International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

Speech Recognition using Neural Networks

Mr. Hardik Dudhrejia
Department of Computer Engineering
G H Patel College of Engineering & Technology
Vadodara, India

Mr. Sanket Shah
Department of Computer Engineering
G H Patel College of Engineering & Technology
Anand, India

***Abstract*—Speech is the most common way for humans to interact. Since it is the most effective method for communication, it can be also extended further to interact with the system. As a result, it has become extremely popular in no time. The speech recognition allows system to interact and process the data provided verbally by the user. Ever since the user can interact with the help of voice the user is not confined to the alphanumeric keys. Speech recognition can be defined as a process of recognizing the human voice to generate commands or word strings. It is also popularly known as ASR (Automatic speech recognition), computer speech recognition or speech to text (STT). Speech recognition activity can be performed after having a knowledge of diverse fields like linguistic and computer science. It is not an isolated activity. Various techniques available for speech recognition are HMM (Hidden Markov model)[1], DTW(Dynamic time warping)-based speech recognition[2], Neural Networks[3], Deep feedforward and recurrent neural networks[4] and End-to-end automatic speech recognition[5]. This paper mainly focusses on Different Neural networks used for Automatic speech recognition. This research paper primarily focusses on different types of neural networks used for speech recognition. In addition to this paper also consist of work done on speech recognition using this neural networks.**

Keywords— Speech recognition; Recurrent Neural network; Hidden Markov Model; Long Short term memory network

I. INTRODUCTION

Throughout their life-span humans communicate mostly through voice since they learn all the relevant skills in their early age and continue to rely on speech communication. So, it is more efficient to communicate with speech rather than by using keyboard and mouse. Voice Recognition or Speech Recognition provides the methods using which computers can be upgraded to accept speech or human voice assist input instead of giving input by keyboard. It is extremely advantageous for the disabled people.

Speech is affected greatly by the factors such as pronunciations, accents, roughness, pitch, volume, background noise, echoes and gender. Preliminary method of speech processing is the process of studying the speech signals and the methods of processing these signals.

The conventional method of speech recognition insist in representing each word by its feature vector and pattern matching with the statistically available vectors using neural networks. On the contrary to the antediluvian method HMM, neural networks does not require prior knowledge of speech process and do not need statistics of speech data. [3]

Types of speech recognition: Based on the type of words speech recognizing systems can recognize, the speech recognition system is divided into the following categories:

- *Isolated Word*:
  Isolated word requires each utterance to have quiet on both sides of sample window. At a time only single words and single utterances are accepted and it is having "Listen and Non-Listen state".
- *Continuous Word:*
  Continuous speech recognisers provide the users a facility to speak in a continuous fashion and almost naturally and at the same time the computer determines the content of the speech. Recognisers rendering the facility of continuous speech capabilities are pretty much difficult to create because they require some special and peculiar methods in order to determine the boundaries of the utterances.
- *Connected Word:*
  Connected words are very much alike the isolated words but they allow separate utterances to be executed with "minimal pauses" in between them.
- *Spontaneous speech:*
  At an elementary level, spontaneous speech can be considered as a speech that is coming out naturally and not a rehearsed one. An Automatic Speech Recogniser must be able to handle a wide range of speech features like the words being run together.

Classification of speech sounds:
In this modern time, the process of classification of speech sounds is commonly done on the basis of 2 process based on how the classification process is looked upon:

Based on the process of obstruction and non-obstruction sounds

The process of classifying the sounds with respect to the process of obstruction and non-obstruction relies upon the conception of bodily air. While generating human sounds, the air coming out of the body has two functions; it is obstructed in the mouth or throat somewhere or it doesn't get obstructed, but the air comes out very easily. Correspondingly, the sounds that are produces as a result of obstructions and non-obstructions are not same excluding some of their qualities that are trivial.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

For e.g.- all the vowels (a,e,i,o,u) are non obstruction speech sounds and all the consonants (b,c,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z) are obstruction speech sounds.

