Detection of Attributes for Bengali Phoneme in Continuous Speech using Deep Neural Network

Tanmay Bhowmik
Centre for Educational Technology
IIT Kharagpur
India

Sankar Mukherjee
Laboratoire Parole et Langage
CNRS, Aix-Marseille University
France

Shyamal Kumar Das Mandal
Centre for Educational Technology
IIT Kharagpur
India

***Abstract*— Hidden Markov Model (HMM) has contributed greatly in the area of speech recognition during last two decades. However, in recent days, detection-based, bottom-up speech recognition techniques achieve high success rate. In this detection-based, bottom-up approach of speech recognition, first step is detection of speech attributes like place and manner of articulation of the phonemes. This paper describes about the detection of attributes which leads to identification of place and manner of articulation of Bengali phonemes using Deep Neural Network (DNN).**

Keywords—speech attribute; place of articulation; manner of articulation; deep neural network; deep belief network.

I. Introduction

State-of-art Automatic Speech Recognition (ASR) systems are based on Hidden Markov Model (HMM) where a pattern matching framework is used for isolated word recognition [1]. The approach of ASR is to train the acoustic models usually at the subword level and to use dynamic programming techniques to find the best word sequence for a given spoken utterance [2].

However, after decades of research and many successful innovations, the performance of ASR has not reached the performance level of Human Listeners due to some limitations of HMM. Major limitations are the assumptions that successive observations are independent and the probability of a given state at time t only depends on the state at time t-1, are inappropriate for speech sounds where dependencies extend through different states [1].

As a result, some alternative research interests of designing ASR has grown up. A recently proposed bottom-up framework is Automatic Speech Attribute Transcription (ASAT) based on phonological features i.e. speech attribute detection [3]. In this approach, first the speech attributes are detected, then using the probabilistic model, attribute to phone, phone to syllable and syllable to word is recognized in order. ASAT has been applied to a number of tasks including rescoring of word lattices generated by HMM systems [4], continuous phoneme recognition [5], cross-language attribute detection and phoneme recognition [6]. Majority of the phonological-based approaches to speech recognition are supported by abstract representation of articulation [7] which is concerned with how to extract phonetic features and use them in ASR [5][6]. ASR can be improved by using more linguistically motivated features [7] as they have several nice properties, such as robustness to noise and cross-speaker variation [8], portability across different languages [6] etc. Different approaches have been developed for feature extraction. Artificial Neural Networks (ANN) is used [8] to score speech attributes. A bank of feature detectors can be used to capture articulatory information. This can be used in a lattice rescoring process to correct utterances with errors in Large Vocabulary Continuous Speech Recognition (LVCSR) [4]. Recent trends of using Deep Neural Network (DNN) achieved high success rate in phonetic feature extraction and recognition. A range of DNN architectures has been used for boosting attributes and phone estimation accuracies for detection-based speech recognition [9]. Context-dependent pre-trained deep architecture has been used successfully with a very good performance in Large Vocabulary Speech Recognition (LVSR) [10], conversational speech transcription [11] also.

In this study it has been tried to detect the phonological features i.e. the speech attributes. This type of ASAT model is an attempt to copy the human speech recognition capabilities with various speech event detection followed by bottom-up knowledge integration. Here, the speech attributes are the place and manner of articulation of phonemes and the attributes are detected for the phonemes which are used in Bengali speech. For this detection purpose, the conventional single hidden layer Neural Network approach has been extended to the DNN with three hidden layers with 300 hidden units in each layer.

II. Phonological Features in Bengali

In Bengali language, phonological features i.e. the acoustic-phonetic speech attributes are based on place of articulation and manner of articulation. During the articulation the airstream through the vocal tract is obstructed. The place where the obstruction takes place is called the place of articulation whereas manner of articulation is concerned with airflow, the paths it take and the degree to which it is impeded by vocal tract constrictions. Consonants are classified depending on the place of articulation and manner of articulation. Vowels may be specified in terms of the position of the tongue and the position of the lips. Depending on these classifications the

978-1-4799-5991-4/15/$31.00 ©2015 IEEE

phonological features of the Bengali phonemes and their corresponding attributes are listed in Table I [12].

