Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 8, August 2015, pg.301 – 307

REVIEW ARTICLE

Review on Speech Recognition with Deep Learning Methods

Rubi, Chhavi Rana

M.Tech. Software Engineering (2nd Sem), UIET

M.D. University Rohtak (Haryana), India

rubimalik95@gmail.com

ABSTRACT - The most common mode of communication between humans is speech. As this is the most preferred way, humans would like to use speech to interact with machines also. That is why, speech recognition has gained a lot of popularity. Many approaches for speech recognition exist like Dynamic Time Warping (DTW), Hidden Markov Model (HMM). The main objective of this paper is defined a three stage neural integrated model speech signal enhancement and use the decomposition integrated HMM model for speech feature transformation. For the feature extraction of speech Discrete wavelength transform (DWT) has been used which gives a set of feature vectors of speech waveform. The work has been done on MATLAB and experimental results show that system is able to recognize words at sufficiently high accuracy.

KEYWORDS - Speech recognition, Deep Neural Network (DNN), Discrete Wavelength Transform (DWT), Support Vector Machine (SVM).

1. INTRODUCTION

A. Speech Recognition

Speech recognition is one of the most complex biometric feature recognition method applied on speech signals to identify the speech text. In this work, a feature adaptive neural network model is presented for speech signal recognition. In first stage of this model, the speech improvement is provided using probabilistic estimation. In second stage, the decomposition model is applied along with HMM approach to transform the input signal set to feature set. In final stage, the neural network model is applied for speech signal recognition.

1) Problem Definition

Speech signal is the biometric signal or feature used as the communication medium. The speech can be captured from standard mic or some other microphone devices. Speech is considered in many applications for biometric authentication. It can be used as online or offline recognition model. But because of the environmental and device complexities, the speech signal can have different kind of integrated noise. In this work, noise robust intelligent model is presented for speech recognition. The presented work is defined as three stage model. In first stage of this model, the window adaptive probabilistic mapping is applied for speech signal filtration. This stage is defined to achieve the speech signal enhancement. Once the improved signal is obtained, the speech signal features are extracted as the second stage of recognition model.

In this work, a decomposition integrated HMM model is applied for feature extraction. The window adaptive decomposition modeling is applied for speech feature extraction. The window adaptive sampling with multiple statistical parameters is defined transform the input speech signal to the featured form. In final stage of this work model, the neural network is applied for speech signal recognition. The presented work model is implemented in matlab environment. The result shows that the proposed work model has improved the recognition rate.

2) Significance of work

The significance of work is define here under

- The presented work model is noise robust so that the improved speech signal recognition is applied.
- The feature adaptive neural model has improved the accuracy of work.

B. Research Methodology

In this present work, an improved feature adaptive neural network model is presented for speech signal recognition. This presented work model is applied on noisy speech signal.

1) Signal Enhancement

As the raw signal is extracted, it can have different kind of noise in the signal. This noise can be background noise, signal noise, instrumentation noise etc. So that to improvement the recognition rate, at the earlier stage, the improvement to the speech signal is applied. In this work, a frame adaptive probabilistic model is applied for speech signal enhancement. According to this model, the frame window is setup and the window is moved over the speech signal observed the comparative derivation. This window based comparative derivation identified the signal variation analysis. As some abnormal aspect is identified, the signal averaging and the thresholding is applied to obtained the improved signal. The probabilistic constraints are applied for speech signal improvement. This speech signal improvement is able to identify the high level as well as low level signal noise. The noise identifications at first applied based on window specific analysis applied over the signal. The signal magnitude and the speech signal improvement under the phase variation is applied for the signal adjustment. The Fourier transformation based signal improvement is provided for noise suppression.

