Separating Speaker and Environmental Variability Using Factored Transforms

Michael L. Seltzer, Alex Acero

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

Abstract

Two primary sources of variability that degrade accuracy in speech recognition systems are the speaker and the environment. While many algorithms for speaker or environment adaptation have been proposed to improve performance, far less attention has been paid to approaches which address for both factors. In this paper, we present a method for compensating for speaker and environmental mismatch using a cascade of CMLLR transforms. The proposed approach enables speaker transforms estimated in one environment to be effectively applied to speech from the same user in a different environment. This approach can be further improved using a new training method called speaker and environment adaptive training method. When applying speaker transforms to new environments, the proposed approach results in a 13% relative improvement over conventional CMLLR.

Index Terms: speaker adaptation, environment adaptation, robustness, factored transforms

1. Introduction

The performance of speech recognition systems degrades when there is mismatch between the acoustic models of the recognizer and the speech seen in deployment. Two primary sources of this mismatch are the speaker and the environment. One way in which this mismatch can be mitigated is to adapt the acoustic models to the current conditions. While many algorithms for speaker or environmental adaptation have been proposed, e.g. [1, 2], far less attention has been paid to approaches which address both factors. Nevertheless, it would be advantageous to be able to adapt a recognizer to the speaker and environment in a way allows these sources of variability to be separated. For example, transforms estimated for speaker adaptation in one environment could be applied to speech from the same speaker in a new environment.

A method of joint environment and speaker adaptation was proposed in which Jacobian adaptation for noise compensation was combined with MLLR for speaker adaptation [3]. This approach was recently improved by using Vector Taylor Series (VTS) adaptation to update both the means and variances of MLLR-compensated acoustic models [4]. The VTS noise parameters and the MLLR transforms were jointly estimated using an iterative approach. By combining methods that use different adaptation strategies improved separation of the speaker parameters and the environmental parameters can be achieved. This separation was called acoustic factorization in [5]. In this work, a product of MLLR transforms was proposed where one transform captured the environmental variability and one captured the speaker variability. The environmental transform was estimated using a cluster-adaptive training approach and the speaker transform was estimated with conventional MLLR.

In this paper, we present a method for compensating for speaker and environmental mismatch using a cascade of CMLLR transforms. Because this cascade of transforms is itself a CMLLR transform, we refer to it as a factored transform. We propose a method for estimating factored transforms in order to identify the two constituent transforms that best capture the speaker and environmental variability in the adaptation data. The goal of this work is to estimate speaker transforms that can be applied to speech from the same speaker in a different environment. We believe there are several benefits to the proposed approach. First, because both transforms are estimated using a data-driven approach, no assumptions about the underly acoustic model or features have to be made, unlike VTS, which assumes a clean acoustic model trained using mel cepstral or similar features. The proposed method can use any features and acoustic model. In addition, using linear transforms makes the use of adaptive training straightforward and does not require the more complicated and computationally-expensive noise-adaptive training approaches recently proposed [6, 7]. Finally, the proposed approach is more efficient than the MLLR approach in [5], as CMLLR can be implemented using a transformation of the features rather than the model parameters.

The remainder of this paper is organized as follows. In Section 2, we introduce the concept of factored transforms. Section 3 shows how these transforms can be estimated from adaptation data or training data. Adaptive training using the proposed factored transforms is discussed in Section 4. Finally, experiments to evaluate the performance of the proposed approach are described in Section 5 and some concluding remarks are made in Section 6.

2. Factored transforms

The basis of the adaptation in this work is Constrained MLLR (CMLLR) which applies the same linear transform to both the Gaussian means and variances. The advantage of CMLLR is that it can be implemented as a feature transform which means that no changes to the acoustic models are required at runtime if only a global transformation is used. If regression classes are used, then the determinant of the transform needs to be accounted for when computing acoustic likelihoods. In conventional CMLLR, the features are transformed using a linear transform as

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

Let's assume that there exists a linear transform that compensates for environmental variability $\mathbf{W}_e = \{\mathbf{A}_e, \mathbf{b}_e\}$ and a second transform that compensates for speaker variability

$\mathbf{W}_s = \{\mathbf{A}_s, \mathbf{b}_s\}$. Applying these in succession leads to

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

It is straightforward to show that this is equivalent to a single transform $\mathbf{y} = \mathbf{A}'\mathbf{x} + \mathbf{b}'$ where $\mathbf{A}' = \mathbf{A}_s\mathbf{A}_e$ and $\mathbf{b}' = \mathbf{A}_s\mathbf{b}_e + \mathbf{b}_s$. The speaker and environment transforms can also be applied in the reverse order (speaker transform first). This is an equivalent representation though obviously the transforms learned will be different as the relationship is not commutative.

