



HAL
open science

Exploring Gaussian mixture model framework for speaker adaptation of deep neural network acoustic models

Natalia Tomashenko, Yuri Khokhlov, Yannick Estève

► **To cite this version:**

Natalia Tomashenko, Yuri Khokhlov, Yannick Estève. Exploring Gaussian mixture model framework for speaker adaptation of deep neural network acoustic models. 2020. hal-02551714

HAL Id: hal-02551714

<https://hal.science/hal-02551714v1>

Preprint submitted on 23 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Exploring Gaussian mixture model framework for speaker adaptation of deep neural network acoustic models

Natalia Tomashenko^a, Yuri Khokhlov^b, Yannick Estève^a

^a*LIA, University of Avignon, France*

^b*STC-innovations Ltd, Saint-Petersburg, Russia*

Abstract

In this paper we investigate the GMM-derived (GMMD) features for adaptation of deep neural network (DNN) acoustic models. The adaptation of the DNN trained on GMMD features is done through the maximum a posteriori (MAP) adaptation of the auxiliary GMM model used for GMMD feature extraction. We explore fusion of the adapted GMMD features with conventional features, such as bottleneck and MFCC features, in two different neural network architectures: DNN and time-delay neural network (TDNN). We analyze and compare different types of adaptation techniques such as i-vectors and feature-space adaptation techniques based on maximum likelihood linear regression (fMLLR) with the proposed adaptation approach, and explore their complementarity using various types of fusion such as feature level, posterior level, lattice level and others in order to discover the best possible way of combination. Experimental results on the TED-LIUM corpus show that the proposed adaptation technique can be effectively integrated into DNN and TDNN setups at different levels and provide additional gain in recognition performance: up to 6% of relative word error rate reduction (WERR) over the strong feature-space adaptation techniques based on maximum likelihood linear regression (fMLLR) speaker adapted DNN baseline, and up to 18% of relative WERR in comparison with a speaker independent (SI) DNN baseline model, trained on conventional features. For TDNN models the proposed approach achieves up to 26% of relative WERR in comparison with a SI baseline, and up to 13% in comparison with the model adapted by using i-vectors. The analysis of the adapted GMMD features from various points of view demonstrates their effectiveness at different levels.

Keywords: Acoustic model adaptation, Deep Neural Networks (DNN), Automatic Speech Recognition (ASR), Gaussian Mixture Models (GMM), Speaker adaptation, GMM-derived (GMMD) features

1. Introduction

Adaptation of DNN acoustic models is a rapidly developing research area. The aim of acoustic model (AM) adaptation is to reduce mismatches between

training and testing acoustic conditions and improve the accuracy of the automatic speech recognition (ASR) system for a target speaker or channel using a limited amount of adaptation data from the target acoustic source. In the recent years DNNs have replaced conventional Gaussian mixture models (GMMs) in most state-of-the-art ASR systems, because it has been shown that DNN Hidden Markov Models (HMMs) outperform GMM-HMMs in different ASR tasks. Many adaptation algorithms that have been developed for GMM-HMM systems [Gales, 1998; Gauvain & Lee, 1994] cannot be easily applied to DNNs because of the different nature of these models. Among the adaptation methods developed for DNNs only a few take advantage of robust adaptability of GMMs [Seide et al., 2011; Rath et al., 2013; Kanagawa et al., 2015; Lei et al., 2013; Liu & Sim, 2014; Murali Karthick et al., 2015; Parthasarathi et al., 2015]. However, none of them suggests a universal method for efficient transfer of all adaptation algorithms from the GMM models to DNN framework.

In the past, there were different attempts to integrate GMM and DNN models into a single structure. One of the common approaches is to use features generated by neural networks, such as *tandem* [Hermansky et al., 2000] or *bottleneck* (BN) [Grézl et al., 2007; Yu & Seltzer, 2011; Paulik, 2013] features in order to train a GMM model. Other approaches include *deep GMMs* [Demuyne & Triefenbach, 2013] and a softmax layer with hidden variables [Tüske et al., 2015b,a], which use the concept of log-linear mixture models. In [Variani et al., 2015], a GMM layer is used as an alternative to the softmax layer in a DNN model.

In this paper we investigate a recently introduced GMM framework for adaptation of DNN-HMM acoustic models [Tomashenko & Khokhlov, 2014, 2015; Tomashenko et al., 2016b,a,c]. Our approach is based on using features derived from a GMM model for training DNN models [Tomashenko & Khokhlov, 2014, 2015; Tomashenko et al., 2016b; Pinto & Hermansky, 2008] and GMM-based adaptation techniques. In the previous works it was shown that GMM log-likelihoods can be effectively used as features for training a DNN HMM model, as well as for the speaker adaptation task.

The first objective of this paper is to propose a universal way of integration of the GMM adaptation framework into the most commonly used neural network AMs, such as DNN (Section 3.1) and time delay neural network (TDNN) AMs (Section 3.2) using MAP adaptation (Section 3.3) as an example.

Another important objective is to present an extensive experimental analysis of the proposed adaptation approach on the standard TED-LIUM corpus [Rousseau et al., 2014] for different types of neural network AMs. These experiments include: adaptation of both cross-entropy (CE) and sequence trained DNN acoustic models (Section 4.4.3); adaptation of TDNN AMs (Section 4.5.3); complementarity of the proposed approach with the two most popular adaptation techniques, such as fMLLR (Section 4.4.3) and i-vectors (Section 4.5.3); discovering the best possible way of information fusion (Section 4.2) from the AMs trained with GMM-derived (GMMD) features and the baseline conventional AMs, both for DNN and TDNN AMs, in order to improve the overall recognition accuracy.

The final goal is to look more deeply into the nature of the GMM features and adaptation techniques associated with them for better understanding their properties, strengths and weaknesses and the potential for improvement. For this purpose we perform a series of experiments on TDNN AMs (Section 5) using lattice-based features (Section 5.1) by means of t-distributed stochastic neighbor embedding (t-SNE) visual analysis (Sections 5.3, 5.4), using Davies-Bouldin (DB) index (Sections 5.2, 5.4), and different distributions for lattice-based features statistics.

2. Review on neural network acoustic model adaptation

Various adaptation methods have been developed for DNNs. These methods can be categorized in two broad classes, *feature-space* and *model-based* methods.

Model-based adaptation methods rely on direct modifications of DNN model parameters. In [Swietojanski & Renals, 2014; Swietojanski et al., 2016], learning speaker-specific hidden unit contributions (LHUC) was proposed. The main idea of LHUC is to directly parametrize amplitudes of hidden units, using a speaker-dependent amplitude function. The idea of learning amplitudes of activation functions was also studied in [Trentin, 2001]. The adaptation parameters estimation via maximum a posteriori (MAP) linear regression was proposed in [Huang et al., 2014], and a hierarchical MAP approach was studied in [Huang et al., 2015b]. Other model-based DNN adaptation techniques include linear transformations, adaptation using regularization techniques, subspace methods and others.

Feature-space adaptation methods operate in the feature space and can either transform input features for DNNs, as it is done, for example, in fMLLR adaptation [Seide et al., 2011] or use auxiliary features.

2.1. Linear transformation

One of the first adaptation methods developed for DNNs was linear transformation that can be applied at different levels of the DNN-HMM system: to the input features, as in linear input network transformation (LIN) [Neto et al., 1995; Gemello et al., 2006; Li & Sim, 2010] or feature-space discriminative linear regression (fDLR) [Seide et al., 2011; Yao et al., 2012]; to the activations of hidden layers, as in linear hidden network transformation (LHN) [Gemello et al., 2006]; or to the softmax layer, as in LON [Li & Sim, 2010] or in output-feature discriminative linear regression [Yao et al., 2012].

2.2. Regularization techniques

In order to improve generalization during the adaptation, regularization techniques, such as L2-prior regularization [Liao, 2013], Kullback-Leibler divergence regularization [Yu et al., 2013; Huang & Gong, 2015; Tóth & Gosztolya, 2016] conservative training [Albesano et al., 2006] and others [Ochiai et al., 2014] are used.

2.3. Multi-task learning

The concept of multi-task learning (MTL) has recently been applied to the task of speaker adaptation in several works [Li et al., 2015; Huang et al., 2015a; Swietojanski et al., 2015] and has been shown to improve the performance of different model-based DNN adaptation techniques, such as LHN [Huang et al., 2015a] and LHUC [Swietojanski et al., 2015]. A slightly different idea was proposed earlier in [Price et al., 2014] in the form of special hierarchy of output layers, where tied triphone states are followed by monophone states.

2.4. Subspace methods

Subspace adaptation methods aim to find a speaker subspace and then construct the adapted DNN parameters as a point in the subspace. In [Dupont & Cheboub, 2000] an approach similar to the eigenvoice technique [Kuhn et al., 2000], was proposed for the fast speaker adaptation of neural network AMs.

In [Wu & Gales, 2015] a *multi-basis adaptive neural network* is proposed, where a traditional DNN topology is modified and a set of sub-networks, referred as *bases* were introduced. This DNN has a common input layer and a common output layer for all the bases. Each basis has several fully-connected hidden layers and there is no connections between neurons from different bases. The outputs of bases are combined by linear interpolation using a set of adaptive weights. The adaptation to a given speaker can be performed through optimization of interpolation weights for this speaker. The idea of this approach was motivated by the cluster adaptive training (CAT), developed for GMM AMs. Paper [Tan et al., 2015] also investigates the CAT framework for DNNs. A subspace learning speaker-specific hidden unit contributions (LHUC) adaptation was proposed in [Samarakoon & Sim, 2016b].

2.5. Factorized adaptation

Factorized adaptation [Li et al., 2014; Yu et al., 2012; Qian et al., 2016; Tran et al., 2016; Samarakoon & Sim, 2016a] takes into account different factors that influence the speech signal. These factors can have different nature (speaker, channel, background noise conditions and others) and can be modeled explicitly before incorporating them into the DNN structure, for example, in the form of auxiliary features [Li et al., 2014], such as i-vectors, or can be learnt jointly with the neural network AM. The first case, when factors, such as noise or speaker information, are estimated explicitly from the training and testing data, and are then fed to the DNN AM, is also known as *noise-aware* or *speaker-aware training* correspondingly [Yu & Deng, 2014]. In paper [Yu et al., 2012] two types of factorized DNNs were introduced: *joint and disjoint models*. In [Tran et al., 2016] an extension of the LIN adaptation, so-called *factorized LIN* (FLIN), has been investigated for the case when adaptation data for a given speaker include multiple acoustic conditions. The feature transformations are represented as weighted combinations of affine transformations of the enhanced input features.

