Factored Adaptation for Separable Compensation of Speaker and Environmental Variability

Michael L. Seltzer, Alex Acero

Microsoft Research
Redmond, WA 98052 USA
{mseltzer,alexac}@microsoft.com

***Abstract*—While many algorithms for speaker or environment adaptation have been proposed, far less attention has been paid to approaches which address both factors. We recently proposed a method called factored adaptation that can jointly compensate for speaker and environmental mismatch using a cascade of CMLLR transforms that separately compensate for the environment and speaker variability. Performing adaptation in this manner enables a speaker transform estimated in one environment to be be applied when the same user is in different environments. While this algorithm performed well, it relied on knowledge of the operating environment in both training and test. In this paper, we show how unsupervised environment clustering can be used to eliminate this requirement. The improved factored adaptation algorithm achieves relative improvements of 10-18% over conventional CMLLR when applying speaker transforms across environments without needing any additional a priori knowledge.**

I. Introduction

Because speech recognition systems are statistical pattern classifiers trained from data, any mismatch between the speech seen in deployment and that used to train the speech recognizer will cause a degradation in performance. Two of the biggest sources of acoustic mismatch are the environment and the speaker. In both of these cases, acoustic model adaptation has been proposed as way to reduce the mismatch and improve performance.

Historically, methods for environmental or speaker adaptation have been developed independently. Environmental adaptation is commonly performed using methods that utilize a parametric model that explains how clean speech is corrupted by additive and convolutional noise. These methods such as parallel model combination (PMC) [1] and vector Taylor series (VTS) adaptation [2] have the advantage that they can adapt all parameters of the recognizer based on a small observation of the noise. In order for such approaches to operate, the acoustic model must be trained from clean speech or using specialized noise-adaptive training techniques [3], [4].

In constrast, the most common methods for speaker adaptation are data-driven approaches in which the model parameters are transformed in a manner which maximized the likelihood of the adaptation data. For example, the various versions of MLLR or CMLLR use one or more affine transforms to modify the Gaussian parameters of the recognizer [5]. These methods are not particular to speaker adaptation and can be used to compensate for any mismatch, including environmental noise [6]. While this is a benefit, it can also be a drawback, as it is unknown exactly what mismatch the estimated transforms are compensating. This prevents these transforms (and the adaptation data in general) from being reused in situations where the same speaker may be in a different acoustic environment.

Because of this problem, there has been some interest in approaches in which the speaker and environmental variability can be compensated in a way that allows the sources of variability to be separated. A method of joint environment and speaker adaptation was proposed in which Jacobian adaptation for noise compensation was combined with MLLR for speaker adaptation [7]. This approach was recently improved by using Vector Taylor Series (VTS) adaptation to update both the means and variances of speaker-adapted models whose means were compensated using MLLR [8]. The VTS noise parameters and the MLLR transforms were jointly estimated using an iterative EM approach. Combining methods that use different adaptation strategies enables straightforward separation of the speaker transform parameters and the environmental transform parameters. This desirable separability was called acoustic factorization in [9]. In that work, MLLR adaptation was combined with cluster adaptive training such that the speaker variability was captured by the MLLR transforms while the environmental variablity was captured by the cluster weights in the acoustic model.

Most recently, we proposed a technique called factored adaptation [10]. In this approach, a cascade of linear transforms was used to jointly compensate for the speaker and the environment. Because the cascade of transforms is itself a linear transform, it is computationally equivalent to any transform-based adaptation method at runtime. In this work, adaptation was performed using a cascade of CMLLR transforms which is appealing because a CMLLR transform can compensate both the Gaussian means and the variances and can be efficiently applied to the features, rather than the model parameters.

In our previous work, we relied on *a priori* knowledge of the speaker identity and the environment. For example, we assumed that the user may be identified using a device hardware code, caller ID on a phone, or a login name. Similarly, in many "situated" applications, the environment can also be easily known, e.g., in an in-car voice control system or a living room game console. While these assumptions may be reasonable in some cases, there are many more situations where this

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

will not be true. In particular, for speech applications on mobile phones, it is still reasonable to assume knowledge of the speaker, but the acoustic environment cannot be known *a priori*.