TABLE I. BENGALI SPEECH ATTRIBUTES AND ASSOCIATED PHONEMES

		Attributes	**Phonemes**
Consonants	*Place of Articulation*	Velar	/k/, /kʰ/, /g/, /gʰ/, /ŋ/
		Post-Alveolar	/ʧ/, /ʧʰ/, /ʤ/, /ʤʰ/, /ɽ/, /ʃ/
		Alveolar	/ʈ/, /ʈʰ/, /ɖ/, /ɖʰ/, /r/, /s/
		Dental	/t/, /tʰ/, /d/, /dʰ/, /n/, /l/
		Bilabial	/p/, /pʰ/, /b/, /bʰ/, /m/,
		Glottal	/h/
		Palatal	/ɳ/, /j/
	Manner of Articulation	Plosive	/k/, /kʰ/, /g/, /gʰ/, /ʈ/, /ʈʰ/, /ɖ/, /ɖʰ/, /t/, /tʰ/, /d/, /dʰ/, /p/, /pʰ/, /b/, /bʰ/
		Stop	
		Affricate	/ʧ/, /ʧʰ/, /ʤ/, /ʤʰ/
		Fricative	/ʃ/
		Nasal Murmur	/m/, /n/, /ŋ/, /ɳ/
		Lateral	/l/
		Trill	/r/
		Retroflex flap	/ɽ/, /ɽh/
		Approximant	/j/, /w/
		Unvoiced	/k/, /kʰ/, /ʈ/, /ʈʰ/, /t/, /tʰ/,/p/, /pʰ/, /ʃ/, /s/, /h/
		Voiced	/g/, /gʰ/, /ɖ/, /ɖʰ/, /d/, /dʰ/, /b/, /bʰ/, /m/, /n/, /ŋ/, /ɳ/, /l/, /r/, /ɽ/, /ɽh/, /j/, /w/
		Unaspirated	/k/, /g/, /ʈ/, /ɖ/, /t/, /d/, /p/, /b/
		Aspirated	/kʰ/,/gʰ/,/ʈʰ/, /ɖʰ/,/tʰ/,/dʰ/,/pʰ/,/bʰ/
Vowels	*All vowels are voiced*	Back, close, rounded	/u/
		Back, close-mid, rounded	/o/
		Back, open, rounded	/ɔ/
		Front, open, unrounded	/a/
		Front, open-mid, unrounded	/æ/
		Front, close-mid, unrounded	/e/
		Front, close, unrounded	/i/
Others		Silence	Considered as '/sil/' in this experiment.
		Diphthongs	Monosyllabic vowel-vowel combination.

III. DEEP NEURAL NETWORK

A DNN is described as a multi-layer perceptron with more than one hidden layer. In general, each hidden unit uses the sigmoid function to map from input layer to output layer [13]. If each hidden unit is represented by j, the input x_j and the output y_j then,

$$y_j = sigmoid\left(x_j\right) = \frac{1}{1+e^{-x_j}},\ x_j = b_j + \sum_i y_i w_{ij}, \qquad (1)$$

where b_j is the bias of unit j, i is an index of input layer, which is in the bottom of the hidden layer and w_{ij} is the weight of the association to unit j from the unit i.

In this experiment, an attribute detector has been designed, based on pre-trained, deep neural network model. Due to the deeper and more expressive neural network architecture, more efficient training is desired. In this case, unsupervised Deep Belief Network (DBN) pre-training is employed to ensure more effective training [14]. A widespread observation of DBN illustrates that each layer of a DBN is greedily initialized by treating each pair of layers as Restricted Boltzmann Machine (RBM) before the joint optimization of all the layers [15].

