International Journal of Speech Technology (2019) 22:111–119
https://doi.org/10.1007/s10772-018-09577-3

A comparative study of deep neural network based Punjabi-ASR system

Virender Kadyan[1] · Archana Mantri[2] · R. K. Aggarwal[3] · Amitoj Singh[4]

Received: 21 January 2018 / Accepted: 27 November 2018 / Published online: 15 December 2018

Abstract
HMM is regarded as the leader from last five decades for handling the temporal variability in an input speech signal for building automatic speech recognition system. GMM became an integral part of HMM so as to measure the efficiency of each state that stores the information of a short windowed frame. In order to systematically fit the frame, it reserves the frame coefficients and connects their posterior probability over HMM state that acts as an output. In this paper, deep neural network (DNN) is tested against the GMM through utilization of many hidden layers which helps the DNN to successfully evade the issue of overfitting on large training dataset before its performance becomes worse. The implementation DNN with robust feature extraction approach has brought a high performance margin in Punjabi speech recognition system. For feature extraction, the baseline MFCC and GFCC approaches are integrated with cepstral mean and variance normalization. The dimension reduction, decorrelation of vector information and speaker variability is later addressed with linear discriminant analysis, maximum likelihood linear transformation, SAT, maximum likelihood linear regression adaptation models. Two hybrid classifiers investigate the conceived acoustic feature vectors: GMM–HMM, and DNN–HMM to obtain improvement in performance on connected and continuous Punjabi speech corpus. Experimental setup shows a notable improvement of 4–5% and 1–3% (in connected and continuous datasets respectively).

Keywords Deep neural network (DNN) · Gaussian mixture model (GMM) · Hidden markov model (HMM) · Maximum likelihood linear transformation (MLLT) · Cepstral mean and variance normalization (CMVN) · Maximum likelihood linear regression (fMLLR)

1 Introduction

Speech recognition systems typically model the input uttered signal with their corresponding phones in two different steps of feature extraction and modeling classifiers. To increase the performance of the recognition system, researchers try to integrate or refine the extracted feature vector in first phase before employing them for classification.

Studies explored stochastic modeling techniques (SMT) e.g. HMM that requires the prior knowledge and experience for dealing with the issues in learning of classifiers complexity (Rabiner 1989) from integrated, refined or baseline feature vectors point of view. SMT faces the challenge of reliable parameter evaluation, modeling of speech units and prediction of omission observation that must be independent from each other. Implementation of HMM alone for training of the system did not generate promising results. So, researchers purposed many new or hybridization of classifiers (Palaz and Collobert 2015; Mitra et al. 2014; Sivasankaran et al. 2015) for calculation of its state emission probability (Juang et al. 1986). Machine learning algorithm in last four decades introduced a powerful tool of expectation maximization (EM)—a algorithm for training of HMM. EM helps to connect GMM that built a strong relation between HMM states and acoustic signa. Extensive efforts have also

✉ Amitoj Singh
amitoj.ptu@gmail.com

[1] Department of Computer Science & Engineering, Chitkara University Institute of Engineering & Technology, Chitkara University, Rajpura, Punjab, India

[2] Department of Electronics & Communication Engineering, Chitkara University Institute of Engineering & Technology, Chitkara University, Rajpura, Punjab, India

[3] Department of Computer Engineering, N.I.T. Kurukshetra, Kurukshetra, Haryana, India

[4] Department of Computer Application, M.R.S. P.T.U, Bathinda, Punjab, India

been made to avoid over fitting on available training data space (Hinton et al. 2012). The accuracy of GMM–HMM system is increased by embedding tandem and bottleneck features through neural networks (Hermansky et al. 2000; Bourlard and Morgan 1993). Still there is a vast scope for the improvement in the speech recognition systems because GMM shows serious shortcomings of inefficient modeling of data on or near to nonlinear manifold in existing data space. Consequently, many researchers replaced GMM with DNN that is capable of attaining better gains (Schmidhuber 2015). However, the current progress in DNN (Hinton et al. 2012) in parallel with hardware technology [through graphical processing unit (GPU)] for numeral calculations enables tackling of huge training task in large vocabulary ASR system. Earlier, the successful implementation on large vocabulary continuous speech recognizer using TIMIT phone corpora is has been demonstrated through DNN that displayed its power over GMM. The shared view research group of Microsoft, University of Toronto, IBM circle and others adopted DNN as an efficient modeling approach (Hinton et al. 2012). The significant gain of DNN is presented in Chen and Cheng (2014) that worked well for handling large child speech corpus, and in native or non-native Mandrian speech corpora. The later part of the paper is structured as follows: Sect. 2 describes the overview of Punjabi Speech recognition status and overview of some of the techniques like speaker adaptation, low rank LDA feature approach and acoustic classification approaches. A detailed description of proposed system overview is presented in Sect. 3. Finally, performance evaluations and conclusion are provided in Sects. 4 and 5.

