ON COMBINING DNN AND GMM WITH UNSUPERVISED SPEAKER ADAPTATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Shilin LIU, Khe Chai SIM

School of Computing, National University of Singapore, Singapore
{shilin,simkc}@comp.nus.edu.sg

ABSTRACT

Recently, context-dependent Deep Neural Network (CD-DNN) has been found to significantly outperform Gaussian Mixture Model (GMM) for various large vocabulary continuous speech recognition tasks. Unlike the GMM approach, there is no meaningful interpretation of the DNN parameters, which makes it difficult to devise effective adaptation methods for DNNs. Furthermore, DNN parameter estimation is based on discriminative criteria, which is more sensitive to label errors and therefore less reliable for unsupervised adaptation. Many effective adaptation techniques that have been developed and proven to work well for GMM/HMM systems cannot be easily applied to DNNs. Therefore, this paper proposes a novel method of combining DNN and GMM using the Temporally Varying Weight Regression framework to take advantage of the superior performance of the DNNs and the robust adaptability of the GMMs. This paper addresses the issue of incorporating the high-dimensional CD-DNN posteriors into this framework without dramatically increasing the system complexity. Experimental results on a broadcast news large vocabulary transcription task show that the proposed GMM+DNN/HMM system achieved significant performance gain over the baseline DNN/HMM system. With additional unsupervised speaker adaptation, the best GMM+DNN/HMM system obtained about 20% relative improvements over the DNN/HMM baseline.

Index Terms— Gaussian mixture model, Deep Neural Network, Speaker Adaptation

1. INTRODUCTION

Context-dependent Deep Neural Network (CD-DNN) [1] has been reported to outperform various conventional Gaussian Mixture Models (GMM) [2] based Hidden Markov Model (HMM) [3] systems by a large margin for many large vocabulary continuous speech recognition (LVCSR) tasks [4, 5]. DNN uses a long span of acoustic features as input so that both rich inter-frame and intra-frame information can be modelled for better discrimination. The multiple layers of nonlinear transformation allows the complex relationship between the acoustic features and the context-dependent HMM states to be effectively learned. However, unlike the GMM approach where each triphone state is represented by a GMM, a single DNN is used to simultaneously predict the posterior probabilities of *all* the states. It is difficult to interpret the DNN parameters in a meaningful manner. There is no clear and effective way of adapting the DNN parameters. Moreover, the DNN parameters are typically estimated discriminatively using the cross-entropy criterion which is more sensitive to label errors. By contrast, many advanced adaptation techniques, such as Maximum Likelihood Linear Regression (MLLR) [6] and Maximum A Posteriori (MAP) [2], have been developed and shown to work well for the GMM/HMM systems. In particular, these methods are based on the generative training paradigm, which is more robust for unsupervised adaptation.

Various indirect approaches have been proposed to take advantage of both the GMMs and DNNs. Some researchers suggested using the DNN to extract better discriminative features, such as the tandem features [7, 8, 9, 10]. In order to develop feasible tandem features for GMM training, the high dimensional CD-DNN posteriors have to be projected to a lower dimension, which inevitably causes information loss. Performance degradation of the tandem systems using ML training has been observed in multiple reports [8, 10]. Further, discriminative training and unsupervised speaker adaptation have been successfully applied to the tandem systems, which achieved superior performance compared to the hybrid DNN/HMM system [8]. Others have also proposed to use the adapted acoustic features based on the fMLLR transformation obtained from the GMM system [11] or another adaptation NN using speaker code information [12].

In this paper, GMM+DNN/HMM is proposed as a novel system that combines the GMM and DNN using the Temporally Varying Weight Regression (TVWR) framework [13, 14]. Based on this framework, a regression model is trained to transform the DNN posteriors into the time-varying scaling factors for the Gaussian weights. However, directly incorporating the high-dimensional CD-DNN posterior features will lead to a substantial increase in the number of regression parameters. This paper will present some solutions to address this issue.

The rest of this paper is organized as follows. Section 2 reviews the prior works about GMM and DNN. Section

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

3 introduces the proposed GMM+DNN/HMM system. Section 4 presents solutions to address the issue of incorporating the high-dimensional CD-DNN posteriors efficiently without dramatically increasing the model complexity. Finally, the experimental results are reported in Section 5.

