Exploring GMM-derived Features for Unsupervised Adaptation of Deep Neural Network Acoustic Models

Natalia Tomashenko, Yuri Khokhlov, Anthony Larcher, Yannick Estève

Exploring GMM-derived Features for Unsupervised Adaptation of Deep Neural Network Acoustic Models

Natalia Tomashenko[1,2,3], Yuri Khokhlov[1], Anthony Larcher[3], and Yannick Estève[3]

[1]STC-innovations Ltd, Saint-Petersburg, Russia
{tomashenko-n,khokhlov}@speechpro.com
[2]ITMO University, Saint-Petersburg, Russia
[3]University of Le Mans, France
{anthony.larcher,yannick.esteve}@univ-lemans.fr

Abstract. In this paper we investigate GMM-derived features recently introduced for adaptation of context-dependent deep neural network HMM (CD-DNN-HMM) acoustic models. We present an initial attempt of improving the previously proposed adaptation algorithm by applying lattice scores and by using confidence measures in the traditional maximum a posteriori adaptation (MAP) adaptation algorithm. Modified MAP adaptation is performed for the auxiliary GMM model used in a speaker adaptation procedure for a DNN. In addition we introduce two approaches - data augmentation and data selection, for improving the regularization in MAP adaptation for DNN. Experimental results on the Wall Street Journal (WSJ0) corpus show that the proposed adaptation technique can provide, on average, up to 9.9% relative word error rate (WER) reduction under an unsupervised adaptation setup, compared to speaker independent DNN-HMM systems built on conventional features.

Keywords: speaker adaptation, deep neural networks (DNN), MAP, fMLLR, CD-DNN-HMM, GMM-derived (GMMD) features, speaker adaptive training (SAT), confidence scores

1 Introduction

Nowadays, deep neural networks (DNNs) have replaced conventional GMM-HMMs in most state-of-the-art automatic speech recognition (ASR) systems, because it has been shown that DNN-HMM models outperform GMM-HMMs in different ASR tasks. However, various adaptation algorithms that have been developed for GMM-HMM systems cannot be easily applied to DNNs because of the different nature of these models. Many new adaptation methods have recently been developed for DNNs, and a few of them [1–5] take advantage of robust adaptability of GMMs. However, there is no universal method for efficient transfer of all adaptation algorithms from the GMM framework to DNN

models. The purpose of the present work is to make a step in this direction using GMM-derived features for training DNN models.

Most of the existing methods for adapting DNN models can be classified into several types: (1) linear transformation, (2) regularization techniques, (3) auxiliary features, (4) multi-task learning, (5) combining GMM and DNN models. **Linear transformation** can be applied at different levels of the DNN system: to the input features, as in linear input network transformation (LIN) [6] or feature-space discriminative linear regression (fDLR); to the activations of hidden layers, as in linear hidden network transformation (LHN) [6]; or to the softmax layer, as in LON or in output-feature discriminative linear regression. The second type of adaptation consists in re-training the entire network or only a part of it using special **regularization techniques** for improving generalization, such as L2-prior regularization [7], Kullback-Leibler divergence regularization [8], conservative training [9]. The concept of **multi-task learning** (MTL) has recently been applied to the task of speaker adaptation and has been shown to improve the performance of different model-based DNN adaptation techniques, such as LHN and learning speaker-specific hidden unit contributions [10]. **Using auxiliary features** is another approach in which the acoustic feature vectors are augmented with additional speaker-specific or channel-specific features computed for each speaker or utterance at both training and test stages. An example of effective auxiliary features is i-vectors [11]. Alternative methods are adaptation with speaker codes [12] and factorized adaptation [13].The most common way of **combining GMM and DNN models** for adaptation is using GMM-adapted features, for example fMLLR, as input for DNN training [1]. In [2] likelihood scores from DNN and GMM models, both adapted in the feature space using the same fMLLR transform, are combined at the state level during decoding. The authors of [5] propose combining the GMM and DNN models using the temporally varying weight regression framework.

In this work we investigate a novel approach for SAT of DNNs based on using GMM-derived features as the input to DNNs [3, 4]. We present an initial attempt of improving the previously proposed scheme for DNN adaptation by using recognition lattices in MAP adaptation and by the data augmentation and data selection approaches.

2 SAT for DNN-HMM based on GMM-derived features

Construction of GMM-derived features for adapting DNNs was proposed in [3,4], where it was demonstrated, using MAP and fMLLR adaptation as an example, that this type of features makes it possible to effectively use GMM-HMM adaptation algorithms in the DNN framework.

