International Journal of Control and Automation
Vol. 12, No. 4, (2019), pp. 11-19

Accelerating Speech Recognition System by Adam optimization and CNN for Real Time System using GPU

Rajkumar S. Bhosale[1], Narendra S. Chaudhari[2]
[1]*Amrutvahini College of Engineering, Sangamner; phone: 7588546775; bhos_raj@rediffmail.com*
[2]*Indian Institute of Technology, Indore; Department of Computer Science; nsc@iiti.ac.in*

Abstract

In today's digital world of computer and Mobile system, Automatic Speech recognition in real time is promising application. The fast responsive speech commands gives flexibility to the user for rapid access of smart systems, Laptop and other devices efficiently instead of typing commands. Along with great state of accuracy and learning by example, deep learning makes machine learning classification at highest peak. In deep learning, Convolutional Neural Network (CNN) class gives promising results in speech recognition. In previous systems bind speech recognition to only two convolutional layers. The proposed model works on five layers of convolutional neural network. The proposed deep neural network classification system applied on 65000 WAVE Google's Tens or flow dataset and AIY commands. Mel Spectrogram extract from the input speech and Adam optimization algorithm perform training of convolutional neural network (CNN). The Convolutional Neural Network proves to outperform than other models and can achieve accuracy of 95.1% for 6 labels. For better performance of system, we added Background Noise in data. If noise is added, the network not only recognizes different spoken words but also detects input contains any silence or background noise. The Data augmentation support for augmenting the data can increase the effective size of the training data and help prevent the network from over fitting. Training essentially consider very crucial so CPU or GPU (NVDIA Tesla K40 C GPU) is used for training purpose for time efficiency. We can test our newly trained speech command detection network on streaming audio from microphone. Confusion matrix will be calculated for evaluation of system and prediction of the unknown speech words. The proposed system outperformed for 11 labels with Google TebsorFlow and AIY teams, it contains 105,000 wave audio files and five layer model which achieve accuracy of 94.9% in less training time of 4.5116 sec using GPU.

Keywords—*Graphics processing unit (GPU), Adam Optimization, Convolutional Neural Network (CNN), linear discriminant analysis algorithm (LDAA).*

1. INTRODUCTION

Sound recognition (SR) is the art and science of having machine to identify speech. In the current decade real time speech recognition moving towards hand free working on smart digital systems. Voice technology gives increasing attention for interacting with machine and mobiles using speech signal commands. Google offers one of the best real time speech command technology to search data by recognizing voice. Most of the android phones now provide hands free experience. The fact is that, it is very convenient and effortless for mobile users compared to typing by hands. Since real time speech command recognition can run smoothly on Smartphone, tablet and small device. Hidden Markov models (HMMs) and Gaussian mixture models (GMMs) are popular techniques to recognize speech with less inaccuracy [1]. In current era, tremendous generation of data requires deep learning methods. From deep learning methods deep neural networks found

ISSN: 2005-4297 IJCA

promising results in terms of accuracy and time in automatic speech recognition from previous decades [1,8]. The Convolutional Neural Network architecture is composed of several filter layers which votes best classification results in speech processing [2]. The main aim of our project is to use novel work on keyword spotting system to detect different predefined speech commands in real time application using CNN deep neural network [3] and ADAM optimization [6] on GPU. For performance evaluation we have used GPU because it is a computation powerhouse with accelerate processing for large datasets. We are using dataset provided by Google Tensor Flow and AIY team; it contains 105,000 wave audio files. Such large dataset with 11 different classes is not used in previous speech recognition research. One of the contribution to implement this large dataset recognition and increase the performance, we have performed our experimentations on GPU.

In our paper, proposed working is performed in three categories like splitting of Data into Training, Validation, and Testing of datasets. On processing through different phases computation of Speech Spectrograms is done. To prepare the data for efficient training of a convolutional neural network, convert the speech waveforms to log-Mel spectrograms. Compute the log-Mel spectrograms for the training, validation, and test sets. To obtain data with a smoother distribution, take the logarithm of the spectrograms. At the end Confusion matrix has been calculated for evaluation of system and prediction of the unknown speech words. Section 2 describes literature findings in speech classification. Section 3 describes preprocessing and spectrogram details. Proposed work implementation algorithms specified in section 4. Discussions about results are elaborated in section 5.