Based on the process of voice and voiceless sounds

Voiced sound is produced when the vocal chords vibrate when the sound is produced. Whereas in the voiceless sound no vocal cord vibration is produced. To test this, place your finger on your throat as you say the words. A vibration will be felt when the voiced sounds are uttered and no vibration will be felt while uttering a voiceless sound. Many a times it is difficult to feel the difference between them. So in order to distinguish between them another test can be performed by putting a paper in front of our mouth and the paper should move only by saying the voiceless sounds. All the vowels are voiced whereas some of the consonants are voiced as well as voiceless.

Voiced consonants are :- b,d,g,v,z,th,sz,j,l,m,n,ng,r,w,y
Voiceless consonants are :- p,t,k,f,s,th,sh,ch,h

II. SPEECH RECOGNITION PROCESS

Speech Recognition is truly a ponderous and tiresome process. It consists of 5 steps:-

1. Speech
2. Speech Pre-Processing
3. Feature Extraction
4. Speech Classification
5. Recognition

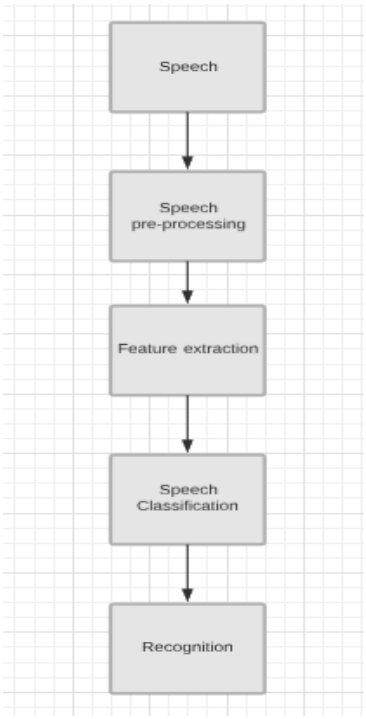

Figure-1 Speech recognition process

Speech

Speech is defined as the ability to express one's thoughts and feelings by articulate sounds. Initially the speech of a person is received in the form of a waveform. Also there are numerous tools and software's available which record the speech delivered by the humans. The phonic environment and the equipment device used have a significant impact on the speech generated. There is a possibility of having background or room reverberation blended with the speech but this is completely undesirable.

Speech Pre-Processing

The solution of the problem described above is the "Speech Pre-Processing". It plays an influential role in cancelling out the trivial sources of variation. The speech pre-processing typically includes reverberation cancelling, echo cancellation, windowing, noise filtering and smoothing all of which conclusively improves the accuracy of speech recognition.

Feature Extraction

Each and every person has different speech and different intonation. This is due to the different characteristics ingrained in their utterance. There should be a probability of identifying speech from the theoretical waveform, at least theoretically. As a result of an enormous variation in speech there is an imminent need to reduce the variations by performing some feature extraction. The ensuing section depicts some of the feature extraction technologies which are extremely used nowadays.

LPC (Linear Predictive Coding):- It is an extremely useful speech analysis technique for encoding quality speech at low bit rate and is one of the most powerful method. The key idea behind this method is that a specific speech sample at current time can be approximated as a linear combination of past speech samples. In this method the digital signal is compressed for competent storage and transmission. The principle behind the use of LPC is to reduce the sum of squared distance between the original speech and estimated speech over a finite duration. It can be further used to provide unique set of predictor coefficients. Gain (G) is also a crucial parameter.

MFCC (Mel Frequency Cepstarl Coefficients):- This is the standard method feature extraction. It is preliminary based on the frequency domain which is based Mel scale based on human ear scale. They are more accurate than time domain features ever since they fall into the category of frequency domain features. The most conspicuous impediment is its sensitivity to noise as it is highly dependent on the spectral form. Techniques utilizing the periodicity of speech signals could be used to overcome this drawback although speech also encompasses aperiodic content.

