Speaker and Noise Factorisation for Robust Speech Recognition

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

***Abstract*—Speech recognition systems need to operate in a wide range of conditions. Thus they should be robust to extrinsic variability caused by various acoustic factors, for example speaker differences, transmission channel and background noise. For many scenarios, multiple factors simultaneously impact the underlying "clean" speech signal. This paper examines techniques to handle both speaker and background noise differences. An acoustic factorisation approach is adopted. Here separate transforms are assigned to represent the speaker (maximum likelihood linear regression (MLLR)), and noise and channel (model-based vector Taylor series (VTS)) factors. This is a highly flexible framework compared to the standard approaches of modelling the combined impact of both speaker and noise factors. For example factorisation allows the speaker characteristics obtained in one noise condition to be applied to a different environment. To obtain this factorisation modified versions of MLLR and VTS training and application are derived. The proposed scheme is evaluated for both adaptation and factorisation on the AURORA4 data.**

I. Introduction

To be applicable to many real-life scenarios, speech recognition systems need to be robust to the *extrinsic* variabilities in the speech signal, such as speaker differences, transmission channel and background noise. There has been a large amount of research into dealing with individual factors such as speaker [2] or noise [3]. Schemes developed to adapt the speech recognisers to specific speakers are often known as *speaker adaptation*, while schemes designed to handle the impact of environment are referred to as *environmental robustness*. It is possible to combine the above techniques to adapt the speech recogniser to the target speaker and environment. Normally, this is done via feature enhancement or model compensation to remove the effect of noise, followed by speaker adaptation. However, these approaches typically model the two distinct acoustic factors as a combined effect. Thus in the standard schemes there is no distinction between the transforms representing the speaker characteristics and the noise characteristics. The transforms are simply estimated sequentially, with the, typically linear, speaker transforms modelling all the residual effects that are not modelled by the noise transforms. This paper proposes a new adaptation scheme, where the impacts of speaker and noise differences are modelled separately. The proposed scheme sits in a fully model-based framework, which allows two different model transforms, i.e., a model-based VTS transform and an MLLR mean transform, to be estimated in a *factorised* fashion and applied independently. This allows, for example, the speaker characteristics obtained in one noise condition to be applied to a different environment. This is important for some applications, where the speaker characteristics are known to be relatively constant while the background environment changes.

A variety of schemes have been proposed for speaker adaptation, e.g., [4], [5], [6], [7], [8], [9]. For adaptation with limited data, linear transform-based schemes are the most popular choices. In these schemes, a set of linear transforms, e.g., MLLR [5], [6] and constrained MLLR (CMLLR) [6], are used to adapt the mean and/or covariances of Gaussian components in the acoustic models, such that the target speaker can be better modelled. These *adaptive* techniques modify the acoustic models to better match the adaptation data, and do not rely on an explicit model of speaker differences. Hence they can be also used for the purpose of general adaptation, e.g., environmental adaptation [10], [11]. Furthermore, to train acoustic models on *found* data which is inhomogeneous in nature, adaptive training [12] has been proposed, where "neutral" acoustic models are estimated on multi-style data and the differences among speakers are "absorbed" by speaker transforms. This adaptive training framework has also been extended to train neutral acoustic models on data from different environment, e.g., [13], [14].

Approaches for handling the effect of background and convolutional noise can be broadly split into two categories. In the first, *feature compensation*, category, schemes attempt to denoise (or clean) the noise corrupted feature vectors. These enhanced feature vectors are then treated as clean speech observations. Schemes fitting into this category include ETSI advanced front-end (AFE) [15], SPLICE [16], model-based feature enhancement (MBFE) [17], and feature-space Vector Taylor Series (VTS) [18]. In the second, *model compensation*, category, the back-end acoustic models are compensated to reflect the noisy environment. Normally, the impact of channel and background noise is expressed as a mismatch function relating the clean speech, noise and noisy speech. Using a mismatch function as an explicit distortion model will be referred to as *predictive* approaches. Examples of predictive approaches include Parallel Model Combination (PMC) [19], model-space VTS [20], [21], joint uncertainty decoding (JUD) [13] and joint compensation of additive and convolutive distortions (JAC)[22]. Both feature compensation and model-based approaches achieve good acoustic model robustness.

The authors are with the Engineering Department, Cambridge University, Cambridge CB2 1PZ, U.K. (e-mail: yw293@cam.ac.uk).

This work was partially supported by Google research award and DARPA under the GALE program. The authors would also like to thank Dr. F. Flego for making VTS code available. This work is an extension of its conference version presented in [1].

Model-based approaches are more powerful than standard enhancement schemes, as they allow a detailed representation of the additional uncertainty caused by background noise. Recently, adaptive training framework have been successfully extended to handle variations in the training data environment, e.g.,[13], [14], [23]. Here noise-specific transforms are estimated for each environmental homogeneous block of data, allowing "clean" acoustic models to be estimated from multi-style data that are corrupted by different noises. Experimental results demonstrated that adaptively trained acoustic models are more amenable to be adapted to the target acoustic conditions.

Speaker adaptation can be combined with environmental robustness to adapt the speech recogniser to both speaker and environment factors. There are generally two approaches in the literature for joint speaker and environment adaptation. The first one is to use feature enhancement techniques to denoise the observation before back-end model adaptation, e.g., [24]. The other approach, discussed in [25], is a fully model-based approach: acoustic models are first compensated for the effect of noise, then linear transform-based adaptation can be performed to reduce the residual mismatch, including the one caused by speaker differences. Little work has been done to separate the speaker and environmental differences. Two notable works are [26] and [27]. In [26], component-specific biases based on Jacobian compensation with speaker-dependent Jacobians were used to clean the observation prior to the speaker adaptation and only the mean vectors are compensated for the effect of noise. This work will also use speaker-dependent Jacobians, but in a full model-based framework. The proposed scheme is based on the concept of "acoustic factorisation" in [27], and uses the structured transform in [28]. In acoustic factorisation, transforms are constructed in such a way that each transform is related to only one acoustic factor. Note that in [28], though multiple transforms are used, they are not constrained to be related with one specific acoustic factor. Ideally, different sets of transforms should be "orthogonal", i.e., the impact of each set of transforms should be able to be applied independently. This will yield a highly flexible framework for using the transforms. To achieve this orthogonality, the transforms need to be different in nature to each other. In this work, a model-based VTS transform [20] is associated with each utterance, while a block-diagonal MLLR mean transform [5], [6] is used for each speaker who may have multiple recordings. The amount of data required to estimate an MLLR transform is far greater than that required for a VTS transform: VTS transform can be robustly estimated on a single utterance, while MLLR transform requires multiple utterances. Thus when estimating the speaker transform, the system must be able to handle changing background noise conditions. As these two transforms are different in nature, and are estimated on different adaptation data, it is now possible to decouple them, thus achieve the factorisation.

This paper is organised as follows. The next section introduces the general concept of acoustic factorisation. Speaker and noise compensation schemes and the ways to combine them are discussed in section III. Estimation of transform parameters is presented in section IV. Experiments and results are presented and discussed in section V with conclusions in section VI.

