial intelligence and artificial psychology, as well as a hot research area in signal processing and pattern identification. The research has been widely implemented in sectors like as human-computer interaction, interactive teaching, entertainment, and security.

The voice emotion processing and recognition system was divided into three parts: speech signal acquisition, feature extraction, and emotion recognition. Framework for the system in this system, the accuracy of speech emotion recognition was directly affected by the quality of feature extraction. In the feature extraction process, the entire emotion sentence was usually used as a unit for feature extraction, and the extraction contents were four aspects of emotion speech, which were several acoustic characteristics of time construction, amplitude construction, fundamental frequency construction, and formant construction. Then, using these four elements, compare emotion speech to no emotion sentence, learning the law of emotional signal distribution, and classifying emotion speech according to the law. Deep

neural networks (DNN) have achieved extraordinary success in the fields of voice recognition and picture recognition [3], but no research on deep neural networks has been conducted in the subject of speech emotion processing. We discovered that the deep belief network (DBN) of DNN has a significant advantage in voice emotion processing [4]. As a result, this research developed a method for automatically extracting emotional aspects from sentences. It trained a 5-layer-deep network with DBNs to extract voice emotion features. It combines the emotion features of more consecutive frames of speech to create a high latitude characteristic, and it employs an SVM classifier to identify the emotional speech. We compared this method to other classic feature extraction methods and determined that the rate of voice emotion recognition reached 86.5%, which was 7% higher than the original method. Speech can represent emotion because it contains characteristic qualities that can reflect emotion information. We can extract and examine the change of distinctive parameters to measure the emotional changes in speech. The essential point above is obtaining speech emotion characteristic characteristics from voice data. The accuracy of speech emotion recognition is directly affected by the quality of feature extraction. Meanwhile, speech signals contain not just emotional feature information but also crucial information about the speaker, so study on how to extract and which speech emotion characteristic parameters to extract is critical. Traditional emotional feature extraction relied on the examination and comparison of a wide range of emotion characteristic parameters, with emotional characteristics with high emotional resolution being chosen for feature extraction. Traditional emotional feature extraction, in general, focuses on the analysis of emotional aspects in speech based on time construction, amplitude construction, fundamental frequency construction, and signal feature extraction.

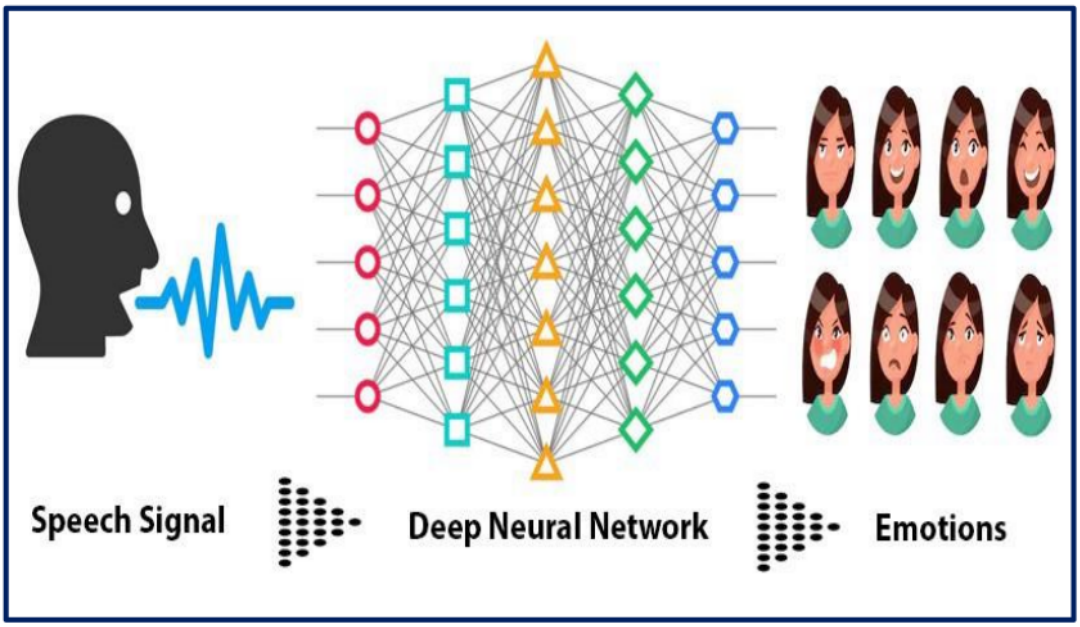


Figure 1: Emotion detection *result on sample images*

1.2 Problem statement

It is visible in call centers. If you've ever noticed, call center employees never speak the same way; their approach to pitching/speaking to customers varies depending on the customer. Now, this occurs with ordinary people as well; employees recognize customers' emotions through their speech, allowing them to improve their service and convert more customers. They are utilizing speech emotion recognition in this manner.

1.3 Objective

SER is a critical component of Human-Computer Interaction systems that is widely used in a variety of sectors including healthcare, robotics, automated call centers, and distance education. Speech emotion recognition entails analyzing the signal in detail and identifying the appropriate emotion using extracted features from a trained database. Method/Statistics.

1.4 Motivation

The main motivation for me is to improve my ability to use programming in facilitating and developing the process of customer services through a program that detects customer's satisfaction level from each call because normal surveys won't be as accurate as my program, It'll make any company have a clearer vision of how their customers actually feel and also due to the rapid growth in the field of automation and the abundance of customer service oriented jobs and applications , automating the response of customer service application is an ever growing demand.

1.5 Thesis layout

A lot of study has gone into Automatic Speech Emotion Recognition in the last decade (SER). The main goal of SER is to improve the human-machine interface. It can also be utilized in lie detectors to monitor a person's psychophysiological condition. Speech emotion detection has recently found applications in health and forensics. In this work, pitch and prosody traits are used to distinguish seven emotions. The majority of the speech features examined in this study are temporal in nature.

Chapter 2: Background and Literature Review

2.1 Background

Emotion detection is a critical marketing strategy in today's world. You could personalize various items for an individual based on their interests. As a result, we decided to work on a project that would allow us to detect a person's emotions simply by listening to their voice, allowing us to manage a variety of AI-related applications. Several examples include requiring call centers to play music when a customer is irate during a call. Another possibility is that a smart car will slow down when a driver is angry or fearful. As a result, this type of application has a lot of potential in the world, both for businesses and for consumer safety.