A. *Restricted Boltzmann Machine (RBM)*

An RBM consists of a layer of visible units which represent the input data and a layer of hidden units which study the considerable dependencies between the visible units [15]. There are undirected connections between the visible and the hidden units but no interlayer connections i.e. no connection between visible-visible or hidden-hidden units. In that sense, it is termed as "Restricted" Boltzmann Machine. Due to no visible-visible and hidden-hidden connections, the learning is faster in RBM [13] [16].

B. *Deep Belief Network (DBN)*

Each of the RBM is an undirected model; however the DBN is a hybrid generative structure whose top two layers are undirected and the lower layers have top-down directed structure [13] [19]. The top two layers form the final RBM.

DBN is distinguished from other multilayer, directed generative structure in such a way that in case of DBN, the states of the layers of the hidden units can be derived in a single forward pass through a fairly accurate way [13]. So the DBN is learnt by training a stack of RBMs and the generative weights are used in the reverse direction to initialize entire feature detecting layers of a deterministic feed-forward DNN. Thus the DNN is pre-trained generatively as a DBN.

IV. EXPERIMENTAL SETUP

In this experiment, an attribute detector has been designed for each of the attributes mentioned in Table I. It detects the presence of the corresponding attributes in the input speech signal. A functional block diagram of the attribute detector is given on Fig. 1.

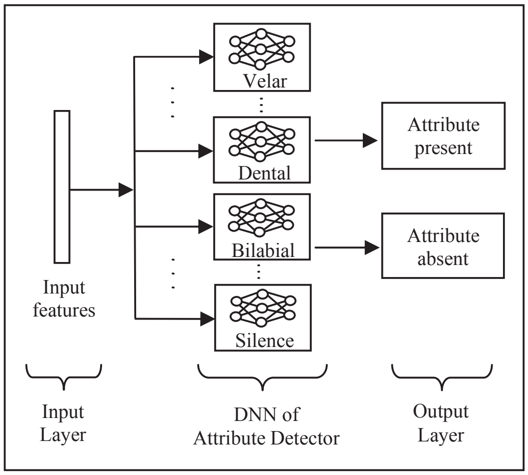

Fig. 1. Functional block diagram of Attribute Detector

The Bengali read speech corpus from the Center for Research on Bangla Language Processing (CRBLP) of BRAC University, Bangladesh [20] and the speech corpus from Centre for Development of Advanced Computing (CDAC) has been used for continuous spoken input speech data. The CRBLP speech corpus contains recordings of a professional male speaker's voice of age 27 and the CDAC speech corpus contains four female and five male speaker's voice of different age group varies from 25 to 40. Audio and text have been time-aligned and labeled at the sentence-level. The sampling frequency of all the recording sentences in the speech corpus is 44.1 KHz. All the recordings were resampled to 16 KHz. Resampling reduces irrelevant information which is not produced by human voice, in the sound files [17] [18].

The experiment was implemented separately for two cases. In the first case, 300 sentences were taken randomly from the 'CRBLP' and CDAC speech corpus out of which 200 sentences are used for training purpose, 50 sentences for testing and 50 sentences for cross validation. For the second case, 500 sentences were taken out of which 300 sentences were for training, 100 for testing and 100 for cross validation. The sentences for testing purpose are first transcribed with HMM toolkit (3 states HMM). Then the phonetic alignment is done manually for these sentences which are used to fine tune the DNN. Hamming window of 10 ms frame and a frame shift of 5 ms has been taken for the spectral analysis of the input data. 10ms framing yields 15346 speech frames in test data for 50 sentences and 31533 speech frames in test data for 100 sentences.

The input to each attribute detector can be any speech features. In this experiment, we adopt MFCC features. That is why, for each frame, 12 MFCC features plus the 0th cepstral coefficient is computed. 13 delta and delta-delta coefficients are also computed to yield a 39 dimensional input feature vector. Delta and delta-delta coefficients are computed using the equation $\delta_1[y(n)] = y(n+1) - y(n-1)$ and $\delta_2[y(n)] = 0.5y(n+1) - 2y(n) + 0.5y(n-1)$ respectively. The number of possible outputs for each detector is two: attribute present or absent.