II. Acoustic Factorisation

Model-based approaches to robust (in the general sense) speech recognition have been intensively studied and extended in the last decade. In this framework, intrinsic and extrinsic variability are represented by a canonical model $\mathcal{M}_c$ and a set of transforms $\mathcal{T}$, respectively. Consider a complex acoustic environment, in which there are two acoustic factors, s and n, simultaneously affecting the speech signal. The canonical model is adapted to represent this condition by the transform $\mathcal{T}^{(sn)}$:

$$\mathcal{M}^{(sn)} = \mathcal{F}(\mathcal{M}_c, \mathcal{T}^{(sn)}) \tag{1}$$

where $\mathcal{M}^{(sn)}$ is the adapted acoustic model for condition (s, n), $\mathcal{T}^{(sn)}$ the transform for that condition, and $\mathcal{F}$ is the mapping function. The transform is normally estimated using the ML criterion:

$$\mathcal{T}^{(sn)} = \arg\max_{\mathcal{T}} \left\{ p(\mathcal{O}^{(sn)} | \mathcal{M}_c, \mathcal{T}) \right\} \tag{2}$$

where $\mathcal{O}^{(sn)}$ is a sequence of feature vectors observed in the acoustic condition (s, n). It is possible to combine different forms of transforms to obtain the final transformation, $\mathcal{T}^{(sn)}$. However the amount of data required to estimate the parameter of final transformation is determined by the need to robustly estimate parameters of all transforms. Thus in the case considered in this work, combining MLLR and VTS, sufficient data in the target condition (s, n) is required to estimate the MLLR transform, as the VTS transform can be rapidly estimated on far less data. When the transforms are estimated to model the combined condition it will be referred to as *batch-mode* adaptation in this paper.

To more effectively deal with complex acoustic environments, the concept of acoustic factorisation was proposed in [27], where each of the transforms is constrained to be related to an individual acoustic factor. In the above example, this requires that the transform $\mathcal{T}^{(sn)}$ can be factorised as:

$$\mathcal{T}^{(sn)} = \mathcal{T}^{(s)} \otimes \mathcal{T}^{(n)} \tag{3}$$

where $\mathcal{T}^{(s)}$ and $\mathcal{T}^{(n)}$ are the transforms associated with acoustic factors s and n, respectively.

The factorisation attribute in Eq. (3) offers additional flexibility for the models to be used in a complex and rapid changing acoustic environment. This can be demonstrated by considering a speaker (s) in a range of different noise (n) conditions. For r acoustic conditions, $(s, n_1), \cdots, (s, n_r)$, it is necessary to estimate a set of transforms $\mathcal{T}^{(sn_1)}, \ldots, \mathcal{T}^{(sn_r)}$, using the data, $\mathcal{O}^{(sn_1)}, \ldots, \mathcal{O}^{(sn_r)}$, from each of these conditions. Using factorisation only a single speaker transform, $\mathcal{T}^{(s)}$, and a set of noise transforms $\mathcal{T}^{(n_1)}, \ldots, \mathcal{T}^{(n_r)}$ are required. As noise transforms can be robustly estimated from a single utterance, it is only necessary to have sufficient data of a specific speaker over all conditions to estimate the speaker transform.

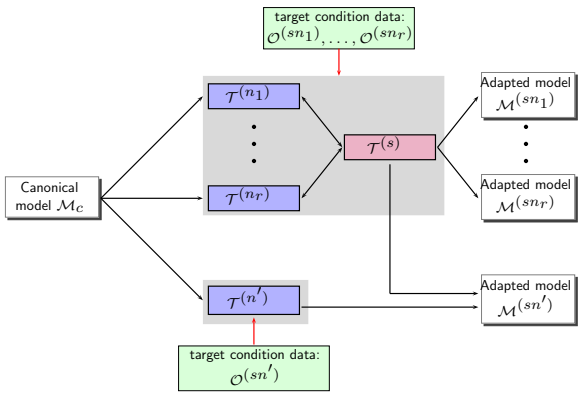

Fig. 1. Speaker and noise adaptation in the factorisation mode.

Furthermore, for a new condition (s, n') it is only necessary to estimate the noise transform $\mathcal{T}^{(n')}$ and combine this transform with the existing speaker transform. This form of combination relies on the "orthogonality" of the transforms: the speaker transform only models speaker attributes and the noise transform the noise attributes. Figure 1 shows the concept of acoustic factorisation for speaker and noise factors.

The following procedure illustrates how the speaker and noise adaptation can be performed in this factorisation framework. Note that the canonical model, $\mathcal{M}_c$, is assumed to have been trained.

1) Initialise the speaker transform to an identity transform, i.e., $\mathcal{T}^{(s)} = [\mathbf{I}, \mathbf{0}]$, and obtain initial estimates (for example using voice activity detection) for the noise transforms.
2) Estimate the noise transform for each condition as
$$\mathcal{T}^{(n_i)} = \arg\max_{\mathcal{T}} \left\{ p(\mathcal{O}^{(sn_i)} | \mathcal{M}_c, \mathcal{T}^{(s)} \otimes \mathcal{T}) \right\} \quad (4)$$
3) Estimate the speaker transform $\mathcal{T}^{(s)}$ using
$$\mathcal{T}^{(s)} = \arg\max_{\mathcal{T}} \left\{ \prod_{i=1}^{r} p(\mathcal{O}^{(sn_i)} | \mathcal{M}_c, \mathcal{T} \otimes \mathcal{T}^{(n_i)}) \right\} \quad (5)$$
4) Goto (2) until converged.

Having obtained the speaker and noise transforms for the training data, the transform for a new acoustic condition (s, n'), can be obtained simply by estimating the noise transform
$$\mathcal{T}^{(n')} = \arg\max_{\mathcal{T}} \left\{ p(\mathcal{O}^{(sn')} | \mathcal{M}_c, \mathcal{T}^{(s)} \otimes \mathcal{T}) \right\} \quad (6)$$
Given the speaker transform and the noise transforms, the acoustic model is adapted to the test condition using the transform $\mathcal{T}^{(sn')} = \mathcal{T}^{(s)} \otimes \mathcal{T}^{(n')}$.

III. Speaker and Noise Compensation

To achieve acoustic factorisation, the speaker transform $\mathcal{T}^{(s)}$ and noise transform $\mathcal{T}^{(n)}$ must have different forms to yield a degree of orthogonality. In this work, a linear transform, MLLR, and a nonlinear one, model-based VTS compensation, are used for speaker and noise adaptation respectively. This section describes the forms of VTS compensation and the options for combining it with MLLR-based speaker adaptation.

