VOICE FOCUS

Project Report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

By

Samson Shukla (161005)

under the guidance of

Mr. Alok Kumar

JAYPEE UNIVERSITY OF INFORMTION TECHNOLOGY, WAKNAGHAT

May 2020

TABLE OF CONTENTS

DECLARATION

I hereby declare that the work reported in the B.Tech Project Report entitled **“ Voice Focus ”** submitted at **Jaypee University of Information Technology, Waknaghat, India** is an authentic record of my work carried out under the supervision of **Mr. Alok Kumar**.I have not submitted this work elsewhere for any other degree or diploma.

Samson Shukla

Samson Shukla

161005

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Alok Kumar

Mr. Alok Kumar

Date:

Head of the Department/Project Coordinator

ACKNOWLEDMENT

I would like to thank every person that has conferred to the development of this project, which is the final stage of my Bachelor Education in Electronics and Communication Engineering at Jaypee University of Information Technology, Waknaghat, Solan.

I want to thank my supervisor Mr. Alok Kumar, for his advice and supervision during the development of the project. I would also like to thank Mr. Kamlesh for helping me in laboratory as well and most importantly I would like to thank my parents who provided me with the opportunity to study in this prestigious university and enlightened my life and career.

Samson Shukla(161005)

LIST OF ACRONYMS AND ABBREVIATIONS

ASR	-	Automatic Speech Recognition
AI	-	Artificial Intelligence
ML	-	Machine Learning
HD	-	High Definition
DSP	-	Digital Signal Processing
RL	-	Reinforcement Learning
IDLE	-	Integrated Development Environment
GUI	-	Graphical User Interface
SVM	-	Support Vector Machine
GMM	-	Gaussian mixture model
MASS	-	Multichannel Audio Source Separation
LGM	-	Gaussian Signal Model
NMF	-	Non-negative Matrix Factorization
MIR	-	Music Information Retrieval
FT	-	Fourier Transform
CASA	-	Computational Auditory Stream Analysis
ICA	-	Independent Component Analysis
NMPCF	-	Non-Negative Matrix Partial Co-Factorization
AV	-	Audio-Visual
AO	-	Audio-Only

LIST OF FIGURES

ABSTRACT

In daily life, human speech is regularly contaminated by other interfering speakers, environmental noise andreverberation. Although humans have a tendency to isolate noise and focus on preferred audio signals which they want to listen to (this is called “Cocktail Party Effect”) but some people lose this ability due to hearing disorders. So, making machine mimic this technique is what this project is about.

Audio Source Separation & speech enhancement is the field of science which deals with these problems. Many methods have been devised in this field concerning Cocktail Party but only a few have provided any significant results. This is primarily due to approach towards tackling this problem. Most of the researchers dealt with this problem considering it to be audio only which made the problem extremely difficult to solve. In this project I’ve mentioned a new approach - Audio-Visual method which gives significantly better results than many Audio-only techniques and provides nearly state of the art results.

There are numerous applications in this field. In fact, Audio Source separation and speech enhancement techniques are building blocks to many cutting edge technologies like robust remote-microphone, Automatic Speech Recognition (in Google and Microsoft virtual apps), Virtual private assistants (like Google Assistant or Cortana), car navigation systems (Google Maps), televisions, video game consoles (VR headset like Oculus), scientific dictation gadgets and meeting transcription devices (Microsoft Dictate).

In this project, I’ve mentioned about existing audio source separation techniques which give decent results and then compared them with new Audio-Visual approach which is independent of the speaker unlike other Audio-Visual methods which work only on specific users.

CHAPTER – 1

INTRODUCTION

1.1An insight of the overlooked

Human brain has the tendency to recognize a wide variety of sounds and voices effortlessly, whether it is dog's bark, glass breaking or any friend calling. Also, we have the ability to filter out sounds we want to hear and just focus on them. This is called Cocktail Party Effect. Naturally we possess some highly useful but often overlooked powers. But if just we could enhance our this ability or maybe use it as an inspiration and come out with some AI (Artificial Intelligence) algorithm to help people who have some kind of hearing problem, that would be something special.

As the technology advances, we are seeing significant work being done in the field of Image Processing using ML (Machine Learning). Google's Lens is one of the examples. But field of sound is often neglected due to its unstructured and complicated nature. Big tech giants like Google, Amazon and Apple are doing some work on their virtual assistants but that still lacks security and is not very reliable. Some other person with similar voice features can interact with your virtual assistant very easily and steal your data due to complexity of your voice and similarity with that other person's voice.

Consider a scenario. We all love to go to parties, have a nice chit-chat with our friends and enjoy ourselves, but you must have encountered an obvious problem. With the loud music playing in the background and others talking out loud it's hard to understand what the person you are trying to talk to is telling you and vice versa. In short it's hard to have a nice conversation in the parties or wedding ceremonies. This kind of problem is called Cocktail Party Problem (the problem of perceiving speech in noisy social settings).

Now question arises, is there any kind of gadget which can solve this problem?

Well the answer is, not yet. But there are theories and researches being done in the audio field which can definitely lead us to the ultimate solution (a device which can help us listen better). This project is based on this very concept of improving a person's hearing power by devising a method so that the device can focus on the target speaker and separate its audible speech and suppressing the background sounds (or noise).

1.2 Motivation

Real world speech signals are often contaminated by interference of other people's voice signals or noise from the surroundings. So, it gets difficult for the people with hearing problems where the patient is unable to retain its attention towards the speaker. Human minds can concentrate their auditory attention, i.e. an effect of selective attention in the brain, on a particular stimulus while filtering out a collection of other stimuli like a person can focus on a single chat in a noisy room, for example in a party. The phenomenon is Cocktail Party effect. But the problem is, it is hard to successfully perform this when surroundings are too noisy and full of various frequencies.

Consider a real life problem we all face during a marriage ceremony. We tend to fail in listening to the person talking to us due to very loud music, especially in Indian marriages. Take another example of sports matches like FIFA cup. We all would love to listen to the arguments of the players with their coaches and other fellow team players. Cristiano Ronaldo leading his team against France, fighting against the cheaters or Zidane supporting his team from the bench. Another application can be in military or to spies who can listen to distant important conversations of the enemies. There are innumerable applications of this simple idea of enhancing our abilities of Selective Auditory Attention which we all naturally possess.

We all have used portrait effect, a.k.a bokeh effect on our smartphones during taking beautiful portraits. The idea of this project is to devise an analogue of this effect in sounds. Automatic Speech Recognition systems can be improved using this terminology, personal assistants like Siri or Alexa can be improved, Cinematic user experience can improved, etc.

1.3 Existing Technology

Big tech giants like Apple, Samsung, Google and Amazon have dived in the sound processing field. Their significant products are their personal assistant applications like Apple's Siri, Google Assistant, Amazon Alexa and Samsung's Bixby but no one, successfully, has launched any product related to Audio Focus technology. Although, Samsung and Nokia (with its OZO program) have R&D teams working on this since a long time. Parts of this approach can be seen in some of their products (mostly smartphones) with Samsung's Audio Zoom on it's flagship smartphone Samsung Galaxy Note 10 and Nokia's OZO Audio in Oppo Reno 10X Zoom.

a. Samsung Audio Zoom

The camera in Samsung Galaxy Note 10 comes with a form of a mic that's capable of zooming in on sounds to hone in on the correct audio when you're recording video with the phone.
Essentially the mic is able to use the sound it picks up and pair it with the subject of the video that's currently in frame to determine which audio to focus on. It amplifies this audio and then works to drown out any noise in the background that isn't in frame so your video only has the sound you want.

b. Nokia's OZO

OZO by Nokia is an advanced camera which is able to record stereoscopic (3-Dimensional) 360° video. This device is made up of Aluminium (Al) alloy & contains 8 lenses & microphones which combined records stereoscopic (3D) 360-degree video & audio. Some features of this OZO technology is Audio 3D, Audio Zoom, Audio Focus and Audio Windscreen.

- 3-Dimensional Audio technology captures & delivers a natural spatial sound experience & degree of accuracy in high fidelity-audio is nearly 1. Depth, direction and detail can be captured with complete richness using Audio 3D.

- “Audio Focus” allows users to isolate (or suppress) & dynamically track an audio source while removing all the background noise. Some of the advanced features for users is that they can adjust its elevation angle or rotate area of focus for mesmerizing audio playback.

- Audio Zoom is the feature which allows users to dynamically select or focus and amplify audio while in a zoomed video. So this is a zooming lens for audio that lets you get up close to the action.

- Audio Windscreen is yet another feature which lets users capture crystal-clear audio quality which are far from the ideal setting. Sophisticated wind recognition algorithms identify & automatically adjust suppression processing to suit the wind noise conditions. Therefore providing clear audio, even when the circumstances aren’t.

This gadget was announced in July 2015 & first released in November 2015.

The main purpose of this device to improve audio experience of consumers in movie theatres by providing Ultra HD ambient surround sound.

1.4 Objective

The idea is to come out with a final solution, to build a software for a gadget capable of successfully implementing the artificial Selective Auditory Attention. Although, this would be a tedious task because of the complexities in the calculations needed for successfully implementing the idea.

For this software, the programming language we'll use is Python as it is easy to program in, it supports many no. of libraries and it fits to be a perfect language for some tedious calculations. Audio samples taken into account for selected audio source separation can initially be taken as pre-recorded parts and later on can be recorded in real time.

The general approach to implement this project is first, by understanding the pros and cons of every method that has been devised to achieve something similar to this, like in music classification, and music source separation. Those methods will form a basic layout of what we want to achieve. Then, we'll see if audio only method gives good enough results or we need to use Visual as second attribute to get better results as claimed by new studies.

1.5 Audio Processing

Widely speaking, the domain related to the study of signals, various operations and generation of signals (visual or audible) is Signal Processing. Practically speaking, an application related field is Digital Signal Processing or DSP.

DSP is defined as the mathematics, algorithms and the techniques used to operate ondigitally converted signals. It has wide number of uses. For example, speech generation, speech recognition, enhancing pictures, data compression, data storage, etc.

The sub field of DSP concerned with audio related stuff is Audio processing. It covers countlessof the miscellaneous fields, all concerned in presenting sound to theaddressees. 3 prominent areaof audio processing are:

(1) Highlyreliable Music Reproduction, for example, in audio CDs,
(2) Voice telecoms, also known as telephone networks, &
(3) Synthetic Speech:Here machines like computer recognize, learn and generate human like voise patterns.

While these implementationshave diverseproblems & goals, they are correlated by a generalsomething common i.e the human ear.

Music

Musician does incredible work, leading soothing audible music from the musician's microphone to the audiophiles'loudspeaker. It's very important for the musician to prevent degradation of stored music content by representing the data digitally. Someone having significant experience and knowledge in this field will be familiar with the difference in musical quality of cassete tapes & compact disks. A melodic piece is commonly recorded in a sound account studio on a few channels or tracks. Now and again, this suggests in any event, recording each instrument and vocalists independently. This gives the sound architect better adaptability in making of the last item. The way toward consolidating singular tracks into a last item is complex and is known as a mixdown. Digital Signal Processing providesvarious important functions at the time of mix-down, such as signal addition & subtraction, filtering, signal editing, and much more.

Amongst some of the veryfascinating DSP functions in music preparation is Artificial Reverberation. The resulting piece sounds frail and dilute, much as if the musicians were playing outdoors when the individual channels are simply added together. This happens because the listeners getdeeplyinclined by the resonance or reverberation part of the music, which is generallyreduced in the sound studio. Digital Signal Processing allows synthetic echoes &Reverberation to be added during mix-upfor simulation of various perfecthearing environments. Echoes with few hundred ms gapsgive the impression of cathedral like areas. Adding some echoes with delays of ten - twentyms provide the illusion of more diffident size listening rooms.

Speech generation

Speech production&speech acknowledgment are used for communicationamong human beings and machines. Instead of using our hands & eyes, we make use ofour mouth & ears. This becomes very suitable when our hands & eyes are busyinsomewhat else, like : steering a car, performing surgical operation, or (unfortunately) firing our weapons at the rivals. There are 2proceduresthat can be used for PC generated speech:1. Digital Recording &2. Vocal Tract Simulation. Voice of a person is digitized & then stored, usually in a compacted form in digital recording. The stored data are uncompacted& thentransformed back into an analog signal during playback. A complete hour of recorded verbal communication requires only around3 MB(megabytes) of storage, well within the capability of even small PCsystem. This is the very primefamiliar method of digital speech synthesis used these days.

Vocal Tract Simulators try to imitate the physical mechanism by which us, human beings create speech. The Human Vocal Tract is a auditory cavity with reverberate frequencies known by its size and state of the chambers. Sound starts in the vocal tract in 1 of 2 different ways, first being voiced and the other being fricative sounds. Vocal rope throb creates some close to occasional beats of air into the verbal cavities with verbal sounds. When thought about, fricative sounds start off from the loud air choppiness at thin tightening influences, as in the teeth and lips. Vocal Tract Simulators work by generating digital signals that look a lot like these 2 types of excitation. Thedescription of the resonationchamber are replicated by allowing the excitation signal throughout a Digital Filter with similar resonances. This approach was in usein one of the very early Digital Signal Processing success stories “ the Speak & Spell “, a broadly sold electronic education aid for kids.

Speech recognition

The automatic recognition of human being’sverbal communicationis to a great extent more complicated than speech generation. Speech recognition is a standard example of things that the human being’s brain does very well.Although, computerized PCs do this badly. Computerized PCs can easily store and review huge measures of information, perform numerous scientific figurings at extraordinary rates and do repeating errands easily without getting exhausted hardened or inadequate. Sadly, the present PCs' complete calculation with crude tactile information deficiently. Training a computer to forwardus a monthly electric bill is very straightforward whereas instructing the same computer to recognize our voice is foremost undertaking.

DSPcommonly approaches the problem of voice identificationin 2 steps: 1. feature extraction and then 2. feature matching. Each and every word in the received audio signal is secluded& then analyzed to recognize the type of excitation &resonate frequencies. These specifications are then comparedwords to identify the closest match with previous examples of spoken words. So frequently these frameworks are constrained to just a couple of 100s words and can just acknowledge discourse with unmistakable stops between words. Likewise, must be retrained for each individual speaker. While this may appear to be sufficient for some industrial applications, these constraints are lowering when contrasted with the capacities of person's hearing. There is a lot of work to be done here. With enormous monetary awards for the individuals who produce effective plug items.