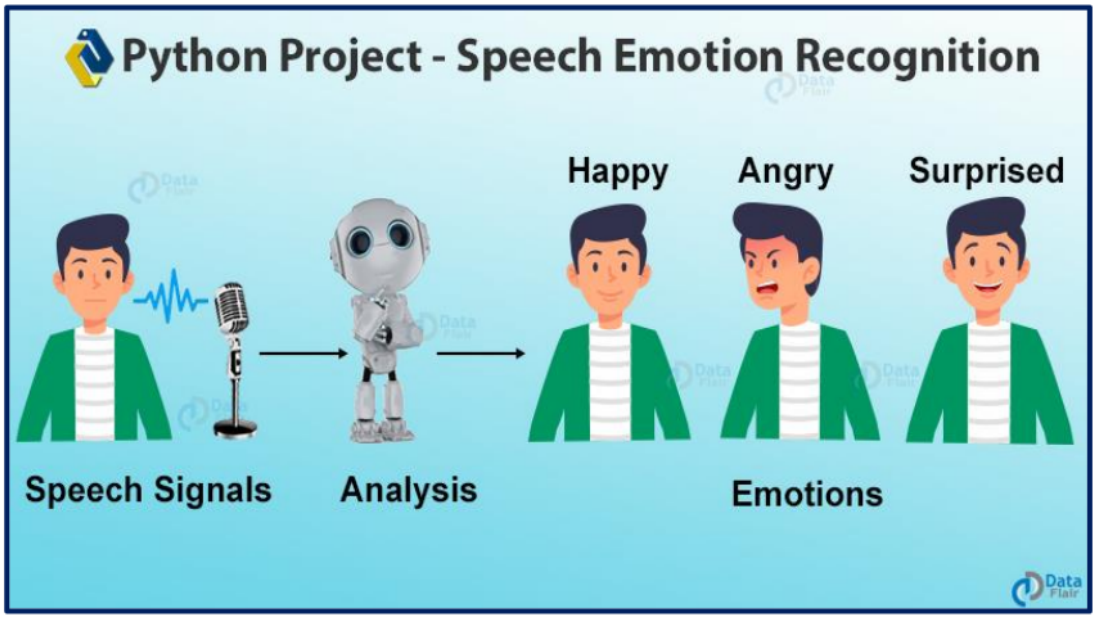


Figure 2: Annotation interface of objects in the image

2.1.1 Machine Learning

M Machine learning is a fundamental tool for employing artificial intelligence (AI) technology. Machine learning is frequently referred to as AI because of its learning and decision-making capabilities, although it is actually a subset of AI. It was a part of AI's evolution until the late 1970s. Then it split out and began to evolve on its own. Machine learning has emerged as a critical response tool for cloud computing and eCommerce, as well as a number of other cutting-edge technologies. For many firms today, machine learning is an essential part of modern business and research. It makes use of algorithms and neural network models to help computers improve their performance over time. Without being specifically taught to make those decisions, machine learning algorithms automatically develop a mathematical model utilizing sample data – often known as "training data" – to make decisions. Machine learning is based on a model of brain cell interaction to some extent. Donald Hebb developed the concept in his book The Organization of Behavior, published in 1949. (PDF). Hebb's theories on neuron stimulation and communication between neurons are presented in this book. “When one cell assists another in firing, the first cell's axon generates synaptic knobs (or enlarges them if they already exist) in contact with the soma of the second cell," Hebb noted. When applied to artificial neural networks and artificial neurons, Hebb's model can be described as a method of changing the interactions between artificial neurons (also known as nodes) and the changes to individual neurons. When two neurons/nodes are triggered at the same moment, their link strengthens; when they are activated separately, it weakens. These connections are referred to as "weight" relationships. Strong positive weights are assigned to nodes/neurons that tend to be both positive and negative. Those nodes with opposite weights acquire strong negative weights (for example, 11=1, -1x-1=1, -1x-1=1, -1x-1=-1). In the 1950s, IBM's Arthur Samuel created a computer software for playing checkers. Samuel launched alpha-beta pruning since the application only had a minimal amount of computer memory available. A scoring function based on the locations of the pieces on the board was added in his design. The scoring function attempted to calculate each team's chances of winning. A minimax strategy, which subsequently evolved into the minimax algorithm, is used by the software to determine its next move. Samuel also devised a variety of strategies to help his

programmer improve. In what Samuel referred to as rote learning, his programmer recorded/remembered all previous places and combined them with the reward function's values. The term "machine learning" was coined by Arthur Samuel in 1952. The perceptron was invented in 1957 by Frank Rosenblatt at the Cornell Aeronautical Laboratory, who coupled Donald Hebb's idea of brain cell interaction with Arthur Samuel's machine learning techniques. The perceptron was designed as a mechanism rather than a programmer at first. The software was put on a custom-made machine called the Mark 1 perceptron, which was intended for picture recognition and was initially planned for the IBM 704. As a result, the software and algorithms might be transferred and used by other devices.

The Mark I perceptron, dubbed the "first successful neuro-computer," had some issues with expectations that were not met. Although the perceptron appeared promising, it was unable to distinguish a wide range of visual patterns (including faces), leading to frustration and a halt in neural network research. The frustrations of investors and financing agencies would last for several years. The field of neural networks and machine learning stagnated until the 1990s, when it saw a rebirth. The nearest neighbor method, which was the start of rudimentary pattern recognition, was created in 1967. This technique was used to map routes and was one of the first algorithms to solve the challenge of identifying the most efficient path for travelling salespeople. A salesman uses it to enter a desired city and have the programmer tour nearby cities until all have been visited. The "nearest neighbor rule" is said to have been invented by Marcello Pelillo. He, in turn, gives credit to the 1967 Cover and Hart paper (PDF).

Artificial intelligence research in the late 1970s and early 1980s concentrated on logical, knowledge-based approaches rather than algorithms. Furthermore, computer science and AI experts have abandoned neural network research. Artificial intelligence and machine learning developed a schism as a result of this. Machine learning had previously been employed as an AI training programmer. The machine learning

industry, which employed many researchers and technicians, was separated into its own field and struggled for nearly a decade. The industry's emphasis has switched from artificial intelligence training to solving actual difficulties in service delivery. Its focus switched away from AI-inspired ideas and toward probability theory and statistics methodologies and tactics. The ML industry remained focused on neural networks at this time, and then boomed in the 1990s. The majority of this success was due to Internet expansion, which benefited from the ever-increasing availability of digital data and the ability to share its services via the Internet.

-Supervised learning: -

Supervised learning, often known as supervised machine learning, is an artificial intelligence and machine learning subcategory. Its use of labelled datasets to train algorithms that accurately classify data or predict outcomes defines it. As input data is fed into the model, the weights are adjusted until the model is properly fitted, which happens during the cross-validation phase. Organizations can use supervised learning to tackle a range of real-world problems at scale, such as spam classification in a distinct folder from your email. A training set is used in supervised learning to teach models to produce the desired output. This training dataset contains both correct and incorrect outputs, allowing the model to improve over time. The loss function is used to assess the algorithm's correctness, and it is adjusted until the error is suitably minimized. An algorithm is used to classify test data and allocate it to certain groups. It recognizes certain entities in the dataset and makes educated guesses about how those entities should be labelled or defined. Linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest are some of the most common classification algorithms. To explore the relationship between dependent and independent variables, regression is used. It's widely used to produce forecasts, such as for a company's sales revenue. Popular regression algorithms include linear regression, logistical regression, and polynomial regression. Linear regression is a statistical technique for determining the relationship between a dependent variable and one or more independent variables, and it is commonly used to forecast future results. Simple linear regression is used when there is only one independent variable and one dependent variable. Multiple linear regression is used when the number of

independent variables is increased. It aims to plot a line of greatest fit, which is derived using the least squares method, for each type of linear regression. When shown on a graph, however, this line is straight, unlike other regression models. When the dependent variables are continuous, linear regression is used; however, when the dependent variables are categorical, such as "true" and "false" or "yes" and "no," logistical regression is used. While both regression models aim to discover correlations between data inputs, logistic regression is more commonly utilized to handle binary classification problems like spam detection. The KNN algorithm, also known as the K-nearest neighbor algorithm, is a non-parametric algorithm that classifies data points based on their proximity and correlation with other data. This technique assumes that data points that are comparable can be discovered close together. As a result, it attempts to determine the distance between data points, which is commonly done using Euclidean distance, and then assigns a category based on the most common category or average. Data scientists prefer it because of its ease of use and short computation time, but as the test dataset grows larger, the processing time increases, making it less appealing for classification jobs. KNN is commonly used in picture recognition and recommendation algorithms.

2.1.2 Neural Network and Deep Learning

A neural network is a network or circuit of biological neurons, or, more recently, an artificial neural network made up of artificial neurons or nodes, A neural network is thus either a biological neural network composed of biological neurons or an artificial neural network used to solve artificial intelligence (AI) challenges. Artificial neural networks model biological neuron connections as weights between nodes. A positive weight indicates an excitatory link, while a negative weight indicates an inhibitory connection. All inputs are weighted and added together. This activity is known as a linear combination. Finally, an activation function governs the output amplitude. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1. These artificial networks can be used for predictive modelling, adaptive control, and other applications that need training using a dataset. Self-learning can emerge as a result of experience within networks, which can draw inferences from a complicated

and seemingly unconnected set of data. Many models are employed; they are developed at various degrees of abstraction and model various elements of brain systems. They span from models of individual neurons' short-term behavior to models of the dynamics of brain circuitry deriving from interactions between individual neurons to models of behavior arising from abstract neural modules that represent entire subsystems. Models of long-term and short-term plasticity of neural systems, as well as their relationship to learning and memory, are included, ranging from the individual neuron to the system level. Scientists stated in August 2020 that bi-directional connections, or the addition of appropriate feedback connections, can accelerate and improve communication between and within modular neuronal networks of the brain's cerebral cortex, as well as reduce the threshold for successful communication. They demonstrated that adding feedback connections between resonance pairs can support the successful transmission of a single pulse packet across the whole network.

2.1.3 Long Short-term memory

In Long short-term memory (LSTM) is a deep learning artificial recurrent neural network (RNN) architecture. Unlike traditional feedforward neural networks, LSTM has feedback connections. It is capable of processing not just single data points (such as photos and audios), but also complete data sequences (such as speech or video). For example, LSTM can be used for tasks like unsegmented, connected handwriting recognition, speech recognition, and anomaly detection in network traffic or intrusion detection systems (intrusion detection systems). A cell, an input gate, an output gate, and a forget gate comprise a typical LSTM unit. The cell stores values for arbitrary time intervals, and the three gates control the flow of information into and out of the cell.

