# DEEP NEURAL NETWORKS WITH AUXILIARY GAUSSIAN MIXTURE MODELS FOR REAL-TIME SPEECH RECOGNITION

*Xin Lei* *Hui Lin* *Georg Heigold*

Google Inc., 1600 Amphitheatre Parkway
Mountain View, CA 94043 USA
{xinlei,linhui,heigold}@google.com

## ABSTRACT

We present a framework that improves real-time speech recognition performance using deep neural networks (DNNs) with auxiliary Gaussian mixture models (GMMs). The DNNs and the auxiliary GMMs share the same hidden Markov model (HMM) state inventory. First, online incremental feature-space adaptation is performed using the GMM acoustic model. The speaker-adapted features are used to improve the recognition performance of both GMM and DNN models. Second, the acoustic scores from GMMs and DNN are combined at the state-level during decoding. Experiments on a large vocabulary speech recognition task show that both approaches improve recognition performance consistently and that the gains are mostly additive, resulting in about 5% relative improvement over the competitive DNN baseline in both Portuguese and English systems.

***Index Terms***— DNN, GMM, speaker adaptation, system combination.

## 1. INTRODUCTION

Over the past few years there have been significant advances in automatic speech recognition (ASR) using deep neural networks (DNNs) for acoustic modeling [1]. Instead of using traditional Gaussian mixture models (GMMs) to model the probability distribution of acoustic feature vectors associated with each state of an hidden Markov model (HMM), DNNs with many hidden layers, are used to produce posterior probabilities over HMM states. The DNNs have been shown to outperform GMMs by a large margin in several competitive large-vocabulary continuous speech recognition (LVCSR) systems [2, 3, 4].

Although DNNs have been shown to have superior power for discriminating HMM states, the GMMs have the advantages of easily parallelizable training, fast and efficient speaker adaptation [5], and being less computationally demanding. Furthermore, the GMM system exhibits different error patterns compared to the DNN systems. This suggests DNNs and GMMs may be complementary. If so, they could be combined to achieve even better speech recognition performance. In addition, since the target distributions for DNN training are typically obtained from forced alignments generated by a baseline GMM system, the bootstrap GMMs can be used directly at run-time without training overhead.

Given by the differences and similarities between DNN and GMM acoustic models, this paper investigates how to improve the performance of DNNs in real-time LVCSR systems by adding auxiliary GMMs that share the same HMM state inventory. Two approaches are studied: online incremental speaker adaptation, and state-level score combination. Recent studies [2, 6] have shown multi-pass batch-mode adapted features improve accuracy on offline Switchboard transcription benchmarks. In this paper we focus on online incremental adaptation that is practical for real-time speech recognition tasks such as mobile Voice Search. Different from conventional system combination at the hypothesis-level [7], we explore the effectiveness of combining DNN and GMM acoustic scores at the state-level without introducing extra latency in decoding.

The rest of this paper is organized as follows. In Section 2, online feature-space adaptation for DNN is described. Section 3 presents state-level score combination of DNN and GMM. Section 4 shows the experimental results on adaptation and combination. Finally, Section 5 concludes the paper and discusses future work.

## 2. ONLINE INCREMENTAL FEATURE-SPACE ADAPTATION FOR DNN

Speaker adaptation is important in reducing the mismatch between training and decoding conditions. Various adaptation techniques have been proposed for GMM-based acoustic models. These techniques can be roughly divided into two categories: model-space adaptation, and feature-space adaptation. Feature-space adaptation does not require modifying the entire acoustic model, hence, it is well suited for real-time ASR server systems. Constrained maximum likelihood linear regression (CMLLR) [5], also called feature-space MLLR (fMLLR), is one of the most popular feature-space adapta-

tion techniques adopted in research and commercial ASR systems [8, 9].

Unsupervised online fMLLR estimates and updates a speaker-specific affine transformation of the feature vectors during decoding. Let $o_t$ be the $n$-dimensional feature vector at time frame $t$, the transformed feature $\hat{o}_t$ is,

$$\hat{o}_t = Ao_t + b, \tag{1}$$

where $A$ is the $n \times n$ rotation/scaling matrix, $b$ is the $n \times 1$ bias term. The transform parameters $W = [A \quad b]$ are estimated by optimizing an auxiliary Q-function and can be solved iteratively [5].

For GMM systems, fMLLR can improve the recognition accuracy with as little as 5 seconds of speech data. With sufficient amount of adaptation data, fMLLR can typically achieve 10% to 20% relative improvement in terms of word error rate (WER), compared to a speaker independent baseline. Since fMLLR adapts the input feature vectors, it is a natural extension to use it to transform features for DNNs, if they share the same feature space as the underlying GMMs. Batch-mode fMLLR transformed features have been shown effective for English Broadcast News and Switchboard offline systems [2, 6]. Here we explore the efficacy of online incremental fMLLR adaptation for DNNs in real-time LVCSR systems. It is also worth mentioning that in [2] the authors proposed a feature-space discriminative linear regression (fDLR) technique for DNNs using back-propagation. This technique achieves similar gains to fMLLR, but it is more computationally expensive and it is uncertain how well it works for online incremental adaptation scenarios.