1.6 Audio Processing Features

Sound being way more complicated than image has a large number of features which distinguishes one sound signal from the other.

Main features of sound processing taken into consideration are :

1. **Loudness**

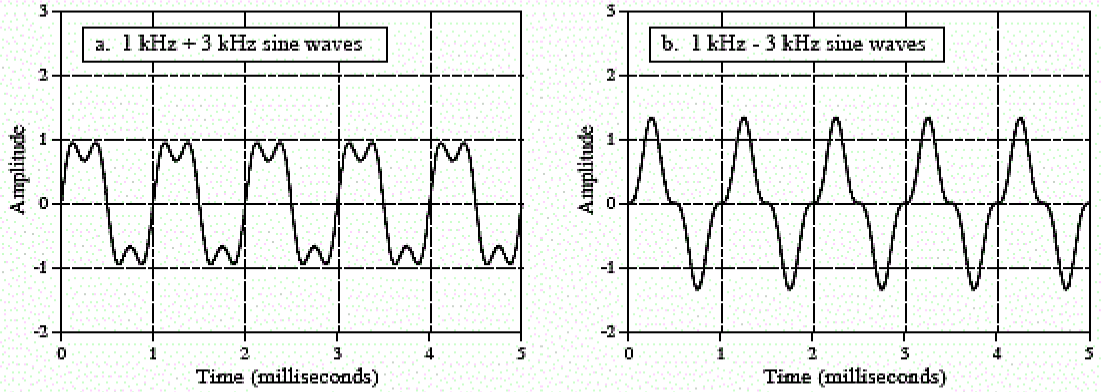
   In acoustics, loudnessis atrait of sound which determines the power of auditory sensation produced. More properlyit maybe defined as "That attribute of auditory sensation in terms of sounds which can be ordered on a scale increasing from quiet to loud." The intensity of sound, as professed by human ears,is almostcomparative to the log (logarithmic operation) of sound strength.When the poweris too small, the sound is not perceptible and when it is too great, it becomes hurting&unsafe to the ear. The sound intensity that our ear can endureis around 1012 times greater than the amount that is just perceptible. This range varies with the frequency of the sound from person to person.

2. **Pitch**

   Pitch is anintuitivefeature of sounds which allows their order on a frequency- related scale or more frequently, Pitch is the featurewhich makes it possible for a system to judge sounds as "higher" &"lower" in the sense related with musical melodies. Pitch can be intented only in sounds that have a frequency whichis clear &steady enough to discriminate from noise. Pitch is a chief auditory characteristic of musical tones, alongside with duration, loudness& timbre.

3. **Timbre**

   Timbre , a.k.atone shading or tone quality (from psychoacoustics) is the apparent sound nature of a melodic note, sound or any tone [4]. Timbre separates various kinds of sound creation, for example, ensemble voices and instruments, like, string instruments, wind instruments, and percussion instruments. Likewise it empowers audience members to separate different various instruments in a similar class (for e.g, an oboe and a clarinet are both Wood Wind instruments).



Phase detection of the human ear. The human ear is very insensitive to the relative phase of the component sinusoids. For example, these two waveforms would sound identical, because the *amplitudes* of their components are the same, even though their relative *phases* are different.

Figure 1.1

Timbre is far more complicated as it is being resoluted by the symphonious content of the signal. Following figure represents2 waveformseach one formed by totalling a 1 kHz sine signal with amplitude of 1 to 3 kHz sine wave with an magnitude of 1/2. The distinctionamong the 2 waveforms is that the one shown in (Fig b) has the greater frequency reversed before the calculation. Putting ina different way, the 3rd harmonic (3 kHz) is phase shifted by 180 degrees as in comparision to the 1st harmonic (1 kHz). In spite of very unlike time domain waveforms, these 2 signals sound same. This happens in light of the fact that perceptible range depends on the extent of the frequencies and is truly apathetic to their part. The state of the time space is just in a roundabout way identified with hearing and as a rule isn't considered in sound framework.

1.7 Machine Learning for Audio Processing

Arthur Samuel wasone of thefounders of artificial intelligence. Hedevised the term “Machine Learning”. He phrased machine learning as – “Field of study that gives computers the capability to learn without being explicitly programmed”.

ML is the field of learning which gives digital computers the potential to be trained without being externally programmed. ML is one of the most thrilling technology that someone would ever have come across. As it is clear from the name ML gives the computer the ability which makes it more analogous to human beings .[2]. Machine learning (ML) is widely being used today, perhaps in many more places than one imagine.

In a layman terms, Machine Learning(ML) can be simplified as automating &recuperating the knowledge process of computers based on their experiencefrom past exclusive of being actually programmed andexclusive of any human interference or help. The first procedure starts with feeding good featured data & then training the machines (computers) by constructing ML models, using the data &various algos. The variety of algorithms majorly depend on thekind of data we have and thetype of task we are trying to automate.

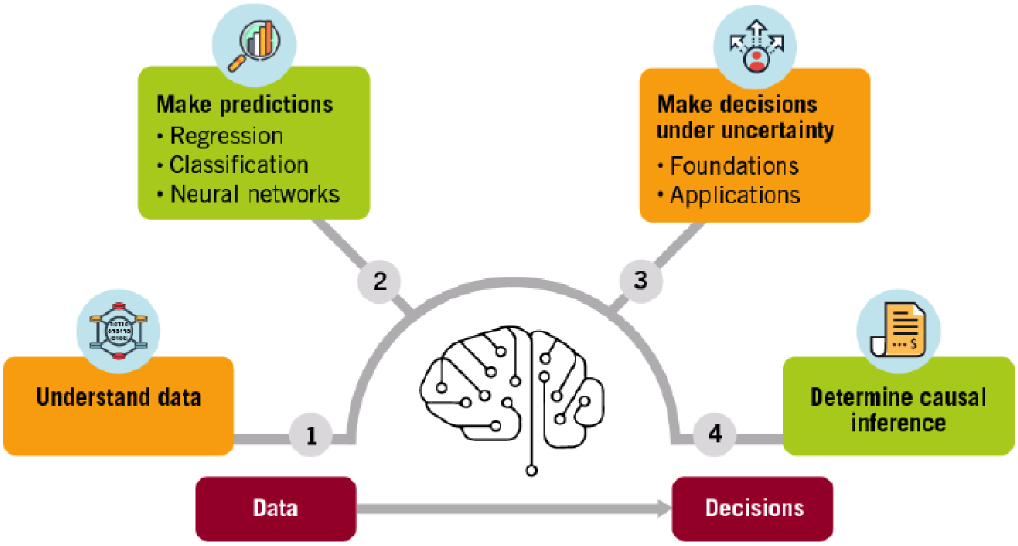


Figure 1.2

○ Distinction between Machine Learning&Traditional Programming

Traditional Programming : DATA (Input) + PROGRAM (logic) is fed, run on machine and we get the output.

Machine Learning :DATA(Input) + Output is fed, run on the machine and model is trained, thenthe machine creates its own program(logic), which can be evaluated in testing process.

○ How ML works?

• The model gathers past data in any form suitable for processing. Better the quality of data, improved and more apt it will be for modelling.

• Data Processing : Many times the data composed is in the raw form & it is required to be pre – processed before working on.

Example: Few tuples might have some missing values for somenumber of attributes, so, in this situationit has to be crammed with suitable values so as to execute machine learning (ML) or any other form of data mining techniques.

Missing data points for numerical featureslike the price of the house can be restored with mean value of the features whereas misplaced values for categorical featurescan be replaced with the featureshaving the highest mode. This invariablemostlydepends on the kinds of filters we usein our model. If the data is in the type of text or maybe pictures, then transformingit to numerical type will be vital.Eitherit is a list or an array or a matrix. Simply, data is to be preparedsignificant, useful, reliable& consistent. It is to be transformedinto a configuration which is clear by the machine.

• Partition the input data as training,cross-validation & test sets. The ratio of respective sets should be 6:2:2.

• Constructing models with right algos& techniques on the training setis required.

• Evaluatingthisconceived model with the data that was not feeded to the model when training & thentestingits performance using F1 score, precision & recall.

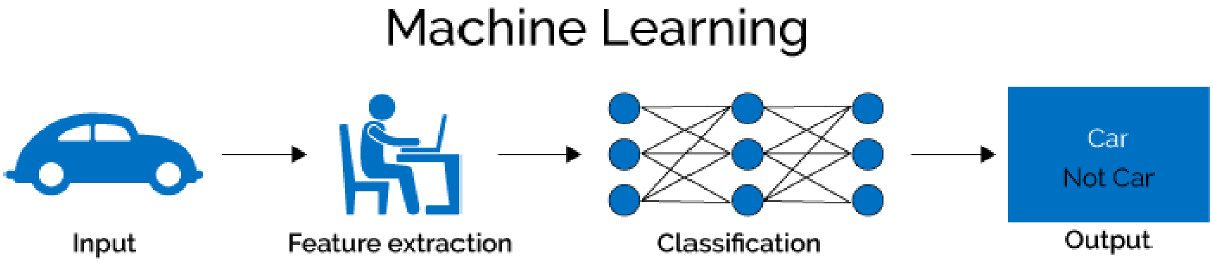
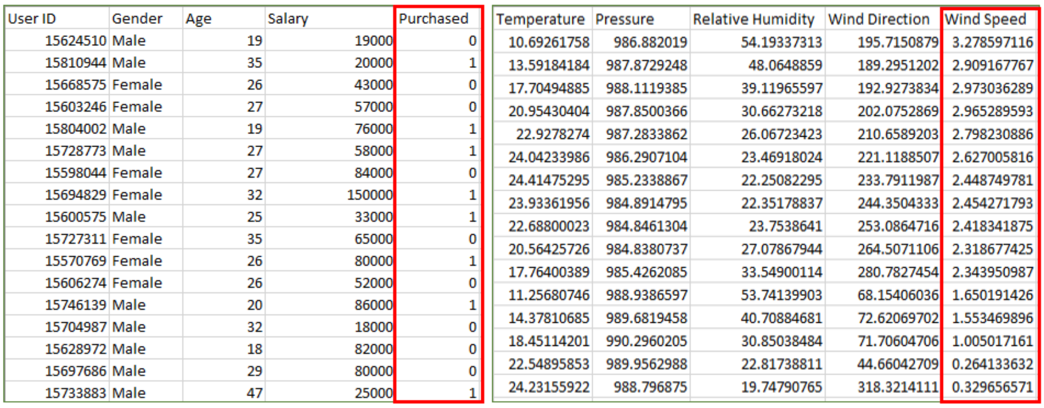


Figure 1.3

○ Types of Learning

1. Supervised learning

A learning is called Supervised realizing when the model is getting readied on a stamped dataset. Named dataset is the one which has the two data sources and yield parameters. In this kind of learning, both training &rationale datasets are labelled like shown in the below figures.



User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Figure 1.4

Theshown figures above have labelled data set –

*Figure A*has dataset of a shopping store.Itpredictsif the customer will purchase a product according to their considerationaccording to their gender, age & salary.

Input is Gender, Age & Salary

Output iswhether person will purchasethe product where value iseither 0 (No) or 1 (Yes).

*Figure B*is a meteorological dataset thatused for predicting wind speedsaccording toa variety ofparameters.

Input here is Dew Point, Temperature, Pressure, Relative Humidity & Wind Direction

Output is the Wind Speed.

Training the ML system:

While preparing the model, information is commonly parted in the proportion of 4:1 for example 80% is preparing information and rest is trying information. In preparing information we feed contribution just as yield for 80% of the information.The model realizes from training data only. We use different machine learning (ML) algorithms to build our own model. By learning it means that the model will build some kind of logic onits own.

Once the model gets ready, then it is good to be tested. At the time of testing, input is fed from remaining 20% of the data which the model has never seen before. The model will predict some value & we will compare it with the actual output & calculate the accuracy.

○ Supervised learning problems can further be grouped into regression & classification problems.

• Classification: A classification problem is said to be when the output variable is a category, such as “red” or “blue” or “disease” & “no disease”.

• Regression: A regression problem is said to be when the output variable is a real value, such as “dollars” or “weight”.

Some other common types of problems built on top of classification and regression include recommendation & time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

a. Linear regression for regression problems,

b. Random forest for classification and regression problems, and

c. Support vector machines for classification problems.

2. Unsupervised Learning

Unsupervised learning is the preparation of machine utilizing data which is neither characterized nor marked and permitting the calculation to follow up on that data with no direction [4]. Here, the job of the deviceis to group unsorted instructions according to similarity, pattern&differentiation without any prior training of data.

In contrast to supervised learning, no educator is given to the algo that implies no earlier preparing will be given to the machine. Accordingly, machine is confined to locate the concealed structure in unlabeled information without anyone else.

For instance, suppose the machineis given an image having both dogs & cats which it has not seen ever.Therefore, the machine has no idea about the features of dogs & cats, so we can't categorize it in dogs and cats. But it can categorize them according to their similarities, patterns, & differences i.e, we can easily categorize those pictures into two parts. Firstly first part may contain all pictures having dogs in it and second part may contain all pictures having cats in it. Here you didn't make it learn anything before, means no training data or examples.

○ Unsupervised learning problems may be further grouped into two, clustering and association problems.

• Clustering: A clustering problem is said to be so where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.

• Association: An association rule learning problem can be said so where you want to discover rules that describe large portions of your data, like people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are :

• k-means, for clustering problems.

- Apriori algorithm, for association rule learning problems.

3. Semi-Supervised Learning

The most basic disadvantage of any kind of Supervised Learning algorithm is that the dataset has to be hand-labelled either by a Machine Learning (ML) engineer or a Data Scientist. This is really a very costly process, especially when you have to deal with large volumes of data. And the most basic disadvantage of any Unsupervised Learning is that it's application spectrum is limited.

