**A Project Based Seminar Report**

**on**

**“Speech Recognition using Convolutional Neural Network”**

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

In

Information Technology

By

**Shivani Dandir**

(T150388694/ 315021 & A)

Under the guidance of

**Prof. Aditi Jahagirdar**

****

Department Of Information Technology

Maharashtra Institute of Technology College of Engineering

Kothrud, Pune, Maharashtra, India

**2017-2018**

****

**CERTIFICATE**

This is to certify that the project based seminar report entitled **“Speech Recognition using**

**convolutional neural networks”** being submitted by **Shivani Dandir (T150388694 / 3154021 &**

1. is a record of bonafide work carried out by him/her under the supervision and guidance of **Prof.** **Aditi Jahagirdar** in partial fulfillment of the requirement for **TE (Information Technology Engineering) – 2015 course** of Savitribai Phule Pune University, Pune in the academic year 2017-2018

Date:

Place: MITCOE, Pune

Prof. Aditi Jahagirdar

Guide

Dr. Anil S Hilwale

Head of the Department



This project based seminar report has been examined by us as per the Savirtibai Phule Pune University, Pune, requirements at MIT College of Engineering on . . . . . . . . . . . . . .

(Name & Signature)

Internal Examiner

(Name & Signature)

External Examiner

I

**ACKNOWLEDGEMENT**

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them I respect and thank Dr. Anil.S.Hiwale, HOD-IT Department for providing me an opportunity to do the project work in [IT DEPARTMENT] and giving us all support and guidance which made me complete the project duly. I owe my deep gratitude to our project guide Prof. Aditi Jahagirdar, who took keen interest on our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system. I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of [Information Technology Department] which helped us in successfully completing our project work. Also, I would like to extend our sincere esteems to all staff in laboratory for their timely support.

Shivani Dandir

II

**Abstract**

Nowadays, many speech recognition system uses HMM to deal with variability in speech, this model can be outrun by CNN.CNN uses feed forward neural network and error rate reduction can be obtained. Deep neural networks (DNNs) that have many hidden layers and are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks. In this seminar, we present a description of CNN and how it is used for speech recognition. Further, the design of CNN which includes pooling, connectivity and weight sharing is described. Experimental results show that the proposed CNN method can achieve over 10% relative error reduction in the core TIMIT test sets when comparing with HMM using the same number of hidden layers and weights.

III

**Contents**

Certificate

Acknowledgement

Abstract

List of Figures

I

I

III

IV

**Sr.**

**Chapter**

**Page No**

1. Introduction  **7**

1.1 Introduction to Project

1.2 Motivation behind Project Topic

1.3 Objective of the work

1.4 Introduction to Speech Recognition using CNN

1.4.1 Aim and Objectives of Seminar

1. Literature Survey of Speech Recognition using CNN 10
2. CNN and their use in ASR13

3.1 Convolutional Ply

3.2 Pooling ply

3.3 Weights in CNN

3.4 Treatment of energy Resource

3.5 CNN Architecture

3.6 Benefits of CNNs

1. Applications of CNN 18

4.1 Speech Data and Analysis

4.2 Drug Discovery

1. Conclusion 19
2. References 20

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Figure Name** | **Page No.** |
|  |  |  |
| 1 | CNN Layer | 14 |
|  |  |  |
| 2 | CNN –full weight sharing | 17 |
|  |  |  |



**CHAPTER 1**

**INTRODUCTION TO “Depression Detection using Speech Feature Extraction”**

1.1 Introduction to Project

Depression is a common but a serious mood disorder that can affect people of genders, all ages and any background. It is a common mental disorder that mostly presents with depressed mood, loss of interest, feelings of guilt or low self-worth, disturbed sleep or appetite, low energy and poor concentration and could result in unbearable pain if proper treatment not given. It usually occurs as result of adverse life events, but it can also occur due to no apparent cause as well. Major lifestyle changes can also lead to depression or feelings of dissatisfaction or dullness. Depression is much more common in women than men, with female/male risk ratios roughly 2:1. Depression classification is one of the most challenging tasks in a speech signal processing domain. There is no laboratory test for detecting and procedures for diagnosis of depression. Rather it has been diagnosed as a part of complete mental health evaluation. Symptoms of depression often first appear during adolescence at the time when the voice is changing in both males and females. The severity of symptoms can vary, but the good news is that with the right treatment and support, people can often make a full recovery. The proposed methodology offers an objective measure of assessment of risk for depression. It is also a low cost and relatively easy to implement approach. The achieved sensitivity, specificity and accuracy values make it suitable for mass-screening tests of large populations of adolescents.