Speech Classification

These systems are used to extract the hidden information from the input processing signals and comprises of convoluted mathematical functions. This section describes some commonly used speech classification techniques in brief.

HMM (Hidden Markov Model):- This is the most strongly used method in order to recognize pattern in the speech. It is safer and possesses a secure mathematical foundation as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

compared to the template based and knowledge based approach. In this method, the system being modelled is assumed to be a Markov process having hidden states. The speech is distributed into smaller resounding entities each of which represent a state. In simpler Markov Model, the states are clearly visible to the user and thus the state transition probabilities are only the parameters. On the other hand, in hidden Markov Model, the state is not directly visible, but the output, which is dependent on the state, is evident. HMM are specifically known for their application in reinforcement learning and pattern recognition such as speech, handwriting and bioinformatics.

DTW (Dynamic Time Warping):-In time series analysis, DTW is a kind of algorithm which measures the similarity or affinity between two temporal sequences that vary in speed or time. It correlates the speech words with reference words. This method changes the time dimension of the undiscovered words unless and until they are matched with the reference word. A well-known application of DTW is the automatic speech recognition, in order to cope up with different speaking speeds. Various other applications are online signature recognition and speaker recognition.

VQ (Vector Quantization):- This method is preliminary based on block coding principle. This technique allows the modelling of probability density functions by the circulation of prototype vectors. It was formerly used for data compression. It performs the mapping of the vector from a vast vector space to a finite number of regions in that space. Every region is known as cluster and can be depicted by its centre called code word. Vector Quantization is used in lossy data compression, lossy data correction, and clustering and pattern recognition.

Recognition

After the above four phases speech recording, speech pre-processing, feature extraction and speech classification the final step that is remaining is the speech recognition. Once all the above mentioned steps are completed successfully then the recognition of speech can be done by three approaches.

1. Acoustic phonetic approach[6]
2. Pattern recognition approach[7]
3. Artificial intelligence approach

This paper is mainly concern with Artificial intelligence approach for speech recognition. This is a combination of acoustic phonetic and pattern recognition approach. In this approach system created by neural networks are used to classify and recognize the sound.. Neural networks are very powerful for recognition of speech. There are various networks for this process. RNN, LSTM, Deep Neural network and hybrid HMM-LSTM are used for speech recognition.

III. NEURAL NETWORKS

Traditionally neural networks referred to as neurons or circuit. At present the term neural networks refers to as Artificial Neural Network, consisting of artificial neurons or nodes. It is a network of elementary elements known as artificial neurons, which receives an input, changes the state according to that input and generates an output. An interconnected group of natural or artificial neurons uses a mathematical model for information processing based on connectionist approach to communication. Neutral Networks can be treated as simple mathematical models defining a function *f*: X Y or a distribution over X or both X and Y [8], but many a times models are intimately connected with a particular learning rule. The term "artificial neural network" refers to inter-connections in among the neurons in the different layers of each system. Mathematically a neuron's network function *f(x)* is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. A most commonly used type of composition is the nonlinear weighted sum, where $f(x) = K (\sum_i w_i g_i(x))$, where K is a predefined function. Referring the collection of functions g_i as a vector g simply would be more advantageous. The first view being the functional view; the given x is then changed to a 3-D vector h, which is then finally transformed into *f*. This view is frequently encountered in the optimization context. Probabilistic view is the second view and the arbitrary variable $G = g(H)$, depends on H = h(X), which in turn depends upon the X(random variable). In the perspective of graphical models this view most commonly comes across. Networks described such as the above one are often called feed forward as its graph is a DAG (Directed Acyclic Graph). Networks which have cycles in it is called recurrent neural networks.

Neural Network Models:

Language modelling and acoustic modelling are both vital aspects of modern statistically-based speech recognition systems. The ensuring section illustrates various methods for Speech Recognition.

