ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Speech Recognition of Continuous Tamil phoneme using DBN

Banumathi.A.C[1], Dr. E. Chandra[2]

Research Scholar, Department of Computer Science, Mother Teresa Women's University, KodaiKanal, India[1]

Professor & Head, Department of Computer Science, Bharathiar University, Coimbatore, India[2]

ABSTRACT: Speech Recognition has been an interesting field of research for the past few decades. Instruction to the machine through the speech has become inevitable for communication in this current state of Technology. The main issue of the continuous speech recognition algorithm is that the complexity is more to find the best match for the given pattern of speech. Our proposed work brings the efficient application of the ZCR algorithm for the feature Extraction of the input speech, since it has proved more effective in separation of voiced and unvoiced speech. The next part is the Speech Recognition achieved by the Deep Belief Network (DBN). This is a popular architecture in Machine learning, which uses a stack of Restricted Boltzmann Machines to create a powerful model using training data. Deep Belief Networks (DBNs) have been proposed for phone recognition and were found to achieve highly competitive performance. Our paper brings a study and novel approach to the speech recognition of Tamil continuous phoneme using segmentation of word into small acoustic units and speech Recognition using Deep Belief Network.(DBN).

KEYWORDS: Zero cross Rate (ZCR), Deep Belief Network (DBN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Restricted Boltzman Machine (RBM).

I. INTRODUCTION

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them into an understandable form. Speech is recognized by considering its various attributes such as its energy, trajectory of utterance, speaking mode and different speaker utterance. The same phoneme is made completely complex by different speakers, speaking mode and context. The fundamental unit of speech is segment. Segmenting is the process of subdividing the word into smaller units of speech called segmentation.

Speech Recognition of limited simple words can be done by traditional method, Whereas when we have to recognize more words an efficient method is required. Deep Belief Network can be used for such purpose. In Automatic Speech Recognition, translation of spoken words into text is still a challenging task due to the high viability in speech signals. Neural Network are rarely successful for continuous recognition tasks so Deep learning comes as a rescue for continuous recognition of words.

In speech recognition first phase is preprocessing which deals with a speech signal which is an analog signal at recording time, which varies with time. To process the signal by digital means, it is necessary to sample the continuous signal into a discrete valued (digital) signal. The preprocessing stage in speech recognition systems is used in order to increase the efficiency of subsequent feature extraction and classification stages and therefore to improve the overall recognition performance. Commonly the preprocessing includes the sampling step, a windowing and a noise removing step. Sampling is the process at the end of the preprocessing. The compressed and filtered speech frames are forwarded to the feature extraction stage. The next stage is to extract the set of features from speech signal. The extracted features are send to the Deep Belief Network algorithms and compared with the traditional method and finally conclude that Deep Belief Algorithm has a better performance.

For eg. Tamil இனிய காலை வணக்கம் is divided into இனிய / காலை / வணக்கம்

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

II. RELATED WORK

Automatic speech recognition is a process by which a machine identifies speech from the continuous speech [1,2,3]. Speech can be divided into numerous voiced and unvoiced regions. It can be classified as speech signal of voiced, unvoiced Speech Processing, Extraction and Classification.

Md. Mijanur Rahman1, et al, in their research, has developed a system that automatically segments continuous Bangla speech and clusters speech segments into some predefined clusters. The developed system may be used to develop large vocabulary continuous speech recognition system.

G.M.Bhandari1, et al, has presented comparative analysis on feature extraction using segmentation techniques. Different parameters such as audio type, accuracy, recall factor and precision have been evaluated for pure speech, silence etc.

George E. Dahl et.al described a context-dependent DNN-HMM (deep neural network- hidden Markov model) model for LVSR(Large Vocabulary Speech Recognition) that achieves substantially better results than strong, discriminatively trained CD-GMM-HMM(Gaussian mixture model Hidden Markov Model), baselines on a challenging business search dataset.

