Open access • Proceedings Article • DOI:10.1109/ICASSP.2007.366931

Adaptive Training with Joint Uncertainty Decoding for Robust Recognition of Noisy Data — Source link

Hank Liao, Mark J. F. Gales

Institutions: University of Cambridge

Published on: 15 Apr 2007 - International Conference on Acoustics, Speech, and Signal Processing

Topics: Noise, Acoustic model, Signal-to-noise ratio and Speech coding

ADAPTIVE TRAINING WITH JOINT UNCERTAINTY DECODING FOR ROBUST RECOGNITION OF NOISY DATA

H. Liao and M. J. F. Gales

Cambridge University Engineering Department
Trumpington St., Cambridge, CB2 1PZ, UK
{hl251,mjfg}@eng.cam.ac.uk

ABSTRACT

Standard noise compensation techniques for automatic speech recognition assume a clean trained acoustic model. What is thought of as "clean" data, may still have a variety of speakers, different channels and varying noise conditions. Hence it may be more reasonable to consider such data multi-conditional for multistyle training. This paper shows that multistyle models benefit from VTS compensation or `Joint` uncertainty decoding by reducing the mismatch between training and test. An EM-based noise estimation procedure that produces ML VTS or `Joint` noise models is also described. Alternatively, adaptive training with `Joint` uncertainty transforms factors out the noise from the data. The uncertainty variance bias de-weights observations in the training data where the SNR is low. This property allows data with a wide SNR range to be used and produces canonical models that truly represent clean speech, whereas multistyle trained models must account for all acoustic variation associated with different noise conditions. This paper presents `Joint` adaptive training including formula for estimating the transforms and canonical model parameters. Experiments are conducted on the Resource Management and Broadcast News corpora.

***Index Terms*—** Speech recognition, Robustness

1. INTRODUCTION

Conventional approaches to improve recognition robustness of noisy speech presume the acoustic model is clean trained. Despite being considered "clean", the training data may contain a wide variety of speakers, accents, channels and noise conditions. Hence, clean data may be considered multi-conditional and the models trained on this data in a multistyle fashion. Little research has been conducted on applying noise compensation techniques to multistyle systems. This paper examines VTS and `Joint` compensation as general approaches to reducing the mismatch between the training and test condition for both clean and multistyle trained acoustic models.

Alternatively, adaptive training may be applied to remove these unwanted factors, such as speaker differences or the acoustic environment, from being included in the acoustic models [1, 2]. Rather than force the acoustic model to represent all these factors, as expected in multistyle training, transforms are used instead to model the variation from different factors. MLLR transforms can only normalise low levels of noise, hence is unsuitable for adaptive training with data that has large variations in SNR. This motivates a novel model training framework called `Joint` adaptive training (JAT), based on noise normalisation using `Joint` transforms for training models on noisy data. JAT takes into account the SNR of the data when estimating the canonical model parameters. When the noise subsumes the speech, the uncertainty variance bias ensures those observations do not contribute to the parameter update. In this way, JAT weights the training data using uncertainty due to noise. Hence, JAT explicitly handles a large range of SNR in the training data, producing a final acoustic model that is truly noise-free.

Experiments are reported on the large vocabulary Broadcast News transcription task and an artificially corrupted 1000-word Resource Management corpus.

2. MODEL-BASED NOISE COMPENSATION

In speech recognition, there are two main approaches to compensating noisy speech: cleaning the features, e.g. CMN or SPLICE, or modifying the model parameters. The latter approach, often called model-based compensation, tends to give better results than the former, feature-based approach [3]. It is frequently assumed that a noise corrupted speech observation, $\boldsymbol{o}_t = [\boldsymbol{y}_t^{\mathsf{T}} \ \Delta\boldsymbol{y}_t^{\mathsf{T}} \ \Delta^2\boldsymbol{y}_t^{\mathsf{T}}]^{\mathsf{T}}$, at time t is conditionally independent of all other observations given the clean speech $\boldsymbol{s}_t$ and the noise $\boldsymbol{n}_t$ at that time frame. The clean speech and noise are also assumed to be generated by HMMs with states θ_t^n for the noise[1] and θ_t for the clean speech. Under these assumptions the likelihood of the noisy speech may be expressed as

