ISSN:0975-9646

Automatic Speech Recognition using different Neural Network Architectures – A Survey

Lekshmi.K.R[#1], Dr.Elizabeth Sherly[*2]

[#]*Research Scholar, Bharathiar University*
Coimbatore, India

[*]*Professor, Indian Institute of Information Technology Management-Kerala*
Trivandrum,Kerala,India

***Abstract*— Speech is the vocalized form of communication based on lexical syntax. Each spoken word is a phonetic combination of vowels and consonants. Automatic Speech Recognition can be defined as computer-driven transcriptions of speech into human readable text. As it is an emerging technique many researchers are attracted to this and achieved progress to a certain extent in recent years. This survey paper aims at explaining the architecture of Deep Neural Network, Convolutional Neural Network and Recurrent Neural Network and their performance in the field of Automatic Speech Recognition. We also summarise main contributions of various researchers during 2010 – 2016 on Acoustic Modeling and Language Modeling (main components of Automatic Speech Recognition) using these architectures and pointing out their impact in ASR. We conclude this paper with a comparative study regarding the advantages of the architectures discussed during the survey with respect to Word Error Rate (WER), Phone Error Rate (PER) etc. in the area of Automatic Speech Recognition (ASR).**

***Keywords*— Automatic Speech Recognition, Recurrent Neural Network, Hidden Markov Model, Long Short Term Memory, Connectionist Temporal Classification, Deep Neural Network, Convolutional Neural Network**

I. INTRODUCTION

Speech is the ability to express thoughts and feelings by articulate sounds. Each uttered word is a combination of vowel and consonant speech segment. Automatic Speech Recognition (ASR) is the ability of a machine to identify the words spoken by a human being in a machine understandable format. The automatic speech recognition process involves the process of converting a speech signal to a sequence of words. Speech recognition varies on many factors like types of speech, the dependency of speaker, size of vocabulary, types of recognition like an isolated word, connected word and continuous word etc. There exist many speech recognition techniques like Acoustic phonetic approach, Pattern recognition approach, Support Vector machine approach and Artificial Intelligence. The successful model for speech recognition is Hidden Markov Model (HMM) with an acoustic model based on Gaussian mixtures. Figure 1 shows a general architecture of an Automatic Speech Recognition system. ASR consists of Feature extraction, Acoustic Modeling, Language Modeling, Lexicon and Decoder. Acoustic Modeling deals with the acoustic feature related to a speech signal. The Language model is the core component of any speech recognition system.

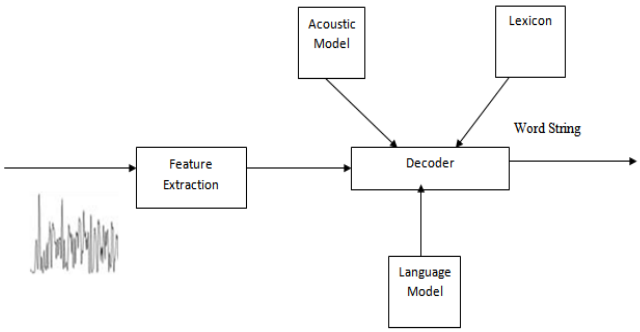

Fig1: A typical Automatic Speech Recognition System

The paper is organised as follows, in Section 2, different Neural Network Architectures are explained in detail. Significant contributions of various researchers towards ASR are discussed in Section 3.This section also presents the database used by various researchers and their achievement in the field. In Section 4 we point out our future research plan in ASR together with the concluding remarks.

II. DIFFERENT NEURAL NETWORK ARCHITECTURES

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Deep Neural Network (DNN)

An Artificial Neural Network (ANN) with multiple hidden layers of units between the input and output layers are called Deep Neural Network (DNN). DNN can model complex non-linear relationships. A DNN is a typically designed feed forward neural network (with more than one hidden layer) discriminatively trained with standard backpropagation algorithm. The bottom layer (first layer) of DNN is the input layer and the topmost layer is the output layer. DNN architecture is shown in Figure 2. In this we can see there exist many hidden layers - makes it Deep Neural Network.