1.2 Motivation behind project topic

The study is aimed to reduce depression level. Since it is affecting adolescents (I.e. Those age is 13-20 years) enlarge amount of number, it results in the range of serious outcomes like increase in suicide attempts and deaths. Therefore, the main motive of this project topic is early identification of depression in adolescents which is important and which could significantly reduce the burden of the disease. From psychological point of view the signs of person being depressed are the way emotions are expressed in his/her speech. This is based on assumption that the emotional state of a person suffering from some depressive disorders affects the acoustic qualities of their speech and therefore depression could be detected through the changes in the acoustical properties of the speech.

An acoustic analysis of speech will provide clinicians with an additional quantitative measure to compliment and strengthen the current diagnostic techniques.

1.3 Aim and Objective(s) of the work

The principle aim of this group project is to design and develop a standardized application for Detection of level of Depression by Extraction of Speech Features. The application must be capable of detecting the depression level and thus provide the necessary remedies to the user if he/she is suffering from Depression. Level of depression is classified according to the values of the attributes that are extracted using the proposed methodology.

A broader aim of the project takes into account the approach of social welfare as this would provide a standardized method for Depression Detection which is considered as a fatal problem if not treated well in time. The concept of this project has been worked on for a number of years, and it has proved a considerable task to develop a fully functioning Depression Detection application. Thus, in the interest of future projects, the work carried out during this project will be guided towards producing a system that can readily be further developed by subsequent project groups. To make this a reality the project will be documented as closely and as accurately as possible, thus providing subsequent groups with all the knowledge required continuing the development of this system.

**Project objectives-**

* To recognize the speech features.
* To extract the speech features.
* To learn various speech features and their effects.
* To learn various speech recognition models.
* To study factors of speech that determines depression.
* To enhance speech recognition and extraction by using new models.
* To evaluate a person’s depression level.

1.4 Introduction to Speech Recognition using CNN

Automatic speech recognition (ASR) can be defined as the independent, computer‐driven transcription of spoken language into readable text in real time (Stuckless, 1994). In a nutshell, ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text.

In a hidden Markov model, there are "hidden" states, or unobserved, in contrast to a standard

Markov chain where all states are visible to the observer. Hidden Markov models are used for machine learning and data mining tasks including speech, handwriting and gesture recognition.HMM is successful in handling variable length sequences as well as modeling the temporal behavior of speech signals using a sequence of states.

Very recently, HMM models that use artificial neural networks (ANNs) instead of GMMs have witnessed a significant resurgence of research interest. In machine learning, a convolutional neural network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural network. CNNs have been applied to acoustic modeling before, in which convolution was applied over windows of acoustic frames that overlap in time in order to learn more stable acoustic features.

.

1.4.1 Aim and Objectives of Seminar

The aim of this Seminar is to understand the speech recognition with better working models (CNN) that will help in error reduction as compared with the traditional HMM model. The main objective is to increase accuracy level of the input speech by using neural networks.

**CHAPTER 2**

**LITERATURE SURVEY OF Speech recognition using CNN**

Geoffrey Hinton in their paper, Deep Neural Networks for Acoustic Modeling in Speech Recognition[1]: The Shared Views of Four Research Groups, proposed that current speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. An alternative way to evaluate the fit is to use a feed-forward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. Deep neural networks (DNNs) that have many hidden layers and are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin.

Alex Graves et. al. talk about Speech recognition with deep recurrent neural networks[3], Recurrent neural networks (RNNs) are a powerful model for sequential data. End-to-end training methods such as Connectionist Temporal Classification make it possible to train RNNs for sequence labeling problems where the input-output alignment is unknown. The combination of these methods with the Long Short-term Memory RNN architecture has proved particularly fruitful, delivering state-of-the-art results in cursive handwriting recognition. However RNN performance in speech recognition has so far been disappointing; with better results returned by deep feed forward networks. Their paper investigates deep recurrent neural networks, which combine the multiple levels of representation that have proved so effective in deep networks with the flexible use of long range context that empowers RNNs. When trained end-to-end with suitable regularization, they find that deep Long Short-term Memory RNNs achieve a test set error of 17.7% on the TIMIT phoneme recognition benchmark.

