An Explicit Independence Constraint for Factorised Adaptation in Speech Recognition

Y.-Q. Wang and M.J.F. Gales

Engineering Department, Cambridge University
Trumpington St. Cambridge University, CB2 1PZ, U.K.
{yw293,mjfg}@eng.cam.ac.uk

Abstract

Speech signals are usually affected by multiple acoustic factors, such as speaker characteristics and environment differences. Usually, the combined effect of these factors is modelled by a single transform. Acoustic factorisation splits the transform into several factor transforms, each modelling only one factor. This allows, for example, estimating a speaker transform in a noise condition and applying the same speaker transform in a different noise condition. To achieve this factorisation, it is crucial to keep factor transforms independent of each other. Previous work on acoustic factorisation relies on using different forms of factor transforms and/or the attribute of the data to enforce this independence. In this work, the independence is formulated in mathematically, and an explicit constraint is derived to enforce the independence. Using factorised cluster adaptive training (fCAT) as an application, experimental results demonstrates that the proposed explicit independence constraint helps factorisation when imbalanced adaptation data is used.

Index Terms: speaker adaption, robust speech recognition, acoustic factorisation

1. Introduction

For real-life applications, speech recognition systems must be able to handle complex acoustic environments where there may be multiple acoustic factors simultaneously affecting the speech signal. For example, a typical speech recogniser is often required to operate in a wide range of environments for a large number of possible users. Hence it must have the ability to adapt to the target speaker and environment condition rapidly. The conventional approach is to build a model transform for the combined speaker and noise, target condition. This requires adaptation data for all possible operating conditions. An alternative approach, acoustic factorisation first proposed in 2001 [1], has been adopted by a number of sites very recently e.g., [2, 3, 4, 5, 6]. In parallel with the factorisation approach in speech recognition, there is also work along this line in speech synthesis, e.g., [7, 8], where the goal is to synthesis the effect of multiple factors, such as speaker, language and emotion. The idea underlining acoustic factorisation is to divide the model transform into a set of *factor transforms*, each associated with only one distinct acoustic factor. This is illustrated in Figure 1, using speaker and environment adaptation as an example. Assuming that the impact of speaker and environment on the acoustic model can be represented by model transforms whose parameters are $\boldsymbol{\lambda}_{\mathrm{s}}$ and $\boldsymbol{\lambda}_{\mathrm{n}}$ respectively, two ellipses in figure 1 illustrate the speaker and environment coverage in the adaptation data, where point b and c represent two observed conditions. To adapt the model to the target condition a, only the speaker transform $\boldsymbol{\lambda}_{\mathrm{s}}^{(b)}$ and the noise transform $\boldsymbol{\lambda}_{\mathrm{n}}^{(c)}$ are needed, which can be estimated from adaptation data. This *factorised adaptation* is possible due to the *independence* between two factor transforms, as explained in the magnifier: when the operating condition is moved from point b towards point a, $\boldsymbol{\lambda}_{\mathrm{n}}$ get a small update $\Delta\boldsymbol{\lambda}_{\mathrm{n}}$ to reflect the environment transition; due to the independence between $\boldsymbol{\lambda}_{\mathrm{s}}$ and $\boldsymbol{\lambda}_{\mathrm{n}}$, speaker transform will not be affect by $\Delta\boldsymbol{\lambda}_{\mathrm{n}}$.

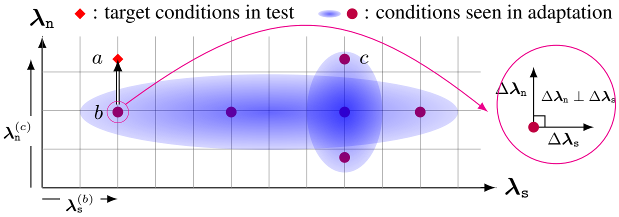
$\boldsymbol{\lambda}_{\mathrm{n}}$
◆ : target conditions in test
● : conditions seen in adaptation
a
b
c
$\boldsymbol{\lambda}_{\mathrm{n}}^{(c)}$
$\boldsymbol{\lambda}_{\mathrm{s}}^{(b)}$
$\boldsymbol{\lambda}_{\mathrm{s}}$
$\Delta\boldsymbol{\lambda}_{\mathrm{n}}$
$\Delta\boldsymbol{\lambda}_{\mathrm{n}} \perp \Delta\boldsymbol{\lambda}_{\mathrm{s}}$
$\Delta\boldsymbol{\lambda}_{\mathrm{s}}$

Figure 1: Factorised adaptation to the target speaker and environment: transforms independence.