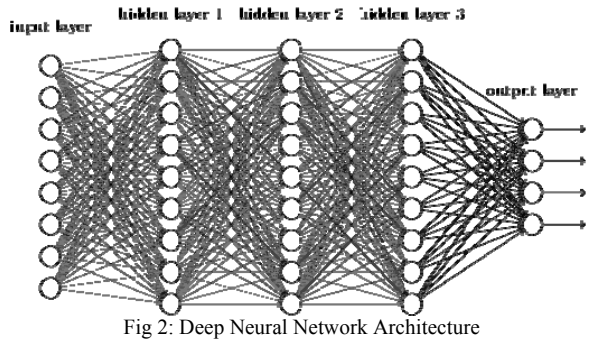

Fig 2: Deep Neural Network Architecture

B. *Convolutional Neural Network (CNN)*

The Convolutional Neural Network is a feed-forward artificial neural network, which consists of convolutional and pooling layers as shown in Figure 3.

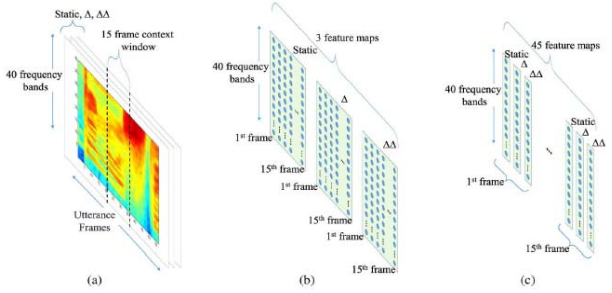

Fig 3: Convolutional Neural Network Architecture

Two different ways can be used to organize speech input features to a CNN. In Figure 3 40MFSC features plus first and second derivatives with a context window of 15 frames for each speech frame are considered.

After extracting the feature from the input a feature map is generated. Once the input feature map is generated the convolution and pooling layers apply their respective operations to generate the activations of the units in those layers. According to CNN terminology, one CNN "layer" means a pair of convolution and pooling layers in succession.

C. *Recurrent Neural Network (RNN)*

In ANN, all the inputs (and outputs) are independent to each other. There should be a connectivity perception for each word. This is the shortcoming of a traditional neural network. Recurrent Neural Network (RNN) can handle this issue. RNNs are networks that perform the same task for every element of a sequence, with the output being dependent on the previous computations.

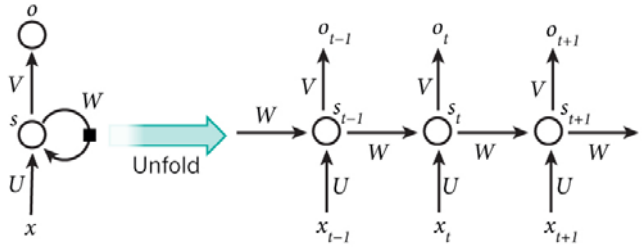

Fig 4: A recurrent neural network and the unfolding in time of computation

The RNN is unfolded into a full network. In RNN for a sentence of n -words, the unfold network would be unrolled into a n layer Neural Network with one layer for each word. RNN architecture is described in Figure 4.

One of the most commonly used types of RNN is – Long Short Term Memory (LSTM) - capable of learning long term dependencies. Hochreiter, S., & Schmidhuber, J. (1997) [12] introduced LSTM, the solution to the vanishing gradient problem. In standard RNN, there will be a single tanh layer. But in LSTM, it is a chain like structure, but repeating modules have a different structure.