Deep belief nets (DBNs) are used in image processing and generally are applied on two-dimensional image data but are rarely tested on 3-dimensional data and object recognition [4]. The other applications of DBNs are: hand-written character recognition [3, 4], information retrieval [5, 6], modeling and capturing data motions [7,8], machine transliteration [9], modeling EEG (Electroencephalography) for classification of waveforms and anomaly detection [10], document classification [8], music emotion recognition [12], phone recognition [13] and etc.

Deep Networks have been successfully used for audio analysis, speech recognition, and natural language processing [15]. Non-linear deep networks such as Boltzmann Machines [16] and Neural Networks [17] have the ability to learn a rich feature representation in an unsupervised manner, making them very powerful.

III. OUR PROPOSED SYSTEM ARCHITECTURE

Our proposed speech Recognition comprises of the following steps.

1. Segmentation of speech phoneme
2. Feature Extraction using ZCR (Zero Crossing Rate)
3. Applying deep learning algorithms to speech recognition and compare the speech recognition performance with conventional GMM-HMM based speech recognition method.
4. Deep Belief Network (DBN) has a better performance for speech recognition.

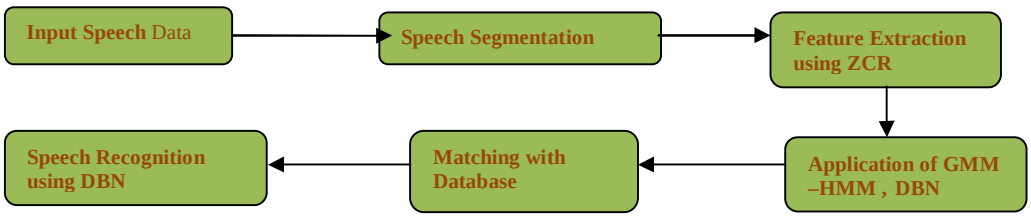

Figure 1: The *Architecture of Speech Recognition using DBN.*

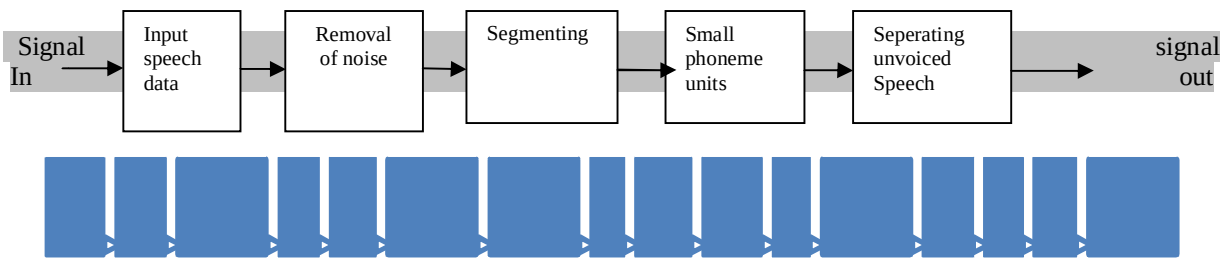

Figure. 2. *Speech Spectrum is divided into a sequence of segments.*

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Segmentation is the process of uniquely identifying meaningful speech units such as phonemes, syllables, words or sub-words and processes them to extract features. Segmentation plays an important role in speech recognition to reduce memory size and minimize the computation complexity for large vocabulary systems. In general, Automatic speech segmentation methods can be classified in many ways, but very common classification is the division of **Blind and Aided segmentation** algorithms.

The main difference between aided and blind method is that in how much the segmentation algorithm uses previously obtained data or external knowledge to process the expected speech. For our research we take the blind algorithm and divide the speech word into discrete units. Due to the lack of external information, the blind segmentation depends entirely on the acoustical features present in the signal. A Blind segmentation is further classified into two kinds of segmentation. One is phonemic segmentation [18], which segments speech into phonemes and other is syllable-like unit segmentation [19], which segments speech into syllables, sub-words or words. For both phonemic and syllabic unit segmentation, most of the approaches are based on the thresholds of the parameters used to segment the speech data. Based on various studies the thresholds can be set for feature segmentation. As it is difficult to identify the proper phoneme from continuous speech the end-point detection technique is used in our research. Segmentation is done by proper start and end point of speech event.