Deep Neural Network (Hidden Markov Models HMM is a generative model in which observable acoustic features are assumed to be generated from a hidden Markov process that transitions between states S = {s1, s2 …., sk}.

HMM was the most widely used technique for Large Vocabulary Speech Recognition (LSVR) for at least two decades. The decisive parameters in the HMM are the initial state probability distribution f = {p(q0=si)}, where qt is the state at time t; the transition probabilities aij= p(qt=sj |qt-1=si) ; and a model to estimate the observation probabilities p(xt |si). The conventional HMM used in the process of Automatic Speech Recognition had their observation probability modelled by using Gaussian Mixture Model (GMM). Even the GMM had a vast number of advantages, the issue was that, they were statistically inept for modelling data that lie on or near the non-linear diverse in the space. For instance, modelling of the points residing very close to surface of the sphere hardly requires any parameter using a suitable model class, but it requires large number of diagonal Gaussians or a fairly huge number of full-covariance Gaussians. Because of this other types of models may work better than GMM the exploitation of information embedded in a large window of frames is done well. On the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

other hand ANN (Artificial Neural Network) can handle the data residing on or near a non-linear model more effectively and learn much better models of data. Since the past few years, outstanding advances have been made both in machine learning algorithms and computer hardware which ultimately has led to more efficient methods of learning having many layers and a large layer of output. The output layer must hold a great number of HMM states that arise on each phone being modelled by a different number of triphone. By employing various new learning methods, vast number of research group have shown that Deep Neural Network has better performance than GMMs at acoustic modelling for recognition of speech that includes massive vocabularies and behemoth dataset. An Artificial Neural Network has more than one layer of hidden units between its inputs and outputs whereas Deep Neural Network is a feed-forward network. Each hidden unit in DNN, j, uses the logistic function to map all of its input from the layer below, xj , to the scalar state, yj that it sends to the above lying layer.

$$Y_j = \log(x_j) = 1/(1+e^{-xj})$$

$$X_j = b_j + \sum_i y_i w_{ij}$$

Where b_j is the bias of unit j, i is an index over units in the layer below, and $w_{ij}$is the weight on a connection to unit j from unit i in the layer below. For multiclass classification, output unit j converts its total input, x_j , into a class probability, p_j, by using the softmax non-linearity

$$p_j = \exp(x_j) / \sum_k \exp(x_k)$$

where k is an index over all classes.
Deep Neural Network DNN‘s can be trained by back-propagating derivatives of a cost function that measures the divergence between the target outputs and the actual outputs produced. When working with softmax function, the natural cost C can be calculated as

$$C = -\sum_j d_j \log p_j$$

where p is the output of the softmax d represents the target probabilities.
DNNs having many hidden layers are difficult to optimize. The optimal way is not to choose the gradient descent from a arbitrary starting point near the origin in order to find a good set of weights and unless the initial scales of the weight are carefully chosen, in different layer there will be varying magnitudes of the back propagated gradients. In addition to these issues, generalization of test data may be done poorly by DNN. Layers of DNN are quite flexible and each with a large number of parameters. As a result of this it makes DNN capable of modelling very intricate and non-linear relationships between outputs and inputs. There is a possibility of having severe over fitting and this issue can be eliminated by early stopping or weight penalties but it is only possible by reducing considerable amount of power. Large amount of dataset can reduce the over fitting and preserving the modelling power simultaneously but it increases the computational cost. Therefore there is a prominent need of using the information in the process of training set to build various layers of non-linear feature detectors.

Recurrent Neural Network: Recurrent Neural Network is a kind of Artificial Neural Network, which is represented in the form of directed cycle where each and every node is connected to the other nodes. Two units become dynamic as soon as the communication takes place in between them. Since RNN uses internal memory just like the feed-forward networks to process the sequence of arbitrary input, it makes them ideal choice for speech recognition. The key feature of RNN is that the activations flow round in a loop as the network contains at least one feed-back connection thus allowing the networks to do temporal processing and learn sequences.