2. PRIOR WORKS

For decades, GMM has been used as the representation of the HMM state distribution due to its efficient training and decoding algorithms [2]. In the conventional GMM/HMM system, the state emission probability is given as:

$$p(\mathbf{o}_t|j) = \sum_{m=1}^{M} c_{jm} p(\mathbf{o}_t|j,m) \quad (1)$$

where j is the HMM state, $\mathbf{o}_t$ is the observation at time t, M is the number of Gaussian components per state, c_{jm} is the static component weight and $p(\mathbf{o}_t|j,m)$ is a Gaussian distribution. Due to the observation independence assumption and the use of diagonal covariance matrices for the Gaussian components (for better efficiency), the inter-frame and intra-frame correlations are poorly modelled by the GMM/HMM systems, which limits the performance to some extent. Nevertheless, effective adaptation techniques, such as the MLLR [6], have been developed to achieve reliable unsupervised speaker adaptation.

On the other hand, DNN is a general purpose machine learning model that is capable of learning the complex non-linear function to map a long span of acoustic features into high quality CD state posterior probabilities. The state probability of a DNN/HMM hybrid system is given as:

$$p(\boldsymbol{\xi}_t|j) \propto \frac{p(j|\boldsymbol{\xi}_t)}{P(j)}, \quad \boldsymbol{\xi}_t = \{\mathbf{o}_{t-\delta} \dots \mathbf{o}_t \dots \mathbf{o}_{t+\delta}\} \quad (2)$$

where $P(j)$ and $p(j|\boldsymbol{\xi}_t)$ are the prior and posterior probability of state j respectively and $\boldsymbol{\xi}_t$ is the long span acoustic features. Unfortunately, the MLLR adaptation method cannot be easily applied to adapt the DNNs. There are currently several workarounds reported in the literatures including 'borrowing' MLLR transforms estimated using the GMM/HMM systems to adapt the acoustic features [11], introducing speaker code to adapt the parameters in the first few layers of a DNN [12] and using the DNN posterior features to train the tandem systems [7, 8, 9, 10] which can then be adapted using standard techniques like MLLR.

3. COMBINING GMM AND DNN

As previously mentioned, the DNNs are able to predict high quality discriminative CD posteriors while the GMMs can be reliably adapted in an unsupervised manner using MLLR. Therefore, to get the best of both worlds, this paper proposes

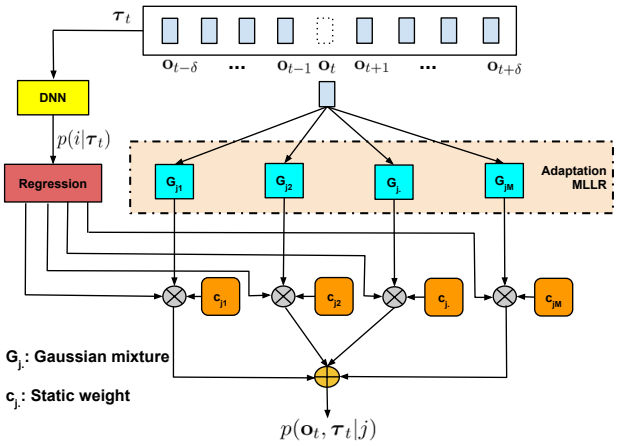

Fig. 1. An schematic diagram showing the state output probability function of the proposed GMM+DNN/HMM system.

combining the GMMs and DNNs using the TVWR framework [13, 14]. According to this framework, the state output probability of the long span acoustic features is given as:

$$p(\boldsymbol{\xi}_t|j) \propto \sum_{m=1}^{M} c_{jm} \sum_{i=1}^{N} P(i|\boldsymbol{\tau}_t) P(i|j,m) p(\mathbf{o}_t|j,m) \quad (3)$$