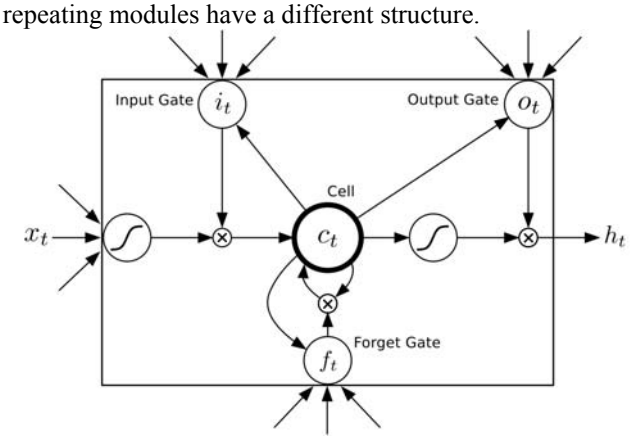

Fig 5: The Long Short Term Memory Architecture Diagram

The LSTM works through a gated cell as shown in figure 5. Information can be stored in, written to, or read from a cell. The cell makes decisions about, what to store, when (it allows) to read etc., through gates that open and close. These gates act on the signal they receive, that is they block or pass the information through them based on signal strength and weights. The weights in LSTM are adjusted via, the recurrent network learning process. The cells will learn when to allow data to enter, back propagating error and adjusting weights through gradient descent.

Bi-directional RNN or BRNN use a finite sequence to predict or label each element of the sequence based on both the past and the future context of the element. This is done by concatenating the outputs of two RNN, one processing the sequence from left to right, and the other one from right to left. Deep Bidirectional LSTM was recently introduced in speech recognition to get lowest error rate on TIMIT database. This can be achieved with the help of discriminative sequence transcription method with RNN - they are connectionist Temporal Classification (CTC) (Maas, A. et.al., 2012)[14] and Sequence Transduction. These methods do not require forced alignment to presegment acoustic data; they directly optimize the probability of target sequence and can learn an absolute language model from acoustic training data.

III. Literature Survey

Panchal, A., & Kale, O. (2014) [16] conducted a literature survey on various techniques for speech recognition using RNN. Generally, speech recognition is Gaussian Mixture Model (GMM) – HMM framework. Due to the increased modeling power the system will be very robust, so GMM-HMM can be replaced Neural Network (NN) – HMM this is explained by Seltzer, M. L. et.al.(2013, May) [19] and Weng, C. et.al.(2014, May) [22].

Zhang, S. X. et al. (2015, April) [25] coined a new type of deep neural network. This DNN uses a Support Vector Machine (SVM) as a top layer for classification instead of softmax activation layer. They developed two training algorithms at frame-level and sequence-level to learn Support Vector Machine and Deep Neural Network parameters in maximum-margin criteria. In this approach for frame-level maximum margin training, the parameters of the last layer of DNN is first estimated using multi class SVM training algorithm (Crammer, K., & Singer, Y., 2001) [4]. After training with multiclass SVM, most training frames can be classified correctly. The decoding process of Deep Neural Support Vector Machine (DNSVM) system is similar to standard DNN – HMM, but the posterior probabilities are replaced by scores from DNSVM. The proposed model DNSVM yields 8% relative error rate reduction over standard DNN. The DNSVM system have a substantial influence on the computational efficiency in terms of caching, Parallelization.

Abdel-Hamid, O. et.al. (2014) [1] suggested the use of convolutional Neural Network (CNN) - a variant of standard neural network – for acoustic Modeling. For the colour images RGB (Red, Green, and Blue) if CNN is used for image processing, values can be viewed as 2-D feature maps. But when CNN is used for speech recognition, the input "image" is considered as speech "spectrogram" with static, delta and delta- delta features serving the roles of Red, Green and Blue values. Instead of using conventional MFCC features, they computed log-energy directly from Mel-Frequency Spectral Coefficients- with no DCT (Discrete Cosine Transformation) - denoted as MFSC (Mel Frequency Spectral Coefficients) features along with delta and delta-delta. After the input feature maps are formed, the convolution and pooling layers are applied with their respective operations. A deep CNN with more than two convolution and pooling layers are called convolution ply and pooling ply. To combine the features across all frequency bands, one or more fully connected hidden layers are added on top of final CNN layer before feeding to output layer that is softmax layer. The CNN will compute the posterior probabilities for all HMM states. As a substitute of full weight sharing (FWS) scheme, they suggested limited weight sharing (LWS) for ASR. LWS helps to reduce the number of pooling ply. This allows only the convolution units that are attached to the same pooling units share the same weight. The researchers achieved about 6-10% relative error reduction.

