INCREMENTAL PREDICTIVE AND ADAPTIVE NOISE COMPENSATION

F. Flego and M.J.F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge CB2 1PZ, U.K.
{ff257,mjfg}@eng.cam.ac.uk

ABSTRACT

Model compensation schemes are a powerful approach to handling mismatches between training and testing conditions. Normally these schemes are run in a batch adaptation mode, re-recognising the utterance used to estimate the noise model parameters. For many applications this introduces unacceptable latency. This paper examines three forms of incremental mode model-based compensation: vector Taylor series; joint uncertainty decoding; and predictive CMLLR. These predictive schemes can also be combined with adaptive schemes such as CMLLR. By combining the approaches, weaknesses of each can be addressed. The performance is evaluated on in-car recorded data, where the combined incremental scheme shows gains over either individually.

Index Terms: noise robustness, speaker adaptation

1. INTRODUCTION

Speech recognition in noise has been an area of active research for many years. Good performance using model-based compensation schemes, such as vector Taylor series (VTS) [1] and joint uncertainty decoding (JUD) [2], can be obtained. These *predictive* approaches make use of a mismatch function that represents the impact of the background noise on the clean speech. The number of parameters associated with this mismatch function is usually small, the additive noise distribution and an estimate of the convolutional distortion. This is in contrast to *adaptive* approaches to speaker and noise compensation where, normally, linear transforms of the model parameters are estimated. These adaptive approaches make no assumptions about the underlying nature of the mismatch. For non-linear mismatches, such as the impact of noise in the mel-cepstral domain, a large number of transforms and associated parameters must be estimated.

Adaptation is normally run in either a batch or incremental mode. In incremental mode adaptation the data is assumed to arrive in a causal fashion and hypotheses also generated in a causal fashion. For batch adaptation all data to be decoded is available in a single block. Recognition and adaptation can be run repeatedly on the same block before the final output is generated. Model noise compensation is often only described in a batch adaptation, though sometimes the size of the batch may be a single utterance. In some application areas, such is in-car embedded speech recognition, batch mode adaptation would cause too great a latency to be useful. Though the modifications to the theory required for incremental predictive adaptation are small, basically the statistics are accumulated in an incremental fashion, incremental adaptation allows interesting contrasts and combinations of predictive and adaptive schemes to be used. In this work an approach for combining adaptive and predictive noise compensation schemes is described. By combining the two schemes weaknesses of both the schemes for noise robust speech recognition are addressed: rapidness and accuracy for adaptive schemes; accuracy of the mismatch function and speaker modelling for the predictive schemes.

This work was partly funded by Toshiba Research Europe Ltd.

2. ADAPTIVE COMPENSATION

Adaptation is commonly used to compensate for different speakers and noise conditions. The most popular forms of rapid adaptation are based on linear transforms, for example maximum likelihood linear regression (MLLR) [3] and constrained MLLR (CMLLR) [4]. CMLLR is often used as for large systems it is computationally efficient as it can be implemented as a (set of) linear transform of the features. The general form of the transform is

$$p(\boldsymbol{y}_t|\mathtt{s}_m) = |\mathbf{A}^{(r)}|\mathcal{N}(\mathbf{A}^{(r)}\boldsymbol{y}_t + \mathbf{b}^{(r)}; \boldsymbol{\mu}_{\mathsf{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathsf{x}}^{(m)}), \tag{1}$$

where $\boldsymbol{\mu}_{\mathsf{x}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathsf{x}}^{(m)}$ are the mean and covariance matrix of component $\mathtt{s}_m$ of the clean system[1]. In this expression the component is assumed to belong to regression class r. By using multiple regression classes it is possible to handle non-linear transformations. This is important when used to compensate for noise as the impact of noise on the clean speech is highly non-linear, see equation 6.