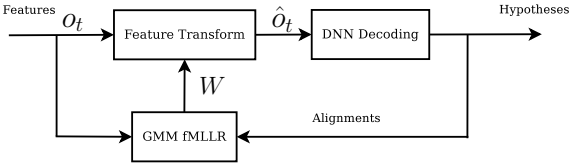


**Fig. 1**. Online incremental fMLLR adaptation for DNN.

As illustrated in Figure 1, for each utterance, the same feature vectors $o_{1:T}$ are used for GMM fMLLR adaptation and transformed to $\hat{o}_{1:T}$ for DNN decoding. The GMM adaptation also utilizes the Viterbi alignments from DNN decoding to accumulate sufficient statistics and estimate feature transform $W$. The feature transform block is initialized with an identity transform and continuously updated to $W$ once there are enough adaptation data. Each updated transform $W$ is then applied to the feature vectors of the following utterances in the speech session.

## 3. STATE-LEVEL COMBINATION OF GMM AND DNN SCORES

In recent DNN-HMM based ASR systems, a DNN classifier is discriminatively trained to predict the targets of context-dependent (CD) HMM states. In the decoding phase, for each observation vector $o'_t$ and CD-HMM state $s_j$, the posterior probability $P(s_j|o'_t)$ is computed with the DNN and then converted to state emission likelihood:

$$p_{\text{dnn}}(o'_t|s_j) = \frac{P(s_j|o'_t)}{P(s_j)} \cdot p(o'_t), \tag{2}$$

where $s_j$ is the $j$-th HMM state, and observation vectors $o'_t$ are acoustic feature vectors augmented with neighbor frames. $P(s_j)$ is the prior probability of state $s_j$, which can be estimated from the frequency of the state in training alignments. Since $p(o'_t)$ is a constant independent of state $s_j$, we ignore it in the likelihood computation.

In conventional GMM-HMM systems, for an acoustic state $s_j$ with $M$ multivariate Gaussian densities, the state emission likelihoods are computed directly:

$$p_{\text{gmm}}(o_t|s_j) = \sum_{m=1}^{M} w_{jm}\mathcal{N}(o_t; \mu_{jm}, \Sigma_{jm}), \tag{3}$$

where $w_{jm}$ is the mixture weight of the $m$-th Gaussian component in state $s_j$, $\mu_{jm}$ is the mean vector, $\Sigma_{jm}$ is the covariance, and $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a Gaussian density with mean $\mu$ and covariance $\Sigma$.

### 3.1. Speaker-independent score combination

We observe the phone classification error patterns of DNN and GMMs are quite different. Therefore, combining DNN and GMM systems may achieve better classification and ASR performance. We perform system combination at the state-level for every frame. This approach is similar to state-level combination described in multi-stream audio-visual speech modeling [10], but here different classifiers are used instead of different knowledge sources. The final acoustic score for frame $t$ and state $s_j$ is a linear combination of DNN and GMM acoustic scores:

$$\log \hat{p}(o_t, o'_t|s_j) = \alpha_j \log p_{\text{dnn}}(o'_t|s_j) + (1-\alpha_j) \log p_{\text{gmm}}(o_t|s_j). \tag{4}$$

The parameter $\alpha_j$ is the state-dependent weight of DNN log likelihood score and is between 0 and 1. The state-dependent weights can be learned by minimizing a discriminative criterion such as phone or state classification error rate. For simplicity, in this paper we use a single weight $\alpha$ for all states. The weight $\alpha$ can be optimized by grid search on a development set. A diagram of the state-level score combination is shown in Figure 2. In this speaker-independent score combination case, different types of acoustic features may be used

for DNN and GMM acoustic models. For example, we can use filter-bank energies (FBE) for DNN and perceptual linear predictive (PLP) features for GMM.

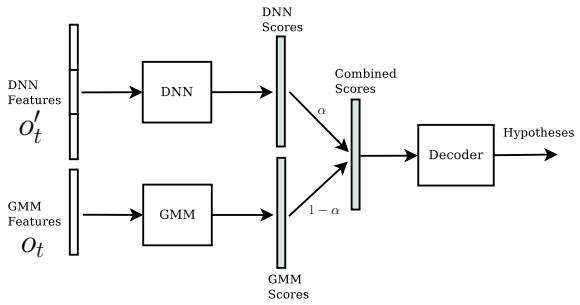


**Fig. 2**. Speaker-independent state-level score combination of DNN and GMM.

### 3.2. Speaker-adapted score combination

The state-level combination of DNN and GMM scores can be used jointly with feature-space speaker adaptation described in Section 2. For the DNN and GMM to share the same feature-space adaptation transform, they need to be trained from the same type of source features. In our adaptation experiments, we use PLP features for both DNN and GMM models.