2. RELATED WORK

Sound recognition is a key strategic technology in embedded software which present in every device as application across every smart phones or speech follower to interact for human such as Google, Amazon Alexa and Apple Siri. Typical approaches in speech recognition are based on perceptive the components of human speech. New machine learning methods can lead to important advances in automatic speech recognition (ASR). For better understanding the difficulties associated with ASR, it is essential to comprehend the production of speech sounds and sources of changeability [7]. Sound Recognition Systems can be categorized into various groups based on the constraints forced on the nature of the input signal speech like number of speakers, vocabulary size, nature of utterance, spectral bandwidth etc[7].

Hidden Markov Models (HMMs) are the widely accepted models used in the area of uninterrupted sound recognition. Hidden Markov Model (HMM) is a powerful technology of machine learning uses for classification of dataset. Using neural networks as acoustic models for HMM-based speech recognition formerly was introduced over 18 to 20 years ago [11, 12]. Much of this unique work implements the basic ideas of hybrid DNN-HMM systems which are used in Modern system. In this use the general techniques like KWS is the Key-word/Filler [13, 14, 15, 16].

In recent times, deep learning-based approaches verified performance improvements over conventional machine learning methods for many different applications [17]. The neural networks built with memory capabilities have made speech recognition mostly accurate [8, 17, 18]. Each label of dataset is getting trained from an HMM model, and a filler model HMM is trained from the non-label segment of the speech signal (filters) [10]. But HMM method is quite computational expensive, since HMM requires Viterbi decoding. However the large-margin formulation model [19, 20] work based or recurrent neural network [21, 22] are promising than HMM technology. But they have relatively long-latency, since they either require to process over the whole speech to find the region of the keyword or take inputs from a long period of time to predict the keyword. Recently era of Keyword spotting (KWS) system at Google [9] applied a deep neural network (DNN),

which left back the traditional HMM system and is also very accurate, less time consuming and requires relatively lower computation. Further art of improvement for the above system is a CNN model can give better performance over a DNN model in a variety of small and large vocabulary tasks [5, 10, 23].

Deep learning-based approach presented for speech command categorization is proposed in [2], which uses 1D and 2D wave frames and spectrograms input speech commands in convolutional networks. The paper findings elaborate that similar results are given by all the models based on 1D convolution neural networks also. Due to outlier detection and variance reduction, embedded ensembles performs better.

Three different models CNN, DNN and Vanilla testing for KWS commands is done in paper [3]. With the use of MFCC feature extraction CNN model do better than the other two models. Second Convolutional Layer uses large number of multiplies is the drawback of the method. So requires new CNN based model to reduce multiplies with improved performance.

Fast and simple deep KWS system is proposed in [4] paper. The deep KWS trained with only KW data is highly performed than any other. This work is limited to small data. Small footprint keyword spotting classification using limited multiplies of CNNs to DNNs are proposed in [5] paper. It shows that CNN gives improvement over a DNN in both clean and noisy conditions.

In next future the exponential growth of speech recognition in which based devices that will become assistant to our daily needy lives. Only required to develop and optimize sound recognition algorithms that support for real time application and work enough to support many different embedded platforms [2].

Here we can say that CNNs model cross over two model DNNs for KWS task for mainly two advantages. First, DNNs applies concept of column vector instead input topology. However, for input speech signals, the spectrum representations show very strong correlations in time and frequency. Due to this modeling local correlations with CNNs get improved and these cause much better performance than DNNs. Second, recognizing parameter is sharing quality of CNNs, CNNs support very few parameters compared to DNNs for the same task, means reducing memory footprint and computational requirement. Thus observation conclude for that CNNs have enhanced routine and summary model size over DNNs and is thus the state-of-the-art technique for KWS task.