Our features are obtained as follows (see Figure 2). First, 39-dimensional Mel-frequency cepstral coefficients (MFCC) with delta and acceleration coefficients are extracted with per-speaker cepstral mean normalization (CMN). Then an auxiliary GMM monophone model is used to transform cepstral feature vectors into log-likelihoods vectors. At this step, speaker adaptation of the auxiliary

speaker-independent (SI) GMM model is performed for each speaker in the training corpus and the new speaker-adapted (SA) GMM model is obtained in order to extract SA GMM-derived features. In the auxiliary GMM, each phoneme is

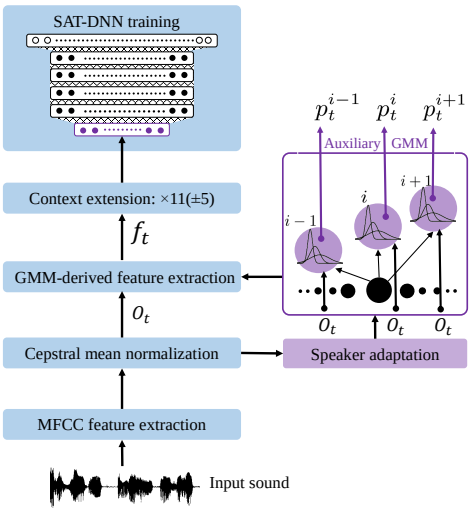


Fig. 1. Using speaker adapted GMMD features for SAT DNN training.

modeled using a three state left-right context-independent GMM. For a given acoustic MFCC-feature vector, a new GMM-derived feature vector is obtained by calculating log-likelihoods across all the states of the auxiliary GMM monophone model on the given vector. Suppose o_t is the acoustic feature at time t, then the new GMM-derived feature vector f_t is calculated as follows:

$$f_t = [p_t^1, \ldots, p_t^n], \tag{1}$$

where n is the number of states in the auxiliary GMM model,

$$p_t^i = \log\left(P(o_t \mid s_t = i)\right) \tag{2}$$

is the log-likelihood estimated using the GMM. Here s_t denotes the state index at time t. In our case n is equal to 132 ($39\times3+3\times5$), coming from: 39 three-state phones, one five-state silence model, and two five-state (speech and non-speech) noise models. Hence this procedure leads to a 132-dimension feature vector per speech frame. After that, the features are spliced in time taking a context size of 11 frames (i.e., ±5). We will refer to these resulting features as GMMD features. The dimension of the resulting features is equal to 1452 (11×132). These features are used as the input for training the DNN. The proposed approach can be considered a feature space transformation technique with respect to DNN-HMMs trained on GMMD features.

3 MAP adaptation using lattices scores

The use of lattice-based information and confidence scores [14] is a well-known method for improving the performance of unsupervised adaptation. In this work we use the MAP adaptation algorithm for adapting the SI GMM model. Speaker adaptation of a DNN-HMM model built on GMMD features is performed through the MAP adaptation of the auxiliary GMM monophone model, which is used for calculating GMMD features. We modify the traditional MAP adaptation algorithm by using lattices instead of alignment from the first decoding pass as follows. Let m denote an index of a Gaussian in SI acoustic model (AM), and μ_m the mean of this Gaussian. Then the MAP estimation of the mean vector is

$$\widehat{\mu}_m = \frac{\tau\mu_m + \sum_t \gamma_m(t) p_s(t) o_t}{\tau + \sum_t \gamma_m(t) p_s(t)}, \tag{3}$$

where τ is the parameter that controls the balance between the maximum likelihood estimate of the mean and its prior value; $\gamma_m(t)$ is the posterior probability of Gaussian component m at time t; and $p_s(t)$ is the confidence score of state s at time t in the lattice obtained from the first decoding pass by calculating arc posteriors probabilities. The forward-backward algorithm is used to calculate these arc posterior probabilities from the lattice as follows:

$$P(l|O) = \frac{\sum_{q \in Q_l} p_{acc}(O|q)^{\frac{1}{\alpha}} P_{lm}(w)}{P(O)}, \tag{4}$$

where α is the language model scale factor (the optimal value for α is found empirically); q is a path through the lattice corresponding to the word sequence w; Q_l is the set of paths passing through arc l; $p_{acc}(O|q)$ is the acoustic likelihood; $P_{lm}(w)$ is the language model probability; and $p(O)$ is the overall likelihood of all paths through the lattice. In a particular case, when $p_s(t) = 1$ for all states and t, formula (3) represents the traditional MAP adaptation. In addition to this frame-level weighting scheme, we apply confidence base selection scheme, when we use in (3) only those observations, which confidence scores exceed the given threshold.

4 Data augmentation and data selection for SAT

In this work we explore two approaches to improve the performance of SAT DNN models with MAP adaptation. The first approach is based on using different values of τ (in formula (3)) when extracting adapted GMMD features for DNN training. In this approach we extract features for all training corpus several times for a set of τ values. And then the DNN models are trained on the union of the obtained features. The intuition behind this approach is similar to that used in data augmentation.