Additive and convolutional noise corrupt "clean" speech, resulting in the noisy, observed, speech. In the Mel-cepstral domain, the *mismatch function* relating the clean speech static $\boldsymbol{x}$ and the noisy speech static $\boldsymbol{y}$ is given by:
$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x} + \boldsymbol{h} + \mathbf{C} \log\left(\mathbf{1} + \exp\left(\mathbf{C}^{-1}(\boldsymbol{n} - \boldsymbol{x} - \boldsymbol{h})\right)\right) \\ &= \mathbf{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}), \end{aligned} \quad (7)$$
where $\boldsymbol{n}$ and $\boldsymbol{h}$ are the additive and convolutional noise, respectively, and $\mathbf{C}$ is the DCT matrix. It is assumed that for the u-th noise condition or utterance: $\boldsymbol{n}$ is Gaussian distributed with mean $\boldsymbol{\mu}_{\mathtt{n}}^{(u)}$ and diagonal covariance $\boldsymbol{\Sigma}_{\mathtt{n}}^{(u)}$; $\boldsymbol{h} = \boldsymbol{\mu}_{\mathtt{h}}^{(u)}$ is an unknown constant . Model-based VTS compensation [20], [21] approximates the mismatch function by a first-order vector Taylor series, expanded at the speech and noise mean, $\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}$, for each component m. Under this approximation,
$$p(\boldsymbol{y}|m, u) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{\mathtt{vts,y}}^{(mu)}, \boldsymbol{\Sigma}_{\mathtt{vts,y}}^{(mu)}) \quad (8)$$
where the compensated mean $\boldsymbol{\mu}_{\mathtt{vts,y}}^{(mu)}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathtt{vts,y}}^{(mu)}$ are given by:
$$\begin{aligned} \boldsymbol{\mu}_{\mathtt{vts,y}}^{(mu)} &= \mathbf{f}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}), \\ \boldsymbol{\Sigma}_{\mathtt{vts,y}}^{(mu)} &= \mathrm{diag}\left(\mathbf{J}_{\mathtt{x}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)} \mathbf{J}_{\mathtt{x}}^{(mu)\mathsf{T}} + \mathbf{J}_{\mathtt{n}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{n}}^{(u)} \mathbf{J}_{\mathtt{n}}^{(mu)\mathsf{T}}\right) \end{aligned} \quad (9)$$
and $\boldsymbol{\mu}_{\mathtt{x}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}$ are the mean and covariance of component m, $\mathbf{J}_{\mathtt{x}}^{(mu)}$ and $\mathbf{J}_{\mathtt{n}}^{(mu)}$ are the derivatives of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$ and $\boldsymbol{n}$ respectively, evaluated at $\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}$. With the continuous time approximation [29], the delta parameters under VTS compensation scheme are compensated by:
$$\begin{aligned} \boldsymbol{\mu}_{\mathtt{vts,\Delta y}}^{(mu)} &= \mathbf{J}_{\mathtt{x}}^{(mu)} \boldsymbol{\mu}_{\mathtt{\Delta x}}^{(m)}, \\ \boldsymbol{\Sigma}_{\mathtt{vts,\Delta y}}^{(mu)} &= \mathrm{diag}\left(\mathbf{J}_{\mathtt{x}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{\Delta x}}^{(m)} \mathbf{J}_{\mathtt{x}}^{(mu)\mathsf{T}} + \mathbf{J}_{\mathtt{n}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{\Delta n}}^{(u)} \mathbf{J}_{\mathtt{n}}^{(mu)\mathsf{T}}\right) \end{aligned} \quad (10)$$
where $\boldsymbol{\mu}_{\mathtt{\Delta x}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathtt{\Delta x}}^{(m)}$ are the mean and covariance matrix of clean delta parameters. The delta-delta parameters are compensated in a similar way. For notational convenience, only the delta parameters will be considered in the following.

To adapt the speaker independent model to the target speaker s, the MLLR mean transform [5] in the following form is often used:
$$\boldsymbol{\mu}^{(sm)} = \mathbf{A}^{(s)} \boldsymbol{\mu}^{(m)} + \boldsymbol{b}^{(s)}, \quad (11)$$
where $[\mathbf{A}^{(s)}, \boldsymbol{b}^{(s)}]$ is the linear transform for speaker s, $\boldsymbol{\mu}^{(m)}$ and $\boldsymbol{\mu}^{(sm)}$ the speaker independent and speaker dependent mean for the component m respectively.

A. "VTS-MLLR" scheme

The simplest approach to combining VTS with MLLR to yield a speaker and noise adapted model is to take the VTS compensated models and apply MLLR afterwards. Considering block-diagonal transforms[1], the following transform to the

[1]It is possible to use full-transforms, however in this work to be consistent with the factorisation approach only block-diagonal transforms are considered.

speaker and noise condition is obtained.

$$\begin{aligned}\boldsymbol{\mu}_{\mathtt{y}}^{(smu)} &= \mathbf{A}^{(s)}\mathbf{f}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}) + \boldsymbol{b}^{(s)} \\ \boldsymbol{\mu}_{\Delta\mathtt{y}}^{(smu)} &= \mathbf{A}_{\Delta}^{(s)}\mathbf{J}_{\mathtt{x}}^{(mu)}\boldsymbol{\mu}_{\Delta\mathtt{x}}^{(m)} + \boldsymbol{b}_{\Delta}^{(s)}\end{aligned} \tag{12}$$

and

$$\boldsymbol{\Sigma}_{\mathtt{y}}^{(smu)} = \boldsymbol{\Sigma}_{\mathtt{vts,y}}^{(mu)}, \qquad \boldsymbol{\Sigma}_{\Delta\mathtt{y}}^{(smu)} = \boldsymbol{\Sigma}_{\mathtt{vts},\Delta\mathtt{y}}^{(mu)} \tag{13}$$

where $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)}, \boldsymbol{b}^{(s)}]$ and $\mathbf{W}_{\Delta}^{(s)} = [\mathbf{A}_{\Delta}^{(s)}, \boldsymbol{b}_{\Delta}^{(s)}]$ are the speaker s's linear transform for the static and delta features, respectively. The combined MLLR transform will be written as $\mathbf{K}^{(s)} = (\mathbf{W}^{(s)}, \mathbf{W}_{\Delta}^{(s)})$ This scheme will be referred to as "VTS-MLLR".

B. "Joint" scheme

In "VTS-MLLR", the speaker linear transform is applied on top of the noise-compensate models. This means that it will represent attributes of both speaker and noise factors, as the VTS compensated model will depend on the noise condition. Thus the "VTS-MLLR" scheme may not have the required factorisation attribute, i.e., the linear transform in "VTS-MLLR" does not solely represent the speaker characteristics. To address this problem a modified scheme, named as "Joint", is proposed where the speaker transform is applied to the underlying "clean" speech model prior to the application of VTS. The speaker transform should therefore not depend on the nature of the noise.

As the speaker adaptation in the "Joint" is applied to the clean speech models, this adaptation stage can be expressed for speaker s as

$$\boldsymbol{\mu}_{\mathtt{x}}^{(sm)} = \mathbf{A}^{(s)}\boldsymbol{\mu}_{\mathtt{x}}^{(m)} + \boldsymbol{b}^{(s)}, \quad \boldsymbol{\Sigma}_{\mathtt{x}}^{(sm)} = \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}, \tag{14}$$

where $\boldsymbol{\mu}_{\mathtt{x}}^{(sm)}$ and $\boldsymbol{\Sigma}_{\mathtt{x}}^{(sm)}$ are the compensated clean speech distribution parameters for component m of speaker s.

For standard VTS compensation scheme above, the compensation and Jacobian are based on the speaker independent distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)})$. For the "Joint" scheme these terms need to be based on the speaker compensated distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathtt{x}}^{(sm)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(sm)})$. Substituting the speaker dependent mean $\mathbf{W}\boldsymbol{\xi}_{\mathtt{x}}^{(m)}$ (for clarity of notation, the speaker index s will be dropped if there is no confusion) into Eq. (9) yields a new, "Joint", compensation scheme:

$$\begin{aligned}\boldsymbol{\mu}_{\mathtt{y}}^{(mu)} &= \mathbf{f}(\mathbf{W}\boldsymbol{\xi}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}), \\ \boldsymbol{\Sigma}_{\mathtt{y}}^{(mu)} &= \mathrm{diag}\left(\mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)\mathsf{T}} + \mathbf{J}_{\mathtt{n},\mathbf{w}}^{(mu)}\boldsymbol{\Sigma}_{\mathtt{n}}^{(u)}\mathbf{J}_{\mathtt{n},\mathbf{w}}^{(mu)\mathsf{T}}\right)\end{aligned} \tag{15}$$

where $\boldsymbol{\xi}_{\mathtt{x}}^{(m)} = [\boldsymbol{\mu}_{\mathtt{x}}^{(m)\mathsf{T}}, 1]^{\mathsf{T}}$, and

$$\mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|_{\mathbf{W}\boldsymbol{\xi}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}}, \quad \mathbf{J}_{\mathtt{n},\mathbf{w}}^{(mu)} = \mathbf{I} - \mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)}. \tag{16}$$

In this work, the MLLR mean transform is constrained to have a block diagonal structure, where the blocks corresponding to the static and delta parameters. With this block diagonal structure [2] and the continuous time approximation, the compensated delta parameters are given by:

$$\begin{aligned}\boldsymbol{\mu}_{\Delta\mathtt{y}}^{(mu)} &= \mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)}(\mathbf{A}_{\Delta}\boldsymbol{\mu}_{\Delta\mathtt{x}}^{(m)} + \boldsymbol{b}_{\Delta}), \\ \boldsymbol{\Sigma}_{\Delta\mathtt{y}}^{(mu)} &= \mathrm{diag}\left(\mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)}\boldsymbol{\Sigma}_{\Delta\mathtt{x}}^{(m)}\mathbf{J}_{\mathtt{x},\mathbf{w}}^{(mu)\mathsf{T}} + \mathbf{J}_{\mathtt{n},\mathbf{w}}^{(mu)}\boldsymbol{\Sigma}_{\Delta\mathtt{n}}^{(u)}\mathbf{J}_{\mathtt{n},\mathbf{w}}^{(mu)\mathsf{T}}\right)\end{aligned} \tag{17}$$

where $\boldsymbol{\mu}_{\Delta\mathtt{x}}^{(m)}$, $\boldsymbol{\Sigma}_{\Delta\mathtt{x}}^{(m)}$ are the m-th component parameters for the clean delta features respectively, and $\boldsymbol{\Sigma}_{\Delta\mathtt{n}}^{(u)}$ is the variance of $\Delta\boldsymbol{n}$, the noise delta.