Each DBN layer is pre-trained for 20 epochs as a RBM with mini-batch of size 10. Parameters are updated with a learning rate of 0.002 and a momentum of 0.98. Mean squared error is used as the objective function. In DNN training, the number of epochs was taken as 30 with batch size of 25. The learning rate is taken as 0.2 and the dropout fraction was taken as 0.5.

A DELL precision T3600 workstation is used for this experiment. This workstation is a 6 core computer with a CPU clock speed of 3.2 GHz, 12 MB of L3 cache memory and 64 GB DDR3 RAM and an NVIDIA Quadro 4000 General Purpose Graphical Processing Unit (GPGPU).

V. Results

Distribution of different speech attributes of Bengali phonemes in the input data are given in Fig. 2. To calculate the number of occurrences of attributes, first the numbers of associated phonemes are counted. As an example, say we have to count how many times the attribute 'Affricate' is observed in the input dataset. To count this, first thing is to do, to find the associated phonemes. Associated phonemes can be observed from Table I and they are /ʧ/, /ʧʰ/, /ʤ/, /ʤ^{h}/. So the total number of occurrences of phonemes /ʧ/, /ʧʰ/, /ʤ/, /ʤ^{h}/ will be the count of the total number of 'Affricate' in the input dataset. From Fig. 2 we found that the number of affricate is 406 and 796 respectively for two distinct number of test data frames.

The number of possible output for each attribute detector is two: attribute present or absent. There are some frames which are misclassified. Table II shows the detection accuracies of the correctly detected attributes of the Bengali phonemes along with the number of occurrences of each attributes in input test data. Detection accuracy for each attribute is calculated with the formula,

$$accuracy = \left(\frac{\sum inputframes - \sum misclassifiedframes}{\sum inputframes} \right) \times 100\% \qquad (2)$$