The second approach, which we call data selection strategy, consists in splitting training data for each speaker in the training corpus into several parts and

then performing MAP adaptation independently on each of the part. In this paper we use a simple implementation of this strategy - we randomly separate training data for each speaker into several subsets so that the total amount of data in each subset is approximately equal to the average amount of data per speaker in the test set. This strategy serves as a regularization and is supposed to make adaptation more robust to the size of the adaptation set. Hence, the

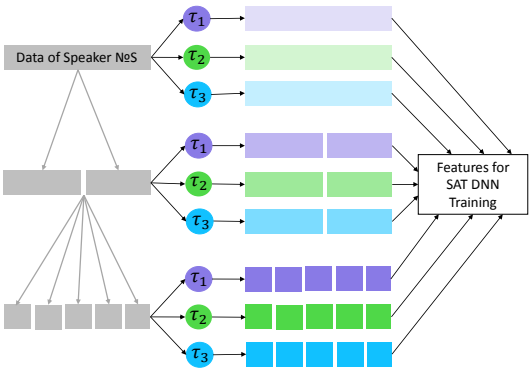
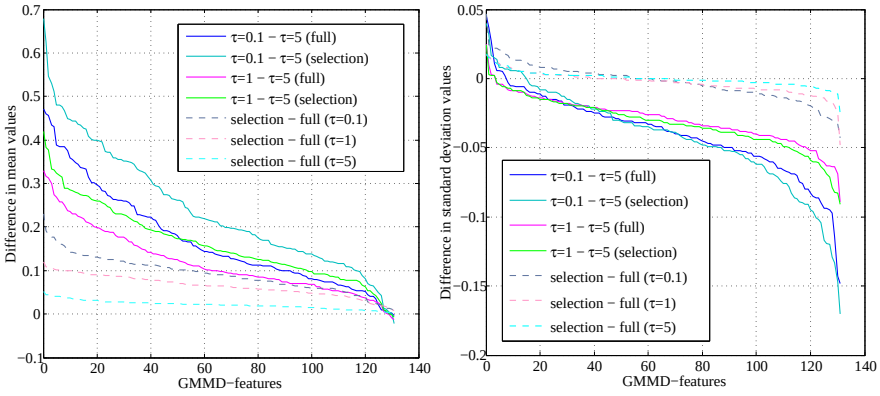


Fig. 2. Data augmentation and data selection scheme for SAT

original data from the training corpus are used in AM training several times with different values of τ and inside different subsets of data chosen for adaptation. The motivation for these two approaches lies in obtaining more robust SAT DNN models for MAP adaptation, especially when the training corpus is relatively small.

The GMMD feature dynamic in the training corpus for different values of τ and for different data selection strategies is shown in Figure 3. In both pictures "full" means that during the SAT training for a given speaker all data of that speaker from the training corpus are used for MAP adaptation, whereas "selection" means that data selection strategy is applied and training data for this speaker is randomly spitted into two subsets so that MAP adaptation is performed for each subset independently. Let denote T_1 and T_2 two types of features, (or more precisely, to GMMD features extracted with different parameters). Every curve in Figures 3.a and 3.b, marked as "T_1–T_2", corresponds to the average differences between T_1 and T_2 features and is calculated as follows. First, we subtract coordinatewise features T_2 from T_1 on the training corpus. Then we found mean (Figure 3.a) and standard deviation values (Figure 3.b) for each feature vector coordinate. Finally, we sort the obtained values for each feature vector dimension by descending order. We can see that GMMD features calculated for various τ and with (or without) data selection strategy have different amplitude and dynamic characteristics, therefore they can contain complemen-

tary information. Hence data augmentation might improve AM by making them more robust to τ and to the size of the adaptation set.



a. Mean values. b. Standard deviation values.

Fig. 3. Differences in GMMD-features.

5 Experimental results

The experiments are conducted on the WSJ0 corpus [15]. For AM training we use 7138 utterances of 83 speakers from the standard SI-84 training set, which correspond to approximately 15 hours of data, recorded with the Sennheiser microphone, 16 kHz. AMs are trained using the Kaldi speech recognition toolkit [16], following mostly Kaldi WSJ recipe (except for GMMD-features and adaptation). We use conventional 11×39MFCC features (39-dimensional MFCC (with CMN) spliced across 11 frames (± 5)) as baseline features and compare them to the proposed GMMD features. We train four DNN models: SI model on 11×39MFCC; SI and two SAT models on GMMD features. These four DNNs have identical topology (except for the dimension of the input layer) and are trained on the same training dataset. An auxiliary monophone GMM is also trained on the same data.