The above "Joint" scheme uses a speaker transform, $\mathbf{K} = (\mathbf{W}, \mathbf{W}_{\Delta})$ to *explicitly* adapt the models to the target speaker. In contrast to the "VTS-MLLR" scheme, the speaker transform is applied *before* the noise transform.

IV. Transform Estimation

There are two sets of transform parameters to be estimated in the "Joint" and "VTS-MLLR" schemes: the linear transform $\mathbf{K}$ and the noise model parameters $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}^{(u)}\}$, where $\boldsymbol{\Phi}^{(u)}$ is the noise model parameters of u-th utterance, $\boldsymbol{\Phi}^{(u)} = (\boldsymbol{\mu}_{\mathtt{n}}^{(u)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\Sigma}_{\mathtt{n}}^{(u)}, \boldsymbol{\Sigma}_{\Delta\mathtt{n}}^{(u)})$. These parameters can be optimised using EM. This yields the following auxiliary function for both forms of compensation [3]:

$$\mathcal{Q}(\mathbf{K}, \boldsymbol{\Phi}) = \sum_{u,m,t} \gamma_t^{(mu)} \log \mathcal{N}(\boldsymbol{o}_t^{(u)}; \boldsymbol{\mu}_{\mathtt{o}}^{(mu)}, \boldsymbol{\Sigma}_{\mathtt{o}}^{(mu)}), \tag{18}$$

where the summation over u involves all the utterances belonging to the same speaker, $\gamma_t^{(mu)}$ is the posterior probability of component m at time t of the u-th utterance given the current transform parameters $(\hat{\mathbf{K}}, \hat{\boldsymbol{\Phi}})$, $\boldsymbol{o}_t^{(u)} = [\boldsymbol{y}_t^{(u)\mathsf{T}}, \Delta\boldsymbol{y}_t^{(u)\mathsf{T}}]^{\mathsf{T}}$ is the t-th observation vector of the u-th utterance, and

$$\boldsymbol{\mu}_{\mathtt{o}}^{(mu)} = \begin{bmatrix} \boldsymbol{\mu}_{\mathtt{y}}^{(mu)} \\ \boldsymbol{\mu}_{\Delta\mathtt{y}}^{(mu)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathtt{o}}^{(mu)} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathtt{y}}^{(mu)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\Delta\mathtt{y}}^{(mu)} \end{bmatrix} \tag{19}$$

are the adapted mean and covariances, obtained by Eq. (15) and Eq. (17) for "Joint" or Eq. (12) and Eq. (13) for "VTS-MLLR".

To estimate $\mathbf{K}$ and $\boldsymbol{\Phi}$ for both "Joint" and "VTS-MLLR" schemes, a *block coordinate descent* strategy is adopted: first, for each speaker, $\mathbf{W}$ and $\mathbf{W}_{\Delta}$ are initialised as $[\mathbf{I}, \mathbf{0}]$, and $\boldsymbol{\Phi}$ as the standard VTS-based noise estimates for each utterance; then $\mathbf{K}$ is optimised at the speaker level while keeping the noise model parameter fixed at the current noise estimates $\hat{\boldsymbol{\Phi}}$; finally, given the speaker transform updated, $\hat{\mathbf{K}}$, the noise parameter $\boldsymbol{\Phi}$ is re-estimated. This process is repeated N_{EM} times.

A. Transform estimation for "VTS-MLLR"

In the"VTS-MLLR" scheme, the VTS-compensated static and dynamic parameters are transformed independently by $\mathbf{W}$ and $\mathbf{W}_{\Delta}$ respectively. Hence the estimation of $\mathbf{K} = (\mathbf{W}, \mathbf{W}_{\Delta})$ can be done separately. Given the noise estimates for each utterance, $\boldsymbol{\Phi}^{(u)}$, the transform $\mathbf{W}$ needs to be estimated at the speaker level, involving multiple utterances thus associated with different noise conditions. The transform estimation

[2] It is possible to extend the theory to handle full transforms, however this is not addressed in this paper.

[3] This section does not discuss multiple speakers. The extension to multiple speakers is straightforward.

statistics in [6] are modified to reflect the changing noise conditions:

$$\begin{aligned}\boldsymbol{k}_i &= \sum_u \sum_m \sum_t \frac{\gamma_t^{(mu)} y_{t,i}^{(u)}}{\sigma_{\mathtt{vts},i}^{(mu)2}} \boldsymbol{\xi}_{\mathtt{vts,y}}^{(mu)}, \\ \mathbf{G}_i &= \sum_u \sum_m \frac{\gamma^{(mu)}}{\sigma_{\mathtt{vts},i}^{(mu)2}} \boldsymbol{\xi}_{\mathtt{vts,y}}^{(mu)} \boldsymbol{\xi}_{\mathtt{vts,y}}^{(mu)\mathsf{T}},\end{aligned} \tag{20}$$

with $y_{t,i}^{(u)}$ being the i-th element of $\boldsymbol{y}_t^{(u)}$, and $\sigma_{\mathtt{vts},i}^{(mu)2}$ the i-th diagonal item of $\boldsymbol{\Sigma}_{\mathtt{vts,y}}^{(mu)}$, $\gamma^{(mu)} = \sum_t \gamma_t^{(mu)}$. Given these statistics, the i-th row of $\mathbf{W}$, $\mathbf{w}_i^{\mathsf{T}}$, is obtained by $\mathbf{w}_i^{\mathsf{T}} = \boldsymbol{k}_i^{\mathsf{T}} \mathbf{G}_i^{-1}$. Estimating of $\mathbf{W}_{\Delta}$ is done similarly.