In this paper, we present an improved method for factored adaptation that eliminates the need for environment labels. In the proposed algorithm, unsupervised clustering is used to characterize and then compensate the environmental variability. Unlike the previous algorithm that relied on labeled environments, this method can operate in both environments seen in training and those that are unseen. We demonstrate through a series of experiments that the factored adaptation algorithm can perform joint environment and speaker compensation that a manner which enables the effective reuse of the speaker transforms across multiple environments.

The remainder of the paper is organized as follows. In Section II, we introduce the concept of factored adaptation and factored transforms. In Section III, we show how these transforms can be estimated from adaptation data or training data. The method used for unsupervised environment clustering is described in Section IV. In Section V, we show how adaptive training can be performed using factored transforms. Finally, experiments that demonstrate the efficacy of the proposed algorithm are described in Section VI and some concluding remarks are made in Section VII.

II. Factored adaptation using a cascade of transforms

In this work, we assume that environmental variability and speaker variability can be jointly compensated for by a single CMLLR feature transform,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{1}$$

and that this transform can be decomposed into two distinct transforms that capture each of these sources of variability separately. This can be expressed as

$$\mathbf{y} = \mathbf{A}_s(\mathbf{A}_e\mathbf{x} + \mathbf{b}_e) + \mathbf{b}_s \tag{2}$$

where $\mathbf{W}_s = \{\mathbf{A}_s, \mathbf{b}_s\}$ and $\mathbf{W}_e = \{\mathbf{A}_e, \mathbf{b}_e\}$ represent the speaker and environment transforms, respectively. The equivalence between (1) and (2) can be seen by letting $\mathbf{A} = \mathbf{A}_s\mathbf{A}_e$ and $\mathbf{b} = \mathbf{A}_s\mathbf{b}_e + \mathbf{b}_s$.

Because the relationship between the environmental and speaker transforms is linear, the transforms in (2) can be also applied in the reverse order where the speaker transform is applied first. This is an equivalent model for factored adaptation but the transforms learned would be different, as matrix operations are not commutative in general.

III. Factored Transform Estimation

Let us assume that adaptation data exists from many speakers in one or more different environments. Let $\mathbf{\Lambda}_S$ be the set of speaker transforms for S different speakers in the data. Similarly, let $\mathbf{\Lambda}_E$ be the set of environmental transforms for the E different environments in the data. Given this adaptation data, the goal is to estimate the set of transforms $(\mathbf{\Lambda}_E, \mathbf{\Lambda}_S)$ by maximizing the likelihood of the data. If we define i, t, and k as the indices for the utterance, the frame and the Gaussian component, respectively, we can the write the following auxiliary function

$$\mathcal{Q}(\mathbf{\Lambda}_E, \mathbf{\Lambda}_S) = \sum_{i,t,k} \gamma_{tk}^{(i)} \log(p(\mathbf{y}_t^{(i)}|k)) \tag{3}$$

where $\mathbf{y}_t^{(i)}$ is defined according to (2) and $p(\mathbf{y}_t^{(i)}|k))$ is a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{\Sigma}_k$.

Because any linear transform defined in (1) can be *arbitrarily* factored into two transforms as in (2), it is impossible to separate the environmental and speaker variability without additional constraints. Thus, we make some assumptions about the nature of the adaptation data. First, we assume that each utterances has an associated speaker label and environment label. These labels may represent the true speaker and/or environment or some other identifier such as cluster membership. In addition, we assume that there is a significant diversity of speakers in each environment of interest. Using these assumptions, each of the transforms in $(\mathbf{\Lambda}_E, \mathbf{\Lambda}_S)$ is optimized using a distinct (but overlapping) set of data.