To counter all these disadvantages, the concept of Semi-Supervised Learning was introduced. In this kind of learning, the algorithm is trained upon a combination of labelled and unlabelled data. Typically, this combination contains a very small amount of labelled data and a very large amount of unlabelled data. The basic procedure involved is that firstly, the programmer has to cluster similar data using an unsupervised learning algorithm and then use the existing labelled data to label the rest of all the unlabeled data. The typical use cases of such kinds of algorithms have a common property among them which isthe acquisition of unlabelled data is relatively cheaper while labelling the said data is very expensive.

Intuitively, you may imagine the 3 types of learning algorithms as Supervised learning where a student is under the supervision of a teacher at both home & school, Unsupervised learning where a student has to figure out a concept all by himself / herself and Semi-Supervised learning where a teacher teaches a some concepts in class and gives questions as homework which are based on similar kind of concepts.

A Semi-Supervised algorithm assumes the following about the data –

- Continuity Assumption: This kind of algorithm assumes that the points which are closer to each other are more likely to have the same output label.
- Cluster Assumption: Here the data can be divided into discrete clusters & points in the same cluster are more likely to share an output label.

• Manifold Assumption: In this type, the data lies approximately on a manifold of much lower dimension than the input space. This assumption allows the use of distances& densities which are then defined on a manifold.

Now probably it is clear that in,

Supervised Learning,all data is labelled& the algorithms learn to predict the output from the input data.

Unsupervised, all data is unlabeled & the algorithms learn to inherent structure from the input data.

Semi-supervised, some data is labelled but most of it is unlabelled & a mixture of supervised & unsupervised techniques can be used.

4. Reinforcement Learning

Reinforcement learning is one of the zone of Machine Learning. It is tied in with making reasonable move to amplify rewards given a specific circumstance. It is utilized by different programming projects and machines to locate the most ideal conduct or the way it should take in a particular circumstance. Reinforcement learning (RL) varies from the directed learning in the way that is, in regulated learning the preparation information has the appropriate response key inside it, so the model is prepared with the right answer itself yet in support learning, there is no earlier answer given however the support specialist chooses what move to make to play out the given assignment. In the absence of training dataset, it is bound to learn from its own experience.

Structure of Reinforcement learning –

• Input is an initial state from which the model starts.

• Output are many possible things as there is variety of solutions to any particular problem.

• Training is based on the input. The model returns a state & the user decides to reward or punish the model based on its output.

• The model keeps continues to learn.

• Then the best solution is decided, based on the maximum reward.

1.8 Limitations of ML and the Solution : Deep Learning

The conventional **Machine Learning** algos are not asdifficult as they may appear, they are very much device like. They require a lot of human interference&fieldproficiency and therefore arejustable of what they are specifically intended for. This is where Deep Learning is more promising for AI Engineers, designers and users.

So, what is Deep Learning?

For all intents and purposes, Deep Learning is a field subset of Machine Learning that accomplishes more force and greater adaptability by figuring out how to speak to the world as somebody in other settled progressive system of ideas where every idea is characterized according to other less difficult ideas and increasingly unique portrayals are processed as far as less theoretical ones.

ADLmethod learnsthrough its hidden layer architecture, which defines low-level categories such as letters and then advanced level categories such as words and then superior level divisions like sentences. Taking the case of picture recognition, what it does is recognizes light or dim zones before arranging lines and then it shapes to permit face acknowledgment. Every neuron or hub in the system speaks to one part of the entire and together they give a full portrayal of the picture [6]. Every hub or shrouded layer is invigorated a weight that speaks to the of its relationship with the yield and as the model creates, the loads are balanced as needs be.

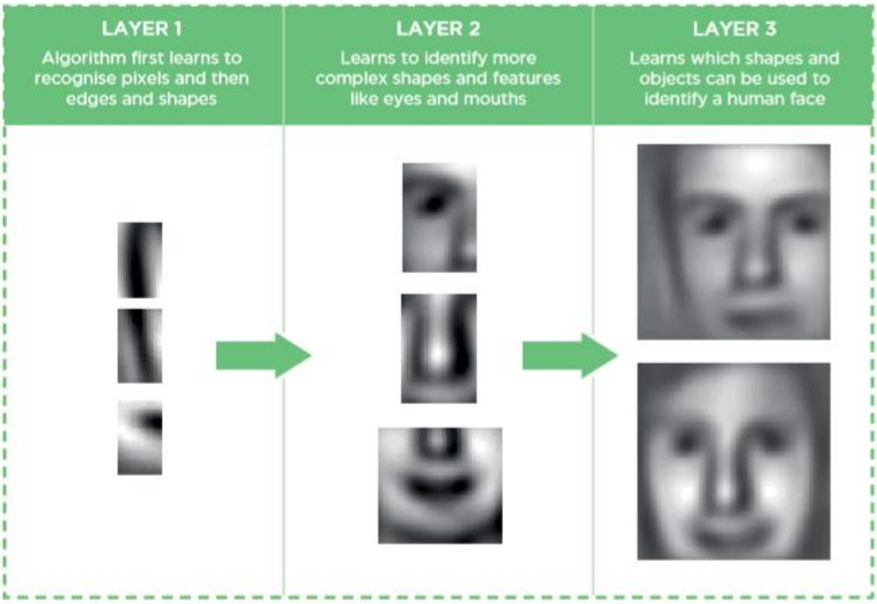

LAYER 1
Algorithm first learns to recognise pixels and then edges and shapes
LAYER 2
Learns to identify more complex shapes and features like eyes and mouths
LAYER 3
Learns which shapes and objects can be used to identify a human face

Figure 1.5

Why Deep Learning?

In conventional ML procedures, the greater part of the applied highlights are to be recognized by some space master to diminish the unpredictability of the information and to improve designs noticeable to the learning calculations for them to work. The greatest bit of leeway Deep Learning calculations have are their capacity to take in elevated level highlights from information in a steady way. Along these lines the need of area ability and hard core feature component extraction is eliminated.

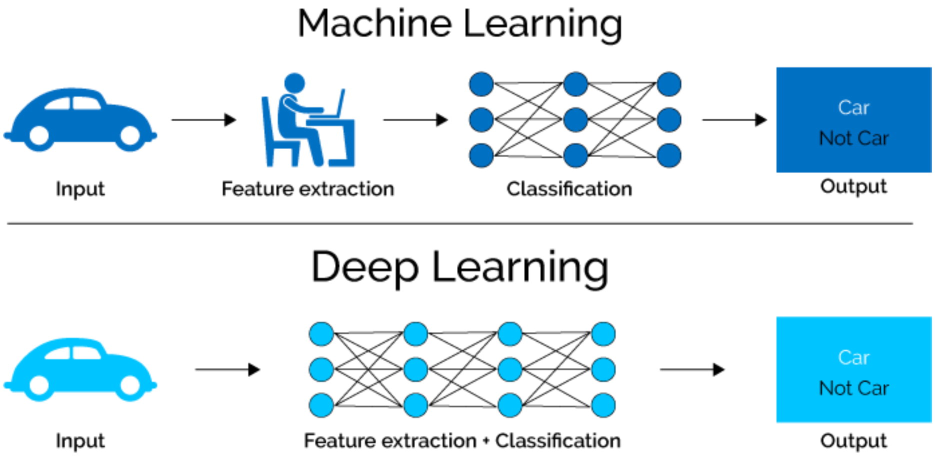


Figure 1.6

Another major distinguishing factoramong **Deep Learning** & **Machine Learning** methodologyis their dilemma solving proposition. **DL** methodology tendsto explain the problem end to end while **ML** methodologybreak downdifficulty statements to dissimilarbitssolves them first and then their results are combined at the final stage.

Generally, a **DL** algorithm takes longer to train due to the huge no. of parameters. For Example, Popular **ResNet** algorequiresaround 2 weeks to train entirely from scratch. But, conventional **ML**algos take somewhere from a few secs to a few hrs to train. This scenario is entirely overturnin thetest phase. At theanalysis time, **DL** algorithms takevery feweroccasion to run. Comparing it with k-nearest neighbour ML algorithm, test time increases when the size of dataincreases.

So, when to use DL over others?

1. DL outperforms other methodology when the data size is huge. But when dealing withtiny data size, ML algorithms are just fine.

2. DLmethodologyrequire to include a high end framework for trainingin reasonable amount of time.

3. Using DL is a better option where a lack of domain understanding for attributeintrospection is there. Here, DL methodologyoutshines others as here we don’t need to worry about feature engineering.

4. DL gives best results with Unstructured Data.

5. DL truly shines when complex problemslikeimage classification, Speech Recognition& NLP are concerned.

1.9 About Deep Learning

Deep Learning (DL) models are able enough to spotlight on just the aptattributesby themselves with lessor no regulation from the programmer. These prototypes also tend to solve the dimensionality predicament, partially.

The idea behind DL algorithmis to build learning algorithms that mimic the human brain.

Deep Learning (DL) is a subset of ML that uses a prototype of computing whichis very much encouraged by the arrangement of brain.

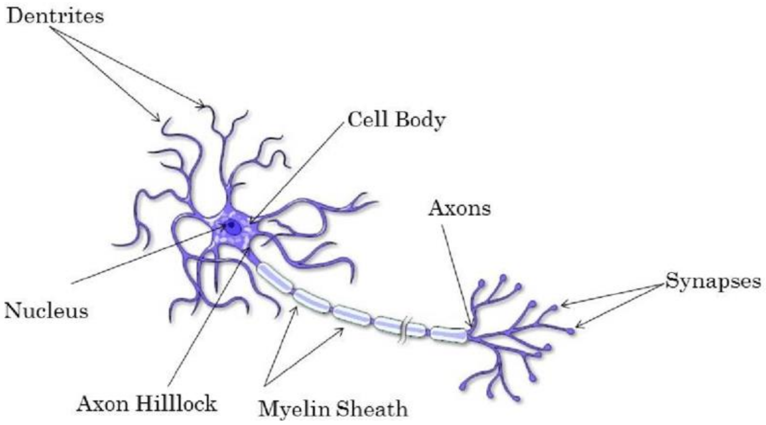


Figure 1.7

Definition : DL is a compilation of arithmeticalMLmethods which are used to learn attributerankingwhich are based on the perception of ANN.

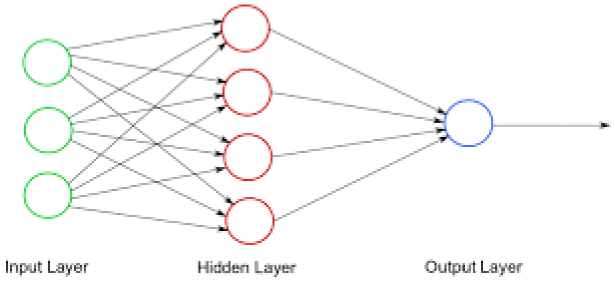


Figure 1.8

○ Types of Neural Networks

There are various kinds of Artificial Neural Networks & they are used according to the needs, type of data & set of specifications to establish the output.

• Feed Forward Neural Network

This is perhaps the least complex type of Artificial Neural Network (ANN). Here, the information (the information), goes in just single heading. The information goes through the info nodes & exits on the yield nodes. This type of neural network can have the hidden layers. Simply, it doesn’t have**Back Propagation,**only has a front propagated wavewhich usually uses a categorizing activation function.

Following is a Single layer feed forward system. In this, the total of the results of information sources and the loads are determined which is taken care of to the yield. The yield output is viewed as just in the event that it is over a specific limit esteem (typically 0) and the neuron fires it with an initiated yield (which is normally 1) and on the off chance that it doesn't fire, at that point deactivated esteem is transmitted (which is generally - 1).

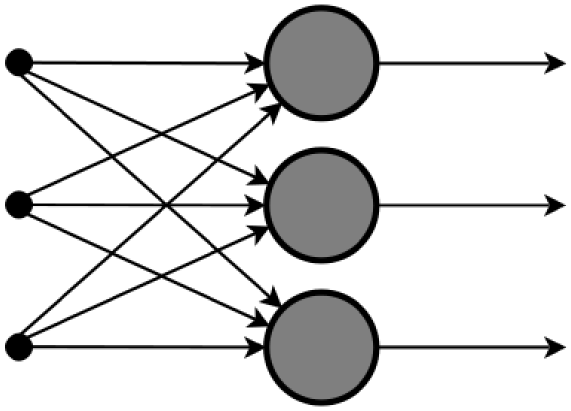


Figure 1.15

Use of Feed forward neural systems : They are found in computer vision and discourse acknowledgment framework where grouping the objective classes are entangled. These sort of NN are extremely receptive to uproarious information and simple to keep up.

♦ **Types of Feed Forward Neural Networks**

1. Perceptron : So, an synthetic neuron or Perceptron is a linear representationwhich is used for binary categorization. It prototypes a neuron which has a some inputs, each of which is assigned a precise weights. The neuron determinesa few functions on these weighted inputs & then gives the output.

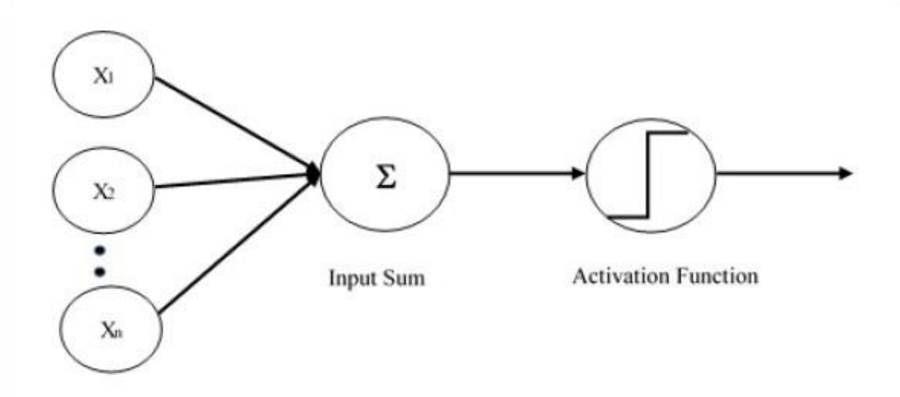


Figure 1.16

A single-layer neural network can compute a continuous output instead of a step function. A common choice is the so-called logistic function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

** Single Layer Perceptron has no hidden layer.

2. Multilayer Perceptron : A solitary neuron clearly can't perform exceptionally complex undertakings. In this manner, we use gatherings of neurons to produce the ideal yields. A multi layer perceptron has structure like that of a solitary layer perceptron however it is with at least one concealed layers and is in this manner thought about a profound neural system. Each layer has different neurons and all the neurons in each layer are associated with the various neurons in the following layer. These systems can be called as Fully Connected Networks.

