Deep Learning Model to Recognize Punjabi Language Speech Commands

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Pranav Kaushal
(Roll No. 801732034)

Under the supervision of
Dr. Maninder Singh
Professor & Head, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

June 2019

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Deep Learning Model to Recognize Punjabi Language Speech Commands"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Maninder Singh and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature:
Pranav Kaushal
(Pranav Kaushal)

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

(Dr. Maninder Singh)
Professor & Head, CSED

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude to Thapar Institute of Engineering & Technology, Patiala for providing me the platform to do my thesis work at such an esteemed institute.

I wish to express my respect, deep sense of gratitude and indebtedness to my guide Dr. Maninder Singh, Professor & Head of Computer science and Engineering Department of Thapar institute of Engineering & Technology, Patiala for his valuable and enthusiastic guidance, useful suggestions, unfailing patience and sustained encouragement throughout this research work and want to thank him for helping me by providing necessary facilities for completion of thesis work.

I am indebted to all the faculty members and staff of the department for valuable suggestions, friendly support and full cooperation rendered by all of them.

I am grateful to my family and friends for their constant encouragement, support and confidence. I would also like to express my deepest gratitude to my parents for everything they have done for me and I am thankful for their blessings and moral support.

In last, I would thankful to god, without his grace this would never come to today's reality and thanks Him for providing me with strength and ability to complete my thesis.

Pranav Kaushal

(Pranav Kaushal)

ABSTRACT

This dissertation deals with the deep learning model to recognize Punjabi language speech commands. The goal of this dissertation is to recognize the Punjabi spoken words using deep neural network model for classification. In the past research work, the main hindrance in the speech recognition is the achieving the high accuracy. To recognize the speech only GMM-HMM model had been explored. With the new adoption of Deep learning, it has over powered the earlier traditional model. With the advancement in the GPU computing power as well usage of cloud computing where hardware is not feasible for running these huge deep neural networks for training with huge dataset. Deep Learning has flow throughout and has achieved the accurate results that has asserted the fact of it outperforming the GMM-HMM model.

In this research work, convolutional neural network (CNN) deep learning model has been implemented. The dataset comprises of the audio data (.wav files) captured words like " ਹਾਂ ", " ਨਾ ", " ਉੱਪਰ ", " ਥੱਲੇ "in Punjabi Language. Data set has been targeted to achieve a good balance of distribution among the captured audio files with noise for training of the network, to attained the highest accuracy. The training data sets (speech waveforms) are converted into log-mel spectrograms by calculating the duration in each speech clip, duration of each frame, number of Mel filters and the time between each column of spectrogram.

The CNN model architecture comprises of five stack of convolutional layer, ReLU unit and Max pooling unit and further the output from these stacks is passed on to the one fully connected layer. Validation accuracy has been achieved as ≈87% with speech dataset distribution ratio 80:10:10 of the training dataset. The results obtained from this work has shown better performance as compared to the existing work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

In this chapter of dissertation, the speech recognition has evolved around the years with several models, methods and algorithms in the past years. The main challenging in speech recognition of the spoken words, is the high viability in speech signals. This is due to the speakers having different dialects, accents, pronunciations, speaking styles differently with different rates in many emotional states. The additional variability in the automatic speech recognition are reverberation, presence of environmental noise and different recording devices. The representation and feature learning in unsupervised are the areas of machine learning is referred as Deep learning. The speech recognition with Gaussian mixtures for feature detection are successfully swapped at the gradually scale by the mainstream technologies using the deep learning for speech recognition.

In recent times not only academic paper are published, but industries are also working on deploying of deep learning techniques in designing various speech recognition systems. Speech recognition is used in various fields like text analysis, voice search, internet of things, smart devices which are voice enabled like mobile-phone, tablets and many other electronic devices. In this dissertation, words which are commonly used in daily life of Punjab region are identified in Punjabi language. In the Indo-Aryan language Punjabi is the one of the largest speaking volumes and broadly spoken in various part of world predominantly in Indian Sub-continent. Speech is an audio processing of specific frequencies representing features to be recognized. The measurements of Mel Frequency cepstral coefficients of these localized specific frequencies become the target the areas in frequency where its recognition is most relevant. In deep learning, spectrograms are used classify the speech or spoken words audio files into their corresponding classes. Which converts 1D audio files or audio signals into 2D images with time localized frequencies i.e. spectrograms and then used as input to deep learning convolutional neural network to recognize the spoken language.

1.1. Speech Recognition

Firstly, what is speech? Speech is waves of changing air pressure which is realized through excitation from vocal cords and modulated by the vocal tracts which are modulated by the articulate like lips, teeth, tongue. In automatic speech recognition system, speech recognition is the mechanism that enabled human-human and human-computer interactions.

1.2 Classification of speech system

Speech recognition systems are differentiated in various categories on the basis of speaker's representation, variety of speech signals, variety of channel and identify the types of lexis by capability of the speaker.

Types of speech signals are:

- **Vocalization signals**

The vocalization signals of the spoken words in the speech has important significance in the recognition system. Utterances of different words, many words, a sentence, or number of sentences are considered as vocalization signals in the speech recognition system.

- **Speech system for Isolated words**

Isolated word articulation attains the different words or sound at an instance and requires additional gap between the words. Utterance is able to identify the isolated words in the Speech systems. It is relatively easier to implement due to vocabulary tends that are well-defined its benefit and this way utterance limitations are also obtainable. The picking of these diverse confines that affects the consequences becomes the primary drawbacks.

- **Connected word speech system**

The use of articulation of the isolated words together with lesser quantity and in time distribution are called connected words. The particular denotation to the computer is the word of vocalization in the speech.

- **Continuous speech system**

A method used to identify the word of the human beings is communicating naturally, referred as Continues speech. From the computer transcription point of view an additional co pronunciation is included, words running without silences and having separation in the nearby words.

- **Spontaneous speech system**

In spontaneous speech, the method used to recognize the lexis in the vocalizations with stumble and breaches. A typically situation that has been adjusted earlier, or spontaneous speech situation the disfluencies are created, due to disfluencies present it is difficult to distinguish the dialogue. The mispronunciation, repetition, contraction filled pause, false-begins and exclamations are in the category of spontaneous speech.

1.3 Types of speaking style for Punjabi speech recognition system

Speech identification systems are broadly classified into different types based on the unusual voice tones, typical quality of physical body of the speaker.

Speaker dependent recognition system

Speaker dependent methods are usually exacting speaker more precisely without taking account of other qualities of the speakers. Most of the meticulous type of speakers are urbanized. These types of systems are not upgraded to the other models due to its lower cost.

Speaker independent recognition system

The diversity of speakers is identifying by the speaker independent model. It is urbanized to function for any particular type of speaker. The voice reaction method is used to create interactive tones. The accomplishment becomes more complex in the speaker independent system.

Speaker adaptable recognition system

The working of the speaker adaptive system to confirmed with the dependent data and conversions are applied on predictable error rate.

1.4 Speech Recognition Techniques

The ability to listen, understand and after facts of the spoken information are characteristics of the speech recognition system. The classification the speech recognition techniques is based on the different phases to recognition of the speech. Figure 1.1 describe the classification of the speech recognition techniques.

The classification of these techniques is accommodating the all phases of the speech recognition process. The different phases of the process include the analysis, feature extraction, modelling and matching of the speech signals.

Technical aspect of Speech Recognition System

The computational parameters are the technical feature in the optimal word sequence W in speech recognition prediction, X is the speech signal. A posteriori probability is calculated by equation 1.1

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p_{\wedge,r}(W \mid X) \quad \ldots,(1.1)$$

Where $\wedge$ and r are parameters of the acoustic model and language model.

Accordingly, the Bayes rule

$$p_{\wedge,r}(W \mid X) = p_{\wedge}(X \mid W) p_r(W) \mid p(X) \quad \ldots,(1.2)$$

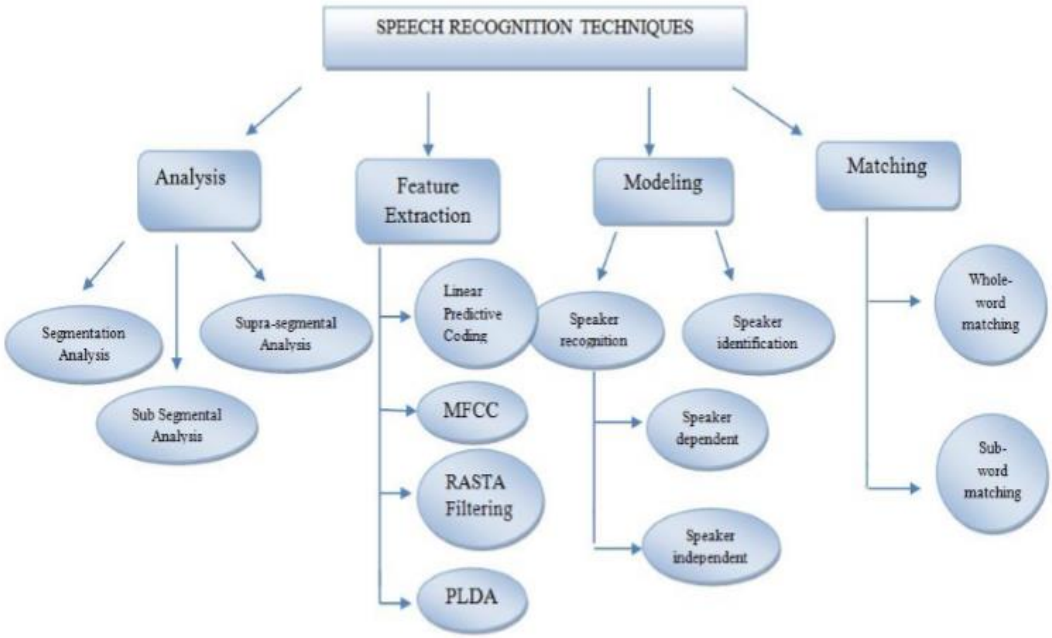


Figure 1.1: Classification of the speech recognition techniques.

Equation 1.2 can be re-written as:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p\wedge(X|W)pr(W) \quad \ldots, (1.3)$$

where $p\wedge(X|W)$ - AM likelihood

$pr(W)$ - LM probability

As the time increases values of it are generates the hidden states by using hidden Markov models (HMMs). The set of state sequences are represented by transcription W. The acoustic feature of the speech signal is computed by the feature extraction method. The probability of the signals\features to be consideration for computations are passed to the acoustics and then language model. The word sequence as the out having highest probability from the acoustic and language models. In speech recognition system, acoustic and language model are called as back end of the system.

1.5 Deep Learning

Artificial intelligence (AI) is a generic field to automate intellectual tasks normally performed by human beings. The domain of the AI encompasses machine and deep learning, includes other approaches having no learning involvement. A machine learning system is explicitly programmed to get trained in comparison with the traditional AI. The rules for automating the

task are presented with many statistical structure examples related to a given task and finds its solutions.

Machine learning is also based on the mathematical statistics, but it is not related to statistics. Machine learning deals with complex and large datasets for which classical statistical analysis becomes impractical. The subfield of machine learning is a deep learning; that follows hierarchy of features to learn from the input data. Figure 1.2 depicts the relationship of artificial intelligence to machine learning and deep learning.

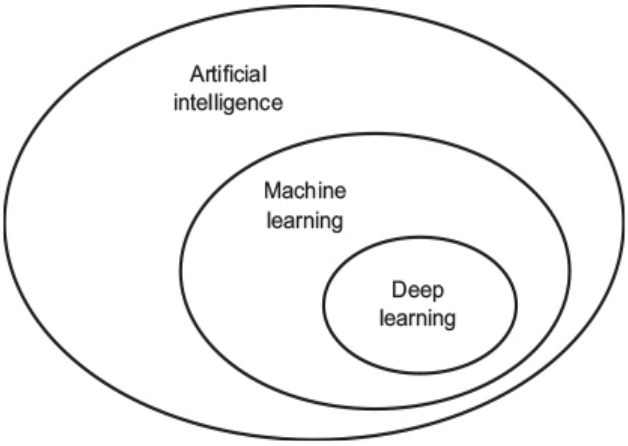


Figure 1.2: Artificial intelligence, machine learning, and deep learning.

Since deep learning attempts to make a better analysis and is able to learn massive amounts of unlabeled data. To model the high-level abstractions in the data sets deep structured learning techniques are used, referred as hierarchical learning. These developed algorithms have layered deep structures and having hierarchical architecture of learning and representing data. The architecture of the hierarchical learning is inspired by machine learning matching the deep layered learning process like in the human brain processing. The features and abstractions are extracted from given data set. The deep learning algorithms are capable of solving the representation of naturally learned and unsupervised data sets.

Deep learning emphasis on successive of meaningful representations of the data taken as the learning representations from data. The deep in deep learning is a reference to deeper layers in the applied in the approach to get the meaningful representation of the features in the data. The depth of the model is defined by number of layers participated in the model.

Deep learning models consists of hundreds of layer representations and get trained automatically form the training data. The machine learning approaches are called shallow learning that involves the learning with one or two layers of representations of the data sets.

The layered representation is structured stack wise on each other and modelled by the neural networks. The deep learning is a mathematical framework for learning representations from data.

1.6 Deep Neural Networks (DNNs)

A deep neural network is also stimulated like biological nervous systems. It contains number of nonlinear processing layers and parallel operated. It structured as an input layer having multiple hidden layers and give the results as an output layer. All the nodes, or neurons in the layers are interconnected with each hidden layer neurons and with the output layer. The architecture of deep neural network is shown in figure 1.3.

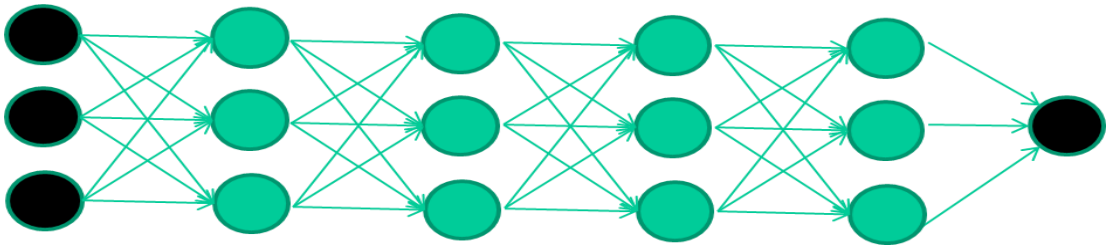


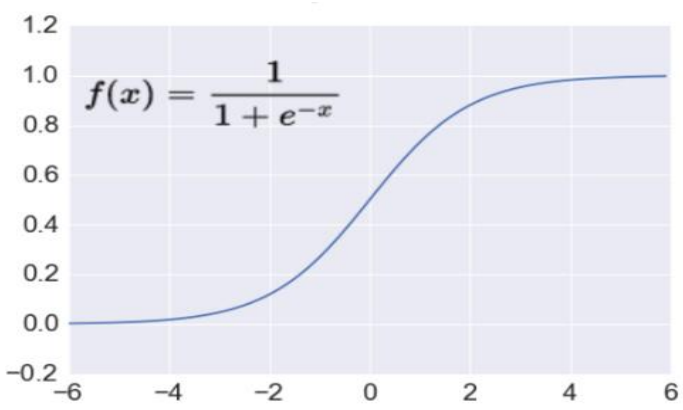
Figure 1.3: Deep neural network

The working example of DNNs is explained here, by considering a set of images in which each image has single of four categories of the object. The training of deep learning network is automatically recognizing the object in each image. The labelled images are used in order to training network for given data. The network starts to understand the training data of the specific object features and mapped to the related category.

