Joint Adaptation and Adaptive Training of TVWR for Robust Automatic Speech Recognition

Shilin LIU, Khe Chai SIM

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

Abstract

Context-dependent Deep Neural Network has obtained consistent and significant improvements over the Gaussian Mixture Model (GMM) based systems for various speech recognition tasks. However, since DNN is discriminatively trained, it is more sensitive to label errors and is not reliable for unsupervised adaptation. Moreover, DNN parameters do not have a clear and meaningful interpretation, therefore, it has been difficult to develop effective adaptation techniques for the DNNs. Nevertheless, unadapted multi-style trained DNNs have already shown superior performance to the GMM system with joint noise/speaker adaptation and adaptive training. Recently, Temporally Varying Weight Regression (TVWR) has been successfully applied to combine DNN and GMM for robust unsupervised speaker adaptation. In this paper, joint speaker/noise adaptation and adaptive training of TVWR using DNN posteriors are investigated for robust speech recognition. Experimental results on the Aurora 4 corpus showed that after joint adaptation and adaptive training, TVWR achieved 21.3% and 11.6% relative improvements over the DNN baseline system and the best system in currently reported literatures, respectively.

Index Terms: Adaptation, Adaptive Training, Temporally Varying Weight Regression, Deep Neural Network

10.21437/Interspeech.2014-152

1. Introduction

In practice, many speech variabilities exist due to the acoustic environment, reverberation, background noise, communication channel and speaker difference. Training an acoustic model to be robust against these variabilities has always been important and challenging. When the Gaussian Mixture Model (GMM) was a dominant observation distribution representation of the Hidden Markov Model (HMM) state, many techniques were developed to make it robust to various acoustic conditions. Since the GMM is a generative model whose parameters represent the statistics of the data, it is easier to manipulate the parameters to adapt to a specific condition. For example, Maximum a Posterior (MAP) [1] and Maximum Likelihood Linear Regression (MLLR) [2] are commonly used for speaker adaptation. On the other hand, Vector Taylor Series (VTS) is widely used to compensate acoustic features or models for different noise conditions. All these methods modify the GMM parameters to better represent the data statistics in the respective acoustic conditions.

The DNN parameters do not represent the statistics of the speech data, but are obtained using the error back-propagation algorithm [3] to minimize the cross-entropy objective function. Hence, existing model-based techniques, such as MAP, MLLR and VTS, cannot be applied to the DNN models. In order to handle speaker variability, discriminative learning of speaker code has been proposed to adapt a DNN-based system [4]. In this work, the speaker code is learned as the additional parameters in the first two layers of the DNN using the error back-propagation algorithm. Another simple but effective method to "adapt" the DNN is to borrow the fMLLR feature transform from a GMM-based system [5]. However, both methods require adaptive training to achieve consistency between training and testing. Moreover, speaker code based method may be not reliable for unsupervised speaker adaptation.

Currently, techniques that improve the noise robustness of the DNN-based acoustic models can be divided into two categories: multi-style training [6] and feature engineering [5, 7, 8, 9, 10]. The former simply lets the model learn the appropriate parameters to handle the variabilities in the multi-style data. Although multi-style training has been found to greatly improve the average performance of a DNN-based system, it sometimes degrades the performance on the clean environment. Feature engineering aims to either extract more robust acoustic features against variabilities or remove variabilities to get normalized acoustic features. If a DNN-based system is trained using the clean data, feature enhancement or denoising techniques, such as feature-based VTS [11, 7, 8], can be performed to cope with the noisy environment. If the stereo training data (a parallel clean and noisy speech) is available, SPLICE [12], MMSE [10] or DRADE (Deep Recurrent Denoising Autoencoder) [9] can be applied to estimate the "clean" speech. In practice, it is much more expensive to prepare the stereo training data than the multi-style training data.