Previous work on acoustic factorisation relies on two schemes to achieve the independence. In [2], two rather different forms of factor transforms (vector Taylor series[9], a non-linear transform, for the environment factor and maximum likelihood linear regression, MLLR [10], a linear transform, for the speaker factor) were used. It is hoped that by using different forms of factor transforms, each models the specific factor to which it is tuned. It was demonstrated that this achieves the independence to some extent while the MLLR transform is still modelling the noise effect. In [4], two constrained MLLR[11] transforms were cascaded to represent the speaker and environment distortion respectively. As both factors are modelled by linear transforms, there is no built-in mechanism to avoid a factor transform learning the effect of other factors. To achieve factorisation, the speaker transform is estimated on one speaker's data with a range of environments in a balanced manner. In this way, the speaker transform is hoped to be independent to the environment, thus achieving the independence. However, this requires data from the same speaker are distributed in an environment-balanced manner. In practice, it is quite common that the majority of a user's data is collected from the same environment. In this case, as there is no independence guaran-

This work was partially supported by Google research award and DARPA under the GALE and RATS program. The paper does not necessarily reflect the position or the policy of US Government and no official endorsement should be inferred.

tee, the estimated “speaker” transform will tend to model both speaker and the dominate environment. Work in [7, 8] also used linear transform to represent various acoustic factors and relies on balanced data to enforce the independence. The idea of factorisation has been also used in speaker recognition, e.g., joint factor analysis (JFA) [12] seeks to separate the speaker and the session variability, where both factors are represented by bias vectors in subspaces. However, there is no guarantee that the two subspaces are orthogonal, thus JFA also relies on the data balance to factor out the speaker variability. It is observed in [13] the speaker factor obtained by JFA still contains the session information.

In contrast to the *implicit* constraint approaches adopted in previous work, in this paper the dependence is analysed mathematically and an *explicit* constraint derived in section 2. As an application of this constraint, the factorised CAT (fCAT) model proposed in [8] is modified in section 3. Experiment results are presented and discussed in section 4.

2. An Explicit Constraint for Transform Independence

In this work, it is assumed there are two acoustic factors simultaneously affecting the speech signal: speaker characteristics $\mathtt{s}$ and environment differences $\mathtt{n}$. The proposed approach sits in the model-based framework [14], in which the intrinsic, phoneme variability is represented by a canonical model $\mathcal{M}_{\mathtt{c}}$ while the extrinsic, speaker and environment variability is represented by a set of transforms $\mathcal{T}$. The canonical model is adapted to the target i-th speaker and j-th noise condition by :

$$\mathcal{M}^{(i,j)} = \mathcal{F}(\mathcal{M}_{\mathtt{c}}, \mathcal{T}^{(i,j)}) \tag{1}$$

where $\mathcal{T}^{(i,j)}$ is the model transform that adapts the canonical model $\mathcal{M}_{\mathtt{c}}$ to $\mathcal{M}^{(i,j)}$, and $\mathcal{F}$ is the mapping function. To effectively deal with the complex acoustic environments, the concept of acoustic factorisation was proposed in [1], where the transform $\mathcal{T}^{(i,j)}$ is factorised into two components, each associated with one distinct acoustic factor, i.e.,

$$\mathcal{T}^{(i,j)} = \mathcal{T}_{\mathtt{s}}^{(i)} \otimes \mathcal{T}_{\mathtt{n}}^{(j)} \tag{2}$$

where $\mathcal{T}_{\mathtt{s}}^{(i)}$ and $\mathcal{T}_{\mathtt{n}}^{(j)}$ are the factor transforms for i-th speaker and j-th environment respectively. This provides additional flexibility for using the model-based framework in complex environments [2].