$$p(\boldsymbol{o}_t|\mathcal{M},\check{\mathcal{M}},\boldsymbol{\theta}_t) = \int p(\boldsymbol{o}_t|\boldsymbol{s}_t,\check{\mathcal{M}})p(\boldsymbol{s}_t|\mathcal{M},\theta_t)d\boldsymbol{s}_t \qquad (1)$$

where

$$p(\boldsymbol{o}_t|\boldsymbol{s}_t,\check{\mathcal{M}}) = \int p(\boldsymbol{o}_t|\boldsymbol{s}_t,\boldsymbol{n}_t)p(\boldsymbol{n}_t|\check{\mathcal{M}},\theta_t^n)d\boldsymbol{n}_t \qquad (2)$$

and $\check{\mathcal{M}}$ denotes the compensation model parameters, which may or may not have an explicit model of the noise[2]. Typically the uncompensated "clean" acoustic model $\mathcal{M}$ consists of Gaussian components each defined by a prior, c_m, mean, $\boldsymbol{\mu}_s^{(m)}$, and variance, $\boldsymbol{\Sigma}_s^{(m)}$.

PMC [5] and VTS compensation [6] approximate the integrals in equation 1 for each acoustic model component. This assumes that the frame/state alignment of the clean speech does not change with noise. In the cepstral domain, the relationship between the static clean speech $\boldsymbol{x}$, additive noise $\boldsymbol{n}$, channel $\boldsymbol{h}$ and static corrupted speech $\boldsymbol{y}$ is often written as [5, 6]

$$y_i = x_i + h_i + \boldsymbol{c}_i \log(\boldsymbol{1} + \exp(\boldsymbol{C}^{-1}(\boldsymbol{n} - \boldsymbol{x} - \boldsymbol{h}))) \qquad (3)$$

where matrices $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ are the discrete cosine transform matrix (DCT) and its inverse. The vector $\boldsymbol{c}_i$ denotes the i^{th} row of the DCT. The log and exp functions operate at an element level on the resultant filterbank vectors. VTS compensation approximates this nonlinear equation with a first-order vector Taylor series. While VTS has shown to be more efficient than PMC [7] and a better approximation than the log-normal [6] it is still computationally expensive

Hank Liao thanks Toshiba Research Europe Ltd. for funding this work.

[1] A single state is assumed for the noise model in this paper.

[2] If the compensation model parameters $\check{\mathcal{M}}$ are single-pass retrained, as in [4], then no noise model is explicitly estimated.

as every model component must be individually adapted with respect to the noise. This involves the computation of noisy speech gradients with respect to the noise and clean speech.

In contrast, model-based `Joint` uncertainty decoding [4], shares parameters as `Joint` transforms estimated per cluster or class of model components—analogous to how MLLR transforms may be estimated and applied. The number of clusters R is usually several order of magnitudes less than the total number of Gaussians M in the system. In uncertainty decoding, the corrupted speech likelihood for a component m takes this form

$$p(\boldsymbol{o}_t|\mathcal{M},\check{\mathcal{M}},m)=|\boldsymbol{A}^{(r)}|\mathcal{N}\Big(\boldsymbol{A}^{(r)}\boldsymbol{o}_t+\boldsymbol{b}^{(r)};\boldsymbol{\mu}_s^{(m)},\boldsymbol{\Sigma}_s^{(m)}+\boldsymbol{\Sigma}_{\mathsf{b}}^{(r)}\Big) \tag{4}$$

where the parameters $\boldsymbol{A}^{(r)}$, $\boldsymbol{b}^{(r)}$ and $\boldsymbol{\Sigma}_{\mathsf{b}}^{(r)}$ are simply estimated from the corrupted/clean speech conditional. For `Joint` uncertainty decoding, this conditional is estimated from the joint distribution. In comparison, front-end uncertainty decoding estimates a joint distribution for each component of a front-end GMM representing the observed, corrupted acoustic space; this form however suffers from a fundamental problem and is less efficient than the model-based approach [8].

