International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

SPEECH RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Nidhi A. Kulkarni[1], Satish P. Deshpande[2]

[1]*Student, Dept.of Electronics & Communication, KLSGIT College Belgaum, Karnataka, India*
[2]*Assistant Professor, Dept.of Electronics & Communication, KLSGIT College Belgaum, Karnataka, India*

Abstract - *Language is one of the important means of communication; Speech is its main medium by which two people can communicate. In this present era, speech technologies are widely used as it has unlimited uses. Speech recognition, as the man-machine interface plays a vital role in the field of artificial intelligence where accuracy is a major challenge. In this paper, a review of the speech recognition process, its basic models, and its application is done. Discussion is made on different techniques and approaches of the speech recognition process using the Convolutional Neural Network (CNN). This paper also summarizes some of the wells known methods used in various stages of the speech recognition process. A convolutional neural network is one that reduces spectral variations and models spectral correlations present in the signals. The main objective of this review is to bring to light the progress made in the field of speech recognition that uses a convolutional neural network of different languages and technological viewpoint in different countries.*

Key Words: **Speech Recognition, Man-Machine, Artificial Intelligence, Convolutional Neural Network, Spectral Variation.**

1.INTRODUCTION

Speech is one of the most important media for communication between two people and his environment. The speech recognition system is capable to translate spoken words into text. Text can be either in terms of words or sequences of words or it could be in syllables form. Speech recognition using the convolutional neural network is one of the challenging tasks. We know that human speech signals are highly variable due to different speaker attributes, different speaking styles and can include environmental noises so on; hence speech recognition is done by using a convolutional neural network. In the speech recognition tasks, many parameters will affect the accuracy of the recognition. These parameters are as follows: a continuous word or discrete word recognition, vocabulary, environmental conditions, models, etc. problems such as one may pronounce the same word in a different manner, expressing a word by two different speakers may vary. Resolving these problems is the main step towards the aim. The main advantage of using speech recognition is that a person can save his time by directly speaking to the devices rather than typing continuously.

1.1 CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network is a class of deep neural network which does little preprocessing, that means that the neural network learns the filter before doing the real classification. It consist of single or more than one layer, CNN can do lot of things when they are fed with bunch of signals as input to it. For computer system it is difficult to consider whole signal hence by using Convolutional neural network it uses a part of signal instead of considering entire signal, CNN is a type of neural network where input variables are related spatially to each other. CNN were developed specially to take spatial positions into account.

Filtering is the main mathematics behind the matching, matching is done by considering the features that are lined up with this patch signal then the pixels are compared and multiplied one by one, next these features are then added and divided with the total number of pixels. This step is repeated for all pixels. The act of convolving signals with a bunch full of filters or bunch of features is called as convolutional layer. It is a layer in which operations depends on the stack that is in convolution one signal becomes a stack of filtered signals. Convolutional layer is one among the layer.

Pooling is another important and main building block of CNN. Its function is to reduce the amount of parameters and computation used in the network. Pooling layer operates on each feature independently, the most common approach used is max pooling. Once pooling is done next step is normalization. In normalization, if the pixel value is negative, then these negative values will be replaced with zeros. This process is done to each and every filtered signal. The last step is to stack up all three layers so that present output will become the input for the next. The final layer is the fully connected layer.

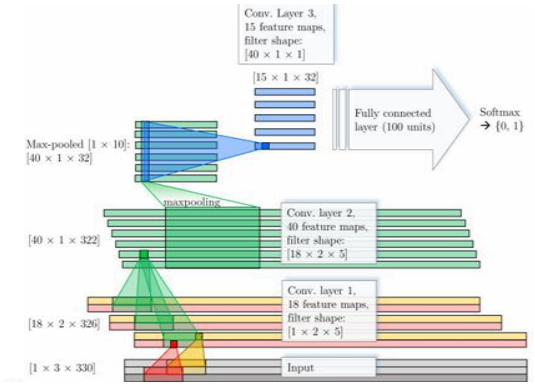

Figure1.1: CNN STRUTURE

2. LITERATURE SURVEY

D. Nagajyothi and P. Siddaiah made a brief survey on speech recognition using a convolutional neural network. This system was developed natively for Telugu language. The database was created based on the frequently asked questions in an airport inquiry. CNN was used for training and testing of the database. It results with the best results as CNN is best for weight sharing, local connectivity and pooling. Experiments performed on speech signals resulted in improvement in the performance of the systems as compared to traditional systems. In this system, they have used tanh as an activation function for the first layer, while the rectified linear unit is used for the next two layers. The input data was normalized to increase the learning speed. The network structure was improved by using Max pooling in the last layer. Later they concluded that CNN based system is able to achieve good performance than a conventional traditional system.