The transform parameters are normally estimated in a maximum likelihood (ML) fashion. In this work the majority of the transforms considered are diagonal as when adapting clean models more diagonal transforms provide better results than a limited number of full transforms. This is not the case for standard speaker adaptation [5] or when adapting multi-style models. For the diagonal case the transform can be estimated non-iteratively using [4]

$$\begin{bmatrix} b_i^{(r)} & a_i^{(r)} \end{bmatrix} = \left(\alpha\mathbf{p}_i + \mathbf{k}_i^{(r)}\right)\mathbf{G}_i^{(r)\text{-}1} \tag{2}$$

where $\mathbf{p}_i = [0\ \ 1]$, α satisfies the quadratic expression

$$\alpha^2\mathbf{p}_i\mathbf{G}_i^{(r)\text{-}1}\mathbf{p}_i^{\mathsf{T}} + \alpha\mathbf{p}_i\mathbf{G}_i^{(r)\text{-}1}\mathbf{k}_i^{(r)\mathsf{T}} - \beta^{(r)} = 0 \tag{3}$$

where $\beta^{(r)} = \sum_{m\in\mathtt{r}_r}\sum_{t=1}^{T}\gamma_{\mathsf{y}}^{(m)}(t)$ and

$$\mathbf{G}_i^{(r)} = \sum_{m\in\mathtt{r}_r}\frac{1}{\sigma_{\mathsf{x}i}^{(m)2}}\sum_{t=1}^{T}\gamma_{\mathsf{y}}^{(m)}(t)\begin{bmatrix}1 & y_{ti}\\ y_{ti} & y_{ti}^2\end{bmatrix} \tag{4}$$

$$\mathbf{k}_i^{(r)} = \sum_{m\in\mathtt{r}_r}\frac{\mu_{\mathsf{x}i}^{(m)}}{\sigma_{\mathsf{x}i}^{(m)2}}\sum_{t=1}^{T}\gamma_{\mathsf{y}}^{(m)}(t)\begin{bmatrix}1 & y_{ti}\end{bmatrix} \tag{5}$$

[1] In this work a distinction will be made between the parameters of a "clean" model, for example $\boldsymbol{\mu}_{\mathsf{x}}^{(m)}$, and the parameters associated with (or modified to reflect) the noise corrupted models,for example $\boldsymbol{\mu}_{\mathsf{y}}^{(m)}$.

$\gamma_{\mathtt{y}}^{(m)}(t)$ is the posterior probability that component $\mathtt{s}_m$ generated the observation $\boldsymbol{y}_t$ at time t given the complete observation sequence $\mathbf{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$. For incremental adaptation, the current transform is used to recognize the next utterance. Statistics are then accumulated for this recently decoded utterance and a new transform estimated.

3. PREDICTIVE NOISE COMPENSATION

The previous section has described CMLLR, which in this work will be referred to as an adaptive transform. This section will describe predictive transforms. In predictive transforms the relationship between the clean and corrupted speech distributions can be parameterised by a mismatch function and underlying noise models. For noise robust speech recognition, the standard form of static mismatch function in the mel-cepstral domain will be used

$$\begin{aligned}\boldsymbol{y}_t^{\mathtt{s}} &= \boldsymbol{x}_t^{\mathtt{s}} + \boldsymbol{h} + \mathbf{C}\log\left(\mathbf{1} + \exp(\mathbf{C}^{-1}(\boldsymbol{n}_t^{\mathtt{s}} - \boldsymbol{x}_t^{\mathtt{s}} - \boldsymbol{h}))\right) \\ &= \boldsymbol{x}_t^{\mathtt{s}} + \boldsymbol{h} + \boldsymbol{f}(\boldsymbol{n}_t^{\mathtt{s}} - \boldsymbol{x}_t^{\mathtt{s}} - \boldsymbol{h}) \end{aligned} \quad (6)$$

where $\mathbf{C}$ is the DCT matrix. The complete features used in this work and many other systems is then $\boldsymbol{y}_t^{\mathsf{T}} = [\boldsymbol{y}_t^{\mathtt{s}\mathsf{T}}\ \boldsymbol{y}_t^{\Delta\mathsf{T}}\ \boldsymbol{y}_t^{\Delta^2\mathsf{T}}]$. Once the noise model parameters, the noise mean and covariance matrix $\boldsymbol{\mu}_{\mathtt{n}}$, $\boldsymbol{\Sigma}_{\mathtt{n}}$, and the convolutional noise $\boldsymbol{\mu}_{\mathtt{h}}$, are known it is possible to transform the model parameters, or obtain transformations of the features. In practice these values are seldom known in advance so must be estimated from the test data. All these parameters can be estimated using Maximum Likelihood (ML) noise estimation [6]. In this work the parameters are estimated in an incremental fashion, rather than the standard batch mode in [6].