3.PREPROCESSING and SPEECH SPECTROGRAMS OF DATASET

A. Speech Dataset

The deep learning model detects the presence of specific words in audio file. This research work uses the Speech Dataset developed by Google to train a convolutional neural network to recognize a speech word. The working concepts of dataset are proposed by splitting of Data into Training, Validation, and Test Sets. We downloaded the Google Speech dataset; it contains 105,000 wave audio files. The dataset takes for different labels which are near about 11 and each label has been consider for count of 2300 to 2400 times and unknown word count is near about 8193. The key word class, labeled separately, contains a set of disturbed words. In this project, the dataset consist of various labels like “Down, go, Left, No, Off, ON, Right, Stop, Up, Yes” and Unknown keywords. To prepare the data for efficient training of a convolutional neural network, we converted the speech waveforms to log-Mel spectrograms. Compute the log-Mel spectrograms for the training, validation, and test sets. To obtain data with a smoother distribution, take the logarithm of the spectrograms. The strange word audio clips confine the words, which are not concerned about this class is differed from the obtained key word class. In further process to make system to be more robust for number of parameter, original signal get supplement with robust noise which is proportional to the training set is added. Such noisy environments audio signal applied for recognition and it is possible by using CNN along with Adam optimization techniques. Throughout our system the supplementary

noise of volume and frequency inserted in proper ratio, so system can work in noise environment also. Finally the pre-processing network must be able not only to recognize different spoken words but also to detect if the input contains silence or background noise.

B. SPEECH SPECTROGRAM

To prepare the data for efficient training of a convolutional neural network, convert the speech waveforms to log-Mel spectrograms. In our paper perform three different phases training, validation and testing are used when data applied for input to output classes. Visual illustration of the spectrum of frequencies of a signal changing with time is called as spectrogram. In this paper, log-Mel spectrograms computation is done and then applied for the training, validation, and testing. Logarithm of the spectrograms used to obtain data with a smoother distribution. On the basis of recorded input signal we plotted the waveforms and corresponding spectrograms of few training samples. The spectrogram has been plotted for the actual dataset used in our paper. The spectrogram and playing of particular audio clips file for few dataset has been shown on sample basis as example shown in figure 2.1.

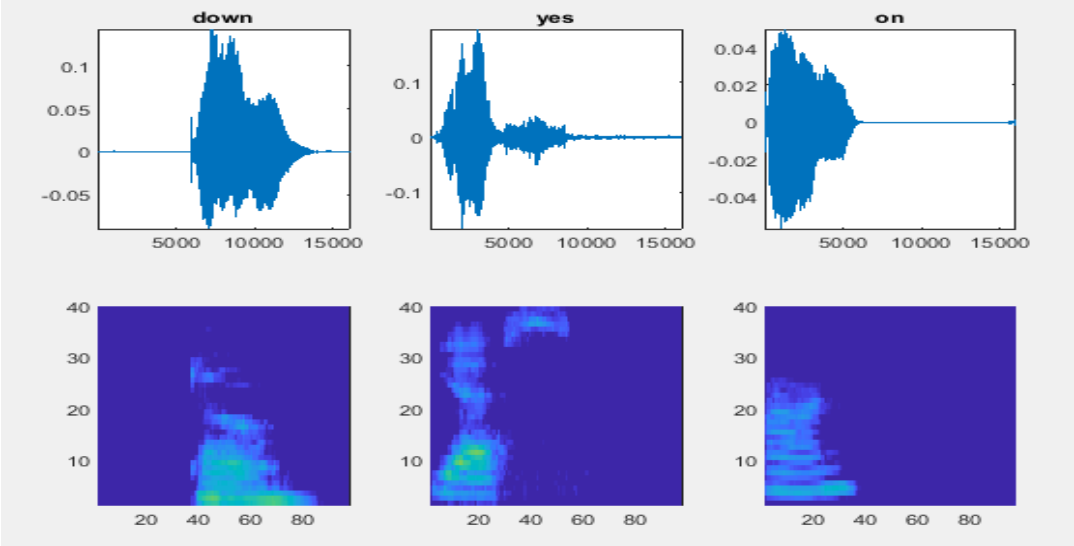

Figure2: Spectrograms of sample words

4. PROPOSED METHODS AND ALGORITHMS

The proposed method uses CNN and Adam Optimization algorithm[5] which is being trained by slandered Graphics processing unit (GPU) NVDIA Tesla K40 C GPU. The GPU is the high processing unit on which training of sample achieved in less time and its support to improve our system for training and testing.

4.1 TRAINING AND VALIDATION OF DATA

Training is the second phase of speech recognition, the input speech wave files function through this and creates trained sample from preprocessed data. Particularly training is fundamental point for every researcher and it must perform in a limited time period. Proposed method suggested about 11 keyword dataset and these are applied for the training phase. We are using different hardware than previous work called as GPU machine which has tremendous speed of execution and reduces the time of training the input speech wave file with large dataset. The numbers of Training samples are taken for extracting features of input speech signal. CNN training is simplest when the inputs to the neural network have a practically smooth distribution and are normalized. We Plotted histograms of values of training samples for checking smoothness in the data distribution.