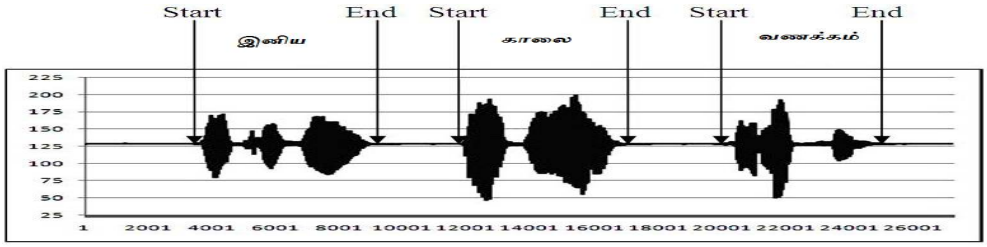

Figure. 3. *The start and end points of words in the speech sentence.*

For evaluation of segmentation systems, some measures are used which are the comparison between detected change points and real change points in the corpus under investigation. The most important measures used, are %FD and %FR which are calculated as shown in Equations 1,2 [20].

$$\%\text{False Detection (FD)} = \frac{\text{False Detection}}{\text{Total Amount of Detection}}$$

$$\%\text{False Rate (FR)} = \frac{\text{missed Detection}}{\text{Total Amount of True Change Points}}$$

False_detections: number of points which are not real change points in the reference corpus; but, are detected by the system as change points. These points are called False Alarm (FA).
Total_amount_of_detections: total number of points detected by the system as change points.
Missed_detections: number of points which are real change points in the reference corpus; But, are not detected by the system as change points. These points are called Missed Detection (MD).
total_amount_of_true_change_points: total number of points correctly detected by the system as change points. To determine the accuracy of segmentation method, F measure is defined as shown in

$$F = \frac{2 * (1-FD) * (1-FR)}{2 - FD - FR}$$

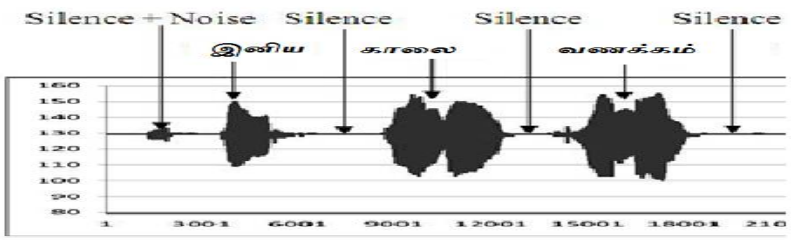

Figure. 4. *Speech sentence that contains noise and silences.*

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Three important issues are considered in the detection of the word boundaries in continuous speech. The first is that the two successive words are Merged omitting the phonemes and the making the sentence long vocalic and difficult to find out the boundaries. Second is the effect due to co-articulation which are much stronger in continuous speech. Third are stresses in articulation, particular words in a sentence and even some particular syllables in a word are emphasized, while others are poorly articulated. Remembering these complexities, an acceptable system may be designed for segmentation using end point detection technique if the articulation of continuous speech is such that there is sufficient pause between speech units.

IV. FEATURE EXTRACTION USING ZERO CROSSING RATE (ZCR)

The main goal of the feature extraction step is to compute a sequence of feature vectors providing a compact representation of the given input signal. Commonly LPC, MFCC, ZCPA, DTW and RASTA are used as feature extraction techniques for speech recognition system.

