**MACHINE LEARNING BASED SPEECH EMOTION RECOGNITION**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Topic** | **Page No.** |
|  | Abstract | I |
|  | List of Figures  List of Tables  Acronyms | Ii  iv  v |
| 1.0 | Introduction | 1 |
|  | 1.1 Literature Survey | 2 |
|  | 1.2 Problem Identification and its Solutions  1.2.1 Background  1.2.2 Solution Overview  1.3 Application requirement | 3  3  3  4 |
| 2.0 | Emotion and Classification  2.1 Definition  2.2 Categorization of Emotions  2.3 Sensory Modalities  2.3.1 Facial Expressions  2.3.2 Speech  2.3.4 Physiological Signals | 6  6  6  7  7  7  8 |
| 3.0 | Machine Learning  3.1 Evolution of Machine Learning  3.2 Importance of Machine Learning  3.3 Types of Machine Learning  3.3.1 Supervised Learning  3.3.2 Unsupervised Learning  3.3.3 Reinforcement Learning  3.4 Deep Learning  3.5 Machine Learning Vs Deep Learning  3.6 Convolutional Neural Network | 9  9  10  10  10  11  12  13  15  16  16 |
| 4.0 | Software Requirements  4.1 Python  4.2 Librosa  4.3 IPython  4.4 Matplotlib  4.5 Numpy  4.6 Pandas  4.7 Keras  4.8 Sklearn  4.9 Colab | 18  18  18  18  19  19  20  20  20  21 |
| 5.0 | Methodology  5.1 Proposed Methods  5.1.1 CNN Model with Spectrogram input  5.1.2 CNN Model with MFCC input  5.1.3 Combined CNN Model based Spectrogram  and MFCC models | 22  23  23  26  26 |
|  | 5.2 Algorithm | 27 |
|  | 5.3 Dataset  5.4 Modules | 30  31 |
| 6.0 | Results and Discussions | 33 |
| 7.0 | Advantages,Disadvantages and Applications  7.1 Advantages  7.2 Disadvantages  7.3 Applications | 37  37  37  37 |
| 8.0 | Conclusion and Future Scope  8.1 Conclusion  8.2 Future Scope  References | 39  39  39 |
|  |  |  |
|  |  |  |

**ABSTRACT**

Nowadays, monitoring and understanding a human’s emotional state plays a key role in the field of human computer interaction. On the other hand, this monitoring and analysis should be as unobtrusive as possible, since in our era the digital world has been smoothly adopted in everyday life activities. There are many approaches that have been proposed that make use of either sensors placed in the users’ environments (e.g., cameras and microphones), or even sensors placed on the users’ bodies (e.g., physiological or inertial sensors). These approaches typically feel that their privacy is violated and of course, the use of body sensors for long time periods is a factor that may also cause discomfort.

Thus, in this project, we characterize human speech emotion recognition (SER) by their speech or voice. To recognize human emotion through speech, various speech features were extracted. We also use MFCC, Chroma and Mel Frequency Cepstrum as speech features rather than raw waveform which may contain unnecessary information that doesn’t help on the classification. Here emotion recognition is done for different emotions like neutral, happy, sad, anger, fear, etc.

This can be done using python .(version 3.7) and some of the python libraries Librosa (version 0.6.3), Numpy Soundfile (version 0.9.0), Numpy PyAudio (version 0.2.11).

Speech Emotion Recognition has found increasing applications in practice, e.g., in security, medicine, entertainment, education.

i

**LIST OF FIGURES**

**S. No Fig. No Name of the Figure Page No.**

1 3.0.1 Machine Learning as subfield of AI 9

2 3.3.1 Supervised Learning 11

3 3.3.2 Unsupervised Learning 12

4 3.3 Reinforcement Learning 13

5 3.4 Neural Network 15

6 3.5.1 Machine Learning 16

7 3.5.2 Deep Learning 16

8 5.0 Architecture 22

9 5.1.1.1 Spectrogram of a sample audio signal 24

10 5.1.1.2 Spectrogram/MFCC based CNN model 25

11 5.1.3 Representative CNN architecture 26

12 5.2 Algorithm 27

13 5.2.1 Schema of MFCC Extraction 29

14 5.2.2 Process for computing the ST 29

representation

15 5.3 Histogram of Labels 30

16 3.3 Convolutional Neural Network 31

17 6.1 Time Domain Plot 33

ii

18 6.2 Frequency Domain Plot 33

19 6.4 Model Accuracy 35

20 6.5 Model Loss 35

iii

**LIST OF TABLES**

**S. No Table No. Name of the Table Page No.**

1 6.3 Model Summary 34

2 6.6 Actual Values Vs Predicted Values 36

iv

**ACRONYMS**

|  |  |  |
| --- | --- | --- |
| S No. | Abbreviations | Full Form |
| 1 | CNN | Convolutional Neural Network |
| 2 | SVM | Support Vector Machine |
| 3 | SER | Speech Emotion Recognition |
| 4 | GUI | Graphical User Interface |
| 5 | PAD | Pleasure Arousal Dominance |
| 6 | EEG | Electroencephalogram |
| 7 | ECG | Electrocardiogram |
| 8 | EMG | Electromyogram |
| 9 | AI | Artificial Intelligence |
| 10 | ReLU | Rectified Linear Unit |
| 11 | HTML | Hypertext Markup Language |
| 12 | MFCC | Mel Frequency Cepstral Coefficients |
| 13 | JSON | Javascript Object Notation |
| 14 | GPU | Graphics Processing Unit |
| 15 | SQL | Structured Query Language |
| 16 | SVR | Support Vector Regression |
| 17 | STFT | Short Term Fourier Transform |
| 18 | FFT | Fast Fourier Transform |
| 19 | DCT | Discrete Cosine Transform |
| 20 | LPC | Linear Prediction Coefficient |
| 21 | MSF | Modulation Spectral Features |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

v

1. **INTRODUCTION**

Speech is the main and direct means of transmitting information. It contains a wide variety of information, and it can express rich emotional information through the emotions it contains and visualize it in response to objects. Speech is the main and direct means of transmitting information. It contains a wide variety of information, and its scenes or events can express rich emotional information through the emotions it contains and visualize it in response to objects.

The automatic recognition of emotions by analyzing the human voice and facial expressions has become the scenes or events subject of numerous researches and studies in recent years. The fact that automatic emotion recognition contains emotions by analyzing the human voice and facial expressions has become the systems that can be used for different purposes in many areas has led to a significant increase in the number of studies on the subject of numerous researches and studies in recent years. The fact that automatic emotion recognition systems can be used for different purposes in many areas has led to a significant increase in the number of studies on this subject. The following systems can be cited as an example of the areas in which these studies are used and their intended use:

**Education**: a course system for distance education can detect bored users so that they can change the style or level of material provided in addition, provide emotional incentives or compromises.

**Automobile**: driving performance and the emotional state of the driver are often linked internally. Therefore, these systems can be used to promote the driving experience and to improve driving performance. ∙ Security: They can be used as support systems in public spaces by detecting extreme feelings such as fear and anxiety.

**Communication**: in call centers, when the automatic emotion recognition system is integrated with the interactive voice response system, it can help improve customer service.

**Health**: It can be beneficial for people with autism who can use portable devices to understand their own feelings and emotions and possibly adjust their social behavior accordingly.

It is known that some physiological changes occur in the body due to people's emotional state. Some variables such as pulse, blood pressure, facial expressions, body movements, brain waves, and acoustic properties vary depending on the emotional state. Pulse, blood pressure, brain waves, and so forth. Although changes cannot be detected without a portable medical device, facial expressions and voice signals can be received directly without connecting any device to the person. For this reason, most studies on this topic have focused on automatic recognition of emotions using visual and auditory signals.

However, acoustic signals are the most used data after facial signs to identify a person's emotional state. There are different methods and classifications available, such as k-Nearest Neighbor, Artificial Neural Networks, Hidden Markov Model, Gaussian Mixture Model, Support Vector Machines, Convolutional Neural Networks and others have been developed to classify human emotions according to learning data sets.

**1.1 Literature Survey**

Emotion recognition systems using Traditional Technique undergo three phases or the signals are converted into emotions with three fundamental components.

1. Signal Processing: Acoustic Pre - processing like noise cancellation and Segmentation Is done at this level.
2. Feature Extraction: To identify the relevant features present in the signal. Further this level is split into two sub parts
   * Extraction of the Feature
   * Selection of the Feature
3. Classification:Once relevant features are identified; this feature vectors are then mapped with relevant emotions.

CNN is another kind of Deep learning strategy dependent on feed forwarding design for grouping. CNNs basically are ordinarily utilized for design acknowledgment and give better information arrangement.Those systems contain minute size neurons available on each layer in the structured model design that procedure the information as responsive fields.

An enactment work for a non-direct strategy should be used to achieve the yields from convolution layers. It ought to be noticed that information sources are little districts of the first volume. Down inspecting is done at every sampling layer to be included maps and reduce the params in the system. This,controlling the overfitting and lifts the preparation procedure. The pool procedure is done over p\*p components for bordering spread of every element map. At the Last stage, the layers should be completely associated as in other neural systems. Later layers take the past low level and mid level highlights and create elevated level deliberation structure of the info discourse information. The last layer otherwise called SVM as well as SoftMax is used to additionally produce the score of arrangement in probabilistic keywords to identify a specific class.