Previously, the joint distribution was estimated using stereo data [4, 8]. It may be predicted given the clean speech and noise model using VTS or PMC [3, 7], resulting in noise compensating `Joint` transforms. However, the `Joint` transform may model other factors, such as speaker differences, if they are accounted for in the mismatch function during generation of the joint distribution. Furthermore, the joint distribution may be considered simply a joint model of training and test conditions. The "clean" speech models may be multistyle or adaptively trained and the compensation applied as a mechanism to reduce the mismatch between training and test. In this way `Joint` transforms are similar to MLLR transforms as they reduce mismatch, howver with the addition of a variance bias and they may be predicted given some prior models.

2.1. The Clean Speech Class Model

To determine this joint distribution per class r, an a priori model of the clean speech $\mathcal{N}\big(\boldsymbol{\mu}_x^{(r)},\boldsymbol{\Sigma}_x^{(r)}\big)$ is needed; this is derived from the full acoustic models. The class and component posteriors, and mean and variance of each class r are

$$\begin{gathered}L^{(m)}=\sum_{t=1}^{T}\gamma_t^{(m)},\quad L^{(r)}=\sum_{m\in r}L^{(m)},\quad \boldsymbol{\mu}_x^{(r)}=\frac{1}{L^{(r)}}\sum_{m\in r}L^{(m)}\boldsymbol{\mu}_x^{(r)}\\ \boldsymbol{\Sigma}_x^{(r)}=\frac{1}{L^{(r)}}\sum_{m\in r}L^{(m)}\Big(\boldsymbol{\Sigma}_x^{(m)}+\boldsymbol{\mu}_x^{(m)}\boldsymbol{\mu}_x^{(m)\mathsf{T}}\Big)-\boldsymbol{\mu}_x^{(r)}\boldsymbol{\mu}_x^{(r)\mathsf{T}}\end{gathered} \tag{5}$$

where $c^{(m)}$ is the component weight, $\boldsymbol{\mu}_x^{(m)}$ the mean, $\boldsymbol{\Sigma}_x^{(m)}$ the variance, and $\gamma_t^{(m)}$ the component posterior alignment probability at time frame t. An approximation to this is to use the diagonal form of $\boldsymbol{\Sigma}_x^{(m)}$ although $\boldsymbol{\Sigma}_x^{(r)}$ is full. For low numbers of classes R compared to the number of model components M in the acoustic model, this should be a good approximation, since the between class covariance should dominate over the within class covariance.

2.2. Noise Model Estimation

In order to apply these predictive compensation schemes to either clean or multistyle trained models, the noise parameters must be estimated to best reduce the mismatch between training and test. Non-speech regions may be used to estimate the additive noise [7], however this does not easily provide an estimate of the channel, nor can this strategy accommodate changes in the noise over long speech utterances. Hence consider an approach where a clean acoustic model is compensated using VTS, with the noise model iteratively improved using EM to maximise the likelihood of the test condition. Such an EM framework for estimating the additive and convolutional noise was presented in [9], but in this work estimation is conducted in the cepstral domain. Also, a simple iterative 1st-order gradient search is used to find an MLE of the noise variance.