Recent work [6] has shown that the standard DNN system using filter-bank features and multi-style training achieved comparable performance to a GMM-based system with joint speaker and noise adaptive training [13]. Therefore, it is interesting to investigate whether combining these two systems will further improve the performance. In our recent work [14], Temporally Varying Weight Regression (TVWR) has been presented as an approach to combine GMM and DNN for speaker adaptation. VTS has also been applied to compensate TVWR using the GMM-based posterior features [15]. In this paper, joint speaker/noise adaptation and adaptive training will be studied for robust TVWR using the DNN posteriors.

The rest of this paper is organized as follows. Section 2 introduce the sparse regression for incorporating TVWR with high-dimensional DNN posteriors. Section 3 introduces the proposed joint speaker/noise adaptation and adaptive training for robust TVWR. Last, experimental results on Aurora 4 corpus are discussed in Section 4.

2. Sparse Regression of DNN Posteriors

As previously mentioned, the DNNs are able to predict high quality discriminative CD posteriors while the GMMs can be reliably adapted in an unsupervised manner. Therefore, to get

the best of both worlds, combining the GMMs and DNNs using the TVWR framework has been proposed for robust speaker adaptation [14]. In the TVWR framework [16, 17], the output probability of state j is given as:

$$p(\boldsymbol{\xi}_t|j) \propto \sum_{m=1}^{M}\sum_{i=1}^{N}\frac{p(i|\boldsymbol{\tau}_t)}{P(i)}P(i|j,m)P(m|j)p(\mathbf{o}_t|j,m)$$
$$= \sum_{m=1}^{M} c_{jm}\mathcal{R}(\boldsymbol{\tau}_t;\mathbf{w}_{jm})\mathcal{N}(\mathbf{o}_t;\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm}) \quad (1)$$

where the linear regression function is given as

$$\mathcal{R}(\boldsymbol{\tau}_t;\mathbf{w}_{jm}) = \sum_{i=1}^{N}\tilde{h}_{it}w_{jmi} \quad (2)$$

$\boldsymbol{\xi}_t = \{\mathbf{o}_t, \boldsymbol{\tau}_t\}$ is the long span acoustic feature, $\boldsymbol{\tau}_t = \{\mathbf{o}_{t-\delta}\dots\mathbf{o}_{t-1}, \mathbf{o}_{t+1}\dots\mathbf{o}_{t+\delta}\}$ denotes the contexts of the current observation $\mathbf{o}_t$, $p(i|\boldsymbol{\tau}_t)$ is the posterior probability of a latent variabe, i, given the acoustic contexts, $c_{jm} = P(m|j)$ is the prior of the Gaussian component m and $w_{jmi} = P(i|j,m)$ is the regression parameter, which describes the prior probability of i being associated with the component m in state j. M and N correspond to the number of Gaussian components and the number of latent variables, respectively. To incorporate the DNN into TVWR, the latent variables are associated with the high dimensional normalized posteriors of the context-dependent phone states, such that: $\tilde{h}_{it} = p(i|\boldsymbol{\tau}_t)/P(i)$. In order to address the issue of incorporating the high-dimensional CD-DNN posteriors without dramatically increasing the model complexity, sparse regression [14] is applied in this paper. The idea of sparse regression is to select only a relevant subset of latent variables for each state. Hence, the linear regression can be efficiently computed by:

$$\tilde{\mathcal{R}}(\boldsymbol{\tau}_t;\mathbf{w}_{jm}) = \sum_{i\in I_j}\tilde{h}_{it}w_{jmi} \quad (3)$$