**1.2 Problem Identification and its Solutions**

**1.2.1 Background**

As human beings speech is amongst the most natural ways to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with the messages.

As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today’s digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them.

We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this study we attempt to detect underlying emotions in recorded speech by analysing the acoustic features of the audio data of recordings.

**1.2.2 Solution Overview**

There are three classes of features in a speech namely, the lexical features (the vocabulary used), the visual features (the expressions the speaker makes) and the acoustic features (sound properties like pitch, tone, jitter, etc.).The problem of speech emotion recognition can be solved by analysing one or more of these features. Choosing to follow the lexical features would require a transcript of the speech which would further require an additional step of text extraction from speech if one wants to predict emotions from real-time audio. Similarly, going forward with analysing visual features would require the excess to the video of the conversations which might not be feasible in every case while the analysis on the acoustic features can be done in real-time while the conversation is taking place as we’d just need the audio data for accomplishing our task. Hence, we choose to analyse the acoustic features in this work.

Furthermore, the representation of emotions can be done in two ways:

* Discrete Classification: Classifying emotions in discrete labels like anger, happiness, boredom, etc.
* Dimensional Representation: Representing emotions with dimensions such as Valence (on a negative to positive scale), Activation or Energy (on a low to high scale) and Dominance (on an active to passive scale)

Both these approaches have their pros and cons. The dimensional approach is more elaborate and gives more context to prediction but it is harder to implement and there is a lack of annotated audio data in a dimensional format. The discrete classification is more straightforward and easier to implement but it lacks the context of the prediction that dimensional representation provides. We have used the discrete classification approach in the current study for lack of dimensionally annotated data in the public domain.

**1.3 Application Requirement**

Speech recognition depends on a lot of different context variations and environmental conditions. Ideally speech recognition should be speaker independent, accept natural language and unrestricted lexicon. This perfect scenario must be replaced in a real scenario context considering that human hearing is much more complex than the signal processing techniques currently implemented.

Some issues to be taken into account are :

• Background noise: fans, computers, machinery running.

• Speech interference: TV, radio background conversation.

• Sound reflections due to the room geometry.

• Non stationary events: door slams, irregular road noise, car horns.

• Signal degradation : microphone and transmission system distortions.

• Unknown words: improper English grammar, unfamiliar accent, out-of-vocabulary words.

• Unusual circumstances: stressed speaker.

• Speaker sound artifacts: speaker lip smacks, heavy breathing, mouth clicks and pops.

Background noise and speech interference as well as acoustic reflection can be reduced by choosing a proper microphone set close to the speaker and pointed to the desired source, while non stationary noise is difficult to handle and will probably lead to a bad trial that must be repeated. The signal degradation due to the material is neglected if the same microphone and system is used for the training and for the testing. This condition is necessary to obtain reasonable performances and also helps taking into account the previously cited acoustic interference. The latter points are more connected to the speaker, and their influence depends on the application and on the techniques used to handle them.

A speech recognition system should work with various speakers and natural unconstrained language. In practice, to build a successful application for speech, some considerations on the features of the application have to be done. From this point of view the following characteristics are important:

• Application has to be designed for the real benefit of the user : increasing productivity, ease of use, good graphical user interface (GUI).

• Application has to be a user friendly system : feedback on failures, effective means of communication.

• Application must be accurate : achieve a specified level of well-defined performances.

• Application must work in real time : system must output the recognized word within maximum 250 ms from the end of the pronounced speech. One recurrent problem is the bearing of errors. The next points proposed a summary of the possible solutions:

• Failure soft method : the error is accepted and will be corrected by a later manual or automatic processing.

• Self-detection of errors : the output space is constrained with some rules to correct the decision. This correction can be performed with a list of expected words.

• Verification : before proceeding the user is asked when there is a doubt on which result is the correct one the first, second or third choice, for example.

• Rejection : pass on to operator when the score is too low

**2.0 EMOTION AND CLASSIFICATION**

This section is concerned with defining the term emotion, presenting its different models. Also for recognizing emotions, there are several techniques and inputs that can be used. A brief description of all of the techniques is presented here.

**2.1 Definition**

A definition is both important and difficult because the everyday word “emotion” is a notoriously fluid term in meaning. Emotion is one of the most difficult concepts to define in psychology. In fact, there are different definitions of emotions in the scientific literature. In everyday speech, emotion is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure . Scientific discourse has drifted to other meanings and there is no consensus on a definition. Emotion is often entwined with temperament, mood, personality, motivation, and disposition. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes. These changes influence thought and behavior. According to other theories, emotions are not causal forces but simply syndromes of components such as motivation, feeling, behavior, and physiological changes. In 1884, in What is an emotion?, American psychologist and philosopher William James proposed a theory of emotion whose influence was considerable. According to his thesis, the feeling of intense emotion corresponds to the perception of specific bodily changes. This approach is found in many current theories: the bodily reaction is the cause and not the consequence of the emotion. The scope of this theory is measured by the many debates it provokes. This illustrates the difficulty of agreeing on a definition of this dynamic and complex phenomenon that we call emotion. “Emotion” refers to a wide range of affective processes such as moods, feelings, affects, and wellbeing. The term “emotion” has been also referred to an extremely complex state associated with a wide variety of mental, physiological, and physical events.

**2.2 Categorization of emotions**

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous). In the first approach, emotions are described with a discrete number of classes. Many theorists have conducted studies to determine which emotions are basic . A most popular example is Ekman who proposed a list of six basic emotions, which are anger, disgust, fear, happiness, sadness, and surprise. He explains that each emotion acts as a discrete category rather than an individual emotional state. In the second approach, emotions are a combination of several psychological dimensions and identified by axes. Other researchers define emotions according to one or more dimensions. Wilhelm Max Wundt proposed in 1897 that emotions can be described by three dimensions: (1) strain versus relaxation, (2) pleasurable versus unpleasurable, and (3) arousing versus subduing. PAD emotional state model is another three-dimensional approach by Albert Mehrabian and James Russell where PAD stands for pleasure, arousal, and dominance. Another popular dimensional model was proposed by James Russell in 1977. Unlike the earlier three-dimensional models, Russell’s model features only two dimensions which include (1) arousal (or activation) and (2) valence (or evaluation) . The categorical approach is commonly used in SER. It characterizes emotions used in everyday emotion words such as joy and anger. In this work, a set of six basic emotions (anger, disgust, fear, joy, sadness, and surprise) plus neutral, corresponding to the six emotions of Ekman’s model, were used for the recognition of emotion from speech using the categorical approach.

**2.3 Sensory modalities for emotion expression**

There is vigorous debate about what exactly an individual can express nonverbally. Humans can express their emotions through many different types of nonverbal communication including facial expressions, quality of speech produced, and physiological signals of the human body. In this section, we discuss each of these categories.

**2.3.1 Facial expressions**

The human face is extremely expressive, able to express countless emotions without saying a word. And unlike some forms of nonverbal communication, facial expressions are universal. The facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same across cultures.

**2.3.2 Speech**

In addition to faces, voices are an important modality for emotional expression. Speech is a relevant communicational channel enriched with emotions: the voice in speech not only conveys a semantic message but also the information about the emotional state of the speaker. Some important voice feature vectors that have been chosen for research such as fundamental frequency, mel-frequency cepstral coefficient (MFCC), prediction cepstral coefficient (LPCC), etc.

**2.3.3 Physiological signals**

The physiological signals related to the autonomic nervous system allow us to assess objectively emotions. These include electroencephalogram (EEG), heart rate (HR), 4 Social Media and Machine Learning electrocardiogram (ECG), respiration (RSP), blood pressure (BP), electromyogram (EMG), skin conductance (SC), blood volume pulse (BVP), and skin temperature (ST) . Using physiological signals to recognize emotions is also helpful to those people who suffer from physical or mental illness thus exhibit problems with facial expressions or tone of voice.

**3.0 MACHINE LEARNING**

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome.

Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

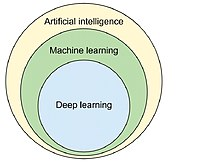


Fig 3.0.1 Machine Learning as a subfield of AI

**3.1 Evolution of Machine Learning**

It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. they learn from previous computations to produce reliable, repeatable decisions and results. This is a science that is not new but which has gained fresh momentum.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data - over and over, faster and faster - is a recent development.

**3.2 Importance of Machine Learning**

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage.

All of these mean it’s possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results - even on a very large scale. By building precise models, an organization has a better chance of identifying profitable opportunities - or avoiding unknown risks.

Machine learning has several very practical applications that drive the kind of real business results – such as time and money savings – that have the potential to dramatically impact the future of your organization. At Interactions in particular, we see tremendous impact occurring within the customer care industry, whereby machine learning is allowing people to get things done more quickly and efficiently. Through Virtual Assistant solutions, machine learning automates tasks that would otherwise need to be performed by a live agent – such as changing a password or checking an account balance. This frees up valuable agent time that can be used to focus on the kind of customer care that humans perform best: high touch, complicated decision-making that is not as easily handled by a machine. At Interactions, we further improve the process by eliminating the decision of whether a request should be sent to a human or a machine: unique Adaptive Understanding technology, the machine learns to be aware of its limitations, and bail out to humans when it has a low confidence in providing the correct solution.

