Punjabi Speech Recognition of Isolated Words Using Compound EEMD & Neural Network

AnchalKatyal, Amanpreet Kaur, Jasmeen Gill

Abstract - Automatic Speech recognition and conversion of speech to text is a work which has proved its importance for decades. A lot of work has already been done in this contrast. This paper focuses on the Punjabi speech and the conversion of speech to text using advanced system voice recognition pattern. This paper also focuses on the optimization of the EEMD process by combining EEMD process with the Neural Network. Neural Network has been found to be friendly in the contrast of compounding different algorithms to it and it produces significant results. This paper also focuses on the future works to be considered in the same field.

Keywords - ASR, EEMD, Neural Network, Acoustical Models, Neural Identifier, Data Acquisition.

I.INTRODUCTION

Speech Recognition is a process in which the listener understands the words spoken by the speaker. Normal human being is habituated to this process but in the world of modernization, this process is getting adopted by the robotic world which is again a production of the human brain. Speech recognition is also known as automatic speech recognition or computer speech recognition which means understanding voice of the computer and performing any required task or the ability to match a voice against a provided or acquired vocabulary. The task is to getting a computer to understand spoken language [2]. By "understand" we mean to react appropriately and convert the input speech into another medium e.g. text. Speech recognition is therefore sometimes referred to as speech-to-text (STT) [1] [2]. As the state-of-the-art speech recognizers can achieve a very high recognition rate for clean speech, the recognition performance generally degrades drastically under noisy environments. Noise-robust speech recognition has become an important task for speech recognition.

1.1Automatic Speech Recognition (ASR)

Automatic speech recognition is the process of mapping an acoustic waveform into a text/the set of words which should be equivalent to the information being conveyed by the spoken words. This challenging field of research has a most made it possible to provide a PC which can perform
as a stenographer, teach the students in their mother language and read the newspaper of reader's choice. The advent and development of ASR in the last 6 decades has resolved the issues of the requirements of certain level of literacy, typing skill, some level of proficiency in English, reading the monitor by blind or partially blind people, use of computer by physically challenged people and good hand-eye co-ordination for using mouse.

In addition to this support, ASR application areas are increasing in number day by day. Research in Automatic Speech Recognition has various open issues such as Small/ Medium/ Large vocabulary, Isolated/ Connected/Continuous speech, Speaker Dependent/ Independent and Environmental robustness [1]. The below figure represents the general model of the automatic speech recognition and matching process which identifies the category of the speech with two systematic approaches. There are two sections in any automatic speech recognition process:

i. Training
ii. Testing

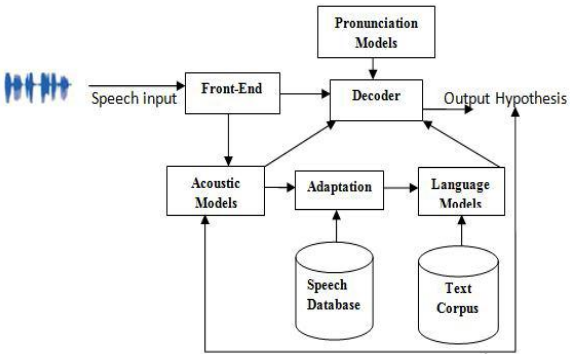

Figure1: The general speech recognition process

A training set is must for the algorithm to work properly as the system can't identify any speech without a proper category in the database. This procedure may also be termed as knowledge discovery [2] [6]. A testing set is a set of voice samples which are to be tested. Every sample of the testing set is matched to the training categories on the basis of which the training models have been defined [4]. The training section may possess the following procedures:

i. Preprocessing
ii. Feature Extraction
iii. Storing data

There are several algorithms for the completion of the training purpose like EEMD, KNN, EMD, EEMD etc [13] [14].