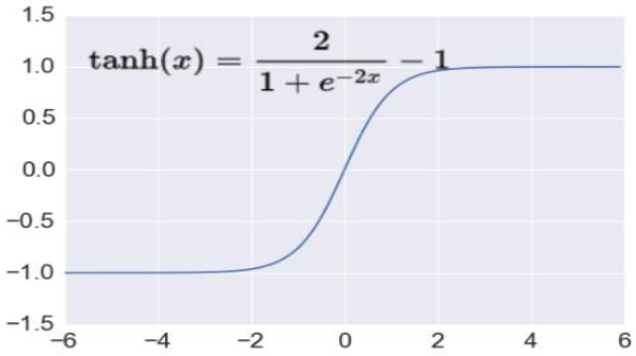
The layers in the network takes input data from the previous layer in the network and transforms the same to the next layer. The complexity of the network increases due the detail of the data flows from layer to layer. Deep neural network learns from the input data without considering the effect of the features are being learnt during the training.

Activation functions: In Deep neural network the activation functions are used to train the network models as in the traditional artificial neural networks.

Activation Sigmoid: Takes a real-valued number and squashes it into range between 0 and 1.



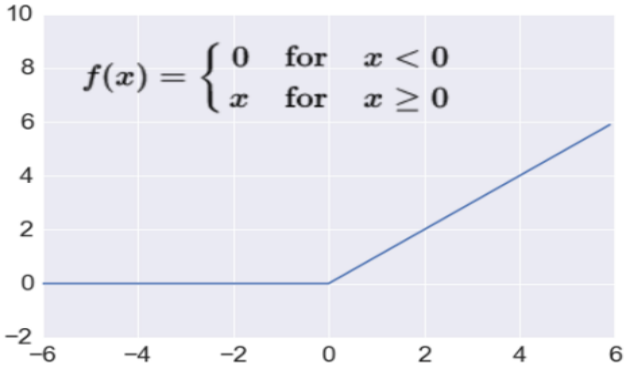
Activation: Tanh: Takes a real-valued number and "squashes" it into range between -1 and 1.



Activation: Rectified Linear Unit (ReLU): Takes a real-valued number and thresholds it at zero.

Nowadays, the Deep Neural Networks uses ReLU activation which trains the network faster and convergence accelerates by using the its linear, non-saturating forms. The operations are implemented by thresholding a matrix that reduces its operation expenses in comparison to the sigmoid/tanh (exponentials etc.) functions.

$$f(x) = \max(0, x)$$



A deep neural network architecture having several layers is as shown in figure1.4 for classify the digit. The transformation an image of a digit is performed in order to recognize what digit it is. The Deep representations learned and transform an image in the model are shown in figure 1.5. In this figure each layer having all input information about representation the digit.

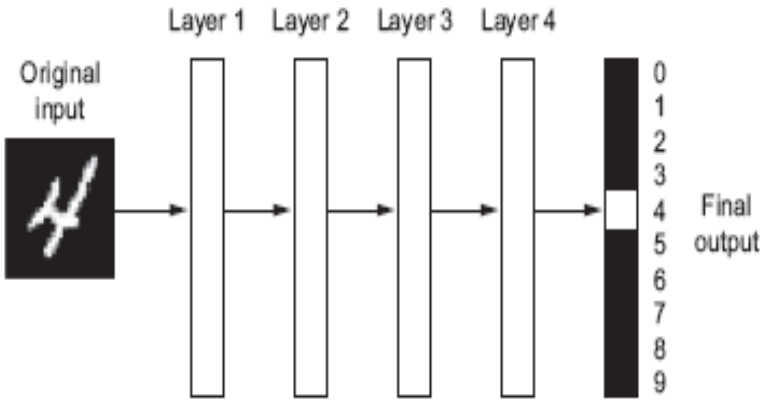


Figure 1.4: Deep neural network for digit classification

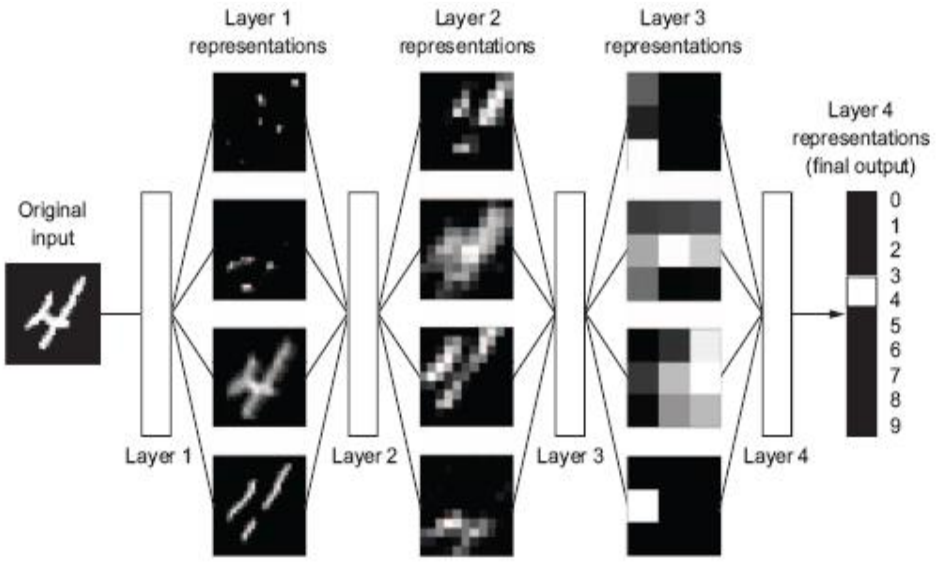


Figure 1.5: Deep representations learned by a digit-classification model.

1.7 Deep Learning Algorithms

Deep learning algorithms are investigated extensively in the recent years. The researchers explored the deep learning for solving challenging problems, other related models used to solve real life problems. Restricted Boltzmann machines (RBMs) and Convolutional neural networks (CNNs) are the main algorithms in the deep learning.

Restricted Boltzmann Machines (RBM)

RBM is probabilistic multiplicative model having visible and hidden units. The input in the model is a visible unit in the form input vector and features represented in the hidden layer. Visible unit is connected to hidden unit each other. Figure 1.6 illustrates the graphical model of restricted Boltzmann machine.

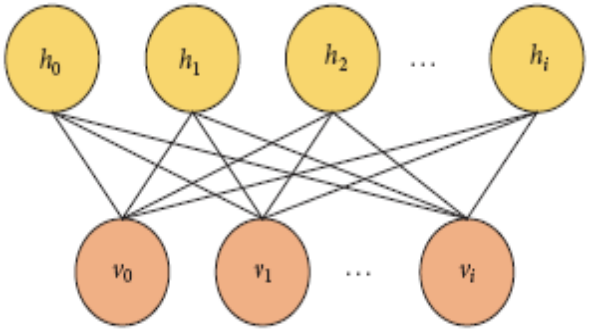


Figure1.6: Restricted Boltzmann machine.

The energy function of a RBM as given by equation

$$\text{Energy } (v, h; Ø) = -b^T v - c^T h - h^T W v \quad \text{..., (1.4)}$$

where $Ø = \{W, b, c\}$ are the parameters of RBM. W represents weights of the visible layer and hidden layer; the bias between the input layer and hidden layer is denoted by b and b, respectively. The energy function is represented here by due to the result of the lack of hidden-hidden and input-input interactions.

Deep neural networks (DNNs) are build and train using RBM. The procedure of a DNN building and training has two steps. The stacking of restricted Boltzmann machines (RBMs) used to configure a belief deep network. The initial model of the deep neural networks is pretrained with DBN.

Convolutional Neural Network (CNN)

The main classes of deep neural networks are convolutional neural network used for image processing tasks. It is mainly used in the computer vision applications. The convolution, subsampling, and full connection layers are three layers of the convolution neural network. Figure 1.7 depicts the architecture of convolutional neural network.

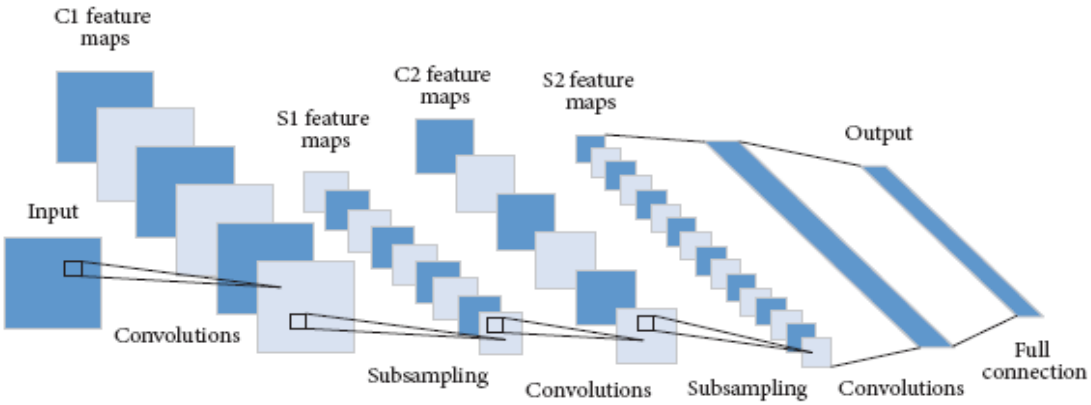


Figure 1.7: Architecture of convolutional neural network.

Convolution Layer: The digital image and convolution matrix are the input and convolution layer. The convolution layer takes the convolution of the input image to compute and results the convolution matrix are generated. The output image is generated by convolution matrix and input of convolution of the input image.

Subsampling Layer: The subsampling layer is used to maintain invariance and robustness of the convolutional neural network. The max pooling is used for subsampling layer in image related tasks. The image consists of blocks and pixel value are matching with the output. The purpose to use this layer is to get faster convergence, translation and rotation in the input pattern.

Full Connection Layer: It is a traditional feed-forward neural layer. This makes the neural network fed with vectors of predefined length.

Deep belief networks (DBNs)

Deep Neural Network RBMs is stacked and trained in a greedy manner called deep belief networks (DBNs). DBNs models are represented as graphical and learn to find deep hierarchical representation in the training data. The architecture of the deep belief networks is shown in figure 1.8. A DBN model with l layers and distributed in the observed vector v and ℓ hidden layers hk as follows:

$$p(v, h^1, \ldots, h^l) = \left(\prod_{k=0}^{l-2} p(h^k | h^{k+1})\right) p(h^{l-1}, \ldots, h^l)$$

where $v = h^0$, $P(h^{k-1} \mid h^k)$ is a conditional distribution for the visible units. The hidden units of the RBM at level k and $P(h^{\ell-1},h^{\ell})$ is the visible-hidden joint distribution with conditioned on the top-level RBM.

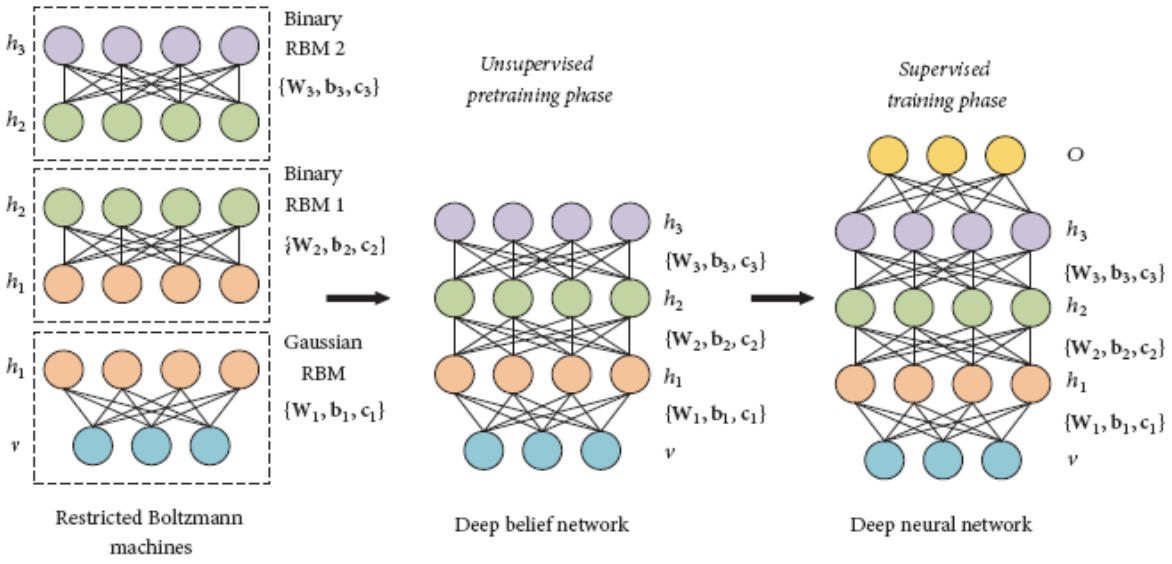


Figure1.8: Architecture of deep belief networks.

1.8 Thesis organization

- Chapter 1: In the first half of this chapter, it gives introduction about speech recognition, that how a speech is produce, different types of speech and components of speech are explained, basically it presents how a raw audio is processed by speech recognition system. second part deals with evolution of deep learning from artificial intelligence and which model are used for speech recognition.
- Chapter 2: This chapter presents the findings of work done in the past in the field of speech recognition using traditional approach. It also presents finding done in the field of deep learning with speech recognition.
- Chapter 3: This chapter defines the research problem and presents the gaps between the current domain of research and earlier research done on it. It also describes the objectives of this research topic.
- Chapter 4: This chapter explained methodology adopted for this thesis work. Firstly, how a dataset for a particular language is collected and how it is further refined on certain parameter which is than, used for the training the network. It also describes architecture of deep learning model which is used to trained the network on the dataset.

- Chapter 5: This chapter presents the result and discussions of the experimentation performed on the proposed model.
- Chapter 6: This chapter states the conclusions on the finding done in this thesis work and its scope for future works.

CHAPTER 2

LITERATURE SURVEY

2.1. Literature Review on traditional method and Deep learning in speech recognition

The detailed literature survey has been carried out for the research work related exiting research work in deep learning for audio and speech applications:

Nur Farhana Hordri et al. (2016), presented a survey of the different applications of the deep learning and provided the references to researcher's community for develop new models with deep learning in future research. There are seven applications that have been applied with deep learning were identified, namely speech recognition, image recognition, natural language processing, drug discovery and toxicology, customer relationship management, recommendation systems and bioinformatics [1].