2.6. Auxiliary features

Using auxiliary features, such as i-vectors [Saon et al., 2013; Karanasou et al., 2014; Gupta et al., 2014; Senior & Lopez-Moreno, 2014], is another widely used 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. Originally i-vectors were developed for speaker verification and speaker recognition tasks [Dehak et al., 2011], and nowadays they have become a very common technique in these domains. I-vectors can capture the relevant information about the speaker in a low-dimensional fixed-length representation [Dehak et al., 2011]. They were first applied for adaptation of GMM-HMM models [Karafiát et al., 2011], and later for DNN-HMMs [Saon et al., 2013; Senior & Lopez-Moreno, 2014; Karanasou et al., 2014; Gupta et al., 2014]. Another example of auxiliary features is the use of speaker-dependent bottleneck (BN) features obtained from a speaker aware DNN used in a far field speech recognition task [Liu et al., 2014]. Alternative methods include adaptation with speaker codes [Abdel-Hamid & Jiang, 2013; Xue et al., 2014].

We will describe i-vectors in more details here because this is one of the most popular methods for DNN adaptation, and we will use this technique as a baseline for comparison with the proposed approach in our experiments in Section 4.5.3.

I-vector extraction

The acoustic feature vector $\mathbf{o}_t \in \mathbb{R}^d$ can be considered as a sample, generated with a *universal background model* (UBM), represented as a GMM with K diagonal covariance Gaussians [Dehak et al., 2011; Saon et al., 2013]:

$$\mathbf{o}_t \sim \sum_{k=1}^K c_k \mathcal{N}(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (1)$$

where c_k are the mixture weights, $\boldsymbol{\mu}_k$ are means and $\boldsymbol{\Sigma}_k$ are diagonal covariances. The acoustic feature vector $\mathbf{o}_t(s)$, belonging to a given speaker s is described with the distribution:

$$\mathbf{o}_t(s) \sim \sum_{k=1}^K c_k \mathcal{N}(\cdot; \boldsymbol{\mu}_k(s), \boldsymbol{\Sigma}_k), \quad (2)$$

where $\boldsymbol{\mu}_k(s)$ are the means of the GMM, adapted to the speaker s . It is assumed that there is a linear dependence between the speaker-dependent (SD) means $\boldsymbol{\mu}_k(s)$ and the speaker-independent (SI) means $\boldsymbol{\mu}_k$, which can be expressed in the form:

$$\boldsymbol{\mu}_k(s) = \boldsymbol{\mu}_k + \mathbf{T}_k \mathbf{w}(s), \quad k = 1, \dots, K, \quad (3)$$

where $\mathbf{T}_k \in \mathbb{R}^{D \times M}$ is a *factor loading matrix*, corresponding to component k and i-vector corresponding to speaker s is estimated as the mean of the distribution

of $\mathbf{w}(s)$. Each \mathbf{T}_k contains M bases, that span the subspace of the important variability in the component mean vector space, corresponding to component k .

The detailed description of how to estimate the factor loading matrix, given the training data $\{\mathbf{o}_t\}$, and how to estimate i-vectors $\mathbf{w}(s)$, given \mathbf{T}_k and speaker data $\{\mathbf{o}_t(s)\}$, can be found, for example, in [Dehak et al., 2011; Saon et al., 2013].

Integration of i-vectors into a DNN model

Various methods of i-vector integration into a DNN AM have been proposed in the literature.

The most common approach [Saon et al., 2013; Senior & Lopez-Moreno, 2014; Gupta et al., 2014] is to estimate i-vectors for each speaker (or utterance), and then to concatenate it with acoustic feature vectors, belonging to a corresponding speaker (or utterance). The obtained concatenated vectors are introduced to a DNN for training. In the test stage i-vectors for test speakers also have to be estimated, and input in a DNN in the same manner.

Unlike acoustic feature vectors, which are specific for each frame, an i-vector is the same for a chosen group of acoustic features, to which it is appended. For example, i-vector can be calculated for each utterance, as in [Senior & Lopez-Moreno, 2014], or estimated using all the data of a given speaker, as in [Saon et al., 2013]. I-vectors encode those effects in the acoustic signal, to which an ASR system is desired to be invariant: speaker, channel and background noise. Providing to the input of a DNN the information about these factors makes it possible for a DNN to normalize the acoustic signal with respect to them.

An alternative approach of i-vector integration into the DNN topology is presented in [Miao et al., 2015, 2014], where an input acoustic feature vector is normalized through a linear combination of it with a speaker-specific normalization vector obtained from an i-vector. Similar approaches have been studied in [Lee et al., 2016; Goo et al., 2016]. Also i-vector dependent feature space transformations were proposed in [Li & Wu, 2015].

2.7. Adaptation based on GMMs

The idea of integrating generative models into discriminate classifiers is not new. In the past, one solution to this problem was to use so-called *kernels* that are calculated using generative models [Jaakkola & Haussler, 1999; Longworth & Gales, 2009; Gales & Flego, 2010; Ragni & Gales, 2011]. Kernel methods were designed to allow classifiers (such as support vector machines (SVMs) as in [Gales & Flego, 2010; Longworth & Gales, 2009; Smith & Gales, 2002]) to handle sequential data and to map variable length data sequences into a fixed dimensional representation. There are several papers devoted to these approaches in different domains, for example, for biosequence analysis [Jaakkola & Haussler, 1999], speaker verification [Longworth & Gales, 2009], and for noise robust speech recognition [Gales & Flego, 2010; Ragni & Gales, 2011]. Using generative models to compute kernels also allows to use compensation and adaptation techniques for classifiers through the adaptation of generative kernels [Gales & Flego, 2010].

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 [Seide et al., 2011; Rath et al., 2013; Kanagawa et al., 2015; Parthasarathi et al., 2015]. In [Lei et al., 2013] 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. Similar ideas are also presented in [Swietojanski et al., 2013]. Other methods include temporally varying weight regression [Liu & Sim, 2014] and GMMD features [Tomashenko & Khokhlov, 2014; Tomashenko et al., 2016d; Tomashenko & Khokhlov, 2015].

3. Hybrid DNN-HMM systems with GMMD features

In a conventional GMM-HMM ASR system, the state emission log-likelihood of the observation feature vector \mathbf{o}_t at time t for certain tied state s_i of HMMs is modeled as

$$\log p(\mathbf{o}_t | s_i) = \log \sum_{m=1}^{M_i} w_{im} \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}), \quad (4)$$

where M_i is the number of Gaussian distributions in the GMM for state s_i ; w_{im} is the mixture weight of the m 'th component of in the mixture for state s_i ; $\boldsymbol{\mu}_{im}$ is the corresponding mean vector, and $\boldsymbol{\Sigma}_{im}$ is the covariance matrix.

In a DNN-HMM system, outputs of a DNN are the state posteriors $p(s_i | \mathbf{o}_t)$, which are transformed for decoding into pseudo (or scaled) likelihoods as follows

$$p(\mathbf{o}_t | s_i) = \frac{p(s_i | \mathbf{o}_t)p(\mathbf{o}_t)}{p(s_i)} \propto \frac{p(s_i | \mathbf{o}_t)}{p(s_i)}, \quad (5)$$

where the state prior $p(s_i)$ can be estimated from the state-level forced alignment on the training speech data, and probability $p(\mathbf{o}_t)$ is independent on the HMM state and can be omitted during the decoding process. Hence, log-likelihoods $\log p(\mathbf{o}_t | s_i)$ can be estimated as $\log p(s_i | \mathbf{o}_t) - \log p(s_i)$.

The use of log-likelihoods from a GMM model for training an MLP recognizer was investigated in [Pinto & Hermansky, 2008]. Construction of GMMD features for adapting DNNs was proposed in [Tomashenko & Khokhlov, 2014, 2015; Tomashenko et al., 2016b], where it was demonstrated, using MAP and fMLLR adaptation as an example, that this type of features provide a solution for efficient transferring GMM-HMM adaptation algorithms into the DNN framework. The same idea of using adapted GMMD features as input to DNNs was later applied to the task of noise adaptation [Kundu et al., 2016].

The GMMD features can be used directly to train DNN acoustic models, as in [Tomashenko & Khokhlov, 2014, 2015], or in combination with other conventional features. In this paper, we present incorporation of the adapted GMMD features into several state-of-the-art recipes for neural network AM training.

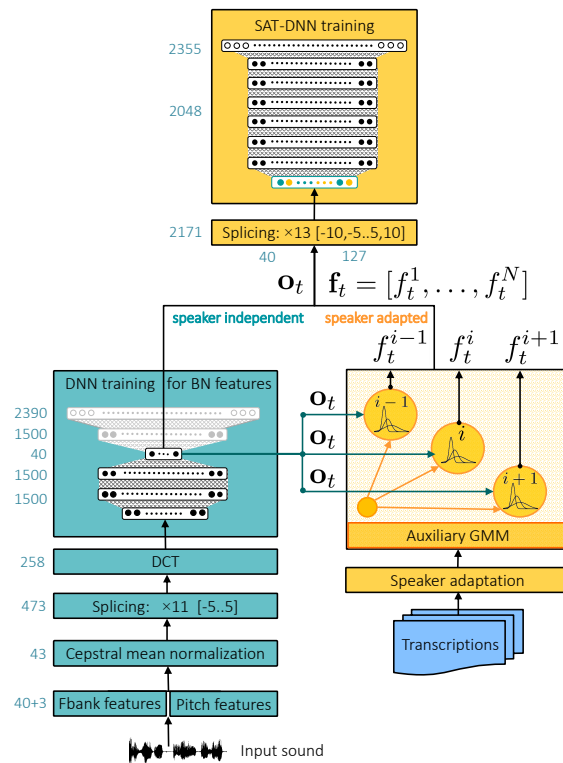


Figure 1: Using speaker adapted BN-based GMM features for speaker adaptive training (SAT) of a DNN-HMM.

3.1. Training DNN acoustic model with GMMD features

The scheme for training DNN models with GMM adaptation framework is shown in Figure 1.

First, 40-dimensional log-scale filterbank features, concatenated with 3-dimensional pitch-features, are spliced across 11 neighboring frames (5 frames on each side of the current frame), resulting in 473-dimensional (43×11) feature vectors. After that, a DCT transform is applied and the dimension is reduced to 258. Then a DNN model for 40-dimensional bottleneck (BN) features is trained on these features. An auxiliary triphone or monophone GMM model is used to transform BN feature vectors into log-likelihoods vectors. At this step, speaker adaptation of the auxiliary speaker-independent (SI) GMM-HMM model is performed for each speaker in the training corpus and a new speaker-adapted (SA) GMM-HMM model is created in order to obtain SA GMMD features.