Given the current linear speaker transform $\hat{\mathbf{K}}$, the parameters of the noise transform can be updated. This requires the noise estimation approaches in, for example [13], [21], [30] to be modified to reflect that the compensated model will have the speaker transform applied. To estimate the additive and convolutional noise mean, a first-order VTS approximation is made, e.g., the mean and covariance for the static feature are approximated as follows:

$$\begin{aligned}\boldsymbol{\mu}_{\mathtt{y}}^{(mu)} \approx & \hat{\boldsymbol{\mu}}_{\mathtt{y}}^{(mu)} + \hat{\mathbf{A}}\hat{\mathbf{J}}_{\mathtt{h}}^{(mu)}(\boldsymbol{\mu}_{\mathtt{h}}^{(u)} - \hat{\boldsymbol{\mu}}_{\mathtt{h}}^{(u)}) + \hat{\mathbf{A}}\hat{\mathbf{J}}_{\mathtt{n}}^{(mu)}(\boldsymbol{\mu}_{\mathtt{n}}^{(u)} - \hat{\boldsymbol{\mu}}_{\mathtt{n}}^{(u)}) \\ \boldsymbol{\Sigma}_{\mathtt{y}}^{(mu)} \approx & \operatorname{diag}\left(\hat{\mathbf{J}}_{\mathtt{x}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)} \hat{\mathbf{J}}_{\mathtt{x}}^{(mu)\mathsf{T}} + \hat{\mathbf{J}}_{\mathtt{n}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{n}}^{(u)} \hat{\mathbf{J}}_{\mathtt{n}}^{(mu)\mathsf{T}}\right)\end{aligned} \tag{21}$$

where $\hat{\mathbf{J}}_{\mathtt{x}}^{(mu)}$,$\hat{\mathbf{J}}_{\mathtt{h}}^{(mu)}$, $\hat{\mathbf{J}}_{\mathtt{n}}^{(mu)}$ and $\hat{\boldsymbol{\mu}}_{\mathtt{y}}^{(mu)}$ are the Jacobian matrices and the compensated mean based on the current noise estimation $\hat{\boldsymbol{\mu}}_{\mathtt{h}}^{(u)}, \hat{\boldsymbol{\mu}}_{\mathtt{n}}^{(u)}$ and the current linear transform $\hat{\mathbf{K}}$. Because of this VTS approximation, the auxiliary is now a quadratic function of the noise means. Hence $\boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}$ can be obtained via solving a linear equation, in a similar fashion as the one in [30]. After the noise mean estimation, the noise variance $\boldsymbol{\Sigma}_{\mathtt{n}}^{(u)}$, $\boldsymbol{\Sigma}_{\Delta\mathtt{n}}^{(u)}$ and $\boldsymbol{\Sigma}_{\Delta\mathtt{n}^2}^{(u)}$ are estimated via the second order method, in the same way as [13]. At each iteration a check that the auxiliary function increases is performed and the estimates backed-off if necessary [30][4].

B. Transform estimation for "Joint"

For the "Joint" scheme, estimating the noise parameters, given the current speaker transform $\hat{\mathbf{K}}$ is a simple extension of VTS-based noise estimation in [21], [30]: prior to the noise estimation, the clean speech mean is transformed to the speaker-dependent clean speech mean. However, estimating the speaker transform $\mathbf{K}$ is not straight-forward, since the transform is applied to the "clean" speech and then VTS compensation applied. To address this non-linearity, a first-order vector Taylor series approximation can again be employed to express $\boldsymbol{\mu}_{\mathtt{y}}^{(mu)}$ and $\boldsymbol{\Sigma}_{\mathtt{y}}^{(mu)}$ as functions of the current, $\hat{\mathbf{W}}$, and new,$\mathbf{W}$, estimates of the speaker transform,

$$\begin{aligned}\boldsymbol{\mu}_{\mathtt{y}}^{(mu)} \approx & \mathbf{f}(\hat{\mathbf{W}}\boldsymbol{\xi}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}^{(u)}, \boldsymbol{\mu}_{\mathtt{n}}^{(u)}) + \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)}(\mathbf{W} - \hat{\mathbf{W}})\boldsymbol{\xi}_{\mathtt{x}}^{(m)} \\ \boldsymbol{\Sigma}_{\mathtt{y}}^{(mu)} \approx & \operatorname{diag}\left(\mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)} \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)\mathsf{T}} + \mathbf{J}_{\mathtt{n},\hat{\mathbf{w}}}^{(mu)} \boldsymbol{\Sigma}_{\mathtt{n}}^{(u)} \mathbf{J}_{\mathtt{n},\hat{\mathbf{w}}}^{(mu)\mathsf{T}}\right)\end{aligned} \tag{22}$$

[4]Since the second order optimisation assumes the approximation in Eq. (21), there is no guarantee that the auxiliary function in Eq. (18) will be non-decreasing.

while for the delta parameters,

$$\begin{aligned}\boldsymbol{\mu}_{\Delta\mathtt{y}}^{(mu)} \approx & \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)} \mathbf{W}_{\Delta} \boldsymbol{\xi}_{\Delta\mathtt{x}}^{(m)} \\ \boldsymbol{\Sigma}_{\Delta\mathtt{y}}^{(mu)} \approx & \operatorname{diag}\left(\mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)} \boldsymbol{\Sigma}_{\Delta\mathtt{x}}^{(m)} \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)\mathsf{T}} + \mathbf{J}_{\mathtt{n},\hat{\mathbf{w}}}^{(mu)} \boldsymbol{\Sigma}_{\Delta\mathtt{n}}^{(u)} \mathbf{J}_{\mathtt{n},\hat{\mathbf{w}}}^{(mu)\mathsf{T}}\right)\end{aligned} \tag{23}$$

Due to the approximation in Eq. (23), the optimisation of $\mathbf{W}$ and $\mathbf{W}_{\Delta}$ again becomes two separate but similar problems. The estimation of $\mathbf{W}$, given the current noise estimation $\hat{\boldsymbol{\Phi}}$ and the VTS approximation in Eq. (22), uses the following, approximate, auxiliary function (up to some constant term):

$$q(\mathbf{W}; \hat{\mathbf{W}}) = \sum_{u,m,t} \gamma_t^{(mu)} \log \mathcal{N}(\boldsymbol{z}_t^{(mu)}; \mathbf{W}\boldsymbol{\xi}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{full}}^{(mu)}) \tag{24}$$

where

$$\begin{aligned}\boldsymbol{z}_t^{(mu)} &= \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)-1}(\boldsymbol{y}_t^{(u)} - \hat{\boldsymbol{\mu}}_{\mathtt{y}}^{(mu)} + \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)} \hat{\mathbf{W}} \boldsymbol{\xi}_{\mathtt{x}}^{(m)}) \\ \boldsymbol{\Sigma}_{\mathtt{full}}^{(mu)} &= \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)-1} \hat{\boldsymbol{\Sigma}}_{\mathtt{y}}^{(mu)} \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)-\mathsf{T}}\end{aligned}$$

and $\hat{\boldsymbol{\mu}}_{\mathtt{y}}^{(mu)}$, $\hat{\boldsymbol{\Sigma}}_{\mathtt{y}}^{mu)}$ are the compensated parameters using the current transforms $\hat{\mathbf{W}}$ and $\hat{\boldsymbol{\Phi}}^{(u)}$. As $\boldsymbol{\Sigma}_{\mathtt{full}}^{(mu)}$ is a full matrix (in terms of the static parameters), this optimisation is equivalent to the MLLR estimation with full covariance matrices [31]. Let $\mathbf{p}_i^{(mu)\mathsf{T}}$ be the i-th row vector of $\boldsymbol{\Sigma}_{\mathtt{full}}^{(mu)-1}$, $p_{ij}^{(mu)}$ the j-th element of $\mathbf{p}_i^{(mu)}$, and

$$\begin{aligned}\boldsymbol{k}_i &= \sum_{u,m,t} \gamma_t^{(mu)} \mathbf{p}_i^{(mu)\mathsf{T}} \boldsymbol{z}_t^{(mu)} \boldsymbol{\xi}_{\mathtt{x}}^{(m)} - \sum_{j \neq i} \mathbf{G}_{ij} \mathbf{w}_j, \\ \mathbf{G}_{ij} &= \sum_{m,u} \gamma^{(mu)} p_{ij}^{(mu)} \boldsymbol{\xi}_{\mathtt{x}}^{(m)} \boldsymbol{\xi}_{\mathtt{x}}^{(m)\mathsf{T}}.\end{aligned} \tag{25}$$

Differentiating the auxiliary with respect to $\mathbf{w}_i^{\mathsf{T}}$ yields

$$\frac{\partial q(\mathbf{W}; \hat{\mathbf{W}})}{\partial \mathbf{w}_i^{\mathsf{T}}} = -\mathbf{w}_i^{\mathsf{T}} \mathbf{G}_{ii} + \boldsymbol{k}_i^{\mathsf{T}}. \tag{26}$$

The update formula for $\mathbf{w}_i$ depends on all the other row vectors through $\boldsymbol{k}_i$. Thus an iterative procedure is required [31]: first $\mathbf{G}_{ij}$ is set as $\mathbf{0}$ for all $j \neq i$ to get an initial $\mathbf{w}_i$; then $\mathbf{w}_i$ and $\boldsymbol{k}_i$ are updated on a row-by-row basis. Normally, one or two passes through all the row vectors is sufficient.