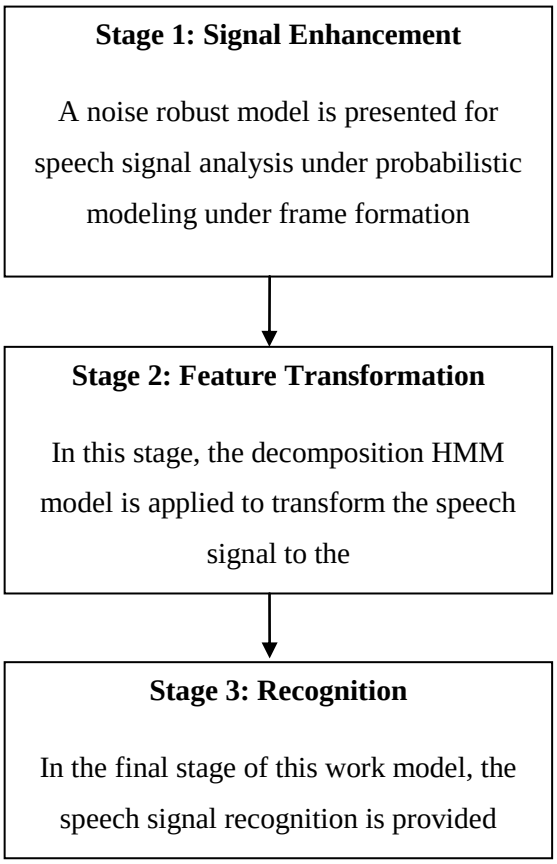

Figure 1: Work Model

2) Signal Feature Extraction

The signal is high and low-pass filtered along the rows and the results of each filter are down- sampled by two. Those two sub-signals correspond to the high and low frequency components along the rows and are each of size N by N/2. Each of those sub-signals is then again high and low-pass filtered, but this time along the column data. The results are again down-sampled by two. The DWT is the decomposition approach used to extract the effective signal features and to maintain the signal effectiveness. In this work, DWT is applied on filtered speech signal to explore the signal features. Here two levels DWT is applied using sym6 function. The function divided the signal in High and low frequency bands. The information preserving signal exploration is performed using sym6.

3) Recognition

SVM is the classifier used in this work to identify the speech signal. This classifier is ability to provide the high accuracy in classification process whiling working with high dimensional data. This classifier also provides the modeling for diverse sources of data. SVM classifier actually comes under kernel based algorithmic approaches. In this approach, the data dependency is also identified as the functional computation to generate the feature space. The advantages of the work are here defined in terms of non linear decision boundaries defined under method specification for linear classification. This classifier is defined along with the specification of kernel function that itself provides the fixed dimensional vector representation with sequence generation and structural representation. This method effectively process on different decision functions to process on data values to identify the data criticality. The data usage based analysis can be obtained under environmental specification. The kernel specification is here defined to control the classification process. There is different classification method along with the specification of relative kernel parameters. The simplest form of SVM is called linear classifier that can be applied on balanced dataset.

C. Neural Network

In general there are so many factors that can affect artificial Neural Network forecasting ability.
These factors are:

1 Number of input neurons
2 Training period
3 Learning rate
4 Momentum

1) Number of input neurons

This layer has as many neurons as there are input categories. The number of input variables is important parameter that affects Neural Network forecasting capability. The number of input neurons is one of the easiest parameter to select once the independent variables have been preprocessing because each independent variable is represented by its own input neurons. If an inadequate number of neurons are used, the network will be unable to model complex data, and the resulting fit will be poor.
If too many neurons are used, the training time may become excessively long, and, worse, the network may *over fit* the data. When over fitting occurs, the network will begin to model random noise in the data. The result is that the model fits the training data extremely well, but it generalizes poorly to new, unseen data. Validation must be used to test for this.

Too few neurons prevent the network from correctly mapping inputs into outputs. It also take very long training time because few neurons are used to train the network. You specify the minimum and maximum number of neurons you want it to test, and it will build models using varying numbers of neurons and measure the quality using either cross validation or hold-out data not used for training. This is a highly effective method for finding the optimal number of neurons, but it is computationally expensive, because many models must be built, and each model has to be validated.