Various data label along with exact count extracted from Google dataset for training are shown in table 1.

Table 1: Training Data for CNN

Sr. No	Label	Count
1.	Down	2359
2.	Go	2372
3.	Left	2353
4.	No	2375
5.	Off	2357
6.	On	2367
7.	Right	2367
8.	Stop	2380
9.	Unknown	8193
10.	Up	2375
11.	Yes	2377

Plot the distribution of the different class labels in the training and validation sets. The test set has a very similar distribution to the validation set.

In the process of training augmented data store is created for automatic augmentation. Resizing of the spectrograms is done in training dataset. Spectrogram randomly translated up to 100 ms of 10 frames forwards or backwards in time. Translating is followed by the spectrogram scaling along the time axis up or down by 20 percent. Scaling of data can increase the effectiveness of the training data. It also helps to prevent over fitting of network. The augmented image data store creates augmented images in real time during training and inputs them to the network. No augmented spectrograms are saved in memory. So we are using an Adam optimization algorithm, which is used for training of CNN. Adam optimization is computationally efficient with less memory requirement than other optimization algorithms. This optimization is best suited for problems that are huge in terms of data along with parameters. It is also suitable for problems with very noisy/or sparse gradients. So Adam optimization gives more efficiency in training CNN. Hyper-parameters have intuitive interpretation and typically require little tuning. Training and validation of label with respective to its distribution is shown in fig.2. All Experiments are carried out on CPU or GPU (NVDIA Tesla K40 C GPU).

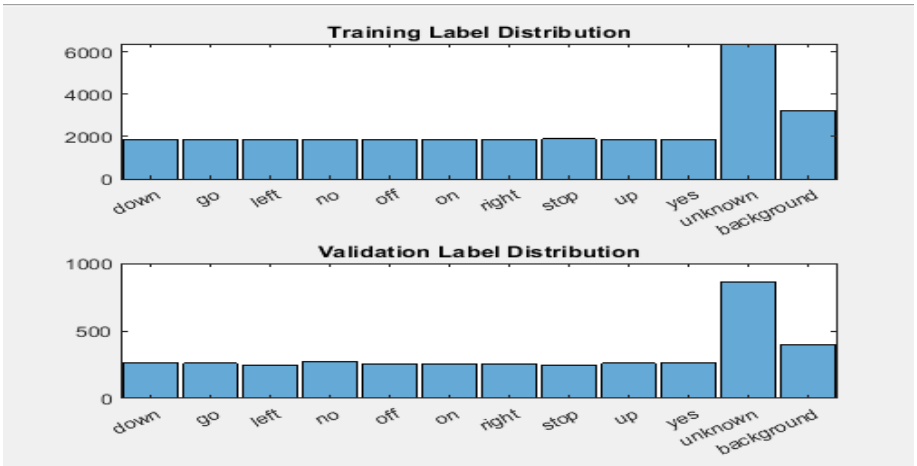

Figure: 3 Training and validation label distribution

4.1.1 Neural Network Architecture

Create simple network architecture as an array of layers. Use convolutional and batch normalization layers, and down sample the feature maps "spatially" (that is, in time and

frequency) using max pooling layers. 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 network has following Sequence of layers in CNN. It uses five convolutional layers with few filters. Sequence of number of layers proposed in CNN are been given as follow.

a. Data Input Layer
b. Convolution 2d Layer
c. Batch Normalization Layer
d. Relu Layer
e. maxPooling2dLayer
f. convolution2dLayer
g. Batch Normalization Layer
h. Relu Layer
i. maxPooling2dLayer
j. convolution2dLayer
k. Batch Normalization Layer
l. ReluLayer
m. maxPooling2dLayer
n. convolution2dLayer
o. batch Normalization Layer
p. Relu Layer
q. convolution2dLayer
r. batch Normalization Layer
s. Relu Layer
t. maxPooling2dLayer
u. dropout Layer
v. fully Connected Layer
w. Softmax Layer
x. weighted Classification Layer