2 Background

Punjabi language is considered as one of the most widely spoken languages of the modern Indo-Aryan Language group. Punjabi language is the 10th most widely spoken language worldwide. Punjabi language, in spite of being spoken by large community, can still be called as *less resource* language (LRL) due to the non-availability of resources like standard keyboard, recorded speech corpora, commercial speech recognition system etc therefore Punjabi language fails to compete with the other most widely spoken languages of the world. Research on Punjabi speech recognition is also not well traversed. In case of the Punjabi language some initiatives were taken by researchers who worked on speaker dependent/independent ASR system for small vocabulary (Kumar and Singh 2017). Dua et al. (2012) constructed a speaker dependent as well as speaker independent 115 isolated lexicon Punjabi ASR system that employed eight speakers with HMM technique for creation of its acoustic model. They added extension in their research work by connecting lexicons of previous 115 words on similar front and back end approaches (Dua et al. 2018). Ghai and Singh (2013) reported a continuous automatic speech recognition system using a baseline 100 sentences whose repetition was generated by nine speakers with a triphone modeling unit. Lata and Arora (2013) examined the impact of /h/ sound with the help of Praat and Matlab software on a particular dialect of Punjabi language such as malwa dataset. Mittal and Sharma (2014) investigated three mechanisms of data processing specifically for read, lecture and conversation utterances and it was analyzed with the help of continuous density HMM on HTK toolkit. Singh et al. (2015) has explained five Punjabi tonemes set up on their position and observed them using 150 exclusive lexicons gathered from ten speakers. Kadyan et al. (2017) proposes an approach for the generation of HMM parameters using two hybrid classifiers such as GA + HMM and DE + HMM. The proposed technique focuses on refinement of processed feature vectors after calculating its mean and variance In this paper, we complement two different contributions: in front-end approaches (MFCC or GFCC) different combination of speaker adaptive techniques are employed and secondly two hybrid HMM classifiers are applied to analyze the effect on the performance of the recognition system. Results of DNN–HMM classifier were compared with traditional GMM–HMM approach using monophone and triphone based context modeling. As expected, an efficient modeling classifier was obtained that achieved better results using robust feature extraction approach. The robust feature vectors are concatenated with feature reduction approach of LDA, speaker adaptation model such as adaptation and transformation technique of SAT like CMVN, MLLT and fMLLR. Overall, it is found that combined feature vector on DNN based hybrid HMM classifier improved the performance of Punjabi speech recognition system.

2.1 Acoustic speaker variability techniques

Various adaptation models can handle speaker variability. It is possible through two categories such as normalization and its adaptation. The first category projects the feature vector normalization and its transformation methods that include maximum likelihood linear transformation (MLLT) (Gales 1998), cepstral mean and variance normalization (CMVN) (Liu et al. 1993; Acero and Stern 1992), and linear discriminate analysis (LDA) (Haeb-Umbach and Ney 1992) for low rank feature projection. The second category involves model-space transformation approach such as maximum likelihood linear regression (MLLR) (Gales and Woodland 1996a).

2.1.1 MLLT

MLLT is used for decorrelation of features. It can be applied on acoustic features as a linear transformation to encapsulate the correlation among the components of its feature vectors. The transformation matrix $\left(W_j\right)$ is computed by maximizing its auxiliary function using Eq. (1).

$$Q\left(M,\hat{M}\right) = K - \frac{1}{2}\sum_{s=1}^{J}\sum_{t=1}^{T}\gamma_s(t) \left[K_s + \log\left(\left|\Sigma_s\right|\right) + \left(o_{t-}\hat{\mu}_s\right)\widehat{\sum_s^{-1}}\left(o_{t-}\hat{\mu}_s\right)^T\right] \tag{1}$$

The model parameters as well as its transformation parameters are optimized on training data through an objective function of maximum likelihood calculations. MLLT is implemented on top vectors of LDA approach and experimented on context independent HMM models through its training data. It is also known as speaker independent adaptation method (Matsoukas et al. 1997).

2.1.2 CMVN

Raw input speech signals are processed to generate acoustic features that are normalized using cepstral mean normalization and cepstral variance normalization approaches. CMVN initiates the transformation process from the feature without the necessity of the transcription or any other model parameters.

It helps in generation of zero mean and unit variance through sphere of the data. Consider a set of D-dimensional cepstral features (CO) that consist of a list of OL observation vectors (co_1, co_2, co_3,..,co_{ol}...., co_{OL}), such that m dimension of oLth frame of a mean normalized feature is calculated through relation of Eq. (2).

$$co_{lD}(m) = co_{lD}(m) - \mu(m) \tag{2}$$

The mean of observation vector $\mu(m)$ of oLth frame is calculated as in Eq. (3).

$$\mu(m) = \frac{1}{OL}\sum_{oL=1}^{oL} co_{lD}(m) \tag{3}$$

2.1.3 MLLR

A baseline acoustic model is necessary for any adaptive training dataset. It depends upon the mean parameter of a Gaussian mixture using a transformation matrix (W_j), its extended mean vector φ_j through relation of Eq. (4):

$$\mu_j = A\mu_j + b = W_j\varphi_j \tag{4}$$

It causes an issue of computational complexity for a full covariance matrix due to W_j, another important feature transformation technique is constrained on MLLR is fMLLR. It transforms the feature using maximum likelihood approach such as EM algorithm. The output of fMLLR is considered as an input to DNN as this approach has an advantage over MLLR/fMLLR adaptation of (Parthasarathi et al. 2015) due to:

- Faster parameter estimation using few EM iterations.
- Easy processing of a few minutes of audio files.
- Compensation of acoustic mismatch upto some level.
- Easy processing of a file with transcription error.