3.1. VTS

Vector Taylor series model-based compensation is a popular approach for model-based compensation [1, 6, 7]. There are a number of possible forms that have been examined. In this work the first-order VTS scheme described in [6] is used. A brief summary of the scheme is given here. The static mean, $\boldsymbol{\mu}_{\mathtt{y}}^{\mathtt{s}}$, and covariance matrix, $\boldsymbol{\Sigma}_{\mathtt{y}}^{\mathtt{s}}$, of the corrupted speech distribution are given by [7][2]

$$\boldsymbol{\mu}_{\mathtt{y}}^{\mathtt{s}} = \boldsymbol{\mu}_{\mathtt{x}}^{\mathtt{s}} + \boldsymbol{\mu}_{\mathtt{h}} + \boldsymbol{f}(\boldsymbol{\mu}_{\mathtt{n}}^{\mathtt{s}} - \boldsymbol{\mu}_{\mathtt{x}}^{\mathtt{s}} - \boldsymbol{\mu}_{\mathtt{h}}) \quad (7)$$

$$\boldsymbol{\Sigma}_{\mathtt{y}}^{\mathtt{s}} = \mathrm{diag}\left(\mathbf{J}\boldsymbol{\Sigma}_{\mathtt{x}}^{\mathtt{s}}\mathbf{J}^{\mathsf{T}} + (\mathbf{I}-\mathbf{J})\boldsymbol{\Sigma}_{\mathtt{n}}^{\mathtt{s}}(\mathbf{I}-\mathbf{J})^{\mathsf{T}}\right) \quad (8)$$

where matrix $\mathbf{J}$ above is the partial derivative, $\partial\boldsymbol{y}^{\mathtt{s}}/\partial\boldsymbol{x}^{\mathtt{s}}$, evaluated at $\overline{\boldsymbol{\mu}}^{\mathtt{s}} = \boldsymbol{\mu}_{\mathtt{n}}^{\mathtt{s}} - \boldsymbol{\mu}_{\mathtt{x}}^{\mathtt{s}} - \boldsymbol{\mu}_{\mathtt{h}}$. This may be expressed as

$$\mathbf{J} = \partial\boldsymbol{y}^{\mathtt{s}}/\partial\boldsymbol{x}^{\mathtt{s}} = \mathbf{C}\mathbf{F}\mathbf{C}^{-1} \quad (9)$$

where $\mathbf{F}$ is a diagonal matrix with elements given by $\mathbf{1}/(\mathbf{1} + \exp(2\mathbf{C}^{-1}(\overline{\boldsymbol{\mu}}^{\mathtt{s}}))$. For best performance all the model parameters need to be estimated . To obtain the expression for the compensated dynamic parameters (Δ, Δ^2) the *continuous time approximation* is used in this work providing

$$\boldsymbol{\mu}_{\mathtt{y}}^{\Delta} = \mathbf{J}\boldsymbol{\mu}_{\mathtt{x}}^{\Delta}; \quad \boldsymbol{\Sigma}_{\mathtt{y}}^{\Delta} = \mathrm{diag}\left(\mathbf{J}\boldsymbol{\Sigma}_{\mathtt{x}}^{\Delta}\mathbf{J}^{\mathsf{T}} + (\mathbf{I}-\mathbf{J})\boldsymbol{\Sigma}_{\mathtt{n}}^{\Delta}(\mathbf{I}-\mathbf{J})^{\mathsf{T}}\right) \quad (10)$$

where $\boldsymbol{\mu}_{\mathtt{y}}^{\Delta^2}$ and $\boldsymbol{\Sigma}_{\mathtt{y}}^{\Delta^2}$ have similar form. As each component is compensated separately the likelihood calculation involves

$$p(\boldsymbol{y}_t|\mathtt{s}_m) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathtt{y}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{y}}^{(m)}), \quad (11)$$

[2]The dependence on the noise corrupted speech mean and clean speech mean on the component have been dropped for clarity.

3.2. Joint Uncertainty Decoding

Though VTS has been shown to yield large reductions in word error rate (WER) the scheme is computationally expensive as each component is compensated individually. To reduce the computational load joint uncertainty decoding [2, 6] has been proposed. Here compensation parameters are computed at the regression base-class level. The approximate joint Gaussian distribution of the corrupted speech, $\boldsymbol{y}$ and the clean speech $\boldsymbol{x}$ is computed for the regression class. This can then be used in an uncertainty decoding frame work to yield