Because there might be lags of undetermined duration between critical occurrences in a time series, LSTM networks are well-suited to categorizing, processing, and making predictions based on time series data. LSTMs were created to address the vanishing gradient problem that might occur when training regular RNNs. In many cases, LSTM outperforms RNNs, hidden Markov models, and other sequence learning algorithms due to its relative insensitivity to gap length. Classic (or "vanilla") RNNs, in theory, may maintain track of arbitrary long-term dependencies in input sequences.

The problem with vanilla RNNs is computational (or practical) in nature: when back-propagating a vanilla RNN, the long-term gradients can "vanish" (that is, tend to zero) or "explode" (that is, tend to infinity) due to the computations involved in the process, which use finite-precision numbers. Because LSTM units allow gradients to flow unmodified, RNNs utilizing LSTM units partially solve the vanishing gradient problem. However, the expanding gradient problem can still affect LSTM networks. An RNN with LSTM units can be trained supervised on a set of training sequences, using an optimization algorithm like gradient descent combined with backpropagation through time to compute the gradients required during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the LSTM network's output layer) with respect to the corresponding weight.

2.2 Previous Work

2.2.1 Speech Recognition

2.2.1.1 Strategy & Structure

Some models have a significant disadvantage in that they can identify semantic segmentation but not instance segmentation within the same class. As a result, in this study, the researchers attempted to develop a model that can do semantic and instance segmentation using a high-order loss function called Coverage Loss. The researchers apply heuristics to limit the search space by performing hierarchical segmentation of the image.

2.2.1.2 Data

Their method suggested that they employed RAVDESS for the needed objects, which we were able to retrieve through Speech recognition. Python includes a built-in library for creating such databases. After bringing in the library, we created occurrences for each subject, with a keyword argument and a restriction number as arguments. Following the creation of their dataset, they generated their explanation files. For each audio file in their explanation files, they made a wav file. A few other libraries, such as MFCC and ZCR, are imported here. For audios, they employ the MFCC and ZCR Selector libraries.

Method Evaluation

Using the CNN architecture to extract additional feature moreover the segmentation tree, improves the performance of the model at the instance segmentation task as a qualitative result. However, the model struggles in case of larger numbers of objects inthe same scene. So, the perfect coverage score has not accomplished yet [16].

2.2.1.3 Results Evaluation

It has a total of 64 layers, 8 of which are hidden, an input or pre-processing layer, and the remaining eight levels shape the neural net's yield layer. Sequential functions are used to create the yield or audio box. The neural network's many layers have a sequence of nodes that shape the layers. Calculating approximations and correct nesses is the responsibility of these nodes. We may use the confusion matrix created for comprehending the various states of the programmer to better assess the demonstration.

Chapter 3: Material and Methods

3.1 Materials

3.1.1 Data

Source This dataset was intended to aid in the evaluation of deep-learning algorithms for emotion recognition in academic studies. We received audio datasets containing approximately 5000 wav audio files. from the following websites[10][11][12]:

The first website contains speech data which is available in audio format. We went with the Audio only zip file because we are dealing with finding emotions from speech. The zip file consisted of around 7442 audio files which were in wav format.

The second website contains around 1440 audio speeches from 24 different actors with different emotions.

The Third website contains around 2800 audio speeches from 8 different actors with different emotions.

We tested out one of the audio files to know its features by plotting its waveform and spectrogram.

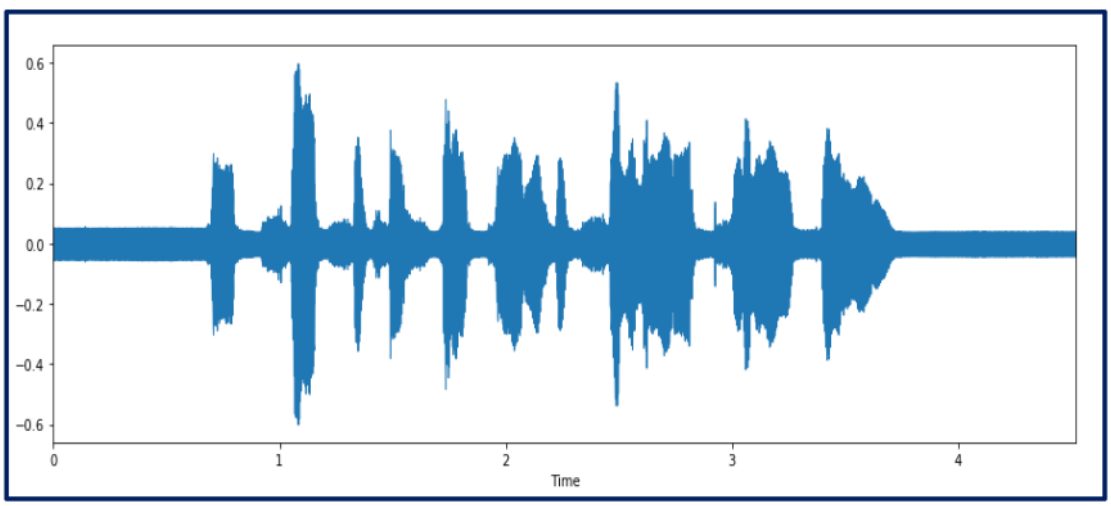


Figure 3: wave Dataset samples

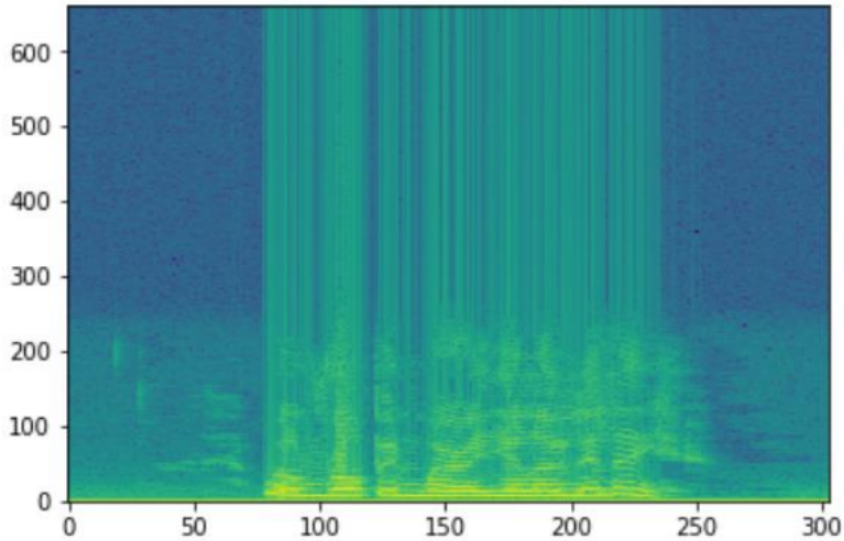


Figure 4: Chromogram Dataset samples

3.1.2 Tools

- Anaconda: is a platform for efficient developing and applying AI and machinelearning models.
- Python 3: An open-source programming language that enables developers to workand integrate their systems quickly and effectively.
- Spyder: is a platform I use it for the GUI an interface for my project
- Google Collaboratory: it is an open-source programming Visual machine used for training the model and testing it and to generate the features of the audio files

3.1.3 Environment

- Local CPU, intel i5 processor with 6 cores.
- Google Collaboratory GPU

3.2 Methods

3.2.1 System architecture Overview

First of all, the office-home dataset is used after filtration as an input to the model. As a preprocessing step, the data is loaded to our model and a fixed size is used by resizing or padding the images.