2.2 Low rank feature vector projection

The task of the feature extraction process in a speech recognition is to generate compact feature space that captures relevant information discrimination characteristics. All the information of a processed feature vector is stored in a fixed feature dimension. Each dimension removes less discriminate information within the vectors that can affect the classification and knowledge generation process in decoding module. This technique can be used to sustain the power of information discrimination while projecting the feature vector in subspace of lower dimensions. The process of generation of a new low dimensional feature space R^{M} after transformation of its original feature vector is calculated using Eq. (5).

$$F_M(T) = \theta_T F_n(T) \tag{5}$$

where $F_m(T)$ depicts the feature vector in a transformed space of feature, $F_n(T)$ denotes the original feature vector in real feature space, and M indicates the required dimensionality of its feature space. Numerous techniques are practiced in the past for feature decorrelation and its dimension reduction using PCA, LDA, HLDA in Kumar and Andreou (1998). In this study, LDA is employed for feature reduction. It is a statistical technique that increases the separability among different classes. It perform the linear transformation that convert n dimension into m dimensions such that m < n to increase the distance between inter class. The separated sub vector does not contain any classification information and its intra class variance is equal to each other.

2.3 Acoustic classifier approaches

2.3.1 GMM–HMM model

In practical, the acoustic model is trained using HMM technique to generate the observation probabilities. These probabilities are modeled using multivariate GMM technique (Povey et al. 2011). The learning of HMM for a word in vocabulary is done through basic unit such as phone. The phones are combined to generate the specific word information. These words are also used to generate the sentence and phrase related information after concatenating them. The multivariate Gaussian density function for a dimension D of observation feature vector is calculated through Eq. (6).

$$N\left(o_l, \emptyset_{pn},\ C_{pn}\right) = \frac{1}{\sqrt{(2\pi)^D\left|C_{pn}\right|}} \exp\{-\frac{1}{2}(O_l - \ \emptyset_{pn})^T \times C_{pn}^{-1}\left(O_l - \ \emptyset_{pn}\right)\} \tag{6}$$

where $\emptyset_{pn}$ depicts the mean and C_{pn} indicates the covariance matrix on nth Gaussian component of a pth state respectively. A number of other model parameters of HMM such as weight of mixtures as well as state transition probabilities are also trained using Baum–Welch re-estimation algorithm (Rabiner and Juang 1993).

2.3.2 Deep neural network model

The statistical GMM technique earlier faces the challenge of on or near the non-linear manifold issue in data space. It helps in modeling of the data. A single hidden layer can't resolve the issue in artificial neural network. DNN consists of many hidden layers (of various non-linear hidden units) connected to a number of output layer for tackling the issue of acoustic variability.It is a feed forward network where each hidden layer j employed the logistic function in Eq. (7).

$$y_j = logistic\left(x_j\right) = \frac{1}{1 + e^{-xj}}, \quad x_j = b_j + \sum_i y_i w_{ij} \tag{7}$$

where b_j depicts the bias for unit j using index i through their weight w_{ij} on unit i to j in layer below. These output layer will facilitate a number of HMM states. Each state modeled the phone on either side using the triphone modeling. A thousand of tied state will be output from the triphone HMM on a large set of dataset. It process large number of training set by operating on small batch instead of whole dataset before updating the value of weight according to the gradient. The value of gradient can be improved by a momentum coefficient value lies between $0 < \alpha < 1$ that helps in smoothening of gradient used for calculation of a minibatch (t). On other hand, DNN also helps in handling the issue of over-fitting for large training dataset by terminating the learning process before its performance become worse through large weight values. Initially small values are assigned to the weight of initial layer than its hidden layer to avoid attaining of same gradient for all the unit of hidden layer. A DNN generates an output of probabilities in the form of (HMM state by acoustic input) p_j as a class probability.

2.3.3 DNN–HMM model

DNN is adopted to calculate the posterior probabilities for senone (through adoption of context dependent tied state model) in an HMM based speech recognition. Consider a feature vector x_t of a context dependent window frame and applies non linear transformation on it through hidden layers of DNN. The technique of force-alignment is used to gain the senone label over a training data through conventional GMM–HMM system. The DNN parameters are handled by gradient descent function with the help of back propagation algorithm. It uses softmax output layer that consists of many node (number of classes) equal to its number of senones. The multi-class classification in DNN is possible through a softmax non-linearity relation of Eq. (8).

$$p\left(x_t\right) = \frac{\exp\left(x_t\right)}{\Sigma_k \exp\left(x_p\right)} \tag{8}$$

where p depicts the index for all classes. The senone likelihood $\left(p\left(x_t\right)\right)$ is employed in HMM through sequential characteristics of the speech modeling information. Finally the process of decoding is performed on converting the posteriors to scaled likelihood on test dataset. The output produces the recognized hypothesis on scaled likelihood.