Danfeng Xie et al (2017), discussed the solutions of problems related to classification of images, speech recognition, signal processing, and natural language processing using deep learning algorithms. In their study, computer vision application based on deep learning algorithms is also discussed. The deep learning algorithms has been implemented to specific areas such as road crack detection, fault diagnosis, and human activity detection. They also explored the challenges of designing and training of deep neural networks [2].

Szu-Wei Fu et al. (2018) explained to clean the noisy data with the speech enhancement model. An objective function is applied in the input to calculate the model parameters. The utterance has been computed by apply the speech enhancement framework based on fully convolutional neural networks (FCNN). The temporal associated information in the speech signal has been explained by calculation of objective function based on the perception. The results of MSE in their approach has been observed better for the targets evaluated [3].

Yong Xu et al, (2015) described pre-processing methodology to smoothen the speech signals by using the supervised method in the deep neural networks (DNNs). The speech and type of noise in the training data has been considered in the model. The capability of the model has been enhanced by the regression function in the model. They also explained the techniques to improve over smoothing of the training model with the noise training strategies. The experimental results observed had been in higher accuracy range than the conventional MMSE based technique [4].

De Liang Wang (2018), studied speech separation problem in the signal processing. They explained the speech separation as a supervised learning problem. The training data set contains

the speech of the speakers and background noise. An overview of the deep learning based supervised speech recognition has been explained. They also discussed main components required in the training the model with specific acoustic features. The important issue of generalization, unique to supervised learning is also discussed [5].

Li Deng and Xiao Li (2013) discussed the machine learning techniques used in the automatic Speech Recognition system. They also explored latest techniques of the machine learning speech recognition system and considered the relevant features to achieve better accuracy. The learning paradigms included in their research work presented has been based on generative, discriminative, supervised, unsupervised, semi-supervised, and active learning. They also explained the advancing ASR technology based on sparse representations using deep learning [6].

A. E. Omer (2017), described the importance of the feature selection in the speech signals to designed automatic speech recognition technology. Speaker recognition has been studied concerned feature extraction and speaker modelling. The MFCC and FFT algorithms has been computed the recognition accuracy and text dependency [7].

J. Martinez et al. (2012), explained that speech recognition is achievable by using mfcc feature extraction method from audio signal and then computing the values in recognizing the audio by vector quantization method [8].

Singh, Satyanand, et al. (2011), has been explained the extraction of features used in the approach. This technique has been modelled with standard acoustic feature set of the human auditory system for speech related applications. The performance of the model has been computed by MFCC and inverted MFCC with traditional filters with recorded speech signals [9].

Anjali Jain et al. (2013), discussed the brief survey on Automatic Voice Recognition in their study. Various stages of voice recognition system are based on the technological perspective and an appreciation of the fundamental in area of voice communication [10].

Bharti, R and Bansal, P (2015), discussed Automatic Speaker Recognition (ASR) system based on the mathematic algorithm and recognition of the speaker on a real-time. The biometric type of the features played a vital to recognition of the speaker. They also considered the parameters based discriminative nature of the speaker vocal exist in speech signals. And performed clustering as set of acoustic vectors present related to the speech signals [11].

Majeed, S. A et al. (2015), discussed the Mel Frequency Cepstral Coefficients (MFCCs) are the main components features of the speech signals used to recognition feature in speaker and speech recognition applications. They explored different design methods of the filters required

in calculation of the MFCC coefficients to improve the recognition accuracy speech recognition systems [13].

Irakli Kardava et al. (2016), Explained the problem facing for speech recognition system in language pronunciation in different accent [33].

Waghmare K., et al. (2015), described the usage of Mel frequency cepstral coefficients for accent identification from speech database of Hindi languages which consists of Hindi, Marwadi, Urdu, Marathi speakers [31].

Arshdeep Singh, et al. (2016), described the different dialects of Punjabi by identifying similar dialect with help of words, which are found in other dialect of Punjabi language [32].

Yong Xu et al, (2015), described the mapping functions of the noisy and clean speech signals by applying the supervised technique based in the deep neural networks (DNNs) [4].

Y Zhang, (2013), discussed the Gaussian coding features in the deep learning algorithm as unsupervised feature learning. The speech to text conversion has been performed their research of automatic continuous speech recognition [22].

Tiken Moirangthem et al. (2018), described the deep neural network is used for recognition model. The model is experiment on digits of Assamese language using Mel Frequency Cepstral Coefficients for extracting speech features and Long Short-Term Memory is used as primary layer in neural network model [37].

Shyam Agrawal, *et al.* (2012) describes the method to collect the text and speech data in Indian English and hindi languages has been used in the mobile communication. The personal communication in the native languages of data has been collected as from the messages in the communication. The collection of data has been taken from different speaking population in the related age groups, sex and dialects. The database contains in 12 domains of Hindi and English language. The language grammar rules were used to clean text raw data based on slangs and unconventional. A special software has been used to create the records containing 630 phonetically rich sentences. The prompt data set was recorded of the 100 speakers and simultaneously in three channels. The collection of utterances as the speech database was also recorded from the text. The Indian English consonants and vowels as the phonetic lexicon of Hindi and Hindi consonants and vowels as the lexicon of the English has also been recorded in the database [23].

F. Ordonez et al, (2016) described the heuristic processes to solved the engineered features in the human activity recognition system. Researcher discussed the automate feature extraction from the sensor inputs. The capturing temporal dynamics in human activities are considered the motor movements as complex sequences. The sensor inputs are explored to automate the

feature extraction using deep convolutional neural networks. Based on the time series domains the success of recurrent neural networks frame work is discussed the deep framework for recognition with both convolutional and LSTM recurrent models. The public activity recognition challenges are used to evaluate the framework. The results of the study have been applied to improve performance of fused multimodal sensors and homogeneous sensor modalities. The activity recognition from wearable sensors based on convolutional and LSTM recurrent deep architecture has been demonstrated in their research. The framework performed the high accuracy in the results for given dataset [24].

Shweta Bansal et al. (2016), designed and developed speech database of annotated data set for Punjabi language. The developed corpus has been explored for synthesis in phonetic study in the speech systems. The scripts of Gurmukhi write/transcribe the in India and Shahmukhi as well as in Pakistan. The Punjabi dialects includes Majhi, Doabi, Malwai, Powadhi, Pothohari, and Multani. Their study explored the dialect of the Punjabi language: in the Malwai dialect of 35 letters in the Gurmukhi alphabet. The most commonly dialects out of the spoken in Punjab, India are Majhi, Malwai and Doabi. In the design of the corpus, vowel and pure consonants has been recorded in the data sets. The classification of consonants as well as the articulatory of phonemes and tonemes of the Punjabi language has been considered in their research work. The recorded database has been segmented and labelled phonemically and each phoneme of the Punjabi language [25].

G. Hinton et. al. (2012), discussed the comparison in the machine learning algorithms that made a significant advancement in automatic speech recognition. The most commonly applied hidden Markov models (HMMs) has been considered in their comparative study. The coefficients of the features have been determined by manipulated the Gaussian models of a frames in the window in the data sets. Expectation-Maximization (EM) algorithm has been developed for speech recognition systems for real world tasks. The relationship of the features coefficients computed by Gaussian mixture models (GMM)and HMM states has been used as acoustic input. The training of the feedforward neural network has been performed by taking coefficients as input and output produces HMM states. New methods have been trained with number of hidden layers to outperform the accuracy of the speech recognition system [28].

Yogesh Kumar and Navdeep Singh, (2017) developed spontaneous Punjabi speech corpus for the Punjabi language. The corpus has been employed in the study of spontaneous speech recognition system. Java programming has been used as interfaces for live Punjabi speech system. 6012 Punjabi words and 1433 Punjabi sentences have been considered to train the

automatic speech system, the recognition performance accuracy was measured in terms for 93.79% for Punjabi words and 90.8% for Punjabi sentences [29].

CHAPTER 3

Problem Statement

3.1 Problem Definition

In any language, the different linguistic variations are used to speak that language. Among the different speakers' different variations are categories by regional variations as word selection and grammar (dialects), in pronunciations (accents), and by sociological variations as in different speaking styles due to age, situation and gender. In our daily life, Speech recognition systems are increasingly becoming indispensable part.

In speech recognition system, there are number of problems to recognition the specific natural language. The most common problems of speech recognition are created due to differences in pronunciation, in accent and intonation, while speaking the original language. In the literature survey it has been observed that gaps in development of speech recognition system based on the speech variances for different speakers based on the dialects or accents of the spoken language. The accuracy of the recognition is the main constraint in the speech recognition system for its different applications. There are number of speech recognition systems having some solutions but they do not solve all problems.

The general gaps in present research work are representation of the data that it is given as features that effects the performance of the system and ability to extract patterns from the raw data.

Another scenario that observed in the literature survey that it is difficult to capture high-level abstract feature from input data, to know which features need to be extracted. It is also difficult to capture data by a system how exactly the human brain captures information to learn.

3.2 Research Gaps

Major gaps that need to be solved are.

- To achieve the better accuracy of the model that takes the inputs captured from the real world scenario like data captured with noisy environment.
- The accuracy in recognition of Punjabi language data set needs to be improved.

In speech recognition systems the acoustic input is typically represented by Perceptual Linear Predictive coefficients (PLPs) or concatenating Mel Frequency Cepstral Coefficients (MFCCs) are computed from the raw input waveform, and their first- and second-order temporal differences. These non-adaptive pre-processing of the waveform having large amount of

information in waveforms that is considered to be irrelevant for discrimination and considered as major challenge in achieving the better accuracy in speech recognition. The Punjabi language has been considered in the present research work. A deep learning neural network technique is proposed to solve the speech recognition system for Punjabi language.

3.3 Research Objectives

The research objectives are:

- To study and explore various speech recognition system based on different languages.
- To design and implement a deep learning based speech recognition system for Punjabi language.
- To test and validate the results.

CHAPTER 4

Research Methodology

4.1 Data Collection

A method of organizing speech signals corresponding to the sequence to be expressed by the spoken utterance is required in this speech recognition technique. In the recent, Deep Neural Networks (DNNs) have been employed at on different applications, including automatic speech recognition, visual object detection and image classification. The main features of DNNs is their capability to learn features in the speech signals and classify them. To achieve a high computational power of the DNNs requires larger amount of data to process. In the present research work, there is no publicly available data set of the Punjabi language for speech recognition system, so the large dataset preparation is a task itself. The data set has been recorded/created using tools like Audacity and Matlab from different human being speaking Punjabi language within the Punjab state of India. For the experimental purposes data sets of audio file are recorded in a non-controlled environment. The created data sets are used in this speech recognition as training dataset of speech signals and testing dataset for recognition. In the training level, input information has been fed in the input module of the system. In the testing level, mapping of the computed features with input has been evaluated by output module of the system. The features based on the speaker's dependency, vocalization manners, pronunciations, dialect, style of speech and detection surroundings typically playing a role in the correctness of recognition system.

4.1.1 Recording

The data set comprises of one-second period of audio files in wave format of the single spoken Punjabi words from different part of the Punjab state. These words are spoken by different speakers and from a small set of commands. The audio files are arranged in different directories based on the name of the word spoken. The data set has been designed to train, test and validate the deep learning models. This recording has 200 samples for each person. Data for 50 males and 50 females is recorded in order to achieve a good balance. These audio files are recoded using audacity software and matlab program. The audio files are captured with sample frequency rate is set to 16 KHz, in a 16-bit little-endian PCM-encoded WAVE file format. The audio files are arranged into labels by name of the directories. The main goal is to capture sound of the Punjabi speaking peoples as single-word, instead of sentences used in

conversations. The words recorded as command by fifty persons, each speaker spoken each word by five times. The Punjabi words are " ਹਾਂ ", " ਨਾ ", " ਉੱਪਰ ", " ਥੱਲੇ ", " ਖੱਬੇ ", "ਸੱਜੇ ", " ਉਤੇ ", " ਬੰਦ ", " ਠਹਿਰੋ ", " ਜਾਓ ", " ਸਿਫ਼ਰ ", " ਇੱਕ ", " ਦੋ ", " ਤਿੰਨ ", " ਚਾਰ ", " ਪੰਜ ", " ਛੇ ", " ਸੱਤ ", " ਅੱਠ", and " ਨੌਂ ". The original audio files are collected in uncontrolled locations by people and design of the data set examples of the sort of speech data that may likely to encounter in consumer and robotics applications, where it is not possible do have much control over the recording equipment or environment.

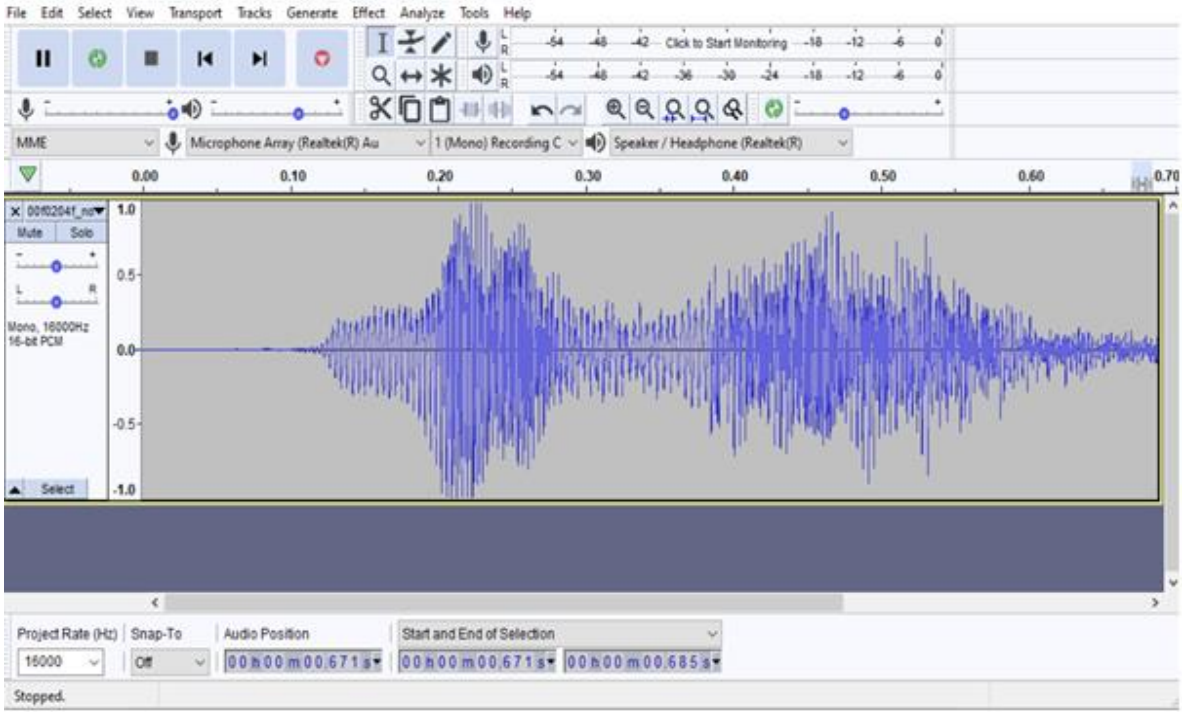


Figure.4.1: Audacity tool for recording data

4.1.2 Labelling

The recorded audio files are labelled by names of the files in the random names using number of digits for multiple recorded file of the same speaker. A very simple approach has been followed to label the data files with the help of the labelling the process Ant-renamer 2.12 software. The names of the audio files are mapped as the labels. The digit after the second dash i.e. − is the correct label for an audio file. The labels are assigned from the name of the audio itself hence, each audio is named in this fashion and these labels have been hot encoded. One hot encoded is a group of bits which represent a unique label. These are combination of 8 number values which comprise of single high bit and the other bits are low. e.g. 800210 -0 -1.wav: The digit after the second dash is - is 3. So, the audio wav for the word has been labelled as 3. Figure 4.2 depicts the rename the batch by using regular expressions.

