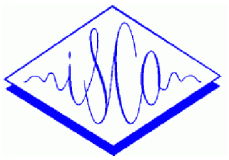
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Factored adaptation using a combination of feature-space and model-space transforms

Michael L . Seltzer, Alex Acero

Microsoft Research, Redmond, WA 99052, USA
{mseltzer,alexac}@microsoft.com

Abstract

Acoustic model adaptation can mitigate the degradation in recognition accuracy caused by speaker or environment mismatch. While there are many methods for speaker or environment adaptation, far less attention has been focused on methods that compensate for both simultaneously. We recently proposed an algorithm called factored adaptation which jointly estimates speaker and environment transforms in a manner which facilitates the reuse of transforms across sessions. For example, a speaker transform estimated in one environment can later be used even if the speaker's environment changes. In this paper, we introduce a new factored adaptation algorithm that uses a combination of feature-space and model-space transforms. We describe an iterative EM algorithm for transform estimation that also incorporates speaker and environment clustering in cases where the speaker or environment labels are unknown. On a large vocabulary voice search task, the proposed method consistently outperforms conventional adaptation.

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

10.21437/Interspeech.2012-488

1. Introduction

It is well-known that the performance of speech recognition suffers when there is a mismatch between the speech used to train the recognizer and that seen in deployment. Two sources of significant acoustic mismatch are the speaker and the environment. This mismatch can be mitigated by adapting the statistical distributions of the recognizer to be more representative of the observed speech. While there are many techniques that enable the recognizer to adapt to new speakers or environments, relatively little attention has been paid to methods that adapt to these two sources of mismatch jointly.

Recently, there has been interest in performing joint compensation of the speaker and the environment in a way which enables the sources of variability to be separated. Doing so would enable the adaptation transforms to be reused even in the presence of speaker or environment changes. For example, transforms estimated for a speaker in one environment could be reused even if the speaker is later in a different environment. An early approach to this used Jacobian adaptation for noise compensation and Maximum Likelihood Linear Regression (MLLR) for speaker adaptation [1]. This was later updated to use Vector Taylor Series (VTS) adaptation to update both the means and variances to the noisy environment [2]. Combining methods that use different adaptation strategies enables straightforward separation of the speaker adaptation parameters and the environmental adaptation parameters. This desirable separability was dubbed acoustic factorization in [3].

We recently proposed a technique called factored adaptation in which a cascade of Constrained MLLR (CMLLR) transforms was used to jointly compensate for the speaker and the environment[4, 5]. Unlike previous methods that used VTS or similar methods for environment adaptation, this approach does not require acoustic models trained from clean speech with mel cepstral features. In this method, unknown environments were clustered based on the noise in the initial silence segment of the utterance.

In this paper, we present a new factored adaptation algorithm which combines a CMLLR transform for speaker adaptation and an MLLR transform for environment adaptation. This approach has two distinct advantages over the previously proposed algorithm. First, it simplifies the use of either the speaker or environment transform in isolation so that the environment transform may be used without any speaker transform or vice versa. This was difficult in the previous algorithm because the CMLLR transforms in the cascade cannot be permuted. Second, iterative estimation of a cascade of CMLLR transforms requires computation equivalent to CMLLR estimation with full covariance Gaussian distributions. In contrast, using a combination of CMLLR and MLLR enables standard transform estimation algorithms to be used.

The proposed algorithm also improves upon previous methods in [2, 5] by using a more flexible definition of "environment" that goes beyond the type and level of additive noise to include any aspect of the observed speech that is common across multiple speakers. This is accomplished by using a general purpose algorithm for the environment adaptation (MLLR) rather than one designed specifically for mismatched caused by additive or convolutional noise such as VTS. In addition, the acoustic sniffing process previously used is replaced by an iterative clustering scheme that is tightly integrated with the EM-based estimation of the speaker and environment transforms.

The factored adaptation technique proposed in this paper is evaluated using a large vocabulary Bing Mobile voice search task. This represents the first evaluation of this style of joint speaker-environment adaptation on real-world data rather than on corpora that use artificially added noise.

The remainder of the paper is organized as follows. In Section 2, we introduce the proposed algorithm for factored adaptation. We then describe the algorithm used to jointly estimate the speaker and environment transforms. This algorithm is then evaluated through a series of experiments in Section 4 and some concluding remarks are made in Section 5.