The first SAT DNN on GMMD features is trained as described in Section 2 with parameter τ for adaptation equal to 5. The second SAT DNN on GMMD features is trained using data augmentation (with τ equal to 0.1, 1 and 5) and data selection strategy, as described in Section 4. For training SI-DNN on GMMD features, we apply the scheme shown in Figure 1, but eliminate the speaker adaptation step. All four CD-DNN-HMM systems had six 2048-neuron hidden layers and 2355-neuron output layer. The neurons in the output layer correspond to context-dependent states determined by tree-based clustering in CD-GMM-HMM. The DNN is initialized with the stacked restricted Boltzmann machines by

using layer by layer generative pre-training. It is trained with an initial learning rate of 0.008 using the cross-entropy objective function. After that five iterations of sequence-discriminative training with per-utterance updates, optimizing state Minimum Bayes Risk (sMBR) criteria, are performed.

In all experiments further we consider SI DNN trained on 11×39MFCC features as the baseline model and compare the performance results of the other models with it. Evaluation is carried out on the standard WSJ0 evaluation test **si_et_20**, which consists of 333 read utterances (5645 words) from 8 speakers. A WSJ trigram open NVP LM with a 20k word vocabulary is used during recognition. The OOV rate is about 1.5%. The LM is pruned as in the Kaldi [16] WSJ recipe with the threshold 10^{-7}. The adaptation experiments are conducted in an unsupervised mode on the test data using transcripts or lattices obtained from the first decoding pass. For adapting an auxiliary GMM model we use MAP adaptation algorithm. We perform two adaptation experiments: (1) with traditional MAP and (2) with lattice-based MAP using confidence scores, as described in Section 3. For lattice-based MAP the value of confidence threshold is 0.6. The performance results in terms of word error rate (WER) for SI and adapted DNN-HMM models are presented in Table 1. We can see that using confidence scores can give an additional slight improvement in MAP adaptation for DNN models over adaptation, which uses an alignment. The best result is obtained using data augmentation and data selection strategies. For comparison purposes we also train six DNN models with τ values 0.1, 1 and 5 with and without data selection strategies, but in all cases the results are worse than the one obtained combining both strategies, so we do not report other results here.

Type of Features	Adaptation	WER, %	Δ WER, %
11×39MFCC	SI	7.51	baseline
GMMD	SI	7.83	–
	MAP (alignment)	7.09	5.6
	MAP (lattice-based)	6.93	8.4
	MAP (data augmentation & selection)	6.77	9.9

Table 1. Summary of WER (%) results on WSJ0 evaluation set si_et_20. Δ WER - relative WER reduction.

6 Conclusion

In this work we have investigated GMM-derived features recently introduced for adaptation of DNN AMs. MAP adaptation algorithm is performed for the auxiliary GMM model used in a SAT procedure for a DNN. We present an attempt of improving the previously proposed adaptation algorithm by using confidences scores in adaptation. In addition we introduced two approaches, so called data augmentation and data selection strategies, for improving the regularization in

MAP adaptation for DNN. Experimental results on the WSJ0 corpus demonstrate that, in an unsupervised adaptation mode, the proposed adaptation technique can provide, approximately, up to 9.9% relative WER reduction compared to the SI DNN system built on conventional 11×39MFCC features.

Acknowledgements This work was partially financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.579.21.0057, ID RFMEFI57914X0057.

References

1. S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.
2. X. Lei, H. Lin, and G. Heigold, "Deep neural networks with auxiliary gaussian mixture models for real-time speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7634–7638.
3. N. Tomashenko and Y. Khokhlov, "Speaker adaptation of context dependent deep neural networks based on map-adaptation and gmm-derived feature processing," in *Proc. Interspeech*, 2014, pp. 2997–3001.
4. ——, "Gmm-derived features for effective unsupervised adaptation of deep neural network acoustic models," in *Proc. Interspeech*, 2015, pp. 2882–2886.
5. S. Liu and K. C. Sim, "On combining dnn and gmm with unsupervised speaker adaptation for robust automatic speech recognition," in *Proc. ICASSP*. IEEE, 2014, pp. 195–199.
6. R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Adaptation of hybrid ann/hmm models using linear hidden transformations and conservative training," in *Proc. ICASSP*, vol. 1. IEEE, 2006.
7. H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 7947–7951.
8. D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7893–7897.
9. D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio, "Adaptation of artificial neural networks avoiding catastrophic forgetting," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 1554–1561.
10. P. Swietojanski, P. Bell, and S. Renals, "Structured output layer with auxiliary targets for context-dependent acoustic modelling," in *Interspeech*, 2015.
11. A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014, pp. 225–229.
12. O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code," in *Proc. ICASSP*. IEEE, 2013, pp. 7942–7946.
13. J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. ICASSP*. IEEE, 2014, pp. 5537–5541.
14. C. Gollan and M. Bacchiani, "Confidence scores for acoustic model adaptation," in *Proc. ICASSP*. IEEE, 2008, pp. 4289–4292.
15. D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

16. D. Povey, A. Ghoshal *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.