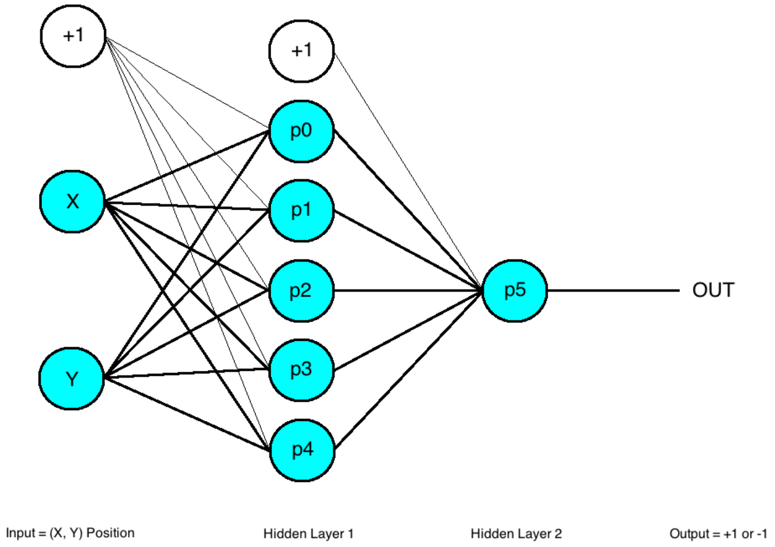


Figure 1.17

• Recurrent Neural Networks

Recurrent Neural Networks are a type of Artificial Neural Network which are designed to recognize patterns in sequences of data like handwriting, text, spoken word, genomes or numerical time series data emanating from stock markets, sensors& government agencies.

In RNNthe yield output from the past step is taken care of as a contribution to the present advance. In conventional NNs, all the sources of info and yields are autonomous of one another, yet in situations when it is required to anticipate the following expression of a sentence, the past words are required and consequently there is a much need to recollect the past words. Subsequently RNN appeared, which comprehended this issue with the assistance of its Hidden Layers. The main & most important feature of RNN is Hidden state, which remembers some information about a sequence.

RNN have a "memory" which recollects the entirety of the data about what has been determined. It utilizes exactly the same parameters for each info since it plays out a similar errand on all the information sources and shrouded layers to create the last yield. This diminishes multifaceted nature of parameters, in contrast to other neural systems.

So, how RNN works?

Check out the following example,

Assume there is a profound system with 1 information layer, 3 shrouded layers and 1 yield layer. At that point like all other neural systems, each shrouded level will have its own arrangement of loads and inclinations. Presently, suppose, for the concealed level1, the loads and inclinations are (w1, b1), (w2, b2) for shrouded level2 and (w3, b3) for concealed level3. This implies every single of these levelare free of each other, for example they don't remember the past yields.

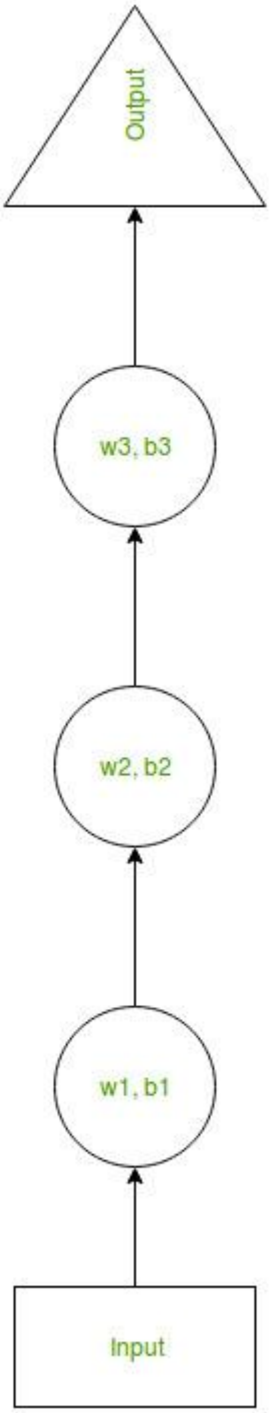


Figure 1.18

Now the RNN does the following:

RNN changes over the independent activations into subordinate activations by giving similar loads and same predispositions to all the layers, along these lines lessening the multifaceted nature when there are expanding parameters and remembering each past yields by giving each yield as contribution to the following concealed layer gets excessively.

Henceforth these 3 layers can be connected together with the end goal that the loads and predisposition of all the obscure layers is the equivalent, into a solitary intermittent layer.Henceforth these 3 layers can be connected together with the end goal that the loads and predisposition of all the obscure layers is the equivalent, into a solitary intermittent layer.

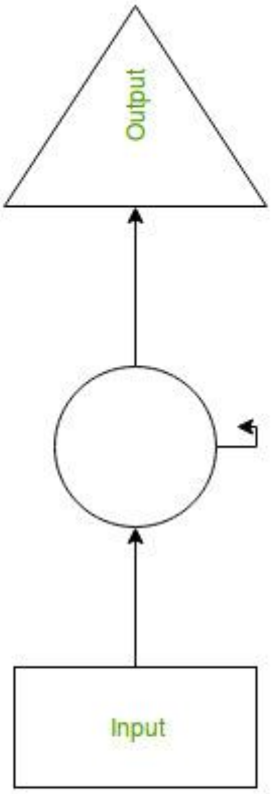


Figure 1.19

→ Formula for calculating current state :

$$h_t = f(h_{t-1}, x_t)$$

where, ht -> current state

ht-1 -> previous state

xt -> input state

→ Formula for applying Activation Function :

$$h_t = \tanh (W_{hh}h_{t-1}+ W_{xh}x_t)$$

Where, whh -> weight at recurrent neuron

wxh -> weight at input neuron

→ Formula for calculating output :

$$y_t = W_{hy}h_t$$

where, Yt -> output

Why -> weight at output layer

• Training using Recurrent Neural Network

1. Only a solitary time venture of the information is given to the system.
2. Then we compute its present state by utilizing a lot of current information and the past state.
3. The current ht then becomes ht-1 for upcoming step.
4. One can go the same number of time as required by the issue and join the data from all the past states.
5. Once all the time steps are finished, the last current state is utilized to compute yield.
6. This yield is then contrasted with the real yield. i.e the objective yield and the blunder is produced.

7. This blunder is then back-proliferated to the system for refreshing the loads and consequently the RNN organize is prepared.

• Advantages of RNN

1. RNN recalls each and every data through time. It is exceptionally helpful in time arrangement expectation just because of its element to recollect past contributions too. We can call it Long Short Term Memory.
2. Recurrent Neural Network are even utilized with Convolutional layers (which we'll see further) to broaden the successful pixel neighborhood.

• Disadvantages of RNN

1. Gradient evaporating and detonating issues.
2. Training RNN is troublesome errand.
3. RNN can't process long arrangements in the event that we are utilizing tanh or ReLU as an initiation work.

○ Convolutional Neural Networks

Convolution Neural Networks (CNNs) or Covnets are Neural Networks which share their parameters. Simply envision you have a picture and it tends to be spoken to as a cuboid having its width, length (as measurements of the picture) and its stature (as picture by and large have red, green, and blue channels).

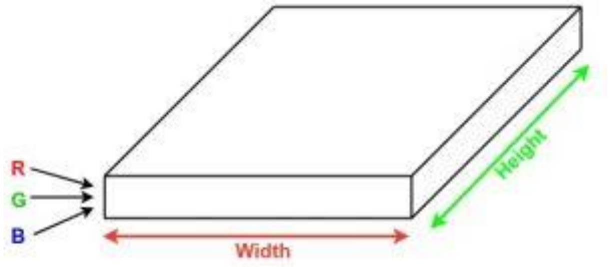


Figure 1.20

Now assume taking a small patch of this image & running a small NN on it, with k outputs (auusme) & represent them vertically. Now slide that neural network across the whole image. As a result, we now get another image with differentheight, depth and width. And instead of just R G B channels now we have more channels with lesser width & height. This operation is called Convolution. Now if the patch size is same as that of the image,it will be a regular neural network. Because of this small patch, we have fewer weights.[6]

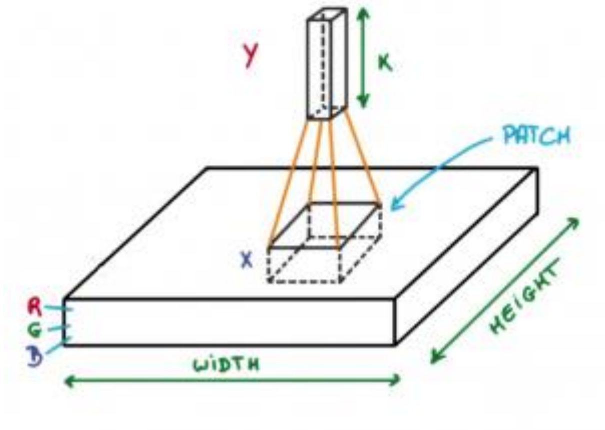


Figure 1.21

Now talking about a bit of maths which is involved in the whole convolution process.

1. Convolution layers comprise of a gathering or set of learnable channels (see fix in the above picture). Each channel has little stature and width and a similar profundity like that of info volume (3 if the information layers is picture input).

2. For model, on the off chance that we need to run a convolution on a picture with measurements as 34x34x3. Conceivable size of channels would then be able to be axax3, where 'a' can be 3, 5, 7, and so forth yet little when contrasted with picture measurement.

3. During the forward pass, we slide each channel across entire info volume bit by bit where each progression is known as step (which can have an incentive as '2' or '3' or possibly '4' for high dimensional pictures) and figure the speck item between the loads of channels and fix from input volume.

4. As we slide our channels we get a 2-D yield for every single channel and we at that point stack them together. Therefore, we get yield volume having profundity equivalent to the quantity of channels. The system at that point learns all the channels.

- **Layers used to build ConvNets**

Covnets is a sequence of layers, & every layer transforms one volume to another through differentiable function.

Following are types of layers:
Take an example by running Covnets on an image of dimension 32 x 32 x 3.
1. Input Layer: This is the layer which holds the raw input of image with width 32, height 32 and depth 3.
2. Convolution Layer: This layer computes the o/p volume by computing dot product between all filters & image patch. Assume, we use a total of 12 filters for this layer, we'll get output volume of dimension 32 x 32 x 12.
3. Activation Function Layer: This layer applies element wise activation function to the output of convolution layer. Some common activation functions are Sigmoid : 1/(1+e^-x), Tanh , RELU : max(0, x), Leaky RELU, etc. The volume remains unchanged so output volume will have dimension 32 x 32 x 12.
4. Pool Layer: This layer is periodically inserted in the CNNs & its main function is to reduce the size of volume making the computation fast which reduces memory & also prevents it from overfitting. 2 common types of pooling layers are - max pooling & average pooling. If we use a max pool with 2 x 2 filters & stride 2, the resultant volume becomes of dimension 16x16x12.

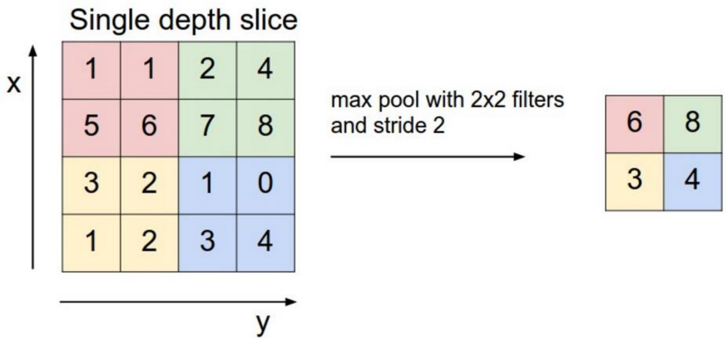


Figure 1.22

5. Completely Connected Layer: In this layer, normal neural system layer takes contribution from the past layer and it registers the class scores and then yields the 1-D exhibit of size equivalent to the no. of classes.

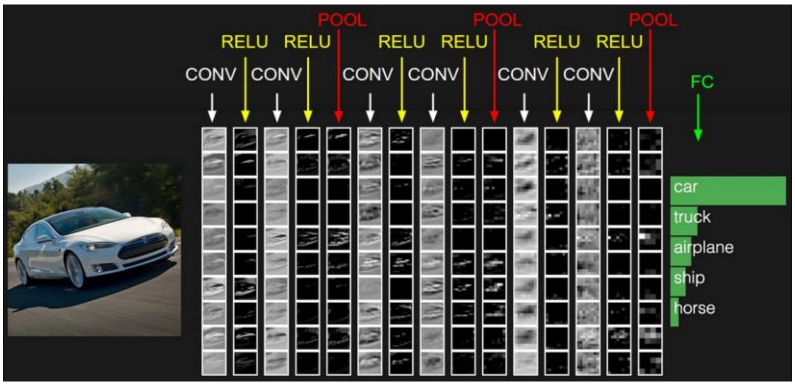

POOL
POOL
POOL
RELU
RELU
RELU
RELU
RELU
RELU
CONV
CONV
CONV
CONV
CONV
CONV
FC
car
truck
airplane
ship
horse

Figure 1.23

1.10A Powerful Tool : Python

Python is an interpreted, high-level, general-purpose programming language created by Guido van Rossum in 1991.

Due to its ease in readability (majorly resembling English like words), simple syntax and vast library support, it is perfect for writing code and supporting machine learning algorithms. Its applications include but are not limited to Machine Learning, Deep Learning and other Artificial Intelligence applications. Some other widespread uses include Web Development, performing mathematics operations, scripting and software development.

Development UI used for coding of this project is Anaconda Navigator 3 and language worked on is Python 3.6.