3 System overview

MFCC and GFCC feature extraction techniques are used for building Punjabi ASR system. Figure 1 consist of two main sub-systems: the wholesystem is tested and trained through MFCC and GFCC approaches on monophone (M1), triphone with delta–delta training (M2), triphone with LDA + MLLT (M3), triphone with MLLT + SAT (M4) modeling units in GMM + HMM or DNN + HMM classifiers. Initially first sub-system is tested using monophone on GMM + HMM classifier, in the next case the output of monophone (M1) is provided as an input to triphone model (M2). To further improve the performance of combined output of M1 and M2, it is processed using LDA + MLLT (M3) method in GMM + HMM classifier.

In second sub-system, the issue of GMM + HMM system is overcame through addition of speaker variability technique

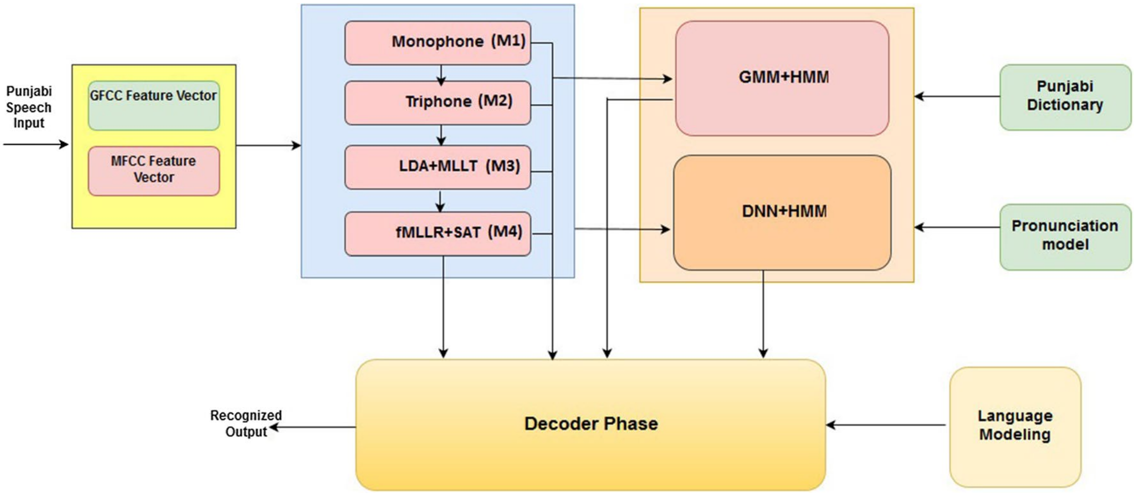

Fig. 1 Block diagram of MFCC and GFCC approaches integrated with monophone (M1), triphone (M2), LDA + MLLT (M3), and fMLLR + SAT (M4) modeling units based on GMM + HMM and DNN + HMM classifiers

such as fMLLR + SAT (M4) with M3 on DNN + HMM classifier.

3.1 System implementation using GMM–HMM based acoustic modeling

The input speech signal is processed through baseline MFCC or GFCC methods to calculate the 13 static features + energy with their first and second order temporal derivatives. These techniques help in filtering the raw signal by discarding irrelevant information present in the speech signal. Moreover, this also prevents the vectors that are used to discriminate in modeling phase. The MFCC and GFCC generate 13 dimensional features sliced in nine frames (with ± 4). On top of it, LDA technique is applied to reduce it into 40 features. Finally the process of MLLT is implemented to do decorrelation among reduced features. The training of these vectors is integrated with GMM–HMM classifiers. A tri-phone based acoustic modeling is employed with state tying decision tree. Each tri-phone model uses three state of HMM through 16 diagonal covariance Gaussian and 32 diagonal covariance matrix is used for silence and short pauses. For an observation sequence O, the most likely model is determined using Baye's rule. It helps in the determination of maximum probability of the observation through hidden states. The maximum probability is calculated over the entire trained model. In decoding phase Viterbi algorithm (Kadyan et al. 2017) is used to traverse the path that has maximum probability through hidden state sequence for desired observation sequence.