George E. Dahl et.al. propose in Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition[2] that, context-dependent (CD) model for large-vocabulary speech recognition (LVSR) that leverages recent advances

in using deep belief networks for phone recognition. They describe a pre-trained deep neural network hidden Markov model (DNN-HMM) hybrid architecture that trains the DNN to produce a distribution over senones (tied triphone states) as its output. The deep belief network pre-training algorithm is a robust and often helpful way to initialize deep neural networks generatively that can aid in optimization and reduce generalization error. They illustrate the key components of our model, describe the procedure for applying CD-DNN-HMMs to LVSR, and analyze the effects of various modeling choices on performance. Experiments on a challenging business search dataset demonstrate that CD-DNN-HMMs can significantly outperform the conventional context-dependent Gaussian mixture model (GMM)-HMMs, with an absolute sentence accuracy improvement of 5.8% and 9.2% (or relative error reduction of 16.0% and 23.2%) over the CD-GMM-HMMs trained using the minimum phone error rate (MPE) and maximum-likelihood (ML) criteria, respectively.

Eric G. Nestler et. al. mentioned in Convolutional neural network[4] ,that Systems and methods of implementing a more efficient and less resource-intensive CNN are disclosed herein. In particular, applications of CNN in the analog domain using Sampled Analog Technology (SAT) methods are disclosed. Using a CNN design with SAT results in lower power usage and faster operation as compared to a CNN design with digital logic and memory. The lower power usage of a CNN design with SAT can allow for sensor devices that also detect features at very low power for isolated operation.

Mihai Constantine Munteanu et. al. in their paper, A convolutional neural network (CNN) [5] for an image processing system comprises an image cache responsive to a request to read a block of N×M pixels extending from a specified location within an input map to provide a block of N×M pixels at an output port. A convolution engine reads blocks of pixels from the output port, combines blocks of pixels with a corresponding set of weights to provide a product, and subjects the product to an activation function to provide an output pixel value. The image cache comprises a plurality of interleaved memories capable of simultaneously providing the N×M pixels at the output port in a single clock cycle. A controller provides a set of weights to the convolution engine before processing an input map, causes the convolution engine to scan across the input map by incrementing a specified location for successive blocks of pixels and generates an output map within the image cache by writing output pixel values to successive locations within the image cache.

Michael Price described in, Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks[6], digital circuit architectures for automatic speech recognition (ASR) and voice activity detection (VAD) with improved accuracy, programmability, and scalability. Our ASR architecture is designed to minimize off-chip memory bandwidth, which is the main driver of system power consumption. A SIMD processor with 32 parallel execution units efficiently evaluates feed-forward deep neural networks (NNs) for ASR, limiting memory usage with a sparse quantized weight matrix format. They argue that VADs should prioritize accuracy over area and power, and introduce a VAD circuit that uses an NN to classify modulation frequency features with 22.3-µW power consumption. The 65-nm test chip is shown to perform a variety of ASR tasks in real time, with vocabularies ranging from 11 words to 145 000 words and full-chip power consumption ranging from 172 µW to 7.78 mW.

**CHAPTER 3**

**CNN and their use in ASR**

The convolutional neural network (CNN) can be regarded as a variant of the standard neural network. Instead of using fully connected hidden layers as described in the preceding section, the CNN introduces a special network structure, which consists of alternating so-called convolution and pooling layers.[8]

3.1. Convolution Ply

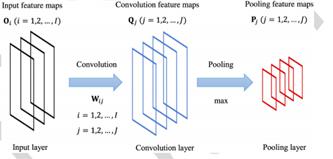


Fig1. An illustration of one CNN “layer” consisting of a pair of a convolution ply and a pooling ply in succession, where mapping from either the input layer or a pooling ply to a convolution ply is based on eq. (2) and mapping from a convolution ply to a pooling ply is based on eq. (3).[7]