Python IDLE – IDLE a.k.aincorporated development environment for Python ispacked with default execution of the programming language since 1.5.2b1. It’s written in Tkinter GUI toolkit and Python, bundled as elective part of Python binding with various Linux distributons.

Anaconda Navigator – Anaconda is an open source GUI (Graphical User Iterface) Python distribution (also supports R language), used for scientific computation, machine learning applications, prediction analytics and much more. Main aim of this software is simplification of package management and deployment in various environments and channels without using command-line commands.

Some applications inbuilt and readily available in Anaconda are :

1. Jupyter Notebook
2. Jupyter Lab
3. Spyder
4. QtConsole
5. R Studio

CHAPTER – 2

LITERATURE REVIEW

1.1 A Survey of Audio-Based Music Classification and Annotation by Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang

This paper gives an outline of highlights and methods utilized for the music arrangement assignments. The key parts of an arrangement framework as indicated by this paper are feature extraction and classifier learning. Highlight extraction tends to the issue of how to speak to the guides to be ordered as far as highlight vectors or pair shrewd similitudes. The reason for classifier learning is to discover a mapping from the component space to the yield marks to limit the forecast blunder [1]. This spotlights on music order dependent on sound signs except if in any case expressed.

This examination utilizes both K-NN and SVM classifiers for single element vector portrayals and pair astute similitude esteems too. In the last case, a portion lattice is worked from pair savvy similitude esteems that can be utilized straightforwardly by the SVM. GMMs are utilized for tune level comparability calculation. This is unique in relation to classifier learning. For the GMM classifier, they fits the Gaussian blend model over the appropriations of tune level highlights in each class.

The undertaking theories classifiers accomplish are Genre Classification, Mood Classification, craftsman distinguishing proof, music comment and instrument acknowledgment.

The problems this research faces are :

- Large-Scale substance based Music Classification with Few Labeled Data
- Music Mining From Multiple Sources
- Learning Music Similarity Retrieval
- Perceptual Features for Music Classification

1.2 A Survey on Sound Source Separation Methods by Ms. Monali R. Pimpale, Prof. Shanthi Therese, Prof Vinayak Shinde.

The detailing of vocalist distinguishing proof framework empowers powerful administration of enormous amount of music information. With this framework, for example vocalist distinguishing proof innovation, tunes completed by a specific artist can consequently be grouped for simple administration and looking. There are numerous calculations which can be utilized for vocalist ID which depend on include extraction idea. This recognizes the suitable artist from the given highlights. In popular music, performing voice is joined with music backup. Consequently, those strategies dependent on the highlights which are removed straightforwardly from the went with vocal sections are troublesome to secure great execution when the backup is more grounded or possibly performing voice is more vulnerable. In order to show signs of improvement execution, a few procedures developed which help in detachment of the performing voice from music backup. There are different sound source partition calculations which are utilized to isolate the performing voice from music backup. Along these lines, Sound source detachment implies that the undertakings to assess the sign created by an individual sound source by a blend of sign comprising of numerous sources. This is basic issue in different sound sign handling undertakings, since examination and preparing of confined or single sources should effectively be possible with a whole lot preferred precision over the preparing of sound blends. "Solo learning" is the term used to portray calculations which attempt to isolate and gain proficiency with the structure of sound sources in blended information because of high unpredictability of sound information and this is done dependent on data hypothetical standards, as factual autonomy between sources, rather than generally refined displaying of the source attributes or human sound-related observation. There are such a large number of solo learning sound source partition calculations out of which some of them incorporate autonomous segment examination (ICA), meager coding, and non-negative network factorization, which has been immensely utilized in source detachment undertakings in a few application regions.Every other method but Non-Negative Matrix Partial Co-Factorization have some significant errors.

Disadvantages of other methods are mentioned as follows :

CASA : The performance of current CASA system is still limited by pitch estimation errors and residual noise.

Beam forming : The amount of noise attenuation increases as the number of microphones.

ICA : A key and primary issue of this method is before an effective source separation the system should estimate the number of unknown sources from the mixed signals.

Pitch Estimation & Tracking : Evaluated essential recurrence of singing is hard to be exact due to the impact of backup. Regardless of whether the assessed key recurrence is right the removed sounds of performing voice are not totally unadulterated on the grounds that the a few music segments of performing voice might be superimposed by pitched instrument.

Even Non-Negative Matrix Factorization has a slight limitation. It imposes only the non-negativity constraint.

So, here comes the savior. A better method which gives some better results. So, Non-Negative Matrix Co-Factorization (NMPCF) is the new technique for source separation that is giving better performance than all existing methods used in source separation. Therefore, NMPCF can be used for singer identification with better performance.

1.3 WAVE-U-NET : A multi-scale Neural Network for End-to-End AudioSource Separation

Till the distribution of this Research Paper "Wave-U-Net" sound source partition used to chip away at the greatness range which disregarded the stage data which thusly made detachment execution dependant on hyper parameters for otherworldly front end [6]. Along these lines, in this examination paper they've detailed a strategy which ends to end source partition in the time space permitting demonstrating stage data and staying away from fixed spectral transformations.

Most of the audio source separation models work on spectrogram render of the audio signals. So, after applying STFT to the i/pblendsign, the composite valued spectrogram is divided into its magnitude & phase elements (which is a mixture phase). But for the parametric model, the input fed is only the magnitude part to separate different audio streams. After that inverse STFT is applied in time domain adding it to the mixture phase which then gives different audio sources.

Wave-U-Net is an adaptation of U-Net architecture to 1-D time domain which divides sources openly in the time sphere& can also take hugesequential contexes into consideration, in contrast to U-Net. Moreover,

1. The Wave-U-Net accomplishes great multi-instrument and performing voice partition, the last of which thinks about well to re-execution of the best in class organize design.
2. Wave-U-Net can process multi-channel audio which is helpful in comparing mono audio to stereo audio.
3. Wave-U-Net also highlights the limitations of using SDR as evaluation metric.

Some previous models are parameter productive yet their memory utilization is high since each component map coming about because of enlarged convolution has the first sound's examining rate as a goals. Wave-U-Net adopts an alternate strategy. It takes more highlights and progressively lower goals to spares memory. In this way, it empowers a huge no. elevated level highlights which needn't bother with test level goals to be valuable like instrument movement.

Wave-U-Net Model

The main goal of this model is to separate a mixture waveform M into K basis waveforms S1, S2, S3, ……. Sk for all K.

L = no. of audio samples, Lm for mixture and Ls for source waveforms

C = no. of audio channels

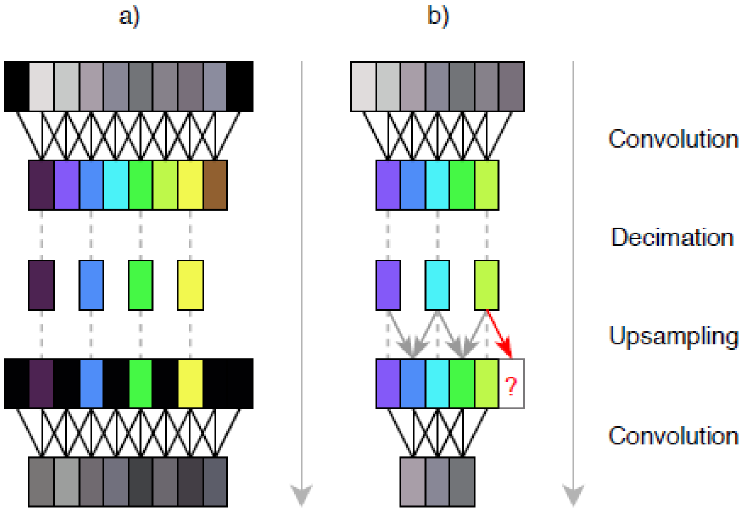
The general design of the model is to initially register an expanding number of more significant level highlights on an unpleasant time scale utilizing downsampling (DS) squares. At that point these highlights are joined with some time ago figured neighborhood and high goals highlights utilizing the upsampling (US) squares, yielding multi-scale highlights which are utilized for building expectations.

Here using in total L levels, each following level operates at 1/2 the time of resolution as compared to the prior one. For total K sources to be approximated, the model returns predictions in out of range (-1,1), one for every audio basissection.

One of the important features of Wave-U-Net architecture is to avoid aliasing artifacts because of upsampling. Instead of transposed strided convolutions , which other models use, this model performs linear interpolation which ensures that in temporal continuity in feature space is kept intact. After that normal convolution is done.

This strategy keeps number of highlights lopsided, so that upsampling doesn't require extrapolating values (red bolt). Here in spite of the fact that yield is littler, curios get maintained a strategic distance from.

Figure 2.1



a) Common model using transposed strided convolutions

b) Wave-U-Net using linear interpolation

There are significant architectural improvements over other previous audio based models to increase model performance. Some points which give a glimpse of how performance is increased can be pointed as follows:

1. Difference in output layer which output1 source approximation for every 'K' source by separately applying 'K' convolutional fi1ters then following by tanh non-linearity to the 1ast feature plot.
2. Does prediction using proper input context & resampling.
3. Support for stereo channels improving the quality of sampling and output audio souurce
4. Different aspects of upsampling from other models by using unstructured feature space.

Finally experimentation is done on primarily 2 tasks. One, singing part voice separation and two, music partition with bass (perfect for real life music concerts and plays).

Dataset used is while experimentation is MUSDB which is a multi track database. Out of this 75 tracks was used as training dataset and the outstanding 25 tracks were used as justification dataset which was used for early halting. Total songs used were 50 and for singing voice separation database used CC Mixter Database.

Now for results the performance metrics generally used in these kinds of problems is Signal to Distortion ratio (SDR). But here one problem arises (as indicated by this study) that when the correct source is quiet or near quiet SDR gets undefined log(0) which can be problematic because in practical cases generally in voice separation of human voice this happens quite often. So, as an improvement to this so as to successfully separate the signals, median absolute deviation (MAD) is used as class based deviation metric over Standard Deviation. The main thing to note is that during the actual application no pre or post-processing was done. [6]

While checking the model after successful training a lengthy temporal context was polished by recurring downsampling & convolution of attribute maps to combine low & high-level characteristics at diverse time-scales was done. It outperformed the state-of-the-art spectrogram-based U-Net design but as we'll further see still lags behind some advanced audio-visual methods.

1.4 Looking to Listen at the Cocktail Party : A Speaker-Independent AudioVisual Model for Speech Separation

This research paper to best of my knowledge is among the first to establish a solid point about audio-visual method being a better approach to solve the cocktail party problem while also being speaker-independent in terms of speech separation in machines. Solving this problem is a very challenging task, especially when only audio signals are considered. The problem solved using this new method was based on a Deep-Network established model that uses both optical& auditory based signals. Visual features were used to concentrate on the speaker whose desired speech is to be refined and then speech separation was done. [7]

AVSpeech is the joint audio-visual database which was used to train this model. Comprised of thousands of hours of videos (more than 2,90,000 high quality videos) from the internet of interviews, TED talks, How-to videos, screaming children, noisy restaurants and bars, etc, from the YouTube. Then out of all this roughly 4700 hours of video clips were extracted in which there were no interfering sounds and just the clean speech. All this was used to create some synthetic cocktail parties including faces, their speech, background noise and all. This database is useful for implementing the model in which just the face of target speaker is required to be pointed and its speech will be isolated. This model is also mentioned to be speaker-independent.

Generalising the topic by getting acquainted with Cocktail Party effect which is the focussing of our auditory concentration towards a single sound source from a mixture of many sounds within a raucoussetting, while suppressing (or muting or de-emphasing) all other voices and sounds. This ability is well knowned in human beings but mimicing it on a machine is a very tedious job. The automatic speech recognition and separation is very well studied in audio processing literature and usually requires prior knowledge of the environment or very specialised microphone configurations to obtain a resonable solution. An important limitation of audio only method is label permutation problem.

The model works to such an extent that a recorded sound blend alongside harvests of recognized faces in each edge is taken from the video as info and the blend is part into isolated sound streams for ech identified speaker. Likewise, here the visual data is utilized as a way to develop the source division quality and to relate the seperated discourse follows noticeable speakers in the video. Along these lines best in class results on discourse partition is accomplished.

Earlier Work

Some work has been done in this field and most of them are based on Audio only methods for speech denoising and speech separation tasks. Wang and Chen have given some comprehensive view on the deep learning models for this as stated in the paper. Issues like the label permutation problem for speaker free and multi-speaker separation in the single channel situation have been halfway fathomed utilizing profound bunching in which discriminatively prepared discourse sections are utilized for isolating groups and isolating various sources.

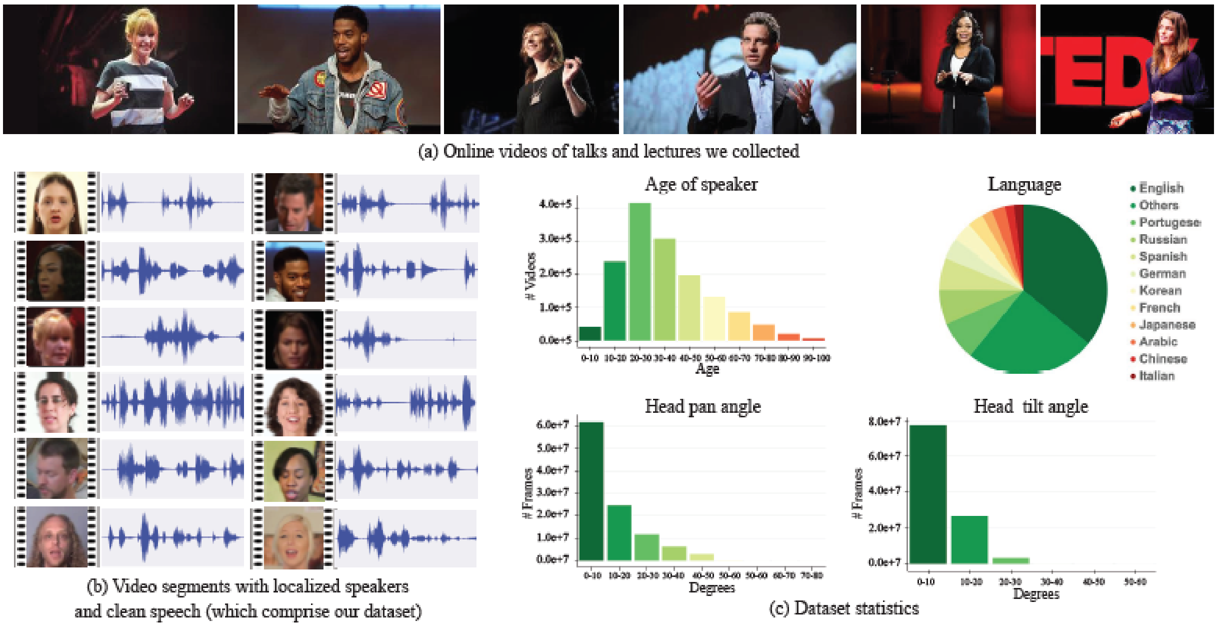
In visual viewpoints, there has been an expanded enthusiasm for taking care of discourse related issues utilizing neural systems for multi-modular combination of sound-related and visual signs. A few works are various media discourse acknowledgment (for foreseeing content or discourse from a quiet video) or lip-perusing. AV (various media) strategies have likewise been utilized for discourse upgrade applications. The primary restriction in past works was that they all were speaker subordinate, for example a committed model must be prepared for every speaker seperately which isn't exceptionally down to earth.