A. EEMD

EEMD stands for Ensemble Empirical mode decomposition. EEMD is a newly developed method aimed at eliminating emotion mode mixing present in the original empirical mode decomposition(EMD)[3]. EEMD which is adaptive and appears to be suitable for non-linear and non-stationary emotional speech signal analysis. It was carried in the time domain to form the basis functions adaptively. The major advantage of EEMD is the basis functions can be directlyderived from the emotional speech signal itself. The emotional intrinsic modes are not necessarily sinusoidal functions [5]. Apparently, EEMD is

Manuscript Received March 2014.

AnchalKatyal, M.Tech (CSE) RIMT-IET Mandi Gobindgarh, India.
Amanpreet Kaur, Assistant Professor BBSBEC Fatehgarh Sahib, India.
Jasmeen Gill, Assistant Professor RIMT-IET MandiGobindgarh, India.

empirical, intuitive, direct, and adaptive. The EEMD decomposes the original emotional signal into a definable set of adaptive basis of functions called the emotional intrinsic mode functions (IMF) [6]. In fact, IMF can be both amplitude and frequency modulated [3] [5] [6].

B. NEURAL NETWORK

A Neural network (NN) is a feed-forward, artificial neural network that has more than one layer of hidden unitsbetween its inputs and its outputs. Each hidden unit, j, typically uses the logistic function1 to map its totalinput from the layer below, xj , to the scalar state, yj that it sends to the layer above[7].

$$y_j = \text{logistic}(xj) = 1/1 + e^{-xj}, x_j = bj + \sum y_i w_{ij}$$

where bj is the bias of unit j, iis an index over units in the layer below, and wij is a the weight on a connection to unit j from unit iin the layer below. For multiclass classification, output unit j converts its total input, xj , into a class probability, pj.

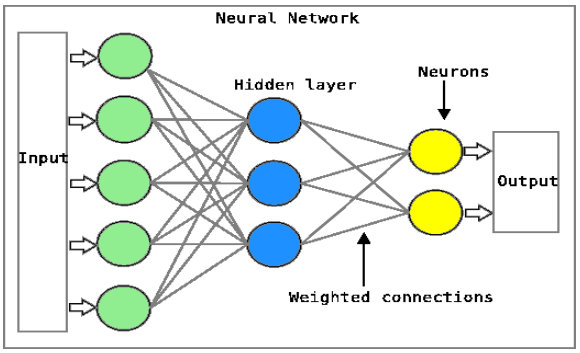

Figure 2: The general architecture of the neural network.

II. PUNJABI PHONEMES

The syllable comprises vowel and consonants. The presence of vowel is must in a syllable. The vowel is the nucleus, presence of consonant is optional. Vowel (V) is always the nucleus part and the left part is onset and the right part is coda that is consonant. The seven types of syllables recognized in Punjabi language are as follows:

V, VC, CV, VCC, CVC, CCVC, CVCC

There are thirty eight consonants, ten non-nasals vowels and same number of nasal vowels in Punjabi language. Consonants can appear with vowels only. Following are the list of consonants in Punjabi language:

ਸ ਹ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ
ਟ ਠ ਡ ਢ ਣ ਤ ਥ ਦ ਧ ਨ ਪ ਫ ਬ
ਭ ਮ ਯ ਰ ਲ ਵ ੜ ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼

List of Non-Nasal Vowels:

ਈ ਇ ਏ ਐ ਅ ਆ ਔ ਊ ਉ ਓ

The number of nasal vowels is same as non-nasal ones and is represented by Bindi or Tippi over the Non-Nasal Vowels.

III.PROPOSED WORK

In our proposed work the output of the EEMD has been stored as a data set which has to be provided as an input to the Neural Network. The second input of the Neural Network becomes the file which has to be uploaded to be tested for its category. Three categories of Punjabi Voice have been taken to be tested. Our research work has mainly following points of problems

i. Training the system using the EEMD algorithm for several voices of Punjabi language
ii. Our problem definition also includes the preprocessing steps like cleaning up the signals and segmentation of the voice signal for better enhancement and recognition
iii. Our problem definition also includes the conversion of speech to text using artificial neural network i.e. custom neural network and EEMD process. The converted speech to punjabi text is the required output.

The simulations have been done in MATLAB 2010 which involves the following steps.