2. Factored adaptation using a combination of feature and model transforms

In this work, we will compensate for speaker and environmental mismatch using a combination of linear transforms. In partic-

ular, a CMLLR transform will be used to adapt the acoustic models to the speaker while an MLLR transform will be used to adapt the models to the environment. If we assume that an utterance $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ comes from speaker s in environment e, we write the log-likelihood of the utterance under the adapted model as

$$\mathcal{L}(X) = \sum_{t,k} \gamma_{tk} \log p(\mathbf{x}_t | k, s, e) \quad (1)$$

where

$$\log p(\mathbf{x}_t | k, e, s) = \log(|\mathbf{A}_s|) + \log(\mathcal{N}(\mathbf{W}_s \mathbf{x}_t^+; \mathbf{W}_e \boldsymbol{\mu}_k^+, \boldsymbol{\Sigma}_k)) \quad (2)$$

where t is the frame index, k is the Gaussian index, and γ_{tk} is the posterior probability of Gaussian k give the observation at time t. The speaker and environment transforms are denoted as $\mathbf{W}_s = [\mathbf{A}_s \ \mathbf{b}_s]$ and $\mathbf{W}_e = [\mathbf{A}_e \ \mathbf{b}_e]$, respectively and the superscript '+' indicates an augmented vector, i.e. $\mathbf{x}_t^+ = [\mathbf{x}_t^T \ 1]^T$. For simplicity, we assume global transforms are used. However, it is straightforward to utilize regression classes and multiple transforms.

Because speaker and environment adaptation are performed using linear transforms, estimating the transforms in a way that results in the desired separability is challenging. Through algebraic manipulation of (1), it can be shown that the two transforms are equivalent to single linear transform applied to the mean (with the speaker transform also applied to the variance). As a result, straightforward maximum likelihood estimation of the transforms in (1) may result in an arbitrary distribution of the speaker and environmental variability between the two transforms rather than the desired outcome where the feature transform adapts to the speaker and the mean transform adapts to the environment. In the next section, we'll show how the transforms can be estimated in a way that enables the sources of variability to be factored in a meaningful way.

3. Joint estimation of speaker and environment transforms

In order to jointly estimate the speaker and environment transforms, we assume that the adaptation (or training) data $\mathcal{X} = \{X^{(1)}, \ldots, X^{(I)}\}$ consists of utterances from many speakers in one or more different environments. We can write the log likelihood of the adaptation data as

$$\mathcal{L}(\mathcal{X}) = \sum_{i,t,k} \gamma_{tk}^{(i)} \log p(\mathbf{x}_t^{(i)} | k, s(i), e(i)) \quad (3)$$

where $s(i)$ and $e(i)$ are the speaker and environment labels for utterance i, respectively, and $\log p(\mathbf{x}_t^{(i)} | k, s(i), e(i))$ is given by (2). The speaker and environment labels may be known *a priori* or can be estimated, as will be discussed in Section 3.3.

We now define $\boldsymbol{\Lambda}_S$ as set of speaker transforms for S different speakers and $\boldsymbol{\Lambda}_E$ as the set of environment transforms for E different environments. Our goal is to find find the set of transforms $(\boldsymbol{\Lambda}_S, \boldsymbol{\Lambda}_E)$ that maximize the likelihood expression in (3).

To avoid an arbitrary factorization of the mismatch between the two transforms, we require that a reasonable diversity of speakers in each of the E environments. Using this assumption, we can achieve a meaningful isolation of the speaker and environmental variability by iteratively estimating each of the speaker and environment transforms using a distinct but overlapping set of adaptation data.