where $\boldsymbol{\tau}_t = \{\mathbf{o}_{t-\delta} \dots \mathbf{o}_{t-1}, \mathbf{o}_{t+1} \dots \mathbf{o}_{t+\delta}\}$ denotes the contexts of the current observation, i is the latent variable to partition the acoustic space, $P(i|\boldsymbol{\tau}_t)$ is the posterior feature, $P(i|j,m)$ is the regression parameter. M and N correspond to the number of Gaussian components and the number of latent variables, respectively. As shown in Figure.1, the long span acoustic feature is decomposed into two parts. Firstly, the regular-sized observation $\mathbf{o}_t$ is modelled by the conventional Gaussian components, where MLLR adaptation can be easily applied. Secondly, the latent variable i is associated with the clustered CD states so that $P(i|\boldsymbol{\tau}_t)$ can be predicted using a DNN. However, directly incorporating the high-dimensional CD-DNN posteriors (in the order of thousands) leads to a large number of regression parameters, $P(i|j,m)$. This will result in expensive computation and may cause over-fitting. In the next section, two solutions will be presented to address this issue without compromising the model efficiency.

4. REGRESSION OF CD-DNN POSTERIORS

In order to maintain a reasonable number of regression parameters, the high-dimensional CD posterior features have to be *projected* down to a lower dimension. In the following, two solutions will be presented to achieve this. The first solution attempts to reduce the number of regression weights via parameter tying. The latent variables, i, are clustered into

groups so that Eq. 3 can be rewritten as:

$$p(\boldsymbol{\xi}_t|j) \propto \sum_{m=1}^{M} c_{jm} \sum_{g=1}^{G} P(g|\boldsymbol{\tau}_t) P(g|j,m) p(\mathbf{o}_t|j,m) \quad (4)$$

where $P(g|j,m)$ is the regression weight for group g and the posterior probability of g is given by:

$$P(g|\boldsymbol{\tau}_t) = \sum_{i \in g} P(i|\boldsymbol{\tau}_t) \quad (5)$$

Therefore, tying the regression weights into groups is equivalent to projecting the posterior probabilities according to Eq. 5. In this work, the groupings are chosen to correspond to the monophone states and the resulting group posteriors are simply the CI posteriors. However, such projection operation may lose valuable context information such that we no longer preserve the superior performance of the CD-DNN model. In order to incorporate richer context information without dramatically increasing the model complexity, multi-stream TVWR [15] is used to integrate multiple sets of posterior features. The resulting state probability is given as:

$$p(\boldsymbol{\xi}_t|j) \propto \sum_{m=1}^{M} c_{jm} \prod_{c=1}^{C} \sum_{g_c=1}^{N_c} P(g_c|\boldsymbol{\tau}_t) P(g_c|j,m) p(\mathbf{o}_t|j,m)$$

where the CD posteriors are now factorized into C groupings and N_c is the number of groups in the c^{th} stream. In this work, we used three streams, one for the centre monophone states (g_2) and the other two for the left (g_1) and right (g_3) contexts. Although multiple DNNs can be trained for each stream, temporal context expansion can be applied to obtain the left and right context stream. Therefore, the posterior probabilities of the left/right contexts are derived from the centre-phone state:

$$P(g_1|\boldsymbol{\tau}_t) = P(g_2|\boldsymbol{\tau}_{t-\Delta_l}) \quad \text{and} \quad P(g_1|\boldsymbol{\tau}_t) = P(g_2|\boldsymbol{\tau}_{t+\Delta_r})$$

where Δ_l and Δ_r are the smallest positive values such that the states corresponding to the largest $P(g_2|\boldsymbol{\tau}_{t-\Delta_l})$ and $P(g_2|\boldsymbol{\tau}_{t+\Delta_r})$ are different from state with largest $P(g_2|\boldsymbol{\tau}_t)$.

The second CD posterior projection method adopts a *sparse* regression model where only a smaller set of CD posteriors is used for each Gaussian component. Considering that the TVWR formulation has the constraint $\sum_{i=1}^{N} P(i|j,m) = 1$, many of the regression parameters could be very small, especially when N is large. Hence, there may be only a small fraction of the parameters that actually contribute to the regression. The objective is to perform the sparse regression using only the most important parameters:

$$p(\boldsymbol{\xi}_t|j) \propto \sum_{m=1}^{M} c_{jm} \sum_{i \in I_j} P(i|\boldsymbol{\tau}_t) P(i|j,m) p(\mathbf{o}_t|j,m) \quad (6)$$

where I_j denotes the set of *active* latent variables for state j. Intuitively, I_j can be chosen such that:

$$I_j = \{i : P(i|j) > \nu\} \quad (7)$$

where $P(i|j)$ is computed as

$$P(i|j) = \frac{1}{|T_j|} \sum_{t \in T_j} P(i|\boldsymbol{\tau}_t) \quad (8)$$

T_j is the set of frames where j is the reference state and $|.|$ denotes the cardinality. ν is a threshold that can be adjusted to control the number of active posteriors. The regression parameters are initialised as $P(i|j,m) = P(i|j)$ if $i \in I_j$ and zero otherwise. In this work, we chose $\nu = 1/N$.