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

$$I_j = \{i : P(i|j) > \nu\},\ P(i|j) = \frac{1}{|T_j|}\sum_{t\in T_j}p(i|\boldsymbol{\tau}_t) \quad (4)$$

where ν is a threshold that controls the number of active parameters, such as $\nu = \frac{1}{N}$, T_j is the set of frames which is associated to state j and $|.|$ denotes the cardinality, the regression parameters are initialized as $P(i|j,m) = P(i|j)$. Hence, the auxiliary function to be optimized for the maximum likelihood estimation [17] is given as:

$$\mathcal{Q}(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Lambda}}) = \sum_{t,j,m}\sum_{i\in I_j}\gamma_{jmi}(t)\log w_{jmi} \quad (5)$$

where $\boldsymbol{\Lambda}$ are the model parameters to be optimized, and $\hat{\boldsymbol{\Lambda}}$ are the current model parameters. The occupancies are given as:

$$\gamma_{jmi}(t) = \gamma_{jm}(t)\frac{\hat{w}_{jmi}\tilde{h}_{it}}{\tilde{\mathcal{R}}(\boldsymbol{\tau}_t;\hat{\mathbf{w}}_{jm})}$$

$$\gamma_{jm}(t) = \gamma_j(t)\frac{\hat{c}_{jm}\tilde{\mathcal{R}}(\boldsymbol{\tau}_t;\hat{\mathbf{w}}_{jm})\mathcal{N}(\mathbf{o}_t;\hat{\boldsymbol{\mu}}_{jm},\hat{\boldsymbol{\Sigma}}_{jm})}{\sum_m \hat{c}_{jm}\tilde{\mathcal{R}}(\boldsymbol{\tau}_t;\hat{\mathbf{w}}_{jm})\mathcal{N}(\mathbf{o}_t;\hat{\boldsymbol{\mu}}_{jm},\hat{\boldsymbol{\Sigma}}_{jm})}$$

$\gamma_j(t)$ is the state occupancy given the current model, $\hat{\boldsymbol{\Lambda}}$ and transcription. The regression parameter update formulae will be given as:

$$w_{jmi} = \begin{cases} \frac{\sum_t \gamma_{jmi}(t)}{\sum_t\sum_{i\in I_j}\gamma_{jmi}(t)} & \text{if } i \in I_j \\ 0 & \text{otherwise} \end{cases}$$

In this paper, the posterior feature is assumed to be context dependent and predicted by a DNN. In order to make the regression robust against noise conditions, DNN is trained using multi-style training data. The next section will present the adaptation and adaptive training of TVWR.

3. Joint Adaptation and Adaptive Training

Different speaker has different speaking style such as female vs. male, native vs. non-native, young vs. elder, etc. Environmental variabilities includes background noises, different microphones and transmission channels. Many adaptation techniques have been developed for each of them. Speaker and noise factorization have also been proposed to consider both of them in the GMM-based system [13]. In this section, the "VTS-MLLR" scheme, which has been found to perform best [13], will be used to exploit the joint adaptation and adaptive training in the TVWR framework. In other words, noise compensation will be first performed using VTS and followed by speaker adaptation using MLLR mean transform for the joint adaptation. It is also assumed that one speaker can speak many utterances but one utterance can only have one speaker. Furthermore, one utterance has one acoustic condition and its own distortion parameters, and one condition can only contain one speaker in the following formulation. Define the clean, noisy speech, additive noise and channel distortion variables as $\mathbf{x}$, $\mathbf{y}$, $\mathbf{n}$, $\mathbf{h}$, respectively. The objective of joint adaptive training is to maximize the likelihood of the multi-style training data, and the auxiliary function can be expressed as in Eq-6 (due to the similar formulation for delta parameters, only the first order delta parameters are discussed here). The forward-backward component occupancy is now given as:

$$\gamma_{jm}^{s,k}(t) = \gamma_j^{s,k}(t)\frac{\hat{c}_{jm}\tilde{\mathcal{R}}(\boldsymbol{\chi}_t^{s,k};\hat{\mathbf{w}}_{jm})\mathcal{N}(\mathbf{y}_t^{s,k};\hat{\boldsymbol{\mu}}_{jm}^{s,k},\hat{\boldsymbol{\Sigma}}_{jm}^{k})}{\sum_m \hat{c}_{jm}\tilde{\mathcal{R}}(\boldsymbol{\chi}_t^{s,k};\hat{\mathbf{w}}_{jm})\mathcal{N}(\mathbf{y}_t^{s,k};\hat{\boldsymbol{\mu}}_{jm}^{s,k},\hat{\boldsymbol{\Sigma}}_{jm}^{k})}$$