Not much was done in Audio-Visual methods due to the lack of database which is now solved by the introduction of AVSpeech. This database is also helpful for the seperation of speech when it comes to tackling the speech in different languages. The vast majority of the prior datasets involved recordings with extremely less number of subjects, talking words from a restricted jargon. One such dataset is CUAVE dataset. It contained around 36 subjects saying every digit multiple times with an absolute 180 models for each digit[7]. Another dataset is video accounts in Mandarin where 320 articulations of Mandarin sentences are conveyed by a local speaker. Every sentence has 10 characters in Chinese with similarly conveyed phenomes.

AVSpeech Dataset

This large scale AV dataset contains verbal communication clips with nil interference of backdrop signals. Fragmentsliein the range starting from 3 to 10 seconds in length and every clip contains noticeable face and audible sound of only 1 person. Totally, the dataset contains around 4700 hours of video segments which contains around 1,50,000 distinct speakers, languages, broadkinds of people and face poses.

This dataset collected was done automatically to avoid any biasness due to human interference and also it wasn't feasible for someone to collect this much data without spending too much time in it.



(a) Online videos of talks and lectures we collected

(b) Video segments with localized speakers and clean speech (which comprise our dataset)

(c) Dataset statistics

Figure 2.2

Verdict

When comparing to audio only methods, since there is no state-of-the-art audio only speech separation techniques yet & very less datasets obtainable for evaluation & the techniques are devised use multiple audio channels. So to compare with them AO baseline was used for speech improvement. It has alikedesign to this AV model, so it’s easier to evaluate. It came out to be near state-of-the-art results for AO when trained & evaluated for ChiME-2 dataset.

Now, comparing with other Audio Visual methods was tricky because there were not any speaker independent methods at the time of experimentation but quantitatively analysis has shown that the results were quite impressive.

CHAPTER - 3

PROPOSED METHODOLOGY

After going through previous advancements for this difficulty of audio source partition& speech improvement, I have some mixed views over it. There are two approaches to solve this problem, first being the audio only method. This method basically views the problem as a mixture of audio segments of the target speaker(s), unwanted sounds from the environment and other noise. Separation of desired audio segment using this approach is a little tedious because voice frequencies of different people lie in a specific range and can variably get mixed and misinterpreted to be of same person. This problem arises mostly when dealing with people of same gender. Female voice are high pitched and lie on the upper spectrum of human voice whereas male's voice is low pitched and lie in the lower spectrum. Second approach is Audio Visual Method where not only audio samples but video frames are also taken in consideration. This approach has given better results till now as compared to AO methodology as stated in the previous mentioned paper "Looking to listen at the cocktail party".

So, the first approach, Audio Only method can be useful in scenarios when visual data is not present, most likely when Military or spy applications are considered. But most of the time for the ordinary people where smartphone is probably the most powerful device anyone has, Audio-Visual application is better suited.

3.1 Voice Focus using Audio-Only Model

Getting to the design aspects, let’s see the audio only model which is theoretically, to my knowledge, is better than any audio only method devised yet. This design is aimed to get good accuracy which makes it a little slower in practicality but this is due to slow internet connectivity and low processing power available in the market. So, check the following flowchart to get better insight of it.

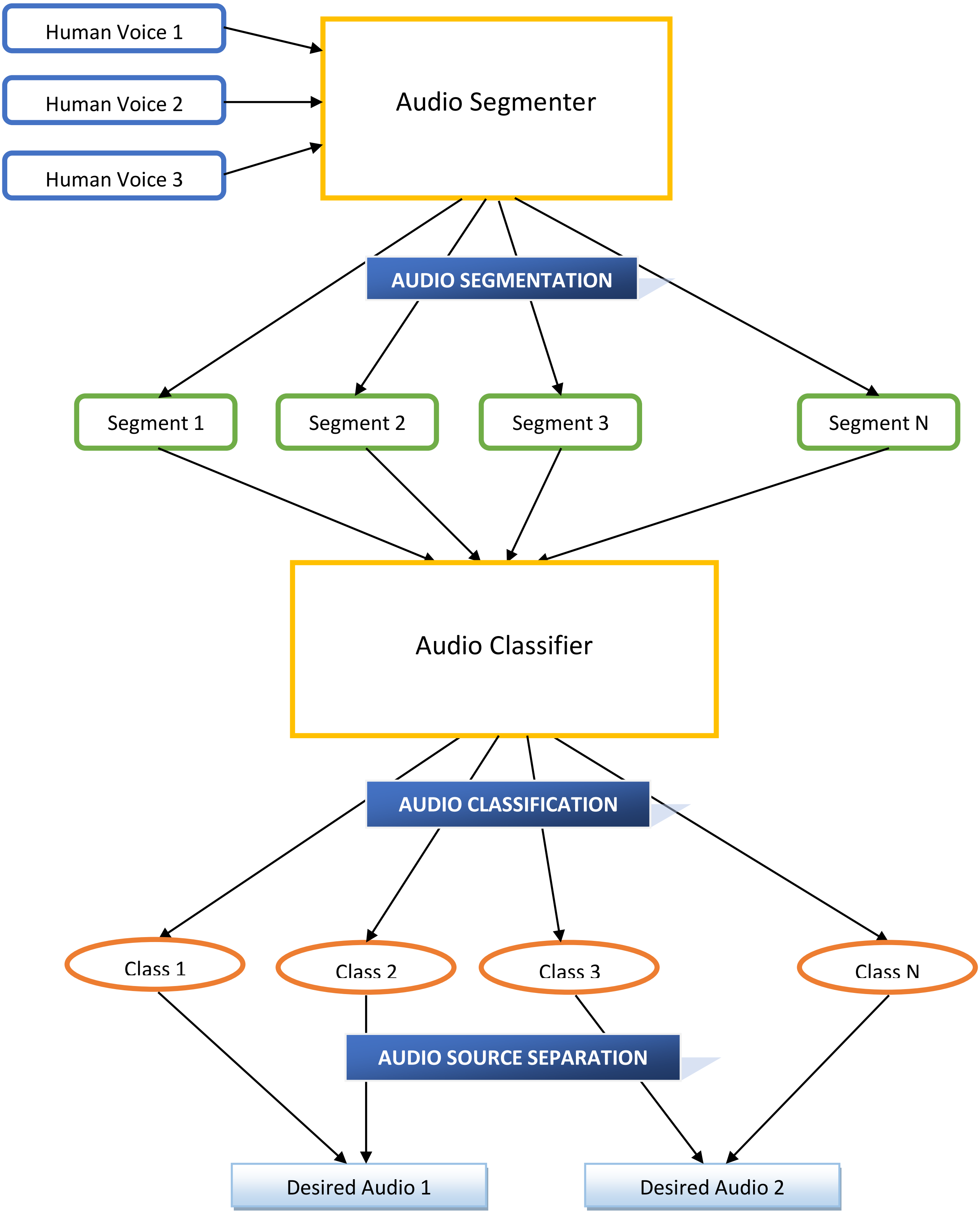

Human Voice 1
Human Voice 2
Human Voice 3
Audio Segmenter
AUDIO SEGMENTATION
Segment 1
Segment 2
Segment 3
Segment N
Audio Classifier
AUDIO CLASSIFICATION
Class 1
Class 2
Class 3
Class N
AUDIO SOURCE SEPARATION
Desired Audio 1
Desired Audio 2

Figure 3.1

3.1.1 Audio Segmentation

Programmed sound division means to separate a computerized sound sign into sections, every one of which contains sound data from a particular acoustic sort, for example, discourse, music, non-verbal human movement sounds, creature vocalizations, natural sounds, clamors, and so forth. The level of detail in sound class examination relies upon the application. For instance, in radio station signals division the intrigue falls in the discovery of the sound parts that contain discourse, music, quietness, and commotions. It's a pre-preparing step that utilized in sound investigation that helps in isolating various sorts of sounds like discourse, music, ecological sounds, commotion, quietness, and blends of all sort of sounds.

In the present scenario there are large number of audio features audio retrieval,audio segmentation, and different environmental sound retrieval. Audio segmentation is the class of theories and algorithms designed to automatically reveal semantically meaningful temporal segments of the audio signal [2]. Automatic detection of *auditory genesis* a critical part in enabling high-level semantic imageries from general audio signals, and can have the advantage in various content-based applications involving both audio and multimodal data sets. This is the Traditional approach to audio segmentation.

Programmed sound division bearing is to have division of a computerized sound sign into fragments, where every one of which contains sound data from a particular acoustic sort, for example, discourse, music, music with various sort of signals, non-verbal human action sounds, creature vocalizations, natural sounds, commotions, etc different things, and so forth. The subtleties in sound investigation rely upon the applications For instance in radio station signals division the premium falls in the discovery of the sound parts that contain discourse, music, quietness, and commotions.

In the area of data innovation handling structures which is for the most part managing sound information, the primary job of the programmed division subsystem is to isolate all the sound signs to the acoustic classes of enthusiasm for request with the end goal that it can additionally be prepared by the comparing frameworks. Such post-preparing frameworks can be utilized as discourse recognizers, speaker recognizers, language recognizers, vocalist recognizers, melody recognizers, sound occasion recognizers and they can remember it on various angles. The idea of such a system and the job of sound division inside it are delineated in following figure that can be seen, the underlying sound stream is into a sound division, which is an open design regarding its distinctive kind can fluctuate in which every sound sort section is crashed into a reasonable piece of post-preparing. in communicate transmissions, discourse parts can be begun into programmed discourse recognizer for phonetic or speaker job handling while music parts can be started into an audio cue assortment library.

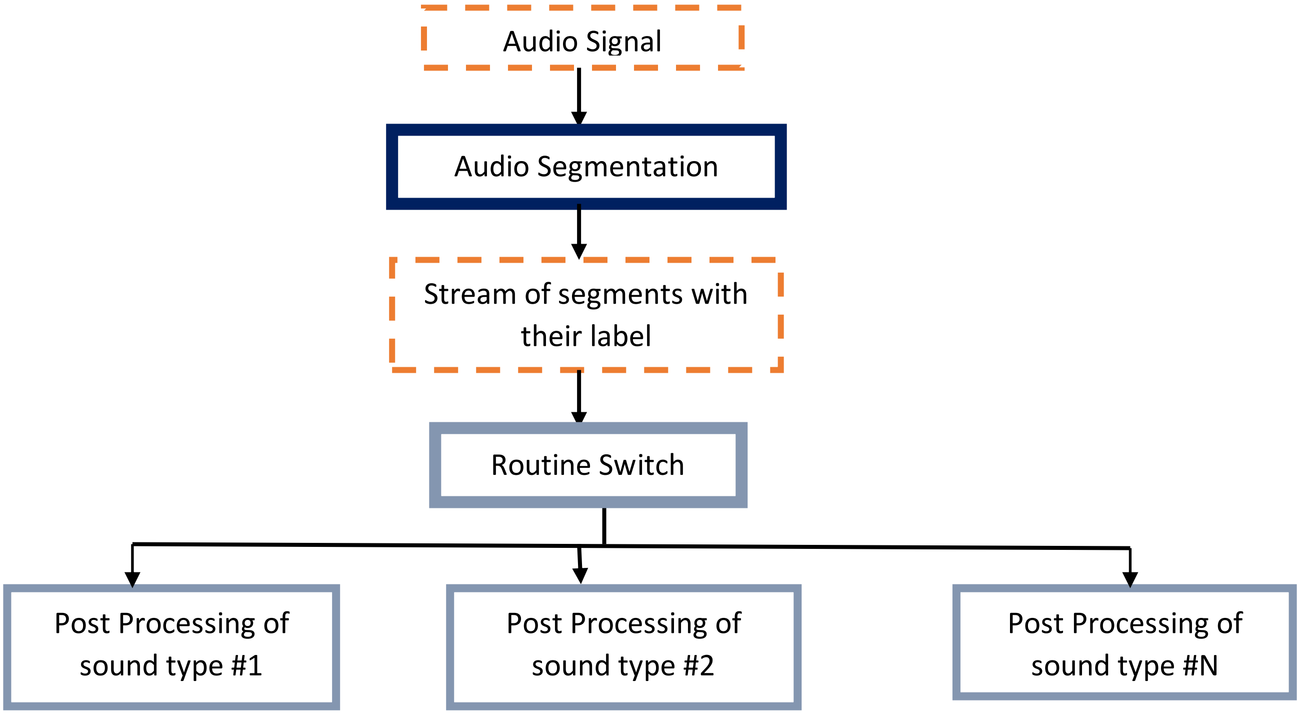

Audio Signal
Audio Segmentation
Stream of segments with their label
Routine Switch
Post Processing of sound type #1
Post Processing of sound type #2
Post Processing of sound type #N

Figure 3.2

3.1.2 Audio Classification

With the improvement of data and interactive media innovation, advanced music has gotten generally accessible from various medias, including Radio Broadcasting, Digital Storage like Compact Disks (CDs), Audio tapes, the Internet, and so on. The tremendous measure of music open to the overall population is utilized for creating instruments to successfully and effectively recover and then deal with the music of individual's enthusiasm to the end clients. Music Information Retrieval (MIR) is inconceivably rising examination zone in sight and sound to adapt up to such need. A key issue in MIR is order, which doles out marks to every single tune dependent on possibly sort, state of mind, specialists, and so on. Music order is an exceptionally fascinating point with numerous potential applications. It gives some significant functionalities to music recovery [1]. This is so in light of the fact that most end clients possibly just inspired by specific sorts of music. Along these lines, a grouping framework would empower them to look for the music they are keen on. Then again, various types of music have various properties. We can oversee them successfully and efficiently once they are ordered into different unmistakable gatherings.