A. Preprocessing of the audio frequency

In this step the noisy spectrum of the audio signal which has been uploaded is rectified and segmented using EEMD algorithm.

B. Feature Extraction of the segmented region

In this step essential features of the audio frequency have been calculated.

C. Classification of Neural Identifier

In this stepthe extracted features as data acquisition first input array has been provided to the custom Neural Network and the second input is the classified dataset of the categories.

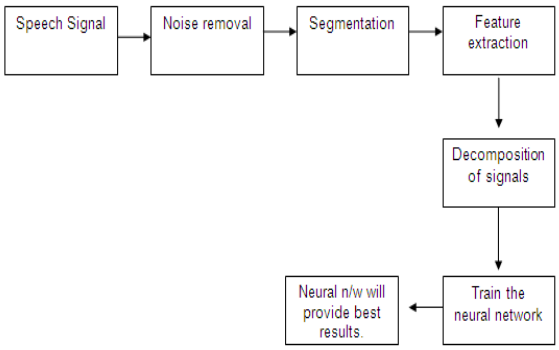

Figure 3: Proposed flow diagram of PSR

The Neural Network matches the data set accordingly and identifies the best possible category of the uploaded voice sample. The Neural Network generates weight for each input as the custom Neural Network is known for. It also generates the weight for the clustered dataset also (stored as an input data set). Then it defines the difference between each data set's weight and the weight of the uploaded data file. The minimum difference of the data set stored to the data set has been uploaded results into the category of the file [12].

IV. IMPLEMENTATION

A. Feature Extraction

In the first stage of the ASR system the raw speech data (signal waveforms) are parameterized into sequences of feature vectors. Since this is a somewhat time consuming task (and since the speech signal files are very large), this feature extraction has been done in advance for the training data set. The feature vector sequences for the training and test data set.

B. Training the Acoustical Models

The parameters of the trained samples are trained using the EEMD-algorithm. The emission pdfs are mixtures of Gaussians (cf. tutorial 'Mixtures of Gaussian'). The EEMD-algorithm is an iterative algorithm, each iteration is invoked in Data acquisition with a call of the function Herest. To find initial parameters the Data acquisition tool HCompV can be used. It scans the set of training feature files, computes the global mean and variance and sets the parameters of all puffs in a given EEMD to this mean and variance values. The invocation of these two tools is implemented in the Perl program train.pl. Its first argument specifies the base names of the new directories that will be created to save the parameters of the trained EEMDs.
The second argument is the training file (use either train1.scp or train2.scp) and the third argument species the number of iterations which are performed to train the EEMDs. For example, if you use the command Perl train.pl ABC train1.scp 2, you will get 3 new directories called ABC EEMD0 (with the initial EEMD parameters), ABC EEMD1 (the EEMD parameters after the first iteration) and ABC EEMD2 (the EEMD parameters after the second and final iteration). In the command line window the commands used by the Perl program to invoke the tools are echoed.

C. Recognizing Test Data and Evaluation of the Recognition Result

The data acquisition tool Hive is a general-purpose Viterbi word recognizer. It matches speech signals against a network of EEMDs and returns a transcription for each speech signal. Results are the Data acquisition performance analysis tool. It reads in a set of label files (typically output from the recognition tool such as Hive) and compares them with the corresponding reference transcription.
The Perl script test.pl first calls Hive to perform speech recognition and obtain a transcription of the
Test speech signals2, and then Result to compute recognition statistics, such as the percentage of correctly recognized words. Its first and only argument is the name of the directory where the EEMD specifications are saved. For example, to use the trained models from the last subsection for recognition, type Perl test.pl ABC EEMD2 Again, the commands for calling the data acquisition[12].

V. PROPOSED ALGORITHM

1. Initialize voice smmple
2. Initialize i=0;featurecount=0;
3. Initialize data_recognized=0;
4. Draw frequency pattern=true
5. For i=1:no.offrequencypat
6. Extract.features.sample(i)
7. Decompose.signal.eemd(i)
8. If decomposition.done==true
9. Match.db.pattern(call.neural)
10. Input.weight[1]=wt.sampled(i)
11. Input.weigth[2]=db.dataset
12. If char.recognized==true
13. Data_recognized=data_recognized+1
14. End
15. Else
16. Docompose.goto(5)
17. End

VI. METHODOLOGY

Record individual samples speech signals (Punjabi language) in .wav file.