To create good machine learning systems Data preparation capabilities, Algorithms - basic and advanced, Automation and iterative processes, Scalability, Ensemble modeling are required.

**3.3 Types of Machine Learning**

**3.3.1 Supervised Learning**

Supervised learning is the most popular paradigm for machine learning. It is the easiest to understand and the simplest to implement.

Given data in the form of examples with labels, we can feed a learning algorithm these example-label pairs one by one, allowing the algorithm to predict the label for each example, and giving it feedback as to whether it predicted the right answer or not. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully-trained, the supervised learning algorithm will be able to observe a new, never-before-seen example and predict a good label for it.

Supervised learning is often described as task-oriented because of this. It is highly focused on a singular task, feeding more and more examples to the algorithm until it can accurately perform on that task.

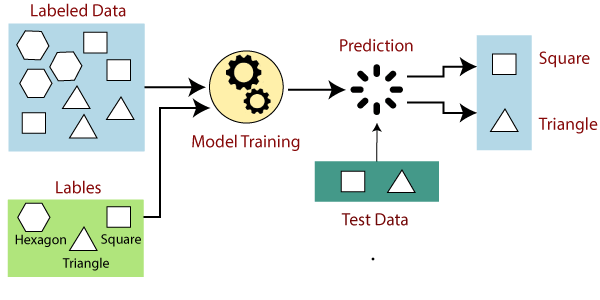


Fig 3.3.1 Supervised Learning

**3.3.2 Unsupervised Learning**

Unsupervised learning is very much the opposite of supervised learning. It features no labels. Instead, our algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and/or organize the data in a way such that a human (or other intelligent algorithm) can come in and make sense of the newly organized data.

What makes unsupervised learning such an interesting area is that an overwhelming majority of data in this world is unlabeled. Having intelligent algorithms that can take our terabytes and terabytes of unlabeled data and make sense of it is a huge source of potential profit for many industries. That alone could help boost productivity in a number of fields.

For example, what if we had a large database of every research paper ever published and we had unsupervised learning algorithms that knew how to group these in such a way so that you were always aware of the current progression within a particular domain of research. Now, you begin to start a research project yourself, hooking your work into this network that the algorithm can see. As you write your work up and take notes, the algorithm makes suggestions to you about related works, works you may wish to cite, and works that may even help you push that domain of research forward. With such a tool, your productivity can be extremely boosted.

Because unsupervised learning is based upon the data and its properties, we can say that unsupervised learning is data-driven. The outcomes from an unsupervised learning task are controlled by the data and the way it's formatted.

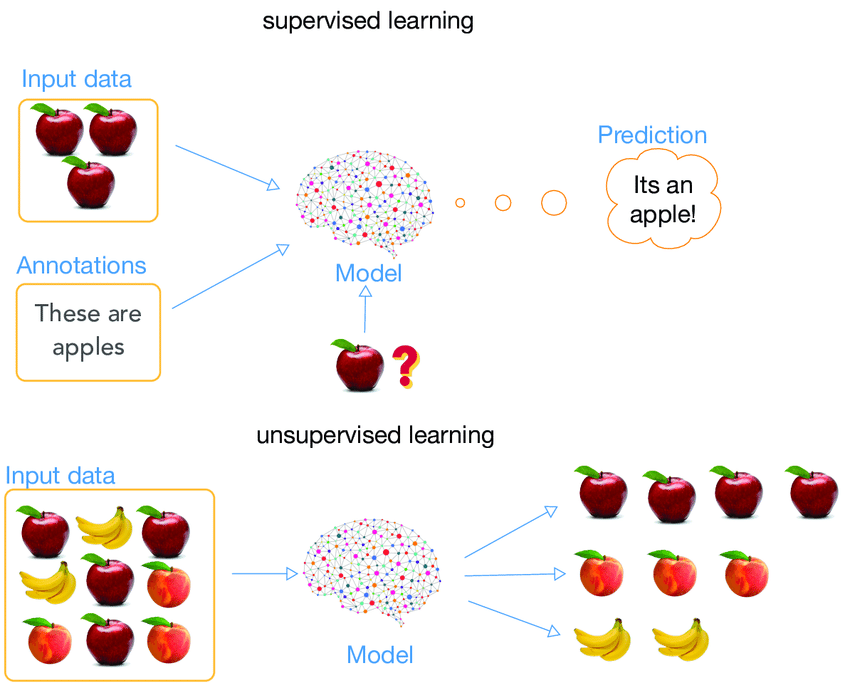


Fig 3.3.2 Unsupervised Learning

**3.3.3 Reinforcement Learning**

Reinforcement learning is fairly different when compared to supervised and unsupervised learning. Where we can easily see the relationship between supervised and unsupervised (the presence or absence of labels), the relationship to reinforcement learning is a bit murkier. Some people try to tie reinforcement learning closer to the two by describing it as a type of learning that relies on a time-dependent sequence of labels, however, my opinion is that that simply makes things more confusing.

I prefer to look at reinforcement learning as learning from mistakes. Place a reinforcement learning algorithm into any environment and it will make a lot of mistakes in the beginning. So long as we provide some sort of signal to the algorithm that associates good behaviors with a positive signal and bad behaviors with a negative one, we can reinforce our algorithm to prefer good behaviors over bad ones. Over time, our learning algorithm learns to make less mistakes than it used to.

Reinforcement learning is very behavior driven. It has influences from the fields of neuroscience and psychology.

For any reinforcement learning problem, we need an agent and an environment as well as a way to connect the two through a feedback loop. To connect the agent to the environment, we give it a set of actions that it can take that affect the environment. To connect the environment to the agent, we have it continually issue two signals to the agent: an updated state and a reward (our reinforcement signal for behavior).

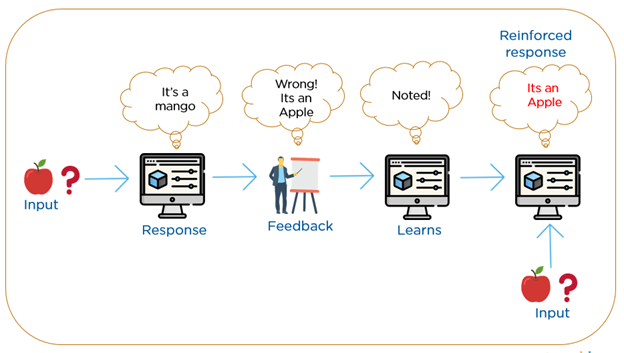
****

Fig 3.3.3 Reinforcement Learning

**3.4 Deep Learning**

Deep learning is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network.

Deep learning has evolved hand-in-hand with the digital era, which has brought about an explosion of data in all forms and from every region of the world. This data, known simply as big data, is drawn from sources like social media, internet search engines, e-commerce platforms, and online cinemas, among others. This enormous amount of data is readily accessible and can be shared through fintech applications like cloud computing.

However, the data, which normally is unstructured, is so vast that it could take decades for humans to comprehend it and extract relevant information. Companies realize the incredible potential that can result from unraveling this wealth of information and are increasingly adapting to AI systems for automated support.

Deep learning unravels huge amounts of unstructured data that would normally take humans decades to understand and process.

Deep learning achieves recognition accuracy at higher levels than ever before. This helps consumer electronics meet user expectations, and it is crucial for safety-critical applications like driverless cars. Recent advances in deep learning have improved to the point where deep learning outperforms humans in some tasks like classifying objects in images.

While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

1. Deep learning requires large amounts of labeled data. For example, driverless car development requires millions of images and thousands of hours of video.

2. Deep learning requires substantial computing power. High-performance GPUs have a parallel architecture that is efficient for deep learning. When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

Most deep learning methods use neural network architectures, which is why deep learning models are often referred to as deep neural networks. The term “deep” usually refers to the number of hidden layers in the neural network. Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150. Deep learning models are trained by using large sets of labeled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.

A neural network is structured like the human brain and consists of artificial neurons, also known as nodes. These nodes are stacked next to each other in three layers:

* The input layer
* The hidden layer(s)
* The output layer

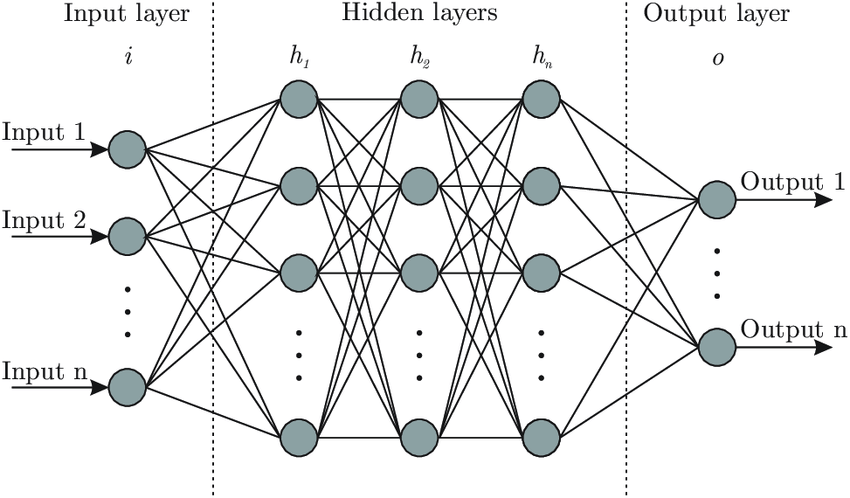


Fig 3.4 Neural Network

Data provides each node with information in the form of inputs. The node multiplies the inputs with random weights, calculates them and adds a bias. Finally, nonlinear functions are applied to determine which neuron to fire.