For a given BN feature vector $\mathbf{o}_t \in \mathbb{R}^d$, a new GMMD feature vector \mathbf{f}_t is obtained by calculating log-likelihoods across all the states of the auxiliary GMM model on the given vector as follows:

$$\mathbf{f}_t = [f_t^1, \dots, f_t^N], \quad (6)$$

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

$$f_t^i = \log(p(\mathbf{o}_t | s_t = i)) \quad (7)$$

is the log-likelihood estimated using the GMM. Here s_t denotes the state index at time t . Formula (7) for the i -th component of the GMMD feature vector \mathbf{f}_t can be rewritten (using the notations from Formula (4)) as follows:

$$f_t^i = \log \sum_{m=1}^{M_i} \frac{w_{im}}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{im}|}} \exp \left\{ -\frac{1}{2} (\mathbf{o}_t - \boldsymbol{\mu}_{im})^T \boldsymbol{\Sigma}_{im}^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_{im}) \right\}. \quad (8)$$

The obtained GMMD feature vector \mathbf{f}_t is concatenated with the original vector \mathbf{o}_t . After that, the features are spliced in time taking a context size of 13 frames: $[-10, -5..5, 10]$ ¹. These features are used as the input for speaker adaptive training (SAT) of a DNN. The proposed approach can be considered a feature space transformation technique with respect to DNN-HMMs trained on GMMD features.

Note, that the proposed GMMD features are very different from i-vectors (Section 2.6) in several aspects despite the fact that both these methods are GMM-related.

First, i-vectors represent the acoustic characteristics of the speaker with respect to the general speaker distribution, which is characterized by a UBM

¹The notation $[-10, -5..5, 10]$ means that for the current acoustic vector \mathbf{o}_t , in the DNN training we use a context vector which consists of the following 13 frames: $\{\mathbf{o}_{t-10}, \mathbf{o}_{t-5}, \mathbf{o}_{t-4}, \dots, \mathbf{o}_t, \dots, \mathbf{o}_{t+4}, \mathbf{o}_{t+5}, \mathbf{o}_{t+10}\}$.

(Formulas (1),(2)). GMMD features, when they are adapted, represent the speaker-adapted distributions of acoustic classes. The better a GMMD vector is adapted, the closer these distributions are to speaker-dependent ones, and the higher likelihoods for the acoustic vectors of a corresponding speaker.

Second, i-vectors do not distinguish between acoustic classes, while in computation of GMMD features this information is explicitly represented by different components of GMMD feature vectors. Each component of a GMMD feature vector is adapted to more closely match to the pronunciation of a given speaker of the corresponding acoustic class.

Also, since we have more classes in GMMD features to adapt individually, this means, that in comparison with i-vectors, they can potentially benefit more from adaptation than the amount of adaptation data increases, especially when we use such adaptation techniques for GMMs as MAP.

Finally, i-vectors are usually computed for a sequences of vectors (per speaker, per utterance, or for a shorter time interval), while a GMMD feature vector is unique for each speech frame.

All these differences can also allow us to suggest that both approaches can be complementary to each other. We will experimentally explore this question in Section 4.

3.2. Training TDNN acoustic model with GMMD features

In addition to the system described in Section 3.1, we aim to explore the effectiveness of using GMMD features to train a time delay neural network (TDNN) [Waibel et al., 1989]. A TDNN model architecture allows to capture the long term dependencies in speech signal. The recently proposed approaches to train TDNN acoustic models [Peddinti et al., 2015] are reported to show higher performance on different LVCSR tasks compared with the standard (best) DNN systems. We aim to incorporate GMMD features into the existing state-of-the-art recipe for TDNN models [Peddinti et al., 2015]. For comparison purposes, we take a Kaldi TED-LIUM recipe with a TDNN acoustic model as a basis. An example of using GMMD features for training a TDNN is shown in Figure 2. Here, as before, we use BN features to train the GMM auxiliary model for GMMD feature extraction. Then, GMMD features are obtained in the same way as described in Section 3.1.

There are several options for obtaining the final features, which are fed to the TDNN model. GMMD features can be combined with the original MFCCs or with BNs, that are used for training the auxiliary GMM model, as shown in Figure 2. In both cases, we can also use speaker i-vectors as complementary auxiliary features. All these possibilities will be explored in Section 4.5.

3.3. MAP adaptation

In this work, we use the MAP adaptation algorithm [Gauvain & Lee, 1994] in order to adapt 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 model which is used for calculating GMMD features. Let m

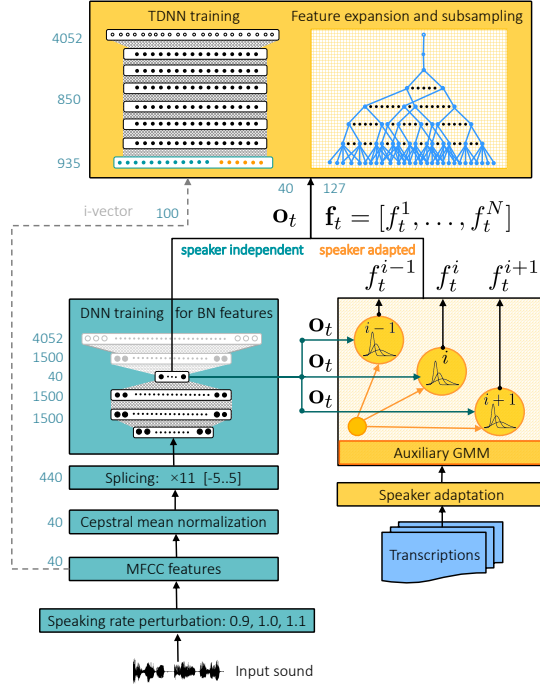


Figure 2: Using speaker adapted BN-based GMMD features for SAT TDNN training.

denote an index of a Gaussian in the SI AM, and $\boldsymbol{\mu}_m$ the mean of this Gaussian. Then the MAP estimation of the mean vector is

$$\hat{\boldsymbol{\mu}}_m = \frac{\tau \boldsymbol{\mu}_m + \sum_t \gamma_m(t) \mathbf{o}_t}{\tau + \sum_t \gamma_m(t)}, \quad (9)$$

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 .

4. Experimental study

4.1. Data sets

The experiments were conducted on the TED-LIUM corpus [Rousseau et al., 2014]. We used the last (second) release of this corpus. This publicly available data set contains 1495 TED talks that amount to 207 hours (141 hours of male, 66 hours of female) speech data from 1242 speakers, 16kHz. For experiments with SAT and adaptation we removed from the original corpus data for those speakers, who had less than 5 minutes of data, and from the rest of the corpus we made four data sets: training set, development set and two test sets. Characteristics of the obtained data sets are given in Table 1.

Table 1: Data sets statistics

Characteristic		Data set			
		Training	Development	Test ₁	Test ₂
Duration, hours	Total	171.66	3.49	3.49	4.90
	Male	120.50	1.76	1.76	3.51
	Female	51.15	1.73	1.73	1.39
Duration per speaker, minutes	Mean	10.0	15.0	15.0	21.0
	Minimum	5.0	14.4	14.4	18.3
	Maximum	18.3	15.4	15.4	24.9
Number of speakers	Total	1029	14	14	14
	Male	710	7	7	10
	Female	319	7	7	4
Number of words	Total	-	36672	35555	51452

For evaluation, two different language models (LMs) are used:

- *LM-cantab* is a publicly available 3-gram language model *cantab-TEDLIUM-pruned.lm3²* with 150K word vocabulary. The same LM was used in experiments presented in [Tomashenko et al., 2016a] and in the Kaldi *tedlium s5* recipe.
- *LM-lium* is a 4-gram LM from TED-LIUM corpus with 152K word vocabulary, which is currently used in the Kaldi *tedlium s5_r2* recipe. We conducted part of the experiments presented here using this LM in order to be compatible with the most recent Kaldi recipe and for comparison purposes with the results of the TDNN acoustic models

4.2. System fusion

In this section, we introduce several types of combination of GMMD features with conventional ones at different levels of DNN architecture. It is known that GMM and DNN models can be complementary and their combination allows to improve the performance of ASR systems [Pinto & Hermansky, 2008; Swietojanski et al., 2013]. We explore the following types of fusion:

- *Feature level fusion* (Figure 3a), where input features are combined before performing classification [Pinto & Hermansky, 2008]. In our case, features of different types – GMMD and cepstral or BN features are simply concatenated and provided as input into the DNN model for training. This type of fusion allows us to combine different adaptation techniques in a single DNN model.
- *Posterior (state) level fusion* (Figure 3b), where the outputs of two or more DNN models are combined on a state level [Parthasarathi et al., 2015; Pinto & Hermansky, 2008; Lei et al., 2013; Swietojanski et al., 2013]. In this work, we

²<http://cantabresearch.com/cantab-TEDLIUM.tar.bz2>

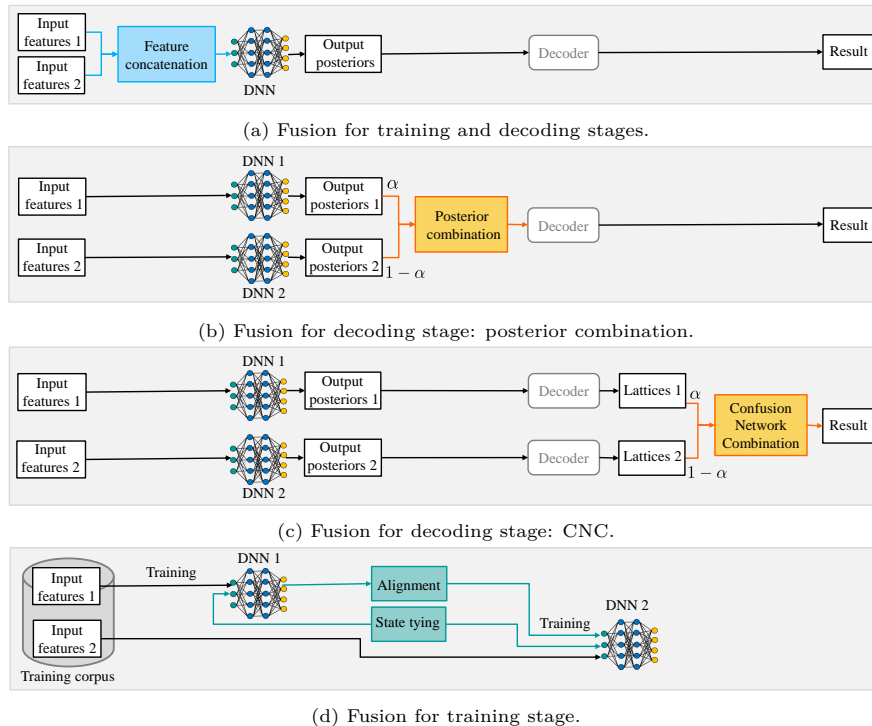


Figure 3: Types of fusion

perform frame-synchronous fusion using a linear combination of the observation log-likelihoods of two models (DNN_1 and DNN_2) as follows:

$$\log(p(\mathbf{o}_t | s_i)) = \alpha \log(p_{DNN_1}(\mathbf{o}_t | s_i)) + (1 - \alpha) \log(p_{DNN_2}(\mathbf{o}_t | s_i)), \quad (10)$$

where $\alpha \in [0, 1]$ is a weight factor that is optimized on a development set. This approach assumes that both models have the same state tying structure.