Geiger, J. T. et.al. (2014) [11] proposed to use Long Short Term Memory RNNs for acoustic Modeling in the hybrid NN-HMM system architecture. In this, the LSTM networks predict the HMM states and use these for Acoustic Modeling. The research group developed LSTM software and is freely available (https://sourceforge.net/p/currennt). They compared two different methods with LSTM network and within HMM framework and were able to predict either phonemes or to predict HMM states.

With the help of a forced alignment of training data, the network is trained to predict the phonemes. The predicted phoneme probabilities can be denoted as p(bt | xt). Wöllmer, M. et.al. (2011, May) [24] proved that from phoneme probability frame-wise discrete phoneme predictions can be obtained. They got a state likelihood p(xt |st) by using a mapping from phonemes to HMM states. The researchers concluded that HMM leads to LSTM confusions, not directly model prediction probabilities of LSTM. The neural network is trained to predict HMM states 's'. By forced alignment of HMM system the training targets are generated. The state likelihoods are obtained using Bayes' rule from the resulting posterior probabilities of the network.

Sak, H. et.al. (2014, September) [17] suggested that for large scale Acoustic Modeling in speech recognition LSTM RNN architecture is efficient. In this, the researchers found that some tasks need a large number of input, output unit and memory cells to store temporal contextual information. So learning of LSTM model will become computationally expensive. They proposed an alternative to standard architecture called Long Short Term Memory Projected (LSTMP). In Figure 6 a separate linear projection layer after LSTM layer is seen.

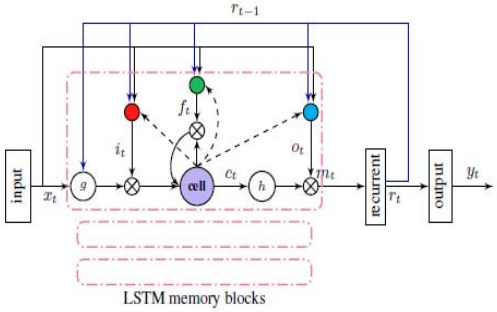

Fig 6: LSTMP RNN Architecture. A single memory block is shown for clarity.

They also suggested deep LSTMP with multiple LSTM layers. Each layer has separate stacked recurrent projection layers to compute the parameter gradients on short subsequences of training utterances. They used truncated backpropagation through time (BPTT) algorithm (Williams, R. J., & Peng, J. 1990) [23]. To optimize the network parameters Asynchronous Stochastic Gradient Descent (AGSD) is used.

Graves, A. et.al. (2013, December) [9] compared the performance of DBLSTM-HMM hybrid with sequence training and measured the possibility of DBLSTM-HMM hybrids for large vocabulary speech recognition. Another advantage of using DBLSTM is that it is able to store the past and future context internally. So the data is presented in a single frame at a time.

Recurrent Neural Network using LSTM architecture has shown state-of-art performance on speech recognition. Zhang et.al (2016) [26] replaced Deep Neural Support Vector Machine for speech recognition (Zhag et.al (2015) [25]) by Recurrent Neural Network. So the top layer – softmax layer- is replaced with Support Vector Machines. The parameters are learned using sequence-level maximum margin criteria as a substitute for cross-entropy. In this work they put forward two training algorithms at frame-level and sequence-level. Using quadratic programming, the

parameters of SVM in the last layer is estimated as the first step. The second step is to update the RNN parameters in all previous layers using sub- gradient approaches. In frame level training and for sequence level training they used a multiclass SVM with RNN features and structured SVM with RNN features respectively. The proposed model known as Recurrent SVM yields 2.8% improvement over standard LSTM.