3.2 System implementation using DNN–HMM based acoustic modeling

The extracted feature vector from MFCC or GFCC is further normalized using CMVN approach. Operations are performed on static feature vectors to produce zero mean and unit variance that helps in reducing speaker variability and additive noise induced in the channel due to environment. The produced vectors are passed through LDA technique for further reduction of feature vectors to 40 dimensions. Reduced feature vectors are accurately modeled using diagonal covariance Gaussian through a feature orthogonal transformation approach of MLLT. The process of normalization on speaker is performed by fMLLR that uses the dimension of 40 × 41 parameters and is calculated on speaker adaptive trainingon GMM dependent system. The DNN technique is implemented with the assistance of Kaldi toolkit and maximum accuracy is obtained at layer 5. Other parameter such as learning rate i.e. a matrix-value is fixed to 0.015 with number of epochs is equal to 20. The minibatch size can be varied as 512, and 1024.

4 Experimental results

System training for the entire dataset is performed with GMM–HMM, and DNN–HMM approaches. Kaldi toolkit 4.3.11 is used to implement two modeling classifiers. Four sets of combinations (M1–M4) on extracted feature vectors are employed using two type of datasets recorded in clean and real environment: connected speech corpus (dataset1) and continuous speech corpus (dataset2). The dataset1 consisted of 21,764 connected words utterances from 13 speakers. The dataset2 consisted of utterances from 13 speakers (including 6 male and 7 female). It is framed on 422 unique phonetically rich sentences and produces a total of 3611 sentences in training phase of the system. It includes both type of speaker that provides good and decays the performance recognition result. The average length of each speaker utterance is around 3–7 s that covers 4–9 words. The corpus is recorded after providing of text transcription that makes it a read speech corpus. The speech output is reported in word error rate (WER) for Punjabi connected and continuous corpora.

4.1 Experimental setup

The evaluation of proposed method with different combination on extracted feature vectors are performed before classification. Extracted features are then processed through LDA, SAT, fMLLR and MLLT methods using triphone and monophone models. Testing of the proposed Punjabi-ASR system is performed using tenfold cross validation where 10% of the development trained data is kept for testing and rest part is involved in training of the system.

4.1.1 Words recognition on different modeling units

Initially the input speech signal is fed to context independent (CI) monophone model in M1 system through GMM–HMM approach. A total of 120 num pdf are used in CI model. The system achieves low accuracy on CI model. To improve the performance on large vocabulary Punjabi-ASR system, context independent models (triphone) are used in different combination from M2 to M4 system. Different numbers of num pdf are used in triphone models which are employed in four sub-systems. We analyze the robust feature vectors on two different acoustic modeling classifiers such as GMM–HMM and DNN–HMM. Exhaustive studies of two different corpuses are performed to analyze word error rate on dataset1 and dataset2 as depicted in Table 1. On large dataset DNN gives better performance than GMM on mismatched train and test conditions. The monophone based model shows high WER on both the dataset. WER of 5.32% is obtained in continuous sentences of dataset2 and 39.73% on dataset1 using connected words. The system is tested with speaker independent corpus where no testing speaker is involved in training of the system.

Table 1 The WER (%) profile obtained through the baseline MFCC and GFCC approaches with integration of M1 to M4 modeling units on two acoustic classifiers

System type	Feature type		
	GFCC	MFCC	MFCC
	Dataset2		Dataset1
M1 + GMM–HMM	46.3	14.11	46.74
M2 + GMM–HMM	82.16	19.14	46.71
M3 + GMM–HMM	73.39	13.08	46.09
M4 + GMM–HMM	34.4	7.01	44.94
M4 + DNN–HMM	24.67	5.22	39.73

The word error rate is calculated for dataset2 using MFCC feature approach at 44 kHz or GFCC at 16 kHz and dataset1 at 44 kHz using MFCC approach

4.1.2 Effect of varying feature dimension with LDA approach

DNN based Punjabi ASR system is found to be more effective than GMM models. The normalization techniques help in WER reduction with these modeling approaches. For analyzing the varying feature dimension, the feature values are varied from 8, 12, 16, 32, and 39 as shown in Fig. 2.

The time-spliced MFCC features are found to be more beneficial than baseline MFCC feature vectors. LDA implemented in the initial stages of HMM training helped in reducing the feature vectors from 117 to 40 entries that increase the different class seprability. Furthermore, MLLT is applied verbosely on LDA feature vectors. It is employed on context dependent feature vector of a triphone acoustic model. The recognition performance on DNN–HMM subsystem is found to be superior at feature dimension value of 16 on front-end combination of triphone with MLLT + SAT. The Fig. 2a conceived low WER at feature value of 16 in MFCC approach on dataset1 and feature value of 32 in MFCC, and feature value of 16 in GFCC on dataset2 of Fig. 2b.

4.1.3 Effect of varying hidden layer

For the similar test and train conditions, a better performance is obtained with an improvement of 4–5% and 1–3% (in connected and continuous datasets respectively) with DNN than GMM–HMM models. Consecutively analysis is carried out by varying the number of hidden units and layer in training of DNN based ASR model. The normalized features are used in DNN–HMM training on non-linear hidden layers using tanh function. A number of hidden layers with 1–7 values are varied with number of unit per layer is equal to 512, 1024, 2048 and 3074 respectively. The training is provided with number of frames of acoustic feature data in its input layer as 7, 11, 15, 27, and 37. It will help in analyzing the system from shallow network to deep network as shown in Table 2. Finally the minibatch size is fixed to a value of 512 at the end that attain maximum word accuracy. The dataset1 receives maximum word accuracy at DNN layer value of 6, but dataset2 attain at a value of 5 in Table 2.