To achieve the factorisation property in Eq. (2), it is crucial that each factor transform models exactly the impact of its associated acoustic factor, thus it will only change when the corresponding acoustic factor changes. In practice, factor transforms are normally maximum likelihood (ML) estimated on their corresponding data. For example, given a set of adaption utterances $\{\mathcal{O}^{(h)}\}$ which are produced by I speaker in J environment conditions, the speaker and environment transforms are obtained by

$$\mathcal{T}_{\mathtt{s}}^{(i)} = \arg\max_{\mathcal{T}_{\mathtt{s}}} \sum_{h:s_h=i} \mathcal{L}(\mathcal{O}^{(h)}; \mathcal{M}_c, \mathcal{T}_{\mathtt{s}}, \mathcal{T}_{\mathtt{n}}^{(n_h)}) \tag{3a}$$

$$\mathcal{T}_{\mathtt{n}}^{(j)} = \arg\max_{\mathcal{T}_{\mathtt{n}}} \sum_{h:n_h=j} \mathcal{L}(\mathcal{O}^{(h)}; \mathcal{M}_c, \mathcal{T}_{\mathtt{s}}^{(s_h)}, \mathcal{T}_{\mathtt{n}}) \tag{3b}$$

where $s_h \in \{1 \dots I\}$ and $n_h \in \{1 \dots J\}$ are the speaker and environment indices of utterance h respectively, $\mathcal{L}(\mathcal{O}^{(h)}; \mathcal{M}_{\mathtt{c}}, \mathcal{T}_{\mathtt{s}}, \mathcal{T}_{\mathtt{n}})$ is the log-likelihood function of the utterance h, given the canonical model, the speaker and the environment transforms. As it is shown in Eqs. (3), an optimal speaker transform is obtained by maximising a likelihood function with the optimal environment transforms as its parameters, thus it is a function of a set of optimal environment transforms, which breaks down the factorisation property in Eq. (2). This becomes more severe when the data distributed unevenly among acoustic factors. In an extreme, if there is only one noise source in a particular speaker's utterances, it is not possible to separate one factor from the other.

To mitigate this problem, it is necessary to keep the independence between factor transforms. This requires

$$\frac{\partial \mathcal{T}_{\mathtt{s}}^{(i)}}{\partial \mathcal{T}_{\mathtt{n}}^{(j)}} = \mathbf{0} \quad , \quad \frac{\partial \mathcal{T}_{\mathtt{n}}^{(j)}}{\partial \mathcal{T}_{\mathtt{s}}^{(i)}} = \mathbf{0} \qquad \forall i, j \tag{4}$$

Note $\{\mathcal{T}_{\mathtt{s}}^{(i)}\}$ and $\{\mathcal{T}_{\mathtt{n}}^{(j)}\}$ are the ML estimators given in Eqs. (3). The independence constraint in Eq. (4) enforces the optimal factor transform will not be affected by other factor transforms, allowing it changes only when the corresponding acoustic factor changes. The constraint in Eq. (4) implies that for any observed utterance $\mathcal{O}$, the optimal factor transforms $\mathcal{T}_{\mathtt{s}}^{*}$ and $\mathcal{T}_{\mathtt{n}}^{*}$ are independent of each other, where $\mathcal{T}_{\mathtt{s}}^{*} = \arg\max_{\mathcal{T}_{\mathtt{s}}} \mathcal{L}(\mathcal{O}; \mathcal{M}_{\mathtt{c}}, \mathcal{T}_{\mathtt{s}}, \mathcal{T}_{\mathtt{n}})$ and $\mathcal{T}_{\mathtt{n}}^{*} = \arg\max_{\mathcal{T}_{\mathtt{s}}} \mathcal{L}(\mathcal{O}; \mathcal{M}_{\mathtt{c}}, \mathcal{T}_{\mathtt{s}}, \mathcal{T}_{\mathtt{n}})$. This constraint can be formulated as a more useful one as the following proposition demonstrates:

Proposition 1. *Let $\boldsymbol{x}^*(\boldsymbol{y})$ be the local maximiser of function $f(\boldsymbol{x}, \boldsymbol{y})$ given $\boldsymbol{y}$. Assuming the function f has its second order derivatives $\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{y}}$ defined everywhere, the derivative of $\boldsymbol{x}^*$ with respect to $\boldsymbol{y}$ is*

$$\frac{\partial \boldsymbol{x}^*(\boldsymbol{y})}{\partial \boldsymbol{y}} = -\left(\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{x}}\right)^{-1} \frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{y}}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{y})} \tag{5}$$

Thus a sufficient condition for $\frac{\partial \boldsymbol{x}^*(\boldsymbol{y})}{\partial \boldsymbol{y}} = \mathbf{0}$ is :

$$\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{y}} = \mathbf{0} \quad \forall \boldsymbol{x}, \boldsymbol{y} \tag{6}$$

This condition also implies $\frac{\partial \boldsymbol{y}^*(\boldsymbol{x})}{\partial \boldsymbol{x}} = \mathbf{0}$.

Proof. Define another function $g(\boldsymbol{y}) = \frac{\partial f(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{y})}$. On one hand,

$$\frac{\partial g(\boldsymbol{y})}{\partial \boldsymbol{y}} = \left[\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{y}} + \frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right]_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{y})}$$

On the other hand, as $\boldsymbol{x}^*(\boldsymbol{y})$ is a local maximiser of $f(\cdot, \boldsymbol{y})$, thus $\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\mathsf{T}}|_{\boldsymbol{x}=\boldsymbol{x}*(\boldsymbol{y})} \succ \mathbf{0}$ and

$$g(\boldsymbol{y}) \triangleq \frac{\partial f(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{x}}|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{y})} = \mathbf{0}$$

Therefore, $\frac{\partial \boldsymbol{x}^*(\boldsymbol{y})}{\partial \boldsymbol{y}} = -\left(\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{x}}\right)^{-1} \frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{y}^\mathsf{T}}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{y})}$ and $\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{y}} = \mathbf{0}$ implies $\frac{\partial \boldsymbol{x}^*(\boldsymbol{y})}{\partial \boldsymbol{y}} = \mathbf{0}$. A similar argument can be made to establish $\frac{\partial^2 f}{\partial \boldsymbol{y} \partial \boldsymbol{x}} = \mathbf{0}$ implies $\frac{\partial \boldsymbol{y}^*(\boldsymbol{x})}{\partial \boldsymbol{x}} = \mathbf{0}$. □

Proposition 1 ensures if the second order derivatives of the log-likelihood function with respect to the speaker transform and noise transform is zero everywhere, the independence constraint in Eq. (2) is satisfied. The log-likelihood function is usually maximised via EM using the auxliary function in the following form:

$$\mathcal{Q} = \sum_{m,t} \gamma_t^{(m)} \log p(\boldsymbol{o}_t; \boldsymbol{\mu}_{\mathtt{c}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{c}}^{(m)}, \mathcal{T}_{\mathtt{s}}, \mathcal{T}_{\mathtt{n}}) \tag{7}$$

where $\boldsymbol{o}_t$ is the observation vector at time t, $\gamma_t^{(m)}$ is the posterior of $\boldsymbol{o}_t$ belonging to m-th component, $\boldsymbol{\mu}_{\mathrm{c}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)}$ are the canonical mean and variance of component m. Assuming $\gamma_t^{(m)}$ does not vary with respect to the change of $\mathcal{T}_{\mathrm{s}}$ and $\mathcal{T}_{\mathrm{n}}$, the following constraint is used to enforce the independence:

$$\frac{\partial^2 \mathcal{Q}}{\partial \mathcal{T}_{\mathrm{s}} \partial \mathcal{T}_{\mathrm{n}}} = \mathbf{0} \qquad (8)$$