The key parts of a characterization framework are : 1. include extraction and 2. classifier learning. Highlight extraction tends to the issue of how to speak to the guides to be grouped regarding some component vectors or pair-wise similitude. The motivation behind Classifier Learning is to discover a mapping from the component space to the yield names to limit the likely expectation mistake. We center around music order dependent on sound signs except if some other information is given.

From the point of view of music understanding, we may isolate sound highlights into 2 levels, one, low-level and mid-level highlights, as delineated in the last two columns of figure alongside other top - level names. Low-level highlights can additionally be partitioned into two classes - timbre and worldly highlights.

Low-level highlights are gotten legitimately from different sign preparing methods like Fourier change (FT), unearthly/cepstral investigation (SA/CA), autoregressive demonstrating, or more. Likewise, each class of low-level highlights comprises of a wide range of highlights, see the figure.

Mid-level highlights give a closer understanding relations and incorporates for the most part 3 classes of highlights, in particular beat, pitch, and amicability. These highlights are by and large removed on low - level ones.

At the astoundingly top level, semantic imprints give data on how people understand and translate music subject to type, perspective, style, etc. This is a hypothetical level considering the way that the imprints can't be speedily gotten from lower level features which is appeared by the semantic opening between mid - level features and names. The inspiration driving substance based music request is to interface the semantic opening by gathering the imprints from low or mid - level features.

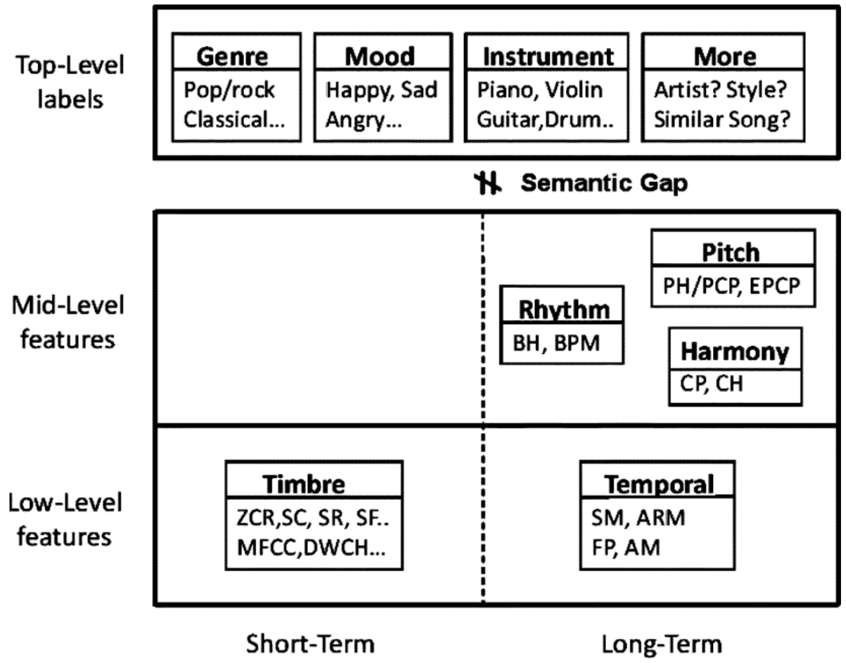


Figure 3.3

Some common low level features used in music classification :

Class	Feature Type
	Zero Crossing Rate (ZCR) Spectral Centroid (SC) Spectral Rolloff (SR) Spectral Flux (SF) Spectral Bandwidth (SB) Spectral Flatness Measure (SFM)

Timbre	Spectral Crest Factor (SCF) Amplitude Spectrum Envelope (ASE) Octave based Spectral Contrast (OSC) Daubechies Wavelet Coef Histogram (DWCH) Mel – frequency Cepstrum Coefficient (MFCC) Fourier Cepstrum Coefficient Linear Predictive Cepstrum Coefficient (LPCC) Stereo Panning Spectrum Features (SPSF)
Temporal	Statistical Moments (SM) Amplitude Modulation (AM) Auto – Regressive Modeling (ARM)

Low level feature extraction is performed in following steps :

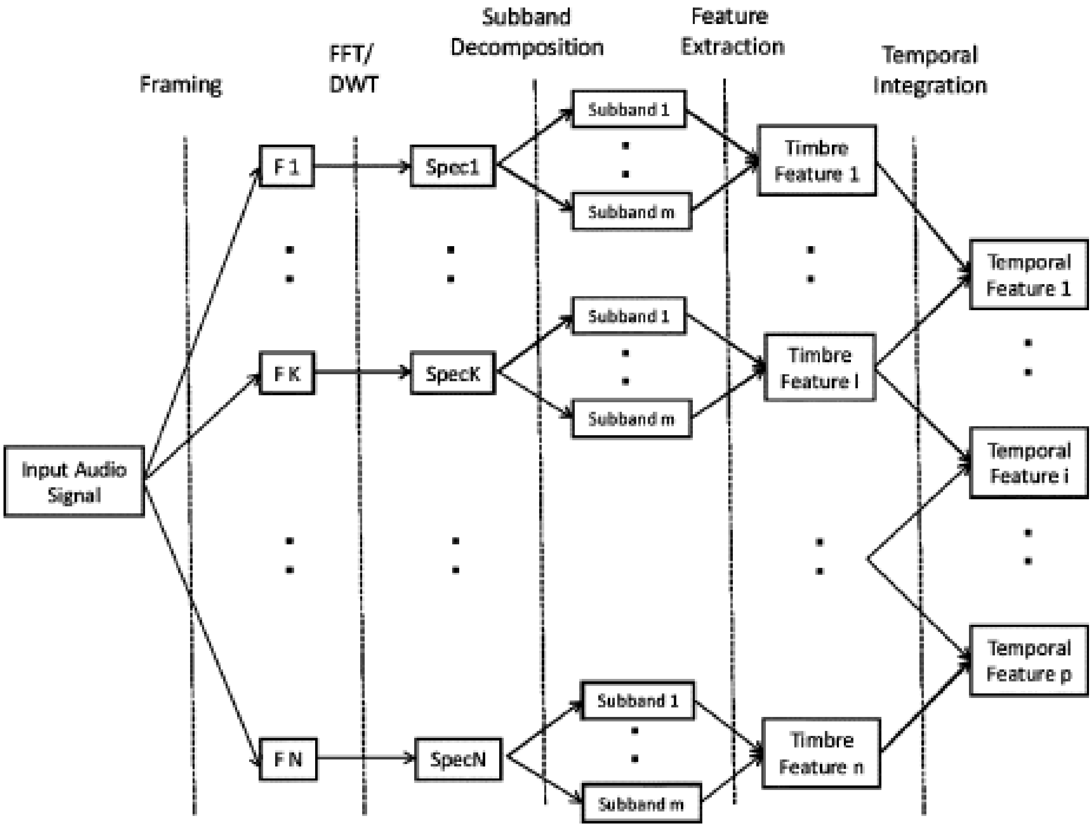


Illustration of steps for low-level feature extraction.

Figure 3.4

3.1.3 Audio Source Separation

The improvement of an artist ID empowers the general powerful administration of a lot of music information. With this, the artist distinguishing proof innovation, melodies performed by some vocalist can be consequently grouped for simple administration and looking. There are a wide range of calculations which right now are utilized for artist recognizable proof. These depend on the idea of highlight extraction which recognizes the proper vocalist from a portion of the acquired highlights. In well known music, performing voice is joined with music backup. Along these lines, those strategies dependent on the highlights removed straightforwardly from the went with vocal fragments are exceptionally hard to secure great execution where backup is more grounded or most likely performing voice is more vulnerable. To show signs of improvement execution there are strategies risen which will in general separate the performing voice from music backup [4]. There are numerous many sound source partition calculations which separate the performing voice from music backup. Sound source partition can be stated, the assignments of assessing the sign delivered by an individual sound source from some blend of signs comprising of different sources. This is an extremely central issue in numerous sound sign handling errands since investigation and preparing of separated or single sources should be possible with much preferable exactness over the handling of blends of sounds. The term solo learning is utilized to portray algos which attempt to isolate and get familiar with the structure of sound sources in a blended information, in light of data - hypothetical standards, for example, measurable autonomy between different sources, rather than exceptionally modern demonstrating of the source qualities or human sound-related discernment.

Source separation is the process in which several signals are readily mixed together to form a combined signal & the main objective of source separation is to obtain or recover the original component signals from that mixed or combined signal. This is a fundamental problem in many audio signal processing tasks as analysis and processing of isolated sources can be done with far better accuracy than the processing of mixtures of sounds.

The well indeed - known case of a source detachment is the mixed drink party issue. In it, different individuals are talking all the while in a space (for example in a mixed drink party), and a person who plays out the job of audience is attempting to catch up one of the conversations. The human cerebrum is so fit for taking care of this sort of sound-related source detachment issue, yet it gets undeniably increasingly hard for this issue to be explained in computerized signal handling [10]. A few methodologies have been proposed for taking care of this issue yet advancement is at present still in progress.

Many approaches of solving this have been used earlier.

Some Supervised Learning methods like Classification, Support vector machines (SVM), Regression Techniques, Gaussian Mixture Model (GMM) and Anomaly Detection.

And Unsupervised Learning methods like Computational Auditory Stream Analysis (CASA), Beam forming, Independent Component Analysis (ICA), Pitch Estimation and Tracking, Sparse Coding, and Non-negative matrix factorization.

All theseprocedures somehow had some disadvantages and were less precise and accurate. So, research performed by Ms. Monali R. Pimpale, Prof. Shanthi Therese , Prof. Vinayak Shinde in their paper " A Survey on : Sound Source Separation Methods " presented in International Journal of Computer Business In Research Trends in Nov 2016 showed us a better approach in tackling Audio Source Separation by the method "Non negative matrix partial co factorization (NMPCF)".

Let us go through Non-negative matrix particularize before checking out this new technique.

Non negative matrix factorization (NMF) is a low-rank approximation method where a non - negative input data matrix is calculated approximately as a product of two non-negative factor matrices. NMF has been used in many applications, such asimage processing, brain computer interface, document clustering, collaborative predictions, and much more [3]. NMF plays important role in the problem of sound source separation.

The calculations dependent on non - negative framework factorization are proficient and powerful for sound source partition, particularly when the sources or segments of sign are reliant on one another. NMF gives 2 yield grids, one contains the all vocal quality and the other framework shows melodic exercises (which is melodic notes). Late advances in lattice factorization techniques recommend aggregate framework factorization or grid co - factorization to join side data where different various networks (target and side data lattices) are at the same time disintegrated, sharing some factor networks. Lattice co-factorization techniques have been created to consolidate name data, interface data and bury - subject varieties.

Main Disadvantage of this method is that it imposes only the non-negativity constraint.

○A Better Method

Non negative framework fractional co factorization (NMPCF) is a joint network disintegration coordinating earlier information on performing voice and backup, to isolate the blend signal into performing voice divide and backup partition. Notwithstanding the objective network to be factorized, framework co - factorizations can be filled in as a helpful apparatus when side data lattices are accessible. NMPCF was risen up out of the idea of aggregate network factorization or joint decay, which makes the numerous information frameworks to be disintegrated into a few factor lattices while some of them are shared, in this manner, it shows a more noteworthy potential in performing voice partition from monaural chronicles.

3.1.4 Practicality of AO method

This methodology, to my knowledge is more accurate than any other audio only method available yet but practical application is problematic. We are not as technologically advanced yet for Deep Learning calculations to be done in real time on small devices. Super computers can definitely do it but practical applications are limited using this option. So, I have found a better approach for dealing with Audio Source Separation & Speech Enhancement problem which is using Audio-Visual Method.

3.2 Voice Focus using Audio-Visual Model

This kind of methodology, as per my knowledge, was introduced by Google Researchers in 2018. It has been established by their theoretical results that this model can achieve a significant accurate results. Here data fed for training and actual practical implementation includes both auditory and visual samples from the dataset. This feature has its perks and cons in some ways. Perks are better determination among different speakers in a frame and cons being need of more hardware devices to achieve just about fair results. But this model has more perks than cons when new Deep Learning packages being introduced due to technological advancements.This model being inspired by previous research is its advancement and the dataset used to implement this model is AVSpeech (same as used in the previous research).

The AV model in this project is an optimization to target the people speaking English language or languages other than English.

3.2.1 Features of AV Speech Separation Model

This model is involved a multi-stream design which takes visual floods of identified countenances and boisterous sound as information, and yields complex spectrogram covers, one for each distinguished face in the video. The uproarious information spectrograms are then increased by the covers to get a segregated discourse signal for every speaker, while stifling all other meddling signs.

Input Features

This model takes both visual and sound-related features as data. Given a video cut containing different speakers, I've used an off-the-rack face locator (for instance Google Cloud Vision API) to find faces in each packaging (75 face thumbnails totally per speaker, tolerating 3-second fastens at 25 FPS). I've used a pretrained face affirmation model to isolate one face embedding per layout for all of the recognized face thumbnails.

Concerning the sound highlights, I've process the brief timeframe Fourier change (STFT) of 3-second sound fragments.

At derivation time, our partition model can be functional to discretionarily long portions of video. At the point when more than 1 talking face is recognized in an edge, this model can acknowledge different face streams as information.