For estimation of $\mathbf{W}_{\Delta}$, another auxiliary function is used:

$$q_{\Delta}(\mathbf{W}_{\Delta}; \hat{\mathbf{W}}) = \sum_{u,m,t} \gamma_t^{(mu)} \log \mathcal{N}(\Delta\boldsymbol{z}_t^{(mu)}; \mathbf{W}_{\Delta}\boldsymbol{\xi}_{\Delta\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{full},\Delta}^{(mu)}) \tag{27}$$

where

$$\begin{aligned}\Delta\boldsymbol{z}_t^{(mu)} &= \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)-1} \Delta\boldsymbol{y}_t^{(u)} \\ \boldsymbol{\Sigma}_{\mathtt{full},\Delta}^{(mu)} &= \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)-1} \hat{\boldsymbol{\Sigma}}_{\Delta\mathtt{y}}^{(mu)} \mathbf{J}_{\mathtt{x},\hat{\mathbf{w}}}^{(mu)-\mathsf{T}}.\end{aligned}$$

This has the same form as the auxiliary function in Eq. (24). Thus the same procedure can be applied to estimate $\mathbf{W}_{\Delta}$.

As a first-order approximation, Eq. (22), is used to derive the approximate auxiliary functions. Optimising $\mathbf{K}$ via $q(\mathbf{W}; \hat{\mathbf{W}})$ and $q_{\Delta}(\mathbf{W}_{\Delta}; \hat{\mathbf{W}})$ is not guaranteed to increase $\mathcal{Q}(\mathbf{K}, \hat{\boldsymbol{\Phi}})$ or the log-likelihood of the adaptation data. To address this problem, a simple back-off approach similar to the one used in [30], is adopted in this work. Note the back-off approach, i.e., step 3 in the following procedure, guarantees that the

auxiliary function is non-decreasing. The estimation of the "Joint" speaker transform is thus:

1 Collect sufficient statistics $\boldsymbol{k}_i$ and $\mathbf{G}_{ij}$ based on the current transform $\hat{\mathbf{W}}$ and $\hat{\mathbf{\Phi}}$. Similar statistics are also collected for $\mathbf{W}_\Delta$.
2 Use the row-iteration method to find the $\check{\mathbf{K}} = (\check{\mathbf{W}}, \check{\mathbf{W}}_\Delta)$ such that $\check{\mathbf{W}} = \arg\max_{\mathbf{W}} q(\mathbf{W}; \hat{\mathbf{W}})$ and $\check{\mathbf{W}}_\Delta = \arg\max_{\mathbf{W}_\Delta} q_\Delta(\mathbf{W}_\Delta; \hat{\mathbf{W}})$
3 Find $\alpha \in [0, 1]$, such that $\mathbf{K} = \alpha\hat{\mathbf{K}} + (1-\alpha)\check{\mathbf{K}}$ satisfy $\mathcal{Q}(\mathbf{K}, \hat{\mathbf{\Phi}}) \geq \mathcal{Q}(\hat{\mathbf{K}}, \hat{\mathbf{\Phi}})$
4 Update current estimate $\hat{\mathbf{K}} \leftarrow \mathbf{K}$, and go to step 1 N_q times. It is observed in the experiments that setting $N_q = 5$ is enough for the auxiliary to converge in most of the cases.

The above procedure allows the speaker transform to be estimated. The noise transforms can then be re-estimated and the whole process repeated. However it is worth noting that there is no unique optimal value for the speaker and noise transforms. There is no way to distinguish between the speaker bias, $\boldsymbol{b}$, from the convolutional noise mean, $\boldsymbol{\mu}_{\mathrm{h}}^{(u)}$. This is not an issue as the parameters of the speaker model are estimated given the set of noise parameters. This ensures that all the convolutional noise means are consistent with one another.

V. Experiments

The performances of the two model-based schemes, and as a contrast, a feature enhancement approach, was evaluated in terms of both adaptation (batch-mode) and factorisation (factorisation mode).

The AURORA4 [32] corpus was used for evaluation. This corpus is derived from the Wall Street Journal (WSJ0) 5k-word closed vocabulary dictation task. 16kHz data were used in all the experiments here. Two training sets, clean and multi-style training datasets, are available. Both these two sets comprise 7138 utterances from 83 speakers. In the clean training dataset, all these 7138 utterances were recorded using a close-talking microphone, whilst for the multi-style data, half of them came from desk-mounted, secondary microphones. The multi-style data had 6 different types of noise added, with the SNR ranging from 20dB to 10dB, averaged 15 dB. There are 4 test sets for this task. 330 utterances from 8 speakers, recorded by the close talking microphone, form `01` (set A). 6 types of noises, as those in multi-style training data, were added to the clean data, with randomly selected SNRs (from 15dB to 5dB, average 10 dB). These form the `02` to `07` (set B). Recordings of these utterances for desk-mounted secondary microphones were also provided in `08` (set C). Noise were added to set C to form `09` to `14` (set D).

All the acoustic models used in experiments were cross-word triphone models with 3140 distinct tied-states and 16 component per state. The standard bi-gram language model provided for the AURORA4 experimental framework was used in decoding. For all the experiments unsupervised adaptation was performed. Where MLLR adaptation was performed, block-diagonal transforms with two regression classes (one speech, one silence) were used. The VTS-based noise estimation was performed on a per-utterance basis, while the speaker adaptation was performed on the speaker level. To minimise differences due to the different forms of adaptation, multiple EM iterations were performed when estimating transforms.

A. Baseline systems

In order to evaluate the effectiveness of the proposed speaker and noise adaptation scheme, a series of baseline systems were build. The first one was the "clean" system, where the acoustic models were trained on the clean training set. A 39 dimensional front-end feature vector was used, consisting of 12 MFCCs appended with the zeroth cepstrum, delta and delta-delta coefficients. Without adaptation, this clean-trained model achieved a WER of 7.1% on clean test set (set A), but the performance was severely affected by the noise: the average WER on all 4 sets was 58.5%, which indicates that the clean-trained model is fragile when operated in noisy conditions. When VTS adaptation was performed, the noise model parameters were initialised using the first and last 20 frames of each utterance. The acoustic models were then compensated using these noise models, and the initial hypotheses generated. With this initial hypothesis, the noise models were re-estimated, followed by the generation of updated hypothesis. This first iteration of VTS was used to provide the supervision for the following adaptation. A second iteration of VTS was also performed to refine the noise models, then the final hypothesis was generated. Note that performing more VTS iterations is possible, but only provided a minimal performance gain. The second system used the same front-end, but is adaptively trained on the multi-style data. A "neutral" model, denoted as "VAT", was estimated using VTS-based adaptive training ([14], [30]), where the differences due to noise were reduced by the VTS transforms. The same procedure for noise model estimation and hypothesis generation as the one used for the clean-trained acoustic models was performed. As a comparison of model compensation versus feature compensation approaches, the ETSI advanced feature (AFE) was used to build the third baseline on the multi-style data. This system is referred to as "AFE".