Sak, H. et al. (2015) [18] investigated that the use of sMBR (state- level MinimumBayesRisk) trained Connectionist Temporal Classification model and introduction of context dependent phone models outperforms conventional LSTM RNN by 8% relative recognition accuracy. CTC is a sequence labelling technique using RNN; where the alignment between the inputs and target labels are unknown. An additional unit for the 'blank' label is used. It is used to estimate the probability of outputting no label at a given time. CTC is implemented with softmax output layer. They used 8-dimensional log Mel-filter bank energy features computed every 10 ms on 15ms window and obtained a significant improvement by increasing number of filter banks. They have shown that they trained the word level acoustic model on medium vocabulary speech recognition to achieve reasonable accuracy without using a language model.

N-gram language model can be replaced by feed forward language model. But RNN cannot train using BPTT because of vanishing gradient problem. A solution to this problem is the use of LSTM. Some basic principles for Neural Network Language Modeling are 1) the input words are encoded by 1-of-K coding where K is the number of words in the vocabulary 2) at the output layer a softmax activation function is used to produce correctly normalized probability values 3) As training criterion the cross entropy error is used which is equivalent to maximum likelihood. Sundermeyer, M. et.al. (2012, September) [21] followed the above mentioned principles with LSTM architecture. They used two hidden layers, the first layer having the interpretation of projecting the input words to a continuous space. Then the LSTM units plugged into the second recurrent layer. They opted to use standard units with sigmoid activation function.The authors showed that the perplexity of the model is constantly lowered by 8% compared to standard RNN. Researchers used RNNLM (Mikolov, T. et.al, 2011, December) [15] toolkit to compare the perplexity results. They suggested that large improvement can be obtained when interpolating LSTM Language Model with a huge Kneser-Ney smoothed backing off model.

Arisoy, E. et.al. (2012, June) [2] proposed Deep Neural Network in language Modeling which outperform conventional n-gram language model in both perplexity and word error rate (WER). Generally Neural Network Language Model (NNLM) is trained with a single hidden layer. Deep Neural Network language Model (DNNLM) has multiple hidden layers which can capture high level discriminative information about input features. In the architecture of NNLM each word in the vocabulary is represented by a N dimensional sparse vector where only the index of the word is '1'and the rest of the entries are '0'. The input to the network is the indices of the previous words. Using linear projection, each word is mapped to its continuous space representations. For DNNLM the researchers used the above architecture and to make the network deep by adding hidden layers followed by hyperbolic tangent nonlinearities. They used cross entropy loss function for training. To evaluate their language odels they used lattice rescoring.

By the Combination of deep bidirectional LSTM recurrent Neural Network and Connectionist Temporal Classification as the objective function a good Word Error Rate can be acquired. Graves, A., & Jaitly, N. (2014) [10] adopted this method and hence they required minimal steps for pre processing. Spectrogram is used for pre processing with bidirectional LSTM network and a CTC as output layer. The network is trained directly on text transcripts without using pronunciation dictionary. To reduce the Word Error Rate (WER) authors introduced a new objective function called Monte – Carlo Sampling. The most probable output transcription y for a given input sequence x was found by picking the single most probable output at every timestamp. With CTC network as output layer they got transition probabilities as output. Researchers also proposed a CTC beam search algorithm to integrate dictionary and language model. The RNN was first decoded without dictionary or language model to calculate WER. They also tried for monogram, bigram and trigram language models.