3.1. Optimizing the speaker transforms

To optimize the set of speaker transforms, we segment the adaptation by speaker and generate an EM auxiliary function per speaker. The adaptation data for each speaker is then used to estimate a CMLLR transform. If we define i_s as the variable that indexes all utterances from speaker s, we can write the EM auxiliary function as

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

where $e(i_s)$ is the environment label for utterance i_s and the posterior probabilities are computed using the current set of speaker and environment transforms. Throughout this paper, a bar on top of a variable, e.g. $\bar{\mathbf{W}}_s$, represents the current estimate of that variable. The log probability in (4) can be expressed as

$$\log p(\mathbf{x}_t^{(i_s)} | k, s, e(i_s)) = \log(|\mathbf{A}_s|) + \log(\mathcal{N}(\mathbf{A}_s \mathbf{x}_t + \mathbf{b}_s; \bar{\boldsymbol{\mu}}_k^{e(i_s)}, \boldsymbol{\Sigma}_k)) \quad (5)$$

where $\bar{\boldsymbol{\mu}}_k^{e(i_s)} = \bar{\mathbf{A}}_{e(i_s)} \boldsymbol{\mu}_k + \bar{\mathbf{b}}_{e(i_s)}$ is the adapted mean using the current estimate of the environment transform for utterance i_s. Thus, the speaker transforms can be estimated using the conventional CMLLR row-by-row method [6] where the Gaussian means have been first transformed using the appropriate environment transforms.

3.2. Optimizing the environment transforms

A similar procedure is used to optimize the set of environment transforms except that now the data is partitioned to create environment-specific auxiliary functions that will span utterances from multiple speakers. If we use i_e to index the utterances from environment e, the auxiliary function can be written as

$$\mathcal{Q}(\mathbf{W}_e, \bar{\mathbf{W}}_e, \bar{\boldsymbol{\Lambda}}_S) = \sum_{i_e,t,k} \gamma_{tk}^{(i_e)} \log(p(\bar{\mathbf{x}}_t^{(i_e)} | k, s(i_e), e)) \quad (6)$$

where $\bar{\mathbf{x}}_t^{i_e} = \bar{\mathbf{A}}_{s(i_e)} \mathbf{x}_t + \bar{\mathbf{b}}_{s(i_e)}$ is the observed feature vector transformed by the current estimate of CMLLR transform for the speaker of utterance i_e. The log probability in (6) can be written as

$$\log p(\bar{\mathbf{x}}_t^{(i_s)} | k, s, e(i_s)) = \log(|\bar{\mathbf{A}}_s|) + \log(\mathcal{N}(\bar{\mathbf{x}}_t^{i_e}; \mathbf{A}_e \boldsymbol{\mu}_k + \mathbf{b}_e, \boldsymbol{\Sigma}_k)) \quad (7)$$

As seen from (6) and (7), estimating the environment transform is equivalent to standard MLLR where the observations are first transformed using the appropriate CMLLR speaker transforms and the posterior probabilities are computed using the current estimate of the speaker and environment transforms.

3.3. Clustering the environments/speakers using maximum likelihood

The previous two sections described how to estimate the speaker and environment transforms given the speaker and environment labels of each utterance. However, in many applications, speaker and/or environment may be unknown *a priori*. Even in scenarios where the general environment is known, such as in an automotive speech application or a living room game console, the conditions within that environment may vary

significantly. Similarly, the speaker identity or a proxy for that speaker, such as a device ID may be known in some cases, but in other cases, such information is unavailable. In these scenarios, speaker or environment clustering may be performed to generate estimated labels.

In previous work, we employed an environmental sniffing method in which a GMM was trained based on the silence segments of the training data and each Gaussian represented a distinct environment. This worked well for the Aurora 2 task, and in general such approaches work well if distinctions among environments can be made solely based on the type and level of additive noise observed. However, in many real world applications, there may be numerous differences between environments, and in general, the notion of "environment" can be more subtle. Ideally, we would like the term "environment" to encompass all the variability that is common across multiple speakers. In this case, using only the silence regions to cluster the environments may be suboptimal.

We instead propose to tightly integrate the environment and speaker clustering into the transform estimation process using the same maximum likelihood criterion used for adaptation. This will produce optimal clusters, in a likelihood sense. In every iteration of the algorithm the data is re-clustered by associating it with the speaker or environment transform that results in the highest likelihood. For example, environment clustering is performed as

$$e(i) = \underset{e}{\text{argmax}} \sum_{t,k} \gamma_{tk}^{(i)} \log(p(\mathbf{x}_t^{(i)}|k, \bar{\mathbf{W}}_s, \bar{\mathbf{W}}_e)) \quad (8)$$

using the current set of speaker clusters, speaker transforms, and environment transforms.