$$p(\boldsymbol{y}_t|\mathtt{s}_m) = |\mathbf{A}_{\mathtt{jud}}^{(r)}|\mathcal{N}(\mathbf{A}_{\mathtt{jud}}^{(r)}\boldsymbol{y}_t + \mathbf{b}_{\mathtt{jud}}^{(r)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)} + \boldsymbol{\Sigma}_{\mathtt{jud}}^{(r)}) \quad (12)$$

where $\{\mathbf{A}_{\mathtt{jud}}^{(r)}, \mathbf{b}_{\mathtt{jud}}^{(r)}, \boldsymbol{\Sigma}_{\mathtt{jud}}^{(r)}\}$ are computed from the joint distribution. In this work diagonal versions of $\mathbf{A}_{\mathtt{jud}}^{(r)}$ and $\boldsymbol{\Sigma}_{\mathtt{jud}}^{(r)}$ are used.

The computational advantage of JUD is that a VTS-like operation is only required at the regression class level. The compensation parameters are then relatively cheap to apply to the model-set. There is a linear transform of the features, similar to CMLLR, and a simple bias on the component variance.

3.3. Predictive CMLLR

Though JUD is significantly faster than VTS, the computational cost of applying the transform is still a function of the number of components in the recognition system as the additive bias must be applied. To address this problem predictive CMLLR (PCMLLR) has been proposed [8]. Here a CMLLR-style transform is computed, but rather than using observations, the mismatch function is used to derive *pseudo* statistics to estimate the transform. Thus the decoding expression is

$$p(\boldsymbol{y}_t|\mathtt{s}_m) = |\mathbf{A}_{\mathtt{pc}}^{(r)}|\mathcal{N}(\mathbf{A}_{\mathtt{pc}}^{(r)}\boldsymbol{y}_t + \mathbf{b}_{\mathtt{pc}}^{(r)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}), \quad (13)$$

The difference to CMLLR is that equations 4 and 5 are now expressed in terms of (only diagonal transforms used in this work)

$$\mathbf{G}_{\mathtt{pc}i}^{(r)} = \sum_{m\in\mathtt{r}_r} \frac{\gamma_{\mathtt{x}}^{(m)}}{\sigma_{\mathtt{x}i}^{(m)2}} \begin{bmatrix} 1 & \mathcal{E}\{y_{ti}|\mathtt{s}_m\} \\ \mathcal{E}\{y_{ti}|\mathtt{s}_m\} & \mathcal{E}\{y_{ti}^2|\mathtt{s}_m\} \end{bmatrix} \quad (14)$$

$$\mathbf{k}_{\mathtt{pc}i}^{(r)} = \sum_{m\in\mathtt{r}_r} \frac{\gamma_{\mathtt{x}}^{(m)}\mu_{\mathtt{x}i}^{(m)}}{\sigma_{\mathtt{x}i}^{(m)2}} \begin{bmatrix} 1 & \mathcal{E}\{y_{ti}|\mathtt{s}_m\} \end{bmatrix} \quad (15)$$

where the component "occupancies", $\gamma_{\mathtt{x}}^{(m)}$, are derived from the occupancy counts in the training data. In [8] the statistics used to derive PCMLLR was obtained using stereo data and applying SPR to clean models. If noise estimates are available though, statistics can be obtained from either the VTS or JUD predictive compensation schemes (other forms of model compensation can also be used). If VTS is used then the expectations are simple to derive as, for example $\mathcal{E}\{y_{ti}|\mathtt{s}_m\} = \mu_{\mathtt{y}i}^{(m)}$. Though this yields simple expressions, the scheme is no more computationally efficient than VTS.

A more efficient form can be derived using JUD. The expectation of y_i for a particular component $\mathtt{s}_m$ can be expressed as

$$\mathcal{E}\{y_{ti}|\mathtt{s}_m\} = \left(\mu_{\mathtt{x}i}^{(m)} - b_{\mathtt{jud}i}^{(r)}\right)/a_{\mathtt{jud}i}^{(r)}. \quad (16)$$

The associated elements of $\mathbf{G}_{\mathtt{pc}i}^{(r)}$ and $\mathbf{k}_{\mathtt{pc}i}^{(r)}$ may then be expressed as

$$\sum_{m\in\mathtt{r}_r} \frac{\gamma_{\mathtt{x}}^{(m)}}{\sigma_{\mathtt{x}i}^{(m)2}}\mathcal{E}\{y_{ti}|\mathtt{s}_m\} = \quad (17)$$

$$\frac{1}{a_{\mathtt{jud}i}^{(r)}}\left(\sum_{m\in\mathtt{r}_r}\frac{\gamma_{\mathtt{x}}^{(m)}}{\sigma_{\mathtt{x}i}^{(m)2}}\mu_{\mathtt{x}i}^{(m)}\right) - \frac{b_{\mathtt{jud}i}^{(r)}}{a_{\mathtt{jud}i}^{(r)}}\left(\sum_{m\in\mathtt{r}_r}\frac{\gamma_{\mathtt{x}}^{(m)}}{\sigma_{\mathtt{x}i}^{(m)2}}\right)$$