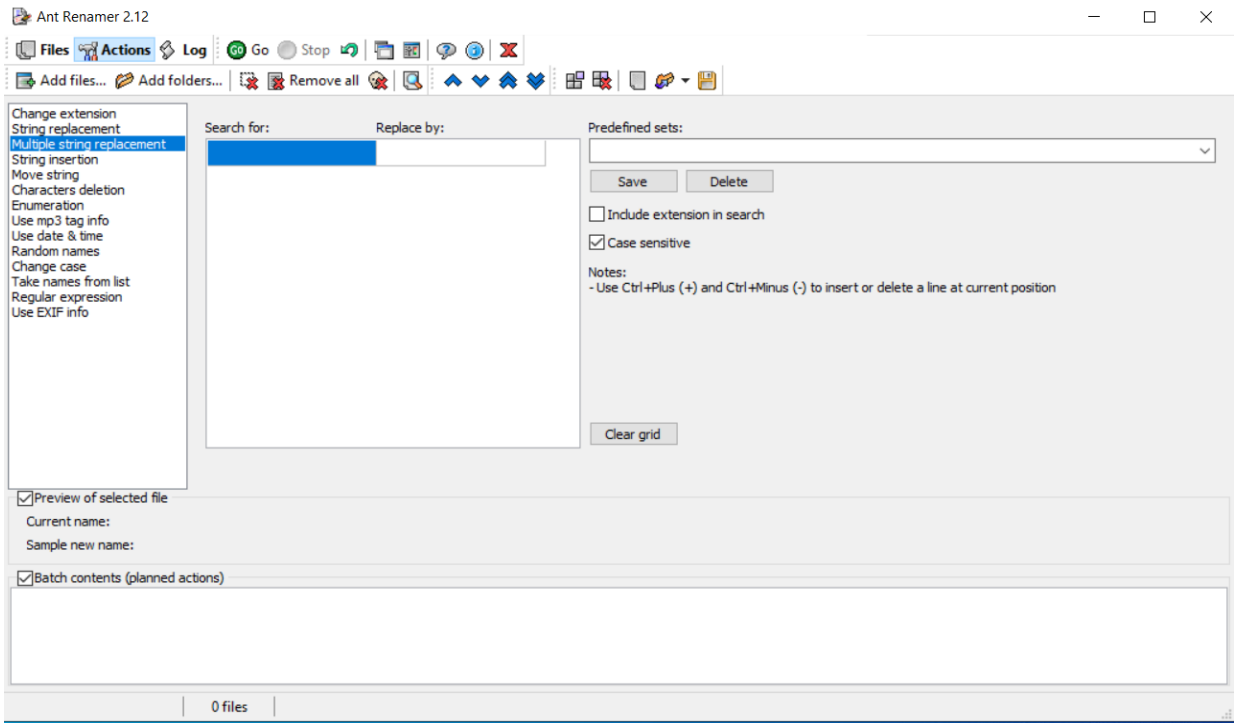


Figure.4.2: Rename the audio files in batch by using regular expressions.

4.2 Data Pre Processing

4.2.1 Reduction and Silence Removal

The noise reduction and silence in the recorded audio files has been done using the Wave Pad software and matLab program. Figure 4.3 depicted the wave file loaded and process of execution on which the noise reduction is to be applied. Also, Figures 4.6 and 4.7 depicts the Audio wave as the wave file input before and after Noise Reduction algorithm applied.

The End-point detection technique has been employed to identifying speech parts in audio signal and removing the silence part in audio files. The isolate word is obtained by removal of silence before and after the isolated word.

Classifying labelling events in speech using three state representation:

Silence: where no speech is produced.

Unvoiced: where vocal cords are not vibrating. This is aperiodic and random in nature.

Voiced: where vocal cords are vibrating. This is almost periodic in nature.

Silence and Unvoiced are classified together because these result in low energy content.

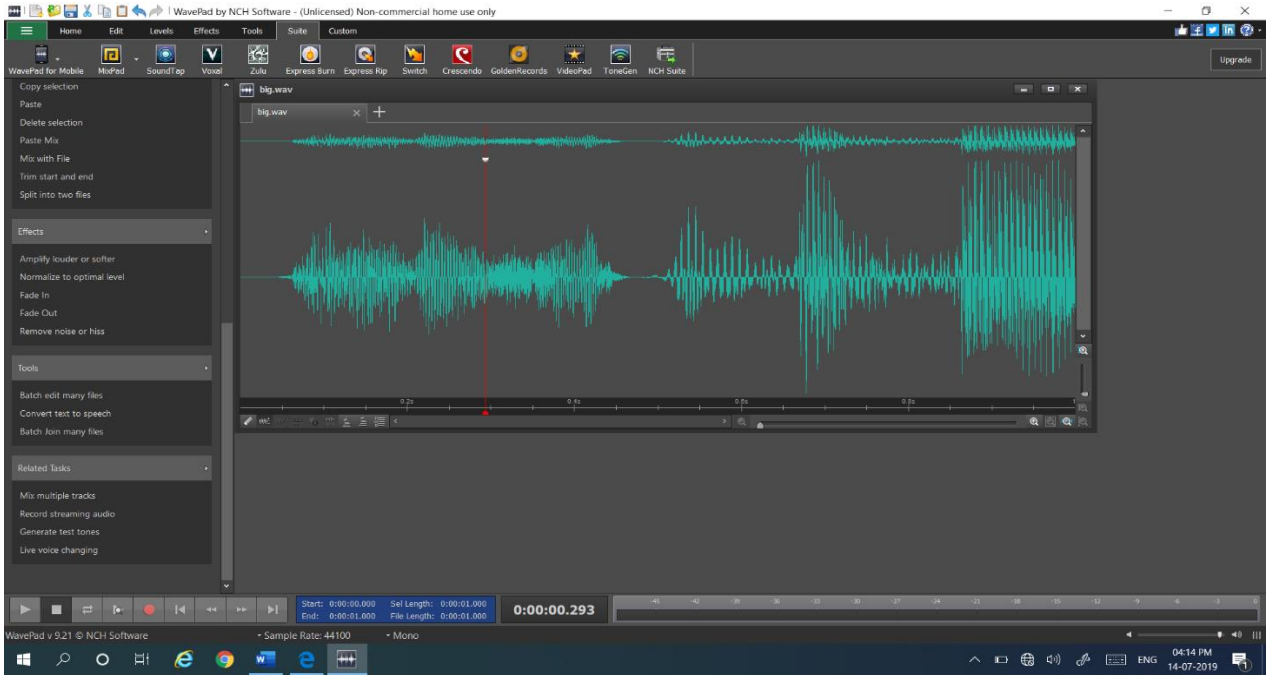


Figure.4.3: The wave file (input) loaded on which the noise reduction is to be applied in Wave Pad.

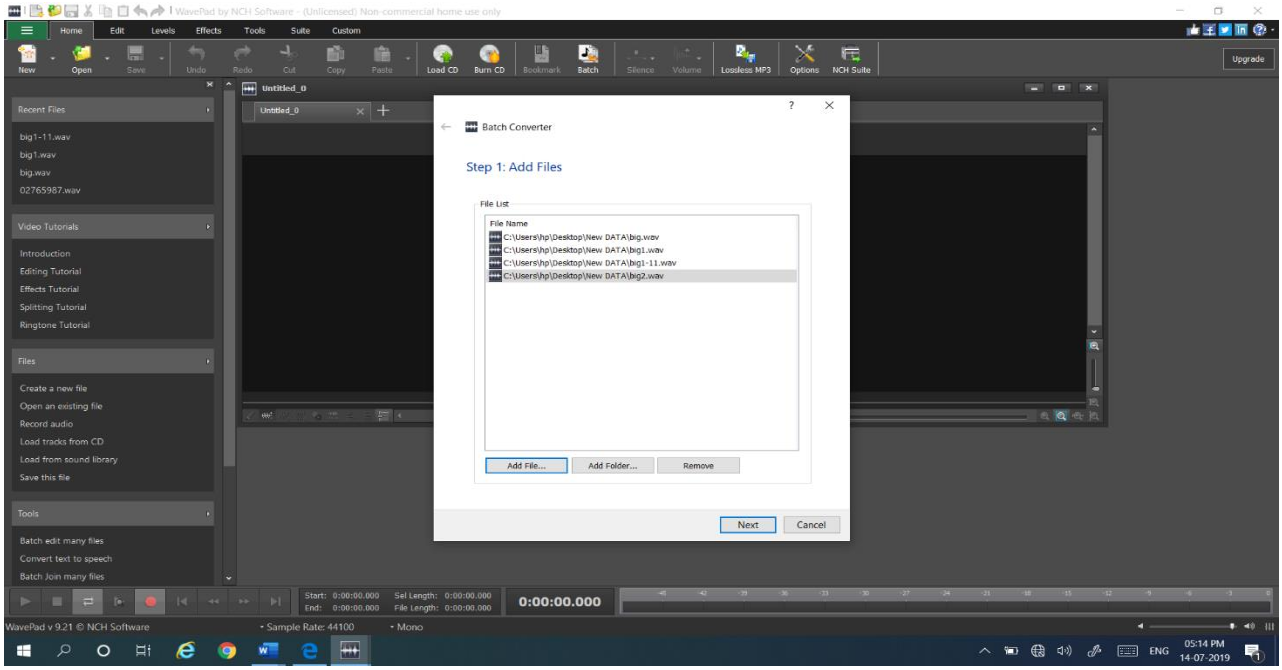


Figure.4.4: The first step of the batch of wave files (input) loaded on which the noise reduction is to be applied in Wave Pad.

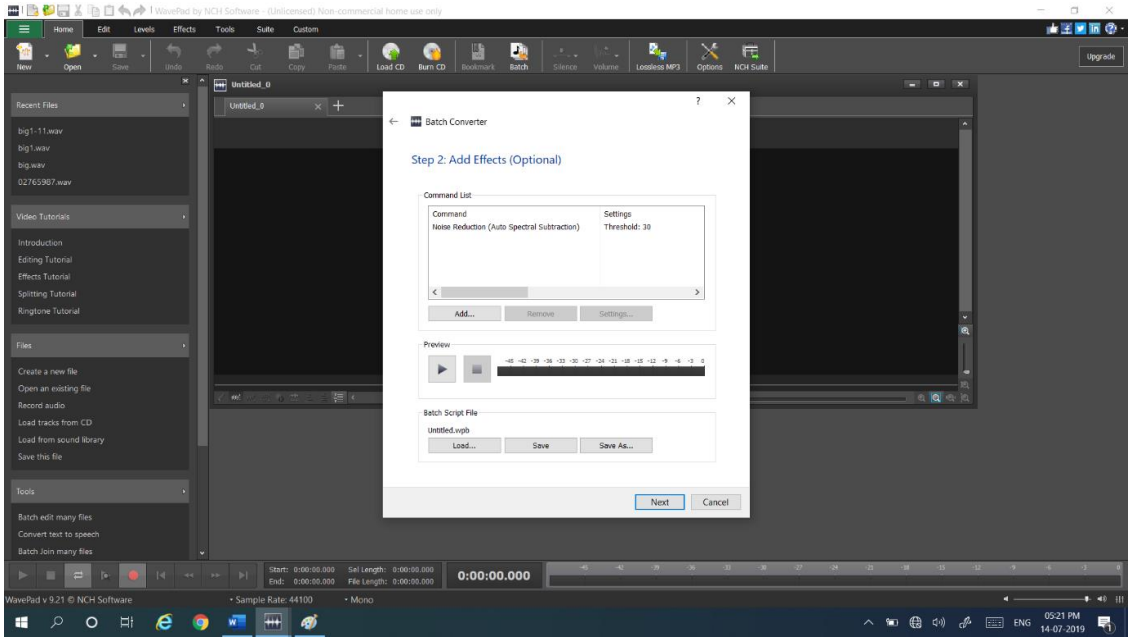


Figure.4.5: The application of Auto Spectral Subtraction Noise Reduction algorithm.

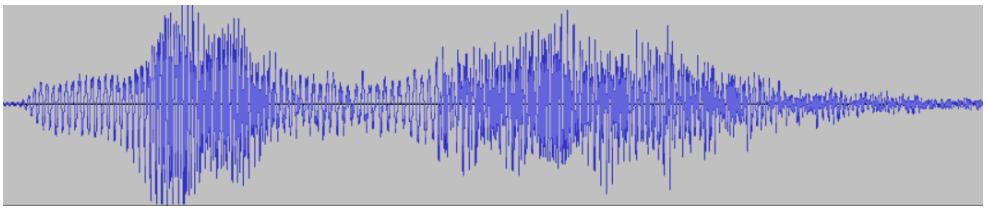


Figure 4.6: Audio wave shown as the wave file input before Noise Reduction.

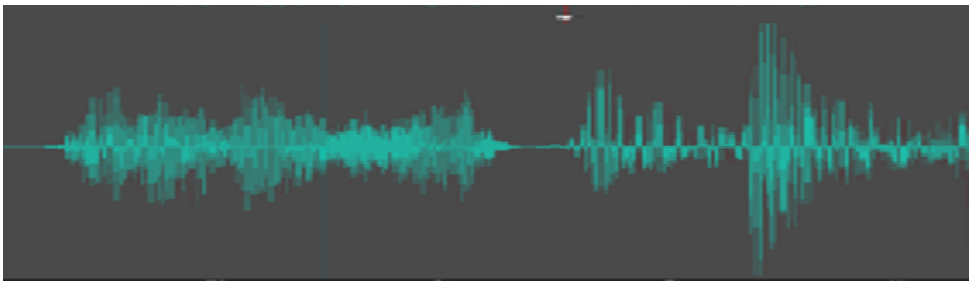


Figure 4.7: Audio wave shown in the wave pad editor after Noise Reduction.