4.2 TESTING PHASE OF DATA USING NETWORK

This is the third phase of real time speech recognition system, which proves with better and very efficient results. In this step calculate the final accuracy of the network on the training set (without data augmentation) and validation set. The network is very accurate on this data set. However, the training, validation, and test data all have similar distributions that do not necessarily reflect real-world environments. This limitation particularly applies to the unknown category also, which contains utterances of only a small number of words with the counter of 8193.
We are calculating confusion matrix for evaluation of system and prediction of the unknown speech words. Plot the confusion matrix. Display the precision and recall for each class by using column and row summaries. Sort the classes of the confusion matrix. For classification linear discriminant analysis algorithm is used. Our system work for both recorded and real time data also. Detect spoken words using live audio from user. We tested our newly trained speech command detection network on streaming audio from microphone. Try saying one of the known commands. Then, try saying one of the unknown words. Specify the audio sampling rate and classification rate in Hz and create an audio device reader that can read audio from your microphone.

5. RESULTS

Firstly due to huge amount of data, we choose GPU to train our network. We have referred the training data from given table 1. Training, validation and prediction error of 11 key datasets of proposed system are shown in table 2. From table we can conclude that when we are applying large Google Tensor flow data by training CNN using Adam optimization on GPU not only performance is increased but error is reduced for training and testing. One of the contribution of the proposed method is that there is large improvement in time for prediction of test data as compared to previous research work done on the speech recognition. Our proposed method requires only 4.51ms time to validate and test large volume of data.

Table 2: Training, Validation and Predictions

Sr. No.	Working Phase	Values
1	Training error	17.7609%
2	Validation error	19.4893%
3	Prediction Time on CPU	4.5116 ms

To elaborate results of the proposed system we plotted the Confusion matrix and it is been calculated for evaluation of system and prediction of the unknown speech words. The Confusion matrix is shown in table 3. Improved result about 94.9% from the existing system in fraction but improves in the training by reducing training time. The current system plot confusion matrix by considering precision and recall factor and result gives as.

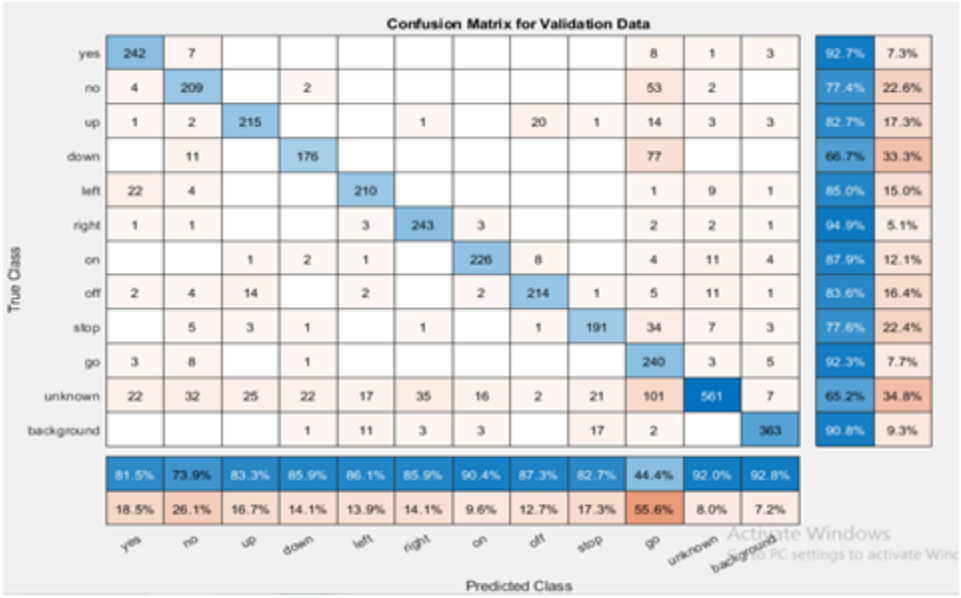

Figure 4: confusion Matrix (precision and recall)

Important benefits of using Adam optimization on other optimization algorithms with the applicability of GPU training perform result is prominent than existing system.

