SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be **University** *u/s 3 of UGC Act, 1956*

18CSE305J – ARTIFICAL INTELLIGENCE – APPLIED DATA SCIENCE WITH VENTURE APPLICATION

Speech Recognition

Submitted by

Triyan Akula-RA2011003010870

Team Members

Register Number	Name
RA2011003010870	Triyan Akula
RA2011030010721	Rishi Reddy Thokala
RA2011003010841	Sai

1.Introduction

Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, AND having it correctly recognize what you are saying.

The following definitions are the basics needed for understanding speech recognition technology.

Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

Speaker Dependance

Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

Vocabularies

Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g."Wake Up"), while very large vocabularies can have a hundred thousand or more!

Accuracy

The ability of a recognizer can be examined by measuring its accuracy - or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

Training

Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy.

1.1 Need of the Project

Although any task that involves interfacing with a computer can potentially use ASR, the following applications are the most common right now.

Dictation
Dictation is the most common use for ASR systems today. This includes medical transcriptions, legal and business dictation, as well as general word processing. In some cases, special vocabulary is used to increase the accuracy of the system.

Command and Control
ASR systems that are designed to perform functions and actions on the system are defined as Command-and-Control systems. Utterances like "Open Netscape" and "Start a new xterm" will do just that.

Telephony
Some PBX/Voice Mail systems allow callers to speak commands instead of pressing buttons to send specific tones.

Wearables
Because inputs are limited for wearable devices, speaking is a natural possibility.

Medical/Disabilities
Many people have difficulty typing due to physical limitations such as repetitive strain injuries (RSI), muscular dystrophy, and many others. For example, people with difficulty hearing could use a system connected to their telephone to convert the caller's speech to text.

Embedded Applications
Some newer cellular phones include C&C speech recognition that allows utterances such as "Call Home". This could be a major factor in the future of ASR and Linux.

1.2 Approach

- Acoustic Phonetic Approach: It is based on the sound waves made by human vocal organs while communicating which shows that a finite distinctive phonetic unit exists in spoken language. These phonetics units are categorized by set of properties of speech signal.

- Pattern Recognition Approach: In this method speech patterns are used directly. It has two steps. First training of speech pattern and second recognition of pattern through comparison. HMM, GMM are the most commonly used pattern recognition approach.

- Artificial Intelligence Approach: It is combination of acoustics phonetic approach and pattern recognition approach. DNN(Deep Neural Network), RNN(Recurrent Neural

1.3 Benefit

- It can help to increase productivity in many businesses, such as in healthcare industries.
- It can capture speech much faster than you can type
- You can use text-to-speech in real-time.
- The software can spell the same ability as any other writing tool.
- Helps those who have problems with speech or sight

1.4 Competition

Alternatives & Competitors to Speech Recognition API

- Krisp. (412)4.8 out of 5.
- Otter.ai. (104)4.1 out of 5.
- Deepgram. (79)4.5 out of 5.
- Microsoft Speaker Recognition API. (17)3.7 out of 5.
- Express Scribe. (30)4.4 out of 5.
- Microsoft Bing Speech API. (22)3.7 out of 5.
- Kaldi. (21)4.1 out of 5.
- Jasper. (17)4.1 out of 5.