**3.5 Machine Learning Vs Deep Learning**

Deep learning is a specialized form of machine learning. A machine learning workflow starts with relevant features being manually extracted from images. The features are then used to create a model that categorizes the objects in the image. With a deep learning workflow, relevant features are automatically extracted from images. In addition, deep learning performs “end-to-end learning” – where a network is given raw data and a task to perform, such as classification, and it learns how to do this automatically.

Another key difference is deep learning algorithms scale with data, whereas shallow learning converges. Shallow learning refers to machine learning methods that plateau at a certain level of performance when you add more examples and training data to the network.

A key advantage of deep learning networks is that they often continue to improve as the size of your data increases.

In machine learning, you manually choose features and a classifier to sort images. With deep learning, feature extraction and modeling steps are automatic.

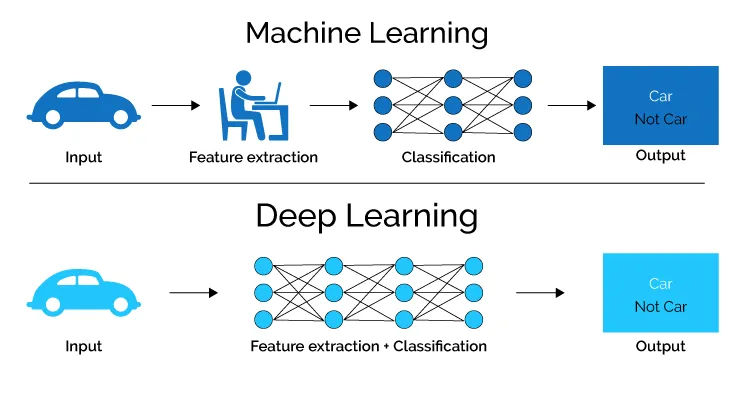


Fig 3.5.1 Machine Learning

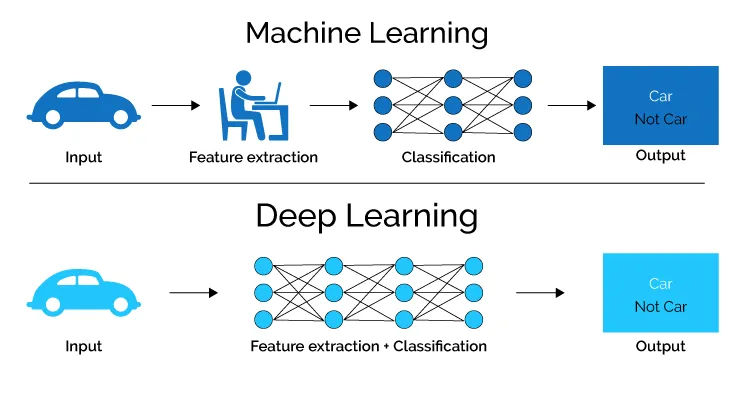


Fig 3.5.2 Deep Learning

**3.6 Convolutional Neural Network**

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. One of the most popular types of deep neural networks is known as convolutional neural networks (CNN or ConvNet). A CNN convolves learned features with input data, and uses 2D convolutional layers, making this architecture well suited to processing 2D data, such as images.

The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area.

CNNs eliminate the need for manual feature extraction, so you do not need to identify features used to classify images. CNN works by extracting features directly from images. The relevant features are not pretrained; they are learned while the network trains on a collection of images. This automated feature extraction makes deep learning models highly accurate for computer vision tasks such as object classification.

CNN’s have multiple layers that process and extract feature from data:

* **Convolution Laye**r: CNN has a convolution layer that has several filters to perform the convolution operation.
* **Rectified Linear Unit (ReLU)**: CNN’s have a ReLU layer to perform operations on elements. The output is a rectified featured map.
* **Pooling Layer**: The rectified feature map next feeds into a pooling layer. Pooling is a down-sampling operation that reduces the dimensions of the feature map. Pooling layer then converts the resulting two-dimensional array from the pooled feature map into a single, long, continuous, linear vector by flattening it.
* **Fully Connected Layer**: A fully connected layer forms when the flattened matrix from the pooling layer is fed as an input, which classifies and identifies the images.

**4.0 SOFTWARE REQUIREMENTS**

**4.1 Python**

Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

**4.2 Librosa**

Librosa is a python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation, Automatic Speech Recognition. It provides the building blocks necessary to create music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques. Librosa can be defined as a package which is structured as a collection of submodules which further contains other functions.

**librosa.display:**

Visualization and display routines using matplotlib.

**4.3 IPython**

IPython (Interactive Python) is a command shell for interactive computing in multiple programming languages, originally developed for the Python programming language, that offers introspection, rich media, shell syntax, tab completion, and history. IPython provides the following features.

* Interactive shells (terminal and Qt-based).
* A browser-based notebook interface with support for code, text, mathematical expressions, inline plots and other media.
* support for interactive data visualization and use of GUI toolkits.
* Flexible, embeddable interpreters to load into one’s own projects.
* Tools for parallel computing

IPython is based on an architecture that provides parallel and distributed computing. IPython enables parallel applications to be developed, executed, debugged and monitored interactively, hence the I (interactive) in IPython. This architecture abstracts out parallelism, enabling IPython to support many different styles of parallelism.

IPython frequently draws from SciPy stack libraries like NumPy and SciPy, often installed alongside one of many Scientific Python distributions. IPython provides integration with some libraries of SciPy stack, notably matplotlib, producing inline graphs when in use with Jupyter notebook. Python libraries can implement rendering of mathematical expressions as rendered LaTex when used within IPython context, and Pandas dataframe uses a HTML representation.

**4.4 Matplotlib**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Pyplot is a Matplotlib module which provides a MATLAB-like interface.[10] Matplotlib is designed to be as usable as MATLAB, with the ability to use Python, and the advantage of being free and open-source.

**4.5 NumPy**

The fundamental package for scientific computing with Python

* Powerful N-Dimensional Arrays: Fast and versatile, the NumPy vectorization, indexing, and broadcasting concepts are the de-facto standards of array computing today.
* Numerical Computing Tools: NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.
* Interoperable: NumPy supports a wide range of hardware and computing platforms, and plays well with distributed, GPU, and sparse array libraries.
* Performant: The core of NumPy is well-optimized C code. Enjoy the flexibility of Python with the speed of compiled code.
* Easy to use: NumPy’s high level syntax makes it accessible and productive for programmers from any background or experience level.
* Open Source: Distributed under a liberal [BSD license](https://github.com/numpy/numpy/blob/master/LICENSE.txt), NumPy is developed and maintained [publicly on GitHub](https://github.com/numpy/numpy) by a vibrant, responsive, and diverse [community](https://numpy.org/community).

**4.6 Pandas**

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the [Python](https://www.python.org/) programming language.

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging,reshaping,selecting, as well as data cleaning, and data wrangling features.

**4.7 Keras**

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Up until version 2.3 Keras supported multiple backends, including TensorFlow[,](https://en.m.wikipedia.org/wiki/Artificial_neural_network) Microsoft Cognitive Toolkit[,](https://en.m.wikipedia.org/wiki/Artificial_neural_network) R[,](https://en.m.wikipedia.org/wiki/Artificial_neural_network) Theano[, and](https://en.m.wikipedia.org/wiki/Artificial_neural_network) PlaidML[.As of version 2.4, only](https://en.m.wikipedia.org/wiki/Artificial_neural_network) TensorFlow [is supported. Designed to enable fast experimentation with](https://en.m.wikipedia.org/wiki/Artificial_neural_network) deep neural networks[, it focuses on being user-friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer is François Chollet, a](https://en.m.wikipedia.org/wiki/Artificial_neural_network) Google [engineer. Chollet also is the author of the XCeption deep neural network model](https://en.m.wikipedia.org/wiki/Artificial_neural_network) works. Keras acts as an interface for the TensorFlow library.

**4.8 scikit-learn**

Simple and efficient tools for predictive data analysis. It is accessible to everybody, and reusable in various contexts. It is built on NumPy, SciPy, and matplotlib. Open source, commercially usable - BSD license.

#### Classification

Identifying which category an object belongs to. Applications - Spam detection, image recognition. Algorithms - SVM, nearest neighbors, random forest.

#### Regression

Predicting a continuous-valued attribute associated with an object. Applications - Drug response, Stock prices. Algorithms - SVR, nearest neighbors, random forest.

#### Clustering

Automatic grouping of similar objects into sets. Applications - Customer segmentation, Grouping experiment outcomes Algorithms - k-Means, spectral clustering, mean-shift.

#### Dimensionality reduction

Reducing the number of random variables to consider. Applications - Visualization, Increased efficiency Algorithms - k-Means, feature selection, non-negative matrix factorization.

#### Model selection

Comparing, validating and choosing parameters and models. Applications - Improved accuracy via parameter tuning. Algorithms - grid search, cross validation, metrics.

#### Preprocessing

Feature extraction and normalization. Applications - Transforming input data such as text for use with machine learning algorithms. Algorithms - preprocessing, feature extraction.