A. Optimizing the speaker transforms

To optimize the speaker transform for speaker s, we define i_s as the index over all utterances from that speaker and rewrite the auxiliary function as

$$\mathcal{Q}(\mathbf{W}_s, \bar{\mathbf{W}}_s, \bar{\mathbf{\Lambda}}_E) = \sum_{i_s,t,k} \gamma_{tk}^{(i_s)} \log(p(\mathbf{y}_t^{(i_s)}|k)) \tag{4}$$

Throughout this paper, a bar on top of a variable, e.g. $\bar{\mathbf{A}}$, represents the current estimate of that variable. Under this objective function, $\mathbf{y}_t$ can be written as

$$\mathbf{y}_t = \mathbf{A}_s(\bar{\mathbf{A}}_{e(i_s)}\mathbf{x}_t^{(i_s)} + \bar{\mathbf{b}}_{e(i_s)}) + \mathbf{b}_s \tag{5}$$

$$= \mathbf{A}_s\bar{\mathbf{x}}_{e,t}^{(i_s)} + \mathbf{b}_s \tag{6}$$

where $e(i_s)$ is the environment for the utterance i_s and $\bar{\mathbf{x}}_{e,t}^{(i_s)}$ is the observation with the transform for environment e applied. Thus, the log probability in (4) can be written as

$$\log(p(\mathbf{y}_t^{(i_s)}|k)) = \log(|\mathbf{\Sigma}_k|) - \log(|\mathbf{A}_s|)^2) + (\mathbf{A}_s\bar{\mathbf{x}}_{e,t}^{(i_s)} + \mathbf{b}_s - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{A}_s\bar{\mathbf{x}}_{e,t}^{(i_s)} + \mathbf{b}_s - \boldsymbol{\mu}_k) \tag{7}$$

From (7), it is clear that the auxiliary function in (4) is equivalent to that of conventional CMLLR where the observations are replaced by the environmental-transformed features and the standard row-by-row optimization procedure can be employed [5].

B. Optimizing the environment transforms

To update the set of environment transforms, we define a similar auxiliary function for each of the environments, in which we define an index i_e that indexes all utterances from that environment.

$$\mathcal{Q}(\mathbf{W}_e, \bar{\mathbf{W}}_e, \bar{\mathbf{\Lambda}}_S) = \sum_{i_e,t,k} \gamma_{tk}^{(i_e)} \log(p(\mathbf{y}_t^{(i_e)}|k)) \tag{8}$$

This is similar to (4) except that the set of utterances is different and the speaker transforms are now assumed fixed. In this case,

$$\mathbf{y}_t^{(i_e)} = \bar{\mathbf{A}}_{s(i_e)}(\mathbf{A}_e\mathbf{x}_t^{(i_e)} + \mathbf{b}_e) + \bar{\mathbf{b}}_{s(i_e)} \quad (9)$$

where $s(i_e)$ is the speaker for utterance i_e. The log probability in (8) can be then be expressed as

$$\log(p(\mathbf{y}_t^{(i_e)}|k)) = \log(|\mathbf{\Sigma}_k|) - \log(|\bar{\mathbf{A}}_s)|^2) - \log(|\mathbf{A}_e|^2) + (\mathbf{y}_t^{(i_e)} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{y}_t^{(i_e)} - \boldsymbol{\mu}_k) \quad (10)$$

Substituting (9) into (10) and rearranging terms gives

$$\log(p(\mathbf{y}_t^{(i_e)}|k)) = \log(|\bar{\mathbf{\Sigma}}_{k,s(i_e)}|) - \log(|\mathbf{A}_e|^2) + (\mathbf{x}_{e,t}^{(i_e)} - \bar{\boldsymbol{\mu}}_{k,s(i_e)})^T\bar{\mathbf{\Sigma}}_{k,s(i_e)}^{-1}(\mathbf{x}_{e,t}^{(i_e)} - \boldsymbol{\mu}_{k,s(i_e)}) \quad (11)$$

where

$$\mathbf{x}_{e,t}^{(i_e)} = \mathbf{A}_e\mathbf{x}_t^{(i_e)} + \mathbf{b}_e \quad (12)$$

$$\bar{\boldsymbol{\mu}}_{k,s(i_e)} = \bar{\mathbf{A}}_{s(i_e)}^{-1}(\boldsymbol{\mu}_k - \bar{\mathbf{A}}_{s(i_e)}\bar{\mathbf{b}}_{s(i_e)}) \quad (13)$$

$$\bar{\mathbf{\Sigma}}_{k,s(i_e)} = \bar{\mathbf{A}}_{s(i_e)}^{-1}\mathbf{\Sigma}_k\bar{\mathbf{A}}_{s(i_e)}^{-1,T} \quad (14)$$