Song, W., & Cai, J. (2015) [20] combined the recent development in both CNN and RNN to develop an end-to-end speech recognition system using purely neural networks. They implemented frame wise classification with the help of Convolutional Neural Network. The log-filter bank features of input are taken which looks like a 2D image. Another advantage that they identified is that CNN's are invariant against translations of the various frequencies across speakers with pitch difference. This helped them to solve the issue occurred due to age or gender variations. The CNN they used consists of 4 convolutional layers. The first two layers have max pooling and the next two densely connected layers with a softmax layer as output. The activation function used was ReLu. They implemented a rectangular convolutional kernel instead of square kernel.

In the second part the authors replaced traditional HMM decoder with suitable Neural network that can predict sequential label on unaligned data. They got inspiration from (Graves, A. et al., 2006, June) [6] to use CTC as loss function. A language model is combined in association with CTC that is CTC with RNN transducer method is explained in (Graves, A. 2012) [7] and (Graves, A. et al., 2013, May) [8]. With this architecture, they were able to predict the phoneme sequence as final output. The authors implemented CNN using Caffe, a highly optimized tool that makes use of GPU's parallelization to speed up training (Jia, Y. et al., 2014, November) 13]. They used Frame Error Rate (FER) and Phone Error Rate (PER) as their evaluation matrices.

Now we compare the impact of various neural network architecture in Acoustic Modeling and language Modeling as shown in Table I

TABLE I
COMPARISON OF DIFFERENT ARCHITECTURE BY MEANS OF WORD ERROR RATE (WER %), PHONE ERROR RATE (PER %), FRAME ERROR RATE (FER %) AND CROSS ENTROPY (CE)

ACOUSTIC MODELING									
System	**Database**	**WER**	**C**	**P**	**Depth**	**N**	**PER**	**FER**	**CE**
GMM+BLSTM (phonemes) (Geiger, J. T. et.al.) [11]	2nd CHiME medium Vocabulary from WSJ	29.6							
BLSTM (states)		25.0							
BLSTM (phonemes) + BLSTM states		22.2							
LSTMP RNN (Sak, H. et.al.) [17]	——	10.7	1024	512	3L	20M			
		10.7	800	512	2L	13M			
DBLSTM (Noise) (Graves, A. et.al.) [9]	TIMIT	12.0					17.99 ± 0.13	27.88 ± 0.16	0.93 ± 0.004
	WSJ dev93							28.2	1.12
DNSVM (Zhang, S. X. et.al.) [25]	TIMIT						21.90 (Frame - level)		
							21.04 (sequence - level)		
RecurrentSVM (Zhang, S. X. et.al.) [26]	Windows Phone short message diction task	20.69(frame-level)							
		19.83(sequence-level)							
LSTM RNN CTC Model (+sMBR) (Sak, H. et.al.) [18]	Google voice search	12.9(Unidirectional)							
		12.2(Bidirectional)							
CNN HMM with LWS (Abdel-Hamid, O. et.al.) [1]	TIMIT and large vocabulary voice search					4.1M	20.36		

LANGUAGE MODELING						
System	**Language Model (LM)**	**WER**	**FER**	**PER**	**Database**	**Perplexity**
———	KN - 4gram	17.6			Quaero French	
	KN - 4gram + LSTM	17.3				
CNN Model	25 frame window, 128-256-384-384 conv, 1024-512 dense		22.1			
CTC Model	CNN Input, 4 × 2048, λ= 1e-3			29.4	TIMIT	
RNN-CTC	TRIGRAM	8.7			WSJ	
RNN-WER	TRIGRAM	8.2				
DNN Model	DNNLM (h=500, d=120 with 3 layers)	20.8				102.8
	4-gram + DNNLM (h=500, d=120 with 3 layers)	20.5				92.6