6. CONCLUSION

As speech recognition is promising towards hand free computation, it not only requires accuracy but time efficiency also crucial. The proposed work improves results by all angles of Training, validation and Testing. We have referred huge Google Speech dataset; it contains 105,000 wave audio files and extracts log-Mel spectrogram for the same. In first phase Adam optimization algorithm is used for training of Convolutional Neural Network. Such training performed through high configures CPU or GPU (NVDIA Tesla K40 C GPU). The proposed system architecture uses five convolutional layers with few filters. Finally accuracy has been calculated for network on the training set which requires

very less time for given dataset. It takes only 4.5116 ms which outperforms all existing methods of speech recognition. Use of Adam optimization on other optimization algorithms makes computationally efficient, little memory requirements and appropriate for problems with very noisy/or sparse gradients on data. Hyper-parameters have intuitive interpretation and typically require little tuning. Due to all supporting algorithm CNN plays important role to improve the performance testing result 94.9% better than existing system in fraction. By consideration of results we conclude that our system is better for real time applications for working on commands by interfacing with hand held device with hands free communication.

RFERENCES

[1].Dong Yu , Li Deng, "Automatic Speech Recognition: A Deep Learning Approach", *Springer Publishing Company, Incorporated*, 2014.
[2].Solovyev, Roman A., et al., "Deep learning approaches for understanding simple speech commands." arXiv preprint arXiv:1810.02364, 2018.
[3].Li, Xuejiao, and Zixuan Zhou , "Speech Command Recognition with Convolutional Neural Network." , Semanticscholar , 2017.
[4].G. Chen, C. Parada, and G. Heigold, "Small-footprint Keyword Spotting using Deep Neural Networks," in Proceedings ICASSP, 2014.
[5].T. Sainath, C. Parada, "Convolutional neural networks for small-footprint keyword spotting", Proceedings Interspeech, pp. 1478-1482, 2015.
[6].Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization.", arXiv preprint arXiv:1412.6980 , 2014.
[7].Gaikwad, Santosh K., Bharti W. Gawali, and Pravin Yannawar. "A review on speech recognition technique.", International Journal of Computer Applications 10.3 (2010): 16-24.
[8].G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82 –97, nov.2012.
[9].G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neural networks for large-vocabulary speech recognition," IEEE Trans. Audio Speech Lang. Processing, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[10].Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, "Speech Recognition With Deep Recurrent Neural Networks," IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, Print ISSN: 1520-6149 Electronic ISSN: 2379-190X DOI: 10.1109/ICASSP.2013.6638947, May 2013.
[11].Herve A Bourlard and Nelson Morgan. "Connectionist speech recognition: a hybrid approach" Springer Science & Business Media, volume 247, 2012.
[12].McClelland, James L., and Jeffrey L. Elman. " The TRACE model of speech perception." Cognitive psychology 18.1 (1986): 1-86.
[13].Rohlicek, J. Robin, et al. "Continuous hidden Markov modeling for speaker-independent word spotting." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.
[14].Rose, Richard C., and Douglas B. Paul. "A hidden Markov model based keyword recognition system." International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1990.
[15].Wilpon, J. G., L. G. Miller, and P. Modi. "Improvements and applications for key word recognition using hidden Markov modeling techniques." International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1991.
[16].Silaghi, Marius-Calin, and Hervé Bourlard. "Iterative Posterior-Based Keyword Spotting Without Filler Models: Iterative Viterbi Decoding and One-Pass Approach." Tech. Rep. , 2000.
[17].LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.
[18].Ian, Yoshua Bengio, and Aaron Courville. "Deep learning." MIT press, 2016.
[19].David Grangier, Joseph Keshet, and Samy Bengio, "Discriminative keyword spotting," Automatic speech and speaker recognition." Large margin and kernel methods, pp. 175–194, 2009.
[20].Tabibian, Shima, Ahmad Akbari, and Babak Nasersharif. "An evolutionary based discriminative system for keyword spotting." International Symposium on Artificial Intelligence and Signal Processing (AISP). IEEE, 2011.

[21].KP Li, JA Naylor, and ML Rossen, “A whole word recurrent neural network for keyword spotting,” in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 1992, vol. 2, pp. 81–84.
[22].Santiago Fernández, Alex Graves, and Jürgen Schmidhuber, “An application of recurrent neural networks to discriminative keyword spotting,” in Artificial Neural Networks–ICANN 2007, pp. 220–229. Springer, 2007.
[23].L. Toth, “Combining Time-and Frequency-Domain Convolution in Convolutional Neural Network-Based Phone Recognition,” in Proc. ICASSP, 2014.