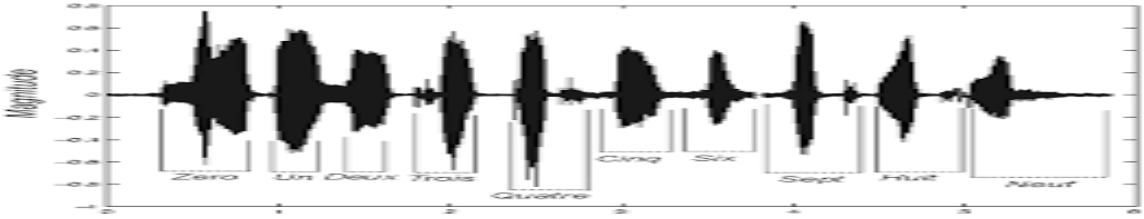

Figure.5. *Feature Extraction of speech using ZCR.*

Speech is received after undergoing several transformations at different levels such as articulatory, linguistic, semantic and syntactic changes. Differences in these transformations appear as differences in the acoustic properties of the speech signal [19]. A well formed recognition system is to be determined for representing the information of the speech signal. This transformation of signal will help us to identify the signal in different domains. The zero crossing Rate (ZCR) is one of the feature extraction method in speech processing to calculate how many times the speech waveform has crossed the zero axis. Below is the general formula for ZCR.

$$ZCR = \frac{1}{2N}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|$$

A zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is simple measure of the frequency content of a signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. Zero-crossing rate is an important parameter for voiced/unvoiced speech data. The energy is concentrated in the signal spectrum and the zero crossing count acts as an indicator of the frequency of the data. The voiced data is produced due to excitation of the vocal tract by the flow of air through the glottis and it shows a low energy count, whereas the unvoiced speech is produced by the constriction of the vocal tract which causes turbulent airflow resulting in noise and it shows a high zero-crossing count. Energy of a speech is another parameter for classifying the voiced/unvoiced parts. The energy will be more in voiced data and less for the unvoiced data.[21].

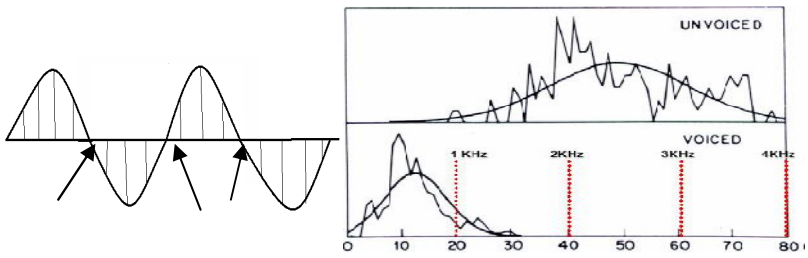

Figure 6. *Zero crossing rate and Energy level of voiced & unvoiced speech.*

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

This is a good measure of the pitch as well as the noisiness of a signal. Zero crossings are calculated by finding the number of times the signal changes sign from one sample to another (or touches the zero axis). [22] The voice obtained from the ZCR process is taken as the input for the DBN.

$Z_n = \sum \text{sgn}[x(m)] - \text{sgn}[x(m-1)] \mid w(n\ m)$

The model for speech production suggests that the energy of voiced speech is concentrated below about 3 kHz because of the spectrum fall introduced by the glottal wave, whereas for unvoiced speech, most of the energy is found at higher frequencies. Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced [23].

V. DEEP BELIEF NETWORK (DBN)

Speech recognition is an established technology, but it fails, in noisy or crowded environments, or when the speaker is far away from the microphone. In order to improve the accuracy of speech recognition, especially in these challenging environments we apply the Deep Belief Network a new way for speech recognition. Deep Learning' means using a neural network with several layers of nodes between input and output and these series of layers between input & output does the feature identification and processing in a series of stages, just as our brains seem to do.