- *Lattice level fusion* (Figure 3c) is the highest level of fusion operates in the space of generated word hypotheses [Fiscus, 1997; Evermann & Woodland, 2000]. In this work, we experiment with the Confusion Network Combination (CNC) [Evermann & Woodland, 2000] approach, where confusion networks built from individual lattices are aligned.

There are other possible ways of combining information from different ASR systems. In this paper, we used phoneme-to-speech alignment obtained by one acoustic DNN model to train another DNN model (Figure 3d). In addition, we used state tying from the first DNN model to train the second DNN. This procedure is important when we want to apply *posterior fusion* for two DNNs and need the same state tying for these models.

4.3. Overview of experiments and questions addressed in the study

We used the open-source Kaldi toolkit [Povey et al., 2011] and mostly followed the standard TED-LIUM Kaldi recipes to train the two baseline systems, corresponding to two different types of acoustic models (DNN and TDNN)³.

The following questions have been addressed in this section.

First, in order to explore and prove the universality of the proposed adaptation approach, we start our study from the classical fully connected DNN topology, and then choose TDNNs as one of the most efficient neural network architecture for AMs.

Second, we are interested in comparing the proposed adaptation technique with the two most popular adaptation approaches for neural network AMs: fMLLR (experiments for DNNs) and i-vectors (experiments for TDNNs).

Third, we study different types of fusion (features-, posterior-, lattice-level) of GMM features with other features (both SI and adapted ones) for DNNs and TDNNs. We aim to explore the complementarity of the proposed adaptation technique to other adaptation approaches: fMLLR (for DNNs) and i-vectors (for TDNNs).

Finally, the impact of the training criterion on the adaptation performance is investigated.

4.4. Experiments with DNN models

This section presents systems and adaptation results for DNN AMs.

4.4.1. Baseline systems

AMs in this series of experiments are DNNs trained on BN features, and for the baseline with speaker adaptation we used fMLLR adaptation. For these models, we also used two different training criteria: cross-entropy (CE) criterion and sequence-discriminative training with Minimum Bayes Risk (sMBR) criterion in order to study the impact of the training criterion on the adaptation performance. Hence, we trained four baseline DNN AMs (see Appendix A for details):

- **DNN_{BN-CE}**: BN features, CE criterion;
- **DNN_{BN-sMBR}**: BN features, sMBR criterion;
- **DNN_{BN-fMLLR-CE}**: fMLLR-adapted BN features, CE criterion;
- **DNN_{BN-fMLLR-sMBR}**: fMLLR-adapted BN features, sMBR criterion.

LM-cantab was used for decoding.

³using "nnet1" and "nnet3" Kaldi setups: <http://kaldi-asr.org/doc/dnn.html>

Table 2: Summary of the adaptation results for DNN models. The results in parentheses correspond to WER of the consensus hypothesis.

#	Features	DNN	WER,%		
			Development	Test ₁	Test ₂
1	BN	CE	13.16	11.94	15.43
2	BN	sMBR	12.14	10.77	13.75
3	BN-fMLLR	CE	11.72	10.88	14.21
4	BN-fMLLR	sMBR	10.64 (10.57)	9.52 (9.46)	12.78 (12.67)
5	GMMD \oplus BN	CE	12.92	11.62	15.19
6	GMMD \oplus BN	sMBR	11.80	10.47	13.52
7	GMMD-MAP \oplus BN	CE	10.46	9.74	13.03
8	GMMD-MAP \oplus BN	sMBR	10.26 (10.23)	9.40 (9.31)	12.52 (12.46)

4.4.2. Proposed systems with speaker-adapted GMMD features

For experiments with speaker adaptation, we trained four acoustic models using the approaches proposed in Sections 3.1.

- **DNN_{GMMD \oplus BN-CE}**: speaker-independent (SI) GMMD features appended with BNs, CE criterion;
- **DNN_{GMMD \oplus BN-sMBR}**: SI GMMD features appended with BNs, sMBR criterion;
- **DNN_{GMMD-MAP \oplus BN-CE}**: SAT DNN model trained on speaker adapted GMMD-MAP features, CE criterion.
- **DNN_{GMMD-MAP \oplus BN-sMBR}**: on speaker adapted GMMD-MAP features, sMBR criterion.

Models **DNN_{GMMD-MAP \oplus BN-CE}** and **DNN_{GMMD-MAP \oplus BN-sMBR}** were trained as described in Section 3.1. The GMMD features were extracted using a monophone auxiliary GMM model, trained on BN features. This GMM model was adapted for each speaker by MAP adaptation algorithm (Section 3.3). See more details in Appendix C.

4.4.3. Adaptation and fusion results

The adaptation experiments were conducted in an unsupervised mode on the test data using transcripts from the first decoding pass obtained by the baseline SAT-DNN model, unless explicitly stated otherwise.

We empirically studied different types of fusion described in Section 4.2 and applied them to DNN models trained using GMMD-features extracted as proposed in Section 3. The performance results in terms of WER for SI and SAT DNN-HMM models are presented in Table 2. The first four lines of the table correspond to the baseline SI (#1, #2) and SAT (#3, #4) DNNs, which were trained as described in Section 4.4.1.

Table 3: Summary of the fusion results for DNN models. The results in parentheses correspond to WER of the consensus hypothesis. Here \downarrow denotes relative WER reduction (for consensus hypothesis) in comparison with AM trained on BN-fMLLR (#4 in Table 2). The bold figures in the table indicate the best performance improvement.

#	Fusion: #4 and #8	WER,%		
		Development	Test ₁	Test ₂
9	Posterior fusion, $\alpha = 0.45$	9.98 (9.91) \downarrow 6.2	9.15 (9.06) \downarrow 4.3	12.11 (12.04) \downarrow 5.0
	Lattice fusion, $\alpha = 0.46$	10.06 \downarrow 4.8	9.09 \downarrow 4.0	12.12 \downarrow 4.4

Parameter τ in MAP adaptation, that controls the balance between the maximum likelihood estimate of the mean and its prior value [Gauvain & Lee, 1994; Tomashenko & Khokhlov, 2014], for both acoustic model training and decoding was set equal to 5.

For comparison purpose with lattice-based fusion we report WER of the consensus hypothesis in parentheses for experiments #4 and #8.

After that we made posterior fusion of the obtained model #8 and the baseline SAT-DNN model (#4). The result is given in Table 3, line #9. Value α in Formula (10) (Section 4.2) is a weight of the baseline SAT-DNN model. Parameter α was optimized on the development set.

Finally, we applied lattice fusion for the same pair of models (line #10). In this type of fusion, before merging lattices, for each edge, scores were replaced by its a posteriori probabilities. Posteriors were computed for each lattice independently. The optimal normalizing factors for each model were found independently on the development set. Then the two lattices were merged into a single lattice and posteriors were weighted using parameter α . As before, value α in Formula (10) corresponds to the baseline SAT-DNN model. The resulting lattice was converted into the CN and the final result was obtained from this CN.

We can see, that both - posterior and lattice types of fusion provide similar improvement for all three models: approximately 4%–6% of relative WER reduction (WERR) in comparison with the adapted baseline model (SAT DNN on fMLLR features, #4), and 12%–18% of relative WERR in comparison with the SI baseline model (#2). For models #7–8 only MAP adaptation was applied. Experiments #9–10 present combination of two different adaptation types: MAP and fMLLR. It is interesting to note that in all experiments optimal value of α is close to 0.5, so all types of models are equally important for fusion. We can see that MAP adaptation on GMMD features can be complementary to fMLLR adaptation on conventional BN features.

4.5. Experiments with TDNN models

In this section, we expand our experimental study to the TDNN topology.

4.5.1. Baseline system

We trained four baseline TDNN acoustic models, which differ only in the type of the input features (see Appendix B for details):

- $\text{TDNN}_{\text{MFCC}}$: high-resolution MFCC features;
- $\text{TDNN}_{\text{MFCC} \oplus \text{i-vectors}}$: high-resolution MFCC features appended with 100-dimensional i-vectors;
- TDNN_{BN} : BN features;
- $\text{TDNN}_{\text{BN} \oplus \text{i-vectors}}$: BN features appended with 100-dimensional i-vectors.

LM-lium was used for decoding.

4.5.2. Proposed systems with speaker-adapted GMMD features

Four TDNNs were trained using GMMD features as proposed in Section 3.2:

- $\text{TDNN}_{\text{MFCC} \oplus \text{GMMD}}$: high-resolution MFCC features appended with speaker adapted GMMD features;
- $\text{TDNN}_{\text{MFCC} \oplus \text{GMMD} \oplus \text{i-vectors}}$: high-resolution MFCC features appended with speaker adapted GMMD features and 100-dimensional i-vectors;
- $\text{TDNN}_{\text{BN} \oplus \text{GMMD}}$: BN features appended with speaker adapted GMMD features;
- $\text{TDNN}_{\text{BN} \oplus \text{i-vectors} \oplus \text{GMMD}}$ is a version of the $\text{TDNN}_{\text{BN} \oplus \text{GMMD}}$ with 100-dimensional i-vectors.

All the four TDNN models were trained in the same manner, as the baseline TDNN model, and differ only in the type of the input features.

4.5.3. Results for TDNN models

In this set of experiments, for adaptation of the proposed models without i-vectors, we used the first decoding output made by the baseline model, which is also without i-vectors (TDNN_{BN}).

Adaptation results for the TDNN models are given in Table 4. The best result was obtained by $\text{TDNN}_{\text{GMMD} \oplus \text{BN} \oplus \text{i-vectors}}$ (line #8). For Development set it gives 7.4% of relative WERR over $\text{TDNN}_{\text{MFCC} \oplus \text{i-vectors}}$, though for the other test sets the result is very close to the baseline (line #2) for the models which are already MAP-adapted. If we compare (#5, #6) or (#7, #8), we can see that i-vectors always give an additional improvement for the model which is already MAP-adapted.

To further investigate the complementarity of the two different adaptation techniques, we performed CNC of recognition results for different TDNN models (Table 5). The best result, obtained by combination of TDNN models #2 and #8, provides approximately 7-13% of relative WERR in comparison with the baseline model $\text{TDNN}_{\text{MFCC} \oplus \text{i-vectors}}$.

Table 4: Summary of the adaptation results for TDNN models. The results in parentheses correspond to WER of the consensus hypothesis. The bold figures in the table indicate the best performance improvement.