1. Customer Validation

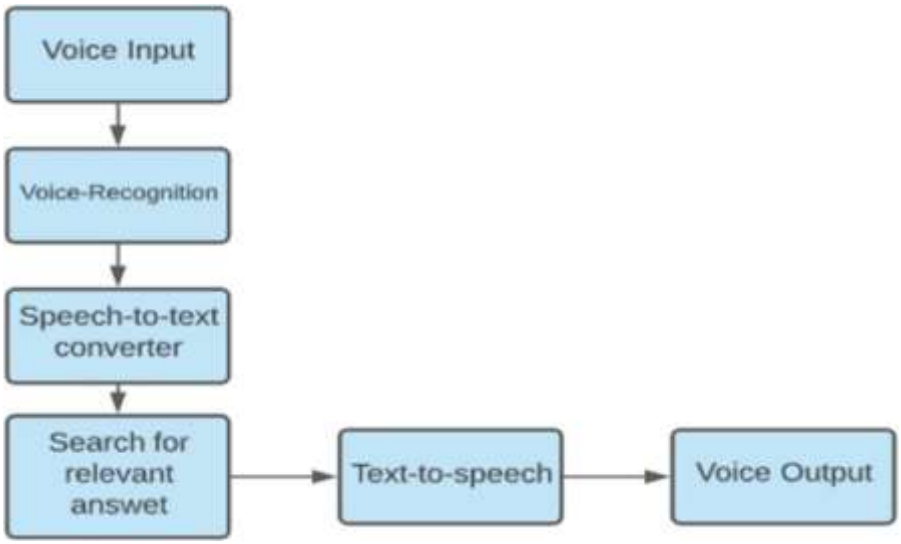
- Add the forms used for customer validation (Problem Sizing, Persona – Each Member should have different persona, Customer validation questions and answers)

- Describe the Outcomes of the customer validation

2. Project Description

Approaches of speech recognition Speech recognition approaches are categorized into three types as shown in Fig. 1. Acoustic Phonetic Approach: It is based on the sound waves made by human vocal organs while communicating which shows that a finite distinctive phonetic unit exists in spoken language. These phonetics units are categorized by the set of properties of speech signal. Pattern Recognition Approach: In this method speech patterns are used directly. It has two steps. First training of speech pattern and second recognition of pattern through comparison. HMM and GMM are the most used pattern recognition approach. Artificial Intelligence Approach: It is a combination of acoustics phonetic approach and pattern recognition approach. DNN (Deep Neural Network), RNN (Recurrent Neural Network) are examples of artificial intelligence approach). Based on Speaker Mode, it is classified into three types. Speaker Dependent: It works for only specific users, easy to develop, less expensive and produces more accuracy. Speaker Independent: It works for any kind of speaker, most difficult, most expensive and having less accuracy. Adaptive Speaker mode: It can adapt characteristics of new speaker. Accuracy improves gradually.

3.1 Illustrate the UI/ Input, output



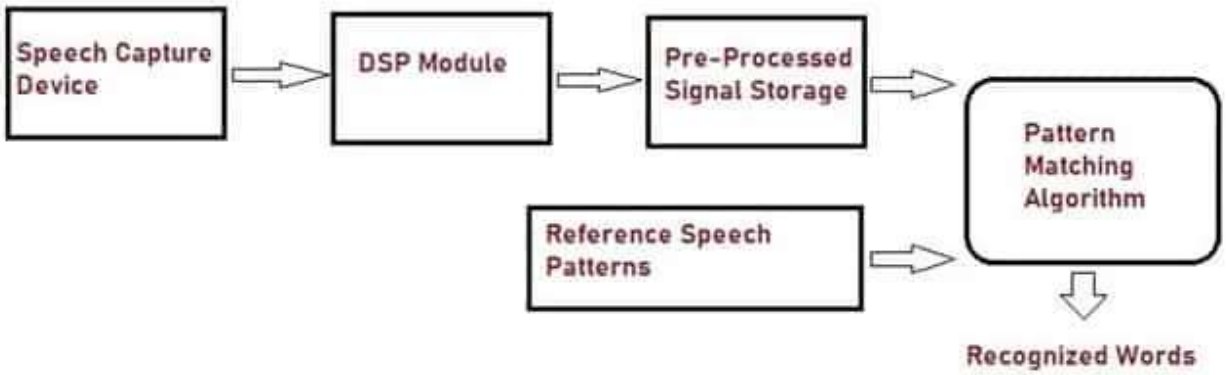
3.2 Technical Components of the project

- We will use librosa, a Python library for analyzing audio and music.
- We will be using Jupyterlab, it's an open-source, web based UI for Project
- Jupyter and it has all basic functionalities of the Jupyter Notebook.
- For this, we will use the RAVDESS dataset.
- RAVDESS dataset is the Ryerson Audio-Visual database of Emotional
- Speech and Song dataset and is free to download.
- This dataset has 7356 files rated by 247 individuals 10 times on
- emotional validity, intensity, and genuineness.
- The entire dataset is 24.8GB from 24 actors.