4.2.2 Feature Extraction

The feature extraction of Punjabi speech involves components of speech signal that relates the audio association with vocalization. These features are computed through by processing of the Mel-frequency cepstral coefficients (MFCC) in the audio waveform. The representation of the signal as vocalizations of cosine transformation algorithm of the signal in the form of output. The articulation of the short term energy spectrum on the scale of Mel frequency is calculated by using the FFT of the speech signal.

The deep learning approach has been used to identify people based on features extracted from signals of the speech. To train the classifier having features as: pitch of the voiced segments of the speech, and the Mel-Frequency Cepstrum Coefficients (MFCC). In the speaker identification, the speech signals under considerations are mapped with computed signals in the nearest mapping classification.

The proposed approach used in the present research work for the features recognition to speaker identification is shown in the Figure 4.8. Various speaker's sound has been recorded for speaking Punjabi language and the components in the signals as pitch and Mel-Frequency Cepstrum Coefficients (MFCC) are considered for the classification of the recorded signals. Deep learning model has been used to train by using these features. The speech signals for the testing model or to be classified are recorded by the same recording parameters

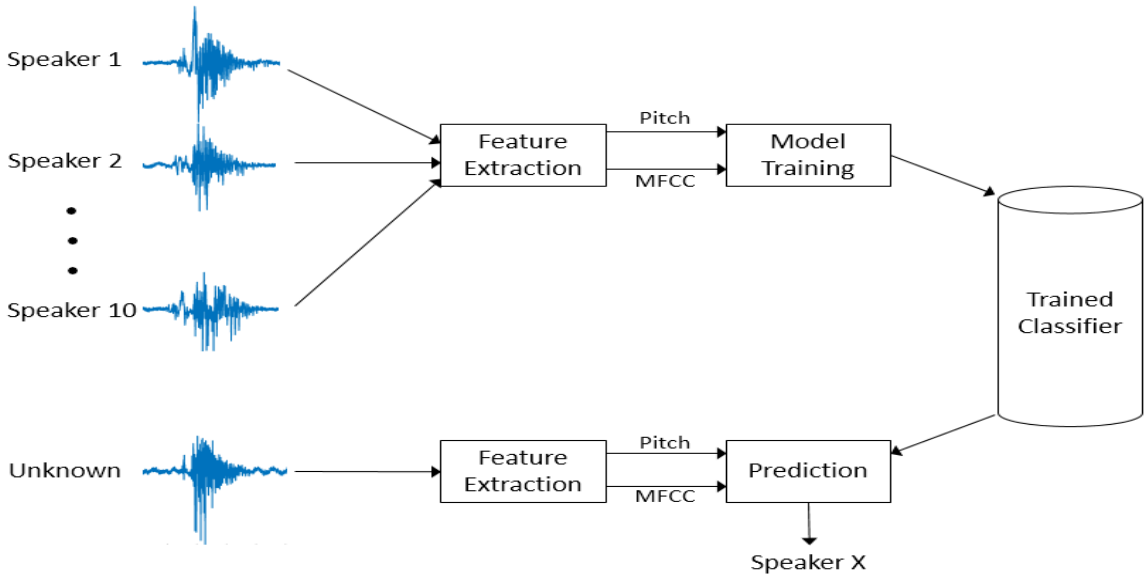


Figure.4.8: Model for features identification to speech recognition

4.2.3 Pitch of the voiced segments of the speech

Speech or the spoken word of the language are mainly classified as voiced and unvoiced. The speech modulated through vocal cards from the lungs is considered as voiced speech with low frequency range. It is called pitch of the speech. When the speech is modulated through a contraction in the vocal tract from the lungs is considered as unvoiced speech as excitation of noise. These excitations in the speech are known as speech filters generated by the vocal tract source. The speech filter and source plays great role in the identification of the speech signals. The voiced and unvoiced speech signals are considered in a time-domain representation. To identify these signals as per the spoken word, the consonant in the language *i.e.* unvoiced

speech considered as noise. The consonant in the language *i.e*. voiced speech considered as fundamental frequency.

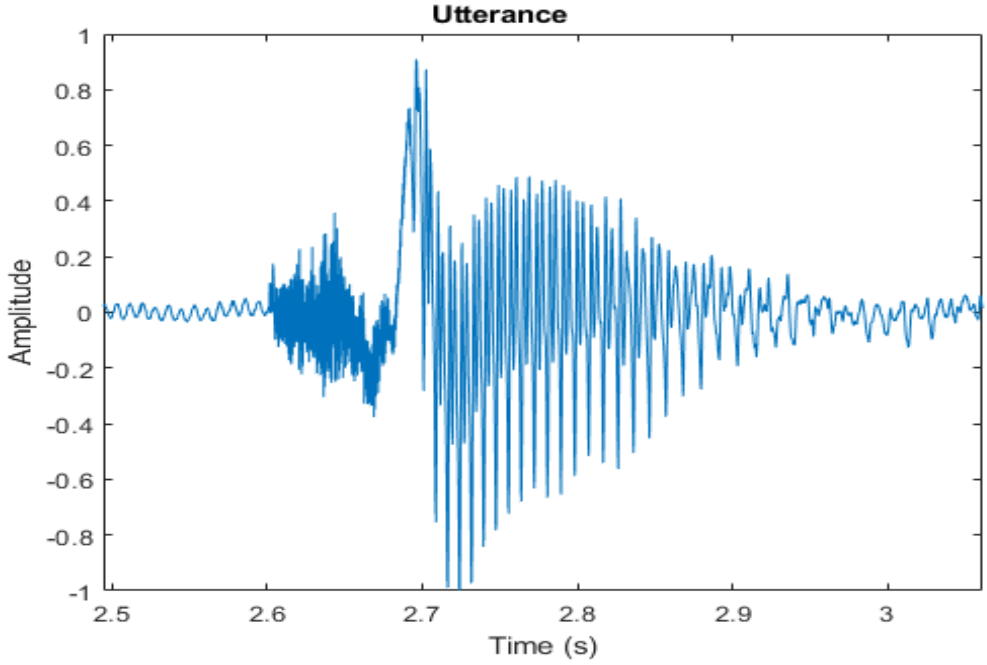


Figure.4.9: Time-domain representation of the word spoken to speaker Recognition.

The zero-crossing rate in the speech signals is used to differentiate the speech signals. The counting of the number of zero crossings in the speech signal indicate the low frequency components present in the speech signals. These frequency components are used to calculate the pitch in the speech signals. The utterance/pitch detection to the word spoken the pitch changes over time is shown in the Figure 4.9. This is the characteristic of a speaker in the form of pitch contour as identification.

4.2.4 Mel-Frequency Cepstrum Coefficients (MFCC)

In speech recognition, the features extracted from speech signals are Mel-Frequency Cepstrum Coefficients (MFCC). The filters in the speech signal based on the MFCCs are modelled as source filter model. During the analysis of the given signal, the response from the vocal tract is horizontally linear and in case of voice speech it is an impulse. The spectral in the speech signal is calculated for the vocal tract as an outcome of the analyzed signal. As per the response of the cochlea, the number of coefficients (MFCC) represent the information of the vocal tract needs to compressed *i.e*. to get the horizontally linear spectrum. Figure 4.10, represents the steps used for the calculation MFCCs.

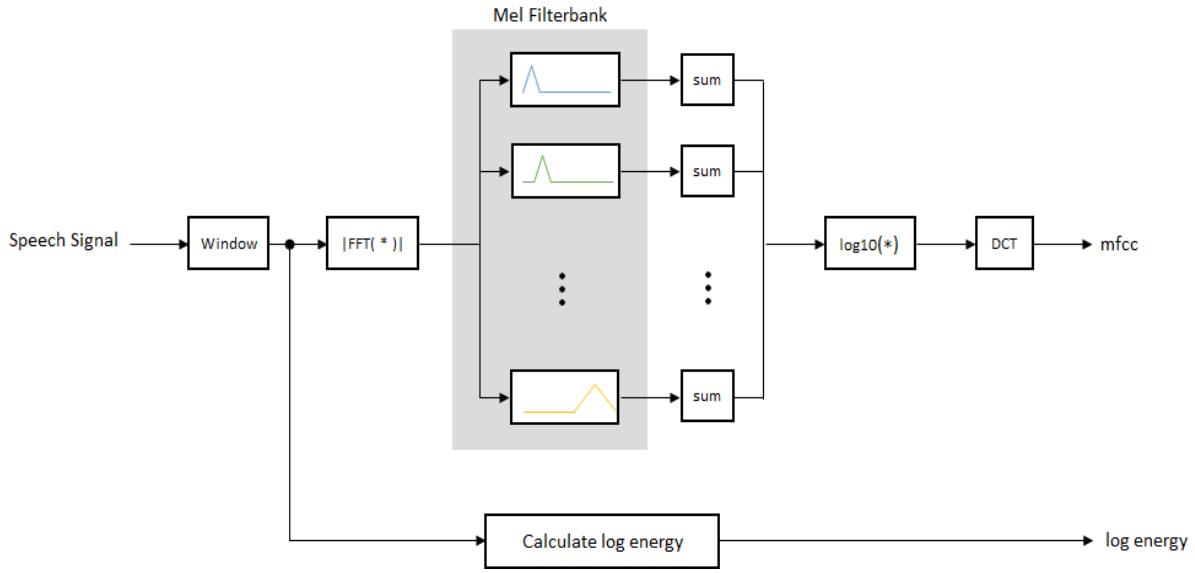


Figure.4.10: Mel-Frequency Cepstrum Coefficients (MFCC) of the word spoken to speech recognition.

The response of the human ear is also represented by the Mel-frequency cepstrum coefficients (MFCCs) and the derivation of its behavior is computed by the MFCC processor. Figure 4.11 depicted the steps to calculate the derivate of MFCCs.

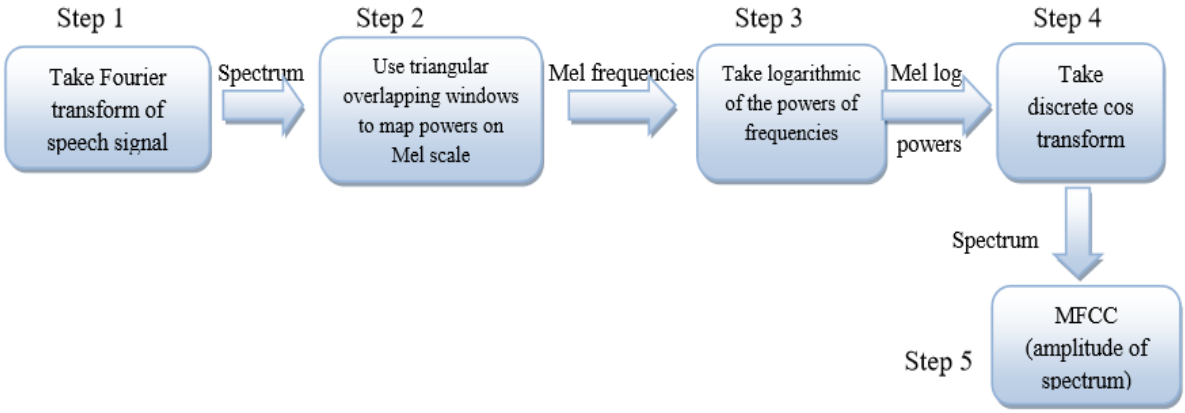


Figure.4.11: Steps to calculate the derivate of MFCCs

4.2.5 Computation of speech spectrograms

The high performance of the model is achieved by train the convolutional neural network with speech waveforms as input data and its conversion into Log-Mel spectrograms. The following components are used for calculation of spectrogram:

segment Duration - Total time of speech signal.

frame Duration - Total time of frame.

hop Duration - Time between columns of the spectrogram.

numBands - number of filters.

4.3. Layers of CNN.

- Convolutional Layer.
- Pooling Layer
- Fully Connected Layer

Convolutional Layer: Convolution are a sequence of sliding and projection operations. The net product of the two functions of displacement and projection is termed as sliding. Convolved features are calculated by the dot product of slide filters and given input data. The input data is a 2D data applied with conv2D model.

The method that has been used in our model is conv2D as the sound data is depicted as a two dimensional data. Before feeding the input tensor into the convolution layer the weights and biases values are set. The first convolutional layer of the first stack receives the input tensor and the weights. The bias and activation function are applied on the RELU layer of model.

Pooling Layer: is a way to scale down a vector and max poling is used this layer. The convolved vectors are calculated by mapping the conv filters with the input data. The feature mapping is performed by convolutional and detection by the pooling. The high performance in classification of CNN is achieved by using the both the conv and pooling layer in the network model. As input the pooling layer needs the output from the conv and RELU layer and the filter size and the stride size. The filter window is the sliding window with which the dot product is taken and the stride value is the amount by which the filter window slides. This layer completes the stack. It can be further passed on to the next layer or the fully connected layer.

Fully Connected Layer: This layer is the basic neural network layer having all the connections of input layer, hidden and output layers. The activation function is calculated in this layer. The output of this layer is a symbolic variable.

4.4. Mechanisms in CNN

Forward Propagation: The input data propagates from the input layer to the next layer in the forward direction by using forward propagation. The initial weights and bias are also given to each nodes in the convolutional layer.

Further the data is processed and the output is provided to the RELU unit. This unit normalizes the values present in the output vector by converting the negative values to zero. The output from this layer is further pushed to the max pooling layer. The output from this layer is passed

on to the next stack of convolutional, RELU, max- pool layer or it can also be passed directly to the fully connected layer. This layer gives out the final values associated with each class. The last unit is the Softmax layer which converts the final values into a range of 0 to 1 and the sum totaling to 1 giving final result as probabilities for each class.

Backpropagation: The error is calculated at the output layer by comparing desired out with the computed output from the network. The value of the error adjusts the values of the initial weights at the input layer. This value of the error controls the change the values of the weights. The change in the weight values is calculated according to the gradient descent. The computing cost for calculating the weight change is the gradient and it effects the training process. The decreasing the values of the gradient at the starting layers is main effects the training of the Deep Learning network. As the gradient values changes at large then network get trained at early and when its values are small the network get trained slowly. These are layers detect the simplest patterns.

Dropout: The neurons in the different layers are selected not to participate in the training process, is called a regularization. The contribution of these neurons is evaluated during forwarding of its outputs to neurons in the next layer and it ignored to updates the weights in the starting layers during backward pass. Now if during training these neurons are dropped randomly, so in order to cope up with this situation other neurons will have participated and achieve the desired targets. Which is required to achieve for the necessary predictions for the missing neurons and it needs to increase the value of the forward propagated neurons.
The adjusting the values of the dropped neurons in the network and these effect the training not get over fitted the training data and results in high accuracy. In order to perform dropout on a layer set some of the values randomly to zero. Drop out is only done during training. According to a paper by G. Hinton tuning of dropout should be done with respect to tuning the size of hidden layer. Until the data fits perfectly turn off the dropout and increase the hidden layer size. Then turn the dropout on and train with the hidden layer size as set in previous step. Finally, turnoff the dropout as soon as the training is over.