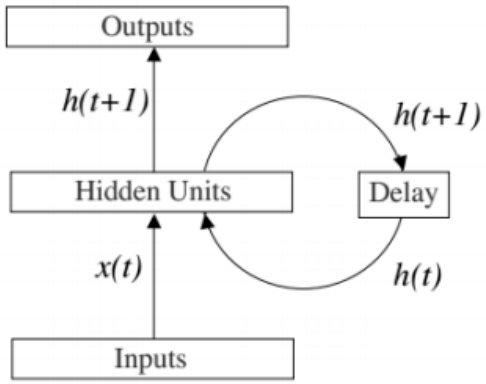

Figure 1 Recurrent Neural Network[9]

The elementary structure is a feed-forward DNN having an input and output layer and certain number of hidden layers with full recurrent connections. Even of basic architectures, learning in RNN can be achieved. In Fig. 2 representation of simplest form of fully recurrent neural network that is an MLP (Multi-Layer Preceptron) having previous set of concealed unit activations (h (t)), feeding back into the network along with the input (h (t+1)). The time scale t refers to the operation of real neurons and as far as artificial systems are concerned any relevant time step size for the given problem can be used. In order to hold the unit until they are processed at the next step, a delay unit is introduced purposely.

LSTM (Long Short Term Memory): LSTM is a RNN architecture that in addition to regular network unit, contains LSTM blocks. LSTM blocks are generally referred to as “smart” network unit that possess the capability of remembering a value having arbitrary length of time. It contains “gates” whose function is to let us know when the input is significant to remember, when to forget and when it should output the value. In LSTM architecture, a set of recurrently connected subnets known as “memory blocks” resides in the recurrent hidden layer. In order to control the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

flow of information, each memory block contains one or more self-connected memory cells and three multiplicative gates. The flow of information in each cell of LSTM is secured by the learned input and output gates. The forget gate is added for the purpose of resetting the cells. A conventional LSTM can be defined as follows:

Given an input sequence x = (x1, x2 , . . . , x_T), a conventional RNN computes the hidden vector sequence h = (h1, h2, . . . , h_T) and output vector sequence y = (y1, y2, . . . , y_T) from t = 1 to T as follows:

$$h_t = H (W_{xh}x_t + W_{hh}h_t{-}1 + b_h)$$
$$t = W_{hy}h_t + b_y$$

Where, the W denotes weight matrices, the b denotes bias vectors and H(·) is the recurrent hidden layer function. The following figure illustrates the architecture of LSTM:-

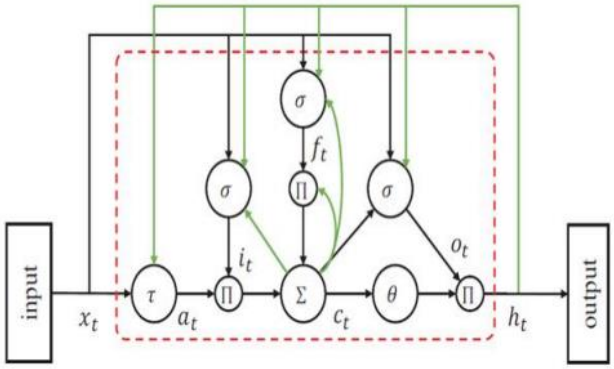

Figure 2 Architecture of LSTM network having a single memory block[10]

Limitations of LSTM:
Because LSTM has the capacity to store only one of its inputs, error will not be reduced which in turn won't moderate the error by solving the sub goal. Full gradient can be used as a solution to this problem.

a. However full gradient has also some limitations: 1) It is very complex 2) Error flow is visible only when truncated LSTM is used.
b. The weight factor increases to 32 as a single hidden unit is replaced by 3 units. Therefore, a single memory has the requirement of two additional cell blocks.
c. The problems associated with the feed forward nets also persists in the LSTM because, it behaves like fed forward neural network trained by back propagation neural network to see the complete input string.
d. Practical problem in all the gradient based approaches is "Counting the time steps" problem.