where $\boldsymbol{\chi}_t = \{\mathbf{y}_{t-\delta}\dots\mathbf{y}_{t-1}, \mathbf{y}_{t+1}\dots\mathbf{y}_{t+\delta}\}$ is the noisy context, s represents a speaker, k is the utterance (or condition) index. Assuming that each speaker uses a global block-diagonal MLLR mean transform [1], the joint speaker and noise adaption for the Gaussian means can be written as:

$$\boldsymbol{\mu}_{\odot jm}^{s,k} \approx \mathbf{A}_{\odot}^{s}\boldsymbol{\mu}_{\odot jm}^{k} + \mathbf{b}_{\odot}^{s} \quad (7)$$

$$\boldsymbol{\mu}_{\Delta jm}^{s,k} \approx \mathbf{A}_{\Delta}^{s}\boldsymbol{\mu}_{\Delta jm}^{k} + \mathbf{b}_{\Delta}^{s} \quad (8)$$

where $\odot$ and Δ indicates the static and delta parameters, respectively, $\mathbf{A}^s$, $\mathbf{b}^s$ are the speaker transforms, VTS [18, 19]

[1] This assumption is made for easier discussion, and it is possible to use full transform and multiple regression classes.

$$\mathcal{Q}(\mathbf{\Lambda},\hat{\mathbf{\Lambda}}) = \sum_{s,k,t,j,m} \gamma_{jm}^{s,k}(t)\left(-\frac{1}{2}\log|\mathbf{\Sigma}_{\odot jm}^{k}| - \frac{1}{2}(\mathbf{y}_{\odot t}^{s,k}-\boldsymbol{\mu}_{\odot jm}^{s,k})^{T}\mathbf{\Sigma}_{\odot jm}^{-1,k}(\mathbf{y}_{\odot t}^{s,k}-\boldsymbol{\mu}_{\odot jm}^{s,k})\right) + \sum_{s,k,t,j,m} \gamma_{jm}^{s,k}(t)\left(-\frac{1}{2}\log|\mathbf{\Sigma}_{\Delta jm}^{k}| - \frac{1}{2}(\mathbf{y}_{\Delta t}^{s,k}-\boldsymbol{\mu}_{\Delta jm}^{s,k})^{T}\mathbf{\Sigma}_{\Delta jm}^{-1,k}(\mathbf{y}_{\Delta t}^{s,k}-\boldsymbol{\mu}_{\Delta jm}^{s,k})\right) \quad (6)$$

expansion functions for noise compensation are given as:

$$\boldsymbol{\mu}_{\odot jm}^{k} \approx \mathbf{f}(\hat{\boldsymbol{\mu}}_{\odot jm},\hat{\boldsymbol{\mu}}_{\odot h}^{k},\hat{\boldsymbol{\mu}}_{\odot n}^{k}) + \mathbf{G}_{jm}^{k}(\boldsymbol{\mu}_{\odot jm}-\hat{\boldsymbol{\mu}}_{\odot jm}) + \mathbf{G}_{jm}^{k}(\boldsymbol{\mu}_{\odot h}^{k}-\hat{\boldsymbol{\mu}}_{\odot h}^{k}) + \mathbf{F}_{jm}^{k}(\boldsymbol{\mu}_{\odot n}^{k}-\hat{\boldsymbol{\mu}}_{\odot n}^{k}) \quad (9)$$

$$\boldsymbol{\mu}_{\Delta jm}^{k} \approx \mathbf{G}_{jm}^{k}\hat{\boldsymbol{\mu}}_{\Delta jm} + \mathbf{G}_{jm}^{k}(\boldsymbol{\mu}_{\Delta jm}-\hat{\boldsymbol{\mu}}_{\Delta jm}) \quad (10)$$