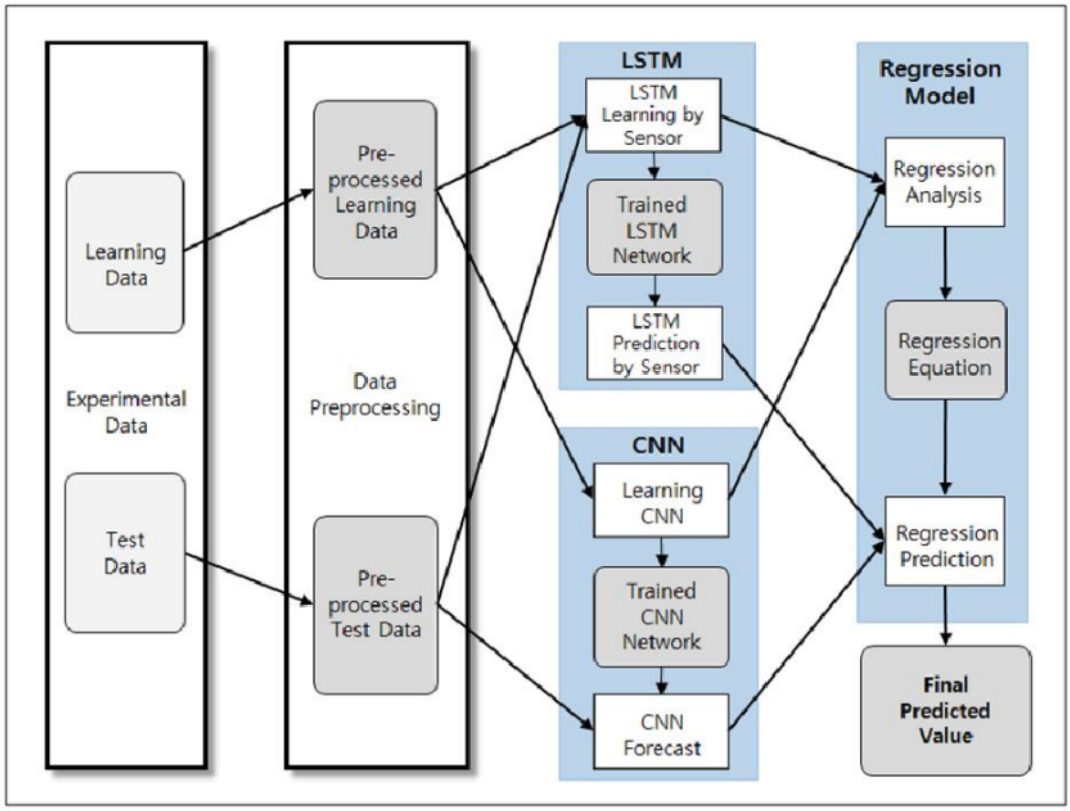


Figure 5: LSTM system architecture overview

Chapter 4: System Implementation

4.1 System Development

The first model was developed to check the ability of the machine to classify any audio file and determine the emotion of the audio happy, sad, surprises, fearful, disgust, angry, neutral.

The dataset is not in a fixed size, and there are some noises. So, the solution is preprocessing the input audios by 5 steps to use this data set

First step is to instance the audio is loaded to 'AudioSegment' object by using 'AudioSegment.form' Function:

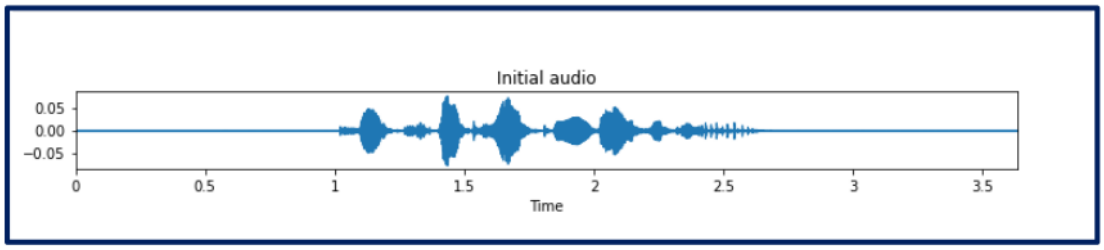


Figure 6: AudioSgement Audio Sample

Second step is to normalize the audio that means transform audio signals to array by using 'effects.normalized' Function:

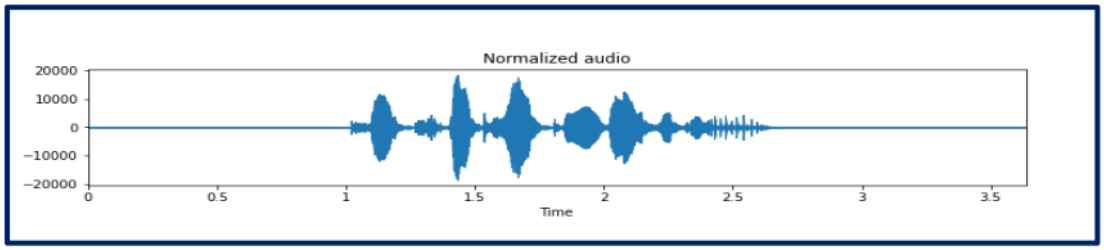


Figure 7: Normalized Audio Sample

Third step to trim silence from the beginning and end to get rid of unnecessary data by using 'librosa.effects.trim' Function:

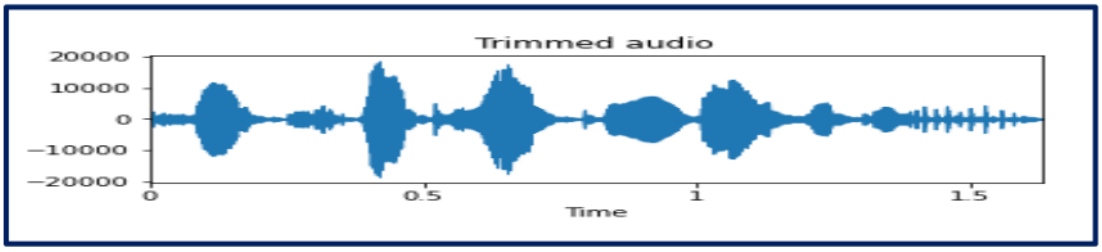


Figure 8: Trimmed Audio Sample

Fourth step is to padding the audio file to the same length for length equalization by using 'pad' Function:

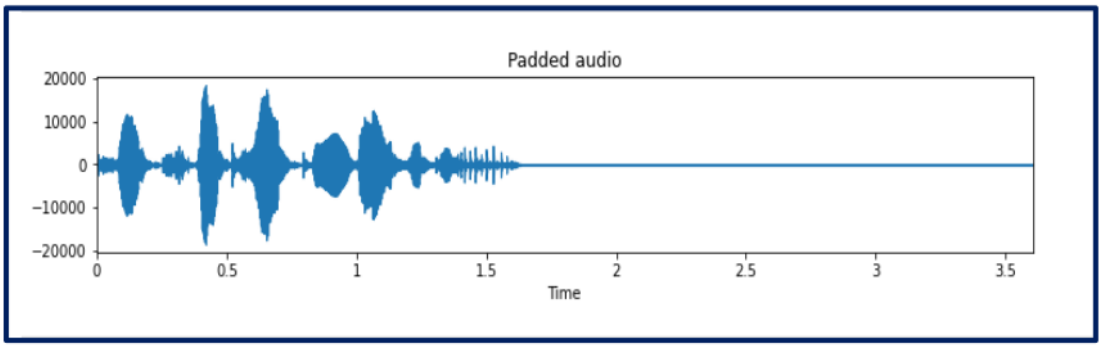


Figure 9: Padded Audio Sample

Last step is to remove the noise by using 'Reduce_noise' Function:

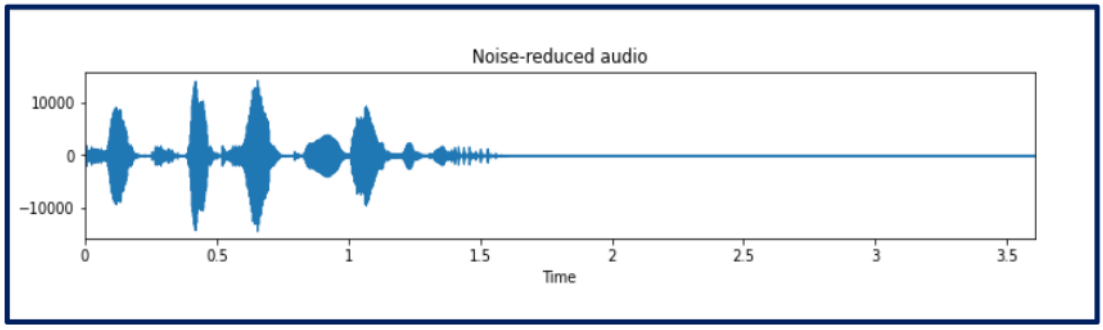


Figure 10: Noise-Reduced Audio Sample

By using Keras APIs a simple classification model of DNN architecture was developed toensure that the model is able to recognize the difference between the emotion of the input audio then output the emotion.