Output

The yield of this model is a multiplicative spectrogram shroud, which depicts the time-repeat associations of clean talk to establishment obstruction.. Multiplicative veils have been seen to work better than choices, for example, direct expectation of spectrogram sizes or direct forecast of time-space waveforms [8]. Numerous sorts of covering based preparing targets exist in the source division writing, principle two are : proportion veil (RM) and complex proportion veil (cRM).

In a perfect world, the proportion veil is characterized as the proportion between the sizes of the clean and boisterous spectrograms, and is accepted to lie somewhere in the range of 0 and 1 while the mind boggling proportion cover is characterized as the proportion of the perplexing spotless and loud spectrograms.

Given numerous detected speakers' facialbits as input, the network outputs a different mask for each& every speaker &1 for background intrusion. It is found by experiments that cRM is better than RM, so this experiment uses it.

3.2.2 Architecture of AV Speech Separation Model

Sound and visual streams : The sound stream some portion of this representation comprises of widened convolutional layers and the visual stream of this model is utilized to process the info face embeddings and comprises of expanded convolutions. Here, the "spatial" convolutions & enlargements in the visual stream are executed over the transient pivot.

To make up for the inspecting rate error amid the sound and video signals, we upsample the yield of the visual stream to coordinate the spectrogram examining rate (100 Hz). This is finished utilizing basic closest neighbor interjection in the worldly component of each visual element.

AV combination : The sound and visual streams are united by connecting the part maps of each stream, which are thus dealt with into a BLSTM followed by three FC layers. The last yield involves a puzzling spread (2-channels, veritable and whimsical) for all of the information speakers. The relating spectrograms are registered [9].

by complex duplication of the loud info spectrogram and the yield covers. The squared mistake (L2) between the force law compacted clean spectrogram and the improved spectrogram is utilized as a misfortune capacity to prepare the system. The last yield waveforms are acquired utilizing ISTFT.

Multiple speakers : This model backings detachment of various obvious speakers in a video, each spoke to by a visual stream. A different, committed model is prepared for each number of obvious speakers. All the visual streams share similar loads across convolutional layers. For this situation, the took in highlights from each visual stream are linked with the scholarly sound highlights before proceeding to the BLSTM.

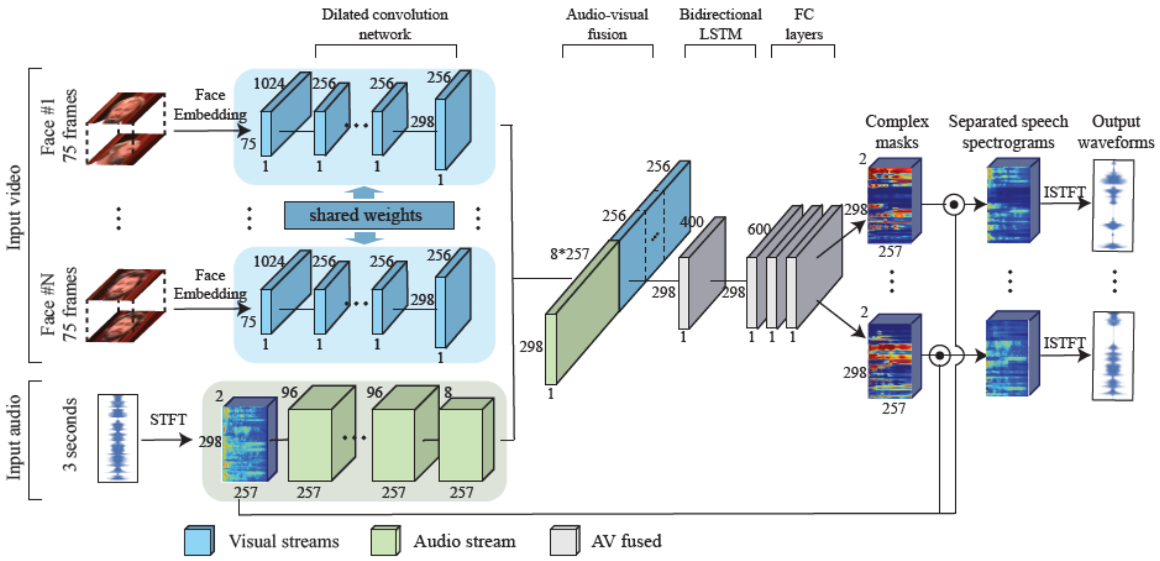


Figure 3.5 : Architecture of AV Speech Separation Model

3.2.3 Implementing AV Speech Separation Model

This framework is executed in TensorFlow and its included exercises are used for performing waveform and STFT changes. ReLU institutions follow all framework layers except for last (spread), where a sigmoid is applied. Group standardization is performed after all convolutional layers. Here, group size of 6 examples is utilized and prepared with Adam enhancer for 5 million stages (clusters) with a learning pace of $3*10^{-5}$, which is decreased considerably every 1.8 million stages.

All sound is resampled to 16 kHz, and sound system sound is changed over to mono by taking just the left channel. At that point, STFT is processed.

The face embeddings are resampled from all recordings to 25 casings for every second (FPS) before preparing and construed by either expelling or repeating embeddings. This outcomes in an information visual stream of 75 face embeddings. Additionally, face discovery, arrangement and quality appraisal is performed.

CHAPTER - 4

OBSERVATIONS AND CONCLUSION

This AV method is tested in different conditions and with different models like state-of-the-art Audio only model & AV speech partition& enhancement.

Audio Only Comparision

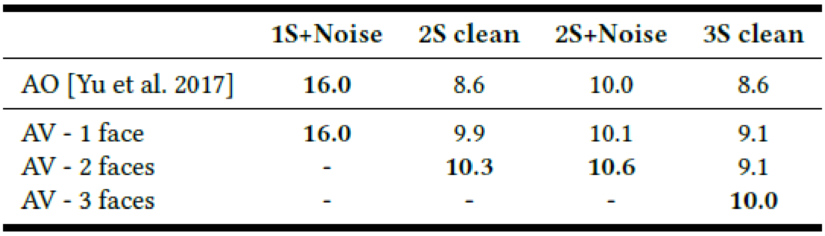
There is no state-of-the-art audio only speech partition or enhancement system publicly available yet & only few datasets are there for training & evaluation. Also, the extensive literature available on this topic require multiple audio channels i.e. multiple microphone setup to achieve some significant results. This won't be pertinent for this model. Hence, this model is executed utilizing an AO pattern for discourse improvement which has a particularly comparable design to the sound stream in the varying media model.

This model when prepared and assessed on the CHiME-2 dataset, broadly utilized for discourse upgrade work, this AO standard accomplished a sign to-bending (SDR) proportion of 14.6 dB, which is close to as great as the best in class single channel aftereffect of 14.75 dB. Along these lines, we can say AO improvement model is in this manner close to best in class pattern.

Audio-Visual Methods Comparison

Since all the current AV discourse partition and improvement techniques are speaker subordinate, it couldn't be effortlessly thought about in the investigations on engineered blends or can run on the chose normal recordings. In any case, quantitative correlations with existing strategies on existing datasets by running this model on recordings should be possible.

Comparison tables of quantitative analysis are listed below.



	1S+Noise	2S clean	2S+Noise	3S clean
AO [Yu et al. 2017]	**16.0**	8.6	10.0	8.6
AV - 1 face	**16.0**	9.9	10.1	9.1
AV - 2 faces	-	**10.3**	**10.6**	9.1
AV - 3 faces	-	-	-	**10.0**

Quantitative Analysis of audio only speech separation & enhancement

Figure 4.1

1S+Noise : 1 speaker + Noise

2S Clean : Clean uninterrupted audio of 2 Speakers

2S+Noise : 2 speakers with some background noise

3S Clean : Clean uninterrupted audio of 3 Speakers

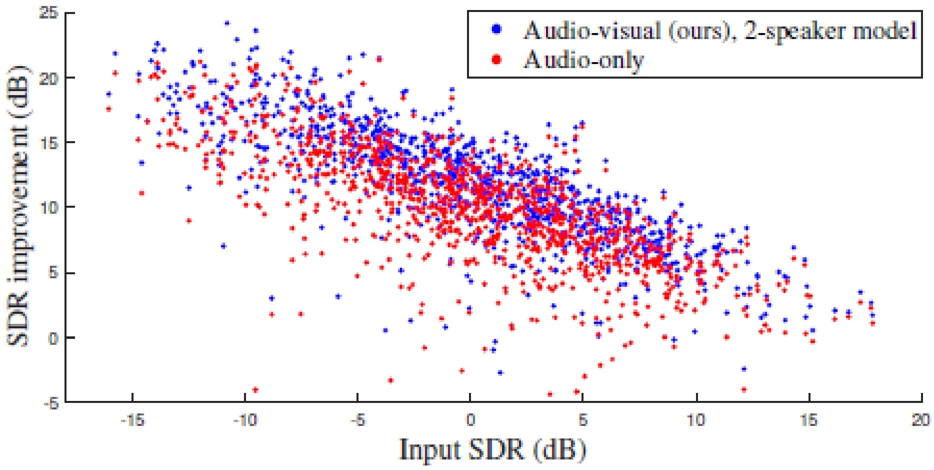
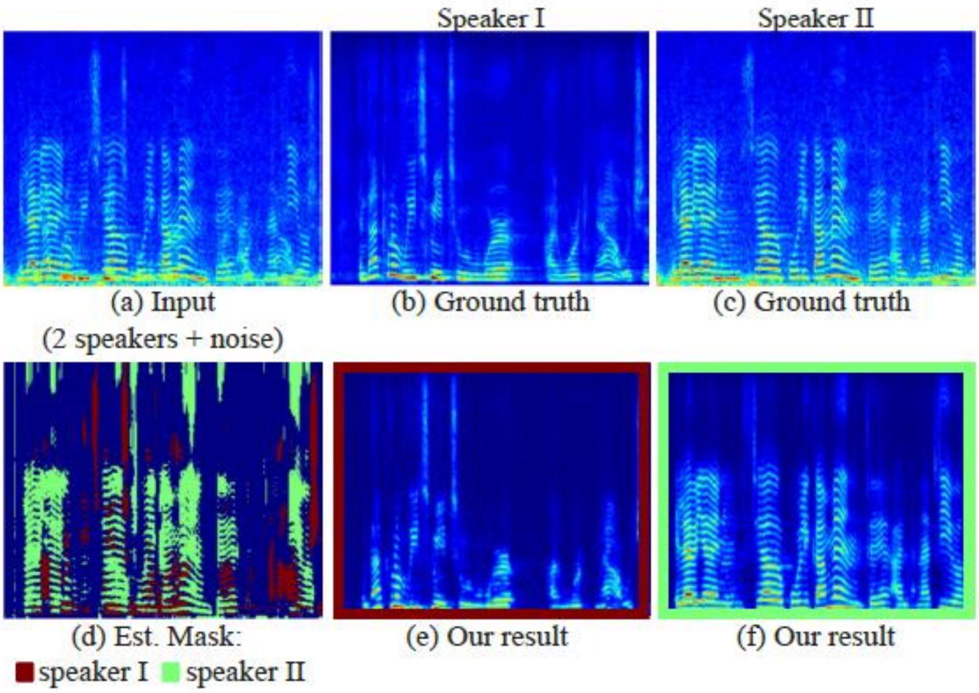


Figure 4.2

Above Figure 4.2 is the scatter plot representating separation performance (SDR improvement) for the task of separation of 2 speakers from original noisy data.



(a) Input
(2 speakers + noise)

(b) Ground truth

(c) Ground truth

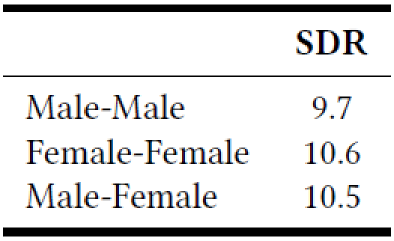
(d) Est. Mask:
🟥 speaker I 🟩 speaker II

(e) Our result

(f) Our result

Figure 4.3

Figure 4.3 shows input and output audio spectrogram for 1 segment of training data which involved 2 speakers + background noise. Here (b) and (c) are spectrograms of each speaker, (d) is the mask to estimate separation of speakers audio signals and (e) & (f) are the output spectrograms for each speaker.



	SDR
Male-Male	9.7
Female-Female	10.6
Male-Female	10.5

Figure 4.4

Figure 4.4 shows Audio separation based on gender. Surprisingly, this model works best in distinguishing female-female mixtures, followed by male-female and male-male (which is also a good result).

CONCLUSION

Audio Visual Methods are significantly better while solving Cocktail Party Problem when the speakers are in the visual frame as compared to Audio Only methods. Hence, there is a significant future of this technology in the upcoming future.

REFERENCES

[1] Z. Fu, G. Lu, K.M. Ting, and D. Zhang. “A Survey of Audio-Based Music Classification and Annotation”,*IEEE Transactions on Multimedia*, VOL. 13, NO. 2, APR 2011

[2] T. Theodorou, I. Mporas, and N. Fakotakis. “ An Overview of Automatic Audio Segmentation”, *MECS*, NOV 2014

[3] M.R. Pimpale, S. Therese , and V. Shinde. “A Survey on: Sound Source Separation Methods”,*International Journal of Computer Engineering In Research Trends*, Volume 3, Issue 11, pp. 580-584, NOV 2016

[4] T. Virtanen. "Unsupervised Learning Methods for Source Separation in Monaural Music Signals" - Tuomas Virtanen

[5] Y. Hu and G. Z. Liu. “Singer identification based on computational auditory scene analysis and missing feature methods”, *J. Intell. Inf. Syst.*, pp. 1–20, 2013

[6] D. Stoller, S. Ewert, S. Dixon. “Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation”, *19th International Society for Music Information Retrieval Conference,* Paris, France, 2018

[7] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman, and M. Rubinstein. “Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation”,*ACM Trans*. Graph. 37, 4, Article 112, AUG 2018

[8] A. Ephrat, T. Halperin, and S. Peleg. “Improved Speech Reconstruction from Silent Video”,*ICCV Workshop on Computer Vision for Audio-Visual Media,* 2017

[9] E. Vincent, R. Gribonval, and C. Fevotte. “ Performance Measurement in Blind Audio Source Separation”,*Trans. Audio, Speech and Lang. Proc.* 14, 4 (2006)