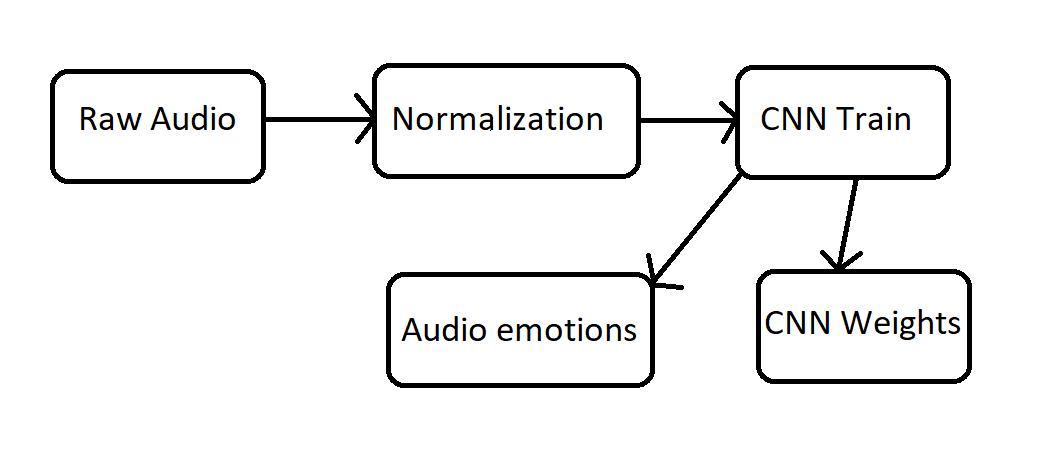
**4.9 Colab**

Google is quite aggressive in AI research. Over many years, Google developed an AI framework called TensorFlow and a development tool called Colaboratory. Today TensorFlow is open-sourced and since 2017, Google made Colaboratory free for public use. Colaboratory is now known as Google Colab or simply Colab.

Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs.

**5.0 METHODOLOGY**

The speech emotion recognition application is executed using convolutional neural networks. Following is the architecture of the system:

Fig 5.0 Architecture

**Training Model and Testing Model**

A training data is fetched to the system which consists of the expression label and Weight training is also provided for that network. An audio is taken as an input. Thereafter, intensity normalisation is applied over the audio. A normalised audio is used to train the Convolutional Network, this is done to ensure that the impact of the presentation sequence of the examples doesn’t affect the training performance. The collections of weights come out as an outcome to this training process and it acquires the best results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final network weights trained it gives the determined emotion. The output is represented in a numerical value each corresponds to either of five expressions.

There are 3 emotions that are being detected based on the person’s bpm value, those are Relaxed/Calm,Joy/Amusement, Fear/Anger. The produced art’s colors and shapes are parallel to the detected emotion based on the principles of “color psychology” and “shape psychology”.

**5.1 Proposed Methods**

**CNN Model based on Speech Features**

Recent research into speech processing has shown the successful application of deep learning methods and concepts such as CNNs and Long Short Term Memory (LSTM) cells on speech features. CNNs have been shown, by extensive research, to be very useful in extracting information from raw signals in various applications such as speech recognition, image recognition, etc. In our work, we use Spectrograms and MFCCs, as they are commonly used to represent speech features, along with CNNs for emotion detection.

**5.1.1. CNN Model with Spectrogram input**

A spectrogram is a representation of speech over time and frequency. 2D convolution filters help capture 2D feature maps in any given input. Such rich features cannot be extracted and applied when speech is converted to text and or phonemes. Spectrograms, which contain extra information not available in just text, gives us further capabilities in our attempts to improve emotion recognition.

The following model uses Mel-frequency Spectrogram as input to a 2D CNN. Spectro grams are generated when Short Term Fourier Transform (STFT) is applied on windowed audio or speech signals. The audio is sampled at 22050Hz. Windowing is then carried out on each audio frame using a “*hann*” window of length 2048. Fast Fourier Transform (FFT) windows of length 2048 are then applied on the said windowed audio samples with an STFT hop-length equal to 512. The obtained Spectrogram magnitudes are then mapped to the Mel-scale to get Mel-spectrograms. 128 Spectrogram coefficients per window are used in this model. The Mel-frequency scale puts emphasis on the lower end of the frequency spectrum over the higher ones, thus imitating the perceptual hearing capabilities of humans. We used the "librosa" python package, along with the above mentioned parameters, to compute the Mel-spectrograms. A sample Spectrogram corresponding to a sample audio is shown below in Fig. 5.1.1.1

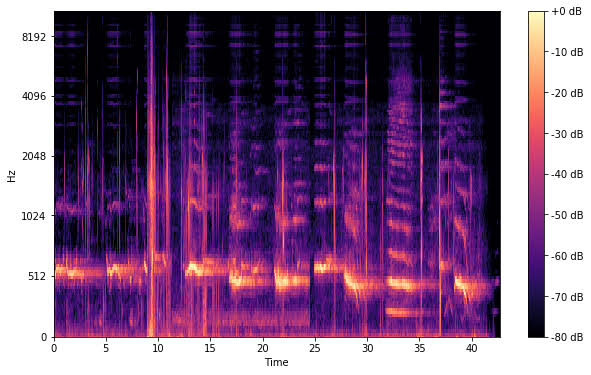


Fig. 5.1.1**.**1Spectrogram of a sample audio file

In our CNN model, we take Spectrogram input with a maximum image width of 256 (number of windows). Since our dataset has audios of varying lengths, we trim long duration audio files to a fixed duration (6 seconds), which covers 75 percentile of all audio data samples of the dataset. This decision was made under the assumption that the frequency variations that characterize the emotionality of the speech data will be present throughout the dialogue and hence will not be lost by this reduction in length. Fig. 4 details the 2D CNN architecture used to detect emotion using Spectrograms. A set of 4 parallel 2D convolutions are applied on the Spectrogram to extract its features. The input shape of the Spectrogram image is 128 x 256 (number of Mels x number of windows). 200 2D-kernels are used for each of the parallel convolution steps. Figuring out the optimal kernel size is a difficult and time taking task, which may depend on several factors all which cannot be clearly defined. To prevent choosing one single kernel size that could possibly be sub-optimal we decided to use kernels of different sizes, each of which is fixed for a single parallel path, to take advantage of the different patterns picked up by each kernel. The sizes of each of the kernels in their respective parallel CNN paths are 12 x 16, 18 x 24, 24 x 32, and 30 x 40.

The features generated in the said convolution layers are then fed to their respective max-pool layers, which extracts 4 features from each filter as the pool size is exactly half along the width and height of the convolution output. The extracted features are fed to the Fully Connected (FC) layer. This model makes use of two FC layers of sizes 400 and 200. Batch normalization is applied to both the FC layers. We experimented with dropout rates varying between 25% and 75% for the first FC layer but excluded it completely from the second FC layer. The activation function used in the convolutional layers and the first FC layer is the Rectified Linear Unit (ReLU). The output of the last FC layer is then fed to a Softmax layer, which classifies the input speech signal among 4 different emotion classes. “Adadelta'' is the optimization technique used during training.

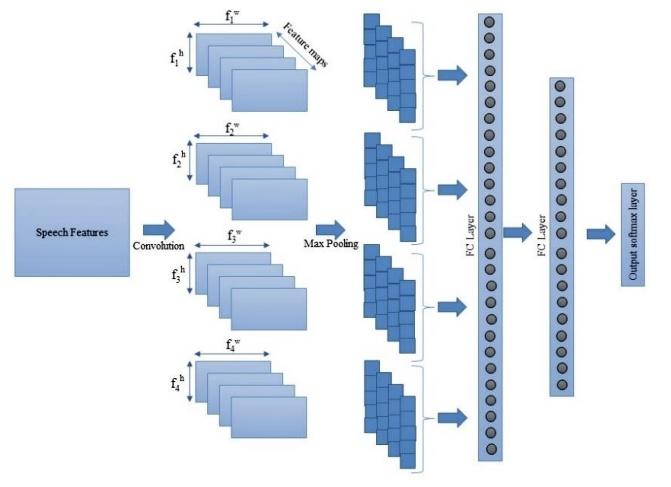


Fig. 5.1.1.2Spectrogram/MFCC based CNN model

We experimented with another variant of this model (refer Model 2B). The architecture of this model is quite similar to Model 2A but is augmented with four additional parallel convolution layers. Spectrograms, which have been down-sampled by 2 in both time and frequency, are taken as the input. The convolutional filters in this model are kept as is in Model 2A. This experiment was conducted in an attempt to extract even higher level features as compared to the ones in the above model.

**5.1.2 CNN Model with MFCC input**

Mel Frequency Cepstrum (MFC), refer Fig. 5, is a representation of the Short-Term Power Spectrum of sound. It is based on a linear cosine transform of a log power spectrum on a non-linear Mel-scale of frequency. As MFCC is a popular speech feature widely used in various speech processing applications, we decided to use the same in our experiments on emotion detection.

The hyper-parameters and the python package (librosa) used for MFCC generation are similar to the ones described for Spectrogram generation. The only difference is that 40 MFCCs per window are generated compared to the earlier mentioned 128 Spectrogram coefficients per window.

This model also consists of 4 sets of parallel convolutional layers, followed by max pooling layers and 2 more FC layers, similar to the one described in the previous section. As the input size is different to that of Model 2A and 2B we experimented with kernels of different sizes and eventually chose the set of 4 x 6, 6 x 8, 8 x 10 and 10 x 12 kernels for this model.

