INTERSPEECH 2009 BRIGHTON

Adaptive Training with Noisy Constrained Maximum Likelihood Linear Regression for Noise Robust Speech Recognition

D. K. Kim[1,2] *and M. J. F. Gales*[1]

[1]Cambridge University Engineering Department, Trumpington St., Cambridge, CB2 1PZ, UK
[2]Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, 500-757, Korea

dkim@chonnam.ac.kr, mjfg@eng.cam.ac.uk

10.21437/Interspeech.2009-364

Abstract

Adaptive training is a widely used technique for building speech recognition systems on non-homogeneous training data. Recently there has been interest in applying these approaches for situations where there is significant levels of background noise. This work extends the most popular form of linear transform for adaptive training, constrained MLLR, to reflect additional uncertainty from noise corrupted observations. This new form of transform, Noisy CMLLR, uses a modified version of generative model between clean speech and noisy observation, similar to factor analysis. Adaptive training using NCMLLR with both maximum likelihood and discriminative criteria are described. Experiments are conducted on noise-corrupted Resource Management and in-car recorded data. In preliminary experiments this new form achieves improvements in recognition performance over the standard approach in low signal-to-noise ratio conditions.

Index Terms: speech recognition, speaker adaptation, noise robustness

1. Introduction

Current large vocabulary speech recognition systems are normally constructed on large amounts of non-homogeneous training data from multiple speakers with different background noise and channel conditions. One of the techniques to build a speech recognition system on this non-homogeneous data is to use adaptive training [1, 2] with, for example, noise and speaker specific linear transforms. These transforms should compensate for the noise, allowing a "clean" acoustic model to be trained on noise corrupted data. These "clean" models can then be adapted to a particular test condition. This yields a pure canonical model of speech compared to multi-style training where the models incorporate all the variability of the acoustic data. Adaptive training is usually based on linear transforms, in particular constrained maximum likelihood linear regression (CMLLR) [2], as minimal changes to the standard code are required for estimating the canonical model. However, these forms of adaptive training do not deal specifically with data with high levels of background noise. For low SNR conditions the estimates of the clean may be highly noisy and should not be treated with the same confidence as high SNR data.

Noise adaptive training schemes, such as [3], make use of feature-enhancement approaches to handle varying noise conditions in the training data. In the same fashion as CMLLR, these techniques do not alter the level of confidence in the estimate of the clean speech to reflect the noise conditions, possibly limiting performance when trying to deal with low-SNR data. To overcome the limitation of the feature-enhancement based approaches, model-based adaptive training schemes have been proposed for data with high levels of background noise using joint uncertainty decoding (JUD) [4] or vector Taylor series (VTS) [5]. Both schemes are closely related to one another as when the number of regression classes used with JUD is equal to the number of components VTS and JUD become equivalent to one another. However these adaptive training schemes differ from one another in how the canonical model is trained. In [4] a second-order gradient descent-based schemes was described. Alternatively an expectation-maximisation (EM) approach is used in [5].

In this paper, a new approach for adaptive training on noise-corrupted training data is presented. First, a modified version of generative model relating the "clean" speech and the observation is proposed. This allows a noise term to be included. This generative model can be expressed as a linear transform of the observed features and a bias on the variance (thus it may be viewed as combining CMLLR with the variance bias described in [6]) and will be referred to as noisy CMLLR (NCMLLR). Second, EM-based canonical model estimation formulae are derived using NCMLLR. Given the relationship between JUD and VTS these are closely related to those given in [5].