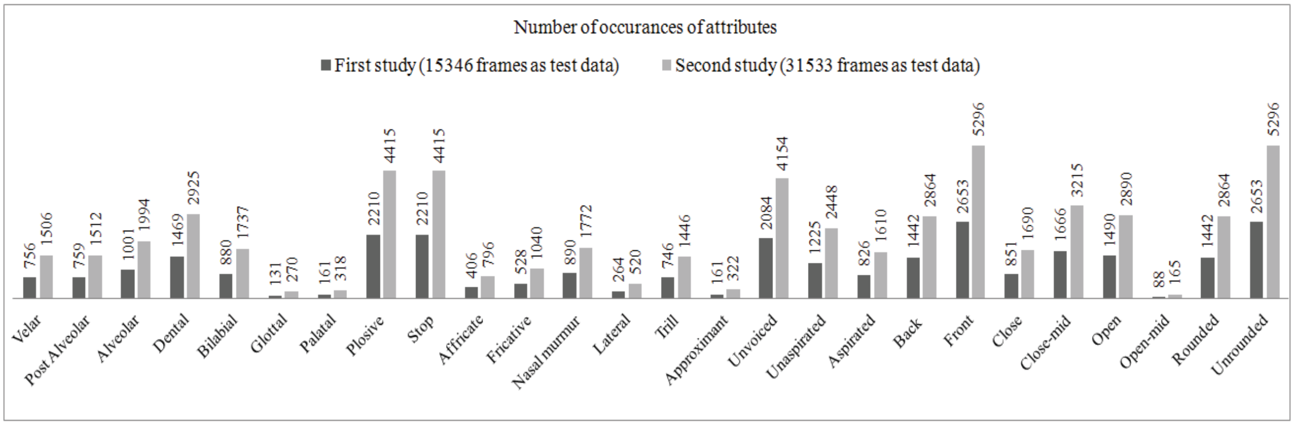

Fig. 2. Distribution of speech attributes of Bengali phoneme in input test data

TABLE II. ACCURACIES FOR THE DETECTED SPEECH ATTRIBUTES ASSOCIATED WITH THE BENGALI PHONEMES

Attribute	**Study 1 (Number of test data frames 15346)**		**Study 2 (Number of test data frames 31533)**	
	Number of occurrences of attributes	***Accuracy (%)***	***Number of occurrences of attributes***	***Accuracy (%)***
Velar	756	×	1506	×
Post Alveolar	759	×	1512	89.25
Alveolar	1001	93.70	1994	93.72
Dental	1469	89.71	2925	87.89
Bilabial	880	×	1737	×
Glottal	131	×	270	×
Palatal	161	×	318	×
Plosive	2210	81.97	4415	82.64
Stop	2210	81.97	4415	82.64
Affricate	406	×	796	94.58
Fricative	528	×	1040	×
Nasal murmur	890	92.93	1772	93.07
Lateral	264	×	520	×
Trill	746	×	1446	×
Approximant	161	×	322	×
Unvoiced	2084	77.57	4154	77.30
Unaspirated	1225	84.31	2448	85.46
Aspirated	826	×	1610	91.57
Back	1442	88.35	2864	88.44
Front	2653	74.45	5296	74.07
Close	851	×	1690	93.14
Close-mid	1666	84.24	3215	84.38
Open	1490	85.76	2890	86.05
Open-mid	88	×	165	×
Rounded	1442	88.35	2864	88.44
Unrounded	2653	74.45	5296	74.07

It is observed from Table II that, in case of the first study when number of test data frames was 15346, the attribute detector couldn't achieve any detection accuracy for some of the attributes like velar, post alveolar, affricate, lateral etc. The attribute detector couldn't identify the misclassified frames properly for these attributes. That is why, no detection accuracy was achieved. Whereas, in the second study, when number of test data frames was 31533, the attribute detector detected some of them, like post alveolar, affricate, aspirated and close vowel. The number of misclassified frames for each of the attributes is observed in Table III. The detector couldn't identify the misclassified frames when the number of occurrences of the attributes in the test data frames is quite low and hence no detection accuracy was achieved. When number of test data frames is increased, number of occurrence of the attributes is also increased and then some of them were detected with high accuracy. Detection accuracy for other attributes remains almost same when the number of test data frames increased from 15346 to 31533.

From Table II, it is also observed that, in case of attribute 'Plosive' and 'Stop', the number of occurrences in input data and accuracy, all are equal in value. Table I shows that these two are the common attributes of 16 phonemes /k/, /kʰ/, /g/, /gʰ/, /ʈ/, /ʈʰ/, /ɖ/, /ɖʰ/, /t/, /tʰ/, /d/, /dʰ/, /p/, /pʰ/, /b/, /bʰ/. Similarly in case of vowels, /a/, /æ/, /e/, /i/ have the common attributes 'Front' and 'Unrounded' and attribute 'Back' and 'Rounded' are the common attributes of the vowels /u/, /o/, /ɔ/. As a result, in case of these attributes same values are also observed.

TABLE III. NUMBER OF MISCLASSIFIED FRAMES FOR THE SPEECH ATTRIBUTES

Attribute	**Number of test data frames 15346**	**Number of test data frames 31533**
	Number of misclassified frames	
Velar	×	×
Post Alveolar	×	3389
Alveolar	1007	1980
Dental	1887	3819
Bilabial	×	×
Glottal	×	×
Palatal	×	×
Plosive	2767	5476
Stop	2767	5476
Affricate	×	1709
Fricative	×	×
Nasal murmur	1085	2186
Lateral	×	×
Trill	×	×
Approximant	×	×
Unvoiced	3441	7158
Unaspirated	2408	4586
Aspirated	×	2659
Back	1720	3646
Front	4085	8179
Close	×	2163
Close-mid	2327	4925
Open	2103	4400
Open-mid	×	×
Rounded	1720	3646
Unrounded	4085	8179

A cross validation set has been generated by extracting 50 and 100 sentences respectively for study 1 and study 2, from total data set to terminate the training. So, for each of the correctly detected attributes a plot of error vs. number of epochs has been generated for the training set and cross validation set.