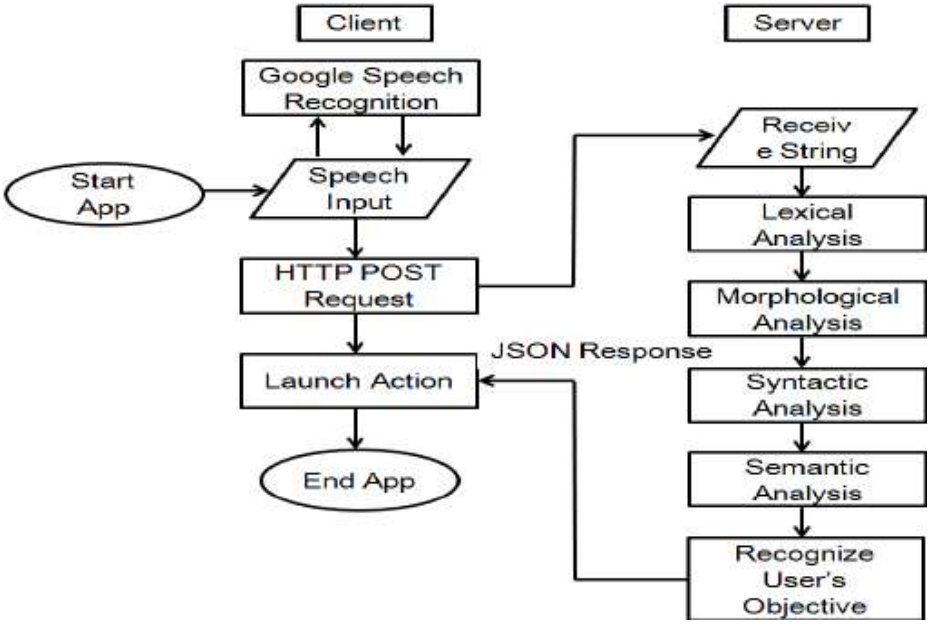
3.3 System Architecture

The architecture of the system consists of following modules:

- Speech Capturing Device
- Digital Signal Processor Module
- Pre-processed Signal Storage
- Reference Speech Patterns
- Pattern Matching Algorithm



3.4 Data Flow in the system



3. Business Plan

The solution utilizes an increasingly popular and accurate speech recognition component known as “telephony.” This solution will provide the orthopaedic surgeons with the ability to input patient data as they give the actual exam. This will fill a major gap not available in the current competitive product offerings.

The value proposition to the physician is the ability to see more patients and decrease their operating costs. This provides higher revenue and greater operating margins.