3. Estimating the transforms

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. Note that while the definition of a "speaker" is clear and well-defined, the definition of "environment" is less so. It can be defined by the noise type that corrupts the speech, some combination of noise and SNR, or some alternate definition. 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

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

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 $\mathbf{\Sigma}_k$.

Obviously, any linear transform defined in (1) can be *arbitrarily* factored into two transforms as in (2). Thus, without additional considerations, it is impossible to have one transform capture environmental variability while the other captures speaker variability. Thus, we make some assumptions about the nature of the adaptation data. First, we assume that we know the identity of the environment and the speaker in each utterance. In addition, we assume that there is a significant diversity of speakers in each environment of interest.

Both of these assumptions are realistic in many practical applications. For example, it is reasonable assume the environment of many "situated" systems such as an in-car voice control system or a living room game console. In addition, the speaker identity can be determined using a device code, caller ID on a phone, or a user login. Using these assumptions, each of the transforms in $(\mathbf{\Lambda}_E, \mathbf{\Lambda}_S)$ is optimized using a distinct (but overlapping) set of data.

3.1. Optimizing the speaker transforms

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

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

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

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

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

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

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

Clearly, 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 [1].

3.2. Optimizing the environment transforms

In order to update the environmental transforms, we define an index i_e that indexes all utterances from a common environment. We then define a similar objective function to (4) for a set of environmental transforms using the following auxiliary function.

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

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 standard row-by-row optimization cannot be performed and other techniques must be used.

3.3. 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 this work, 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 environmental 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.

4. Speaker and Environment Adaptive Training

Because both the environment and speaker transforms are linear operations on the features, performing adaptive training [8] 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 3.3, 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. In contrast to traditional speaker adaptive training (SAT), we are explicitly updating separate transforms that account for speaker variability and environmental variability. As a result, we refer to this training as Speaker and Environment Adaptive Training (SEAT).

5. Experiments and Results

In order to evaluate the proposed method for adaptation using factored transforms, a series of experiments were performed using the Aurora 2 corpus [9]. Aurora 2 consists of data degraded with eight types of noise at SNRs between 0 dB and 20 dB. 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.

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 accuracy on Set A of 92.70%.

Table 1: *Word accuracy for Set A using batch unsupervised CMLLR for each combination of speaker + environment*

Set A	N1	N2	N3	N4	Avg
Clean	99.60	99.58	99.58	99.63	99.60
20 dB	99.60	99.33	99.64	99.48	99.51
15 dB	99.29	99.06	99.25	99.07	99.17
10 dB	98.10	98.46	98.21	97.75	98.13
5 dB	95.79	94.74	93.95	93.12	94.40
0 dB	84.31	76.42	69.85	79.94	77.63
-5 dB	44.46	33.74	24.84	39.93	35.74
Avg	95.42	93.60	92.18	93.87	93.77

The experiments performed were designed to test the ability of the proposed factored transform approach to separate the speaker and environmental variability. This was done by estimating the speaker transforms from speech in one environment and evaluating their effectiveness when applied to speech from the same speaker in different environments. In all experiments, the environment was defined by the type of noise, regardless, of SNR. Thus, in the training data and Set A, there are four environments. These will be referred to as N1-N4, and correspond to subway, babble, car, and exhibition hall, respectively.

In the first experiment, we sought to establish the upper bound in performance using unsupervised CMLLR adaptation to jointly compensate for the combined effects of speaker and environment mismatch. To do so, we estimated a single CMLLR transform for each speaker+environment combination. Each of the four environments in Set A contains speech at 7 different SNRs (including clean speech). There are 100 speakers with 10 utterances per speaker per SNR which means that 70 utterances per speaker were used for adaptation. Using this data, a single CMLLR transform was estimated using the hypothesized transcriptions from the baseline CMN system. The utterances were then re-recognized after applying the estimated transforms. The results are shown in Table 1. This unsupervised batch adaptation using CMLLR results in 93.77% word accuracy, which represents a 14.6% relative reduction in word error rate from the baseline CMN system. This represents an upper bound on performance using batch adaptation with CMLLR where a transform is learned for each speaker + environment combination.