Similarly, speaker clustering is performed using the current set of environment clusters, speaker transforms and environment transforms as

$$s(i) = \underset{s}{\text{argmax}} \sum_{t,k} \gamma_{tk}^{(i)} \log(p(\mathbf{x}_t^{(i)}|k, \bar{\mathbf{W}}_s, \bar{\mathbf{W}}_e)) \quad (9)$$

If speaker or environment labels are available, than one or both of these clustering steps can be skipped. Note that because this clustering procedure can only find a local optimum, initializing the clusters in some sensible way is advisable.

3.4. Iterative EM algorithm for joint estimation of speaker and environment transforms

The steps described in Sections Section 3.1–Section 3.3 comprise an iterative EM algorithm for joint estimation of the speaker and environment transforms. The complete algorithm is shown Algorithm 1. Note that this algorithm can be performed either supervised or unsupervised, using known or estimated transcriptions, respectively.

Algorithm 1 Factored Adaptation using CMLLR and MLLR

Input: Adaptation data $\mathcal{X} = \{X^{(1)}, \ldots, X^{(I)}\}$ with transcriptions, HMM parameters Λ_X,
Output: Speaker and environment transforms $(\mathbf{\Lambda}_S, \mathbf{\Lambda}_E)$
  Initialize all transforms $(\mathbf{\Lambda}_S, \mathbf{\Lambda}_E)$ to identity
  Initialize speaker and environment labels for each utterance
  repeat
    for $i = 1$ to I **do**
      Find maximum likelihood speaker label $s(i)$ for utterance i using (9)
    end for
    for $s = 1$ to S **do**
      Update CMLLR speaker transform $\mathbf{W}_s$ using (4)
    end for
    for $i = 1$ to I **do**
      Find maximum likelihood environment label $e(i)$ for utterance i using (8)
    end for
    for $e = 1$ to E **do**
      Update MLLR environment transform $\mathbf{W}_e$ using (6)
    end for
  until likelihood converges

4. Experiments

To evaluate the effectiveness of the proposed algorithm, we performed a series of experiments using data from the Bing Mobile voice search application. This task consists of internet search voice queries made on a smartphone. The training set consists of approximately 300 hours of speech. Context-dependent HMMs with 3600 tied states and 36 Gaussians per state were trained using maximum likelihood estimation. The input features were 39-dimensional HLDA features derived from 52-dimensional MFCC features (static, plus first, second, and third order delta features). Decoding was performed with a trigram langauge model with a 65K word vocabulary.

A test set was designed specifically to evaluate adaptation. The test set consisted of 5400 utterances with 20 utterances from 270 different speakers, comprising about 5.5 hours of speech. The test set was designed so that all speakers in the test set had at least 20 utterances in the training set. Thus,all speakers in the test set were unseen in training. In all experiments, the device ID associated with the voice query was used as a speaker label. No environment information was available.

In our initial experiments, we evaluated the performance of conventional CMLLR adaptation at creating transforms that are reusable across different utterances from the same speaker. In these experiments, the most recent 20 utterances in the training set from the 270 test speakers were used as adaptation data. Using these utterances, one or more CMLLR transforms were estimated for each speaker using supervised adaptation given the transcriptions. The transforms were then applied to the test utterances from the same speaker. The results are shown in Table 1. As the table shows, estimating a single global transform and applying it to future data from the same speaker provides a 3.6% relative reduction in WER. Interestingly, when two regression classes are used and a separate transform is estimated for the speech and non-speech classes, the performance degrades from the global transform case. Lastly, we repeated the experiment with two regression classes, but effectively turned off the transform for the non-speech class, by resetting it to an identity transform ($\mathbf{A} = \mathbf{I}, \mathbf{b} = \mathbf{0}$). In this case, the relative improvement is almost double that of the global transform. This result confirms that environmental variability (which most directly affects the distributions of the non-speech class) may be hindering the ability of the speaker transforms to generalize across utterances.