A. Adding White Noise to existing Speech signals.
B. Segmentation of the wav file is doneusing EEMD and check peaks of segmented parts.
C. Speech Feature Extraction using EEMD algorithm.
D. Extract features of all the wave file used for the training of the system.
E. Creation of the feature vector which consists of features (Properties of the files) like Frequency, Max and Min, Avg Frequency, Spectral Roll off , Jitter , Shimmer . More number of feature extractions will lead to a better matching algorithm formation.
F. The average feature of all the wave files selected would be stored in the MAT file of the system so that it can be used further in the testing part.
G. Neural network would be used at the time of the testing of the category of the user. The neural network would take two inputs:

1. The file to be tested by the user
2. The database of feature vector which has been created using EEMD algorithm

H. The same features would be extracted of the wave file to be uploaded and the neural network would provide the best matching result. The custom neural network is used for the conversion of speech to text.
I. Compute and compare the results with the base paper.

VII RESULTS AND DISCUSSIONS

The results of the proposed work include conversion of speech to text and it comprises of pre-processing, segmentation and feature extraction of the recorded .wav files. The results of the proposed algorithm (BPA Algorithm for Feed Forward Neural Networks) have been presented below.

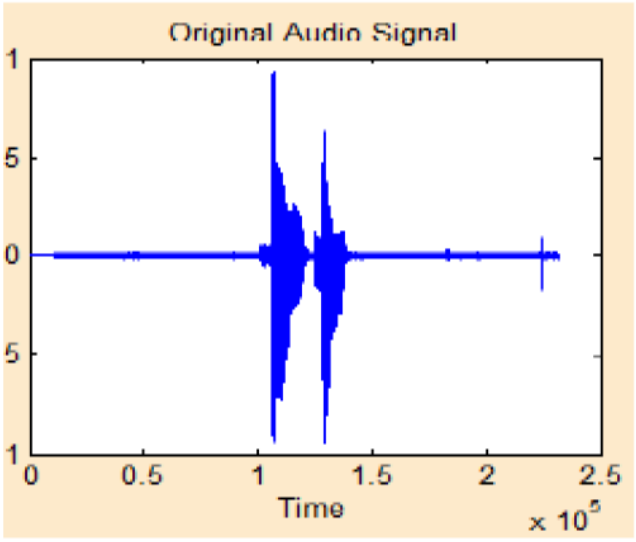

Figure 4: Original signal.

The plot has been drawn between the time and the

amplitude of the signal. The amplitude part of the signal would shift when the noise gets added to it.

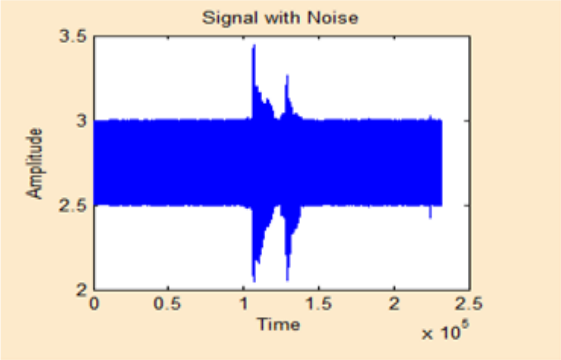

Figure 5: The noisy signal

Random noise has been added to the signal. As it is already said that there would be shift in the amplitude of the signal, it would get shifted to an extent. To add noise to the signal, random bits have been taken.