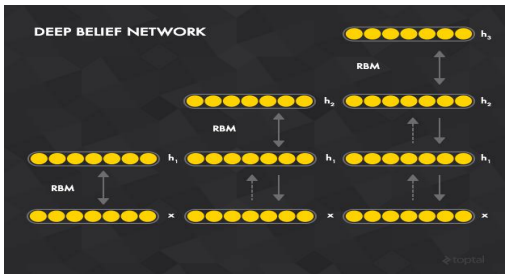

Figure.7(a). *The structure of a DBN*

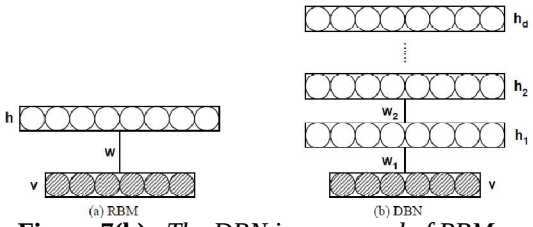

Figure.7(b). *The DBN is composed of RBMs.*

Deep learning [23], sometimes referred as representation learning or unsupervised feature learning, is a new area of machine learning. Deep Belief learning is becoming a important technology for speech recognition [23] and has successfully replaced Gaussian mixtures for speech recognition and feature coding at an increasingly larger scale. Deep Belief Networks (DBNs) are neural networks consisting of a stack of restricted Boltzmann machine (RBM) layers that are trained one at a time, in an unsupervised fashion to induce increasingly abstract representations of the inputs in subsequent layers.

A DBN consists of a stack of RBMs, trained one at a time. Each layer of hidden units learns to represent features that capture higher order correlations in the original input data. In DBNs, subsequent layers usually decrease in size in order to force the network to learn increasingly compact representations of its inputs. The key idea of DBNs is "being deep." Deep acoustic models are important because the low level, local, characteristics are taken care of using the lower layers while higher-order and highly non-linear statistical structure in the input is modeled by the higher layers. Even though the Recurrent Neural Network(RNN) seems to be good for Hand writing recognition, it is not well suited for Speech Recognition. The Deep Belief Network found to produce a better result than the RNN. Deep Belief Learning been successful because it scales well: it can absorb large amounts of data to create highly accurate models. Deep

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Belief Networks are very large and it has many undirected connection between some layers which is helpful to instruct the machine with more number of vocabularies.

VI. RESTRICTED BOLTZMAN MACHINES (RBMs)

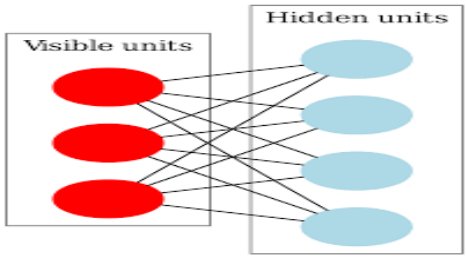

Figure.8 *Restricted Boltzmann Machine Architecture.*

Restricted Boltzmann Machines (RBMS) are non-linear; blocks of neural networks for DBNs are called Restricted Boltzmann Machines (RBM). Each RBM has an input layer (visible layer) and a hidden layer of stochastic binary units. Restricted Boltzmann machines (RBMs) are probabilistic graphical models that can be interpreted as stochastic neural networks. The increase in computational power and the development of faster learning algorithms have made them applicable to relevant machine learning problems. Visible and hidden layers are connected with a weight matrix and no connections exist between units in the same layer. Signal propagation can occur in two ways: recognition, where visible activations propagate to the hidden units; and reconstruction, where hidden activations propagate to visible units. The same weight matrix (transposed) is used for both recognition and reconstruction. By minimizing the difference between the original input and its reconstruction (i.e. reconstruction error) through a procedure called contrastive divergence (CD), the weights can be trained to generate the input patterns presented to the RBM with high probability. The RBM pertaining procedure of a DBN can be used to initialize the weights of a deep neural network, which can then be discriminatively fine-tuned by back-propagating error derivatives. The “recognition” weights of the DBN become the weights of a standard neural network. In cases where the RBM models the joint distribution of visible data and class labels, a hybrid training procedure can be used to fine-tune the generatively trained parameters. Weights are initialized from a normal distribution with zero mean and small standard deviation and it is evaluated by using the gradient descent method. Weight updates are applied after a number of training cycles through the training dataset. RBM is unfolded, and recognition and reconstruction weights are fine tuned with gradient descent method.