Some DNN error plots of the training and validation data for some attributes are given below. Fig. 3 and Fig. 4 depict the DNN error plot for the attribute Dental. These are error vs. number of epochs plot. An epoch should not be mixed up with iteration. An epoch is measured with the number of times when all of the training vectors are used once to update the weights.

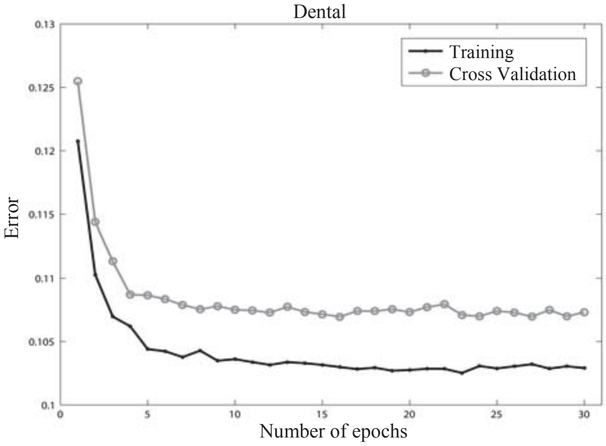

Fig. 3. DNN error on the training and cross validation data for the attribute 'Dental' (Study 1)

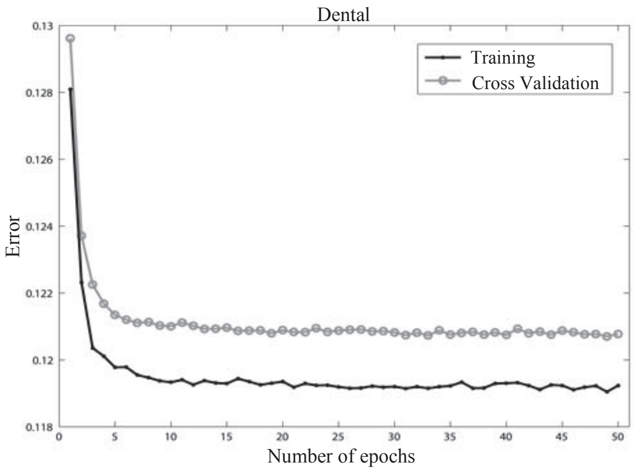

Fig. 4. DNN error on the training and cross validation data for the attribute ‘Dental’ (Study 2)

Fig. 3 and Fig. 4 illustrate the error plot for the attribute ‘Dental’ for study 1 and study 2 respectively. From Table III, we can see that the detection accuracy for the attribute dental is 89.71% and 87.89% for study 1 and study 2 respectively.

Fig. 5, Fig. 6, Fig. 7 and Fig. 8 are the error plots for the attributes unaspirated and rounded respectively for study 1 and study 2. Here the error plots of the attributes dental, unaspirated and rounded are given just because of a random selection of the attributes, one from attribute based on place of articulation i.e. dental, one from attribute based on manner of articulation i.e. unaspirated and one from vowel i.e. rounded.

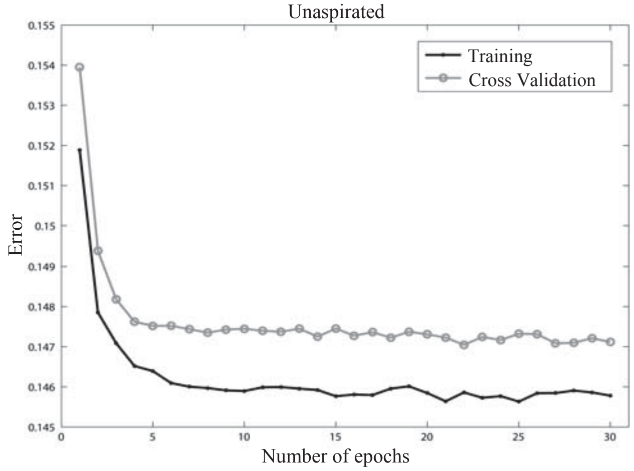

Fig. 5. DNN error on the training and cross validation data for the attribute ‘Unaspirated’ (Study 1)

The accuracy for manner of articulation based attribute, unaspirated is 84.31% and 85.46% respectively for study 1 and study 2 respectively.