5. EXPERIMENTAL RESULTS

The experiments were conducted on the Topic Detection and Tracking - Phase 3 (TDT3) corpus for English broadcast news transcription task. After preprocessing of the data, which includes removing non-speech segments, text normalization and audio segmentation, approximately 100 hours of speech data are retained for acoustic model training. The evaluation task is a 58k open vocabulary broadcast news transcription task taken from the F0 portion of the 1997 Hub-4E Benchmark Test, which consists of about 3 hours of speech data and 49 speakers. The decoding language model is obtained by interpolating two language models trained on the TDT3 transcriptions and the Gigaword English corpus respectively.

The acoustic features are the 39 dimensional PLP coefficients (12 static coefficients, an energy term and the first two derivatives) with utterance-based cepstral mean normalization. The GMM/HMM baseline system is a decision tree state clustered triphone system with 4451 tied states. Each triphone is modelled by a 3-state left-to-right HMM and each state is modelled by a 20-component GMM. All the GMM parameters are trained using the maximum likelihood (ML) criterion. For DNN training, 10 hours of the training data are held out as the cross validation set and the rest are used for training. The input to the DNN is a 15-frame PLP features. The DNN has 5 hidden layers and each hidden layer has 2048 units. The output layer of the DNN corresponds to the 3052 tied states of a GMM/HMM system. The recognition is performed with a bigram full decoding followed by a trigram lattice rescoring. To build the tandem system, principle component analysis (PCA) is used to project the log posterior probabilities of the CI states into 13-dimensional features, which are then appended to the 39-dimensional PLP features. Then, the 52-dimensional tandem system is estimated re-estimation with four Baum-Welch iterations using the ML criterion. Unsupervised speaker adaptation is performed during testing, where one global constrained MLLR transformation is estimated for each speaker using the transcriptions recognised by the unadapted system. The same set of transforms are also used to adapt the acoustic features for the DNN. Note that these transformations were obtained by using only one iteration of Baum-Welch estimation.

Table 1 shows the performance of various baseline systems with or without unsupervised speaker adaptation. With-

System	Adaptation	
	None	MLLR
GMM/HMM	23.4	18.7
Tandem	19.3	16.9
CD-DNN/HMM	**14.5**	18.8

Table 1. *Word Error Rate (WER%) of various baseline systems with or without unsupervised speaker adaptation.*

out adaptation, the CD-DNN/HMM system obtained 8.9% absolute WER reduction over the ML trained GMM/HMM system. Similarly, the tandem system achieved 4.1% absolute improvement over the baseline GMM/HMM system due to the additional tandem features derived from the CI posteriors. However, the tandem system does not perform as well as the CD-DNN/HMM system. This is somewhat expected due to the following reasons: 1) the tandem features are obtained from only the CI posteriors; 2) the information loss due to PCA projection; and 3) the ML parameter estimation of the tandem system. After performing unsupervised speaker adaptation, significant absolute WER reductions of 4.7% and 2.4% were obtained for the GMM/HMM and the tandem systems respectively. However, simply 'borrowing' the constrained MLLR transforms to perform adaptation during decoding is not reliable and a substantial performance degradation of 4.3% absolute is observed. This is probably because we do not consider speaker adaptive training in this work and the MLLR adaptation is only applied to the test data.