Advantages of LSTM:

1. The constant error back propagation in LSTM allows it to bridge extremely long time lags in case of similar problems discussed above. In case of long time lags, LSTM can handle noise, continuous values and distributed representations. However in case of hidden Markov Model LSTM doesn't need to have priori choice of a finite number of states, in that it can deal with infinite state numbers.
2. With respect to the problems discussed in this paper, LSTM can generalize well irrespective of the irrelevant and widely spread inputs in the input sequence. It has the capability of quickly learning about how to differentiate in between two or more widely spread occurrences of a particular element in the input sequence without getting dependent on short time lag training exemplars.
3. It doesn't require parameter tuning at all. It works pretty well with wide range of parameters such as input and output bias gate, learning rate.
4. The LSTM's algorithm complexity per weight and time step is O(1). This is considered to be extremely advantageous and outruns the other approaches such as RTRL.

IV. LITERATURE SURVEY

In early 1920s, speech recognition came into existence. The first machine to recognize speech is named Radio Rex. (Manufactured in 1920). After that, research is begun in Bell Labs in 1936[12].In 1939, Bell labs demonstrated a speech synthesis machine at the world Fair in New York. In 1952 three Bell Labs researchers, Stephen.Balashek , R. Biddulph, and K. H. Davis built a system called "Audrey" an automatic digit recognizer for single-speaker digit recognition. Their system worked by locating the formates in the power spectrum of each utterance. [13] The 1950s era technology was limited to single-sAupeaker systems with vocabularies of around ten word. Michael Price, James Glass, Anantha P. Chandrakasan (2015) [14] described an IC that provides a local speech recognition capability for a variety of electronic devices. With 5,000 word recognition tasks in real-time with 13.0% word error rate, 6.0 mW core power consumption, chip provides search efficiency of approximately 16 nJ per hypothesis. The vowel recognizer of Forgie and Forgie constructed at MIT Lincoln laboratories in 1959 in which 10 vowels embedded in a /b/-vowel/t/ format were recognized in a speaker independent manner.[15]

In late 1960s Raj Reddy was first to take on continuous speech recognition at Stanford university. Early systems are based on pause between each word. This is first continuous speech recognition approach. In the meanwhile, Soviet Union has used DTW algorithm to build a 200-word vocabulary speech recognition machine.

In 1970s, Velichko and Zagoruyko have studied discrete utterance recognition or isolated word in Russia.[17] Same in United states by Itakura and in Japan by Cakoe and Chiba[18].In 1971, striving speech understanding project was funded by Defence Advanced Research Projects Agencies (DARPA). IEEE acoustic, Speech, and Signal Processing group held a conference in Newton, Massachusetts in 1972.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

In mid 1980s, IBM created a voice-activated typewriter called Tangora, which could handle a 20,000-word vocabulary under the lead of Fred Jelinek. [16] In this era, neural networks are emerged as an attractive model for Automatic speech recognition. Speech research in the 1980s was shifted to statistical modelling rather than template based approach. This is mainly known as Hidden Markov model approach. Applying neural networks for speech recognition was reintroduced in late 1980s. Neural networks first introduced in 1950 but for some practical problems they were not that much efficient.
In the 1990s, the Bayes classification is transformed into the optimization problems, which also reduces the empirical errors. A key issue in the design and implementation of speech recognition system is how to choose proper method in the speech material used to train the recognition algorithm. Training can be supervised learning in which class is labelled in the training data and algorithm will predict the label in the unlabelled data. Stephen V. Kosonocky [11] had researched about how neural network can be used for speech recognition in 1995.