VII. APPLYING DBNs FOR SPEECH RECOGNITION

Systems with DBN acoustic models achieve good recognition performance because of three distinct properties of the DBN: it is a neural network which is a very flexible model with many non-linear hidden layers and it is generatively pretrained which acts as a strong, domain-dependent regularize on the weights. To apply DBNs with fixed input and output dimensionality to phone recognition, a context window of n successive frames of feature vectors is used to set the states of the visible units of the lower layer of the DBN which produces a probability distribution over the possible labels of the central frame. To generate speech sequences, a sequence of probability distributions over the possible labels for each frame are fed into a standard decoder.

VIII. PERFORMANCE EVALUATION

For this experiment we have created a database consisting of 60 speech utterance in an elicited context for Tamil (one of the south Indian languages). Our database consists of Two folders “train” and “test” are created for recording the voices of 10 female and 10 male voices of three samples each which is given in the Table. The “test” folder is labeled as samp1 through samp8. This Speech Database of acoustic-phonetic continuous speech corpus dataset in Tamil is used for performance evaluation to build and test the system. The data were normalized to have zero mean and unit variance over the entire corpus. The Word Error Rate for the data corpus is computed using DBN Network.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

IX. RESULTS AND DISCUSSION

The performance of speech recognition system is generally measured in terms of recognition rate and Word error rate. Recognition rate is ratio between the number of words correctly recognized * 100 and total number of words. Word error rate is given as 100 – Recognition rate. To develop real time speaker independent Automatic Speech Recognition (ASR) focus on minimizing the word error rate to zero and recognition accuracy to 100%. For Implementation working platform of MATLAB version R2010 is used. In our experiment ten Tamil words "இனிய காலை வணக்கம்", "வலது புறம் திரும்பு", " இடது புறம் திரும்பு", முன்னாடி செல்", "பின்னாடி செல்". Were taken as input. The Tamil speech sample database is created with 10 speakers utters 5 sentences each with 5 repetitions, total of 250 samples were created. In our experiment ten Tamil words were taken as input. Each speech signal is divided into subsequent 30ms frames and from each frame 12 feature vector MFCC coefficients are extracted. Learning in Deep Models can be achieved effectively and efficiently by a general optimizer without a need for Pre-training. This lead to the diverse range of deep or difficult to –optimize architectures like Recurrent Neural 'network (RNN) and Asymmetric AutoEncoders.[18].

These MFCC features are fed in classifiers for evaluation and corresponding results are presented in Table.1.

S.No	Words in Tamil	Words in English	Different Methods	%WER
1	இனிய காலை வணக்கம்	Iniya kalai Vanakkam	*GMMHMM*	35
			DNN-HMMs	32
			DBN	30
2	வலது புறம் திரும்பு	Valathu puram thirumpu	*GMMHMM*	34
			DNN-HMMs	33
			DBN	31
3	இடது புறம் திரும்பு	Idaathu puram thirumpu	*GMMHMM*	30
			DNN-HMMs	32
			DBN	29
4.	முன்னாடி செல்	Munnnadi sel	*GMMHMM*	33
			DNN-HMMs	31
			DBN	30
5.	பின்னாடி செல்	Pinnadi sel	*GMMHMM*	32
			DNN-HMMs	34
			DBN	30

Table 1. *Word Error Rate of the GMM-HMM, DNN-HMM, DBN*

X. CONCLUSION

In this course project, typical deep learning algorithms, including deep neural networks (DNN), and deep belief networks (DBN) have been learned and understood. Further, a DBN has been implemented for automatic speech recognition. The speech recognition performance evaluations on three speech recognition systems, namely, GMM-HMM, DNN-HMM and DBN, have been performed with corpus dataset in terms of word error rate. The results have shown that the DBN-based speech recognition system beats other two speech recognition systems.