**5.1.3 Combined CNN Model based Spectrogram and MFCC models**

We experimented with 2 different combined models in an attempt to bring together different strengths offered by each of Spectrogram, MFCC and speech transcriptions. Since inputs are different and thus of different dimensions, we use separate CNN channels, as represented in below figure.

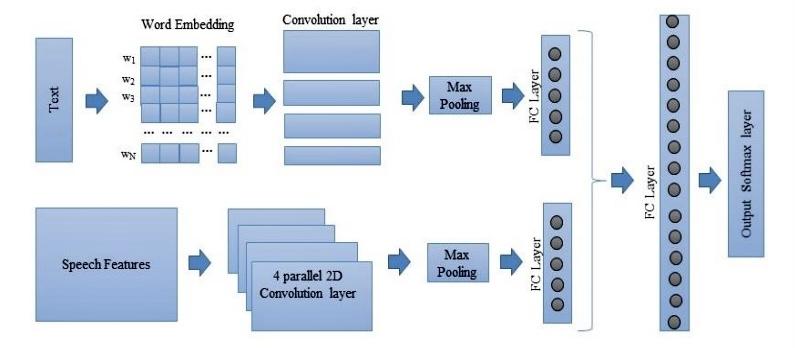


Fig. 5.1.3Representative CNN architecture

In this the Spectrogram channel consists of 4 parallel 2D-CNN layers with kernels of different sizes. Like the Spectrogram channel the MFCC channel also consists of 4 parallel 2D-CNN layers. Outputs from both the channels are fed to one FC layer each. The outputs of both the FC layers, after normalization, are concatenated and fed to the 2nd FC layer. The final step is to feed the outputs of the last FC layer to a softmax layer.

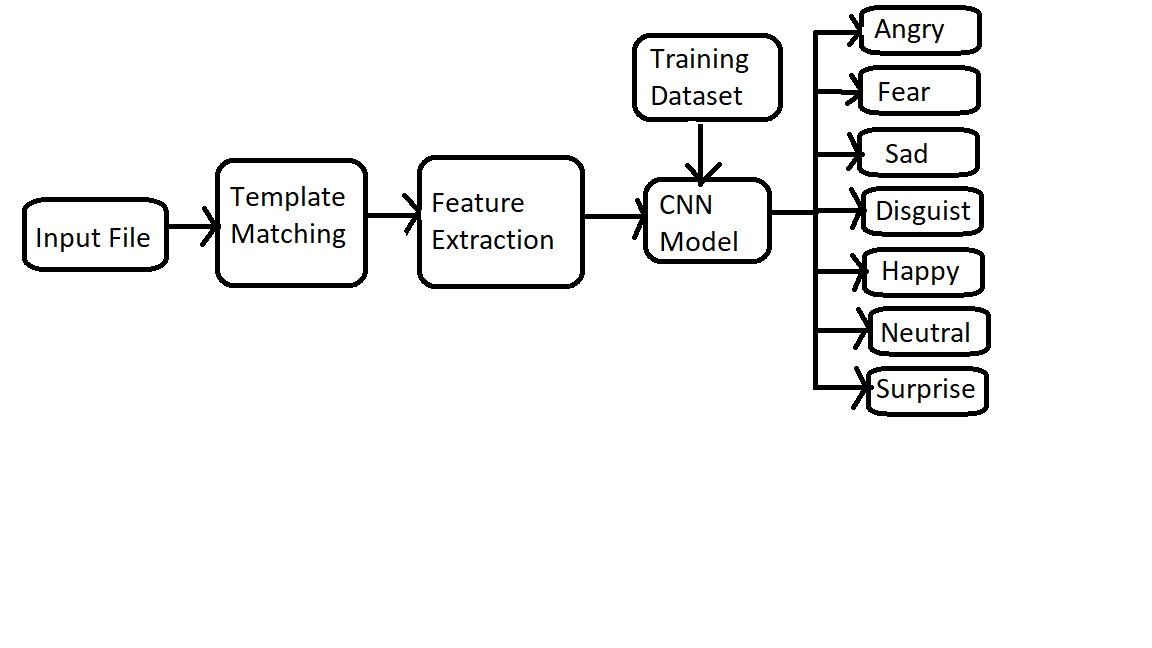
**5.2 Algorithm**

//Colab.google.com in Python language.

Step 1: The sample audio is provided as input. Step 2: The Spectrogram and Waveform is plotted from the audio file. Step 3: Using the LIBROSA, a python library we extract the MFCC (Mel Frequency Cepstral Coefficient) usually about 10–20.

//Processing software

Step 4: Remixing the data, dividing it in train and test and thereafter constructing a CNN model and its following layers to train the dataset. Step 5: Predicting the human voice emotion from that trained data (sample no. - actual value - predicted value)

Fig 5.2 Algorithm

**Template Matching**

When a pronunciation dictionary is not available and there are only a few samples per word, template matching (TM) seems to be the most suitable approach. A template is a collection of vectors of features1 repeating a particular pronunciation and can also be seen as the succession of frames. TM stores during training one or more reference templates per word. During the testing phase each new frame is compared with all of the reference frames and identifies the new utterance as being the word associated with the template with the smallest distance to the new sequence. In TM all information contained in the templates is kept and used to recognize the pronounced word, no a priori assumptions are made and a word can be identified by only a few samples. However, its generalization capabilities are weak and its performances are not as competitive as CNN-based approaches.

The TM approach is subject to the following drawbacks:

1. Separate templates for each word brings dependency with the lexicon size in opposition to smaller units like phones or phonemes.

2. Nonlinear time alignment is crucial (inevitably different speaking rates, even for the same speaker, we have different speaking rates for the same word).

3. Reliability determines the word boundaries.

**Feature Extraction**

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch, formant, and some spectrum features such as linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features. In this work, we have selected modulation spectral features and MFCC, to extract the emotional features.

Mel-frequency cepstrum coefficient (MFCC) is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the Mel-frequency scale. The discrete cosine transform (DCT) of the Mel log energies was estimated, and first 12 DC coefficients.

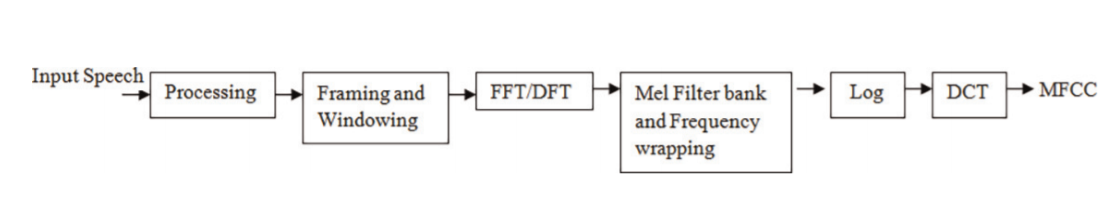


Fig 5.2.1 Schema of MFCC Extraction

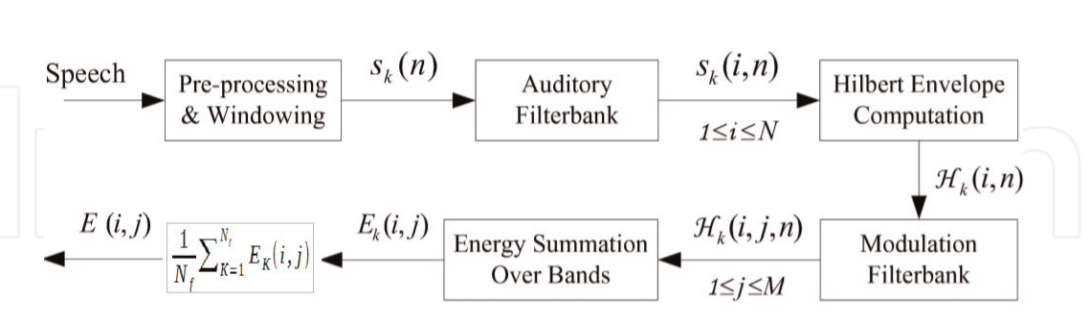


Fig 5.2.2 Process for computing the ST representation

MFCC values used in the classification process. Usually, the process of calculating MFCC is shown in Figure 5.2.1. In our research, we extract the first 12 orders of the MFCC coefficients where the speech signals are sampled at 16 KHz. For each order coefficient, we calculate the mean, variance, standard deviation, kurtosis, and skewness, and this is for the other all the frames of an utterance. Each MFCC feature vector is 60-dimensional. Modulation spectral features (MSFs) are extracted from an auditory-inspired long-term spectro-temporal representation. These features are obtained by emulating the spectro-temporal (ST) processing performed in the human auditory system and consider regular acoustic frequency jointly with modulation frequency. The steps for computing the ST representation are illustrated in Figure 5.2.2. In order to obtain the ST representation, the speech signal is first decomposed by an auditory filterbank (19 filters in total). The Hilbert envelopes of the critical-band outputs are computed to form the modulation signals. A modulation filterbank is further applied to the Hilbert envelopes to perform frequency analysis. The spectral contents of the modulation signals are referred to as modulation spectra, and the proposed features are thereby named modulation spectral features (MSFs). Lastly, the ST representation is formed by measuring the energy of the decomposed envelope signals, as a function of regular acoustic frequency and modulation frequency. The energy, taken over all frames in every spectral band, provides a feature. In our experiment, an auditory filterbank with N ¼ 19 filters and a modulation filterbank with M ¼ 5 filters are used. In total, 95 19 ð Þ 5 MSFs are calculated in this work from the ST representation.