Table 2 shows the performance of the GMM+DNN/HMM systems with different configurations. When using only the CI posteriors without context expansion, the unadapted system gives 16.2% WER, which is 1.7% behind the CD-DNN/HMM system due to the lack of context information. With context expansion, the performance improved to 13.3%, which is 1.2% better than the baseline. On the other hand, applying sparse regression with the CD posteriors achieves 14.6% WER without context expansion and 13.1% WER with context expansion. This results indicate the potential of performing sparse regression by selecting state-dependent active posteriors to reduce the model complexity. Furthermore, performing unsupervised speaker adaptation yields consistent performance improvements of 0.8% – 3.0% for all the GMM+DNN/HMM systems. This shows that the GMM+DNN/HMM systems are able to exploit the adaptability of the GMMs to obtain further improvements. The best performance of 11.6% WER is obtained using the CI posteriors with context expansion, which translates to approximately 20% relative improvement compared to the baseline DNN/HMM system. Although using the CD posteriors with context expansion gives slightly better results than using CI posteriors with context expansion for the updated systems, the former did not achieve as large a performance improvement, probably because the richer information from the DNN

Posteriors	Context Expansion	Regression Parameters	Adaptation	
			None	MLLR
CI	No	120	16.2	13.2
	Yes	360	13.3	11.6
CD	No	137	14.6	12.5
	Yes	437	13.1	12.3

Table 2. *Comparison of the number of regression parameters per Gaussian component and the WER (%) performance of various GMM+DNN/HMM systems with or without context expansion and unsupervised speaker adaptation.*

posteriors has somewhat de-weighted the importance of the GMMs. As a result, there is less impact from applying MLLR adaptation to the GMMs.

Finally, we analyse the number of regression parameters per Gaussian component for the various GMM+DNN/HMM systems. If the raw CD posteriors were directly used, there will be 3052 regression weights per Gaussian component. By using different projection methods, the number of regression weights can be reduced to about an order magnitude smaller. It is worth noting that the preliminary results presented in this paper have considered only the straightforward projection configurations. It may be possible to achieve further performance gains by adjusting the regression model complexity (*e.g.* using groupings other than the CI groups or using a different threshold, ν, for the sparse regression).

6. CONCLUSION AND FUTURE WORK

This paper has proposed combining the GMM and DNN models using the Temporally Varying Weight Regression (TVWR) framework to achieve a high quality and adaptable state probability model for automatic speech recognition. The resulting GMM+DNN/HMM system is different from the tandem systems in that the GMMs are trained directly on the cepstral acoustic features, rather than the DNN-derived tandem features. This paper has focused on addressing the issue of incorporating the high dimensional CD-DNN state posteriors into the TVWR framework without dramatically increase the system complexity. Specifically, projected CI state posteriors, sparse regression and context expansion are introduced to mitigate the problem. Experimental results show that the proposed GMM+DNN/HMM system outperform the baseline DNN/HMM system. In additional, applying unsupervised speaker adaptation can further improve the performance of the proposed system. Future work will consider applying speaker adaptive training and discriminative training to the GMM+DNN/HMM system.

7. REFERENCES

[1] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic Modeling using Deep Nelief Networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[2] J-L Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[3] L. 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.

[4] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[5] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, 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.

[6] Mark JF Gales and PC Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[7] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*. IEEE, 2000, vol. 3, pp. 1635–1638.

[8] Zhi-Jie Yan, Qiang Huo, and Jian Xu, "A Scalable Approach to Using DNN-Derived Features in GMM-HMM Based Acoustic Modeling For LVCSR," in *Interspeech*. ISCA, 2013, pp. 1–1.

[9] František Grézl, Martin Karafiát, Stanislav Kontár, and J Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *ICASSP*. IEEE, 2007, vol. 4, pp. 757–760.

[10] Dong Yu and Michael L Seltzer, "Improved Bottleneck Features Using Pretrained Deep Neural Networks," in *INTERSPEECH*, 2011, pp. 237–240.

[11] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[12] Ossama Abdel-Hamid and Hui Jiang, "Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition based on Discrimative Learning of Speaker Code," in *ICASSP*. IEEE, 2013, pp. 7942 – 7946.

[13] Shilin Liu and Khe Chai Sim, "Implicit trajectory modelling using temporally varying weight regression for automatic speech recognition," in *ICASSP*, 2012, pp. 4761–4764.

[14] Shilin Liu and Khe Chai Sim, "Temporally Varying Weight Regression: A Semi-Parametric Trajectory Model for Automatic Speech Recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[15] Shilin Liu and Khe Chai Sim, "Multi-stream Temporally Varying Weight Regression for Cross-lingual Speech Recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–1.