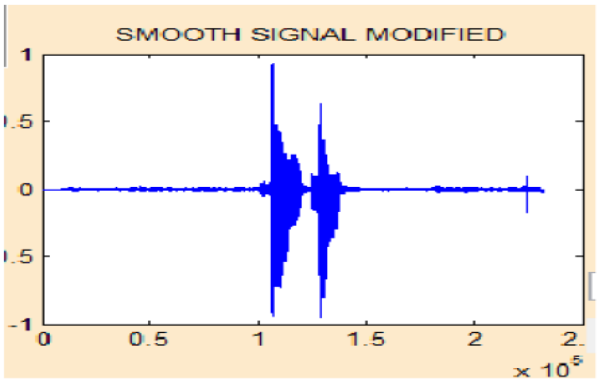

Figure 6: Smooth signal.

A smooth signal is achieved when we apply some sort of filter to it.

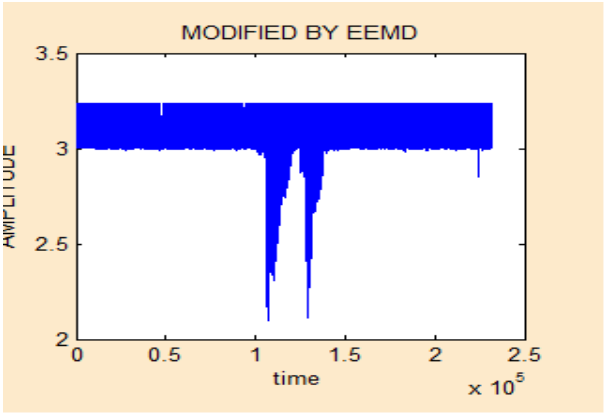

Figure 7 : Modified signal by the EEMD process.

The modified signal would go under the processing of the signal and above the threshold; each value of the signal would be clipped.

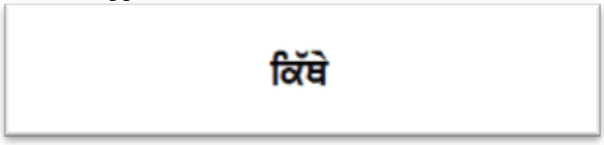

Figure7: The extracted word "Kitthe".

The extraction process has been accomplished by using Neural Network and the results have been tested using different samples of one human being.

Figure8: The extracted word "Shanivar"

The extraction process has been accomplished by using Neural Network and the results have been tested using different samples of one human being.

Figure9: The extracted word "Duja"

The extraction process has been accomplished by using Neural Network and the results have been tested using different samples of one human being.

Figure10: The extracted word "Jamaat"

The extraction process has been accomplished by using Neural Network and the results have been tested using different samples of one human being.

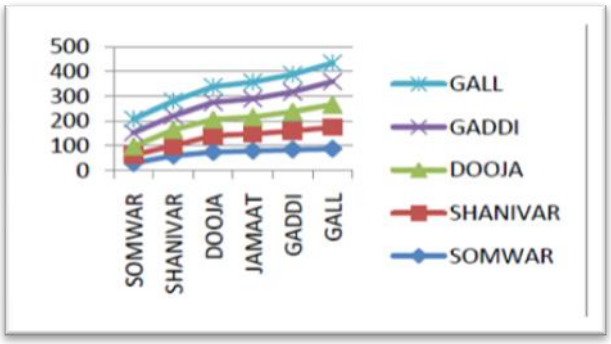

Figure11: The accuracy rate over number of sampled data in the context

It is found that the average accuracy of a word lies between 85 to 94 percent.

WORD	ACCURACY OF RECOGNITION IN %	NO OF USERS IDENTIFIED	NO OF USERS ININPUT SAMPLE
ਕਿੱਥੇ	100	3	5
ਸੌਂਅ੍	85	3	5
ਨੂਮਸ਼ੀ	95	4	6
ਵੀਰਵਰ	93	3	5
ਜਮਾਤ	87	4	6
ਗੱਲ	91	4	6
ਅੱਟਵਰੋਓਰ	92.5	3	5
ਮੁਬਅੁੜਅਅ੍	93	4	6
ਅੰਭਅਲਅ	91	3	5
ਅਅਖੁਪਓ	92	3	5
ਸੌਅ	93	4	6
ਠਓਓਝਅ	94.12	3	5
ਘੱਟੀ	88.32	4	6

Figure12: The accuracy of categories taken in account varies from 88 to 96 percent using the compound algorithm.