REFERENCES

[1]. **Okko Räsänen, "**Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture", Espoo, November 5, 2007
[2]. T. Jayasankar, R.Thangarajan and J. Arputha Vijaya Selvi, "Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis", International Journal of Computer Applications, Vol. 25, No. 1, pp. 31-36, 2011.
[3] Akila A. Ganesh and Dr.E.Chandra Ravichandran, "Syllable Based Continuous Speech Recognizer with Varied Length Maximum Likelihood Character Segmentation", International Conference on Advances in Computing, Communications and Informatics, pp. 935-940, 2013.
[4].V. Nair and G. Hinton, "3-d object recognition with deep belief nets", To appear in Advances in Neural Information Processing Systems 22, 2009.
[5].G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets", Neural Computation, vol. 18, pp. 1527–1554, 2006.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

[6].G. E. Hinton, "Training products of experts by minimizing contrastive divergence", Neural Computation, vol. 14, pp. 1771–1800, 2002.
[7].G. W. Taylor and G. E. Hinton, "Factored conditional restricted boltzmann machines for modeling motion style", in Proc. ICML, 2009.
[8].Lawrence McAfee, "Document Classification using Deep Belief Nets", CS 224n, 6/4/08 Erik M. Schmidt and Youngmoo E. Kim, "Learning Rhythm and Melody Features With Deep Belief Networks", in IEEE, WASPAA, New Paltz, NY, 2011.
[9]. Abdel-rahman Mohamed, George George Dahl, and Geoffrey Hinton, "Deep Belief Networks for phone recognition", NIPS workshop on Deep Learning for Speech Recognition and Related Applications.
[10].Hinton, G., Osindero, S., and Teh, Y. "A fast learning algorithm for deep belief nets", *Neural Computation*, vol. 18, pp. 1527-1554, 2006.
[I1].Doh-Suk Kim, Soo-Young Lee, and Rhee M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments", *IEEE Trans. Speech Audio* Processing. vol. 7, no. I! pp. *5 5 4 9* , Jan. 1999.
[12]. A.K. Jain, R Bolle, and S.Pankanti, "Biometrics: personal identification in networked society", Springer, 1999.
[13]. Rabiner, L. R., and Schafer, R. W., "Digital Processing of Speech Signals", Englewood Cliffs, NewJersey, Prentice Hall, 512-ISBN-13:9780132136037, 1978.
[14]. Michael M. Goodwin and Jean Laroche,: "Audio Segmentation by feature space clustering using linear discriminant analysis and dynamic programming", 2003.
[15].Awni Hannun_, Carl Case, et.al, – Silicon Valley AI Lab, arXiv:1412.5567v2 [cs.CL] 19 Dec 2014
[16]. Yan Zhang, SUNet ID: yzhang5 Instructor: Andrew Ng. "Speech Recognition Using Deep Learning Algorithms. Foundation and Trends in Signal Processing", Vol. 7, Issue 3-4, June 2014.
[17].*Abdel-rahman Mohamed , Dong Yu*2, *Li Deng*, "Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition", Interspeech. 2010.
[18].James Martens, "Deep learning via Hessian-free optimization", University of Toronto, Ontario, M5S 1A1, Canada.

BIOGRAPHY

Banumathi. A.C is a Research Scholar doing her PhD at Mother Teresa University. Kodaikanal. Her area of research is Speech Recognition system and Neural Network.

Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 16 yrs of experience in teaching including 6 months in the industry. At present she is working as Director, School of Computer Studies in Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore. She has published more than 30 research papers in National, International journals and conferences in India and abroad. She has guided more than 20 M.Phil., Research Scholars. At present 3 M.Phil Scholars and 8 Ph.D Scholars are working under her guidance. She has delivered lectures to various Colleges in Tamil Nadu & Kerala. She is a Board of studies member at various colleges. Her research interest lies in the area of Neural networks, speech recognition systems, fuzzy logic and Machine Learning Techniques. She is a Life member of CSI, Society of Statistics and Computer Applications. Currently Management Committee member of CSI Coimbatore Chapter