2. Noisy CMLLR

This section presents a general probabilistic model for the corrupted speech in the feature space. First, assume that the corrupted speech observation $\mathbf{o}_t$ can be written as a generative model of the clean speech vectors $\mathbf{s}_t$ in the form

$$\mathbf{o}_t = \mathbf{H}\mathbf{s}_t + \mathbf{g} + \mathbf{n}_t \tag{1}$$

where $\mathbf{H}$ is a linear transform and $\mathbf{g}$ is a bias on the clean speech[1] and $\mathbf{n}_t$ is a zero-mean Gaussian additive noise with covariance matrix $\mathbf{\Psi}$ such that $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$. Furthermore, the clean speech $\mathbf{s}_t$ is assumed to be generated by state θ_t of an HMM. The speech acoustic model $\mathcal{M}$ consists of Gaussian components each defined by a prior, c_m, mean, $\boldsymbol{\mu}_{\mathbf{s}}^{(m)}$, and diagonal covariance matrix, $\mathbf{\Sigma}_{\mathbf{s}}^{(m)}$, so

$$p(\mathbf{s}_t|\mathcal{M}, \theta_t) = \sum_{m \in \theta_t} c_m \mathcal{N}(\mathbf{s}_t; \boldsymbol{\mu}_{\mathbf{s}}^{(m)}, \mathbf{\Sigma}_{\mathbf{s}}^{(m)}) \; . \tag{2}$$

Thanks to Toshiba for suppling in-car test data

[1]For simplicity of notation a single transform is assumed per the homogeneous block. The extension to multiple transforms using regression class tree is trivial.

The corrupted speech observation $\mathbf{o}_t$ at time t is assumed to be conditional independent of all other observations given the clean speech and the noise at that time. Then the corrupted speech likelihood for a component m can be expressed as

$$p(\mathbf{o}_t|\mathcal{M},m) = \mathcal{N}\left(\mathbf{o}_t; \mathbf{H}\boldsymbol{\mu}_{\mathsf{s}}^{(m)} + \mathbf{g}, \mathbf{H}\boldsymbol{\Sigma}_{\mathsf{s}}^{(m)}\mathbf{H}^{\mathsf{T}} + \boldsymbol{\Psi}\right) . \tag{3}$$

This form of likelihood expression is closely related to forms of the shared factor analysis (FA) model, e.g. [7]. These have the same general form as the generative model in (1). However there are a couple of important differences. First, NCMLLR utilises the shared loading matrices and noise variances as transforms rather than as a method for covariance matrix modelling. Second the dimensions of the observation and latent variable space are required to be the same for NCMLLR. Though a restriction, this allows some computational advantages for NCMLLR. Equation 3 can be rewritten as

$$p(\mathbf{o}_t|\mathcal{M},m) = |\mathbf{A}|\mathcal{N}\left(\mathbf{A}\mathbf{o}_t+\mathbf{b}; \boldsymbol{\mu}_{\mathsf{s}}^{(m)}, \boldsymbol{\Sigma}_{\mathsf{s}}^{(m)}+\boldsymbol{\Sigma}_{\mathsf{b}}\right) \tag{4}$$

where $\mathbf{A} = [\mathbf{H}]^{-1}$, $\mathbf{b} = -[\mathbf{H}]^{-1}\mathbf{g}$ and $\boldsymbol{\Sigma}_{\mathsf{b}} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A}^{\mathsf{T}}$. Normally FA-style models are computationally expensive when calculating the log-likelihood. Depending on the dimensionality of the latent this can be the cost of a full-covariance matrix calculation. In contrast for NCMLLR the variance bias can be restricted to be diagonal, so that if the speech covariance matrix is diagonal, diagonal log-likelihood calculations can be performed.

NCMLLR is also related to CMLLR [2] where

$$p(\mathbf{o}_t|\mathcal{M},m) = |\mathbf{A}|\mathcal{N}\left(\mathbf{A}\mathbf{o}_t+\mathbf{b}; \boldsymbol{\mu}_{\mathsf{s}}^{(m)}, \boldsymbol{\Sigma}_{\mathsf{s}}^{(m)}\right) \tag{5}$$

CMLLR can also be expressed as a generative model where

$$\mathbf{o}_t = \mathbf{H}\mathbf{s}_t + \mathbf{g}; \quad \mathbf{s}_t = \mathbf{A}\mathbf{o}_t + \mathbf{b}$$

Both CMLLR and NCMLLR have an affine transform of the feature, but in NCMLLR (4) there is an additional variance bias, $\boldsymbol{\Sigma}_{\mathsf{b}}$, for modelling the changes in the variance of the corrupted speech due to noise $\mathbf{n}_t$, hence the name. NCMLLR has the same form as the JUD transform. However here the parameters of the NCMLLR transform are estimated in a maximum likelihood (ML) fashion rather than being based on a mismatch function describing the impact of noise on the clean speech and noise model estimates. As it is not necessary to specify a mismatch function allowing more complex forms of front-end processing to be used with NCMLLR than JUD.