$$\mathbf{\Sigma}_{\odot jm}^{k} \approx \mathrm{diag}\left\{\mathbf{G}_{jm}^{k}\mathbf{\Sigma}_{\odot jm}\mathbf{G}_{jm}^{T,k} + \mathbf{F}_{jm}^{k}\mathbf{\Sigma}_{\odot n}^{k}\mathbf{F}_{jm}^{T,k}\right\} \quad (11)$$

$$\mathbf{\Sigma}_{\Delta jm}^{k} \approx \mathrm{diag}\left\{\mathbf{G}_{jm}^{k}\mathbf{\Sigma}_{\Delta jm}\mathbf{G}_{jm}^{T,k} + \mathbf{F}_{jm}^{k}\mathbf{\Sigma}_{\Delta n}^{k}\mathbf{F}_{jm}^{T,k}\right\} \quad (12)$$

where $\mathrm{diag}\{.\}$ is an operator to extract the diagonal elements, $\boldsymbol{\mu}_{\Delta n}=\mathbf{0}$, $\boldsymbol{\mu}_{\Delta h}=\mathbf{0}$ and $\mathbf{\Sigma}_{\odot h}=\mathbf{\Sigma}_{\Delta h}=\mathbf{0}$ are assumed, the distortion function of the static parameters is expressed as:

$$\mathbf{f}(\mathbf{x},\mathbf{h},\mathbf{n}) = \mathbf{x}+\mathbf{h}+\mathbf{C}\log(1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))) \quad (13)$$

$\mathbf{C}$ is the discrete cosine transform, and the Jacobian terms are given as:

$$\mathbf{G}_{jm}^{k} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\Big|_{\hat{\boldsymbol{\mu}}_{\odot jm},\hat{\boldsymbol{\mu}}_{\odot h}^{k},\hat{\boldsymbol{\mu}}_{\odot n}^{k}}, \quad \mathbf{F}_{jm}^{k} = \mathbf{I}-\mathbf{G}_{jm}^{k} \quad (14)$$

Let $\mathbf{\Phi} = \{\boldsymbol{\mu}_{jm},\mathbf{\Sigma}_{jm}\}$ be the canonical model, $\mathbf{\Psi}^{k} = \{\boldsymbol{\mu}_{n}^{k},\boldsymbol{\mu}_{h}^{k},\mathbf{\Sigma}_{n}^{k}\}$ be the noise model, and $\mathbf{K}^{s}=\{\mathbf{A}^{s},\mathbf{b}^{s}\}$ be the speaker transform. By setting $\mathbf{\Phi}=\hat{\mathbf{\Phi}}$ and $\mathbf{\Psi}=\hat{\mathbf{\Psi}}$, the modified MLLR mean statistics [13, 2] after the noise compensation can be accumulated and used to estimate speaker transform, $\mathbf{K}$. Given the current speaker transform $\hat{\mathbf{K}}$, noise and canonical means, $\boldsymbol{\mu}_{n}^{k},\boldsymbol{\mu}_{h}^{k},\boldsymbol{\mu}_{jm}$ can be estimated using the modified noise adaptive training (NAT) formulae [13, 19]. This modification includes the revised Jacobians $\hat{\mathbf{A}}_{\odot}^{s}\mathbf{G}_{jm}^{k}$, $\hat{\mathbf{A}}_{\odot}^{s}\mathbf{F}_{jm}^{k}$, $\hat{\mathbf{A}}_{\Delta}^{s}\mathbf{G}_{jm}^{k}$, $\hat{\mathbf{A}}_{\Delta}^{s}\mathbf{F}_{jm}^{k}$, and jointly adapted Gaussian means, $\hat{\boldsymbol{\mu}}_{\odot jm}^{s,k},\hat{\boldsymbol{\mu}}_{\Delta jm}^{s,k}$. The noise and canonical variance, $\mathbf{\Sigma}_{n}^{k},\mathbf{\Sigma}_{jm}$ estimates are very similar to the conventional NAT formulae [19], except that the compensated means are also adapted to speaker. These three groups of parameters can be estimated in an interleaved manner for joint adaptive training [13], while regression parameters, static weights and transition probabilities can be just estimated using the standard procedures but with the occupancies based on the jointly adapted system.