Although using the MLE noise model derived using VTS compensation may give good results for `Joint` compensation [10], there is a mismatch between the compensation used during noise estimation and that applied during recognition. Hence, it is sensible to generate ML noise parameters explicitly tuned for `Joint` compensation rather than VTS. ML `Joint` noise estimation gave improved results especially for multistyle trained acoustic models [10]. The following auxiliary function is used for ML `Joint` noise estimation

$$\mathcal{Q}_J(\boldsymbol{\Phi};\hat{\boldsymbol{\Phi}})=\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t^{(m)}\log\Big[p_J(\boldsymbol{o}_t|\hat{\boldsymbol{\Phi}},\mathcal{M},m)\Big] \tag{6}$$

except the log probability for the output distribution is now given by

$$p_J(\boldsymbol{o}_t|\hat{\boldsymbol{\Phi}},\mathcal{M},m)=|\hat{\boldsymbol{A}}^{(r)}|\mathcal{N}\Big(\hat{\boldsymbol{A}}^{(r)}\boldsymbol{o}_t+\hat{\boldsymbol{b}}^{(r)};\boldsymbol{\mu}_s^{(m)},\boldsymbol{\Sigma}_s^{(m)}+\hat{\boldsymbol{\Sigma}}_{\mathsf{b}}^{(r)}\Big) \tag{7}$$

The set of R `Joint` transforms, $\mathcal{T}=\Big[\mathcal{T}^{(1)},\mathcal{T}^{(r)},\ldots,\mathcal{T}^{(R)}\Big]$, may be derived from the joint distribution that is estimated from the clean speech class model and the estimated noise parameters $\hat{\boldsymbol{\Phi}}$.

Given the acoustic model $\mathcal{M}$, from which the clean speech class model may be derived, an estimate of the noise parameters $\hat{\boldsymbol{\Phi}}$ that maximises the auxiliary function $\mathcal{Q}_J$ is required. That is find

$$\hat{\boldsymbol{\Phi}}=\Big\{\hat{\boldsymbol{\mu}}_n,\hat{\boldsymbol{\Sigma}}_n,\hat{\boldsymbol{\mu}}_h\Big\}=\arg\max_{\hat{\boldsymbol{\Phi}}}\mathcal{Q}_J\Big(\mathcal{M},\mathcal{T};\mathcal{M},\hat{\mathcal{T}}\Big) \tag{8}$$

where $\hat{\mathcal{T}}$ is computed directly from clean speech class model and $\hat{\boldsymbol{\Phi}}$. With a suitable initial starting point, here the VTS-based MLE noise model, the noise parameters may be iteratively refined using a simple gradient-based optimisation scheme. For example the additive noise mean update is

$$\hat{\mu}_{n,i}=\mu_{n,i}-\zeta\frac{\partial^2\mathcal{Q}_J}{\partial\mu_{n,i}^2}^{-1}\frac{\partial\mathcal{Q}_J}{\partial\mu_{n,i}} \tag{9}$$

where ζ is the learning rate; the additive noise variance and channel mean are similar. The second derivatives need to be conditioned such that they remain negative to ensure the updates converge to a local maximum; when they are not, a simple back-off strategy is to switch to a first-order optimisation. It is also important to ensure that each step improves the auxiliary. More detailed information of this estimation procedure for both VTS and `Joint` MLE noise models is given in [10].

3. JOINT ADAPTIVE TRAINING

Adaptive training is a powerful technique for factoring out unwanted variability due to speaker, channel and environmental mismatch [1, 2]. This yields a pure "canonical" model of speech compared to multistyle training where the models incorporate all the variability of the acoustic data. In adaptive training, both a set of transforms and the acoustic model parameters are iteratively estimated in an EM framework. First, given the current acoustic models $\mathcal{M}$, a new set of transform $\mathcal{T}$ is estimated. Subsequently, the canonical model parameters are updated given this new set of transforms. Multiple iterations of this interleaved training may be performed to optimise an auxiliary function for the noisy speech observations $\boldsymbol{O}$ and state sequence $\boldsymbol{\theta}$ given the transcription. Compared to adaptation with

MLLR, `Joint` uncertainty transforms may explicitly model the effects of noise when a mismatch function for noise, such as VTS, is used to generate the joint distribution.