3. Adaptive Training with NCMLLR

In this section, the ML estimation of the NCMLLR parameters and their use in adaptive training is described. Adaptive training with NCMLLR follows the general adaptive training framework [1, 2]. The canonical acoustic model parameter $\mathcal{M}$ and set of NCMLLR parameters $\mathcal{T}$ are estimated such that they maximise the likelihood of the heterogenous training data comprised of H homogeneous block, $\mathcal{O} = \{\mathbf{O}^{(1)}, \ldots, \mathbf{O}^{(H)}\}$.

In a similar fashion to the EM approach described in [8], the clean speech vectors are considered to be hidden in the NCMLLR model. Let $\mathcal{M}$ and $\mathcal{T}$ are the current model and set of NCMLLR parameters, $\{\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(H)}\}$, and $\hat{\mathcal{M}}$ and $\hat{\mathcal{T}}$ the model and set of NCMLLR parameters to be estimated. Following [9], the auxiliary function may be expressed as[2]

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\mathcal{T};\hat{\mathcal{M}},\hat{\mathcal{T}}) = K - \frac{1}{2}\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\sum_{m=1}^{M}\gamma_{\mathsf{o},t}^{(mh)}\mathbb{E}\Big[\log|\hat{\boldsymbol{\Psi}}^{(h)}| \\ + (\mathbf{o}_t-\hat{\mathbf{H}}^{(h)}\mathbf{s}_t-\hat{\mathbf{g}}^{(h)})^{\mathsf{T}}\hat{\boldsymbol{\Psi}}^{(h)\text{-}1}(\mathbf{o}_t-\hat{\mathbf{H}}^{(h)}\mathbf{s}_t-\hat{\mathbf{g}}^{(h)}) \\ + \log|\hat{\boldsymbol{\Sigma}}_{\mathsf{s}}^{(m)}| + (\mathbf{s}_t-\hat{\boldsymbol{\mu}}_{\mathsf{s}}^{(m)})^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_{\mathsf{s}}^{(m)\text{-}1}(\mathbf{s}_t-\hat{\boldsymbol{\mu}}_{\mathsf{s}}^{(m)})|\mathbf{o}_t,m\Big]\end{aligned} \tag{6}$$

where $\gamma_{o,t}^{(mh)}$ is the posterior probability of component m given the observation sequence $\mathbf{O}^{(h)}$, NCMLLR parameter set $\mathcal{T}^{(h)}$, and model set $\mathcal{M}$ for homogenous block h. This auxiliary function will be iteratively optimised by first updating the NCMLLR transform parameters then the canonical model parameters.

3.1. NCMLLR Transform Estimation

First the NCMLLR transform parameters for each homogeneous block h are estimated. These can either be estimated using $\{\mathbf{A}^{(h)}, \mathbf{b}^{(h)}, \boldsymbol{\Sigma}_{\mathsf{b}}^{(h)}\}$, or the related $\{\mathbf{H}^{(h)}, \mathbf{g}^{(h)}, \boldsymbol{\Psi}^{(h)}\}$. The transform parameters are found by maximising the auxiliary function in (6) with respect to the transform parameters. Rather than optimising $\mathbf{H}^{(h)}$ and $\mathbf{g}^{(h)}$ separately, they are combined into an extended transformation matrix $\hat{\mathbf{V}}^{(h)} = [\hat{\mathbf{g}}^{(h)} \;\; \hat{\mathbf{H}}^{(h)}]$ applied to an extended clean vector $\boldsymbol{\zeta}_t^{\mathsf{T}} = [1 \;\; \mathbf{s}_t^{\mathsf{T}}]$.