4. Experimental Results

In this section, experiments were conducted on the Aurora4 corpus for evaluation of robust TVWR using DNN based posteriors. Multi-style training set with 16kHz sampling rate were used in this experiment. The training dataset comprises 7138 utterances from 84 speakers, or about 12 hours of speech. Half of multi-style training data were recorded by a close-talking microphone, but the rest were recorded by a variety of different microphones. The multi-style training data contain 6 different types of noise. The noise was added at a randomly chosen SNR between 10 and 20 dB, averaged 15 dB. For evaluation, 14 test sets (each with 330 utterances and 8 speakers) are available for covering 14 test conditions, which can be further grouped into 4 sets, named as A, B, C, D. A represents the clean test set; B contains 6 test sets and 6 types of noise, whose SNR is randomly chosen between 5 and 15 dB, averaged 10 dB; C is also the clean test set but recorded with a different microphone; D is similar to B, except that the recording devices are different.

The baseline GMM systems consisted of context-dependent HMMs with 3187 senones (or tied states) and 16 mixtures per state. Multi-style training (MST) was performed using maximum likelihood criterion. The input features for noise adaptation and adaptive training are Mel Frequency Cepstral Coefficients (MFCC), including 12 static parameters, zero-th coefficient, and first two orders of dynamic parameters. Cepstral mean&variance normalization (CMVN) was also used to perform two-model re-estimation for another baseline. The GMM baseline using CMVN features was used to align the training data as the initial senone labels for DNN training. The testing results were obtained by a bigram full decoding using HTK.

The input features for DNN training are 24-dimensional log Mel filter-bank (FBANK) features with two order dynamic parameters and context expansion with window size 11, i.e. 792 visible units. There were 6 hidden layers with sigmoid activation function and 2048 hidden units for each layer. Output layer had 3187 units corresponding to the context-dependent senones and used soft-max as activation function. Layer-by-layer generative pre-training was performed to obtain an initialization of DNN. Cross-entropy criterion was used for back propagation fine tuning. Back propagation was done using stochastic gradient descent with 128 examples per mini-batch. The initial learning rate was 0.015, which was reduced by a factor of 2 if cross validation performance was dropped. Momentum was 0 and maximum training iteration was 25. 1206 multi-style development utterances were extracted from the corpus. Both pre-training and fine-tuning were done by Kaldi toolkit [20]. The trained DNN was then used to re-align the training data, and the second DNN was trained with updated labels, which was the final DNN for the following experiments. The DNN output log posterior features were analyzed by Principle Component Analysis (PCA), and 13 dimensional features were projected and appended to the original CMVN features. The concatenated features were used to train a tandem system.

For the best performance, constrained MLLR (CMLLR) [21] was used for the adaptive training of the GMM-based systems. In particular, one global transform was used per speaker, which was also used for speaker adaptive training of another DNN system. Due to the limitation of current formulation and implementation, only MLLR mean adaptation was used for speaker adaptation and adaptive training in the TVWR framework. To perform unsupervised speaker/noise adaptation, one full bigram decoding was performed first and followed by an EM estimation of noise model/speaker transform. $\nu = 0.0003$ was used as the floor of building the confusion table, and 130 active latent variables in average were obtained for sparse regression.