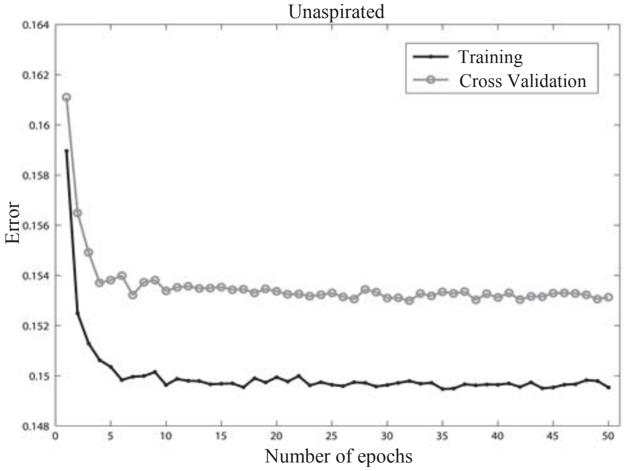

Fig. 6. DNN error on the training and cross validation data for the attribute ‘Unaspirated’ (Study 2)

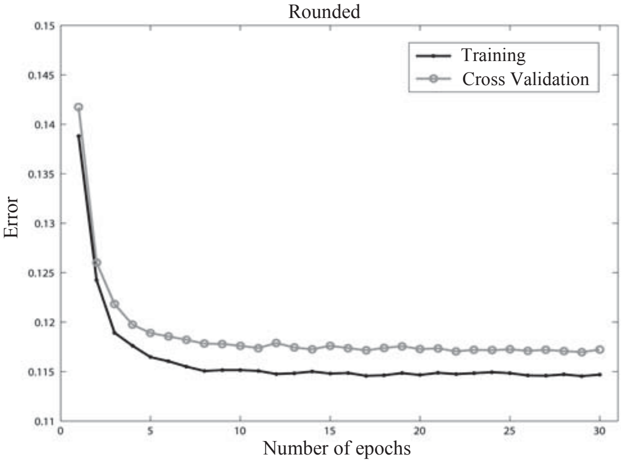

Fig. 7. DNN error on the training and cross validation data for the attribute ‘Rounded’ (Study 1)

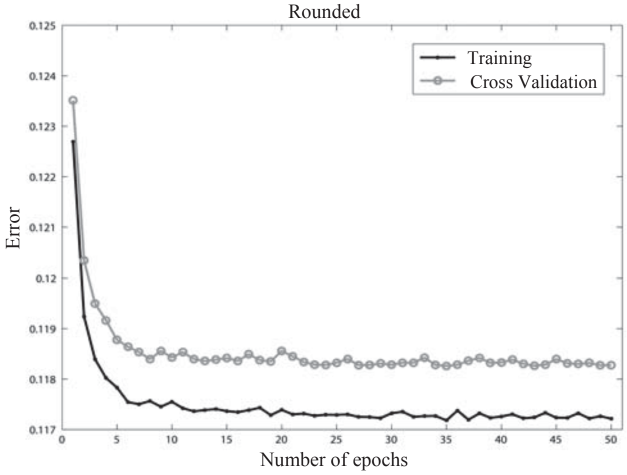

Fig. 8. DNN error on the training and cross validation data for the attribute ‘Rounded’ (Study 2)

The accuracy for the attribute rounded, that is associated with vowel is 88.35% and 88.44% respectively for study 1 and study 2.

VI. Conclusion and Future Work

It is observed from the experimental results that Deep Neural Network can be used to achieve high accuracy in case of place and manner of articulation based attribute detection for Bengali phoneme.

The DNN is pre-trained generatively as a DBN by using the learned weights as the initial weights. Then these weights are fine-tuned by using the back-propagation algorithm. This procedure is very much helpful where limited training data is available as poorly initialized weights can have significant impact on the performance of the final model [21]. Apart from this, a DBN can be efficiently trained in an unsupervised way [13][16]. That is why the DNN approach is very much advantageous than conventional Neural Network framework and HMM based techniques when it needs to train a very large amount of data.