2) Training period: When training is stopped

The question is how many lag periods should be included in predicting the future. Ideally, we desire a small number of lag periods that can unveil the unique features embedded in the data. The inclusion of excessive periods will adversely affect the training time of the network, and the algorithm will likely be trapped in local optimal solutions. On the other hand, if the lag is smaller than required, forecasting accuracy will be jeopardized because the search is restricted to a subspace. Too few or too many lag periods affect either the learning or prediction capability of the network. It is desirable to reduce the number of input nodes to an absolute minimum of essential nodes.

It is to be expected that given a suitable architecture, the error, would reduced from epoch to epoch. The basic idea behind this is to first partition the data set in two subset: training set and test set.

To tune the weights of the network training is divided in to learning and validation. Validation might comprise 20-30 % of the pattern in the training. Learning is the set of pattern that is used to actually train the network using back propagation. At the end of each epoch, the network performance is evaluated using validation.

3) Learning rate:

The learning rate has great impact on the training of the network. The selection of a learning rate is of critical importance in finding the true global minimum of the error distance.

The learning rate parameter controls the step size in each iteration. A very low learning rate results slow learning and very large training time. If the learning rate are small enough, then convergence to the local minimum. Too large a learning rate will proceed much faster, but may simply produce oscillations between relatively poor solutions. Rates that are larger but less than twice the optimal learning rate converges error minimum to but only after much oscillation. Learning rate that are larger than twice the optimal value will diverse from the solution. A very high learning will often result in erratic learning or oscillations.

4) Momentum:

Momentum in the back propagation algorithm can be helpful in speeding the convergence and avoiding local minima. The purpose of the momentum is to accelerate the convergence of error. It is denoted by **α**. It provides supplementing the current the current weight adjustment with fraction of the most recent weight adjustment. Typically **α** is chosen between 0.1 to 0.8. Momentum typically helps to speed up convergence, and to achieve an efficient and more reliable learning profile. Momentum term technique can be recommended for problem with convergence that occur too slowly or for cases when learning is difficult to achieve.
When very low momentum factor is used then there exist very slow training and training time is much more as compared the optimum one. It causes the local minima.
A momentum rate set at the maximum of 1.0 may result to create a risk of overshooting the minimum, which can cause the system to become unstable which is highly unstable and thus may not achieve even local minima, or the network may take an inordinate amount of training time.

2. LITERATURE REVIEW

M.A.Anusuya and S.K.Katti [1, 3] presents a brief survey on Automatic Speech Recognition and discusses the major themes and advances made in the past 60 years of research, so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech communication. The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, data base and performance evaluation. The objective of this review paper is to summarize and compare some of the well known methods used in various stages of speech recognition system and identify research topic and applications which are at the forefront of this exciting and challenging field.

Santosh K.Gaikwad , Bharti W.Gawali and Pravin Yannawar [2] The Speech is most prominent & primary mode of Communication among of human being. The communication among human computer interaction is called human computer

interface. Speech has potential of being important mode of interaction with computer .This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper helps in choosing the technique along with their relative merits & demerits. A comparative study of different technique is done as per stages. This paper is concludes with the decision on feature direction for developing technique in human computer interface system using Marathi Language.

Shanthi Therese, Chelpa Lingam [4] Says that speech has evolved as a primary form of communication between humans. The advent of digital technology, gave us highly versatile digital processors with high speed, low cost and high power which enable researchers to transform the analog speech signals in to digital speech signals that can be scientifically studied. Achieving higher recognition accuracy, low word error rate and addressing the issues of sources of variability are the major considerations for developing an efficient Automatic Speech Recognition system. In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. In this paper, an effort has been made to highlight the progress made so far in the feature extraction phase of speech recognition system and an overview of technological perspective of an Automatic Speech Recognition system are discussed.

Sanjib Das [5] presents a brief survey on speech is the primary and the most convenient means of communication between people. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper helps in choosing the technique along with their relative merits and demerits. A comparative study of different technique is done as per stages. This paper concludes with the decision on feature direction for developing technique in human computer interface system in different mother tongue and it also discusses the various techniques used in each step of a speech recognition process and attempts to analyze an approach for designing an efficient system for speech recognition. The objective of this review paper is to summarize and compare different speech recognition systems and identify research topics and applications which are at the forefront of this exciting and challenging field.