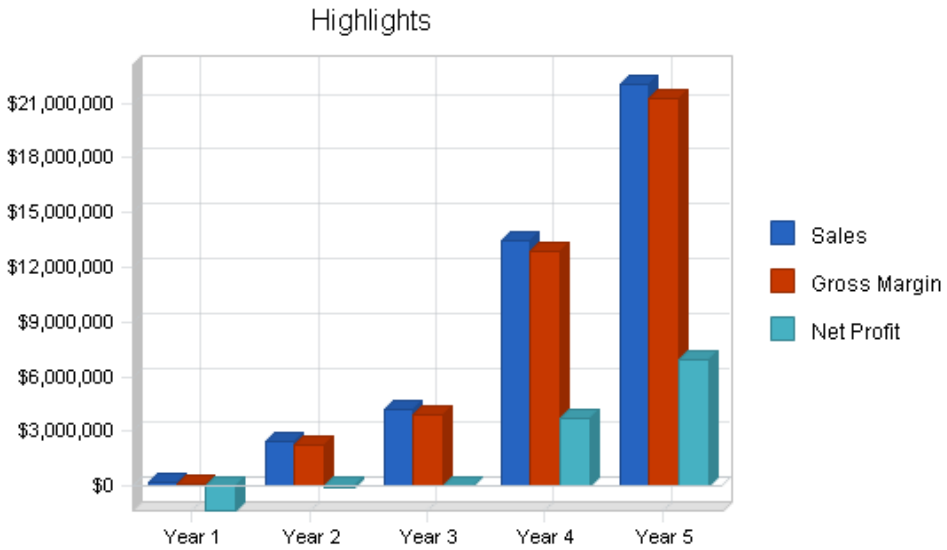
The company will be differentiated from its competitors because of offering the first turnkey solution for electronic medical records utilizing telephony.

Additional market research demonstrated that less than 10% of the industry currently utilizes electronic medical records. Further, 89% of the market plans to purchase electronic medical records within the next five years. Of that, 42% within 18 months, 36% within three years and 11% within five years.

Currently, most physicians collect patient information on paper. That information will include patient demographics, patient examinations, lab reports and referral physician information.

With approximately 15,000 licensed orthopaedist in the United States, this represents a mature market with the need and ability to pay for electronic medical records services. Our plan is to penetrate a minimum of 13% in this market by our fifth year.

The execution of this plan will require initial financing. The cash reserves in the fifth year will be substantial. This will provide the opportunity for the owners to begin buying back the shares of the company required to fund the start-up. The owners currently desire to buy back all outstanding shares during this year.



4.1 Key activities

Speech recognition technology and the use of digital assistants have moved quickly from our cell phones to our homes, and its application in industries such as business, banking, marketing, and healthcare is quickly becoming apparent.

1. In the workplace

Speech recognition technology in the workplace has evolved into incorporating simple tasks to increase efficiency, as well as beyond tasks that have traditionally needed humans, to be performed.

Examples of office tasks digital assistants are, or will be, able to perform:

- Search for reports or documents on your computer
- Create a graph or tables using data
- Dictate the information you want to be incorporated into a document
- Print documents on request
- Start video conferences
- Schedule meetings
- Record minutes
- Make travel arrangements

2. In banking

The aim of the banking and financial industry is for speech recognition to reduce friction for the customer.[8] Voice-activated banking could largely reduce the need for human customer service, and lower employee costs. A personalised banking assistant could in return boost customer satisfaction and loyalty.

How speech recognition could improve banking:

- Request information regarding your balance, transactions, and spending habits without having to open your cell phone
- Make payments
- Receive information about your transaction history

3. In marketing

Voice-search has the potential to add a new dimension to the way marketers reach their consumers. With the change in how people are going to be interacting with their devices, marketers should look for developing trends in user data and behaviour.

4.2 Key Resources

Speech recognizers are made up of a few components, such as the speech input, feature extraction, feature vectors, a decoder, and a word output. The decoder leverages acoustic models, a pronunciation dictionary, and language models to determine the appropriate output.