Akhil Thomas presents a survey on feature extraction technique used for speaker recognition by using Mel frequency Cepstral coefficients (MFCC). Further, he evaluated experiments results conducted along each step of the MFCC process. Later MFCC coefficients were retrieved using DCT. DCT was done using CORDIC Algorithm. One disadvantage of this system is that the performance degrades if there is the presence of noise. The system gave the most accurate results when implemented in an environment where it was really trained. Here the performance was increased with the number of training iterations. The accuracy of detected speech was very high because the MFCC feature extraction technique was used. The advantage of using the CORDIAC algorithm is that it reduces the number of gates and thus it also decreases the gate delay.

Du Guiming, Wang Xia, Wang Guangyan, Zhang Yan and Li Dan presents a brief survey on the convolutional neural network that uses backpropagation to train the network. In this discussion, they have used a group of speech that was recorded personally by some people as training data. The experimental result shows that CNN can efficiently implement word recognition with fewer complexities. The backpropagation algorithm is used to adjust and maintain the parameters of the neural network. First input signals were passed to the input layer, the middle layer and then the output was passed through the output layer, If the output is not equal to the desired output then it results into error signal, then by using Backpropagation algorithms these parameters are adjusted. In this paper the neural network is trained for 30 people, once training is done then it uses the other five people's voice as test data. Lastly, they compared CNN with DNN and concluded that CNN greatly reduces the complexity of the system.

Akhilesh Halageri, Amrita Bidappa, Arjun C, Madan Mukund Sarathy and Shabana Sultana says that speech recognition usually consists of the following steps, The first step is to capture and digitize the sound waves, then it is converted into phonemes. The main purpose of this paper was to review the pattern matching abilities of neural networks on speech signal. For feed-forward networks, the activation function is used basically to stabilize the output layer but this process is not present in recurrent networks. A system is proposed to use some learning algorithms to learn the features without any other assumptions. Algorithms use neural networks to increase the computational power of the proposed system.

Ying Zhang, et al, proposed an end to end speech framework inspired by the advantages of both CNN and CTC approach, CNN and CTC approaches were combined without using any recurrent connections. By evaluating this approach on the TIMIT Phoneme recognition task, they showed that the proposed model was not only computationally efficient but it also exists in baseline systems. In this paper, they showed that in a CNN of sufficient depth, higher layer features have the capacity to capture temporal dependencies with suitable information. This model consists of 10 convolutional layers and 3 fully connected hidden layers. The first convolutional layer is followed by the pooling layer; the pooling layer is 3*1, which means we can only pool over the frequency axis. The filter size 3*5 is used across the layers. Max out with 2 piece linear functions is used as activation function and to optimize the model Adam optimizer is used with learning rate 10^4. Their model has achieved 18.2% error rates on the core test dataset, which is better than the LSTM baseline model. We use the CNN model because it takes less time to train the dataset as compared to the LSTM model.
Narendra D. Londhe, Ghanshyam B. Kshirsagar, and Hitesh Tekchandani proposed a Deep Convolutional Neural Network (DCNN) based ASR for Chhattisgarhi dialect because speaker dependent speech recognition system using conventional machine learning technique was incapable to handle the spectral variations and spectral correlation of acoustic signals. DCNN efficiently handles the spectral variation and spectral correlation of speech signals with a less computational burden. For this experiment, a self-

recorded dataset acquired from 170 subjects was used for word recognition. In this paper, they have used 8 layers of convolution, pooling, and fully connected layer, respectively. Rectified linear unit is used as an activation function. In this experiment, 5-fold cross-validation is used for testing and training of the CNN model. The data was portioned into 4:1 form, i.e. 4 parts were used for training and the remaining one part was used for testing. The implemented algorithm achieved 99.49% accuracy. The proposed DCNN model gave good results as compared to conventional techniques.

Xuejiao Li and Zixuan Zhou aimed to build an accurate, low-latency speech command speech recognition system that was capable of determining predefined words. Speech command dataset provided by Google's Tensorflow and AIY teams is used for training and testing, it consists of 65,000 wave audio files saying thirty different words. Models such as Vanilla single-layer Softmax model, Deep Neural network and convolutional neural network are used, where convolutional neural network proves to outperform the other two models and achieved an accuracy of 95.1%. The first observation made was that vanilla is not a good model because accuracy achieved was about only 56%. The experiment results showed that CNN is more effective than DNN and vanilla, giving 18.6% relative improvement over DNN and 72.3% over Vanilla on precision value. A simple 2-layer ConvLayer CNN network outperforms the Vanilla and DNN and achieves 31.43%, 66.67% relative improvement with regard to DNN and Vanilla in test accuracy and achieved 82% and 94.6% in the loss.