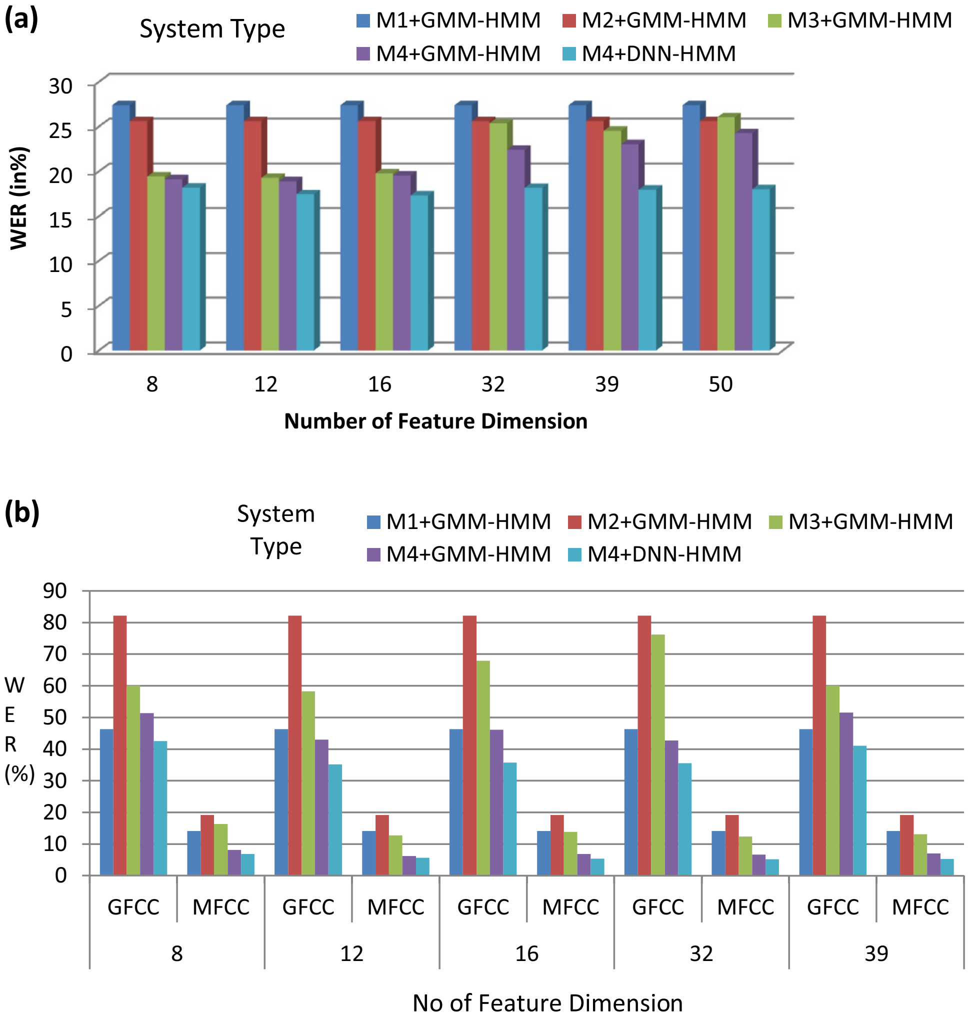

Fig. 2 The WER profile for varying feature dimension obtained through the baseline MFCC and GFCC feature vectors on different system type using two acoustic classifiers. The word error rate is calculated for dataset1 in (**a**) using MFCC approach and dataset2 in (**b**) in both feature approaches

4.1.4 Effect of varying Gaussian mixture

The experiments are also tried on category of varying Gaussian mixture as shown in Fig. 3. It can be observed that initially for all varied value for Gaussian system performance does not change either in context of monophone or triphone model but as the context model is combined with speaker adaptation model its performance get affected. Maximum word accuracy is obtained with a Gaussian mixture of value 16 with DNN–HMM model on dataset2 using MFCC feature extraction technique and GFCC approach produced better result at a very low value of Gaussian mixture. The datset1 achieved a low WER at an optimal value of 64 for Gaussian mixture on MFCC approach.