Initially considering the variance bias. Differentiating (6) with respect to $\hat{\boldsymbol{\Psi}}^{(h)}$ and equating to zero yields

$$\hat{\boldsymbol{\Psi}}^{(h)} = \frac{\sum_{m,t}\gamma_{\mathsf{o},t}^{(mh)}\mathbb{E}\left[(\mathbf{o}_t - \hat{\mathbf{V}}^{(h)}\boldsymbol{\zeta}_t)(\mathbf{o}_t - \hat{\mathbf{V}}^{(h)}\boldsymbol{\zeta}_t)^{\mathsf{T}}|\mathbf{o}_t,m\right]}{\sum_{m,t}\gamma_{\mathsf{o},t}^{(mh)}} \tag{7}$$

If used directly this could yield a full variance bias $\hat{\boldsymbol{\Sigma}}_{\mathsf{b}}^{(h)}$, resulting in a full covariance-matrix likelihood calculation cost. Instead using the equality $\hat{\boldsymbol{\Sigma}}_{\mathsf{b}}^{(h)} = \hat{\mathbf{A}}^{(h)}\hat{\boldsymbol{\Psi}}^{(h)}\hat{\mathbf{A}}^{(h)\mathsf{T}}$, an ML diagonal estimate of the variance bias can be found using[3]

$$\hat{\boldsymbol{\Sigma}}_{\mathsf{b}}^{(h)} = \text{diag}\left(\frac{\sum_{m,t}\gamma_{\mathsf{o},t}^{(mh)}\mathbb{E}\left[(\hat{\mathbf{s}}_t^{(h)} - \mathbf{s}_t)(\hat{\mathbf{s}}_t^{(h)} - \mathbf{s}_t)^{\mathsf{T}}|\mathbf{o}_t,m\right]}{\sum_{m,t}\gamma_{\mathsf{o},t}^{(m)}}\right) \tag{8}$$

where $\hat{\mathbf{s}}_t^{(h)} = \hat{\mathbf{A}}^{(h)}\mathbf{o}_t + \hat{\mathbf{b}}^{(h)}$. Thus $\hat{\boldsymbol{\Sigma}}_{\mathsf{b}}^{(h)}$ is estimated given $\hat{\mathbf{A}}^{(h)}$ and $\hat{\mathbf{b}}^{(h)}$. Differentiating (6) with respect to $\hat{\mathbf{V}}^{(h)}$, and equating to zero, gives

$$\hat{\mathbf{V}}^{(h)} = \left(\sum_{m,t}\gamma_{\mathsf{o},t}^{(mh)}\,\mathbf{o}_t\,\mathbb{E}\Big[\boldsymbol{\zeta}_t^{\mathsf{T}}|\mathbf{o}_t,m\Big]\right)\left(\sum_{m,t}\gamma_{\mathsf{o},t}^{(mh)}\mathbb{E}\Big[\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t^{\mathsf{T}}|\mathbf{o}_t,m\Big]\right)^{\text{-}1} \tag{9}$$

Both (8) and (9) can be expressed in terms of the conditional expectations of the extended clean speech vector

$$\mathbb{E}\left[\boldsymbol{\zeta}_t|\mathbf{o}_t,m\right] = [1 \;\; \mathbb{E}[\mathbf{s}_t^{\mathsf{T}}|\mathbf{o}_t,m]]^{\mathsf{T}} = [1 \;\; \tilde{\mathbf{s}}_t^{(mh)\mathsf{T}}]^{\mathsf{T}} \tag{10}$$

$$\mathbb{E}\left[\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t^{\mathsf{T}}|\mathbf{o}_t,m\right] = \begin{bmatrix} 1 & \tilde{\mathbf{s}}_t^{(mh)\mathsf{T}} \\ \tilde{\mathbf{s}}_t^{(mh)} & \tilde{\boldsymbol{\Sigma}}_{\mathsf{s}}^{(mh)} + \tilde{\mathbf{s}}_t^{(mh)}\tilde{\mathbf{s}}_t^{(mh)\mathsf{T}} \end{bmatrix} \tag{11}$$