Every input feature map (assume I is the total number), Oi(i=1,….,I) , is connected to

many feature maps (assume J in the total number),Qj(j=q,…,J) , in the convolution ply based on a number of local weight matrices ( I\*J in total),wi,j(i=1,…,I;j=1,…J) . The mapping can be represented as the well-known convolution operation in signal processing. Assuming input feature maps are all one dimensional, each unit of one feature map in the convolution ply can be computed as:



……(1)[7]

where oi,m is the m-th unit of the i-th input feature map Oi, qj,m is the m-th unit of the j-th feature map Qj in the convolution ply, wi,j,n is the nth element of the weight vector, wi,j, which connects the ith input feature map to the jth feature map of the convolution ply. F is called the filter size, which determines the number of frequency bands in each input feature map that each unit in the convolution ply receives as input. Because of the locality that arises from our choice of MFSC features, these feature maps are confined to a limited frequency range of the speech signal.

**.....(2)[7]**

where Oi represents the i-th input feature map and wi,j represents each local weight matrix, flipped to adhere to the convolution operation's definition. Both Oi and wi,j are vectors if one dimensional feature maps are used, and are matrices if two dimensional feature maps are used .

A convolution ply differs from a standard, fully connected hidden layer in two important aspects, however. First, each convolutional unit receives input only from a local area of the input. This means that each unit represents some features of a local region of the input. Second, the units of the convolution ply can themselves be organized into a number of feature maps, where all units in the same feature map share the same weights but receive input from different locations of the lower layer.

**3.2 Pooling Ply**

A pooling operation is applied to the convolution ply to generate its corresponding pooling ply[8]. The pooling ply is also organized into feature maps, and it has the same number of feature maps as the number of feature maps in its convolution ply, but each map is smaller. The purpose of the pooling ply is to reduce the resolution of feature maps. The pooling function is applied to each convolution feature map independently. When the max-pooling function is used, the pooling ply is defined as:

…….(3)[7]

Where, G is the pooling size, and s, the shift size, determines the overlap of adjacent pooling windows. Similarly, if the average function is used, the output is calculated as:



……(4)[7]

Where, r is a scaling factor that can be learned.

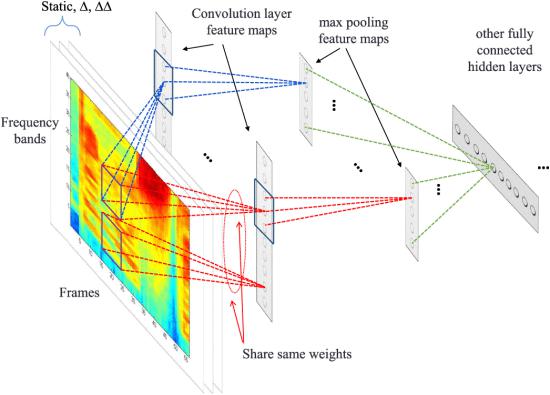


Fig 2: An illustration of the regular CNN that uses so-called full weight sharing. Here, a 1-D convolution is applied along frequency bands.[7]

**3.3Weights in the CNN**

The treatment of shared weights in the convolution ply is slightly different from the fully-connected DNN case where there is no weight sharing. The difference is that for the shared weights here, we sum them in their updates according to:



Where, I and J are the number of feature maps in the input layer and convolution ply, respectively.[7]

Since the pooling ply has no weights, no learning is needed here. However, the error signals should be back-propagated to lower plies through the pooling function.

**3.4. Treatment of Energy Features**

In ASR, log-energy is usually calculated per frame and appended to other spectral features. In a CNN, it is not suitable to treat energy the same way as other filter bank energies since it is the sum of the energy in all frequency bands and so does not depend on frequency. Instead, the log-energy features should be appended as extra inputs to all convolution units.

**3.5. CNN Architecture**

The building block of the CNN contains a pair of hidden plies: a convolution ply and a pooling ply. The input contains a number of localized features organized as a number of feature maps. The size (resolution) of feature maps gets smaller at upper layers as more convolution and pooling operations are applied.