IV. CONCLUSIONS

Speech is a prominent and primary mode of communication among humans and most natural way to efficiently communicate between humans. Speech Recognition (SR) using Neural Network Architecture is an emerging filed or research. Speech Recognition has wide range of applications. SR can substitute for differently abled people to fulfill their lives with the help of Speech-to-Text (STT) and Text-to-Speech (TTS) applications. In this paper we review the recent developments on different architectures of Deep Neural Network on Speech Recognition through major works in this field. Various techniques discussed above for ASR are mainly based on English databases and few French databases. This survey compares different Deep Neural Network architectures. This survey compares different Deep Neural Network Architectures and reveals that the DNN-HMM outperformed traditional GMM-HMM due to increased modeling power. We also analysed the properties of Recurrent Neural Network, Convolutional Neural Network, Bidirectional Neural Network and Deep Bidirectional Neural Network. Convolutional Neural Network gives a 6-10 % relative error reduction compared to that of Recurrent Neural Network's 8-10% relative error rate in terms of Word Error Rate. Our effort will be directed towards developing an appropriate technical method in ASR for Malayalam, the native language of Kerala, one of the highest literate states in India.

REFERENCES

[1] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

[2] Arisoy, E., Sainath, T. N., Kingsbury, B., & Ramabhadran, B. (2012, June). Deep neural network language models. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT (pp. 20-28). Association for Computational Linguistics.

[3] Arisoy, E., Sainath, T. N., Kingsbury, B., & Ramabhadran, B. (2012, June). Deep neural network language models. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT (pp. 20-28). Association for Computational Linguistics.

[4] Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research, 2(Dec), 265-292.

[5] Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8599-8603). IEEE.

[6] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376). ACM.

[7] Graves, A. (2012). Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711.

[8] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.

[9] Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). Hybrid speech recognition with deep bidirectional LSTM. In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on (pp. 273-278). IEEE.

[10] Graves, A., & Jaitly, N. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. In ICML (Vol. 14, pp. 1764-1772).

[11] Geiger, J. T., Zhang, Z., Weninger, F., Schuller, B., & Rigoll, G. (2014). Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic Modeling. In INTERSPEECH (pp. 631-635).

[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[13] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678). ACM.

[14] Maas, A., Le, Q. V., O'neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust ASR.

[15] Mikolov, T., Kombrink, S., Deoras, A., Burget, L., & Cernocky, J. (2011, December). Rnnlm-recurrent neural network language modeling toolkit. In Proc. of the 2011 ASRU Workshop (pp. 196-201).

[16] Panchal, A., & Kale, O. A Literature Survey on Recurrent Neural Network and Various Techniques for Speech Recognition.

[17] Sak, H., Senior, A. W., & Beaufays, F. (2014, September). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In INTERSPEECH (pp. 338-342).

[18] Sak, H., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. .ar*Xiv preprint arXiv:1507.06947*.

[19] Seltzer, M. L., Yu, D., & Wang, Y. (2013, May). An investigation of deep neural networks for noise robust speech recognition. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7398-7402). IEEE.

[20] Song, W., & Cai, J. (2015) End-to-End Deep Neural Network for Automatic Speech Recognition.

[21] Sundermeyer, M., Schlüter, R., & Ney, H. (2012, September). LSTM Neural Networks for Language Modeling. In Interspeech (pp. 194-197).

[22] Weng, C., Yu, D., Watanabe, S., & Juang, B. H. F. (2014, May). Recurrent deep neural networks for robust speech recognition. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5532-5536). IEEE.

[23] Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. Neural computation, 2(4), 490-501.

[24] Wöllmer, M., Eyben, F., Schuller, B., & Rigoll, G. (2011, May). A multi-stream ASR framework for BLSTM modeling of conversational speech. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4860-4863). IEEE.

[25] Zhang, S. X., Liu, C., Yao, K., & Gong, Y. (2015, April). Deep neural support vector machines for speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4275-4279). IEEE.

[26] Zhang, S. X., Zhao, R., Liu, C., Li, J., & Gong, Y. (2016, March). Recurrent support vector machines for speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5885-5889). IEEE.