First, the performance of various baseline systems trained using multi-style data are reported in Table.1. The GMM baseline system using MFCC performed worst among all the systems. Cepstral mean&variance normalization (CMVN) signifi-

System		A	B	C	D	Avg
Model	Feature					
GMM	MFCC	10.5	23.6	29.4	38.1	27.7
GMM	CMVN	8.0	16.2	18.1	32.2	22.6
GMM [13]	AFE	8.8	16.7	19.1	28.6	21.4
DNN	FBANK	**4.8**	**8.8**	**8.8**	**20.5**	**13.6**
GMM	Tandem	6.4	12.4	13.0	28.0	18.7
TVWR	MFCC	5.6	9.1	11.3	21.7	14.4

Table 1: *Recognition results (WER%) of various baseline systems without adaptation on Aurora4.*

cantly improved the recognition results on all test sets. It is also comparable to the Advanced feature enhancement (AFE) [13]. Hence, the initial DNN training label was obtained by forced-alignment using the GMM-CMVN system. The DNN using the filter-bank features achieved best performance. Its performance is also comparable to that reported in [6]. This DNN system was then used to generate the posterior features to train all the subsequent TVWR and tandem systems. The Tandem system significantly outperformed other GMM-based systems due to the additional information from DNN. As dimension reduction was performed to build the tandem features, information loss makes it perform worse than the DNN baseline. On the other hand, TVWR performs sparse regression on the raw posterior features and achieved better performance than other GMM and Tandem systems. However, TVWR still performed worse than the DNN baseline system.

Without speaker/noise adaptation, the GMM performs poorly in noisy data, such that TVWR does not yield better results compared to the DNN baseline system. By performing adaptive training to the GMM components of TVWR, the performance gap between the GMM and DNN can be reduced so that the combination via TVWR may become more effective. Hence, in order to improve the robustness of TVWR, speaker/noise adaptation and adaptive training were performed next. Various system performances based on adaptation and adaptive training are shown in Table.2. In order to apply VTS noise compensation, acoustic features used in this experiment were MFCC. The multi-style trained GMM can benefit from the adaptation techniques and achieved significant improvements over the multi-style GMM baseline. In addition, adaptive training also helped to improve the system performance. With both speaker (MLLR mean) and noise adaptations, GMM [13] obtained comparable performance to DNN using the FBANK features. Adaptation and adaptive training also helped to improve the Tandem system performance, however, it is still worse than the DNN baseline system. After applying either the VTS or MLLR mean adaptation, TVWR obtained 1.8% absolute improvements over the DNN baseline system. If both adaptation were applied, another 0.6% absolute improvements were obtained. After speaker adaptive training (SAT), TVWR with speaker adaptation achieved 2.4% absolute gain over DNN baseline and 1.2% absolute gain over the speaker adaptively trained DNN. Joint adaptation further improved the system performance to 10.9% WER. However, NAT alone does not seem to help much compared to the multi-style training.

Finally, the current best system on Aurora 4 reported in the literature is the DNN-based system trained with the constrained Feature Discriminant Linear Regression (cFDLR) features. This system achieved an overall WER of 12.1%. With either NAT or SAT, TVWR has already outperformed this system

System		Adaptation		A	B	C	D	Avg
Model	Training	Noise	Speaker					
GMM	MST	VTS	-	9.7	14.9	14.0	22.4	17.7
		-	fMLLR	7.1	15.1	15.4	31.4	21.5
	SAT	-	fMLLR	7.0	14.1	12.5	27.6	19.3
	NAT	VTS	-	8.5	13.7	11.4	21.3	16.4
	Joint [13]	VTS	MLLR	5.6	11.0	8.8	17.8	13.4
Tandem	MST	-	fMLLR	5.9	11.5	10.9	26.3	17.4
	SAT	-	fMLLR	5.6	11.0	10.0	25.6	16.8
TVWR	MST	VTS	-	5.6	8.3	8.3	16.8	11.8
		-	MLLR	5.0	8.1	7.3	17.4	11.8
		VTS	MLLR	5.0	7.8	7.7	16.1	11.2
	SAT	-	MLLR	5.0	7.8	**6.7**	16.5	11.2
		VTS	MLLR	5.1	7.6	7.0	15.9	**10.9**
	NAT	VTS	-	4.9	8.5	7.7	17.1	11.9
		VTS	MLLR	4.7	8.0	7.3	16.1	11.2
	Joint	VTS	MLLR	4.4	**7.5**	7.1	**15.6**	**10.7**
DNN	NAT [6]	VTS	-	5.4	8.8	7.8	19.6	13.1
	Drop [6]	-	-	5.1	8.4	8.6	19.3	12.9
	N+D [6]	VTS	-	5.4	8.3	7.6	18.5	12.4
	MST [22]	cFDLR	-	5.1	8.5	8.4	17.6	12.1
	SAT	-	fMLLR	**4.3**	7.9	**6.7**	19.2	12.4