By substituting (11) – (14) into (8), we can see that optimizing the environmental transforms is equivalent to performing CMLLR with adapted Gaussian parameters given by (13) and (14). Note that the adapted covariances have the same stucture as the speaker transforms. If the transforms are full matrices, then so are the covariance matrices. In this case, the row-by-row optimization for full covariances must be used [11].

C. Jointly optimizing the speaker and environmental transforms

Because there is no closed-form for solution to optimizing the full set of transforms jointly, the speaker and environmental transforms are optimized alternately. After choosing initial values for the transforms, the environment transforms are estimated while the speaker transforms are fixed, and then vice versa. This process can be repeated for a fixed number of iterations or until the likelihood of the adaptation data converges.

In this work, the following recipe was used:

1) Initialize the transforms. All $\mathbf{A}$ matrices were initialize to identity and all offset vectors $\mathbf{b}$ were initialized to zero.
2) Fix speaker transforms $\mathbf{\Lambda}_S$ and optimize $\mathbf{W}_e$ for each environment $e = \{1, \ldots, E\}$.
3) Fix environmental transforms $\mathbf{\Lambda}_E$ and optimize the speaker transforms $\mathbf{W}_s$, $s = \{1, \ldots, S\}$.
4) If more iterations desired, go to step 2.

In the experiments reported in this paper, we performed a single iteration of this joint optimization and used full matrices for all transforms. Because we chose to start with the optimization of the environment transforms with the speaker transforms initialized to $\mathbf{A}_s = \mathbf{I}$ and $\mathbf{b}_s = 0$, the environment transforms could be optimized with conventional CMLLR with a diagonal covariance Gaussians, rather than the full covariances indicated by (14). If a second iteration were to be performed full-covariance optimization would be required.

IV. Environment clustering

In order to perform factored adaptation, groups of utterances with common speakers or environments must be identified. While it is in theory possible to label the environments seen in the training data by listening to the utterances, this solution is obviously not scalable to large amounts of data. Alternatively, we propose to automatically cluster the environments in an unsupervised way.

Many methods for environment clustering have been proposed in the literature. These methods have sometimes been referred to as "acoustic sniffing" algorithms [12], [13]. In this work, we use a simple approach that uses a Gaussian mixture model trained on the silence regions of the utterances in the training data. After the mixture model has been trained, each Gaussian in the model is assumed to define a unique acoustic environment. The utterances in the training set are labeled by computing the average posterior probability of the silence segments in the utterance and finding the Gaussian in the mixture with the highest score. All the data associated with each cluster is then used to train the transform for that environment. At run time, the initial silence portion of the utterance is used to estimate the environmental cluster for the utterance. The transform for the cluster with the highest posterior probability is then applied to the utterance.

V. Adaptive training with factored transforms

Because both the environment and speaker transforms are linear operations on the features, they can be combined into a single linear transform that can be applied to the features. As a result, performing adaptive training [14] is quite straightforward. To do so, we simply add the set of HMM parameters $\mathbf{\Lambda}_X$ to the auxiliary function in (3),

$$\mathcal{Q}(\mathbf{\Lambda}_X, \mathbf{\Lambda}_E, \mathbf{\Lambda}_S) = \sum_{i,t,k} \gamma_{tk}^{(i)} \log(\mathcal{N}(\mathbf{y}_t^{(i)}, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)) \quad (15)$$

As in the recipe in Section III-C, the speaker transforms, environment transforms, and acoustic model parameters are each optimized in succession while the other parameters are held fixed. To update the acoustic model parameters, the speaker and environment transforms are combined into a single linear transform (depending on the speaker and the environment of the utterance) and then the acoustic model parameters can be updated using the transformed features. We refer to this training as Speaker and Environment Adaptive Training (SEAT) to reflect that fact that separate transforms are estimated to compensate for the speaker variability and the environmental variability.