3. Factorised Cluster Adaptive Training

In this work, the fCAT model proposed in [8] was used to evaluate the effectiveness of the proposed independence constraint for acoustic factorisation. fCAT is an extension of the standard cluster adaptive training (CAT) [15] or eigenvoice[16], which enables adapting the acoustic models to multiple factors. In fCAT, to compensate the variability of speaker and environment in h-th utterance, the m-th Gaussian component is adapted by:

$$p(\boldsymbol{o}_t^{(h)}; \boldsymbol{\mu}_{\mathrm{c}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)}, m, s_h, n_h) = \mathcal{N}(\boldsymbol{o}_t^{(h)}; \boldsymbol{\mu}^{(mh)}, \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)}) \quad (9)$$

where $\boldsymbol{o}_t^{(h)} \in \mathcal{R}^D$ is the t-th observation vector in h-th utterance, $\boldsymbol{\mu}_{\mathrm{c}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)}$ are the canonical mean and variance, and

$$\boldsymbol{\mu}^{(mh)} = \boldsymbol{\mu}_{\mathrm{c}}^{(m)} + \mathbf{M}_{\mathrm{s}}^{(m)} \boldsymbol{\lambda}_{\mathrm{s}}^{(s_h, q_m)} + \mathbf{M}_{\mathrm{n}}^{(m)} \boldsymbol{\lambda}_{\mathrm{n}}^{(n_h, r_m)}, \quad (10)$$

$\mathbf{M}_{\mathrm{s}}^{(m)} \in \mathcal{R}^{D \times d_{\mathrm{s}}}$, $\mathbf{M}_{\mathrm{n}}^{(m)} \in \mathcal{R}^{D \times d_{\mathrm{n}}}$ are the component m's speaker and environment cluster parameters, d_{s} and d_{n} are the number of speaker and environment cluster respectively; $q_m \in \{1, \dots, Q\}$ ($r_m \in \{1, \dots, R\}$) maps the component index m to the speaker (environment) regression class index; $\boldsymbol{\lambda}_{\mathrm{s}}^{(i,q)}$ is the speaker cluster weight vector associated with q-th speaker base classs for i-th speaker; $\boldsymbol{\lambda}_{\mathrm{n}}^{(j,r)}$ is the environment cluster weight vector associated with r-th environment base class for j-th environment. Among these parameters, $\{\boldsymbol{\mu}_{\mathrm{c}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)}, \mathbf{M}_{\mathrm{s}}^{(m)}, \mathbf{M}_{\mathrm{n}}^{(m)}\}$ form the canonical model parameters $\mathcal{M}_{\mathrm{c}}$, while $\{\boldsymbol{\lambda}_{\mathrm{s}}^{(i,q)} | q = 1 \dots Q\}$ and $\{\boldsymbol{\lambda}_{\mathrm{n}}^{(j,r)} | r = 1 \dots R\}$ are the parameters of i-th speaker transform $\mathcal{T}_{\mathrm{s}}^{(i)}$ and j-th environment transform $\mathcal{T}_{\mathrm{n}}^{(j)}$ respectively. Note this modelling form is similar to the one used in JFA[12], whilst there is an additional diagonal residual term in JFA.

For the auxiliary function $\mathcal{Q}$ in EM, it is easy to show that

$$\frac{\partial^2 \mathcal{Q}(\mathcal{O}; \mathcal{T}_{\mathrm{s}}, \mathcal{T}_{\mathrm{n}})}{\partial \mathcal{T}_{\mathrm{s}} \partial \mathcal{T}_{\mathrm{n}}} = \sum_{t,m} \gamma_t^{(m)} \mathbf{M}_{\mathrm{s}}^{(m)\mathsf{T}} \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)-1} \mathbf{M}_{\mathrm{n}}^{(m)}. \quad (11)$$

Hence a sufficient condition for the independence constraint in Eq. (8) to hold for every possible observation sequence is

$$\mathbf{M}_{\mathrm{s}}^{(m)\mathsf{T}} \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)-1} \mathbf{M}_{\mathrm{n}}^{(m)} = \mathbf{0} \quad \forall m. \quad (12)$$