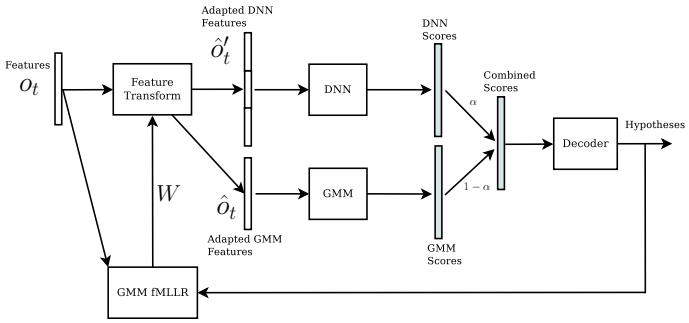


**Fig. 3**. Speaker-adapted state-level score combination of DNN and GMM.

A diagram of the complete system with feature-space adaptation and state-level score combination is shown in Figure 3. The PLP features are incrementally adapted and sent to DNN and GMM scorers to generate likelihood scores. The adapted DNN and GMM scores are then combined for decoding. This system runs in real-time mode and takes advantage of both speaker adaptation and system combination. For modern multi-core CPUs, we can reduce latency by using separate threads for DNN and GMM score computation, and for GMM adaptation.

## 4. EXPERIMENTAL RESULTS

### 4.1. Baseline systems

The experiments are performed with a LVCSR system which transcribes voice search queries, short messages, e-mails, and user actions from mobile devices [11]. We experiment with two systems: the Portuguese system with relatively small amount of training data, and the US English system with a very large training corpus. In both systems, the baseline GMM-HMM models use triphone HMMs with decision-tree clustered states. The acoustic features are 9 contiguous frames of 13-dimensional PLP features spliced and projected to 40 dimensions by linear discriminant analysis (LDA). Semi-tied covariances (STC) [12] are used to further diagonalize the LDA transformed features. Boosted-MMI was used to train the model discriminatively [13].

The GMM-HMM acoustic model is used to force align the training data to obtain senone state labels for DNN training. The input for the DNN is the same LDA+STC transformed PLP features, but augmented with neighbor frames in a context window. All DNNs in the experiments have four hidden layers each with 2560 nodes and logistic activation, and an output layer with softmax activation. At recognition time, the weights in the DNN are quantized to 8 bits and fixed-point SIMD primitives are used to achieve real-time decoding performance [14].

The Portuguese GMM-HMM acoustic model is trained using 100 hours of speech data. It has 2959 senone states and 27K diagonal covariance Gaussians. The same training set is force aligned for DNN training. The DNN is trained from scratch using DistBelief framework [15]. A context window of 26 frames is used: left 20 frames, right 5 frames, plus current frame. The test set is created from a quick data collection procedure [16]. It contains about 10 hours of data, with 763 speakers and an average of 200 utterances per speaker.

The US English GMM-HMM acoustic model is built from a very large corpus and has 7969 states and 580K Gaussians. The GMMs are used to force align a subset of approximately 5780 hours of data for DNN training. The DNN is trained with a context window of 18 frames (left 16 frames and right 1 frame) using GPU similar to [4]. For evaluation, we use a 12-hour test set containing 100 speakers, with between 2 and 30 minutes of data for every speaker. Each speaker's data are collected over a period of usage in very different channel, environment, and noise conditions. Therefore, this test set is a challenging task for adaptation studies.

The performance results of baseline GMM and DNN models in Portuguese (`pt-pt`) and US English (`en-us`) systems are shown in Table 1. We report both word error rate and normalized sentence accuracy (NSACC) results. The `pt-pt` DNN gives about 13% relative improvement over the baseline GMM model. For comparison purpose, we also train one DNN from 11 contiguous frames of 40 log filter-bank energies with no temporal derivatives. This FBE DNN has a

WER of 22.4%, which is slightly worse than the PLP DNN's 22.2% result. Due to time constraint, the US English DNN is not trained fully to convergence. It gives 0.9% absolute WER improvement over the GMM baseline.

**Table 1**. Baseline WER and NSACC (%) results.

| | pt-pt | | en-us | |
|---|---|---|---|---|
| Model | WER | NSACC | WER | NSACC |
| GMM | 25.6 | 62.5 | 18.2 | 55.4 |
| DNN | 22.2 | 67.2 | 17.3 | 57.5 |

### 4.2. Adaptation for GMM and DNN

We then perform online incremental fMLLR adaptation experiments on the GMM and DNN systems. Silence frames are excluded from adaptation statistics accumulation. A minimum adaptation data threshold of 10 seconds is used for all adaptation experiments. The adaptation results are presented in Table 2.