Jui-Ting Huang, Jinyu Li, and Yifan Gong aimed to provide a detailed analysis of CNN's. They showed that edge detector along various directions can be automatically learned by visualizing the localized filter learned in the convolutional layer.CNN provides advantages over fully connected layer in four domains: channel-mismatched training-test conditions, noise robustness, and distant speech recognition and low footprint models. CNN structure is established combined with max out units which gave relative 9.3% WERR. In this paper, all experiments are performed under the context-dependent deep neural network framework, were CNN or DNN is used to classify the input features into classes. CNN architecture uses one convolutional layer followed by one max-pooling layer and four fully connected layers. The training data of about 1000 hours of audio data recorded by the kinetic device is used for this experiment, whereas test data consist of 18683 utterances recorded at 1, 2, or 3 meters away from kinetic devices. CNN with random initialization can be used to learn various sets of edge detectors to extract low-level features. For distant speech recognition, CNN is trained on 1000 hours of Kinect distant speech data and obtains relative 4% of word error rate reduction (WERR).

Ossama Abdel-Hamid et al conducted an experiment on two speech recognition tasks and evaluated the effectiveness of CNN's in automatic speech recognition: small scale phone recognition in TIMIT and large vocabulary voice search (VS) task. The results show that CNN reduces the error rate by 6%-10% as compared with DNN's on both TIMIT phone recognition and the voice search large vocabulary speech recognition. In this paper, the discussion is also made on how to organize speech feature vectors into feature maps which suit for CNN processing. The building block of the CNN consists of a pair of hidden plies: A pooling ply and convolutional ply. The input layer consists of localized features that are organized as features map. In this paper, a hybrid ANN-HMM framework with a softmax output layer is used to compute the posterior probability for all HMM states. The likelihood of all HMM states is passed through a Viterbi decoder to recognize the continuous stream of speech. Finally, the pretraining of CNN's based on convolutional RBMs gives better performance in large-vocabulary voice search as compared to the phone recognition experiment.

William Song and Jim Cai implemented an end-to-end deep learning system utilizing mel-filter features to directly output to spoken phonemes without using traditional hidden Markov model for decoding. In this paper TIMIT, the dataset is used for testing and training, which comprises of 630 speakers with 6300 utterances in 8 dialectics. They tried to implement the CNN model using the caffe framework, an error rate of 22.1% is obtained using the CNN model. The decoded phone sequence achieved an error rate of 29.4%. An end to end system is achieved by replacing the HMM model by RNN and CTC. They also tried to replace the traditional HMM system with a network capable of outputting sequentially labels on input data. The training is done using the CNN model then the last layer of convolutional is fed into the CTC network for training, such that prediction of actual phone sequences is done. This is nothing but the procedure where CNN is trained first and then freezing of CNN parameters is done for fine tuning the network for other tasks. Lastly, they concluded that CNN performs better than the CTC model.

3. SUMMARY

Speech is one of the prominent and primary means of communication between two humans. Speech recognition using the convolutional neural network is an emerging field of research with a wide range of applications. In this paper, a survey is made on various algorithms, speech recognition models used for speech recognition purposes. Survey says that the accuracy of speech recognition depends on which dataset used and the model on which it is trained. The error rate greatly depends on the environmental conditions. The survey points that CNN based system gives better accuracy and boosts the performance of the system due to its unique features like local connectivity and weight sharing, than the conventional traditional system. CNN is widely used in applications related to computer vision, image classification

and pattern matching because it mitigates most of the conventional traditional system problems.

REFERENCES

[1] D. Nagajyothi and P. Siddaiah, "Speech Recognition Using Convolutional Neural Networks", *International Journal of Engineering & technology 7(4.6) (2018)*.

[2] Akhil Thomas, "Speaker Recognition using MFCC and CORDIC Algorithm", *International Journal of Innovative Research in Science, Engineering and Technology Vol. 7, Issue 5, May 2018.*

[3] Du Guiming et al, "Speech Recognition Based on Convolutional Neural Networks*",IEEE International conference on signal and image processing (2016).*

[4] Akhilesh Halageri et al, "Speech Recognition using Deep Learning", *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3206-3209.*

[5] Ying Zhang et al, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks", *arXiv: 1701.02720v1 [cs.CL] 10 Jan 2017.*

[6] Narendra D. Londhe et al, "Deep Convolution Neural Network Based Speech Recognition for Chhattisgarhi", *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN).*

[7] Xuejiao Li, Zixuan Zhou "Speech Command Recognition with Convolutional Neural Network".

[8] Jui-Ting Huang, "An Analysis Of Convolutional Neural Networks For Speech Recognition",
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052.

[9] Ossama Abdel-Hamid, "Convolutional Neural Networks For Speech Recognition*", IEEE/Acm Transactions On Audio, Speech, And Language Processing, Vol. 22, No. 10, October 2014.*

[10] William Son and Jim Cai, "End-to-End Deep Neural Network for Automatic Speech Recognition", *Department of Computer Science Stanford University*