4.2 System Structure

4.2.1 System Overview

This system is composed of 3 essential stages. The first stage is responsible for the task of data preparing and preprocessing. The first step in this stage is dataset loading and reading. The second step is data preprocessing the data to make it apply to read by the 5 step that we talk about them before then the audio goes to Energy (RMS), Zero crossing rate and MFCCs then it will gives you one emotion from the 7 emotions like a output.

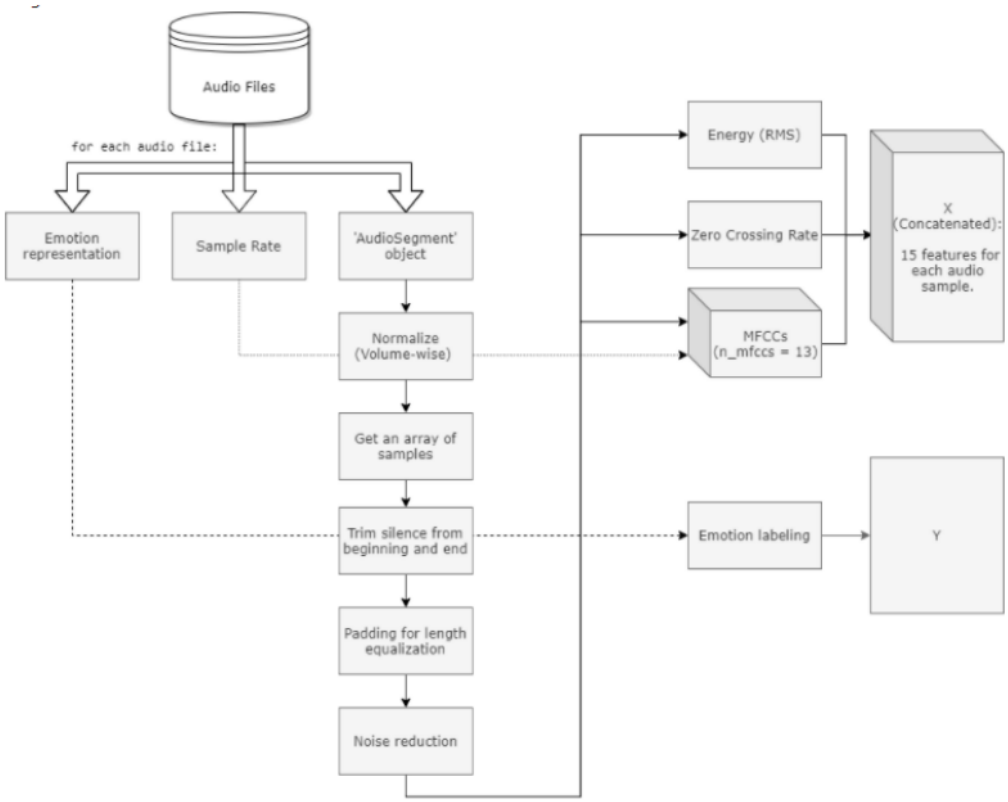


Figure 11: Classification System architecture overview

4.2.2 Class Diagram or Tensor Board

The following figure is the Tensor Board about the whole main components of the DNN architecture for the classification task. It describes its loss for train and validation and the second graph is the accuracy for train and validation.

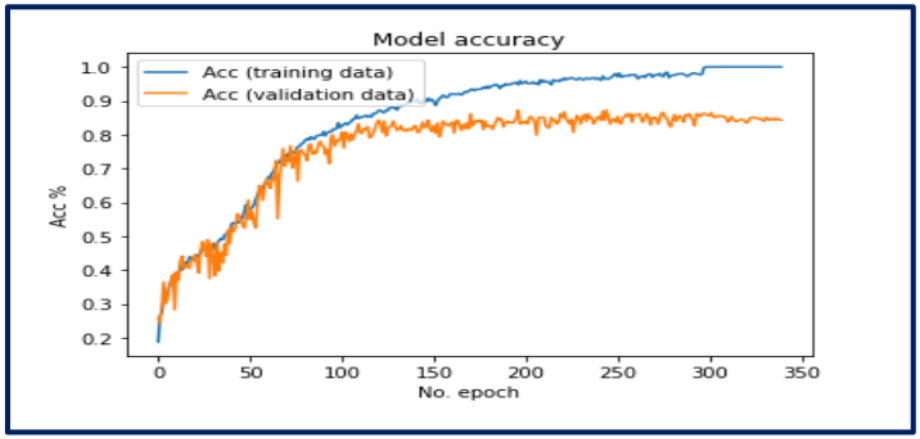


Figure 12: Graph Examples

4.3 System Running

This system takes an audio file from type (.wav) and preprocessing by the 5 steps to use the audio files Audio Segment object, Normalize, Trim Silence from beginning and end, padding for length equalization and last thing is noise reduction. Then we use 4 features to understand the audio and to know how to deal with it the three features are MFCC, RMS.ZCR and Chroma.stft Then this model gives the prediction of this audio file by a char Then it gives you the max emotion that this model has been found from these emotions (neutral, happy, sad, angry, fearful, disgust, surprised)

4.3.1 Component A

The input of this function is the original audio dataset, and the output is the emotion of this audio generated by data rigorization function to know the emotion of the audio file and we try many audios file to be sure that no overfitting in this model and the output like this char.