**Table 2**. WER and NSACC (%) results of adaptation.

| | pt-pt | | en-us | |
|---|---|---|---|---|
| Model | WER | NSACC | WER | NSACC |
| GMM | 25.6 | 62.5 | 18.2 | 55.4 |
| GMM + fMLLR | 23.8 | 64.8 | 17.5 | 56.6 |
| DNN | 22.2 | 67.2 | 17.3 | 57.5 |
| DNN + fMLLR | 21.6 | 68.4 | 17.0 | 57.8 |

We can see that in pt-pt system, speaker adaptation improves the GMM model by 1.8% absolute, and improves the DNN model by 0.6% absolute. In en-us system, adaptation improves GMM baseline by 0.7% absolute and DNN baseline by 0.3% absolute. The improvements from adaptation are much smaller in en-us system, probably due to the more challenging test set. For both systems, the adaptation gain for DNN is approximately 1/3 of that for GMM models. The reason can be two-fold: first, the DNN can already learn some aspects of speaker invariance [17, 2]; second, adapting the DNN using the GMMs has a model mis-match due to the cross-adaptation nature. Similar to speaker adaptive training for GMMs, re-training the DNN using per-speaker normalized features may also reduce the mismatch between training and decoding.

### 4.3. Combination of GMM and DNN

Finally, state-level system combination experiments are performed. We choose a single combination weight $\alpha_j = 0.8$ for all states in all combination experiments. Since GMM models use Gaussian selection to reduce computation at recognition time, some acoustic states are not selected for a given frame and the corresponding GMM score is a constant worst score. For these cases, we find that ignoring the DNN score and outputting the worst score achieve good WER results. We also observe that very aggressive Gaussian selection thresholds hurt the score combination.

**Table 3**. WER and NSACC (%) results of combination.

| | pt-pt | | en-us | |
|---|---|---|---|---|
| Model | WER | NSACC | WER | NSACC |
| DNN | 22.2 | 67.2 | 17.3 | 57.5 |
| DNN + GMM | 21.7 | 67.5 | 16.7 | 58.1 |
| DNN + GMM + fMLLR | 21.0 | 68.5 | 16.5 | 58.5 |

The combination results are summarized in Table 3. Combining DNN and GMM scores achieves 0.5% absolute WER improvement in pt-pt system and 0.6% absolute in en-us system. This gain is mostly additive with the improvement from incremental adaptation. The final system with both combination and adaptation outperforms the baseline DNN system by 1.2% absolute in pt-pt system, and 0.8% absolute in en-us system. In both systems, the WER reduction is about 5% relative over the DNN baseline.

## 5. CONCLUSIONS

In this paper, we have described a framework of combining DNN and auxiliary GMM acoustic models for improved real-time speech recognition. Online incremental feature-space adaptation is performed using auxiliary GMMs. The estimated transform is applied to the features for both the GMMs and the DNN. Then the output likelihood scores from the DNN and GMMs are combined at the state-level for improved decoding performance. On a large vocabulary speech recognition task, we observe consistent improvement from both techniques and the gains are mostly additive, achieving 5% relative WER reduction over the DNN baseline in both Portuguese and English systems.

Future work will include learning state-dependent combination weights with a discriminative criterion. We will also investigate the use of smaller-sized GMM acoustic models to reduce computation and memory footprint.

### Acknowledgments

The authors would like to thank our colleague Andrew Senior for help on training the US English PLP DNN model. Thanks also to Patrick Nguyen and Vincent Vanhoucke for fruitful discussions during this work.

## 6. REFERENCES


[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, November 2012.

[2] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, 2011.

[3] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *Proc. Interspeech*, 2012.

[4] N. Jaitly, P. Nguyen, A. W. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. Interspeech*, 2012.

[5] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep., Cambridge University Engineering Department, May 1997.

[6] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc. ASRU*, 2011.

[7] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, 1997.

[8] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, "Incremental online feature space MLLR adaptation for telephony speech recognition," in *Proc. ICSLP*, 2002.

[9] X. Lei, J. Hamaker, and X. He, "Robust feature space adaptation for telephony speech recognition," in *Proc. Interspeech*, 2006.

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, 2000.

[11] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, pp. 61–90. Springer, 2010.

[12] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[13] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008.

[14] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2011.

[15] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, "Large scale distributed deep networks," in *NIPS*, 2012.

[16] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Proc. Interspeech*, 2010.

[17] T. Schaaf and F. Metze, "Analysis of gender normalization using MLP and VTLN features," in *Proc. Interspeech*, 2010.