The terms in brackets are only functions of the clean model parameters. These can be accumulated and cached once for the system, making this form of compensation highly efficient both in terms of transform estimation and run-time decoding. This form of efficient PCMLLR JUD-based estimation is used in this paper.

4. PREDICTIVE AND ADAPTIVE COMPENSATION

In many ways the adaptive and predictive schemes described in the previous two sections are complementary to one another. Adaptive schemes are applicable to a range of tasks, for example speaker adaptation and noise compensation, but require sufficient adaptation data to obtain robust parameter estimates. In contrast predictive schemes are specifically aimed at noise robust speech recognition. However they only require a small amount of adaptation data as, compared to adaptive schemes, they have very few parameters to be estimated. An additional problem with predictive schemes is the mismatch function must be specified. Inaccuracies and approximations in deriving the mismatch function will impact performance.

Combining the two approaches allows some of the weakness of each to be reduced. For small amounts of data predictive schemes can act as a good "prior" for adaptive schemes. Whereas as the amount of data increases the approximations in the mismatch functions and lack of adaptation to the speaker from predictive schemes can be addressed using the adaptive transforms. The use of priors for MLLR transforms is common [9]. However for CMLLR a conjugate prior, even for the complete data-set, is not possible. Instead count smoothing will be used (this is also simple to apply for MLLR). Here the pseudo counts associated with the predictive transform are combined with the actual observed counts, the transform is then estimated.

Take the example of combining PCMLLR with CMLLR. The observed counts $\mathbf{G}_i^{(r)}$ and $\mathbf{k}_i^{(r)}$ from equations 4, 5 and normalised pseudo-counts from equations 14, 15 are combined to yield

$$\mathbf{G}_{\mathtt{pa}i}^{(r)} = \frac{\mathbf{G}_{\mathtt{pc}i}^{(r)}}{\sum_{m\in\mathtt{r}_r}\gamma_{\mathtt{x}}^{(m)}} + \tau_{\mathtt{sm}}\mathbf{G}_i^{(r)} \tag{18}$$

$$\mathbf{k}_{\mathtt{pa}i}^{(r)} = \frac{\mathbf{k}_{\mathtt{pc}i}^{(r)}}{\sum_{m\in\mathtt{r}_r}\gamma_{\mathtt{x}}^{(m)}} + \tau_{\mathtt{sm}}\mathbf{k}_i^{(r)} \tag{19}$$

and $\beta^{(r)}$ becomes then $\beta_{\mathtt{pa}i} = 1 + \sum_{m\in\mathtt{r}_r}\sum_{t=1}^{T}\gamma_{\mathtt{y}}^{(m)}(t)$. The reason for normalising the pseudo-counts is that it makes their contribution independent of the size of the training data and the smoothing term, τ_{sm} will be related to the minimum transform occupancies normally used to obtain robust transform estimates. The estimated transform is then used in the same fashion as CMLLR in equation 1.

It is possible to combine CMLLR with other forms of predictive scheme. For VTS compensation with CMLLR the decoding would then have the form

$$p(\boldsymbol{y}_t|\mathtt{s}_m) = |\mathbf{A}_{\mathtt{pa}}^{(r)}|\mathcal{N}(\mathbf{A}_{\mathtt{pa}}^{(r)}\boldsymbol{y}_t + \mathbf{b}_{\mathtt{pa}}^{(r)}; \boldsymbol{\mu}_{\mathtt{y}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{y}}^{(m)}), \tag{20}$$

where $\boldsymbol{\mu}_{\mathtt{y}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathtt{y}}^{(m)}$ are obtained from VTS compensation. To get the smoothed pseudo counts for this form of transform, the expressions in equations 14 and 15 must be modified. The combined transform acts on the VTS compensated model parameters. Thus the pseudo counts $\mathbf{k}_{\mathtt{pc}i}^{(r)}$ and counts $\mathbf{k}_i^{(r)}$ become