- Natural language processing (NLP): While NLP isn't necessarily a specific algorithm used in speech recognition, it is the area of artificial intelligence which focuses on the interaction between humans and machines through language through speech and text. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.
- Hidden markov models (HMM): Hidden Markov Models build on the Markov chain model, which stipulates that the probability of a given state hinges on the current state, not its prior states. While a Markov chain model is useful for observable events, such as text inputs, hidden markov models allow us to incorporate hidden events, such as part-of-speech tags, into a probabilistic model. They are utilized as sequence models within speech recognition, assigning labels to each unit—i.e. words, syllables, sentences, etc.—in the sequence. These labels create a mapping with the provided input, allowing it to determine the most appropriate label sequence.
- N-grams: This is the simplest type of language model (LM), which assigns probabilities to sentences or phrases. An N-gram is sequence of N-words. For example, "order the pizza" is a trigram or 3-gram and "please order the pizza" is a 4-gram. Grammar and the probability of certain word sequences are used to improve recognition and accuracy.
- Neural networks: Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold) and an output. If that output value exceeds a given threshold, it "fires" or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. While neural networks tend to be more accurate and can accept more data, this comes at a performance efficiency cost as they tend to be slower to train compared to traditional language models.
- Speaker Diarization (SD): Speaker diarization algorithms identify and segment speech by speaker identity. This helps programs better distinguish individuals in a conversation and is frequently applied at call centers distinguishing customers and sales agents.

4.3 Key Partners

<u>Software provider:</u>

1. **Anaconda**
2. **jupiter notebook**

<u>Resource provider:</u>

3. **Github**
4. **kaggle**

4.4 Value Propositions

4.5 Cost Structure

ITEMS	COST	QUANTITY	TOTAL
Computers	70,000	6	420,000
Software's	3,000	6	18,000
Hard disk	5,000	3	15,000

4.6 Revenue Streams

This speech recognition market is segmented into vehicle types, technologies, verticals, and regions. The type segment is bifurcated into speaker-dependent and speaker-independent. The technology segment is bifurcated into AI-based and non-AI-based. The vertical segment is bifurcated into military, automotive, healthcare, and others. The region segment is bifurcated into the Americas, Asia-Pacific, Middle East & Africa, Europe, and rest-of-the-world.

4.7 Customer Segment

1)Disablity:
Disable people have difficulty to interact with the world or understand that this speech recognition will be helpful for them.
2)Call Center:
Sometimes same type complaints will be occuring to make it easier we can use speech recognition to understand the customer problem and
come up with suitable solution.
3)AI Assitant:
AI makes people's work easier and assit them just by voice and make their work efficent.

4.8 Customer Relationship

Awareness – It is the first touchpoint where prospects try to know more about your brand as a whole.
Discovery – Then you learn and identify the needs of the prospects and share information to fulfill their requirements.
Evaluation – Moving ahead the prospects compare and evaluate your products/services with your competitors.
Intent – Finally your prospect is convinced and made a decision of buying from you.
Purchase – After making the payment the deal is done and the prospect converts into your customer.
Loyalty – Make a follow-up after purchase to determine customer success with your product and ask for referrals.

4. Financial Plan

- **We are going to make money through affiliate links and leads as users take suggested**
- **actions, sort of a cost-per-action model.**
- **By getting the subscription money from companies by selling our product for their**
- **industrial use.**

5.1 Growth Strategy

North America leads the global market for speech recognition technology. Growing implementations of speech recognition applications in cell phones escalate the market value. Besides, the expanding utilization of this technology for obtaining customer consent/ acknowledgment over speech/voice in mobile banking and buyers & IoT gadgets boost the speech recognition market size.