Nidhi Desai1, Prof. Kinnal Dhameliya, Prof. Vijayendra Desai[6,7] survey presents speech is the most natural form of human communication and speech processing has been one of the most inspiring expanses of signal processing. Speech recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Automatic Speech Recognition (ASR) system takes a human speech utterance as an input and requites a string of words as output. This paper introduce a brief survey on Automatic Speech Recognition and discuss the major subjects and improvements made in the past 60 years of research, that provides technological outlook and a respect of the fundamental achievement that has been accomplished in this important area of speech communication. Definition of various types of speech classes, feature extraction techniques, speech classifiers and performance evaluation are issues that require attention in designing of speech recognition system. The objective of this review paper is to summarize some of the well known methods used in several stage of speech recognition system.

Guillaume Gravier, Ashutosh Garg [8, 11] survey presents Visual speech information from the speaker's mouth region has been successfully shown to improve noise robustness of automatic speech recognizers, thus promising to extend their usability into the human computer interface. In this paper, we review the main components of audio-visual automatic speech recognition and present novel contributions in two main areas: First, the visual front end design, based on a cascade of linear image transforms of an appropriate video region-of-interest, and subsequently, audio-visual speech integration. On the later topic, we discuss new work on feature and decision fusion combination, the modeling of audio-visual speech asynchrony, and incorporating modality reliability estimates to the bimodal recognition process. We also briefly touch upon the issue of audiovisual speaker adaptation. We apply our algorithms to three multi-subject bimodal databases, ranging from small- to large vocabulary recognition tasks, recorded at both visually controlled and challenging environments. Our experiments demonstrate that the visual modality improves automatic speech recognition over all conditions and data considered, however less so for visually challenging environments and large vocabulary tasks.

Li Deng and John C. Platt [9] survey presents that deep learning systems have dramatically improved the accuracy of speech recognition, and various deep architectures and learning methods have been developed with distinct strengths and weaknesses in recent years. How can ensemble learning be applied to these varying deep learning systems to achieve greater recognition accuracy is the focus of this paper. We develop and report linear and log-linear stacking methods for ensemble learning with applications specifically to speech-class posterior probabilities as computed by the convolutional, recurrent, and fully-connected deep neural networks. Convex optimization problems are formulated and solved, with analytical formulas derived for training the ensemble-learning parameters. Experimental results demonstrate a significant increase in phone recognition accuracy after stacking the deep learning subsystems that use different mechanisms for computing high-level, hierarchical features from the raw acoustic signals in speech.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Mic [10] survey describe that deep learning is becoming a mainstream technology for speech recognition at industrial scale. In this paper, we provide an overview of the work by Microsoft speech researchers since 2009 in this area, focusing on more recent advances which shed light to the basic capabilities and limitations of the current deep learning technology. We organize this overview along the feature-domain and model-domain dimensions according to the conventional approach to analyzing speech systems. Selected experimental results, including speech recognition and related applications such as spoken dialogue and language modeling, are presented to demonstrate and analyze the strengths and weaknesses of the techniques described in the paper. Potential improvement of these techniques and future research directions are discussed.

3. IMPLEMENTATION & RESULT

The Experiment has been conducted using MATLAB with Neural Network toolbox. In this study we took 40 recorded audio files. These samples were recorded in MATLAB. The sampling frequency for all recording was 44100 Hz. Of these files all 40 samples were used for training sets and 10 out of them were used for testing sets. Then we calculated coefficients which were passed as input to the NN.

Finally, Neural Network toolbox of MATLAB was used to create, train and simulate the networks and mean square error was used to evaluate its performance. As already mentioned NN consists of neurons. These neurons can use any differentiable transfer function *f* to generate their output. The transfer function used in our case for hidden layers is tan-sigmoid, and for the output layer is linear.