4.5. Optimizer

Adam Optimizer: In the training of deep learning models, Adam optimizer is used and it is another form of the for stochastic gradient descent applied in training.

4.6. Architecture of convolutional neural networks

The architecture comprises of convolutional, normalization layers, and max pooling layers. Figure.4.12. depicts the 2dConvolutional neural network architecture. Add a final max pooling layer that pools the input feature map globally over time. This enforces (approximate) time-translation invariance in the input spectrograms, allowing the network to perform the same classification independent of the exact position of the speech in time. Global pooling also significantly reduces the number of parameters in the final fully connected layer. To reduce the possibility of the network memorizing specific features of the training data, add a small amount of dropout to the input to the last fully connected layer. The architecture of 2dConvolutional with 2 layers has been shown in Figure.4.13.

The network is small, as it has only five convolutional layers with few filters. Figure.4.14. also depicts 2dConvolutional 5 layer with softmax layer. The function numF controls the number of filters in the convolutional layers. weighted cross entropy classification loss. weighted Classification Layer(class Weights) creates a custom classification layer that calculates the cross entropy loss with observations weighted by class weights. Specify the class weights in the same order as the classes appear in categories. To give each class equal total weight in the loss, use class weights that are inversely proportional to the number of training examples in each class.

4.7. Basic Structure:

Convolution Layer and Max Pool Layer Stack.

Convolution Layer 1 > Normalization>RELU > Max Pool Layer 1:

The first layer uses 12 filters with the size of 20 × 20

Convolution Layer 2> Normalization > RELU > Max Pool Layer 2:

The second layer uses 24 filters with the size of 24 × 23

Convolution Layer 3> Normalization > RELU > Max Pool Layer 3:

The third layer uses 48 filters with the size of 24 × 23

Convolution Layer 4> Normalization > RELU > Max Pool Layer 4:

The fourth layer uses 48 filters with the size of 24 × 23

Convolution Layer 5> Normalization > RELU > Max Pool Layer 4:

The fifth layer uses 48 filters with the size of 24 × 23

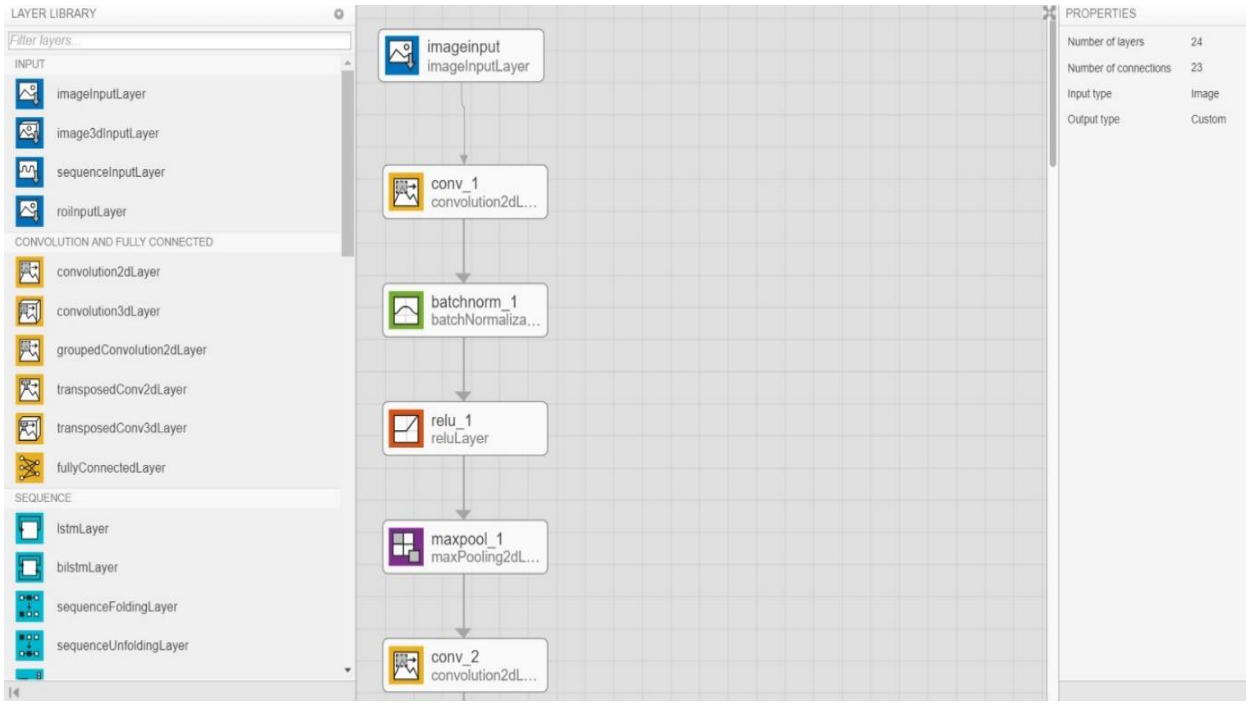


Figure.4.12: 2dConvolutional neural network architecture

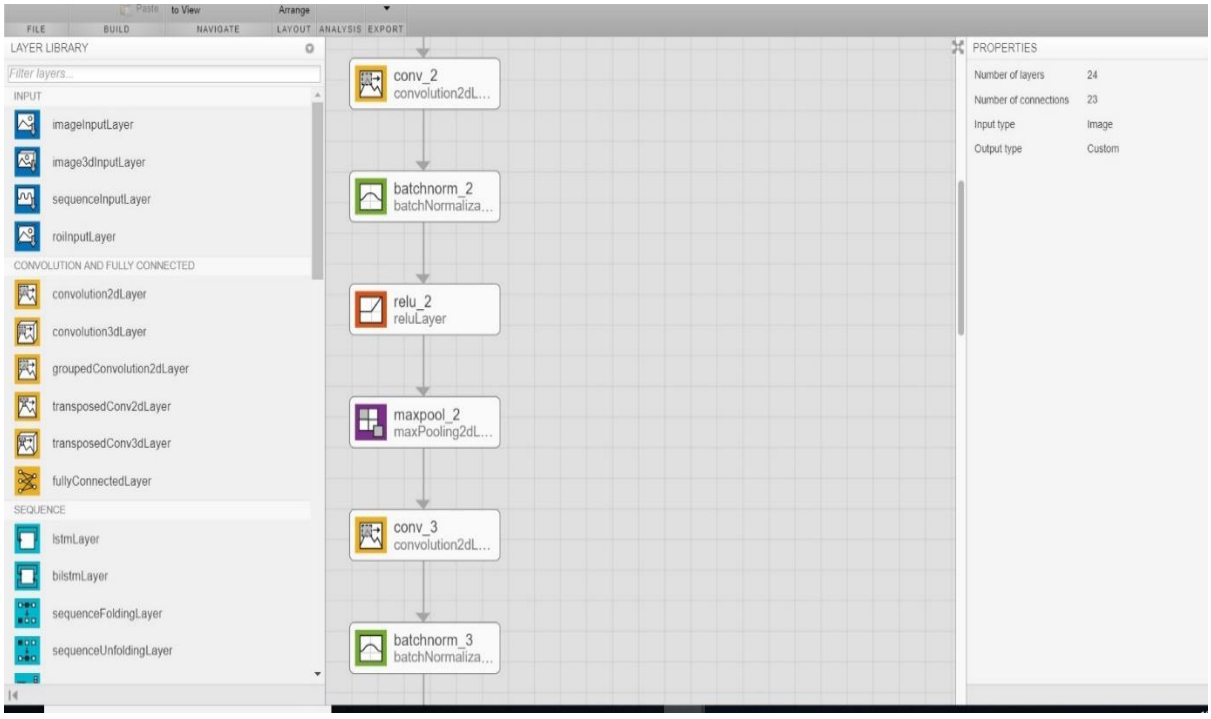


Figure.4.13: 2dConvolutional 2 layer

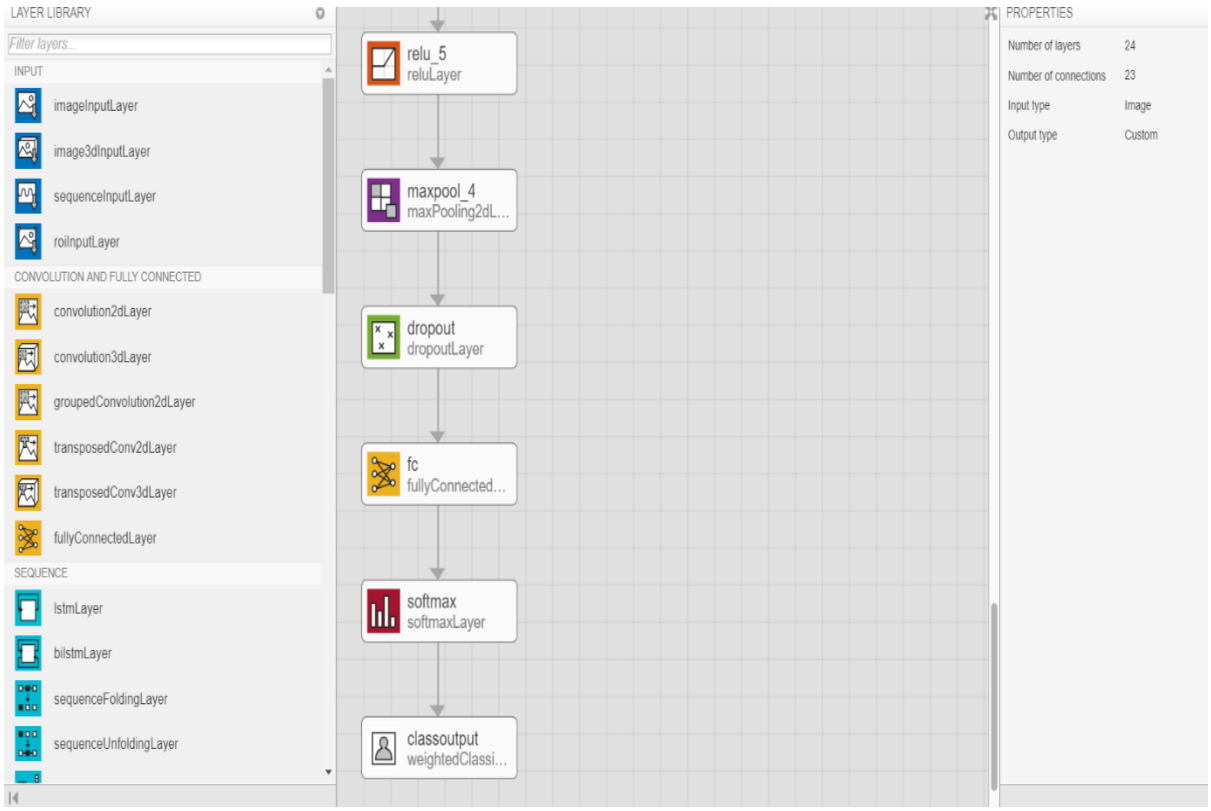


Figure.4.14: 2dConvolutional 5 layer with softmax layer

Fully Connected Layer Stack: The first layer has a dropout of 20%.

Output Layer: The output layer maps the output to the 10 classes (0 -10) and feeds the final output to the softmax Unit which converts the final output values to a range of 0 to 1 and sum total of 1 thus representing the outputs as the probabilities for the predicted class.

CHAPTER 5

Experimentation and Result Discussions

The experiment has been carried out with the design of the deep learning network architecture having normalization and convolutional layers. The input features spatially *i.e.* time and frequency has been added in the network by using max pooling layers. The network has five convolutional layers with filters. The fully connected layer is used to overcome the features dropout to the input. The invariance of time translation in the spectrum input is used to classify location of the speech in time. An addition blocks of the convolutional layers, normalization and activation function as ReLU layer are also used to train the network and calculate the accuracy of the network.

In this experiment, training and validation data sets are audio files collected by recording of Punjabi words consists of words, called as commands are" ਹਾਂ ", " ਨਾ ", " ਉੱਪਰ ", " ਥੱਲੇ ", " ਖੱਬੇ ", " ਸੱਜੇ ", " ਉਤੇ ", " ਬੰਦ", " ਠਹਿਰੋ" and some other words also. In order to obtain an equal balance of all words that are other than the command words are created the group/subset as unknown.

To train the convolutional neural network efficiently, training data sets (speech waveforms) are converted into Log-Mel spectrograms by calculating the duration in each speech clip, duration of each frame, number of Mel filters and the time between each column of spectrogram. The graphical representation of these parameters are also explained in the figure 5.1. The spectrograms are computed for training and validation of the convolutional network and smoothly distributed by using log of spectrogram with some offset value. The training spectrograms of some words are shown in the figure.

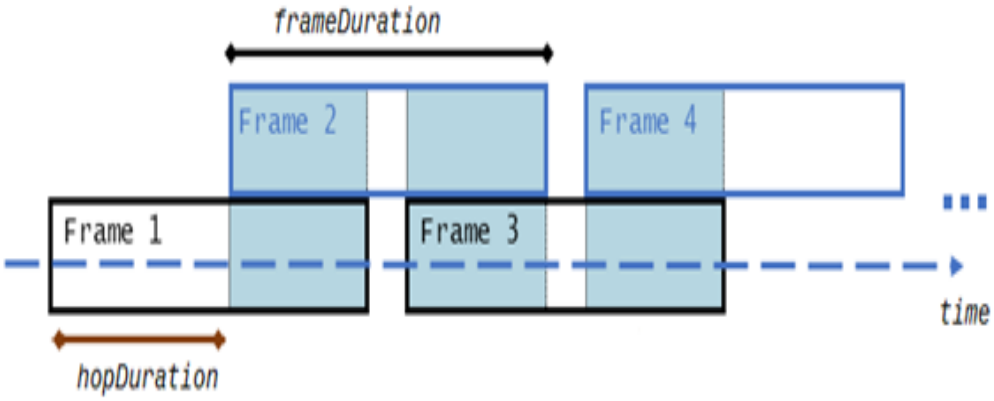


Figure.5.1: Parameters of the spectrogram calculation.