**5.3 DataSet**

The complete dataset is shuffled and split in the ratio 0.8 as training and testing datasets.

Dataset samples : 2556

Training Samples : 2044

Testing Samples : 512

Each sample in the training dataset is labelled with an index for easy recognition and addressing.

* 0 : anger
* 1 : disgust
* 2 : fear
* 3 : happy
* 4 : neutral
* 5 : sad
* 6 : surprise

Each emotion (label) has some number of audio files as follows.

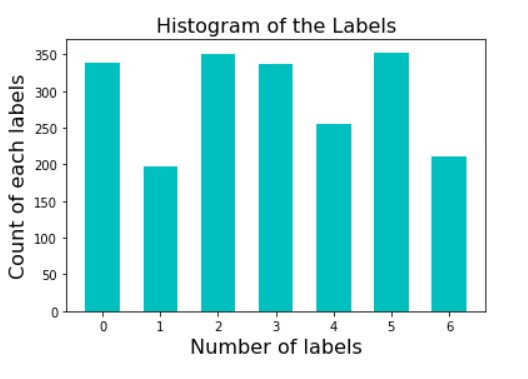
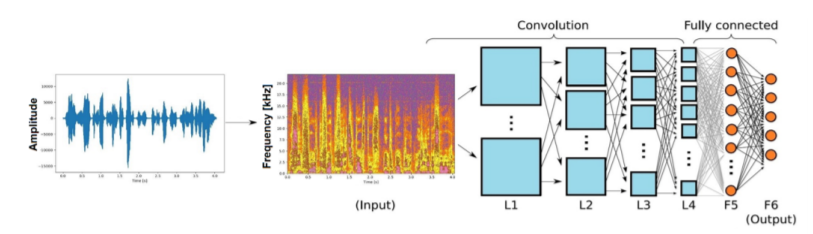


Fig 5.3 Histogram of Labels

**5.4 Modules**

Now we have the data loaded into the memory ready for modelling, we can define, fit, and evaluate a 1D CNN model.

Fig 5.4 Convolutional Neural Network

In our CNN model we have four important layers:

1. **Convolutional layer**: Identifies salient regions at intervals, length utterances that are variable and depicts the feature map sequence.

2. **Activation layer**: A non-linear Activation layer function is used as customary to the convolutional layer outputs. In this we have used a corrected linear unit (ReLU) during our work.

3. **Max Pooling layer**: This layer enables options with the maximum value to the Dense layers. It helps to keep the variable length inputs to a fixed sized feature array.

4. **Dense layer**

➢ Audio Feature Extraction and Visualizations. (module01)

Characteristics extraction is required for classification and depiction. The audio signal is a 3D signal in which 3 axes indicate time, amplitude and frequency.

We will use librosa to analyze and extract characteristics of any audio signal. (.load) function pulls an audio file and decrypts it into a 1D array which is of time series x, and SR is actually sampling rate of x. By default SR is 22 kHz. Here, I will show one audio file display with the use of (IPython.display) function. Librosa.display is important to represent the audio files in various forms i.e. wave plot, spectrogram and colormap.

Wave plots use loudness of the audio at a particular time. Spectrogram displays various frequencies for a particular time with its amplitude.

➢ To train the model for accuracy calculation. (module02)

Within this module we train the model for accuracy estimations. 1st, import necessary modules. Then pull the dataset. We will receive the sampling rate value with librosa packages and mfcc function. Thereafter this value holds other variables. Now audio files and mfcc value hold a variable consequently it will add a list. Then zip the list and hold two variables x & y. Then we have represented (x, y) shape values with the use of numpy package.

➢ Implementation process of CNN model. (module03)

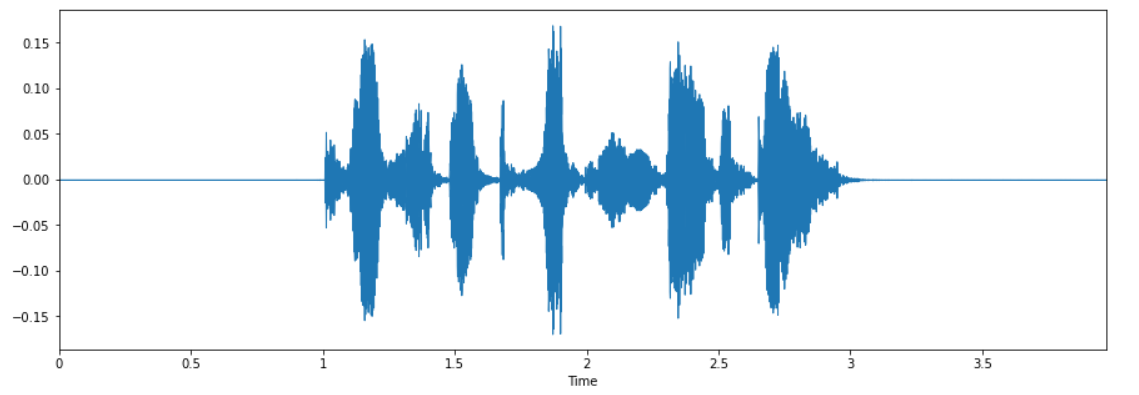
Speech represented in the form of an image with 3 layers. While using CNN, do consider 1st and 2nd derivatives of speech image with time and frequency. CNN can predict, analyze the speech data, CNN can learn from speeches and identify words or utterances.

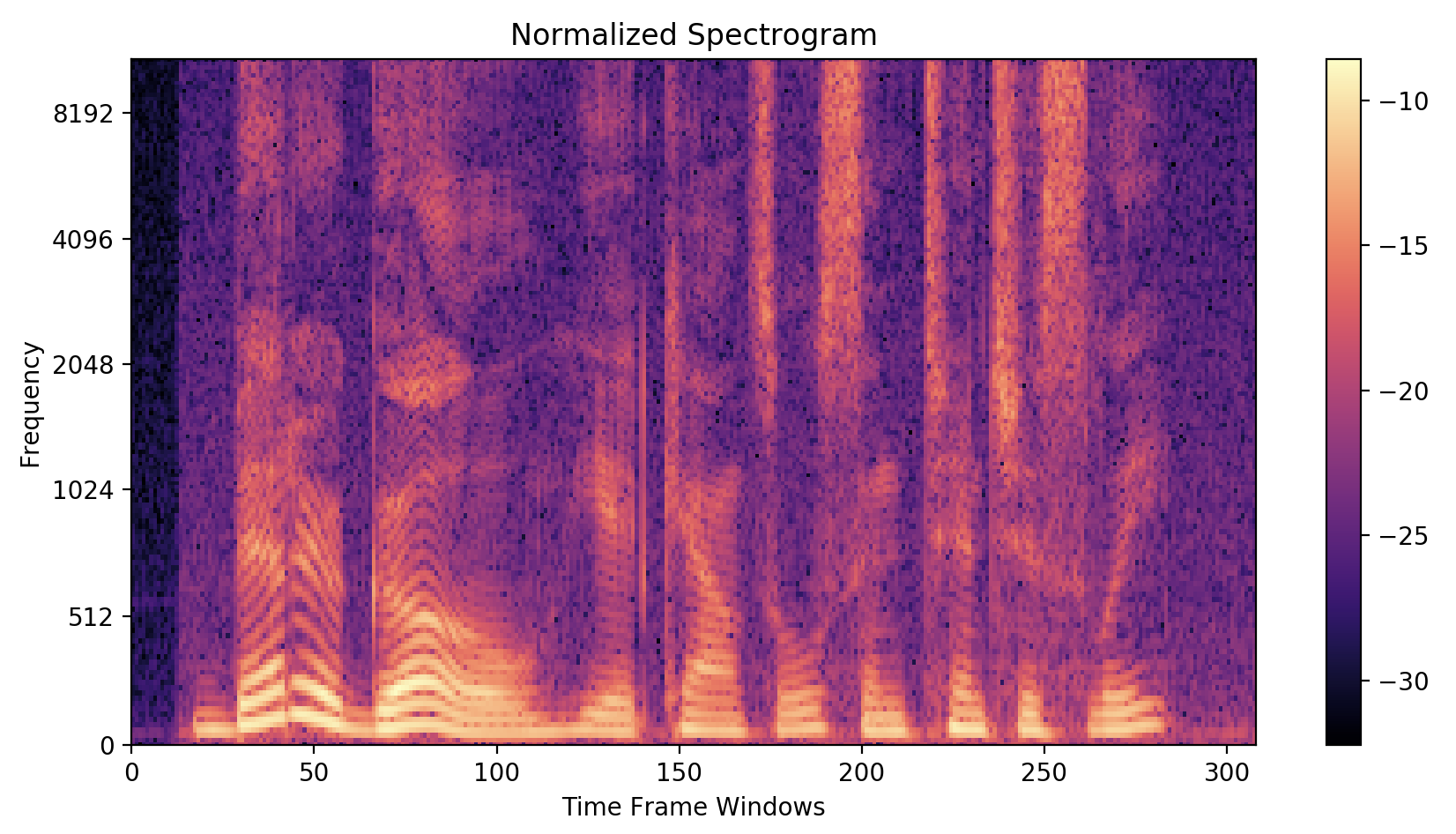
➢ Classification of speech emotions. (module04)

When testing we provide the audio input. Next, we run the audio in order to hear with ipython.display packages. Thereafter plot the audio features with librosa.display.waveplot packages. Extract the Characteristics using librosa.load. It converts one data frame and displays structured form. Further it compares loaded models by predicting function batch size 32. Ultimately it displays the output from the audio file what sort of expression/emotion that audio file has.