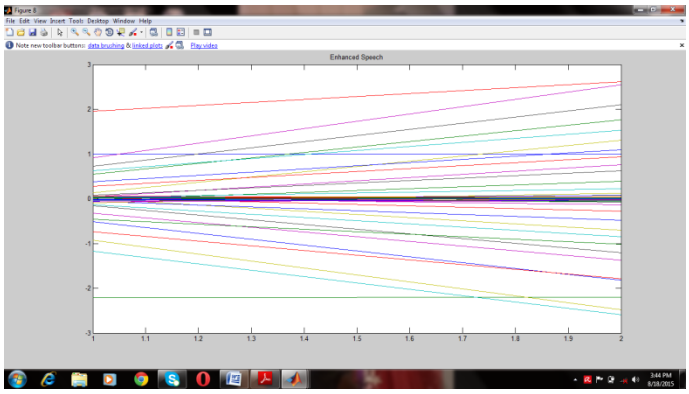

Figure 3.1: Feature Extraction

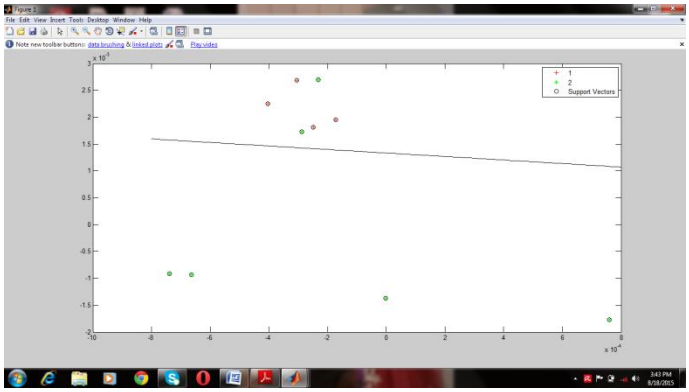

Figure 3.2: Speech Classification

4. CONCLUSION

In this paper, we have used MFCC and Neural Network for speech recognition. The whole experiment has been implemented on MATLAB R2008b using Neural Network toolbox and it successfully recognizes speech. The simulation shows high accuracy in result. Further, improvement can be made in this method which will yield more accurate and precise result.

5. BIBLIOGRAPHY

1. Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank SeideMichael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA 2009

2. M.A.Anusuya and S.K.Katti ,Department of Computer Science and Engineering,Sri Jayachamarajendra College of Engineering, Mysore, India, (IJCSIS) International Journal of Computer Science and Inform4 ation Security,2009.

3. Shanthi Therese ,Chelpa Lingam, International Journal of Scientific Engineering and Technology , June 2013.,Review of Feature Extraction Techniques in Automatic Speech Recognition.

4. Speech Recognition Technique: A Review Sanjib Das Department of Computer Science, Sukanta Mahavidyalaya, (University of North Bengal), India, International Journal of Engineering Research and Applications (IJERA) May-Jun 2012.

5. Nidhi Desai1, Prof.Kinnal Dhameliya2, Prof.Vijayendra Desai3, International Journal of Emerging Technology and Advanced Engineering , December 2013, Feature Extraction and Classification Techniques for Speech Recognition: A Review.

6. Li Deng and John C. Platt, Microsoft Research, One Microsoft Way, Redmond, WA, USA, November 2010, Ensemble Deep Learning for Speech Recognition.

7. Santosh K.Gaikwad, Dr.Babasaheb Ambedkar Marathwada, Bharti W.Gawali, 2011, A Review on Speech Recognition Technique.

8. Samy Bengio and Georg Heigold, Google Inc, Mountain View, CA, USA, feb. 2007, Word Embeddings for Speech Recognition.

9. Audio-Visual Speech Gerasimos Potamianos, Member, IEEE, Chalapathy Neti, Member, IEEE, Guillaume Gravier,, Ashutosh Garg, Student Member, IEEE, and Andrew W. Senior, Member, IEEE 2006, Recent Advances in the Automatic Recognition.

10. Dandan Mo,December 4, 2012, A survey on deep learning: one small step toward AI.

11. Aalto University publication series, Foundations and Advances in Deep Learning, Kyunghyun Cho, 2014.