This paper made an attempt to implement DNN based HMM model to overcome the issue of overfitting on large Punjabi corpora. In proposed Punjabi ASR system, initially a number of experiments are performed with different modeling units (Table 1), analyzing the number of feature dimension in GMM–HMM or DNN–HMM acoustic classifiers (Fig. 2a, b)

Table 2 The WER profile for varying number of hidden layer obtained through the baseline MFCC with integration of M4 modeling unit

Classifier type	Number of hidden layer						
	1	2	3	4	5	6	7
M4 + DNN–HMM (dataset1)	18.01	17.97	17.85	17.58	17.66	**17.53**	17.54
DNN (dataset2)	5.32%	5.41%	5.32%	5.41%	**5.22%**	5.29%	5.51%

All the mentioned experiments are performed on acoustical mismatched trained and test conditions. The word error rate is calculated for dataset1 dataset2 at 44, 16 kHz

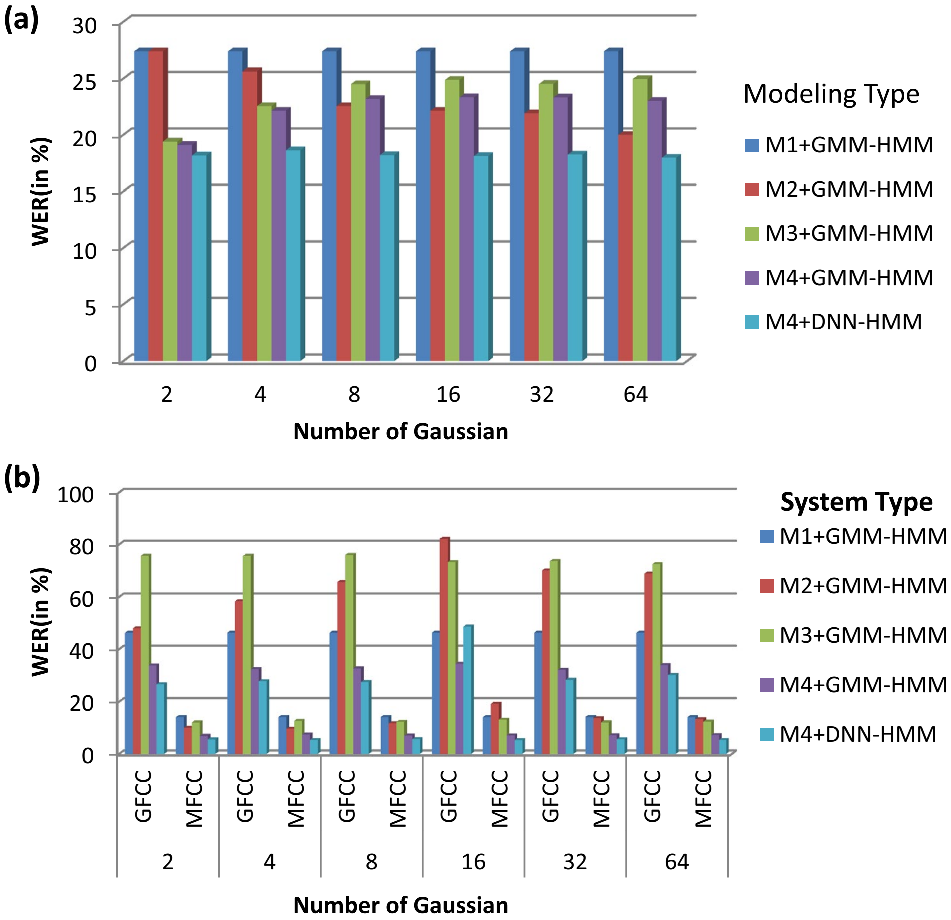

Fig. 3 The WER profile for varying Gaussian mixture obtained through the baseline MFCC and GFCC techniques with integration of context dependent and independent modeling units on two acoustic modeling type for dataset1 in (**a**) using MFCC feature approach only and dataset2 in (**b**) employed both feature vector approaches

along with varying number of hidden layer (Table 2) and their units in DNN based HMM model. Finally number of Gaussian are varied to identify the effect of these variations (Fig. 3a, b). The experiments are performed to obtain an optimal value of parameters. These parameters increase the recognition accuracy after adopting M1–M4 on context model (monophone and triphone) with speaker adaption model on MFCC and GFCC feature vectors. The MFCC technique shows WER performance improvement in comparison to GFCC approach on DNN–HMM classifiers. So after analyzing dataset2 on both the feature vector approaches, the dataset1 is finally demonstrated with MFCC approach that showcase better result on DNN–HMM classifiers.

5 Conclusions

This paper presented the speaker adaptive technique on different acoustic modeling approaches like GMM–HMM and DNN–HMM that are used to reduce the affect of acoustic mismatch between train and test conditions. The issue of overfitting of training data is handled using DNN–HMM model instead of GMM–HMM hybrid modeling. The experiments are performed on two different Punjabi speech corpus i.e. connected words and continuous sentences. Different acoustic modeling approaches on robust feature vectors are analyzed through combination of low rank feature projection and speaker variability techniques. The combination of CMVN normalization and MLLT feature transformation is performed on LDA features in baseline MFCC and GFCC acoustic feature extraction approaches. These approaches yielded performance improvement of 4–5% and 1–3% (in connected and continuous datasets) with DNN–HMM than GMM–HMM approach. Further work can be extended for integration of recurrent neural network approaches in acoustic and language modeling for speech recognition enhancement.