VI. Experiments and results

In order to validate the proposed factored adaptation algorithm, we performed a number of experiments using the Aurora 2 corpus [15]. Aurora 2 consists of data degraded with eight types of noise at a range of SNRs. Evaluation is performed using three test sets that contain noise types seen in the training data (Set A), unseen in the training data (Set B), and additive noise plus channel distortion (Set C). There

are 110 speakers in the training set and 104 speakers in the test set with no overlap between the two sets. In this work, our evaluation is limited to Set A and Set B.

The acoustic models were trained from the multi-condition training set using HTK with the "complex back end" recipe. An HMM with 16 states per digit and 20 Gaussians per state is created for each digit as a whole word model. There is a three-state silence model with 36 Gaussians per state and a one state short pause model tied to the middle state of silence. Standard 39-dimensional MFCC features consisting of 13 static, delta, and delta-delta features were computed from power spectral observations and C0 was used instead of log energy. The baseline system included cepstral mean normalization (CMN) and had a word error rate [1] (WER) of 7.30% on Set A and 7.41% on Set B .

A. Environment compensation in unknown environments

In the first series of experiments, we wanted to assess the performance of CMLLR transforms for compensating environmental distortions both when the environments in training and test are known *a priori* and in the more challenging case when the environments are unknown. In these experiments, the following procedure was followed. The training data was first clustered by environment and then a single CMLLR transform was estimated per cluster using the training data with known transcriptions. During evaluation, each test utterance was assigned to a cluster and the corresponding CMLLR transform learned from the training data was applied prior to decoding. For the experiments where the environments were assumed to be known, the training data was clustered simply using the noise and/or SNR labels associated with each utterance. At test time, cluster assignment was likewise done using the label associated with each test utterance. The results obtained using transforms estimated from labeled environments are shown in Table I. The table shows two different methods of environmental clustering using the labels available in Aurora 2. The first uses four clusters that represent the four noise types N1-N4 in the multi-condition training data. The second uses twenty clusters that represent the combination of four noise types and five SNRs in the training data. As these results indicate, even just using four clusters results in a significant improvement over the baseline while additional improvement is obtained with more granular clusters that utilize on both noise type and SNR. Note that we can only obtain results for Set A, since the noises in Set B are unseen in training.

To evaluate performance when the environmental labels are unknown, the unsupervised GMM-based clustering described in Section IV was performed. A GMM was trained using the first and last 20 frames of all utterances in the multi-condition training set. The GMM was trained using static cepstral features without any preprocessing such as CMN or AGC. All utterances in the training set were then assigned to a cluster by finding the Gaussian in the the mixture with the highest average posterior probability computed from the first 20 frames of the utterance. Once the training data was clustered, a CMLLR transform was then estimated for each Gaussian in the mixture. At runtime, the first 20 frames of each test utterance were used to determine the utterance's cluster assignment and the corresponding environment transform was applied.

[1]In Aurora 2, only SNRs 0-20 are included when reporting average WER

TABLE I
WER SET A AND SET B USING A SINGLE CMLLR TRANSFORM FOR EACH ENVIRONMENT. THE TRANSFORMS WERE ESTIMATED USING THE MULTI-CONDITION TRAINING DATA. THE BASELINE WER FOR SET A IS 7.3% AND FOR SET B IS 7.41%

Environment Clustering	Number of Clusters	Set A % WER	Set B % WER
Supervised (Labels)	4	6.86	–
	20	6.24	–
Unsupervised (GMM)	4	6.86	7.22
	8	6.63	7.22
	16	6.45	7.12
	32	6.43	7.08

The unsupervised results in Table I indicate using the proposed GMM-based clustering method can provide performance that is comparable to that obtained using labeled noise types and close to the performance achieved where the SNR is known as well. In addition, because the technique is unsupervised, it can be applied to Set B. In this case, even though the noises in these utterances were not seen in training, we can still find the CMLLR transform that best matches that environment using the GMM. In addition, because the determining the environment cluster identity only requires the initial silence of the utterance, the appropriate transform can be applied in first-pass decoding.