where

$$\tilde{\mathbf{s}}_t^{(mh)} = \tilde{\mathbf{A}}^{(mh)}\mathbf{o}_t + \tilde{\mathbf{b}}^{(mh)};\; \tilde{\boldsymbol{\Sigma}}_{\mathsf{s}}^{(mh)} = (\boldsymbol{\Sigma}_{\mathsf{s}}^{(m)\text{-}1} + \boldsymbol{\Sigma}_{\mathsf{b}}^{(h)\text{-}1})^{\text{-}1} \tag{12}$$

[2] The dependence on the model, $\mathcal{M}$, and transform parameters, $\mathcal{T}$, are dropped for notational simplicity.

[3] It is not necessary to use diagonal versions of the variance bias, but this makes the likelihood calculations efficient, see [9] for details. This option is not normally possible with FA-based approaches.

and

$$\tilde{\mathbf{A}}^{(mh)} = \left(\boldsymbol{\Sigma}_{\mathtt{s}}^{(m)\text{-}1} + \boldsymbol{\Sigma}_{\mathtt{b}}^{(h)\text{-}1}\right)^{\text{-}1} \boldsymbol{\Sigma}_{\mathtt{b}}^{(h)} \mathbf{A}^{(h)} \quad (13)$$

$$\tilde{\mathbf{b}}^{(mh)} = \left(\boldsymbol{\Sigma}_{\mathtt{s}}^{(m)\text{-}1} + \boldsymbol{\Sigma}_{\mathtt{b}}^{(h)\text{-}1}\right)^{\text{-}1} \left(\boldsymbol{\Sigma}_{\mathtt{s}}^{(m)\text{-}1}\boldsymbol{\mu}_{\mathtt{s}}^{(m)} + \boldsymbol{\Sigma}_{\mathtt{b}}^{(h)\text{-}1}\mathbf{b}^{(h)}\right)$$

Thus $\hat{\mathbf{V}}^{(h)}$ can be found given $\hat{\boldsymbol{\Sigma}}_{\mathtt{b}}^{(h)}$. Hence $\hat{\mathbf{A}}^{(h)}$ and $\hat{\mathbf{b}}^{(h)}$ can be obtained using for example $\mathbf{A}^{(h)} = [\mathbf{H}^{(h)}]^{\text{-}1}$. NCMLLR transform estimation is itself an iterative process, interleaving estimates of $\hat{\boldsymbol{\Sigma}}_{\mathtt{b}}^{(h)}$, and $\hat{\mathbf{A}}^{(h)}$ and $\hat{\mathbf{b}}^{(h)}$.

3.2. Canonical Model Parameter Estimation

After a new set of NCMLLR transform parameters have been estimated, the canonical model parameters must be retrained. The auxiliary function in (6) can again be used. Differentiating (6) with respect to $\hat{\boldsymbol{\mu}}_{\mathtt{s}}^{(m)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathtt{s}}^{(m)}$, leads to the update formulae as follows:

$$\hat{\boldsymbol{\mu}}_{\mathtt{s}}^{(m)} = \frac{\sum_{h,t} \gamma_{\mathtt{o},t}^{(mh)} \mathbb{E}\left[\mathbf{s}_t | \mathbf{o}_t, m\right]}{\sum_{h,t} \gamma_{\mathtt{o},t}^{(mh)}} \quad (14)$$

$$\hat{\boldsymbol{\Sigma}}_{\mathtt{s}}^{(m)} = \text{diag}\left(\frac{\sum_{h,t} \gamma_{\mathtt{o},t}^{(mh)} \mathbb{E}\left[\mathbf{s}_t \mathbf{s}_t^{\mathsf{T}} | \mathbf{o}_t, m\right]}{\sum_{h,t} \gamma_{\mathtt{o},t}^{(m)}} - \hat{\boldsymbol{\mu}}_{\mathtt{s}}^{(m)} \hat{\boldsymbol{\mu}}_{\mathtt{s}}^{(m)\mathsf{T}}\right)$$

The conditional expectations in the above equations are given in (10) and (11), so for example $\mathbb{E}\left[\mathbf{s}_t \mathbf{s}_t^{\mathsf{T}} | \mathbf{o}_t, m\right] = \tilde{\boldsymbol{\Sigma}}_{\mathtt{s}}^{(mh)} + \tilde{\mathbf{s}}_t^{(mh)} \tilde{\mathbf{s}}_t^{(mh)\mathsf{T}}$.