To evaluate the "portability" of conventional CMLLR transforms, the transforms estimated in the previous experiment for each speaker in environment N1 (subway) were applied to the utterances from the same speaker in the other three environments (N2–N4). The results are shown in Table 2. The accuracy on the unseen environments N2–N4 is 92.11% compared to an accuracy of 93.22% on the same environments in the previous experiment. This drop in performance reflects the fact that the transforms estimated in environment N1 are compensating for both the speaker and the environment. When the environment changes, the transforms are no longer optimal. Note that in contrast to the first experiment, the recognition results for environments N2–N4 are obtained with a single recognition pass.

This experiment was then repeated using the proposed factored transform approach. In this experiment, the factored adaptation algorithm described in Section 3 was first applied to the multi-condition training data. Four environmental transforms

Table 2: *Word accuracy for Set A when the CMLLR transforms estimated in environment N1 are applied to the remaining environments*

Set A	N2	N3	N4	Avg
Clean	99.52	99.64	99.60	99.59
20 dB	99.30	99.49	99.44	99.41
15 dB	99.06	99.22	98.95	99.08
10 dB	98.25	98.15	97.28	97.89
5 dB	93.77	92.96	92.32	93.02
0 dB	71.28	65.20	77.48	71.32
-5 dB	30.50	23.47	39.59	31.19
Avg	92.33	91.00	93.09	92.11

Table 3: *Word accuracy for Set A obtained using the proposed factored CMLLR transforms.*

Set A	N2	N3	N4	Avg
Clean	99.52	99.61	99.66	99.63
20 dB	99.33	99.64	99.60	99.61
15 dB	99.03	99.37	99.11	99.17
10 dB	98.31	98.30	97.59	98.06
5 dB	94.26	94.21	93.12	93.86
0 dB	74.58	70.50	79.17	74.74
-5 dB	32.74	25.38	38.82	32.31
Avg	93.10	92.40	93.72	93.07

were estimated (one for each of the four noise types) and 110 speaker transforms were estimated using supervised adaptation. At test time, the first-pass unsupervised transcripts from the baseline model were again used to estimate the speaker transforms using the N1 test data only. However, this time, the environmental transform for N1 learned in training was applied prior to estimating the speaker transforms. Then, these speaker transforms were used in conjunction with the transforms for environments N2–N4 to recognize the test data from those environments. As in the previous experiment, only a single recognition pass was required to obtain these results. As shown in Table 3, the recognition accuracy in the unseen environments improves to 93.07% which is quite close to our upper bound two-pass performance of 93.22%. These results represent a 12% relative reduction in word error rate over the conventional CMLLR approach in the previous experiment.

Finally, the impact of adaptive training on the proposed factored adaptation algorithm was evaluated. The previous two experiments were repeated using SAT with the conventional speaker-specific CMLLR transforms or the proposed SEAT using the factored transforms. As before, the speaker transforms estimated using utterances from environment N1 were applied to speech from environments N2–N4. The results are shown in Table 4. In both cases, the performance improves as expected. However, compared to SAT, a 13% relative reduction of WER is obtained by SEAT, which uses separate environmental and speaker transforms for adaptive training.

Table 4: *Word accuracy on Set A using SAT and the proposed SEAT when speaker transforms from N1 are applied to N2–N4.*

Set A N2–N4	CMLLR + SAT	F-CMLLR + SEAT
Clean	99.66	99.66
20 dB	99.40	99.54
15 dB	99.03	99.23
10 dB	97.95	98.21
5 dB	93.27	94.07
0 dB	72.87	76.36
-5 dB	32.44	33.70
Avg	92.50	93.48

6. Conclusion

In this paper, we have proposed a method for separating the speaker and environmental variability using factored CMLLR transforms. We have shown through a series of experiments that by appropriate selection of the adaptation data, the proposed method can estimate separate transforms for the speaker and the environment, which enables the speaker transforms to be effectively applied to speech from the same user in different 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 in order to estimate the both the environmental transforms and the speaker transforms for adaptation to both speakers and environments not seen in training.

7. References

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

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

[3] 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.

[4] Y.-Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the Aurora4 task," in *Proc. ICASSP*, Prague, Czech Republic, 2011.

[5] M. J. F. Gales, "Acoustic factorisation," in *Proc. ASRU*, Moreno, Italy, 2001.

[6] 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.

[7] 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, 2010.

[8] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. of ICSLP*, Philadelphia, PA, 1996.

[9] H.G. 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.