According to the above constraint, in a normed vector space induced by the inner product function $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}$, the speaker subspace, expanded by $\mathbf{M}_{\mathrm{s}}$, is orthogonal to the environment subspace which is expanded by $\mathbf{M}_{\mathrm{n}}$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\cdots \boldsymbol{\Sigma}_{\mathrm{c}}^{(m)} \cdots)$, $\mathbf{M}_{\mathrm{s}}$ and $\mathbf{M}_{\mathrm{n}}$ are obtained by stacking $\mathbf{M}_{\mathrm{s}}^{(m)}$ and $\mathbf{M}_{\mathrm{n}}^{(m)}$ respectively. This orthogonality guarantees that for a given data point, there is a unique speaker-environment factorisation, thus it is possible to separate speaker factor even if there is only one data point.

fCAT model is trained to maximise the likelihood of training data which consists of various speaker and environment combinations. The parameter updates are the same as those in [8]. The main difference is the constraint in Eq. (11) need to be maintained, thus updating the m-th speaker cluster $\mathbf{M}_{\mathrm{s}}^{(m)}$ requires solving the following constrained optimisation[1]:

$$\begin{aligned} &\max_{\mathbf{M}_{\mathrm{s}}} -\frac{1}{2} \mathrm{tr}\left(\mathbf{M}_{\mathrm{s}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathrm{c}}^{-1} \mathbf{M}_{\mathrm{s}} \mathbf{G}_{\mathrm{s}}\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathrm{c}}^{-1} \mathbf{M}_{\mathrm{s}} \mathbf{K}_{\mathrm{s}}\right) \\ &\text{s. t.} \quad \mathbf{M}_{\mathrm{s}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathrm{c}}^{-1} \mathbf{M}_{\mathrm{n}} = \mathbf{0} \end{aligned} \quad (13)$$

where

$$\mathbf{G}_{\mathrm{s}} = \sum_{t,h} \gamma_t^{(mh)} \boldsymbol{\lambda}_{\mathrm{s}}^{(s_h, q_m)} \boldsymbol{\lambda}_{\mathrm{s}}^{(s_h, q_m)\mathsf{T}}$$

$$\mathbf{K}_{\mathrm{s}} = \sum_{t,h} \gamma_t^{(mh)} \boldsymbol{\lambda}_{\mathrm{s}}^{(s_h, q_m)} (\boldsymbol{o}_t^{(h)} - \boldsymbol{\mu}_c^{(m)} - \mathbf{M}_{\mathrm{n}}^{(m)} \boldsymbol{\lambda}_{\mathrm{n}}^{(n_h, r_m)})$$

and $\gamma_t^{(mh)}$ is the posterior probability of $\boldsymbol{o}_t^{(h)}$ in component m. Using the method of Lagrange multipliers [17], it can be shown that the solution is given by:

$$\mathbf{M} = \left[\mathbf{I} - \mathbf{M}_{\mathrm{n}} \left(\mathbf{M}_{\mathrm{n}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathrm{c}}^{-1} \mathbf{M}_{\mathrm{n}}\right)^{-1} \mathbf{M}_{\mathrm{n}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathrm{c}}^{-1}\right] \mathbf{K}_{\mathrm{s}} \mathbf{G}_{\mathrm{s}}^{-1} \quad (14)$$

Note $\mathbf{M} = \mathbf{K}_{\mathrm{s}} \mathbf{G}_{\mathrm{s}}^{-1}$ if the constraint is removed. Similar equation is adopted for re-estimating for the environment cluster parameter $\mathbf{M}_{\mathrm{n}}^{(m)}$.

There are three main stages involved to train a fCAT model. In the first stage, speaker and environment transforms were initialised. In this work, the eigen-decomposition[16] approach was used for speaker transforms initialisation, while the environment transforms were initialised as one-hot vectors, with the corresponding environment weighted as 1. In the second, standard CAT training stage, the speaker transforms, the speaker cluster parameters and the canonical mean/variance were iteratively updated, while in the third training stage, 5 sets of parameters: the speaker transforms, speaker cluster parameters, environment transforms, environment cluster parameters and the canonical mean/variances were re-estimated iteratively. The overall fCAT training procedure, starting from a well-trained CAT model, is summarised in the following:

1 Initialise the environment transforms by setting the current environment weight as 1, all the other weights as 0; initialise the speaker transforms using the transforms obtained during standard CAT training.