**6.0 RESULTS AND DISCUSSIONS**

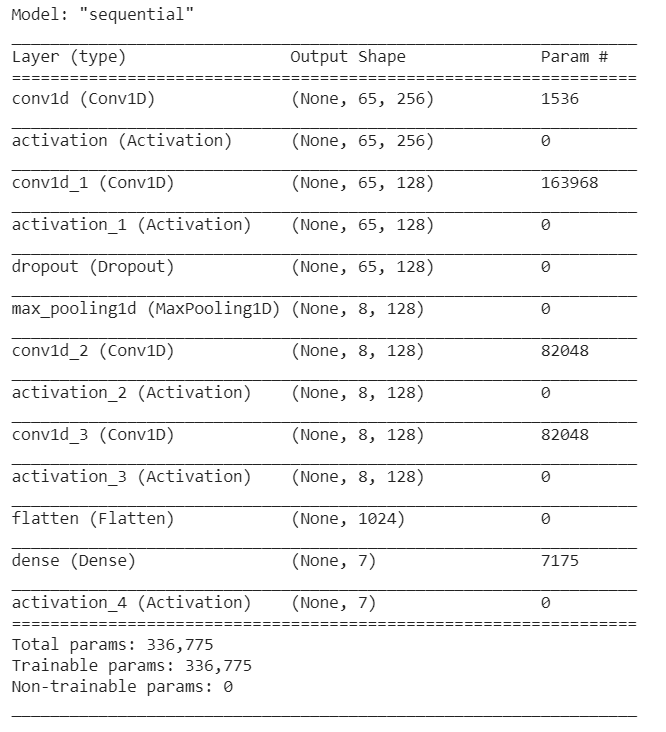
We experimented with an audio file to get its characteristics by plotting the waveform & spectrogram.

Fig 6.1 Time Domain Plot of the Speech Signal

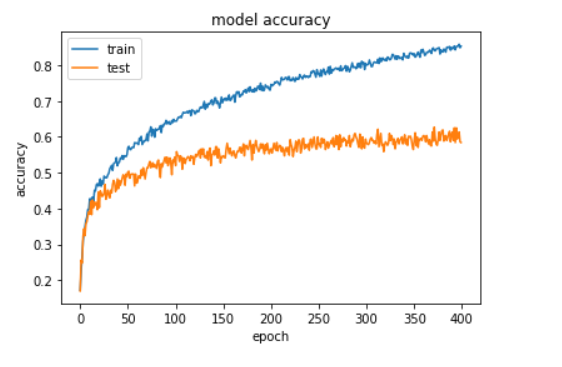
Fig 6.2 Frequency Domain Plot of the Speech Signal

After training various models it came out with the most optimum accuracy of 58% with SoftMax activation layer.

Table 6.3 Model Summary



The below figure shows the training and testing accuracy on our dataset. As the graph says that both “training and testing” accuracy increase as the number of epochs to the training model increases.

Fig 6.4 Model Accuracy

The below figure shows the training and testing loss on our dataset. As the graph says that both “training and testing” errors reduce as the number of epochs to the training model increases.

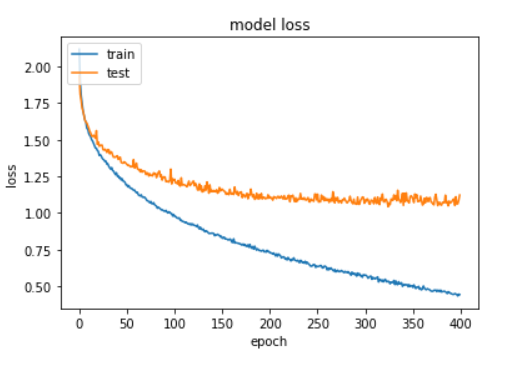
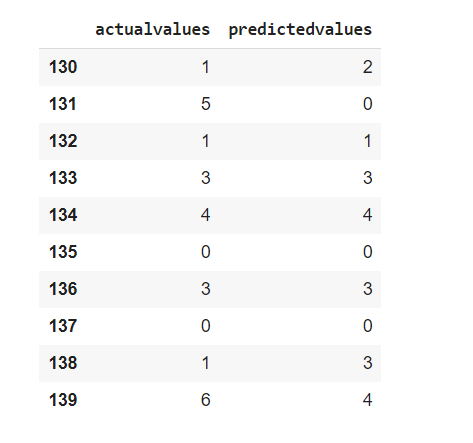


Fig 6.5 Model Loss

The following table displays our depiction with the actual values and the predicted values.

Table 6.6 Actual Values Vs Predicted Values



**7.0 ADVANTAGES, DISADVANTAGES AND APPLICATIONS**

**7.1 Advantages**

* During healthcare, determining patients feeling and comfort level about the treatment
* In the case of autism, struggling to interpret expressions
* In the case of elearning, study the emotions and adjust the learning technique and
* presentation according to the style of learner
* Determining fatigue in the case of driving and alerting in advance.
* Health monitoring of mentally unstable patients
* To ensure safety of people at critical times by sending alert to police when scared.
* Used in psychology studies.
* Used in robots for better understanding of human minds.

**7.2 Disadvantages**

* It cannot recognize while a person is pretending.
* Dataset is complex to construct on our own.
* Overlapping of utterances.
* Presence of background noise.
* If all the speakers know that they are being recorded,the quality will be artificial.

**7.3 Applications**

* **In-Car Systems**- to initiate phone calls, select radio stations or play music from a compatible smartphone.
* **Health Care**- medical documentation ,therapeutic use like radiologic techniques.
* **Military**- In high-performance fighter aircraft like setting audio frequencies, commanding an autopilot system.
* **Helicopters-** control of communication radios, setting of navigation systems, and control of an automated target handover system.
* **Telephony**-Used in contact centers by integrating it with IVR systems.
* **Education-** Helpful for students who are blind and physically disabled or suffer from repetitive strain injury/other injuries.
* **People with disabilities**- people with deaf or hard of hearing this can be used in conferences, classroom lectures.

**8. CONCLUSION AND FUTURE SCOPE**

**8.1 Conclusion**

The applications area for the emotion recognition function in the human voice includes many areas: social assistive robots, artificial brain, intelligent driving, autonomous vehicle, neurofeedback equipment, etc. In the current social context, more and more people are already addicted to intelligent devices, and many of them already have an increasing tendency to interact with intelligent devices the same way they do with other human beings. This is the reason why the intelligent and empathetic device market has a spectacular development in the last few years. Estimates for the coming years are quite daunting.

The convolutional neural network deep learning method for emotion detection from voice described in this is intended for hardware implementation and particular applications in companion robots and pet robots. That’s why we consider that using voice as a source of emotional information is appropriate and the seven basic emotions are a good starting point for a significant feedback based on and similar to human emotions.

We have used the Keras deep learning library and Python language for the implementation. In Speech Emotion Recognition, the CNN classifies the entries in 7 labels corresponding to the following emotions: happy, fear, sadness, disgust, anger, surprise, happy, disgust. The results obtained after training the network with a set of 2556 audio files.

**8.2 Future Scope**

The future scope in this system would be to design a mechanism that would be helpful in music therapy treatment and provide the music therapist that help needed to treat the patients suffering from disorders like mental stress, anxiety, acute depression and trauma.

As a real application, it could be considered a real-time system that can serve like a motor of emotional knowledge in order to understand the autistic children, to accurately describe their internal state and show the real content of their emotions. The system is not only applied to companion robots it could also be applicable to diverse smart sources (smart devices), this could be the case of healthcare, telemedicine or smart well-being systems that can be seen more often. This type of emotional devices working with emotional feedback will have the potential to reveal more about emotional state and the early detection of crisis, balanced lifestyle including and regulated stress level.

**REFERENCES**

[1]. A.Mehriban “Communication without words”, Psychology Today, cilt 2, no. 4, pp. 53-56. ,1968

[2]. B. Fasel ve J. Luettin, “Automatic Facial Expression Analysis: A Survey”, Pattern Recognition, cilt 36, pp. 259-275, 2003.

[3].Pathak, S. and Arun K., "Recognizing emotions from speech." Electronics Computer Technology (ICECT),3rd International Conference on. Vol. 4,2011

[4]. Rao K. S., Kumar T. P., Anusha K., Leela B., Bhavana I. And Gowtham S.V.S.K., “Emotion Recognition from Speech” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2), pages: 3603-3607,2012

[5].Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001.

[6].Utane, Akshay S., and S. L. Nalbalwar. "Emotion recognition through Speech." International Journal of Applied Information Systems (IJAIS) (2013): 5-8.

[7].Hadhami Aouani1, and Yassine Ben Ayed. “Speech Emotion Recognition with deep learning”, 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science 176 (2020) 251–260.

[8].Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, Promod Yenigalla, “Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions”,Samsung R&D Institute India – Bangalore.