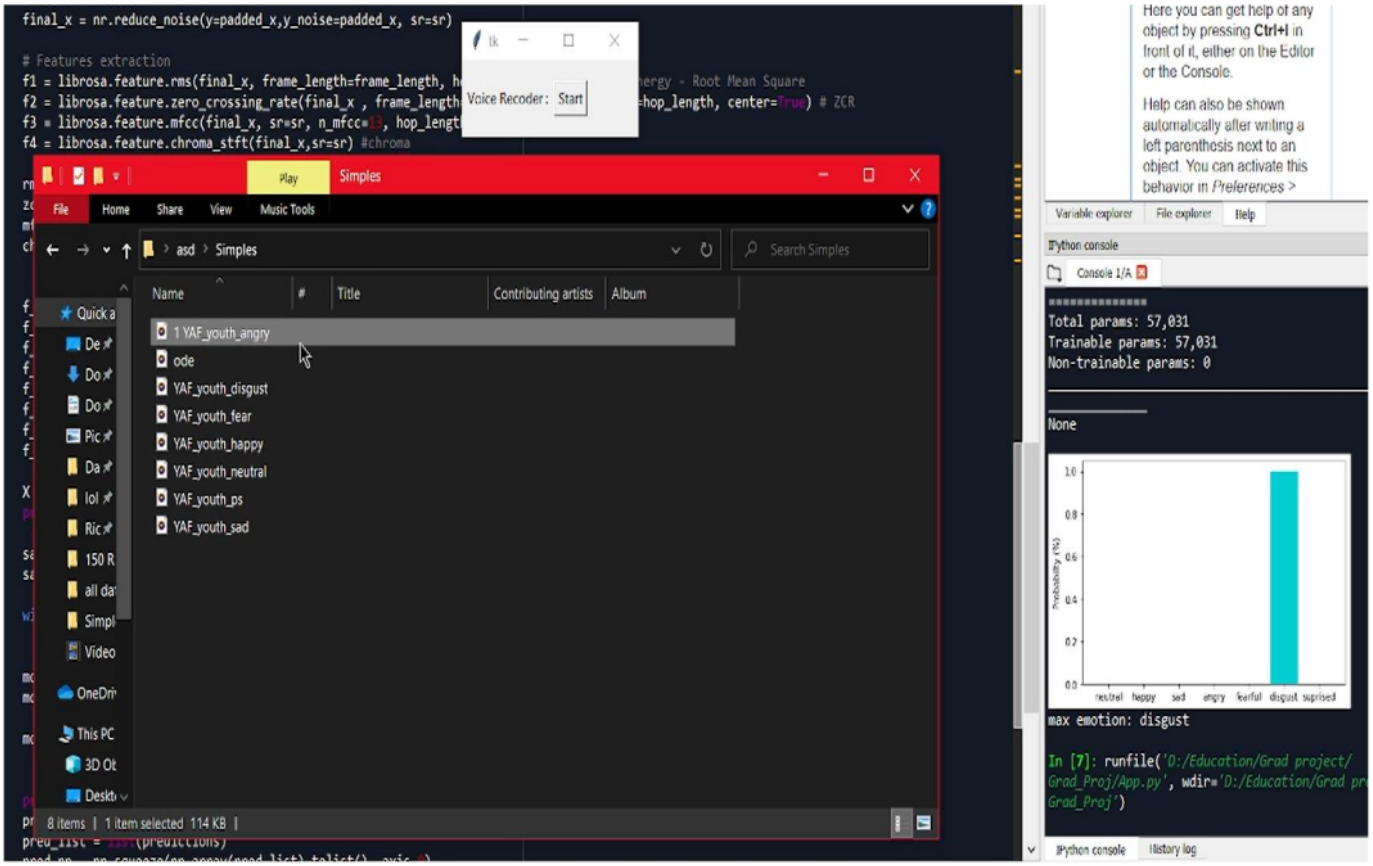


Figure 13: input and output of DNN data augmentation

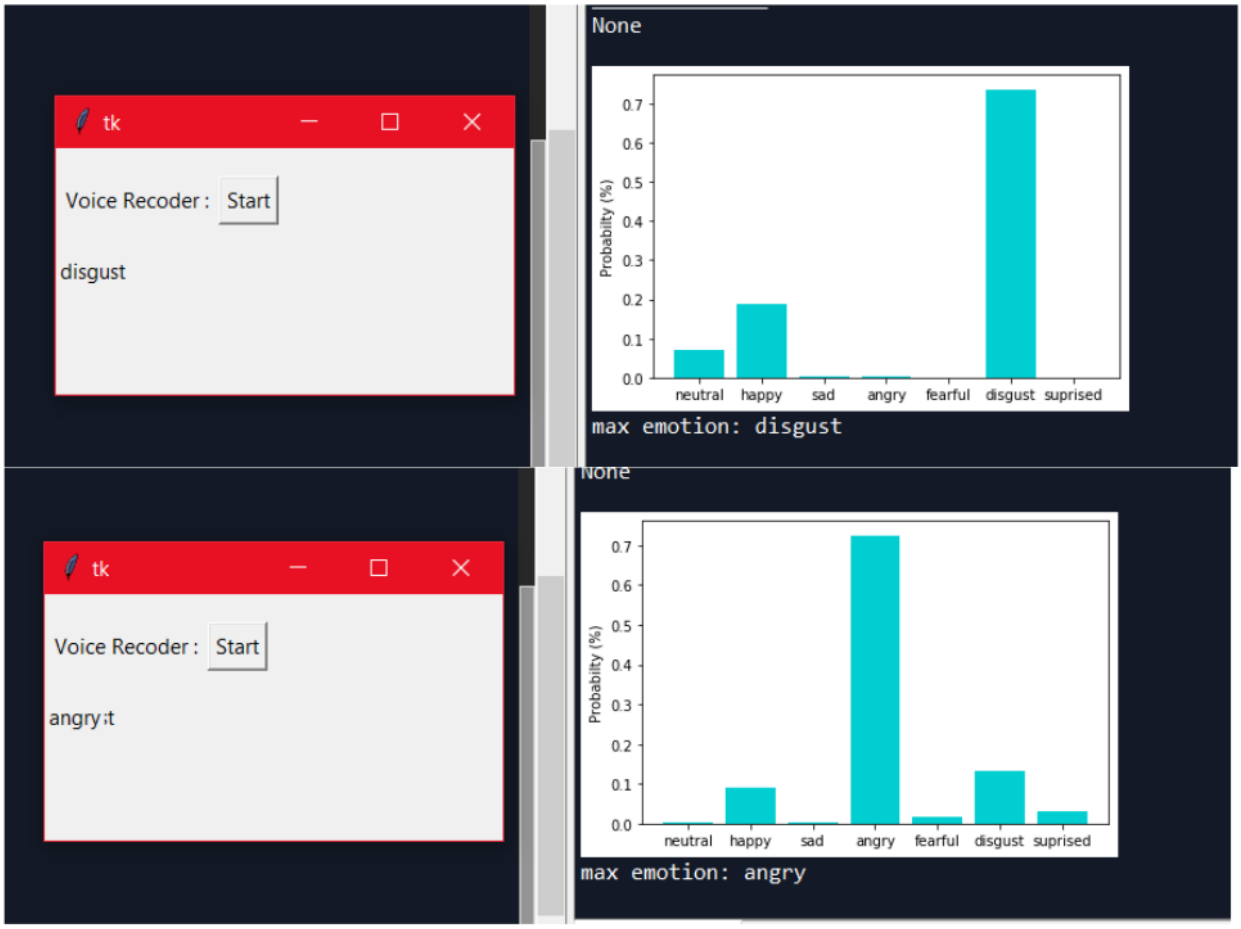
Chapter 5: Results and Evaluation

5.1 Testing Methodology

There are several approaches for evaluating how well the model predicts distinct classes from given datasets. And the methods utilized in the classification LSTM model for emotions such as sad, glad, furious, and so on will be presented in detail in this paragraph. To begin, the dataset was divided into three subsets: training, validation, and testing, and the dataset was fed to the model based on the batch size value supplied. Each of these strategies is assessed and tested independently based on accuracy and loss to determine the best preprocessing strategy to employ for the present classification challenge.