We then performed a series of experiments using the proposed factored adaptation algorithm described in this paper. The same adaptation set was used as in the previous experiments. The training procedure was as follows. All transforms were initialized to the identity matrix and an initial environment clustering was performed using vector quantization of the ini-

Table 1: *Performance of conventional CMLLR adaptation*

CMLLR Adaptation	WER (%)	Rel Imp (%)
none(baseline)	31.6	–
global transform	30.5	3.6
2 classes (speech,non-speech)	30.8	2.5
2 classes (speech,non-speech), non-speech xform $\rightarrow \{\mathbf{A} = \mathbf{I}, \mathbf{b} = \mathbf{0}\}$	29.7	6.2

Table 2: *Performance of proposed factored adaptation using CMLLR and MLLR*

Factored Adaptation CMLLR (spkr) + MLLR (env)	WER (%)	Rel Imp (%)
none(baseline)	31.6	–
global transform	28.9	8.6
2 classes (spch/nonspch)	29.3	7.3
2 classes (spch/nonspch), non-speech *speaker* xform $\rightarrow \{\mathbf{A} = \mathbf{I}, \mathbf{b} = \mathbf{0}\}$	28.6	9.6

tial silence portion of the adaptation utterances. As before, the speaker labels were obtained from the device ID and fixed. In all experiments, two environment clusters were used. Four iterations of factored transform estimation were performed. To estimate the environment label of the test utterances, the same maximum likelihood method used for environment clustering was used. That is, each utterance was decoded using the CMLLR transform associated with given speaker label and each of the environment MLLR transforms. For E environment clusters, this results in E competing hypotheses. The hypothesis that generated the highest likelihood was scored.

As in the first set of adaptation experiments, we performed experiments using global transforms and transforms estimated for two regression classes corresponding to speech and non-speech classes. As before, we also evaluated the performance when the CMLLR speaker transform for the non-speech class was set to the identity transform. However, in factored adaptation, MLLR transforms for both speech and non-speech classes were estimated to compensate for the environment. The results of the factored adaptation experiments are shown in Table 2. As the results in the table indicate, factored adaptation generates improvements of about $7 - 10\%$ relative over the baseline system and $3.5 - 5\%$ relative over conventional CMLLR adaptation. These modest but consistent improvements over conventional adaptation demonstrate the benefit of performing adaptation in a manner that enables the different sources of variability to be compensated separately.

Finally, we performed a set of experiments designed to evaluate the effectiveness of the factored transforms when used in isolation. That is, if an environment transform really adapts the acoustic models to a specific environment, they should provide benefit if applied without speaker adaptation. Similarly, a speaker transform that has been properly estimated should provide benefit regardless of whether an environment transform is used or not.

Table 3 shows the performance obtained by CMLLR speaker transforms estimated in the conventional manner, MLLR environmental transforms estimated using the proposed factored adaptation approach, and CMLLR transforms estimated using factored adaptation. The performance of the baseline system and the joint factored adaptation approach are also shown. In all cases, two regression classes were used and the speaker transform for the non-speech class was set to identity, as this provided the best performance in all algorithms.

Table 3: *Performance of CMLLR transforms vs. separate transforms from factored adaptation (FA)*

Adaptation Transforms	WER (%)	Rel Imp (%)
none (baseline)	31.6	–
CMLLR	29.7	6.2
FA, MLLR env xform only	30.0	5.1
FA, CMLLR speaker xform only	29.2	7.8
FA, CMLLR + MLLR	28.6	9.6

As the results in the table show, doing MLLR adaptation to a small number of environments (2 in this case) provides a 5% relative reduction in WER. Furthermore, applying the CMLLR transform estimating using the factored adaptation approach outperforms the conventional CMLLR transform even when the speaker transform for the non-speech class is unused. These results demonstrate that the separable compensation performed by the factored transforms enable these transforms to provide improvements even when used alone.

5. Conclusion

In this paper, we have proposed a new algorithm for factored adaptation which jointly compensates for speaker and environmental mismatch using a combination of feature-space and model-space transforms. By modeling the speaker and environmental variability separately, the transforms learned are more suitable for reuse across sessions, such as when a speaker is later in a different environment. The transforms are learned using an iterative EM algorithm that includes speaker and environment clustering steps if the speaker or environment labels are unknown. On a large vocabulary voice search task, the proposed algorithm reduced word error rate by up to 10% over the baseline system and 5% over conventional adaptation.

6. References

[1] 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.

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

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

[4] M. L. Seltzer and A. Acero, "Separating speaker and environmental variability using factored transforms," in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[5] M. L. Seltzer and A. Acero, "Factored Adaptation for Separable Compensation of Speaker and Environmental Variability," in *Proc. of ASRU*, Waikoloa, Hawaii, 2011.

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