#	Features	WER, %		
		Development	Test ₁	Test ₂
1	MFCC	11.98 (11.88)	9.42 (9.31)	12.77 (12.66)
2	MFCC \oplus i-vectors	10.16 (10.12)	7.98 (7.90)	11.73 (11.70)
3	BN	11.68 (10.62)	8.83 (8.76)	12.44 (12.41)
4	BN \oplus i-vectors	10.05 (9.98)	8.29 (8.21)	11.90 (11.88)
5	GMMD \oplus MFCC	9.79 (9.70)	8.26 (8.17)	12.21 (12.16)
6	GMMD \oplus MFCC \oplus i-vectors	9.35 (9.32)	8.10 (8.05)	11.98 (11.95)
7	GMMD \oplus BN	9.57 (9.52)	8.18 (8.13)	11.92 (11.87)
8	GMMD \oplus BN \oplus i-vectors	9.41 (9.34)	7.85 (7.74)	11.70 (11.65)

Table 5: Summary of the fusion results (CNC) for TDNN models. Here \downarrow denotes relative WER reduction (for consensus hypothesis) in comparison with the baseline **TDNN_{MFCC \oplus i-vectors}**, α is a weight of TDNN₁ in the fusion. The bold figures in the table indicate the best performance improvement.

#	TDNN ₁	TDNN ₂	α	WER, %		
				Development	Test ₁	Test ₂
9	2	4	0.50	9.34 \downarrow 7.7	7.52 \downarrow 4.8	11.20 \downarrow 4.2
10	4	8	0.31	9.29 \downarrow 8.2	7.68 \downarrow 2.8	11.32 \downarrow 3.3
11	8	6	0.33	9.28 \downarrow 8.3	7.81 \downarrow 1.2	11.54 \downarrow 1.4
12	4	6	0.38	9.22 \downarrow 8.9	7.85 \downarrow 0.7	11.39 \downarrow 2.7
13	2	7	0.49	8.96 \downarrow 11.5	7.33 \downarrow 7.3	10.95 \downarrow 6.4
14	2	6	0.50	8.92 \downarrow 11.8	7.33 \downarrow 7.3	10.99 \downarrow 6.1
15	2	8	0.46	8.84 \downarrow 12.7	7.28 \downarrow 7.9	10.91 \downarrow 6.8

5. Feature analysis

The objective of this section is to analyze the proposed GMMD features and the adaptation algorithm for better understanding their nature and properties at different levels.

5.1. Lattice-based features

Lattice based features or time dependent state posterior scores [Uebel & Woodland, 2001; Gollan & Bacchiani, 2008] are obtained from computing arc posteriors from the output lattices of the decoder. These features contain more information about the decoding process than the posterior probabilities from neural networks because in their extraction also language model probabilities, likelihoods of decoding hypothesis, and other information are taken into account. We use this type of features to analyze the quality of the adaptation of TDNN acoustic models.

Let $\{ph_1, \dots, ph_M\}$ be a set of phonemes and the silence model. For each time frame t , we calculate p_t^m — the confidence score of phoneme ph_m ($1 \leq m \leq M$) at time t in the decoding lattice by calculating arc posterior probabilities. The forward-backward algorithm is used to calculate these arc posterior probabilities from the lattice as follows:

$$p(l|O) = \frac{\sum_{q \in Q_l} p_{ac}(O|q)^{\frac{1}{\lambda}} p_{lm}(w)}{p(O)}, \quad (11)$$

where λ is the scale factor (the optimal value of λ is found empirically by minimizing WER of the consensus hypothesis [Mangu et al., 2000]); q is a path through the lattice corresponding to the word sequence w ; Q_l is the set of paths passing through arc l ; $p_{ac}(O|q)$ 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.

For the given frame \mathbf{o}_t at time t , we calculate its probability $p(\mathbf{o}_t) \in ph_m$ of belonging to phoneme ph_m , using lattices obtained from the first decoding pass:

$$p_t^m = p(\mathbf{o}_t \in ph_m) = \sum_{l \in S_m(\mathbf{o}_t)} p(l|O), \quad (12)$$

where $S_m(\mathbf{o}_t)$ is the set of all arcs corresponding to the phoneme ph_m in the lattice at time t ; $p(l|O)$ is the posterior probability of arc l in the lattice.

The obtained probability $p(\mathbf{o}_t \in ph_m)$ of frame \mathbf{o}_t belonging to phoneme ph_m is the component value p_t^m on the new feature vector \mathbf{p}_t . Thus for a given acoustic feature vector \mathbf{o}_t at time t we obtain a new feature vector \mathbf{p}_t :

$$\mathbf{p}_t = (p_t^1, \dots, p_t^M), \quad (13)$$

where M is the number of phones in the phoneme set used in the ASR system.

Hence for each frame \mathbf{o}_t we have a M -dimensional vector \mathbf{p}_t , each component of which represents the probability of this frame to belong to a certain phoneme.

When some phonemes are not present in the lattice for a certain frame, we set probabilities equal to some very small value ϵ for them in the vector (where ϵ is a minimum value from lattices: $\epsilon \approx 10^{-9}$). In addition to this, we use state index information (position of the state in phoneme HMM: 0, 1 or 2) from the Viterbi alignment from the original transcripts.

5.2. Davies-Bouldin index

We use the Davies-Bouldin (DB) index [Davies & Bouldin, 1979] to evaluate the quality of the phoneme state clusters obtained from the latticed-based features.

$$DB = \frac{1}{K} \sum_{k=1}^K \max_{j \neq k} \left(\frac{\sigma_k + \sigma_j}{\rho_{k,j}} \right), \quad (14)$$

where

K is the number of clusters;

σ_k is the scatter within the cluster k , which is our case the standard deviation of the distance of all vectors corresponding to cluster k , to the cluster center (other possible metric variants are described in [Davies & Bouldin, 1979]);

$\rho_{k,j}$ is a between-cluster separation measure, which in our case is the Euclidean distance between the centroids of clusters k and j .

Smaller values of DB index correspond to better clusters.

5.3. Visual analysis using t-SNE

The lattice-based features were visualized using t-distributed stochastic neighbor embedding (t-SNE) analysis [Maaten & Hinton, 2008]. This technique allows us to visualize high-dimensional data into two or three dimensional space, in such a way that the vectors, which are close in the original space, are also close in the low dimensional t-SNE representation.

We are interested in how well the different acoustic models can cluster different phoneme states. For better visualization, we used data only from inter-word phones and only from the middle state of HMM (State 1). We choose only those phonemes for which we have sufficient amount of data for analysis and perform t-SNE analysis independently on three different groups of phonemes⁴:

- *Vowels* (*UH, OW, AO, EY, ER, AA, AY, IY, EH, AE, IH, AH*);
- *Consonants-1: Liquids* (*L, R*), *Nasals* (*M, N, NG*), *Semivowels* (*W*);
- *Consonants-2: Stops* (*P, T, D, K*), *Affricates* (*CH*), *Fricatives* (*F, V, TH, S, Z, SH*).

⁴The notations are given according to the ARPabet phoneme set: <https://en.wikipedia.org/wiki/Arpabet>

5.4. Analysis for TDNN models

In this set of experiments we compare the following three TDNN acoustic models: $\text{TDNN}_{\text{MFCC}}$, $\text{TDNN}_{\text{BN} \oplus \text{i-vectors}}$ and $\text{TDNN}_{\text{BN} \oplus \text{i-vectors} \oplus \text{GMMD}}$. All the experiments described in this section (except for Figures 8 and 9) are performed using lattice-based features (Section 5.1) on the *Development* data set.

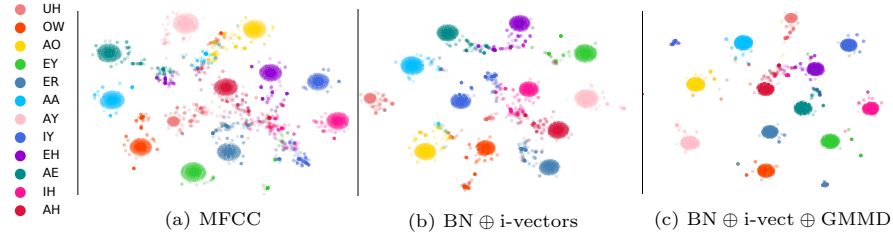


Figure 4: Analysis of lattice-based features for *vowels* using t-SNE for TDNN models trained on different basic features.

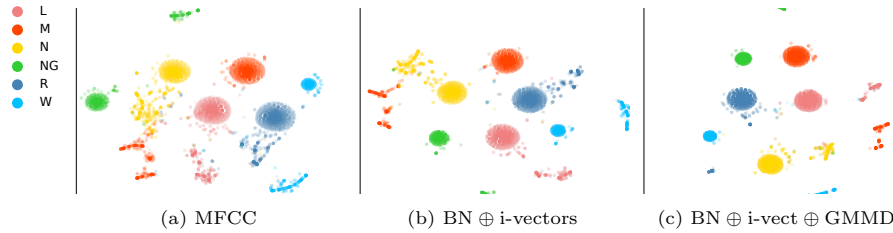


Figure 5: Analysis of lattice-based features for *consonants-1* using t-SNE for TDNN models trained on different basic features.

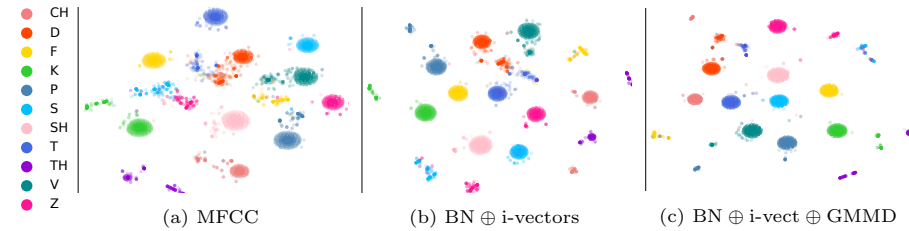


Figure 6: Analysis of lattice-based features for *consonants-2* using t-SNE for TDNN models trained on different basic features.

First we analyzed the adaptation algorithm using t-SNE (Section 5.3). The results of the visual t-SNE analysis are given in Figure 4 for the group of vowels and in Figures 5, 6 – for the two group of consonants. We can observe for

Table 6: *Davies-Bouldin (DB) index for different types of features used in TDNN training. The DB index is calculated on lattice-based features produced by the corresponding model.*

Features	State 0	State 1	State 2
MFCC	1.67	1.52	1.71
BN \oplus i-vectors	1.53	1.36	1.41
BN \oplus i-vectors \oplus GMMD	1.39	1.26	1.27

Table 7: *Statistics for lattice-based features, produced by the corresponding TDNN models. All statistics in the table are calculated only for speech frames (excluding silence). The average log-probability of the correct phoneme is given with the standard deviation.*

Features	FER	Oracle FER	Aver. correct log-prob.
MFCC	5.18	0.72	-0.17 ± 0.83
BN \oplus i-vectors	4.11	0.75	-0.11 ± 0.64
BN \oplus i-vectors \oplus GMMD	3.64	1.23	-0.08 ± 0.52

all groups of phonemes that the adapted features (Figures 4b, 5b, 6b) form more distinct and clear phone clusters than the unadapted features (Figures 4a, 5a, 6a). Also we can note the use of GMMD features helps to further slightly improve cluster separability (Figures 4c, 5c, 6c).