Results of these baselines are presented in Table I. Using VTS-based noise adaptation, the clean-trained model achieved a WER of 17.8%. Compared with other feature-based or model-based noise robustness schemes on AURORA4 (e.g., [33]), it is clear this provides a fairly good baseline on this task. As expected, the use of the adaptively trained acoustic model (the VAT system) gave gains over the clean system on noisy data: the average WER was further reduced from 17.8% to 15.9%. However, a small degradation on the clean set (8.5% of VAT vs. 6.9% of clean) can be seen. This may be explained as VTS is not able to completely remove the effects of noise. Thus the "pseudo" clean speech parameters estimated by adaptive training will have some residual noise effects and so will be slightly inconsistent with the clean speech observation in set A. It is also interesting to look at the performances of the AFE system. With AFE, multi-style training achieved a WER of 21.4%. Note that, the multi-style model using MFCC feature achieved a WER of 27.1%. However, the large performance gap (21.4% vs. 15.9%) between AFE and its counterpart of model-based schemes,

Model	Adaptation	A	B	C	D	Avg.
Clean	VTS	6.9	15.1	11.8	23.3	17.8
VAT	VTS	8.5	13.7	11.8	20.1	15.9
AFE	—	8.8	16.7	19.1	28.6	21.4

TABLE I
PERFORMANCES (WER, IN %) OF THREE BASELINE SYSTEMS.

Adaptation	A	B	C	D	Avg.
VTS	6.9	15.1	11.8	23.3	17.8
VTS-MLLR	5.0	12.1	9.0	19.8	14.7
Joint	5.0	12.1	8.6	19.7	14.6
Joint-MLLR	5.0	11.5	8.1	19.1	14.1

TABLE II
BATCH MODE SPEAKER AND NOISE ADAPTATION OF CLEAN-TRAINED ACOUSTIC MODEL

the VAT system, demonstrates the usefulness of model-based schemes for this task.

B. Batch-mode speaker and noise adaptation

The above experiments built a series of baseline systems, where only noise adaptation was performed. In the following experiments, acoustic models were adapted to both the target speaker and environment. In the first set of experiments, speaker and noise adaptation was performed in a batch mode (referred to as "`bat`"), i.e. the adaptation experiments were run where speaker and noise (utterance-level) transforms were estimated for each speaker for each task[5]. The "Joint" and "VTS-MLLR" schemes were first examined using the clean-trained acoustic model. Following the same procedure used in the baseline systems, one VTS iteration was run for each utterance to generate the supervision hypothesis. The generated noise models were also taken to initialise the noise parameters $\mathbf{\Phi}$. The speaker level transform, $\mathbf{K}$, was initialised as the identity transform. Then, as discussed in Section IV, the block coordinate descent optimisation strategy is applied for "Joint" and "VTS-MLLR". Multiple iterations, $N_{\text{EM}} = 4$, were used to update the speaker transform and noise models. As an additional contrast, an MLLR transform was applied on top of the "Joint", again estimated at the speaker level, yielding another scheme "Joint-MLLR". The results of these batch-mode speaker and noise adaptation experiments are presented in Table II. Significant performance gains[6] were obtained using both "Joint" (14.6%) and "VTS-MLLR" (14.7%), compared to the baseline VTS performance (17.8%). The best performance was obtained using the "Joint-MLLR" scheme (14.1%), which indicates that there is still some residual mismatch after "Joint" adaptation and a general linear transform can be used to reduce this mismatch. These experiments serve as a contrast to the factorisation experiments in the next section.

[5]The speaker transforms were estimated for each speaker on each noise condition (01-14), and were used only in the noise condition where speaker transforms were estimated from. The noise transform was always estimated for every utterance.

[6]All statistical significance tests are based on a matched pair-wise significance test at a 95% confidence level.

C. Speaker and noise transform factorisation

To investigate the factorisation of speaker and noise transforms, a second set of experiments were conducted. Again, the noise transforms were estimated for each utterance. However in contrast to the batch-mode adaptation, the speaker transforms were estimated from either `01` or `04`[7]. These speaker transforms were then fixed and used for all the test sets, just the utterance-level noise transforms were re-estimated. The same setup as the previous experiments was used to estimate the speaker transform from either `01` or `04`[8]. This factorisation mode allows very rapid adaptation to the target condition.

Table III presents the results of the speaker and noise factorisation experiments using clean-trained acoustic models. It is seen that speaker transforms estimated from either `01` (clean) or `04` (restaurant) improve the average performance over all conditions (16.7% and 15.4% compared with 17.8%). This indicates that it is possible to factorise the speaker and noise transform to some extent. For the speaker transform estimated using `01`, the "clean" data, gains in performance (compared with VTS adaptation only) for all the four sets were obtained. Interestingly the average performance was improved by estimating the speaker transform in a noisy environment, `04`. Other than on the clean set A this yielded lower WERs than the clean estimated model for all of the B test sets. This indicates that although the speaker and noise transforms can be factorised to some extent, the linear transform for the speaker characteristics derived from the "Joint" scheme is still modelling some limitations in the VTS mismatch function to fully reflect the noise environment. It is also interesting to compare the results with the batch-mode system from Table II. For test set B the average WER for the batch-mode "Joint" scheme was 12.1%, compared to 12.5% when the speaker MLLR transform was estimated using `04` and then fixed for all the test sets. This indicates that for these noise conditions the factorisation was fairly effective. However for the clean set A, the performance difference between the batch-mode and the factorisation mode was greater. This again indicates that the speaker transform was modelling some of the limitations of the VTS mismatch function. Results of speaker and noise factorisation using "VTS-MLLR" scheme are also presented in Table III. It is clear that the "VTS-MLLR" scheme does not have the desired factorisation attribute, as the linear transforms estimated from one particular noise conditions cannot generalise to other conditions. Hence, "VTS-MLLR" scheme is not further investigated for factorised speaker and noise adaptation.

The above experiments demonstrate the factorisation attribute of "Joint" when the clean-trained acoustic models were used. To examine whether this attribute is still valid for adaptively trained acoustic models, a second set of experiments

[7]In principle, it is possible to estimate speaker transform from any of the 14 test sets. Unfortunately, utterances from three speakers in set C and set D were recorded by a handset microphone which limits the speech signal to telephone bandwidth. This also means it is not useful to estimate speaker transforms from set C or set D.

[8]When the clean-trained acoustic models were adapted by VTS (line 1, Table III), `04` was the worst performed in set B. This trend was also observed in line 1, Table IV.

Models	Adaptation		A	B							C	D							Avg.
	Scheme	Spk. Est.	01	02	03	04	05	06	07	Avg.	08	09	10	11	12	13	14	Avg.	
AFE	—	—	8.8	13.1	15.9	20.0	18.4	15.1	17.4	16.7	19.1	24.1	27.8	31.1	30.9	28.6	29.4	28.6	21.4
	MLLR	bat	7.0	8.9	13.1	16.6	15.3	12.2	14.4	13.4	10.5	14.6	19.8	23.0	21.9	18.5	21.4	19.9	15.5
		01	7.0	18.5	20.6	25.8	24.1	20.7	21.5	21.9	21.0	28.1	32.9	37.7	35.6	32.0	33.1	33.2	25.6
		04	8.7	10.5	14.3	16.6	16.2	13.0	15.1	14.3	16.3	19.2	24.7	26.4	27.0	23.4	26.5	24.5	18.4
VAT	VTS	—	8.5	9.8	13.2	16.0	14.6	12.0	16.4	13.7	11.8	12.4	19.6	23.1	23.2	18.8	23.8	20.1	15.9
	Joint	bat	5.6	6.7	10.5	13.4	12.1	9.5	13.8	11.0	8.8	10.3	17.8	20.7	20.8	16.5	20.8	17.8	13.4
		01	5.6	7.6	12.9	17.7	14.2	11.7	16.0	13.4	11.1	11.5	19.6	24.8	22.6	19.7	23.9	20.3	15.6
		04	6.9	7.4	11.2	13.4	12.6	10.3	14.1	11.5	10.4	11.0	18.2	21.2	20.4	17.9	22.2	18.5	14.1

TABLE IV
FACTORISED SPEAKER AND NOISE ADAPTATION OF VAT AND AFE MODELS.