B. Factored adaptation in unknown environments

Because the goal of this work is to find speaker transforms that can be reused across different environments we next performed a series of experiments to determine how traditional adaptation methods and those proposed in this work can perform in this context. We first evaluated the performance of CMLLR when speaker transforms estimated using the test data from environment N1 in Set A (subway) are applied to the test data from environments N2-N4 in both Set A (car, babble, exhibition hall) and Set B (street, airport, train station). All speaker transforms were estimated in an unsupervised manner. As shown in Table II, this approach provides only a minimal improvement over the CMN baseline, as the transforms estimated are adapting to both the speaker and the subway noise environment. These transforms are sub-optimal when applied to other environments. The table next shows the performance obtained using only the CMLLR environment transforms estimated using the unsupervised clustering approach described in the previous experiment. As the table indicates, more significant improvements are obtained even when the noise environments are different from those used to train the transforms. Note that these results differ from those shown in Table I because we have excluded the results from environment N1 in order to be compatible with the other results in this table.

TABLE II
WER FOR SET A AND SET B USING DIFFERENT ADAPTATION STRATEGIES.

	Set A N2-N4	Set B N2-N4
Baseline CMN	7.91	7.52
CMLLR	7.86	7.28
CMLLR (environment)	6.82	7.23
Factored CMLLR	6.39	6.55

Finally, we evaluated the performance of the proposed factored adaptation approach. For each test utterance from environment N1 in Set A, the best environment transform was determined by finding the appropriate cluster using a 32-mixture environment GMM. Once the appropriate environment transform had been applied to each utterances from a given speaker, the speaker transform was then estimated. These speaker transforms were then used in conjunction with the appropriate environment transforms to decode the test data from the other environments. As shown in Table II, the proposed factored adaptation method results is a significant improvement in performance over the other methods. In particular, an 18% and 10% relative improvement over conventional CMLLR speaker adaptation was obtained on Set A and Set B, respectively. This demonstrates that the factored adaptation approach can provide a significant benefit in scenarios where the users may operate from a wide range of environments, even those unseen in the training data.

As stated in Section V, one benefit of using CMLLR transforms as the basis of factored adaptation is that performing adaptive training is straightforward and efficient as the features need to simply be transformed appropriately prior to conventional HMM training. A final series of experiments was performed to evaluate the effect of adaptive training on factored adaptation. The previous set of experiments comparing CMLLR and factored adaptation were repeated using adaptively trained models in both cases. The results for Set A and Set B are shown in Tables III and IV. Comparing these results to those in Table II, it is clear that adaptive training improves the performance of all methods and the relative improvements of factored adaptation over conventional CMLLR are maintained. These tables also show the results obtained using VTS adaptation in conjunction with Noise Adaptive Training [3]. This approach is representative of state-of-the-art performance in environmental adaptation. It is interesting to observe that a single linear feature transform, computed using factored adaptation can approach, and in some cases exceed, the performance of VTS adaptation, a much more complex algorithm that adapts the means and variances of every Gaussian in the HMM individually for every utterance.

TABLE III
WER ON SET A USING ADAPTIVE TRAINING WHEN SPEAKER TRANSFORMS ESTIMATED FROM ENVIRONMENT N1 IN SET A ARE APPLIED TO ENVIRONMENTS N2–N4. THE PERFORMANCE OF NAT-VTS IS SHOWN FOR COMPARISON.