Acknowledgements This work is partially tested on the sample Punjabi corpus collected for Language Resources for Auditory impaired Person project from IEEE SIGHT. The views and results in the work is as per perspective of the research. The author would like to thank Speech and Multimodel Laboratory members Mandeep, Sashi, Nikhil at Chitkara University Punjab. Special thanks to Dr. Syed who helps in providing valuable input in formation of a baseline DNN system.

References

Acero, A., & Stern, R. M. (1992). Cepstral normalization for robust speech recognition. In *Speech processing in adverse conditions*.

Bourlard, H., & Morgan, N. (1993). *Connectionist speech recognition. A hybrid approach* (Vol. 247). Boston: The Kluwer International Series in Engineering and Computer Science.

Chen, X., & Cheng, J. (2014). Deep neural network acoustic modeling for native and non-native Mandarin speech recognition. In *Proceedings of ISCSLP* (pp. 6–9).

Dua, M., Aggarwal, R. K., & Biswas, M. (2018). GFCC based discriminatively trained noise robust continuous ASR system for Hindi language. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-018-0828-x.

Dua, M., Aggarwal, R. K., Kadyan, V., & Dua, S. (2012). Punjabi automatic speech recognition using HTK. *International Journal of Computer Science Issues, 9*(4), 359–364.

Gales, M., & Woodland, P. (1996a). Mean and variance adaptation within the MLLR framework. *Computer Speech & Language, 10*, 249–264.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language, 12*(2), 75–98.

Ghai, W., & Singh, N. (2013). Continuous speech recognition for Punjabi language. *International Journal of Computer Applications, 72*(14), 422–431.

Haeb-Umbach, R., & Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Acoustics, speech, and signal processing, 1992. ICASSP-92., 1992 IEEE international conference on* (Vol. 1, pp. 13–16). IEEE.

Hermansky, H., Ellis, D. P., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *Acoustics, speech, and signal processing, 2000. ICASSP'00. Proceedings. 2000 IEEE international conference on* (Vol. 3, pp. 1635–1638). IEEE.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97.

Juang, B. H., Levinson, S., & Sondhi, M. (1986). Maximum likelihood estimation for multivariate mixture observations of markov chains (corresp.). *IEEE Transactions on Information Theory, 32*(2), 307–309.

Kadyan, V., Mantri, A., & Aggarwal, R. K. (2017) Refinement of HMM model parameters for Punjabi Automatic Speech Recognition (PASR) System, *IETE Journal of Research*. https://doi.org/10.1080/03772063.2017.1369370.

Kadyan, V., Mantri, V., & Aggarwal, R. K. (2017) Refinement of HMM model parameters for Punjabi Automatic Speech Recognition (PASR) System. *IETE Journal of Research*. https://doi.org/10.1080/03772063.2017.1369370.

Kumar, N., & Andreou, A. G. (1998). Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication, 26*(4), 283–297.

Kumar, Y., & Singh, N. (2017). An automatic speech recognition system for spontaneous Punjabi speech corpus. *International Journal of Speech Technology, 20*(2), 297–303.

Lata, S., & Arora, S. (2013) Laryngeal tonal characteristics of Punjabi—An experimental study. In *2015 2nd international conference on computing for sustainable global development* (pp. 1694–1697).

Liu, F., Stern, R. M., Huang, X., & Acero, R. (1993). Efficient cepstral normalization for robust speech recognition. In *Proceedings of the workshop on human language technology* (pp. 69–74).

Matsoukas, S., Schwartz, R., Jin, H., & Nguyen, L. (1997). Practical implementations of speaker-adaptive training. In *DARPA speech recognition workshop*.

Mittal, S., & Sharma, R. K. (2014). Development of phonetic engine for Punjabi language (Doctoral dissertation), Thapar University, Patiala, India.

Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., & Graciarena, M. (2014). Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In *Fifteenth annual conference of the international speech communication association*.

Palaz, D., & Collobert, R. (2015). Analysis of cnn-based speech recognition system using raw speech as input (No. EPFL-REPORT-210039). Idiap.

Parthasarathi, S. H. K., Hoffmeister, B., Matsoukas, S., Mandal, A., Strom, N., & Garimella, S. (2015). fMLLR based feature-space speaker adaptation of DNN acoustic models. In *Sixteenth annual conference of the international speech communication association*.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., & Rose, R. C. (2011). The subspace Gaussian mixture model—A structured model for speech recognition. *Computer Speech & Language, 25*(2), 404–439.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Upper Saddle River: Prentice-Hall Inc.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Network, 61*, 85–117.

Singh, A., Dipti, P., & Agrawal, S. S. (2015) Analysis of Punjabi tonemes. In *Computing for Sustainable Global Development (INDIACom)* (pp. 1–6).

Sivasankaran, S., Nugraha, A. A., Vincent, E., Morales-Cordovilla, J. A., Dalmia, S., Illina, I., et al. (2015). Robust ASR using neural network based speech enhancement and feature simulation. In *IEEE workshop on automatic speech recognition and understanding (ASRU), 2015* (pp. 482–489).