In 2005, Giuseppe Riccardi [19] developed Variational Bayesian (VB) to solve the problem of adaptive learning in speech recognition and proposed learning algorithm for ASR.

In 2011, Dr.R.L.K.Venkateswarlu, Dr. R. VasanthaKumari, G.VaniJayaSri[20]utilizes Recurrent Neural Network, one of the Neural Network techniques to observe the difference of alphabet from E- set to AH - set. In their research 6 speakers (a mixture of male and female) are trained in quiet environment. The English language offers a ndreumber of challenges for speech recognition. They used multilayer back propagation algorithm for training the neural network. Six speakers were trained using the multilayer perceptron with 108 input nodes, 2 hidden layers and 4 output nodes each for one word, with the noncurvear activation function sigmoid. The learning rate was taken as 0.1, momentum rate was taken as 0.5.Weights were initialized to random values between +0.1 and -0.1 and accepted error was chosen as 0.009. They have compared the performance of neural network with Multi-Layer Perceptron and concluded that RNN is better than Multi-Layer Perceptron. For A-set the maximum performances of speakers 1-6 were 93%, 99%, 96%, 93%, 92% & 94%. For E-set it was 99%, 100%, 98%, 97%, 97% & 95%, and For EH-set 100%, 95%, 98%, 95%, 98% & 98% and lastly for AH-set 95%, 98%, 96%, 96%, 95% & 95% respectively. Results shows that RNN is very powerful in classifying the speech signals.

Song, W., &Cai, J. (2015) [21] has developed end to end speech recognition using hybrid CNN and RNN. They have used hybrid convolutional neural networks for phoneme recognition and HMM for word decoding. Their best model achieved an accuracy of 26.3% frame error on the standard core test dataset for TIMIT. Their main motto is to replace GMM-HMM based automatic speech recognition with the deep neural networks. 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.

V. CONCLUSION

Speech is primary and essential way for communication between humans. This survey is about neural networks are modern way for recognizing the speech. In contrast to traditional approach it does not requires any statistics.A speech recognition system should include the four stages: Analysis, Feature Extraction, Modeling and matching techniques as described in the paper. In this paper, the fundamentals of speech recognition are discussed and its recent progress is investigated. Various neural networks model such as deep neural networks, and RNN and LSTM are discussed in the paper. Automatic speech recognition using neural networks is emerging field now a day. Text to speech and speech to text are two application that are useful for disabled people. Paper mainly focuses on speech recognition of one language, which is English.

VI. REFERENCES

[1] Yu D., Deng L. (2015) Deep Neural Network-Hidden Markov Model Hybrid Systems. In: Automatic Speech Recognition. Signals and Communication Technology. pp 99-116, Springer, London,[Available Online]: Automatic Speech Recognition Using HMM and deep neural network
[2] Zhang, XL, Luo, ZG. & Li, M. J, Journal Of Computer Science and Technology, Springer ,November 2014, Volume 29, Issue 6, pp 1072–1082.https://doi.org/10.1007/s11390-014-1491-0
[3] Zou J., Han Y., So SS. (2008) Overview of Artificial Neural Networks. In: Livingstone D.J. (Eds) Artificial Neural Networks. Methods in Molecular Biology™, vol 458. Humana Press, [Available Online]: https://link.springer.com/protocol/10.1007/978-1-60327-101-1_2
[4] Recurrent deep neural networks for robust speech recognition. / Weng, Chao; Yu, Dong; Watanabe, Shinji; Juang, Biing Hwang Fred. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014. Institute of Electrical and Electronics Engineers Inc., 2014. p. 5532-5536 6854661.
[5] Miao Y., Metze F. (2017) End-to-End Architectures for Speech Recognition. In: Watanabe S., Delcroix M., Metze F., Hershey J. (eds) New Era for Robust Speech Recognition. Springer, Cham.
[6] Schwartz R.M. et al. (1988) Acoustic-Phonetic Decoding of Speech. In: Niemann H., Lang M., Sagerer G. (eds) Recent Advances in Speech Understanding and Dialog Systems. NATO ASI Series (Series F: Computer and Systems Sciences), vol 46. Springer, Berlin, Heidelberg
[7] Rabiner L.R. (1992) Speech Recognition Based on Pattern Recognition Approaches. In: Ince A.N. (eds) Digital Speech Processing. The Kluwer International Series in Engineering and Computer Science (VLSI, Computer Architecture and Digital Signal Processing), vol 155. Springer, Boston, MA
[8] Wikipedia contributors. (2018, September 22). Neural network. In *Wikipedia, The Free Encyclopaedia*. Retrieved 15:37, November 4 , 2017, from https://en.wikipedia.org/w/index.php?title=Neural_network&oldid=860697996