The above expressions have described maximum-likelihood estimation of the canonical model parameters. It is also possible to perform discriminative, here minimum phone error (MPE) [10], training of the model parameters. For MPE training the following criterion is minimised

$$\mathcal{F}_{\mathtt{mpe}}(\mathcal{M}) = \sum_{h=1}^{H} \sum_{\mathcal{H}} P(\mathcal{H}|\mathbf{O}^{(h)}, \mathcal{T}^{(h)}, \mathcal{M}) \mathcal{L}(\mathcal{H}, \mathcal{H}_{\mathtt{ref}}^{(h)}) \quad (15)$$

where $\mathcal{L}(\mathcal{H}, \mathcal{H}_{\mathtt{ref}}^{(h)})$ is the "loss" measured at the phone-level between the hypothesis and reference $\mathcal{H}_{\mathtt{ref}}^{(h)}$. Using a weak-sense auxiliary function [10] the MPE estimate of the mean is given by

$$\hat{\boldsymbol{\mu}}_{\mathtt{s}}^{(m)} = \frac{\sum_{h,t} (\gamma_{\mathtt{no},t}^{(mh)} - \gamma_{\mathtt{do},t}^{(mh)}) \mathbb{E}\left[\mathbf{s}_t | \mathbf{o}_t, m\right] + D_m \boldsymbol{\mu}_{\mathtt{s}}^{(m)} + \tau_{\mathtt{p}} \boldsymbol{\mu}_{\mathtt{p}}^{(m)}}{\sum_{h,t} (\gamma_{\mathtt{no},t}^{(mh)} - \gamma_{\mathtt{do},t}^{(mh)}) + D_m + \tau_{\mathtt{p}}}$$

where $\gamma_{\mathtt{no},t}^{(mh)}$ and $\gamma_{\mathtt{do},t}^{(mh)}$ are the numerator and denominator "posteriors", D_m is the component-specific smoothing constant, and $\boldsymbol{\mu}_{\mathtt{p}}^{(m)}$ and τ_p are the I-smoothing prior and constant respectively. For this work the MMI-estimate is used as the I-smoothing prior. In common with other discriminative adaptive training schemes, e.g. [11], only the canonical models were trained using MPE given ML transform estimates. For further details see [9].

3.3. Implementation Issues

There are a number of issues that must be considered when using NCMLLR, either as a transform in themselves or in adaptive training. When using full linear transforms, $\mathbf{A}^{(h)}$, it is necessary to store full outer-product observation statistics for each component. Equation 9 requires the term $\mathbf{o}_t \mathbb{E}[\boldsymbol{\zeta}_t^{\mathsf{T}} | \mathbf{o}_t, m]$ which from (12) needs functions of $\mathbf{o}_t \mathbf{o}_t^{\mathsf{T}}$ for each component. Terms like this are not required for CMLLR estimation as observation outer-products can be accumulated at the base-class level. As all components tend not to be observed when estimating a particular transform, this is practical even for large vocabulary systems by only generating accumulation space on demand. This is not an issue when using diagonal transforms. Note unlike CMLLR estimation, the transform update formulae are applicable with both diagonal and full covariance canonical models.