To support this visual analysis of cluster separation, we calculated DB index (Section 5.2) for all phonemes, separately for each state type, depending on its position in phoneme HMM (State 0, 1, 2). As we can see in Table 6, DB index decreases for all HMM states when we move from unadapted (MFCC) to adapted (BN \oplus i-vectors) features. That confirms the fact that the clusters are better for adapted features. The acoustic model with the adapted GMMD features (BN \oplus i-vectors \oplus GMMD) shows the best result (the smallest value of DB index).

In order to more deeply investigate the adaptation behavior, we calculated additional statistics for lattices-based features (Table 7). Frame error rate (FER) is calculated on the phoneme level using only speech frames (excluding silence). Oracle FER was calculated also only on speech frames as follows: if the correct phoneme was not present in the list of all candidates in the lattices for a given frame, then it was considered as an error.

We can see that FER decreases when moving from the unadapted features to the adapted ones, and then to the use of the adapted GMMD features, that correlates with the WER behavior (Table 4). It is interesting to note, that Oracle FER, on the contrary, increases with the adaptation. One of the possible explanation for this unusual situation can be phonetic transcription errors which occur in the lexicon. The adapted models, which can be more sensitive to the phonetic transcription errors, can more strictly supplant, during the decoding process, hypotheses that do not match the acoustic signal.⁵

⁵For GMM-HMM models, it is known [Gollan & Bacchiani, 2008] that MAP adaptation

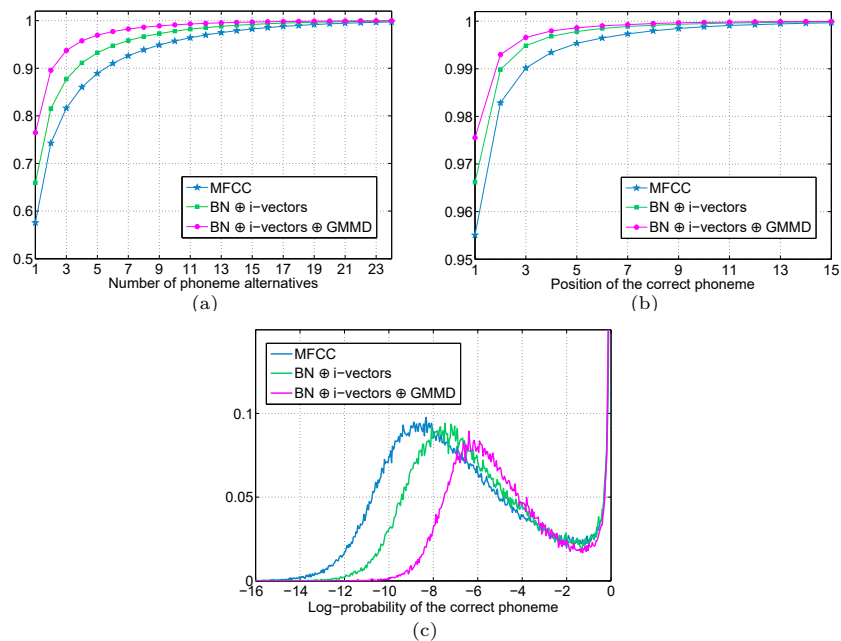


Figure 7: Adaptation analysis based on lattice-based features statistics for the free TDNN models: (a) CDF of the number of phoneme alternatives in the lattices for a certain frame; (b) CDF of the position of the correct phoneme in the list of all phoneme candidates (ordered by the posterior probability) presented in lattices for a certain frame; (c) Log-probability histogram of the correct phoneme (if it exists) in the lattice for a certain frame.

Decoding parameters, such as *decoding beam* and *lattice beam*, were the same for all models, but the adapted models in average have less alternative phoneme candidates for a certain frame, than the unadapted one. This can be seen in Figure 7a, which shows the *cumulative distribution functions* (CDFs) of the number of phoneme alternatives presented in the lattices for a certain frame, estimated only for speech frames. Figure 7b demonstrates CDFs of position of the correct phoneme (if it exists) in lattices for a certain speech frame in the list of all phoneme candidates ordered by their posterior probabilities. We can conclude from this figure that for adapted models (especially for the AM with GMMD features), the correct candidate has less incorrect alternatives with higher probabilities than its own.

Also, the average *correct log-probability* (it is a value from a lattice based features vector, which corresponds to the correct phoneme for a given frame) has a maximum value for $\mathbf{TDNN}_{\mathbf{BN} \oplus \mathbf{i-vectors} \oplus \mathbf{GMMD}}$ model (see the last column of Table 4 and a histogram on Figure 7c).

Hence, if we compare the statistics presented in Table 7 and in Figure 7, we can conclude that the adapted models tend to be more "selective" and "discriminative" in comparison with unadapted models in the sense that: (1) they reduce the number of alternatives in the hypothesis more aggressively; (2) they give higher probability values for correct candidates; and (3) the correct phoneme candidate, if it exists in the lattice for a given frame, has in average, less incorrect competitor alternatives with higher probabilities than its own. The AM trained on the adapted GMMD features most strongly shows the same properties.⁶

This analysis demonstrates that the acoustic models trained with the proposed adapted GMMD features perform better than the baseline adapted model not only by comparing the WER (which is the main metric), but also on the other levels.

Also, what is important, this gives us an understanding of the possible way of the adaptation performance improvement through more careful handling of the transcripts, for example, by automatically estimation their quality and reliability.

In addition, Figure 8 shows the statistics obtained for 42 speakers from *Development*, *Test₁* and *Test₂* data sets for $\mathbf{TDNN}_{\mathbf{BN}}$, $\mathbf{TDNN}_{\mathbf{BN} \oplus \mathbf{GMMD}}$ models. We can observe in Figure 8a that the proposed adaptation approach improves recognition accuracy for 83% of speakers. For the same speakers, Figure 8b illustrates the dependence of relative WER reduction from average likelihood improvement, obtained by MAP adaptation of the auxiliary mono-phone model.

Finally, we explored how sensitive the proposed adaptation is to the quality of the transcriptions from the first decoding pass. This is an important

can reinforce errors present in transcriptions used for adaptation.

⁶In [Woodland, 2001; Woodland et al., 1996] for GMM-HMM acoustic models and MLLR adaptation, it was also noticed that adaptation allows to obtain smaller and more accurate lattices.

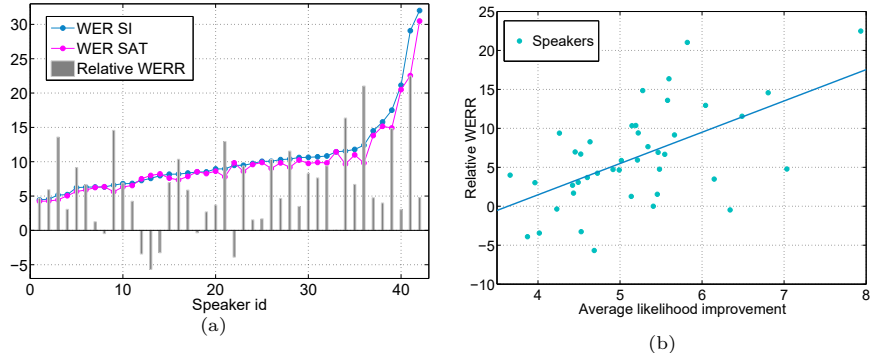


Figure 8: Summary statistics for all speakers from the development and the two test data sets: (a) WERs(%) for two (SI and SAT) TDNN models: SI – TDNN_{BN} , SAT – $\text{TDNN}_{\text{BN} \oplus \text{GMM}}$. Relative WER reduction (Relative WERR, %) is computed for the given WERs. Results are ordered by increasing WER values for the SI model; (b) Dependence of relative WERR (the same as in (a)) on average likelihood improvement, obtained by MAP adaptation of the auxiliary monophone model. The line corresponds to the linear regression model.

aspect of adaptation algorithms. For GMM-HMM models, it is known, that transform-based adaptation approaches, limited in the number of free adaptation parameters, are robust transcription errors [Gollan & Bacchiani, 2008]. For DNN unsupervised adaptation, this problem was investigated for LHUC adaptation in [Swietojanski et al., 2016] where it was concluded that the LHUC algorithm is not very sensitive to the quality of adaptation targets. In our study, we varied (degraded) the quality of the transcriptions used in adaptation by using sub-optimal decoding parameters (*word insertion penalty* and *language model weight*) in the first decoding pass and performed adaptation of $\text{TDNN}_{\text{BN} \oplus \text{i-vectors} \oplus \text{GMM}}$ using these different targets. The results of this experiment are presented in Figure 9. We can observe that changes in WERs in the first decoding pass lead to the changes in the quality of the adapted AM. However, these changes are not so dramatic.

6. Conclusions

In this paper we have investigated the GMM framework for adaptation of DNN-HMM acoustic models and combination of MAP-adapted GMM-derived with conventional features at different levels of DNN and TDNN architectures.

Experimental results on the TED-LIUM corpus demonstrate that, in an unsupervised adaptation mode, the proposed adaptation and fusion techniques can provide approximately, a 12–18% relative WERR on different adaptation sets, compared to the SI DNN system built on conventional features, and a 4–6% relative WERR compared to the strong adapted baseline — SAT-DNN trained on fMLLR adapted features. For TDNN models using the adapted GMM

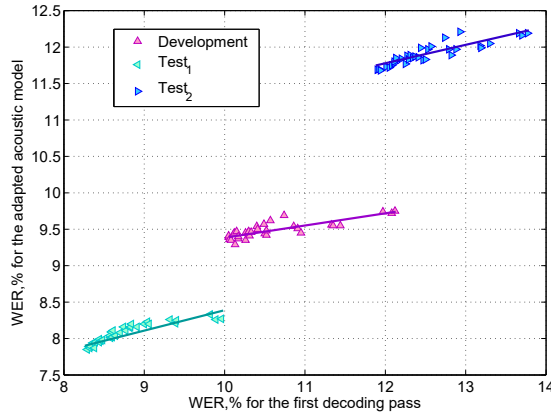


Figure 9: Dependence of WER,% (for the adapted model $\text{TDNN}_{\text{BN}} \oplus \text{i-vectors} \oplus \text{GMM}$) from the quality of the adaptation targets for the development and the two test data sets.

features and fusion techniques leads to improvement of 10–26% WERR in comparison with SI model trained on conventional features and 7–13% WERR in comparison with SAT model trained with i-vectors. Hence, for both considered adaptation techniques, fMLLR and i-vectors, the proposed adaptation approach has appeared to be complementary and provide an additional improvement in recognition accuracy.