DNN based applications is very useful for large volume continuous speech recognition area. Detecting the Bengali Speech attributes by DNN based applications is a very constructive idea for bottom-up, ASAT based ASR design where continuous spoken speech can be directly associated with the phonological features i.e. the speech attributes, overlooking the phonemes.

Till now, the detector couldn't identify many attributes because the misclassified frames are not properly identified. Work is going on to identify those frames which will lead to the detection of the remaining attributes. The misclassified frames can be observed and an estimation of phoneme confusion matrix is very much possible. This kind of bottom-up approach is further can be used for detection of phonemes from the correctly detected attributes and detection of words from correctly detected phonemes which will lead to the design of a Automatic Speech Recognizer for continuous spoken Bengali speech.

References

[1] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[2] I. Bomberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A.M. Daniel, J. Morris, S.M. Siniscalchi, Y. Tsao, and Y Wang, "Detection-Based ASR in the Automatic Speech Attribute Transcription Project," in Proc. Interspeech, Antwerp, Belgium, pp. 1829-1832, August 2007.

[3] C.-H. Lee, M.A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L.R. Rabiner, "An overview on automatic speech attribute transcription (ASAT)," in Proc. Interspeech, Antwerp, Belgium, pp. 1825-1828, August 2007.

[4] S.M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," Speech Communication, vol. 51, no. 11, pp. 1139-1153, 2009.

[5] S.M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in Proceedings of the ASRU, Kyoto, Japan, pp. 566-569, 2007.

[6] S.M. Siniscalchi, D.C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on crosslanguage attribute detection and phone recognition with minimal targetspecific training data," IEEE Transaction on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 875-887, 2012.

[7] S.M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Exploiting deep neural networks for detection-based speech recognition," Neurocomputing 106 (2013), pp. 148-157.

[8] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in Proceedings of ISCLP, Sydney, Australia, 1998, pp. 891-894.

[9] D. Yu, S.M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and Phone Estimation Accuracies with Deep Neural Networks for detection-based speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4169-4172.

[10] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large- vocabulary speech recognition," IEEE Trans. Audio, Speech, and Lang. Proc. Jan. 2012, pp. 30-42.

[11] F. Seide, G. Li and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," Interspeech 2011, pp. 437-440.

[12] S.K. Das Mandal, S. Chandra, S. Lata, A.K. Dutta, "Places and Manner of Articulation of Bangla Consonants: A EPG based study," Interspeech 2011, pp. 3149-3152.

[13] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82-97, 2012.

[14] A.R. Mohamed, T.N. Sainath, G. Dahl, B. Ramabhadran, G.E. Hinton, and M.A. Picheny, "Deep belief networks using discriminative features for phone recognition." in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5060-5063.

[15] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol. 14, no. 8, pp. 1771-1800, 2002.

[16] G. Hinton, "A practical guide to training restricted Boltzmann machines," Neural Networks: Tricks of the Trade, Springer Berlin Heidelberg, pp. 599-619, 2012.

[17] D.R. Feinberg, B.C. Jones, A.C. Little, D.M. Burt, and D.I. Perrett, "Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices," Animal Behaviour, vol. 69, no. 3, pp. 561-568, 2005. attractiveness of human male voices," Animal Behaviour, vol. 69, no. 3, pp. 561-568, 2005.

[18] P. Ladefoged, "Elements of Acoustic Phonetics. 2nd edn," Chicago: University of Chicago Press, 1996.

[19] G.E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[20] F. Alam, S.M. Murtoza Habib, D. A. Sultana, M. Khan, "Development of Annotated Bangla Speech Corpora," Spoken Language Technologies for Under-resourced language (SLTU'10), vol. 1, pp. 35-41, Penang, Malasia, May 3 - 5, 2010.

[21] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation," in Proceedings of the 24th international conference on Machine learning, pp. 473-480. ACM, 2007.