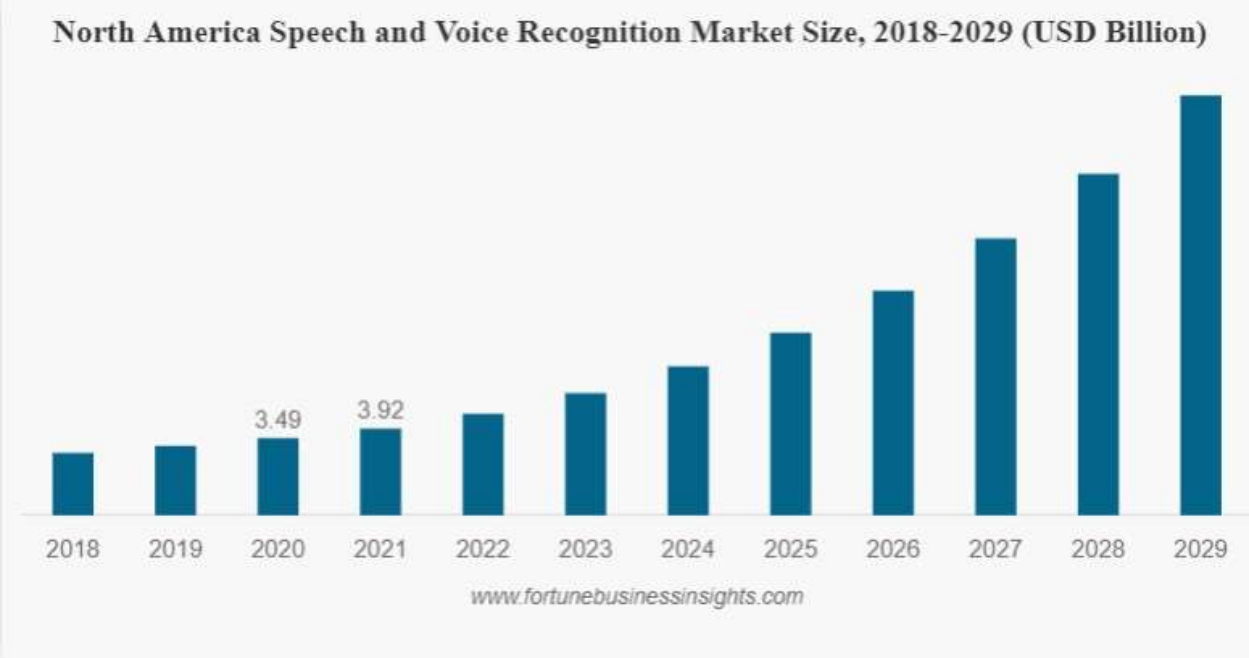
Rapid developments and rising applications of this technology substantiate market revenues. Moreover, the large presence of major technology providers and high spending on automotive infotainment systems increase the market sales of speech recognition.

Europe is another profitable market for speech recognition, headed by the rising interest in discourse and voice acknowledgment. Additionally, voice recognition tech observing significant advances and expanded applications in buyer gadgets and retail areas drives the speech recognition market demand. The vast automotive sector in the region creates substantial market demand, increasingly using patterns of associated devices in robots.

APAC is an emerging market for speech recognition systems. The growing number of developments and use of voice-empowered gadgets in the auto and medical businesses push the market growth. Furthermore, augmenting demand for standalone products of language recognition, speech-to-text, machine translation, and transliteration provides

significant market opportunities. China, Japan, and Singapore are major speech recognition markets supporting the growth of the regional market.

5.2 Traction



5.3 Financials

- Enlist and describe the various heads of income, expenditure
- Draw a table of Financials for three years

	2019	2020	2021	
Users	50,000	400,000	1,600,000	
Jobs	500,000	4,000,000	16,000,000	
Average price per job	75	80	90	
COMPANY REVENUE @15%	5,625,000	48,000,000	216,000,000	
- Cost of Revenue	2,000	7,000	10,000	
Gross Profit	5,625,000	48,000,000	216,000,000	
OPEX				
- Sales & Marketing	5,062,500	38,400,000	151,200,000	70%
- Customer Service	1,687,500	9,600,000	21,600,000	10%
- Product Development	562,500	2,400,000	10,800,000	5%
- Misc.	281,250	2,400,000	4,320,000	2%
TOTAL OPEX	7,596,750	52,800,000	187,920,000	
EBIT	-1,968,750	-4,800,000	28,080,000	13%

6.Conclusion

Through this project, we showed underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc. A few possible steps that can be implemented to make the models more robust and accurate are the following
An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.
Figuring out a way to clear random silence from the audio clip.
Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum.
Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.
Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.