We have looked from the various points of view at the proposed adaptation approach exploring the latticed-based features, generated from the decoding lattices and have demonstrated, that the advantage of using MAP-adapted GMM features manifests itself at different levels of the decoding process. This analysis also shows a possible potential and direction for improvement of the proposed adaptation approach through more careful handling of quality of the phonetics transcripts, used in adaptation. This will be a focus of our future work.

References

- Abdel-Hamid, O., & Jiang, H. (2013). Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7942–7946).
- Albesano, D., Gemello, R., Laface, P., Mana, F., & Scanzio, S. (2006). Adaptation of artificial neural networks avoiding catastrophic forgetting. In *Proc. IJCNN'06* (pp. 1554–1561). IEEE.
- Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE*

- International Conference on Acoustics, Speech and Signal Processing* (pp. 8609–8613).
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*, 224–227.
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 19*, 788–798.
- Demuyne, K., & Triefenbach, F. (2013). Porting concepts from DNNs back to GMMs. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on* (pp. 356–361). IEEE.
- Dupont, S., & Cheboub, L. (2000). Fast speaker adaptation of artificial neural networks for automatic speech recognition. In *Proc. ICASSP* (pp. 1795–1798). IEEE volume 3.
- Evermann, G., & Woodland, P. (2000). Posterior probability decoding, confidence estimation and system combination. In *Proc. Speech Transcription Workshop*. Baltimore volume 27.
- Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU* (pp. 347–354). IEEE.
- Gales, M., & Flego, F. (2010). Discriminative classifiers with adaptive kernels for noise robust speech recognition. *Computer Speech & Language, 24*, 648–662.
- Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech and language, 12*, 75–98.
- Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Trans. Speech and Audio Proc., 2*, 291–298.
- Gemello, R., Mana, F., Scanzio, S., Laface, P., & De Mori, R. (2006). Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training. In *Proc. ICASSP* (p. 1189–1192).
- Gollan, C., & Bacchiani, M. (2008). Confidence scores for acoustic model adaptation. In *Proc. ICASSP* (pp. 4289–4292).
- Goo, J., Kim, Y., Lim, H., & Kim, H. (2016). Speaker normalization through feature shifting of linearly transformed i-vector. In *Interspeech 2016* (pp. 3489–3493).

- Grézl, F., Karafiát, M., Kontár, S., & Cernocký, J. (2007). Probabilistic and bottle-neck features for LVCSR of meetings. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* (pp. IV–757). IEEE volume 4.
- Gupta, V., Kenny, P., Ouellet, P., & Stafylakis, T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *Proc. ICASSP* (pp. 6334–6338). 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* (pp. 1635–1638). IEEE volume 3.
- Huang, Y., & Gong, Y. (2015). Regularized sequence-level deep neural network model adaptation. In *INTERSPEECH* (pp. 1081–1085).
- Huang, Z., Li, J., Siniscalchi, S. M., Chen, I.-F., Weng, C., & Lee, C.-H. (2014). Feature space maximum a posteriori linear regression for adaptation of deep neural networks. In *Proc. INTERSPEECH* (pp. 2992–2996).
- Huang, Z., Li, J., Siniscalchi, S. M., Chen, I.-F., Wu, J., & Lee, C.-H. (2015a). Rapid adaptation for deep neural networks through multi-task learning. In *Proc. INTERSPEECH* (pp. 2329–2329).
- Huang, Z., Siniscalchi, S. M., Chen, I.-F., Li, J., Wu, J., & Lee, C.-H. (2015b). Maximum a posteriori adaptation of network parameters in deep models. In *Proc. INTERSPEECH* (pp. 1076–1080).
- Jaakkola, T., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems* (pp. 487–493).
- Kanagawa, H., Tachioka, Y., Watanabe, S., & Ishii, J. (2015). Feature-space structural MAPLR with regression tree-based multiple transformation matrices for DNN. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 86–92). IEEE.
- Karafiát, M., Burget, L., Matějka, P., Glembek, O., & Černocký, J. (2011). i-vector-based discriminative adaptation for automatic speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on* (pp. 152–157). IEEE.
- Karanasou, P., Wang, Y., Gales, M. J., & Woodland, P. C. (2014). Adaptation of deep neural network acoustic models using factorised i-vectors. In *Proc. INTERSPEECH* (pp. 2180–2184).
- Kuhn, R., Junqua, J.-C., Nguyen, P., & Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8, 695–707.

- Kundu, S., Sim, K. C., & Gales, M. J. (2016). Incorporating a generative front-end layer to deep neural network for noise robust automatic speech recognition. In *Interspeech 2016* (pp. 2359–2363).
- Lee, W., Han, K. J., & Lane, I. (2016). Semi-supervised speaker adaptation for in-vehicle speech recognition with deep neural networks. In *Interspeech 2016* (pp. 3843–3847).
- Lei, X., Lin, H., & Heigold, G. (2013). Deep neural networks with auxiliary Gaussian mixture models for real-time speech recognition. In *Proc. ICASSP* (pp. 7634–7638). IEEE.
- Li, B., & Sim, K. C. (2010). Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In *Proc. INTERSPEECH* (pp. 526–529).
- Li, J., Huang, J.-T., & Gong, Y. (2014). Factorized adaptation for deep neural network. In *Proc. ICASSP* (pp. 5537–5541). IEEE.
- Li, S., Lu, X., Akita, Y., & Kawahara, T. (2015). Ensemble speaker modeling using speaker adaptive training deep neural network for speaker adaptation. In *Proc. INTERSPEECH* (pp. 2892–2896).
- Li, X., & Wu, X. (2015). I-vector dependent feature space transformations for adaptive speech recognition. In *Interspeech* (pp. 3635–3639).
- Liao, H. (2013). Speaker adaptation of context dependent deep neural networks. In *Proc. ICASSP* (pp. 7947–7951). IEEE.
- Liu, S., & Sim, K. C. (2014). On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition. In *Proc. ICASSP* (pp. 195–199). IEEE.
- Liu, Y., Zhang, P., & Hain, T. (2014). Using neural network front-ends on far field multiple microphones based speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5542–5546).
- Longworth, C., & Gales, M. J. (2009). Combining derivative and parametric kernels for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*, 748–757.
- Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.
- Mangu, L., Brill, E., & Stolcke, A. (2000). Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, *14*, 373–400.

- Miao, Y., Zhang, H., & Metze, F. (2014). Towards speaker adaptive training of deep neural network acoustic models. In *Proc. INTERSPEECH* (pp. 2189–2193).
- Miao, Y., Zhang, H., & Metze, F. (2015). Speaker adaptive training of deep neural network acoustic models using i-vectors. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *23*, 1938–1949.
- Murali Karthick, B., Kolhar, P., & Umesh, S. (2015). Speaker adaptation of convolutional neural network using speaker specific subspace vectors of SGMM. In *Proc. INTERSPEECH* (pp. 1096–1100).
- Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., & Robinson, T. (1995). Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. In *EUROSPEECH* (pp. 2171–2174).
- Ochiai, T., Matsuda, S., Lu, X., Hori, C., & Katagiri, S. (2014). Speaker adaptive training using deep neural networks. In *Proc. ICASSP* (pp. 6349–6353). IEEE.
- 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 *Proc. INTERSPEECH* (pp. 3630–3634).
- Paulik, M. (2013). Lattice-based training of bottleneck feature extraction neural networks. In *Interspeech* (pp. 89–93).
- Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTER-SPEECH* (pp. 3214–3218).
- Pinto, J. P., & Hermansky, H. (2008). *Combining evidence from a generative and a discriminative model in phoneme recognition*. Technical Report IDIAP.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al. (2011). The Kaldi speech recognition toolkit. In *Proc. ASRU*.
- Price, R., i. Iso, K., & Shinoda, K. (2014). Speaker adaptation of deep neural networks using a hierarchy of output layers. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 153–158).
- Qian, Y., Tan, T., & Yu, D. (2016). Neural network based multi-factor aware joint training for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*, 2231–2240.
- Ragni, A., & Gales, M. (2011). Derivative kernels for noise robust ASR. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on* (pp. 119–124). IEEE.

- Rath, S. P., Povey, D., Veselý, K., & Cernocký, J. (2013). Improved feature processing for deep neural networks. In *Proc. INTERSPEECH* (pp. 109–113).
- Rousseau, A., Deléglise, P., & Estève, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proc. LREC* (pp. 3935–3939).
- Samarakoon, L., & Sim, K. C. (2016a). Multi-attribute factorized hidden layer adaptation for DNN acoustic models. In *Interspeech 2016* (pp. 3484–3488).
- Samarakoon, L., & Sim, K. C. (2016b). Subspace LHUC for fast adaptation of deep neural network acoustic models. In *Interspeech 2016* (pp. 1593–1597).
- Saon, G., Soltau, H., Nahamoo, D., & Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *Proc. ASRU* (pp. 55–59). IEEE.
- Seide, F., Li, G., Chen, X., & Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Proc. ASRU* (pp. 24–29). IEEE.
- Senior, A., & Lopez-Moreno, I. (2014). Improving DNN speaker independence with i-vector inputs. In *Proc. ICASSP* (pp. 225–229).
- Smith, N., & Gales, M. (2002). Speech recognition using svms. In *Advances in neural information processing systems* (pp. 1197–1204).
- Swietojanski, P., Bell, P., & Renals, S. (2015). Structured output layer with auxiliary targets for context-dependent acoustic modelling. In *Proc. INTERSPEECH* (pp. 3605–3609).
- Swietojanski, P., Ghoshal, A., & Renals, S. (2013). Revisiting hybrid and GMM-HMM system combination techniques. In *Proc. ICASSP* (pp. 6744–6748). IEEE.
- Swietojanski, P., Li, J., & Renals, S. (2016). Learning hidden unit contributions for unsupervised acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24, 1450–1463.
- Swietojanski, P., & Renals, S. (2014). Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *Proc. SLT* (pp. 171–176). IEEE.
- Tan, T., Qian, Y., Yin, M., Zhuang, Y., & Yu, K. (2015). Cluster adaptive training for deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4325–4329). IEEE.
- Tomashenko, N., & Khokhlov, Y. (2014). Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. In *Proc. INTERSPEECH* (pp. 2997–3001).