5.2 Results

5.2.1 Best Results Cases



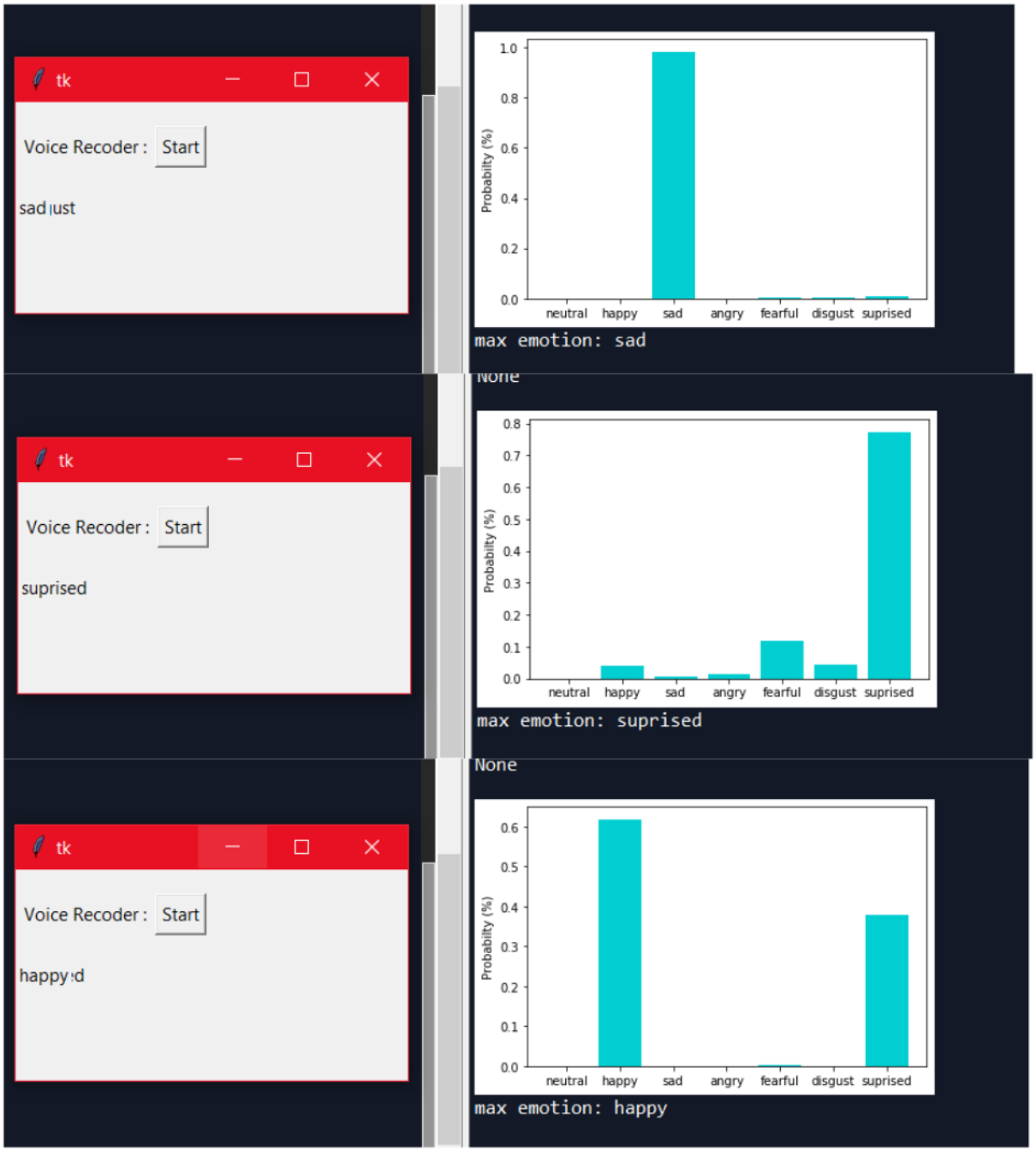


Figure 14: Random Results from the Best DNN Classification Model

The first best result was generated by the following architecture model. The model consists of three convolutional layers with 4 features that used RMS and ZCR and Chroma and MFCC function kernel size. The first two layers with 64 hidden node and without but the third layer without strides. The model was trained for 340 epoch and 24 batch size. The previous figure shows some random results of this model. If these results were observed, it will be recognized that the results are perfect because most of the predicted labels are right even the images that are hard to be recognized such as in

The emotion forms the dataset is always correct but from the record the model confided because all the input audios are their mother language is English but our English is not good as them so it did not get the best result from the model.

5.2.2 Limitations

The chosen dataset was filtered to provide items and real-world photos of chairs, beds, and tables. Following this decision, the number of images was insufficient to train the DNN model, and it was difficult to find another dataset with only one class in each image. As a result, the answer to this problem was data augmentation, which was better for the findings and avoided overfitting. The drawback that could not be overcome is the image quality, since many of them have poor lighting or the angle of the images loses many details and aspects about the existent thing. And if these issues were resolved, the model could produce superior outcomes.

5.3 Evaluation

5.3.1 Accuracy Evaluation

Models Number	No. Nof epochs	Dataset	Activation Features	Features	Batch Size	Training Accuracy	Validation Accuracy
Model 1	100	Small dataset	software	MFCC &ZCR&RMS& chroma	23	43%	54%
Model 2	100	Small dataset	Software	MFCC &ZCR&Chroma	23	29%	40%
Model 3	100	Small dataset	Software	MFCC &RMS&Chroma	23	33%	50%
Model 4	100	Small dataset	Software	Chroma&MFCC	23	30%	46%
Model 5	100	Small dataset	Software	MFCC &ZCR& RMS	23	30%	45%
Model 6	100	Small dataset	Software	MFCC &ZCR	23	35%	46%
Model 7	150	Ravdess	Software	MFCC &RMS	23	23%	48%
Model 8	150	Ravdess	Software	MFCC &ZCR&RMS& chroma	23	52%	49%
Model 9	150	Ravdess	Software	MFCC &ZCR& chroma	23	50%	43%
Model 10	150	Ravdess	Software	MFCC &RMS& chroma	23	48%	27%
Model 12	340	Used dataset	Software	MFCC &ZCR& RMS	23	70%	72%
Model 13	340	Used dataset	software	Chroma&MFCC&RMS%ZCR	23	70%	76%

Table 1: First Summary of the CNN Classification Models with their Accuracy

As shown in the previous table, different implementation to the classification LSTM architecture were tried with different features of parameters such as the like the MFCC,RMS,ZCR and Chroma this is the feature for make a good model and change in the datasets the small dataset take from every dataset 500 simple so it will be 1500 from all dataset but it gives as a small accuracy then the Ravdess is only one dataset 1280 audio use them all the accuracy also is not good at the last step using all the dataset then it gives a good accuracy and we change in the number if epochs Until we get the best accuracy like showing in the table.

The first graph represents the training and validation loss and it could be noticed that the training loss is 0.5276 and the validation loss is 0.9055, and the second graphs represents the training and validation accuracy.

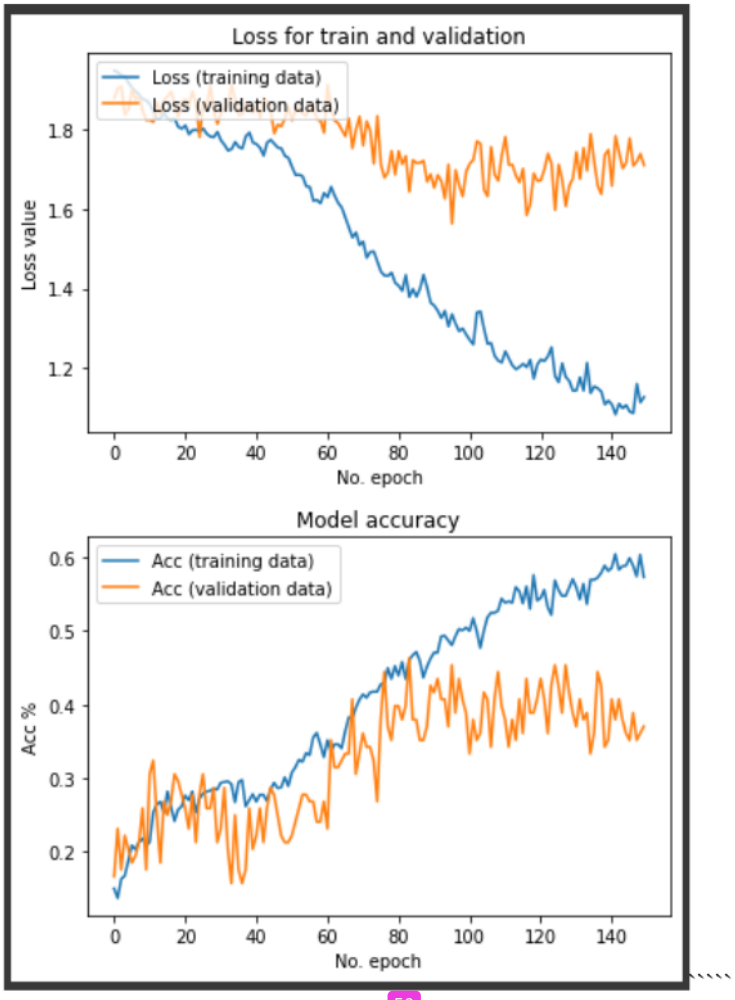
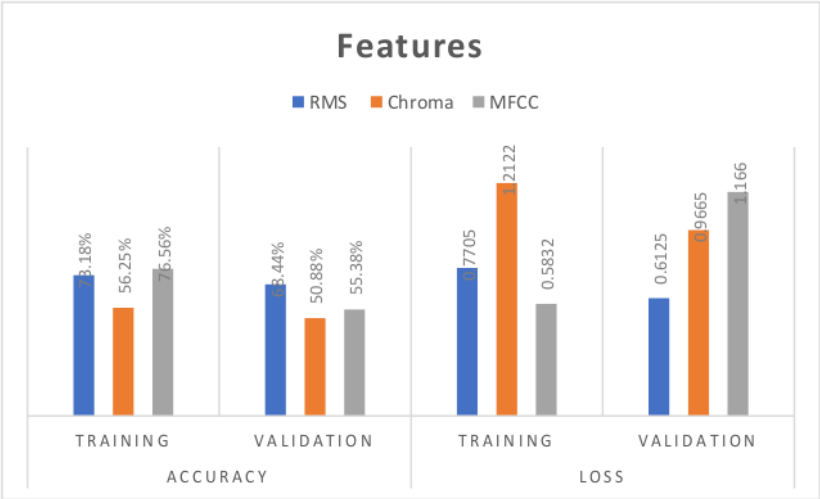


Figure 15: Graphs of the best DNN model loss and Accuracy

The training and validation accuracy and loss of three different activation functions are shown in the figure of the activation functions. In comparison to the MFCC, Chromo, ZCR and RMS activation functions, the sigmoid activation function has the lowest training accuracy.



5.3.2 Time Performance

In terms of time, employing Google Collab GPU resources over Spyder environments and Keras APIs on the local PC CPU provided a substantial gain. When both are compared, the time taken for the same architectures models on the local PC CPU was 76s per epoch and 4s/step, and on Google Collab GPU was 18s per epoch and 1s/step, but my model is too large and will take a long time, so I choose Google Collab to protect my computer.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

To summaries, a fully automatic system for understanding and segmenting the emotion of any voice was built in this thesis. After researching several approaches to determine the most advantageous approach capable of solving this target problem, the decision was made to pursue the deep learning approach and its various algorithms in order to construct a successful system that is regarded as an appropriate system for various emotions.