**3.6. Benefits of CNNs**

The CNN has three key properties: locality, weight sharing, and pooling. Each one of them has the potential to improve speech recognition performance. Locality in the units of the convolution ply allows more robustness against non-white noise where some bands are cleaner than the others. This is because good features can be computed locally from cleaner parts of the spectrum and only a smaller number of features are affected by the noise. This gives a better chance to higher layers of network to handle this noise because they can combine higher level features computed for each frequency band. This is clearly better than simply handling all input features in the lower layers as in standard, fully connected neural networks. Moreover, locality reduces the number of network weights to be learned.

**Chapter 4**

**Application of CNN**

**4.1. Speech Data and Analysis**

The method of speech analysis is similar in the two datasets. Speech is analyzed using a 25-ms Hamming window with a fixed 10-ms frame rate. Speech feature vectors are generated by Fourier-transform-based filter-bank analysis, which includes 40 log energy coefficients distributed on a mel scale, along with their first and second temporal derivatives. All speech data were normalized so that each vector dimension has a zero mean and unit variance.

**4.2. Drug discovery**

CNNs have been used in [drug discovery.](https://en.wikipedia.org/wiki/Drug_discovery) Predicting the interaction between molecules and biological [proteins](https://en.wikipedia.org/wiki/Proteins) can identify potential treatments. In 2015, Atomwise introduced Atom Net, the first deep learning neural network for structure-based [rational drug](https://en.wikipedia.org/wiki/Drug_design) [design.](https://en.wikipedia.org/wiki/Drug_design) The system trains directly on 3-dimensional representations of chemical interactions. Similar to how image recognition networks learn to compose smaller, spatially proximate features into larger, complex structures, AtomNet discovers chemical features, such as [aromaticity,](https://en.wikipedia.org/wiki/Aromaticity) [sp3 carbons](https://en.wikipedia.org/wiki/Orbital_hybridisation) and [hydrogen bonding.](https://en.wikipedia.org/wiki/Hydrogen_bond) Subsequently, AtomNet was used to predict novel candidate [biomolecules](https://en.wikipedia.org/wiki/Biomolecule) for multiple disease targets, most notably treatments for the [Ebola virus](https://en.wikipedia.org/wiki/Ebola_virus) and [multiple sclerosis.](https://en.wikipedia.org/wiki/Multiple_sclerosis)

**CHAPTER 5**

**CONCLUSION**

In this work, we saw various speech recognition models that are being used recently. We studied CNNs to recognize speech in a novel way, such that the CNN’s structure directly accommodates some types of speech variability. We saw comparison of performance between DNN and CNN with similar numbers of weight parameters using CNN we got about 6-10% relative error reduction. The hybrid CNN-HMM approach delegates temporal variability to the HMM, while convolving along the frequency axis creates a degree of invariance to small frequency shifts, which normally occur in actual speech signals due to speaker differences.

Moreover, we studied a new limited weight sharing scheme that can handle speech features in a better way than the full weight sharing that is standard in previous CNN architectures such as those used in image processing. Limited weight sharing leads to a much smaller number of units in the pooling ply, these results in a smaller model size and lower computational complexity than the full weight sharing scheme. And finally, we saw applications of CNN in various fields. The future work of this study will be an application which uses CNN model to recognize speech and its features will be extracted to detect level of depression.

**REFERENCES**

1. Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine 29, no. 6 (2012): 82-97.
2. Dahl, George E., Dong Yu, Li Deng, and Alex Acero. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." IEEE Transactions on audio, speech, and language processing 20, no. 1 (2012): 30-42.
3. Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pp. 6645-6649. IEEE, 2013.
4. Nestler, Eric G., Mitra M. Osqui, and Jeffrey G. Bernstein. "Convolutional neural network." U.S. Patent Application 15/379,114, filed June 15, 2017.
5. Munteanu, Mihai Constantine, Alexandru Caliman, and Corneliu Zaharia. "Convolutional neural network." U.S. Patent 9,665,799, issued May 30, 2017.
6. Price, Michael, James Glass, and Anantha P. Chandrakasan. "A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks." IEEE Journal of Solid-State Circuits (2018).
7. Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. "Convolutional neural networks for speech recognition." IEEE/ACM Transactions on audio, speech, and language processing 22, no. 10 (2014): 1533-1545.
8. https://en.wikipedia.org/wiki/Convolutional\_neural\_network