- Tomashenko, N., & Khokhlov, Y. (2015). GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. In *Proc. INTERSPEECH* (pp. 2882–2886).
- Tomashenko, N., Khokhlov, Y., & Estève, Y. (2016a). A new perspective on combining GMM and DNN frameworks for speaker adaptation. In P. Král, & C. Martín-Vide (Eds.), *Statistical Language and Speech Processing: 4th International Conference, SLSP 2016, Pilsen, Czech Republic, October 11-12, 2016, Proceedings* (pp. 120–132). Cham: Springer International Publishing.
- Tomashenko, N., Khokhlov, Y., & Esteve, Y. (2016b). On the use of Gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models. In *Proc. INTERSPEECH* (pp. 3788–3792).
- Tomashenko, N., Khokhlov, Y., Larcher, A., & Estève, Y. (2016c). Exploration de paramètres acoustiques dérivés de GMM pour l’adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds. In *Proc. 31ème Journées d’Études sur la Parole (JEP)* (pp. 337–345).
- Tomashenko, N., Khokhlov, Y., Larcher, A., & Esteve, Y. (2016d). Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models. In *Proc. International Conference on Speech and Computer* (pp. 304–311). Springer.
- Tóth, L., & Gosztolya, G. (2016). Adaptation of DNN acoustic models using KL-divergence regularization and multi-task training. In A. Ronzhin, R. Potapova, & G. Németh (Eds.), *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings* (pp. 108–115). Springer International Publishing.
- Tran, D. T., Delroix, M., Ogawa, A., & Nakatani, T. (2016). Factorized linear input network for acoustic model adaptation in noisy conditions. *Interspeech 2016*, (pp. 3813–3817).
- Trentin, E. (2001). Networks with trainable amplitude of activation functions. *Neural Networks*, 14, 471–493.
- Tüske, Z., Golik, P., Schlüter, R., & Ney, H. (2015a). Speaker adaptive joint training of gaussian mixture models and bottleneck features. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on* (pp. 596–603). IEEE.
- Tüske, Z., Tahir, M. A., Schlüter, R., & Ney, H. (2015b). Integrating gaussian mixtures into deep neural networks: Softmax layer with hidden variables. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4285–4289). IEEE.
- Uebel, L. F., & Woodland, P. C. (2001). Improvements in linear transform based speaker adaptation. In *2001 IEEE International Conference on Acoustics,*

- Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)* (pp. 49–52 vol.1). volume 1.
- Variani, E., McDermott, E., & Heigold, G. (2015). A gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4270–4274). IEEE.
- Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*, 328–339.
- Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs: A review. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*.
- Woodland, P. C., Pye, D., & Gales, M. J. (1996). Iterative unsupervised adaptation using maximum likelihood linear regression. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (pp. 1133–1136). IEEE volume 2.
- Wu, C., & Gales, M. J. (2015). Multi-basis adaptive neural network for rapid adaptation in speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 4315–4319). IEEE.
- Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L., & Liu, Q. (2014). Fast adaptation of deep neural network based on discriminant codes for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Trans. on*, *22*, 1713–1725.
- Yao, K., Yu, D., Seide, F., Su, H., Deng, L., & Gong, Y. (2012). Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Proc. SLT* (pp. 366–369). IEEE.
- Yu, D., Chen, X., & Deng, L. (2012). Factorized deep neural networks for adaptive speech recognition. In *in Proc. Int. Workshop Statist. Mach. Learn. Speech Process.* Citeseer.
- Yu, D., & Deng, L. (2014). *Automatic speech recognition: A deep learning approach*. Springer.
- Yu, D., & Seltzer, M. L. (2011). Improved bottleneck features using pretrained deep neural networks. In *Interspeech* (p. 240). volume 237.
- Yu, D., Yao, K., Su, H., Li, G., & Seide, F. (2013). KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *Proc. ICASSP* (pp. 7893–7897).

Appendix A. Implementation details for baseline DNN models

This appendix illustrates the details of our experiments with DNN models. We trained four baseline DNN acoustic models:

- **DNN_{BN}-CE** was trained on BN features with CE criterion.
- **DNN_{BN}-sMBR** was obtained from the previous one by performing four epochs of sequence-discriminative training with Minimum Bayes Risk (sMBR) criterion.
- **DNN_{BN-fMLLR}-CE** was trained on fMLLR-adapted BN features.
- **DNN_{BN-fMLLR}-sMBR** was obtained from the previous one by four epochs of sMBR sequence-discriminative training.

For training DNN models, the initial GMM model was trained using 39-dimensional MFCC features including delta and acceleration coefficients. Linear discriminant analysis (LDA), followed by maximum likelihood linear transform (MLLT) and then fMLLR transformation, was then applied over these MFCC features to build a GMM-HMM system. Discriminative training with the boosted maximum mutual information (BMMI) objective was finally performed on top of this model.

Then a DNN was trained for BN feature extraction. The DNN system was trained using the frame-level cross entropy criterion and the senone (tied-state) alignment generated by the GMM-HMM system. To train this DNN, 40-dimensional log-scale filterbank features concatenated with 3-dimensional pitch-features were spliced across 11 neighboring frames, resulting in 473-dimensional (43×11) feature vectors. After that a DCT transform was applied and the dimension was reduced to 258. A DNN model for extraction 40-dimensional BN features was trained with the following topology: one 258-dimensional input layer; four hidden layers (HL), where the third HL was a BN layer with 40 neurons and other three HLs were 1500-dimensional; the output layer was 2390-dimensional. Based on the obtained BN features we trained the GMM model, which was used to produce the forced alignment, and then SAT-GMM model was trained on fMLLR-adapted BN features. Then fMLLR-adapted BN features were spliced in time with the context of 13 frames: [-10,-5..5,10] to train the final DNN model. The final DNN had a 520-dimensional input layer; six 2048-dimensional HLs with logistic sigmoid activation function, and a 4184-dimensional softmax output layer, with units corresponding to the context-dependent states.

The DNN parameters were initialized with stacked restricted Boltzmann machines (RBMs) by using layer by layer generative pre-training. It was trained with an initial learning rate of 0.008 using the cross-entropy objective function to obtain the SAT-DNN-CE model **DNN_{BN-fMLLR}-CE**.

After that, four epochs of sequence-discriminative training with per-utterance updates, optimizing state sMBR criteria, were performed to obtain the SAT-DNN-sMBR model **DNN_{BN-fMLLR}-sMBR**.

Baseline SI DNN models ($\mathbf{DNN}_{\text{BN-CE}}$ and $\mathbf{DNN}_{\text{BN-sMBR}}$) were trained in a similar way as the SAT DNNs described above, but without fMLLR adaptation.

Appendix B. Implementation details for baseline TDNN models

This appendix illustrates the details of our experiments with TDNN models. We trained four baseline TDNN acoustic models, which differ only in the type of the input features:

- $\mathbf{TDNN}_{\text{MFCC}}$ was trained on high-resolution MFCC features.
- $\mathbf{TDNN}_{\text{MFCC} \oplus \text{i-vectors}}$ was trained on high-resolution MFCC features, appended with 100-dimensional i-vectors.
- $\mathbf{TDNN}_{\text{BN}}$ was trained on BN features.
- $\mathbf{TDNN}_{\text{BN} \oplus \text{i-vectors}}$ was trained on BN features, appended with 100-dimensional i-vectors.

The baseline SAT-TDNN model $\mathbf{TDNN}_{\text{MFCC} \oplus \text{i-vectors}}$ is similar to those described in [Peddinti et al., 2015], except for the number of hidden layers and slightly different subsequences of splicing and sub-sampling indexes. The two types of data augmentation strategies were applied for the speech training data: speed perturbation (with factors 0.9, 1.0, 1.1) and volume perturbation. The SAT-TDNN model was trained on high-resolution MFCC features (without dimensionality reduction, keeping all 40 cepstra) concatenated with 100-dimensional i-vectors. The temporal context was $[t - 16, t + 12]$ and the splicing indexes used here were $[-2, 2]$, $\{-1, 2\}$, $\{-3, 3\}$, $\{-7, 2\}$, $\{0\}$, $\{0\}$. This model has 850-dimensional hidden layers with rectified linear units (ReLU) [Dahl et al., 2013] activation functions, a 4052-dimensional output layer and approximately 10.9 million parameters.

The baseline SI-TDNN model $\mathbf{TDNN}_{\text{MFCC}}$ was trained in a similar way as the SAT-TDNN described above, but without using i-vectors.

In addition to these baseline models, for comparison purpose, we trained two other baseline TDNNs (with and without i-vectors) using BN features instead of high-resolution MFCC features. The same BN features we used later for training an auxiliary monophone GMM model for GMMD feature extraction. These BN features were extracted using a DNN trained in a similar way as described in Section 4.4.1 for DNN AM, but on the high-resolution MFCC features (instead of "filter bank \oplus pith" features) and on the augmented (by means of speed and volume perturbation) data base. As in [Peddinti et al., 2015], the i-vectors during the training were calculated every two utterances.

Appendix C. Implementation details for DNNs trained on GMMD features

This appendix presents the details of training DNN models on the proposed GMMD features. In this set of experiments we trained four DNNs, using the approach proposed in Section 3.1:

- **DNN_{GMMD \oplus BN-CE}** is a DNN without performing speaker adaptive training, which in our case means that the auxiliary GMM monophone model was not adapted. This DNN model was trained using the CE criterion.
- **DNN_{GMMD \oplus BN-sMBR}** was obtained from the previous one by performing four epochs of sequence-discriminative training with per-utterance updates, optimizing the sMBR criterion.
- **DNN_{GMMD-MAP \oplus BN-CE}** was a proposed SAT DNN model trained on speaker adapted GMMD-MAP features, with the CE criterion.
- **DNN_{GMMD-MAP \oplus BN-sMBR}** was obtained from the previous one by performing four epochs of sMBR sequence training.

Models **DNN_{GMMD-MAP \oplus BN-CE}** and **DNN_{GMMD-MAP \oplus BN-sMBR}** were trained as described in Section 3.1. The GMMD features were extracted using a monophone auxiliary GMM model, trained on BN features. This GMM model was adapted for each speaker by MAP adaptation algorithm (Section 3.3).

We took the state tying from the baseline SAT-DNN to train all other models. The purpose of using the same state tying is to allow posterior level fusion for these models. We also took an alignment obtained by the SAT-DNN model, because in [Tomashenko et al., 2016a] it was shown to slightly improve the result.

Then we concatenated two types of BN features and spliced them for training the final DNN model. Both DNN models were trained on the proposed features in the same manner and had the same topology except for the input features, as the final baseline SAT DNN model trained on BN features (Section 4.4.1). The other two SI models (**DNN_{GMMD \oplus BN-CE}** and **DNN_{GMMD \oplus BN-sMBR}**) were trained in the same manner but without speaker adaptation.