First, it was implemented a classification algorithm to identify between wave audio of one actor of the key Emotions, which are sad, glad, natural, fearful, furious, surprised, and disgusted. And the LSTM method was chosen for this work because it is a well-known technique for comparable classification tasks. This algorithm has demonstrated its capacity to distinguish between these things.

In this paper, we suggested a method for automatically extracting the emotional characteristic parameter from an emotional voice signal using deep belief networks (DBNs), one of the deep neural networks. We merged deep belief networks (DBNs) and support vector machines (SVMs) and suggested a classifier model based on deep belief networks (DBNs) and SVMs (SVM). During the practical training process, the model has a lower complexity and a 7% higher final recognition rate than typical artificial extracts, and this method can extract emotion characteristic characteristics effectively, significantly boosting the recognition rate of emotional speech recognition. However, training DBNs feature extraction model took 136 hours, which was longer than other feature extraction methods.

6.2 Problem Issues

6.2.1 Technical issues:

- To include as many images as possible for model training and validation, a large fixed size of image padding of 700x700 was chosen, but this affected the batch size because more than 5 batch size caused a system out of resources

crash and it takes more than 12 hours to run this small dataset, so increasing this dataset will take a long time to run.

–

6.3 Future Work

In the future work, deeper pre-trained models will be used in call center, so we need to increase the data and increase the Model value and the accuracy. In addition to that more classes will be added to be segmented and we need to use it in Arabic word also or multi-languages so it will be more accurate and will use in many countries not only in English countries.

References

1. *Kerkeni, L. (2019, March 25). Chapter: Automatic Speech Emotion Recognition Using Machine Learning. Automatic Speech Emotion Recognition Using Machine Learning. https://www.intechopen.com/chapters/65993*
2. *Speech emotion recognition using deep 1D & 2D CNN LSTM networks. (2019, January 1). ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S1746809418302337*
3. *Dan C Cires et al. (2013) Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks". In: Proc. MICCAI. Vol. 2. 2013, pp. 411(2108).*
4. *Dan Cires, Ueli Meier, and Jurgen Schmidhuber. (2012). Multi-column Deep Neural Networks for Image Classi_cation". In: Computer Vision and Pattern Recognition (CVPR), (2012) IEEE Conference on. IEEE February (2012).*
5. *P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear thesision," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, Jul 1997.*
6. *Dakopoulos, D.: TYFLOS: a wearable navigation prototype for blind and visually impaired; design, modelling and experimental results. Ph.D. Dissertation, Computer Science and Engineering, Wright State University (2009)*
7. *Nagabandi, A.; Kahn, G.; Fearing, R.S. and Levine, S. (2018). "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning". In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE. 7559-7566.*
8. *Toronto emotional speech set (TESS). (2019, August 25). Kaggle. https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess*
9. *CREMA-D. (2019, August 30). Kaggle. https://www.kaggle.com/ejlok1/cremad*

10. *RAVDESS Emotional speech audio*. (2019, January 19). *Kaggle*. *https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio*

Speech Emotion Recognition

ORIGINALITY REPORT

33%	32%	20%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.hindawi.com Internet Source	5%
2	www.dataversity.net Internet Source	4%
3	etd.aau.edu.et Internet Source	2%
4	www.ibm.com Internet Source	1%
5	medium.com Internet Source	1%
6	en.wikipedia.org Internet Source	1%
7	www.intechopen.com Internet Source	1%
8	businessyield.com Internet Source	1%
9	www.ncbi.nlm.nih.gov Internet Source	1%

10	dspace.bracu.ac.bd Internet Source	1 %
11	docshare04.docshare.tips Internet Source	1 %
12	ebin.pub Internet Source	1 %
13	eccv18-vlease.github.io Internet Source	1 %
14	Shammy Shikder Shanta, Md. Sham-E-Ansari, Atiqul Islam Chowdhury, Mohammad Munem Shahriar, Md. Khairul Hasan. "A Comparative Analysis of Different Approach for Basic Emotions Recognition from Speech", 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 2021 Publication	1 %
15	handwiki.org Internet Source	1 %
16	iarjset.com Internet Source	1 %
17	D. R. M. R. R. D. R. S. Eheliyagoda, T.K.G. Liyanage, D. C. Jayasooriya, D.P.Y.C.A. Nilmini et al. "Data-driven Business Intelligence Platform for Smart Retail Stores", 2021 3rd	1 %

International Conference on Advancements in Computing (ICAC), 2021
Publication

18	blog.adacore.com Internet Source	1 %
19	repository.tudelft.nl Internet Source	1 %
20	Yingjun Deng, Mingpeiyu Zhang, Tian Wang, Huaming Wu. "Time-Gated Remaining Useful Life Prediction with Non-Periodical Inspection Data", 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), 2021 Publication	<1 %
21	core.ac.uk Internet Source	<1 %
22	link.springer.com Internet Source	<1 %
23	kth.diva-portal.org Internet Source	<1 %
24	repobiblio.cuc.uqroo.mx Internet Source	<1 %
25	Chenchen Huang, Wei Gong, Wenlong Fu, Dongyu Feng. "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM", 'Hindawi Limited', 2014	<1 %

Internet Source

26	www.analyticsvidhya.com Internet Source	<1 %
27	Mohamed Amine Jerbi, Hadhemi Achour, Emna Souissi. "Chapter 9 Sentiment Analysis of Code-Switched Tunisian Dialect: Exploring RNN-Based Techniques", Springer Science and Business Media LLC, 2019 Publication	<1 %
28	jovian.ai Internet Source	<1 %
29	openarchive.ki.se Internet Source	<1 %
30	docplayer.net Internet Source	<1 %
31	ijasre.net Internet Source	<1 %
32	K Saranya, S Jayanthy. "Simultaneous Detection And Analysis Of Multi-Source Data In Prediction Of Human Emotion", 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021 Publication	<1 %
33	ijisrt.com Internet Source	<1 %

34	info.cloudquant.com Internet Source	<1 %
35	kupdf.net Internet Source	<1 %
36	hdl.handle.net Internet Source	<1 %
37	Amogh N. Parab, Dev V. Savla, Jay P. Gala, Kaustubh Y. Kekre. "Stress and Emotion Analysis using IoT and Deep Learning", 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020 Publication	<1 %
38	Mehmet Berkehan Akçay, Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", Speech Communication, 2020 Publication	<1 %
39	atitesting.com Internet Source	<1 %
40	giswin.geo.tsukuba.ac.jp Internet Source	<1 %
41	digital.library.txstate.edu Internet Source	<1 %
	documents.mx	

42	Internet Source	<1 %
43	norma.ncirl.ie Internet Source	<1 %
44	repositories.lib.utexas.edu Internet Source	<1 %
45	www.newworldencyclopedia.org Internet Source	<1 %
46	A. Christy, S. Vaithyasubramanian, A. Jesudoss, M. D. Anto Praveena. "Multimodal speech emotion recognition and classification using convolutional neural network techniques", International Journal of Speech Technology, 2020 Publication	<1 %
47	courses.cecs.anu.edu.au Internet Source	<1 %
48	qspace.library.queensu.ca Internet Source	<1 %
49	www.jetir.org Internet Source	<1 %
50	Chi Nguyen, Adnan Ahmad Cheema. "A Deep Neural Network-Based Multi-Frequency Path Loss Prediction Model from 0.8 GHz to 70 GHz", Sensors, 2021 Publication	<1 %

51	Cramer, Duncan. "EBOOK: ADVANCED QUANTITATIVE DATA ANALYSIS", EBOOK: ADVANCED QUANTITATIVE DATA ANALYSIS, 2003 Publication	<1 %
52	Yongming Huang, Jing Xiao, Kexin Tian, Ao Wu, Guobao Zhang. "Research on Robustness of Emotion Recognition Under Environmental Noise Conditions", IEEE Access, 2019 Publication	<1 %
53	C. Revathy, R. Sureshbabu. "Effective Technique for Noise Removal and Emotion Recognition in Speech Signals Using Cat Swarm Optimized Spiking Neural Networks", Fluctuation and Noise Letters, 2021 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		