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 10, October-2018

[9] International Journal on Recent and Innovation Trends in Computing and Communication Volume: 4
[10] G Gnaneswari, S R VijayaRaghava, A K Thushar, Dr.S.Balaji, Recent Trends in Application of Neural Networks to Speech Recognition, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 4 Issue 1,pp 18 - 25
[11] Kosonocky S.V. (1995) Speech Recognition Using Neural Networks. In: Ramachandran R.P., Mammone R.J. (eds) Modern Methods of Speech Processing. The Springer International Series in Engineering and Computer Science (VLSI, Computer Architecture and Digital Signal Processing), vol 327. Springer, Boston, MA
[12] Wikipedia contributors. (2018, October 2). Bell Labs. In *Wikipedia, The Free Encyclopedia*. Retrieved 13:35, October 5, 2018, from https://en.wikipedia.org/w/index.php?title=Bell_Labs&oldid=862196483
[13] Juang, B. H.; Rabiner, Lawrence R. "Automatic speech recognition–a brief history of the technology development" (PDF): 6. Archived (PDF) from the original on 17 August 2014. Retrieved 17 January 2015
[14] Michael Price, James Glass, Anantha P. Chandrakasan "A 6 mW, 5,000-Word Real-Time Speech Recognizer Using WFST Models", IEEE Journal Of Solid-State Circuits, Vol. 50, No. 1, PP 102-112, January 2015
[15] W. Forgie, James & D. Forgie, Carma. (1959). Results Obtained from a Vowel Recognition Computer Program. The Journal of the Acoustical Society of America. 31. 844-844. 10.1121/1.1936151.
[16] "Pioneer speech recognition", [available online]: http://www03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/
[17] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," Int. J. Man-Machine Studies, 2:223, June 1970
[18] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, speech, signal proc., ASSP 26(1), pp. 43-49, Febreary 1978.*
[19] Giuseppe Riccardi, DilekHakkani-Tür, "Active learning: theory and applications to automatic speech recognition", *IEEE Transaction on Speech and Audio Processing*, vol. 13, no. 4, pp. 504-511, 2005.
[20] Dr.R.L.K.Venkateswarlu, Dr. R. Vasantha Kumari, G.VaniJayaSri, *International Journal of Scientific & Engineering Research Volume 2, Issue 6, June-2011*
[21] Song, W., &Cai, J. (2015) End-to-End Deep Neural Network for Automatic Speech Recognition

Sanket A. Shah was born in Anand City, Gujarat, India in 1997. He is currently pursuing his computer engineering at G.H. PATEL INSTITUTE OF ENGINEERING AND TECHNOLOGY, Bakrol, Gujarat, India. He has attended the national conference, RACST held at his institute in 2016.

Hardik J. Dudhrejia was born in Rajkot City, Gujarat, India in 1997. He is currently pursuing his computer engineering at G.H. PATEL INSTITUTE OF ENGINEERING AND TECHNOLOGY, Bakrol, Gujarat, India. He has attended various workshops as well as conferences including a national conference, IMPRESSARIO at his institute in 2016.

.