Visualizing the audio data

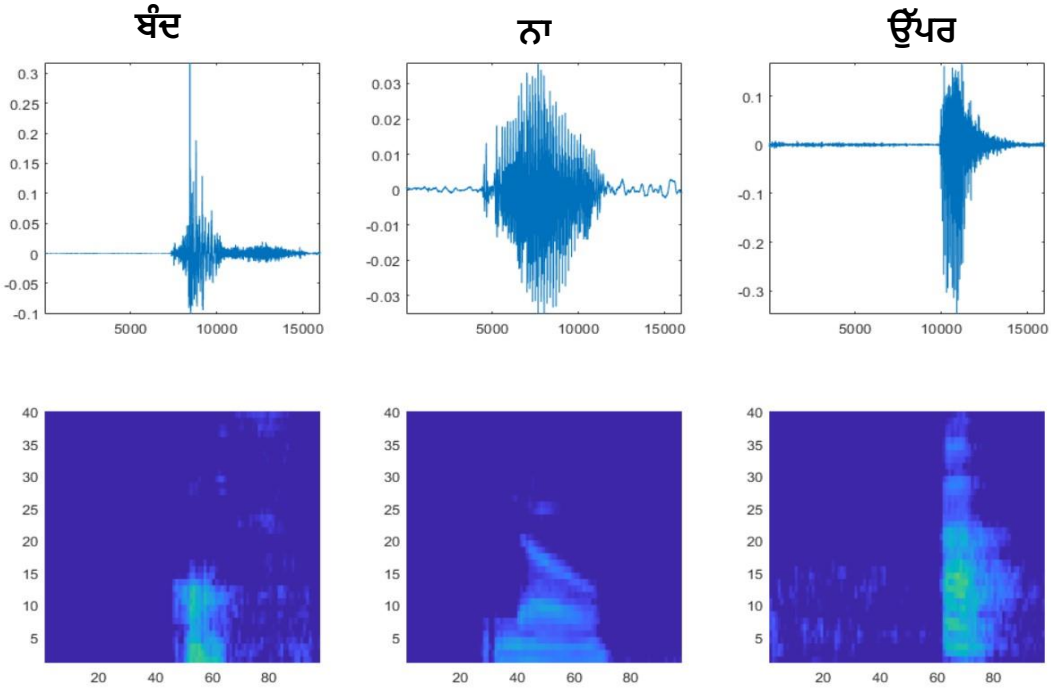


Figure 5.2: Spectrograms of the training data.

To check the smooth distribution of data, probability density has been calculated and the training data's pixel values are plotted as shown in the figure 5.3.

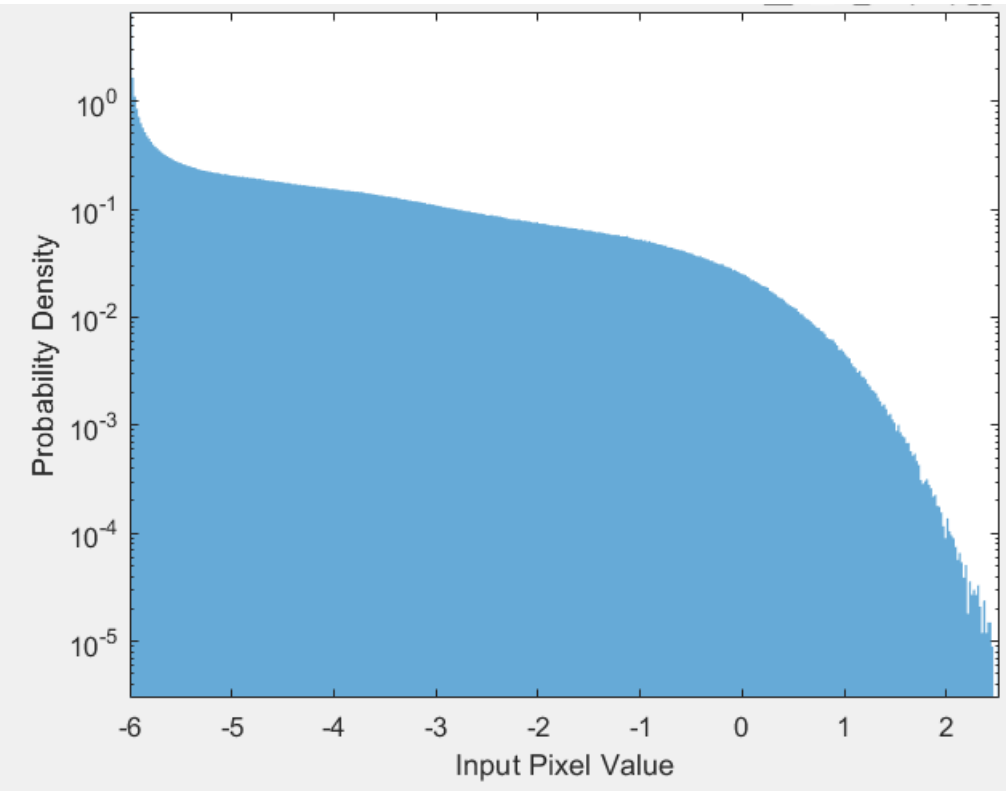


Figure 5.3: Probability density histogram of the training data.

The validation accuracy of the network has also been verified by addition of the background noise data in the training data sets. The network able to detect the difference between the spoken words and background noise data. The imbalance classes of the known (commands) and unknown words in the training data set is reduced by the adding the unknown word having probability. The distribution of the training and validation data sets has been plotted for different labelled classes as shown in the figure 5.4.

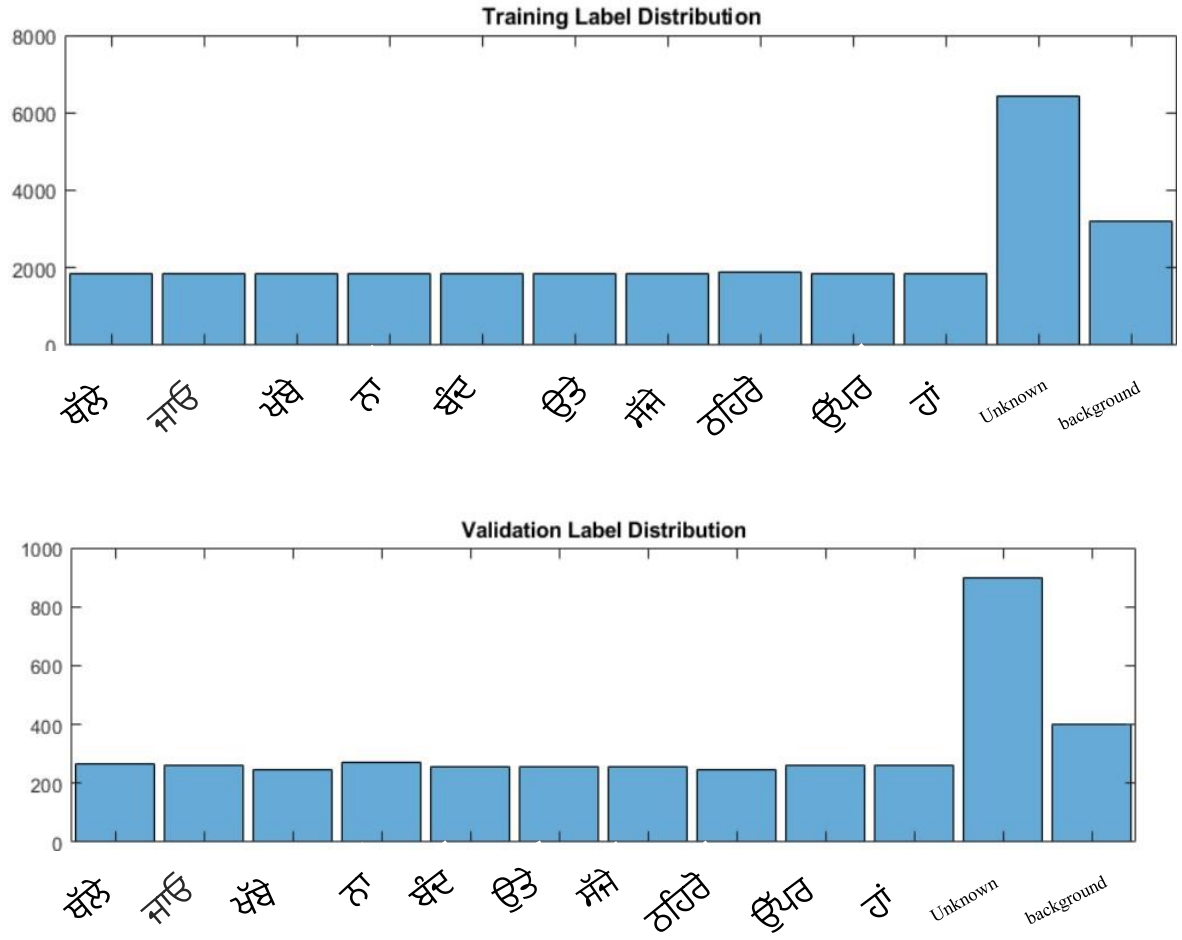


Figure 5.4: Label Distribution of the training and validation data.

The architecture of the proposed model of deep learning network is consists of the number of specified layers of convolutional neural network. The details of the network parameters such as activation function layers, sizes and learnable of the network are shown in the figure 5.5. The proposed network is trained with different training options to achieved the higher accuracy in recognition of the spoken word. The different data set distributed to the training dataset and validation dataset and different number frames of the speech signal.

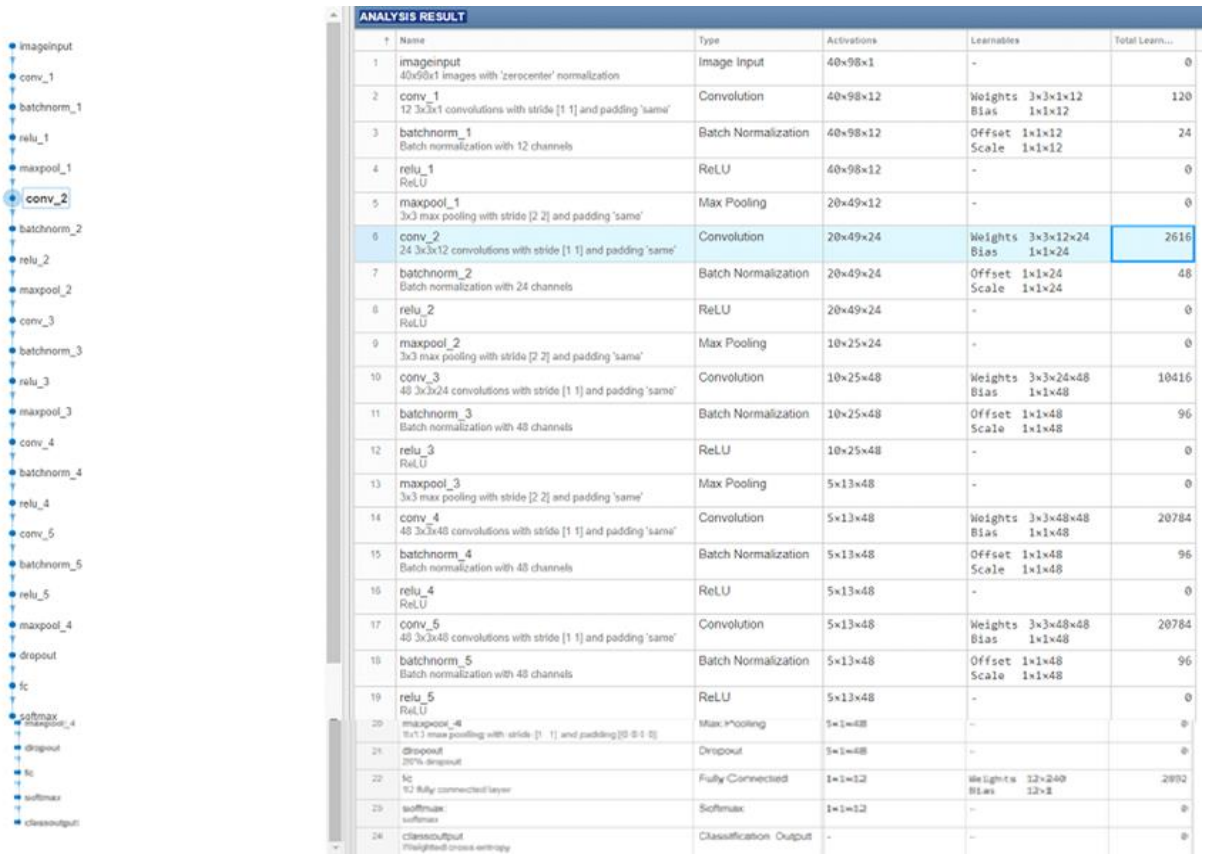


Figure 5.5: Architecture of the proposed Deep Learning Network Model

At each iteration, the training metrics are monitoring during the training progress. The training and validation accuracy and errors are monitoring as shown in the figure.5.6 and figure 5.7. predicts the weighted cross entropy loss between the number of classes and vectors of weights for each class.

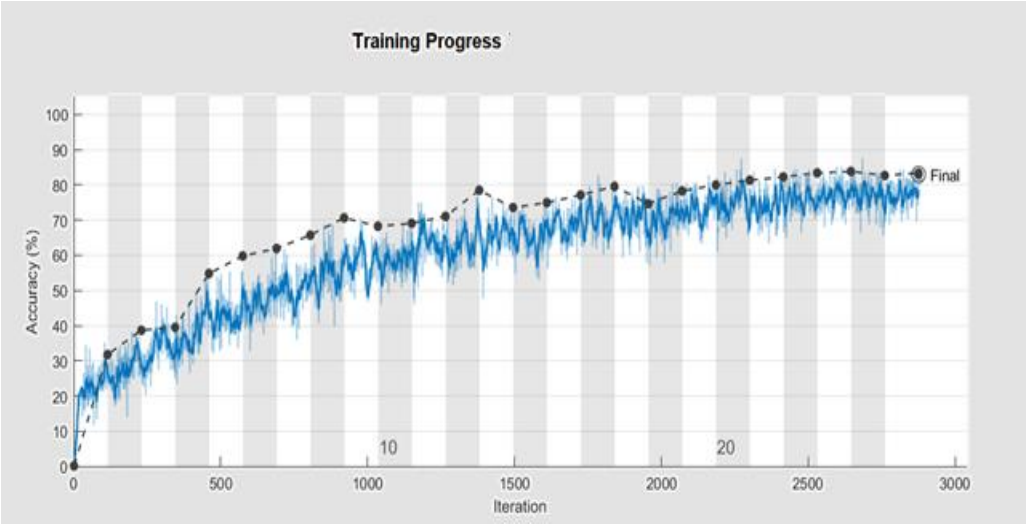


Figure 5.6: The validation accuracy during the training progress.