Table 2: *Recognition results (WER%) of various systems on Aurora4 based on adaptation and adaptive training.*

by 0.9% and 1.2% absolute, respectively. The best TVWR system , with joint adaptive training, achieved 10.7% WER. This system is 1.4% absolute better than the cFDLR DNN system and 2.9% absolute (or 20% relative) better than the baseline DNN system.

5. Conclusions

In this paper, joint speaker/noise adaptive training and adaptation has been proposed for TVWR, a framework that combines the discriminative power of the DNN and the adaptability of the GMM for robust automatic speech recognition. On top of the superior performance of the DNN-based system, we were able to obtain further improvements on the Aurora 4 test sets by applying either the speaker adaptive training or noise adaptive training to the GMM parameters of the TVWR systems. Finally, the best overall word error rate performance of 10.7% was obtained by applying joint adaptive training, which was 11.6% relatively better than the best DNN alone system reported in the literature on the same data set. This shows that TVWR can be an effective approach to combining the best of two worlds: GMM and DNN.

6. Acknowledgements

The authors would like to thank Dr. WANG Yongqiang (Eric) for his help in building the NAT baseline system.

7. References

[1] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

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

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning rep-*

resentations by back-propagating errors. MIT Press, Cambridge, MA, USA, 1988.

[4] O. A-H. and H. J., "Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition based on Discrimative Learning of Speaker Code," in *ICASSP*. IEEE, 2013, pp. 7942 – 7946.

[5] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[6] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013.

[7] B. Li and K. Sim, "Noise adaptive front-end normalization based on vector taylor series for deep neural networks in robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7408–7412.

[8] ——, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 279–284.

[9] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[10] X. Cui, M. Afify, and Y. Gao, "MMSE-based stereo feature stochastic mapping for noise robust speech recognition," in *Proc. ICASSP*. IEEE, 2008, pp. 4077–4080.

[11] P. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, Carnegie Mellon University, 1996.

[12] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the Aurora 2 and 3 tasks," in *Proc. ICSLP*, 2002, pp. 29–32.

[13] Y. Wang and M. Gales, "Speaker and noise factorization for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2149–2158, 2012.

[14] S. Liu and K. Sim, "On Combining DNN and GMM with Unsupervised Speaker Adaptation for Robust Automatic Speech Recognition," in *Proc. ICASSP*. IEEE, 2014, pp. –.

[15] ——, "An Investigation of Temporally Varying Weight Regression for Noise Robust Speech Recognition," in *Proc. Interspeech*. ISCA, 2013, pp. 4761–4764.

[16] ——, "Implicit Trajectory Modelling Using Temporally Varying Weight Regression for Automatic Speech Recognition," in *Proc. ICASSP*. IEEE, 2012, pp. 4761–4764.

[17] ——, "Temporally Varying Weight Regression: a Semi-parametric Trajectory Model for Automatic Speech Recognition," *Speech and Audio Processing, IEEE Transactions on*, pp. –, 2013.

[18] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM Adaptation Using Vector Taylor Series for Noisy Speech Recognition," in *Proc. ICSLP*, 2000, pp. 869–872.

[19] O. Kalinli, M. Seltzer, and A. Acero, "Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition," in *Proc. ICASSP*. IEEE, 2009, pp. 3825–3828.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[21] V. V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 357–366, 1995.

[22] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *OSU-CISRC-6/13-TR14*, 2013.