$$\mathbf{k}_{\mathtt{pc}i}^{(r)} = \sum_{m\in\mathtt{r}_r}\frac{\gamma_{\mathtt{x}}^{(m)}\mu_{\mathtt{y}i}^{(m)}}{\sigma_{\mathtt{y}i}^{(m)2}}\begin{bmatrix}1 & \mu_{\mathtt{y}i}^{(m)}\end{bmatrix} \tag{21}$$

$$\mathbf{k}_i^{(r)} = \sum_{m\in\mathtt{r}_r}\frac{\mu_{\mathtt{y}i}^{(m)}}{\sigma_{\mathtt{y}i}^{(m)2}}\sum_{t=1}^{T}\gamma_{\mathtt{y}}^{(m)}(t)\begin{bmatrix}1 & y_{ti}\end{bmatrix} \tag{22}$$

Similar forms can be obtained for $\mathbf{G}_{\mathtt{pc}i}^{(r)}$ and $\mathbf{G}_i^{(r)}$. From this form it is clear that when $\tau_{\mathtt{sm}} = 0$ the resulting transform will be an identity matrix, as expected. Similar forms of expression can be obtained combining JUD with CMLLR.

5. RESULTS

The proposed schemes were evaluated in an incremental mode on a task with real recorded noise: the Toshiba in-car database. This is a corpus collected by Toshiba Research Europe Limited's Cambridge Research Laboratory. It is a small/medium sized task with noisy speech collected in the office and in vehicles driving at various conditions. This work uses a subset of three of test sets[3] containing digit sequences (phone numbers) recorded in a car with a microphone mounted on the rear-view mirror. Though the task is artificial it is in-car recorded data, so allows an initial investigation of these predictive and adaptive transforms. The ENON set is recorded with the engine idle, and has a 35 dB average SNR. The CITY set, is recorded driving in cities, and has a 25 dB average SNR. The HWY set is recorded on the highway, and has a 18 dB average SNR. The test set comprises 20,19 and 20 speakers respectively, each speaker uttering 30 digit sequences.

The speech recogniser was trained on clean data from the Wall Street Journal corpus. The feature vector dimension is 39 consisting of 12 MFCCs appended with the zeroth cepstrum, and delta and delta-delta coefficients were used. The total number of decision tree clustered states was about 650 with 12 Gaussian components with diagonal covariance matrices each. Cross-word triphones models were obtained with three emitting states per HMM. This system is more compact than the usual form of system built on the WSJ data, but is felt to be more realistic for an embedded application. For JUD, CMLLR, PCMLLR and the combined schemes a regression class tree with 64 classes was used. Thus the number regression classes is approximated a hundredth the number of components in the system. For the standard CMLLR experiments a regression class tree with the same base-classes was used. The minimum occupancy settings were tuned separately for the diagonal (100 frames) and the full (1000 frames). In addition to this clean system a multi-style trained system was built with car noise added at various SNRs. See [10] for more details of the training configuration. For all the predictive schemes, the initial noise model for the first utterance of each sequence was obtained using the first and last 20 frames of the utterance itself. This was then re-estimated for every subsequent utterance using the ML VTS-based scheme described in [6]. Though VTS is used in the noise estimation only a subset of the components (those with non-zero counts) need to be compensated. JUD based noise estimation schemes can be used [10], which allow the computational cost of the noise estimation to be reduced. Table 1 summarises the results obtained for the proposed predictive and adaptive schemes on the three test sets. The performance of the diagonal CMLLR adaptation

[3] Only speakers where all utterance were in the test set were used. This is more appropriate for this incremental mode - avoiding large gaps between utterances.

System	Condition WER (%)		
	ENON	CITY	HWY
Clean	2.91	32.93	66.30
-CMLLR	0.60	7.57	40.62
-CMLLR (full)	0.75	30.27	68.14
-VTS	1.22	3.06	4.11
-JUD	1.10	2.86	5.42
-PCMLLR	1.21	2.93	6.14
MST	2.50	7.59	27.38
-CMLLR	1.40	4.79	8.39
-CMLLR (full)	1.01	4.59	6.41

Table 1. Incremental results on using CMLLR, VTS, JUD and PCMLLR performance on Clean system; diagonal and full CMLLR performance on Multi-style trained (MST) system.

with the clean system was disappointing (though better than the full CMLLR performance which yielded no improvement) on the lower SNR conditions. This is partly because of the large mismatch between the training and test conditions. All the predictive schemes out-performed the adaptive CMLLR approach. The trend was expected, where VTS performed best at the lowest SNR condition, HWY. The difference between VTS and the approximations, JUD and PCMLLR, was very small for the two higher SNR conditions, ENON and CITY. Though the multi-style trained system gave gains over the clean system, and full CMLLR out-performed the diagonal case, the performance was worse than the predictive schemes for the lower SNR conditions.