Set A N2–N4	Baseline CMN	CMLLR + SAT	F-CMLLR + SEAT	VTS + NAT
Clean	0.52	0.34	0.32	0.38
20 dB	0.85	0.60	0.52	0.76
15 dB	1.30	0.97	0.79	1.10
10 dB	2.62	2.05	1.77	2.46
5 dB	7.51	6.86	5.51	6.39
0 dB	27.25	27.13	22.47	21.12
-5 dB	67.67	67.56	64.12	57.77
Avg	7.91	7.52	6.21	6.37

TABLE IV
WER ON SET B USING ADAPTIVE TRAINING WHEN SPEAKER TRANSFORMS ESTIMATED FROM ENVIRONMENT N1 IN SET A ARE APPLIED TO ENVIRONMENTS N2–N4. THE PERFORMANCE OF NAT-VTS IS SHOWN FOR COMPARISON.

Set B N2–N4	Baseline CMN	CMLLR + SAT	F-CMLLR + SEAT	VTS + NAT
Clean	0.52	0.34	0.32	0.38
20 dB	0.84	0.53	0.62	0.67
15 dB	1.38	0.90	0.93	0.99
10 dB	2.77	2.12	2.13	2.04
5 dB	7.31	6.34	6.06	6.02
0 dB	25.30	24.94	21.65	18.78
-5 dB	66.01	65.95	63.22	53.83
Avg	7.52	6.97	6.28	5.70

VII. CONCLUSION

In this paper, we have proposed a method called factored adaptation which can jointly compensate for speaker and environmental variability using CMLLR adaptation. By using this algorithm in combination with appropriate selection of the adaptation data, separate transforms for the speaker and the environment can be estimated. Furthermore we have proposed a method for environment clustering which enables this approach to be used in scenarios in which the environment is unknown. Through a series of experiments, we have shown that factored adaptation enables the speaker transforms learned in one environment to be effectively applied to speech from the same user in different environments, resulting in a relative improvement over conventional CMLLR adaptation of 18% on environments seen in training and 10% on unseen environments. We have also shown how this method can be incorporated into an adaptive training strategy which generates further improvements in performance. In the future, we plan to further develop this approach by estimating multiple transforms using regression classes.

REFERENCES

[1] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Sp. and Audio Proc.*, vol. 4, pp. 352–359, 1996.

[2] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speechrecognition," in *Proc. of ICASSP*, vol. 2, 1996.

[3] O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero, "Noise adaptive training for robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1889 –1901, Nov. 2010.

[4] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proc. of ICASSP*, Honolulu, Hawaii, 2007.

[5] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[6] G. Saon, J. M. Huerta, and E. E. Jan, "Robust Digit Recognition in Noisy Environments: the IBM Aurora 2 System," in *Proc. of Interspeech*, Aalborg, Denmark, 2001.
[7] L. Rigazio, P. Nguyen, D. Kryze, and J.-C. Junqua, "Separating speaker and environmental variabilities for improved recognition in non-stationary conditions," in *Proc. Eurospeech*, Aalborg, Denmakrk, 2001.
[8] Y.-Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the Aurora4 task," in *Proc. ICASSP*, Prague, Czech Republic, 2011.
[9] M. J. F. Gales, "Acoustic factorisation," in *Proc. ASRU*, Moreno, Italy, 2001.
[10] M. L. Seltzer and A. Acero, "Separating speaker and environmental variability using factored transforms," in *Proc. Interspeech*, Florence, Italy, Aug. 2011.
[11] K. C. sim and M. J. F. Gales, "Adaptation of Precision Matrix Models on Large Vocabulary Continuous Speech Recognition," in *Proc. of ICASSP*, Philadelphia, PA, 2005.
[12] M. Akbacak and J. H. L. Hansen, "Environmental sniffing: Noise knowledge for robust speech systems," *Speech and Audio Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 465–477, February 2007.
[13] G. Shi, Y. Shi, and Q. Huo, "A study of irrelevant variability normalization based training and unsupervised online adaptation for lvcsr," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010.
[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. of ICSLP*, Philadelphia, PA, 1996.
[15] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ITRW ASR*, Paris, France, September 2000.