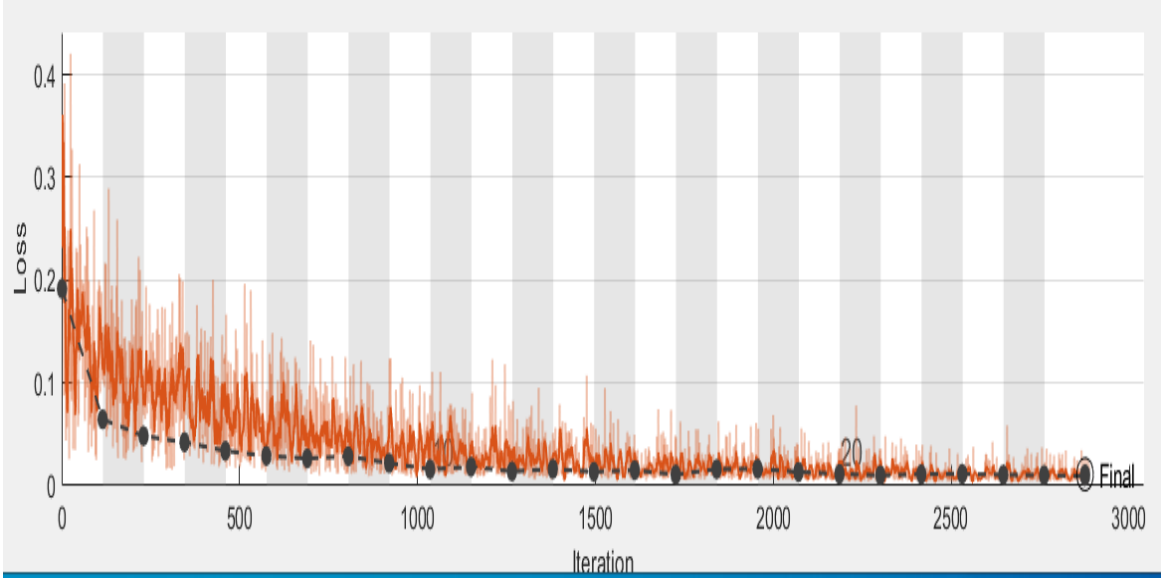


Figure 5.7: The cross entropy loss during the training progress.

The results of testing and validation accuracy for different division of data with noise reduction are given in the table 5.1 as drawn from the present study. Similarly, table 5.2 also have results of testing and validation accuracy for different division of data without noise reduction.

Table 5.1: Results of testing and validation accuracy without Noise Reduction

Training:Validation:Testing Samples	Training error (%)	Validation error (%)	Testing Accuracy (%)	Validation Accuracy (%)
80/10/10	13.126	12.648	79.89	87.36
70/15/15	15.130	17.374	84.67	82.63
60/20/20	15.810	16.131	80.78	83.87

Table 5.2: Results of testing and validation accuracy with Noise Reduction

Training:Validation:Testing Samples	Training error (%)	Validation error (%)	Testing Accuracy (%)	Validation Accuracy (%)
80/10/10	10.489	12.830	86.12	87.16
70/15/15	11.932	13.721	84.29	86.28
60/20/20	15.294	16.402	79.30	83.60

The validation accuracy has been achieved as 87.16% with only speech dataset in ratio of distribution of 80:10:10 in training dataset in comparison as 87.36% with noise data added in the training dataset. The overall accuracy for the true targets and predicted output of the training

and validation for the extended dataset are plotted as confusion matrix as given in the Figure 5.8.

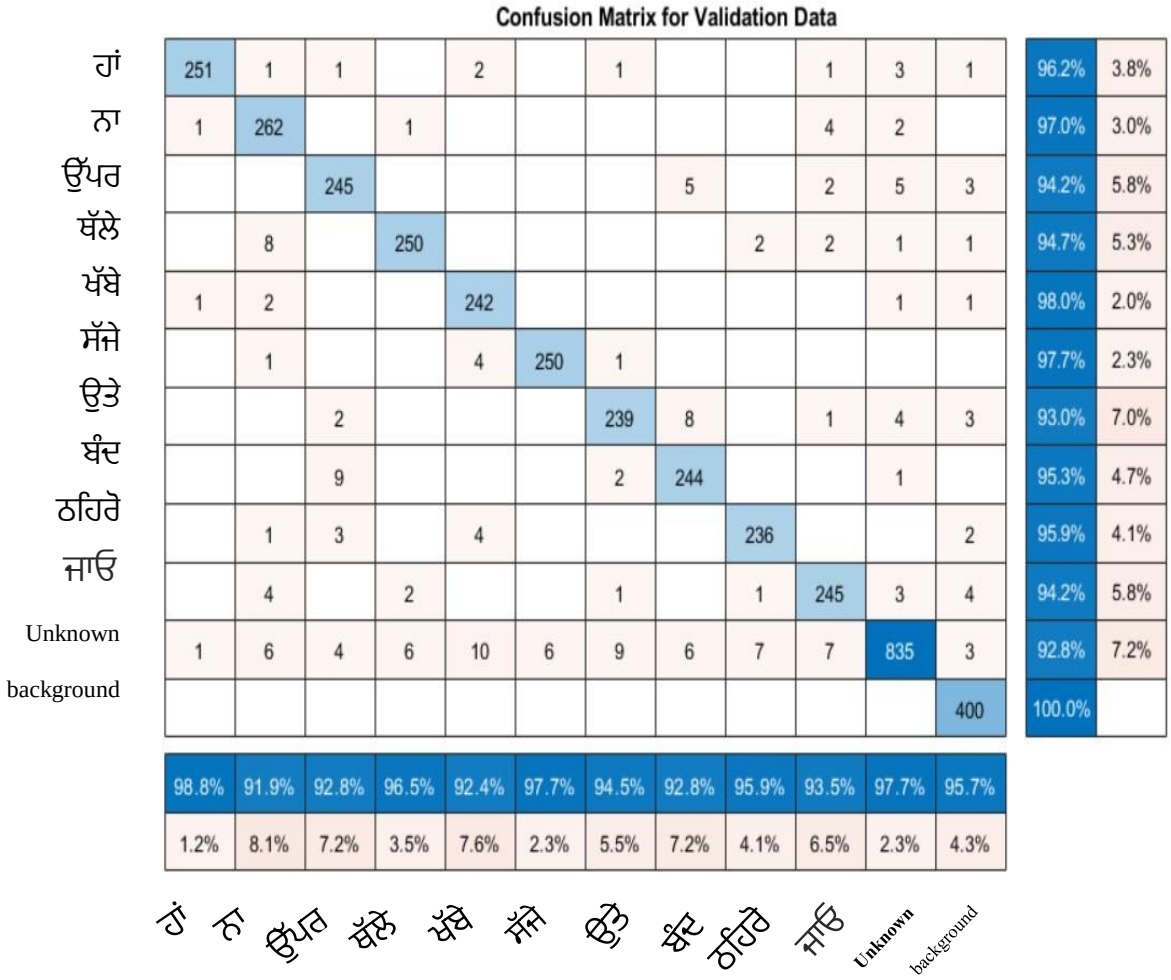


Figure 5.8. Confusion Matrix for the Validation Data.

CHAPTER 6

Conclusions and Future Scope

6.1 Conclusions

In chapter 5, it has been observed that the error rates in the training phase of the network increases as the numbers sample of the dataset included in the training dataset. The error rate of the validation phase also follows the same trend for the error calculations.

The error rate reduction is larger in case of the data set with noise reduction as compared to data set without noise reduction. It is noted that proposed model performs better with the datasets that is closely with real world.

The convergence time of the proposed deep learning model takes higher for the dataset having noise data included in the training data. When the dataset is not equally distributed with the known and unknown commands in the training data, the convergence time on the CPU architecture takes longer time to converge as compared to the convergence run on the GPU architecture.

The basic machine learning used for classification of the simple basic patterns are sufficient to classify the patterns. The complex structure of the patterns in the data, the number of nodes in each layer increases exponentially as the number of possible patterns in the data. The accuracy in the prediction starts suffering and training of the network becomes too expensively. The deep learning tools are single choice to deal such data sets.

From the experimentation it has been found that deep learning convolutional neural networks deals with spatial patterns in depth related to the speech recognition. The performance of this model achieved higher level of accuracy. In case of data sets or patterns are not in the image form then these type of the model does not give an accurate result.

6.2 Scope for Future Work

Improvement in Accuracy: To attain the maximum accuracy from the current accuracy is the main target to achieve by applying the combining the different deep learning models implemented on the higher throughput GPU architectures.

Larger Dataset: The larger data sets may be experimented on the present model by retuning the parameters to attain the higher accurate results.

REFERENCES

1 Nur Farhana Hordri, Siti Sophiayati Yuhaniz, Siti Mariyam Shamsuddin, “Deep Learning and Its Applications: A Review”, I*nt. conf. of Annual Research on Informatics 2016, Univ. Teknologi Malaysia, Kuala Lumpur*.

2 Danfeng Xie, Lei Zhang, and Li Bai, “Deep Learning in Visual Computing and Signal Processing”, *J. Applied Computational Intelligence and Soft Computing*, vol. 2017, pp. 13

3 Szu-Wei Fu , Tao-Wei Wang, Yu Tsao , Xugang Lu, and Hisashi Kawai, “End-to-End Waveform Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(9), 2018, pp 1570-1584

4 Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks”, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(1), 2015, pp 7 – 19.

5 De Liang Wang , *Fellow, IEEE*, and Jitong Chen, “Supervised Speech Separation Based on Deep Learning: An Overview”, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 26(10), 2018, pp 1702-1726..

6 Li Deng and Xiao Li, “Machine Learning Paradigms for Speech Recognition: An Overview” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 21(5), 2013, pp 1060-1089

7 A. E. Omer, "Joint MFCC-and-vector quantization-based text-independent speaker recognition system," *International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), Khartoum*, 2017, pp. 1-6.

8 J. Martinez, H. Perez, E. Escamilla and M. M. Suzuki, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, Cholula, Puebla, 2012, pp. 248-251.

9 Singh, Satyanand and E.G. Rajan “Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC.” *International Journal of Computer Applications, 17 (1), 2011, pp. 1-7.*

10 Anjali Jain, O.P. Sharma, “A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review,” *International Journal of Electronics & Communication Technology (IJECT)-Vol IV Issue SPL IV, Apr. TO June 2013, pp. 26-29*

11 Bharti, Roma & Bansal, Priyanka, “Real Time Speaker Recognition System using MFCC and Vector Quantization Technique.” *International Journal of Computer Applications , 117(1), 2015, pp. 25-31*.

12 Geeta Nijhawan, Dr. M.K Soni, “Speaker Recognition Using MFCC and Vector Quantisation”, *Int. J. on Recent Trends in Engineering and Technology, 11 (1), 2014, pp 32 -36.*

13 Majeed, S. A., Husain, H., Abdul Samad, S., & Idbeaa, T. F., “Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study”, *Journal of Theoretical and Applied Information Technology*, 79(1), 2015, pp. 38-56.

14 Vipul C. Rajyaguru, “Different Methods Used In Voice Recognition Techniques”, *International Research Journal of Engineering and Technology (IRJET), 3 (7), 2016, pp..*

15 Yoseph Linde, Member. IEEE. Andres Buzo, MEMBER, IEEE, A m Robert M. Gray, “An Algorithm for Vector Quantizer Design”, *IEEE Transactions ON Communications, VOL. COM-28, NO. 1, 1980.*

16 Anu L B , Dr Suresh D , Sanjeev kubakaddi, “Voice Identification using MFCC and Vector Quantization”, *IPASJ International Journal of Electronics & Communication (IIJEC), 3 (6), 2015, pp. 66-70*

17 Kishori R. Ghule, R. R. Deshmukh, “Feature Extraction Techniques for Speech Recognition: A Review”, *International Journal of Scientific & Engineering Research, 6 (5), 2015, pp. 42 -46*

18 Koustav Chakraborty, Asmita Talele, Prof. Savitha Upadhya, “ Voice Recognition Using MFCC Algorithm”, *International Journal of Innovative Research in Advanced Engineering (IJIRAE) , 1 (10), 2014, pp 5 -9*

19 Parwinder Pal Singh, Pushpa Rani, "An Approach to Extract Feature using MFCC", *IOSR Journal of Engineering (IOSRJEN), 4 (8), 2014, pp. 21-25.*

20 K Rao, S Nandy and S G Koolagudi, "Identification of Hindi dialects using Speech", *Advances in Intelligent Systems and Computing book series,* 2014, 264, pp. 161-169

21 Aggarwal, R.K., Dave, M, "Integration of multiple acoustic and language models improved Hindi speech recognition system" *Int. J. of Speech Technology* , 15, 2012, pp.165–180

22 https://www.semanticscholar.org/paper/Speech-Recognition-Using-Deep-Learning-Algorithms-Zhang/fbf62fad033af2c083bc3152066fd2cc4544da66

23 Shyam Agrawal, Shweta Sinha, Pooja Singh, and Jesper Olsen, "Development of Text and Speech Database for Hindi and Indian English Specific to Mobile Communication Environment. " *Proc. LREC 2012, Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey*, pp 3415-3421.

24 Francisco Javier Ordóñez and Daniel Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition", *Sensors*, (2016), (16), pp. 115.

25 Shweta Bansal, Shambhu Sharan and S.S. Agrawal, "Corpus design and development of an annotated speech database for Punjabi" *In Proc. IEEE International conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Shanghai, China. 2015.*

26 Naoya Takahashi, Michael Gygli, Beat Pfister and Luc Van Gool, "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition", *Proceedings of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016): Understanding Speech Processing in Humans and Machines, 8-12 September 2016, San Francisco, California, USA,* 2016

27 Mostafa Shahin, Julien Epps and Beena Ahmed, "Automatic Classification of Lexical Stress in English and Arabic Languages using Deep Learning", *Proceedings of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016): Understanding Speech Processing in Humans and Machines, San Francisco, California, USA.* 2016, PP 175-179.

28 G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, 2012, PP. 82-97

29 Yogesh Kumar and Navdeep Singh, "An automatic speech recognition system for spontaneous Punjabi speech corpus", *Int J Speech Techno*., 2017, 20, PP. 297–30

30 Kumar, Y., Singh, N, "An automatic spontaneous live speech recognition system for Punjabi language corpus", *IJCTA*, 2016, 9(20), PP. 259–266.

31 Waghmare K., Chaudari R. and Gawali B, "Accent Identification using MFCC for Hindi Langauge", *Advances In Computational Research*, 2015, 7, pp.170-172.

32 Arshdeep Singh, "Identification of Dialects in Punjabi language", *International Journal of Innovations & Advancement in Computer Science*, 2016, 5, pp. 34-38.

33 Irakli Kardava, Jemal Antidze, and Nana Gulua, "Solving the Problem of the Accents for Speech Recognition Systems", *International Journal of Signal Processing Systems*, 2016, 4(3), pp. 235-238.

34 S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal, Processing*, 1981, 29(2), pp. 254–272.

35 Justin Salamon, Christopher Jacoby and Juan Pablo Bello, "A Dataset and Taxonomy for Urban Sound Research" *Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA*, 2014, PP. 1041-1044.

36 Jianfeng , Zhaoa,b, Xia Maoa, Lijiang ChenJ, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks", *Biomedical Signal Processing and Control*, 2018, pp. 312–323.

37 Tiken Moirangthem, Partha Pratim Barman, Rupjyoti Gogoi, "Speech Recognition Model for Assamese Language using Deep Nerual Network" IEEE-International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering, 2018.

PUBLICATIONS

1. Pranav Kaushal and Maninder Singh, “Speech Command Recognition of Punjabi Language using Deep Learning Model”. To be Communicated.