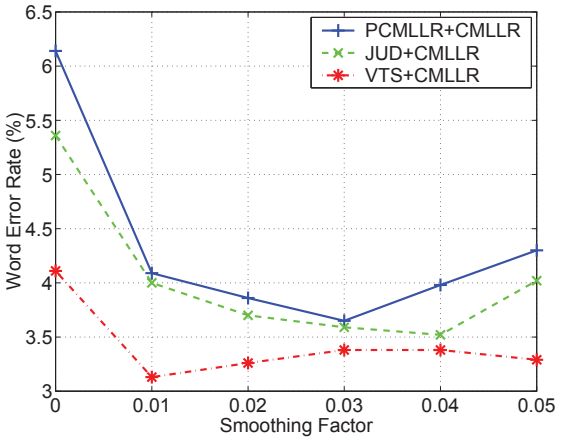

Fig. 1. Result of combined PCMLLR+CMLLR, JUD+CMLLR and VTS+CMLLR schemes varying the smoothing factor τ_{sm} on the HWY test set.

The impact of the smoothing factor on the performance on the predictive and adaptive smoothing scheme for the HWY test set is shown in figure 1. When $\tau_{\text{sm}} = 0$ the performance is the same as that in table 1. Gains are obtained for the range of smoothing factors from $\tau_{\text{sm}} = 0.01$ to $\tau_{\text{sm}} = 0.05$.

The results for all three tasks using $\tau_{\text{sm}} = 0.03$ are summarized in table 2. For all tasks large gains over either the pure predictive, or pure adaptive schemes can be observed. Even at the highest SNR condition, ENON, the combined scheme outperformed the best adaptive scheme, clean plus CMLLR. Another aspect of the combined scheme is that the differences between the various predictive approaches is reduced. This is expected, partly because of the number of utterances for each speaker, but also the adaptive scheme will reduce the impact of the approximations in the faster predictive schemes. It is interesting that PCMLLR+CMLLR, a scheme that can be made highly efficient achieves good performance over all three tasks.

System	Condition WER (%)		
	ENON	CITY	HWY
VTS	1.22	3.06	4.11
+CMLLR	0.51	2.67	3.38
JUD	1.10	2.86	5.42
+CMLLR	0.50	2.30	3.59
PCMLLR	1.21	2.93	6.14
+CMLLR	0.50	2.30	3.65

Table 2. Combined predictive and adaptive compensation on the Toshiba in-car data using the clean system and $\tau_{\text{sm}} = 0.03$.

6. CONCLUSIONS

This paper has discussed the use of incremental model compensation using both predictive schemes, such as VTS and JUD, and adaptive schemes such as CMLLR. In addition the use of incremental predictive CMLLR is discussed. By using statistics derived from JUD compensation, the estimation of PCMLLR parameters can be made efficient, both in terms of estimation and application. Moreover by combining predictive and adaptive schemes together it is possible to obtain gains over either individually on an in-car recorded task.

7. REFERENCES

[1] P. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[2] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech*, 2005.

[3] C. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs.," *Computer Speech and Language*, vol. 9, 1995.

[4] M. J. F. Gales, "Maximum Likelihood Linear Transformations For HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, Jan. 1998.

[5] L. R. Neumeyer, A. Sankar, and V. V. Digalakis, "A comparative study of speaker adaptation techniques," in *Proc. Eurospeech*, 1995.

[6] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for robust large vocabulary speech recognition," Tech. Rep. CUED/F-INFENG/TR552, University of Cambridge, 2006, Available from: mi.eng.cam.ac.uk/~hl251.

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

[8] M. J. F. Gales and R. C. van Dalen, "Predictive linear transforms for noise robust speech recognition," in *Proc. ASRU*, 2007, pp. 59–64.

[9] W. Chou, "Maximum a-posterior linear regression with elliptical symmetric matrix variate priors," in *Proc. ICASSP*, 1999.

[10] H. Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, Cambridge, UK, Sept. 2007.