2 Given the current speaker clusters, and the speaker/environment transforms the environment clusters are estimated to maximise the log-likelihood function while maintaining the independence constraint in Eq. (12).

3 The environment transforms are updated while keeping all the other parameters fixed.

4 The speaker clusters are updated while keeping all the other parameter fixed. Again, the independence constraint in Eq. (12) needs to be maintained.

5 Update speaker transforms while keeping all the other parameters fixed.

6 Goto step 2 $N_1(\sim 2)$ times.

7 The canonical mean and variances are updated given the current speaker/environment transforms and clusters.

8 Goto step 2 $N_2(\sim 5)$ times.

4. Experiments

The effectiveness of the proposed explicit independence constraint for factorised adaptation was evaluated on a noise corrupted wall street journal (WSJ) task. Both WSJ0 and WJS1

[1] For notation simplicity, the index m in superscripts is omitted in Eq. (13) and Eq. (14).

training data were used. There are in total 36,515 utterances in the training set, produced by 284 speakers. To simulated the background noise, 6 types of noise (similar as those used in AURORA4 task) were added to the SI-284 training set to form a multi-conditional training set with 7 environment conditions (including clean condition). As an initial investigation, the training set was created in a speaker-environment balanced manner, i.e., utterances from each speaker were exposed to 7 environment conditions with equal probability. The SNRs in training data ranged from 20dB to 10dB. 3 evaluation sets were defined to simulate different scenarios in which factorised adaptation can be used. The first one, based on the original WSJ1 development set (wsj1_dt), had 10 speakers, each comprising 40 adaptation read utterances and roughly 50 test utterances. These utterances were distorted in the same way as training utterances, except only 6 noisy environment conditions were created with the SNR ranging from 15dB to 5dB. The second one, based on the WSJ0 development set (wsj0_dt), simulated a more realistic operation scenario, where a majority (75%) of adaption utterances (400 in total) were distorted by the same noise source ("restaurant noise"), while the 410 test utterances were distorted by the 6 noise sources with a uniform distribution. The third set, based on the WJS0 evaluation set (wsj0_et) simulated the *practical enrollment* scenario: all the 321 adaptation utterances were distorted by a single noise source ("restaurant noise"), while the environment conditions of the 330 test utterances were uniformly distributed. The SNRs for the adaptation and test utterances in the three evaluation sets were ranging from 15dB to 5dB. Adaptation for the first two sets (wsj1_dt and wsj0_dt) ran in a unsupervised adaptation mode, while wsj0_et set ran in a supervised mode. It was assumed all the speaker and environment identities were known in advance.

A 39 dimensional front-end feature vector was used for the experiments, consisting of 12 MFCCs appended with the zeroth cepstrum, delta and delta-delta coefficients. Cepstral mean normalisation was performed for each utterance. A cross-word triphone acoustic model with 3955 tied states, 16 components per state was trained on the multi-conditional training set. This is known as the multi-styled trained model (MST). Initialised by this MST model, a standard CAT model with 8 clusters were also trained, where a 32-node regression tree was used for CAT adaptation on the speaker level. On top of this CAT model, fCAT model was trained to take the environment variability into account, in which the environment space is consisted of 7 clusters. A 256-node environment regression tree was used. As a contrast, another fCAT model was also trained without the implicit independence constraint. In decoding, A bi-gram language model was used throughout the experiments.

Unadapted decoding was first performed using the MST model on the test data. The first row of table 1 shows the performance on three test sets. The MST model was also used in decoding the adaptation utterances in set wsj0_dt and wsj1_dt to provide the supervision hypothesis for the following adaptation run. For supervised adaptation set wsj0_et, the MST model was used to generate the phone alignment against the reference hypothesis. To run adapted decoding using CAT, speaker transforms were first estimated on the adaptation utterances, and then applied on the test sets. The second row of table 1 shows the performance of adapted decoding using the CAT model. As expected, speaker adaptation improved the performance. To perform factorised adaptation, speaker and environment transforms were iteratively estimated on the adaptation data. In the very first iteration, environment weight vectors were set as the one-of-N vector according to the known environment type, and the speaker transforms were estimated while the component posterior was calculated using the MST model. Given the set of speaker transforms, a set of environment transforms were estimated for each environment condition appeared in the adaptation data. A few(~ 5) iterations [2] were performed. Then the speaker transform estimated from adaptation data in conjunction with the environment transforms obtained during training were used in decoding. This enables the comparison of speaker transforms estimated by different fCAT systems.