Another problem, also observed with JUD compensation, is that the magnitude of the transform matrix and hence the variance bias, both become very large in low SNR regions. Because the corrupted speech distribution is dominated by the noise in the region of low speech energy, the cross-covariance $\boldsymbol{\Sigma}_{\mathtt{so}}$ will be approximately zero. That is, the clean speech and the corrupted speech will be uncorrelated since the clean speech and noise are independent. The cross-covariance term $\boldsymbol{\Sigma}_{\mathtt{so}}^{(m)}$ for component m in these low SNR regions is given by

$$\boldsymbol{\Sigma}_{\mathtt{so}}^{(m)} = \mathbb{E}\left[(\mathbf{s}_t - \boldsymbol{\mu}_{\mathtt{s}}^{(m)})(\mathbf{o}_t - \boldsymbol{\mu}_{\mathtt{o}}^{(m)})^{\mathsf{T}}\right] = \boldsymbol{\Sigma}_{\mathtt{s}}^{(m)} \mathbf{H}^{\mathsf{T}} \approx \mathbf{0}$$

In the NCMLLR scheme, the transform matrix $\mathbf{H}$ will tend to zero in low SNR, hence $\mathbf{A}$ goes to infinity along with the variance bias $\boldsymbol{\Sigma}_{\mathtt{b}}$. To prevent this problem it is sensible to limit the possible values for the compensation parameters. The compensation parameters can then be restricted by enforcing a maximum on the variance bias for dimension i used in (12) and (13) so that

$$\sigma_{\mathtt{b}i}^{(h)2} \leq \rho \cdot \sigma_{\mathtt{s}i}^{(m)2} \quad (16)$$

where ρ is an empirically determined constant. Performance was found to be relatively insensitive to ρ over a range of values for each of the tasks examined.

4. Experiments and Results

The use of NCMLLR transforms for use in adaptation and adaptive training was evaluated on two tasks, noise corrupted Resource Management (RM) and in-car data collected by Toshiba. Though a wide-range of possible contrasts are possible, the tasks were configured so that there was sufficient data for linear transforms to be robustly estimated. Thus NCMLLR was compared with CMLLR as a standard scheme for adaptation and adaptive training.

Initial experiments were conducted on the medium vocabulary speech recognition task, the 1000 word Resource Management (RM) database. The 39 dimensional feature vector consists of MFCCs, including the 0th cepstra, and associated 1st- and 2nd-order coefficients. Cross-word, state-clustered triphone acoustic models with 6 components per state were used along with a simple word pair grammar. Operations Room noise from the NOISEX-92 was artificially added to at the waveform level database to give 20dB and 14dB SNR test sets. All results are averaged across the FEB89, OCT89 and FEB91 test sets. The multi-style model was built from data with Operations Room noise added at the speaker level at SNRs of 8, 14, 20, 26 or 32 dB. This was also used as the initial model to begin adaptive training. 16 regression classes were used for diagonal linear transforms, while a single regression class for a global transform[4]. The variance bias $\boldsymbol{\Sigma}_{\mathtt{b}}$ was restricted to be diagonal. For initial iteration of EM algorithm, $\mathbf{A} = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$ and large diagonal biases were used as the initial values.

Table 1 shows the performance of both NCMLLR adaptation and use in adaptive training compared to CMLLR with both

[4]Using multiple full-transforms did not yield performance gains.

System	Adapt	20db		14dB	
		diag	full	diag	full
Multi-style	—	7.0		15.4	
	CMLLR	5.8	5.6	13.0	13.4
	NCMLLR	6.4	6.0	12.1	11.6
Adaptive Training	CMLLR	5.3	4.3	12.0	10.5
	NCMLLR	5.0	4.5	9.9	9.1

Table 1: Performance of multi-style and adaptive training with 16-diagonal, or 1-full, transform CMLLR and NCMLLR on the RM task.

diagonal and full transforms. For the low SNR condition, 14dB SNR, NCMLLR out-performed CMLLR. This is expected as NCMLLR allows a bias term to model additional uncertainty. If the SNR was increased to 20dB, CMLLR out-performed NCMLLR for the multi-style systems. For multi-style trained systems it is unclear what the variance will model as the acoustic models themselves model speech and noise. For adaptive training, the performance was mixed at 20dB, though the best performance was obtained with the full CMLLR adaptively trained system.