VIII. FUTURE SCOPE

The current work opens a lot of future possibilities. The current work involves the combination of EEMD and Custom Neural Network only where as several other methods of Neural like Back Propagation , Bacterial Forging Optimization are present which can be implemented instead of CNN which may produce some more effective results.

REFERENCES

[1] Preeti Saini, "Automatic Speech Recognition", A Review International Journal of Engineering Trends and Technology-Volume4Issue2- 2013.

[2] WiqasGhai, "Literature Review on Automatic Speech Recognition", International Journal of Computer Applications (0975 – 8887)Volume 41– No.8, March 2012.

[3] Yuqiang Qin,"EEMD-Based Speaker Automatic Emotional Recognition," Chinese Mandarin Appl. Math.Inf. Sci. 8, No. 2, 617-624 (2013).

[4] Akshay S. Utane, "Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine ", International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.

[5] Shing-Tai Pan, "Robust Speech Recognition by DEEMD with A Codebook Trained by Genetic Algorithm", Journal of Information Hiding and Multimedia Signal Processing, October 2012.

[6] Wu, S., Falk, "Automatic recognition of speech emotion using long-term spectro-temporal features," Proc. Internat. Conf. on Digital Signal Processing, 1-6 (2009).

[7] Geoffrey Hinton, Li Deng, "Neural Networks for Acoustic Modeling in Speech Recognition".

[8] H. Hermansky, D. P. W. Ellis, "Tandem connectionist feature extraction for conventional EEMD systems," Proceedings of ICASSP, Los Alamitos, CA, USA, 2000, vol. 3, pp. 1635–1638, IEEE Computer Society.

[9] H. Bourlard, N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach", Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[10] L. Deng, "Computational models for speech production," in Computational Models of Speech Pattern Processing, pp. 199–213. Springer- Verlag, New York, 1999.

[11] L. Deng, "Switching dynamic system models for speech articulation and acoustics," in Mathematical Foundations of Speech and Language Processing, pp. 115–134. Springer-Verlag, New York, 2003.

[12] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

[13] A. Mohamed, G. Dahl, "Acoustic modeling using deep belief networks," IEEE Transactions on Audio, Speech, and Language Processing,, vol. 20, no. 1, pp. 14–22, jan. 2012.

[14] D. E. Rumelhart, G. E. Hinton, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proceedings of AISTATS, 2010, pp. 249–256.

[16] D. C. Ciresan, U. Meier, "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition," Neural Computation, vol. 22, pp. 3207–3220, 2010.

[17] G. E. Hinton , R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[18] H. Larochelle, D. Erhan, "An empirical evaluation of deep architectures on problems with many factors of variation," in Proceedings of the 24th international conference on Machine learning, 2007, pp. 473–480.

[19] J. Pearl, "Probabilistic Inference in Intelligent Systems: Networks of Plausible Inference", Morgan Kaufmann, 1988.

[20] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol. 14, pp. 1771–1800, 2002.

[21] G. E. Hinton, "A practical guide to training restricted boltzmann machines," Tech. Rep. UTML TR 2010-003, Department of Computer Science, University of Toronto, 2010.

BIOGRAPHIES:

Author 1: AmanPreet Kaur, an effectual Assistant Professor has an experience of 10.5 years. She has completed M. Tech and B Tech in Computer Science Engineering and pursuing P.hd. She has published 10 national and international papers and pursuing research work on Punjabi Speech recognition. She has a membership of IEEE.

Author 2: Jasmeen Gill, a dynamic Assistant Professor.has completed M. Tech and B Tech in Computer Science Engineering having an experience of 8 years. She has published 3 national and international papers. Her areas of interest are Artificial Intelligence and Image Processing. She has a membership of IEEE.

Author 3: Anchal Katyal has completed B.Tech and Pursuing M.Tech.She has published 4 national and international papers.She has a membership of IEEE.