System	wsj1_dt	wsj0_dt	wsj0_et
MST	23.3	15.7	14.9
CAT	20.3	13.3	13.0
fCAT(w/o constr.)	19.5	13.6	14.4
fCAT	19.7	12.9	12.3

Table 1: Performances (in WER%) of MST,CAT and fCAT systems.

fCAT systems trained with and without the independence constraint are compared in the second block of table 1. On the first set, wsj1_dt, which is speaker-environment balanced, both systems achieved gains over the speaker-adapted CAT system (19.5% and 19.7% vs 20.3%). This illustrates the benefit of adapting acoustic models to both speaker and environment. However, when used in more realistic scenarios where the speaker-environment distribution of adaptation data is imbalanced, i.e., on the wsj0_dt and wsj1_et sets, factorised adaptation using fCAT model trained without independence constraint degraded performance, compared with the performance of speaker adaptation using CAT (13.6% vs 13.3% and 14.4% vs 13.0%). This demonstrates that the speaker transform which was estimated on imbalanced data modeled not only the speaker but also the environment variability. In contrast, factorised adaptation using fCAT trained with the independence constraint achieved gains over speaker adaptation on both sets (12.9% vs 13.3% and 12.3% vs 13.0%). The gain of this fCAT system on wsj0_et set is more interesting, as there is only one environment condition in the adaptation data. Other adaptation schemes using linear transforms on this set can only learn the combination effect of speaker and noise factor, not able to distinguish one factor from the other. With fCAT, as the subspace expanded by the speaker cluster is orthogonal to the subspace expanded by the environment cluster, speaker transforms learned on adaptation data can only explain the speaker distortion, which allows it can be factorised out.

5. Conclusion

This paper investigated an independent constraint for factorised adaptation. This constraint enforced the factor transforms to be independent of each other, which is the crucial condition for factorisation. Unlike previous work, which mostly relies on data balance to *implicitly* achieve the independence, this work derived an *explicit* constraint. Using fCAT as an application, experimental results demonstrated that with this explicit independence constraint, it is possible to perform factorised adaptation on highly imbalanced data. Future work will examine the application of this constraint to other models, such as JFA.

[2] It is possible to only estimate the speaker transform while borrowing the environment transforms obtained during training as in [4]. In this work, iterative estimation of both transforms was used to investigate what a fCAT model can learn on each factor.

6. References

[1] M. J. F. Gales, "Acoustic factorisation," in *Proc. ASRU-2001*.

[2] Y.-Q. Wang and M. J. F. Gales, "Speaker and noise factorisation for robust speech recognition," *IEEE transactions on audio, speech and language processing*, vol. 20, no. 7, 2012.

[3] Z. Ou and K. Deng, "Combining eigenvoice speaker modeling and VTS-based environment compensation for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4673–4676.

[4] M. L. Seltzer and A. Acero, "Separating speaker and environmental variability using factored transforms," in *Proc. Interspeech*, 2011, pp. 1097–1100.

[5] M. L. Seltzer and A. Acero, "Factored adaptation for separable compensation of speaker and environmental variability," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 146–151.

[6] M. L. Seltzer and A. Acero, "Factored adaptation using a combination of feature-space and model-space transforms," in *Proc. Interspeech*, 2012.

[7] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[8] J. Latorre, V. Wan, M. J. F. Gales, L.-Z. Chen, K.-K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," in *Proc. Interspeech*, 2012.

[9] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP-2000*.

[10] C. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, pp. 171–186, 1995.

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

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on audio, speech and language processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on audio, speech and language processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] M. J. F. Gales, "Model-based approaches to handling uncertainty," in *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, pp. 101–125. Springer, 2011.

[15] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2002.

[16] R. Kuhn, P. Nguyen, J. C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP-1998*.

[17] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.