The Toshiba task is a small/medium sized task with noisy speech collected in the office and in cars driving at various conditions. For this work two in-car test sets consisting of phone numbers were used. This phone-number task comprises unknown-length digit sequences. The performance was evaluated on two different conditions: engine on (ENON) and highway driving (HWY). The average SNRs these tests are 35dB and 18 dB, respectively. Similar features and model topology to the RM system were used, except normalised log energy instead of 0th cepstra. The WSJ SI284 training data was used to train a multi-style system model. Noise-corrupted multi-condition data were generated by adding car noise at the speaker level at average SNRs of 15, 18, 25 and 35 dB. The noise added during train was different to that of the test data. About 6400 distinct states were generated with 16 Gaussian components per state. Only diagonal CMLLR/NCMLLR transforms were used for this task. Both ML and MPE canonical models were trained (there is not sufficient data to robustly train MPE system on the RM tasks). For additional information about the task and training procedure see [9].

System	Adapt (diag)	ENON		HWY	
		ML	MPE	ML	MPE
Multi-style	—	1.2	0.8	6.7	5.0
	CMLLR	0.3	0.3	2.4	2.0
	NCMLLR	0.5	0.6	2.1	1.9
Adaptive Training	CMLLR	0.3	0.2	2.1	1.5
	NCMLLR	0.3	0.2	1.8	1.2

Table 2: Performance of ML and MPE trained multi-style and adaptive training with 16-diagonal transforms CMLLR and NCMLLR compensation on Toshiba in-car phone-number task.

Table 2 shows the results on the phone-number task for ML multi-style and adaptive training models with 16-diagonal transforms. For multi-style systems in the HWY condition, NCMLLR was better than CMLLR. However NCMLLR performed worse than CMLLR in the ENON condition. This is similar to patterns obtained on RM where at high SNR conditions and multi-style. For the systems trained with adaptive training NCMLLR outperformed, or matched, CMLLR for all conditions. For the HWY condition NCMLLR was significantly, using the matched-pair test, different to the CMLLR system.

Table 2 also shows the performance of MPE-trained systems. As expected performance gains were obtained over the equivalent ML-systems. The same general trends as for ML-training can be observed with NCMLLR performing better at the lower SNR conditions. For NCMLLR the gains from using MPE training with a multi-style trained system were small, less than 10% relative on HWY, whereas for adaptive training the gain was over 30% relative. Though not as extreme, the same is true for CMLLR where there a 19% relative reduction using MPE over ML for the multi-style system. The gain was 28% for the adaptively trained systems.

5. Conclusions

This paper has described adaptive training using NCMLLR for handling noise-corrupted training data. NCMLLR is an extension to CMLLR that allows the additional uncertainty that results from noisy data to be modelled. ML training of NCMLLR transforms and their use for adaptive training with both the ML and MPE trained canonical models are described. Two databases were used for assessing NCMLLR, a noise corrupted version of the RM and in-car recorded data. On both tasks model compensation and adaptive training with NCMLLR outperformed the traditional CMLLR scheme at lower SNR. The useful of adaptive training when using discriminative training is also demonstrated. Future work will examine using the ML and MPE canonical model training described here in the joint adaptive training framework, as JUD and NCMLLR have the same form.

6. References

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

[2] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, Jan. 1998.

[3] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 806–809.

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

[5] 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.

[6] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.

[7] R. Gopinath, B. Ramabhadran, and S. Dharanipragada, "Factor analysis invariant to linear transformations of data," in *Proc. ICSLP*, 1998, pp. 397–400.

[8] R. Rose, E. Hofstetter, and D. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. on Speech and Audio Processing*, 1994.

[9] D. Kim and M. Gales, "Noisy CMLLR for noise-robust speech recognition," University of Cambridge, Tech. Rep. TR611, February 2009, available from http://mi.eng.cam.ac.uk/∼mjfg/.

[10] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2003.

[11] L. Wang and P. Woodland, "Discriminative Adaptive Training using the MPE Criterion," in *Proc ASRU*, St Thomas, US Virgin Islands, 2003.