Scheme	Spk. Est.	A	B	C	D	Avg.
VTS	—	6.9	15.1	11.8	23.3	17.8
VTS-MLLR	01	5.0	20.2	16.5	28.0	22.2
	04	10.2	19.7	19.7	28.0	22.5
Joint	01	5.0	14.1	10.4	22.3	16.7
	04	7.0	12.5	11.0	20.4	15.4

TABLE III
FACTORISED SPEAKER AND NOISE ADAPTATION OF CLEAN-TRAINED ACOUSTIC MODELS USING "JOINT" AND "VTS-MLLR".

was run. VAT acoustic models were adapted by "Joint", in both batch and factorisation modes. For the latter, 01 and 04 were again used for speaker transform estimation. Results on all 14 subsets are presented in Table IV. Since the acoustic models are adaptively trained, improved performances are expected, compared with those in Table III. Note that factorisation mode adaptation on 01(04) using the speaker transform estimated from 01(04) is equivalent to the batch-mode adaptation, thus gives identical results to bat on 01(04). The same trends as those observed in the previous experiments can be seen: a batch-mode "Joint" adaptation yielded large gains over VTS adaptation only (13.4% vs. 15.9%, average on all 4 sets), while using the speaker transform estimated on 04 achieved a very close performance, 14.1%. The advantages of using "Joint" scheme were fairly maintained with the adaptively trained acoustic models.

It is also of interest to look at the experiments of the speaker and noise adaptation with the AFE acoustic models. Speaker adaptation for AFE model was done via an MLLR mean transform with the same block diagonal structure, again estimated at the speaker level. The AFE model was first used to generate the supervision hypothesis, following the MLLR adaptation, and then the final hypothesis was generated. Though multiple iterations of hypothesis generation and transform re-estimation could be used, it was found in the experiments the gain was minimal. In the batch-mode adaptation, speaker transforms were estimated for every single set, while for the factorisation mode, speaker transforms were estimated from 01 or 04. Results of these experiments are summarised in the first block of Table IV. It can be seen that the speaker transform estimated from 01 did not generalise well to other noisy sets (WER increased from 21.4% to 25.6%), while the one estimated from 04 can generalise to other noise conditions. This suggests that for feature normalisation style trained acoustic models, the linear transform estimated from one noise data can be applied to other noise conditions for the same speaker. However examining the results in more detail shows that this factorisation using AFE is limited. A 19% relative degradation (18.4% of the factorisation mode vs 15.5% of the batch-mode) was observed. This compares to only 5% relative degradation for the "Joint" scheme. It is worth noting that batch-mode AFE with MLLR (15.5%) is still significantly worse that the "Joint" scheme run in a factorised mode on the 04 data (14.1%)

VI. CONCLUSION

This paper has examined approaches to handling speaker and noise differences simultaneously. A new adaptation scheme, "Joint", is proposed, where the clean acoustic model is first adapted to the target speaker via an MLLR transform, and then compensated for the effect of noise via VTS-based model compensation. Adapting the underlying clean speech model, rather than the noise compensated model, enables the speaker transform and the noise compensation to be kept distinct from one another. This "orthogonality" thus supports acoustic factorisation, which allows flexible use of the estimated transforms. For example, as the one examined in this paper, the same speaker transform can be used in a range of very different noise conditions.

This scheme is compared with two alternatives for handling both speaker and noise differences. The first one, "VTS-MLLR", is a more standard combination of VTS and MLLR where the MLLR transform is applied after VTS compensation. Note this form of scheme is extended in this paper to support inter-leaved estimation of the noise and speaker transforms, rather than estimating them sequentially. The second scheme "AFE-MLLR" uses AFE to obtain de-noised observations prior to adaptation to the speaker.

The AURORA4 data was used for evaluation. Experimental results demonstrate that if operated in a batch mode, both "VTS-MLLR" and "Joint" give gains over noise adaptation alone. However, only "Joint" supports the factorisation mode adaptation, which allows a very rapid speaker and noise adaptation. "Joint" scheme was also compared with the scheme that use feature-based approach to noise compensation, the "AFE-MLLR". Results show "AFE-MLLR" does not achieve the same level of performance as "Joint".

This paper has proposed the "Joint" scheme for speaker and noise adaptation. As speaker and noise factors are modelled separately, it also enables speaker adaptation using a broad range of noisy data. Throughout the paper, it is assumed that the speaker characteristics does not change over the time, and the speaker adaptation is carried out in a static mode. It will be interesting to apply "Joint" in an incremental mode in future work to domains where the adaptation data has large variations in background noise, for example in-car applications.

REFERENCES

[1] Y.-Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the AURORA4 task," in *Proc. ICASSP*, 2011.

[2] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.

[3] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.

[4] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP-1996*, pp. 353–356.

[5] 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.

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

[7] ——, "Cluster adaptive training of hidden Markov models," *IEEE transactions on speech and audio processing*, vol. 8, no. 4, pp. 417–428, 2002.

[8] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[9] 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*.

[10] D. Kim and M. J. F. Gales, "Adaptive training with noisy constrained maximum likelihood linear regression for noise robust speech recognition," in *Proc. Interspeech-2009*, pp. 2382–2386.

[11] P. Nguyen, C. Wellekens, and J. C. Junqua, "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments," in *Proc. Eurospeech-1999*.

[12] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP-96*, pp. 1137–1140.

[13] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proc. ICASSP-2007*, pp. 389–392.

[14] O. Kalinli, M. L. Seltzer, and A. Acero, "Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition," in *Proc. ICASSP-2009*.

[15] E. standard doc., "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced frontend feature extraction algorithm; compression algorithms," ETSI, Tech. Rep. ES 202 050 v1.1.3, 2003.

[16] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP-2000*.

[17] V. Stouten, H. Van hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech communication*, vol. 48, no. 11, pp. 1502–1514, 2006.

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

[19] M. J. F. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Cambridge University, 1995.

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

[21] J. Li, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in *Proc. ASRU-2007*.

[22] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 975 – 983, 2005.

[23] Y. Hu and Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," in *Proc. Interspeech-2007*, pp. 1042–1045.

[24] L. Buera, A. Miguel, O. Saz, A. Ortega, and E. Lleida, "Unsupervised data-driven feature vector normalization with acoustic model adaptation for robust speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 296 –309, 2010.

[25] M. J. F. Gales, "Predictive model-based compensation schemes for robust speech recognition," *Speech communication*, vol. 25, no. 1-3, pp. 49–74, 1998.

[26] L. Rigazio, P. Nguyen, D. Kryze, and J.-C. Junqua, "Separating speaker and environment variabilities for improved recognition in non-stationary conditions," in *Proc. Eurospeech-2001*.

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

[28] K. Yu and M. J. F. Gales, "Adaptive training using structured transforms," in *Proc. ICASSP-2004*.

[29] R. A. Gopinath *et al.*, "Robust speech recognition in noise – performance of the IBM continuous speech recogniser on the ARPA noise spoke task," in *Proc. APRA workshop on spoken language system technology*, 1995, pp. 127–130.

[30] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for robust large vocabulary speech recognition," University of Cambridge, Tech. Rep. CUED/F-INFENG/TR552, 2006.

[31] K. C. Sim and M. J. F. Gales, "Adaptation of precision matrix models on LVCSR," in *Proc. ICASSP2005*.

[32] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Information Process, Mississippi State University, Tech. Rep.

[33] K. Demuynck, X. Zhang, D. Van Compernolle, and H. Van hamme, "Feature verus model based noise robustness," in *Proc. Interspeech-2010*, pp. 721–724.