With JAT, determining the ML transforms and model parameters is not directly possible so an auxiliary function is used

$$\mathcal{Q}_J\Big(\mathcal{M},\boldsymbol{\Phi};\hat{\mathcal{M}},\hat{\boldsymbol{\Phi}}\Big)=\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\sum_{m=1}^{M}\gamma_t^{(m)}\times \log\left[|\hat{\boldsymbol{A}}^{(rh)}|\mathcal{N}\left(\hat{\boldsymbol{A}}^{(rh)}\boldsymbol{o}_t+\hat{\boldsymbol{b}}^{(rh)};\hat{\boldsymbol{\mu}}_s^{(m)},\hat{\boldsymbol{\Sigma}}_s^{(m)}+\hat{\boldsymbol{\Sigma}}_{\mathsf{b}}^{(rh)}\right)\right] \tag{10}$$

where $\gamma_t^{(m)}$ is the posterior probability that the observation o_t is generated by component m on heterogeneous training data segmented into H homogeneous blocks, each of length $T^{(h)}$, for all valid state sequences given the transcription.

The uncertainty transforms in JAT are estimated as described in section 2.2. Note the clean speech class model, described in 2.1, needs to be re-computed every time the canonical model is updated. When estimating new transforms, this creates a disconnect where the initial ML noise model is estimated with a different clean speech class model than the input transform. Nevertheless, it may be possible to begin with the `Joint` transform produced from $\hat{\boldsymbol{\Phi}}$ and $\hat{\mathcal{M}}$. More discussion of this issue may be found in [10].

3.1. Canonical Model Parameter Estimation

After a new set of transforms are estimated, the model parameters are retrained. The auxiliary function, from equation 10, where only terms dependent on the model parameters are shown, given $\hat{\boldsymbol{\Phi}}$, is

$$\mathcal{Q}_J\Big(\mathcal{M};\hat{\mathcal{M}}\Big)=-\frac{1}{2}\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\sum_{m=1}^{M}\gamma_t^{(m)}\times \sum_{i=1}^{D}\left(\log(\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2})+\frac{(\hat{\boldsymbol{a}}_i^{(rh)}\boldsymbol{o}_t+\hat{b}_i^{(rh)}-\mu_{s,i}^{(m)})^2}{\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2}}\right) \tag{11}$$

where diagonal covariance matrices assumed and D is the number of dimensions in the feature vector. Because the joint transform parameters affect the model parameters and are shared over many homogeneous blocks, there is no closed form solution for the model parameters that maximise this auxiliary function. Hence a generalised EM approach is taken, where a second order gradient based optimisation scheme is used to optimise the model parameters

$$\begin{bmatrix}\hat{\mu}_{s,i}^{(m)}\\ \hat{\sigma}_{s,i}^{(m)2}\end{bmatrix}=\begin{bmatrix}\mu_{s,i}^{(m)}\\ \sigma_{s,i}^{(m)2}\end{bmatrix}-\zeta\begin{bmatrix}\frac{\partial^2\mathcal{Q}_J}{\partial\mu_{s,i}^{(m)2}} & \frac{\partial^2\mathcal{Q}_J}{\partial\mu_{s,i}^{(m)}\partial\sigma_{s,i}^{(m)2}}\\ \frac{\partial^2\mathcal{Q}_J}{\partial\sigma_{s,i}^{(m)2}\partial\mu_{s,i}^{(m)}} & \frac{\partial^2\mathcal{Q}_J}{\partial\left(\sigma_{s,i}^{(m)2}\right)^2}\end{bmatrix}^{-1}\begin{bmatrix}\frac{\partial\mathcal{Q}_J}{\partial\mu_{s,i}^{(m)}}\\ \frac{\partial\mathcal{Q}_J}{\partial\sigma_{s,i}^{(m)2}}\end{bmatrix} \tag{12}$$

The learning rate ζ may be less than one, but in this work a value of unity was found to be stable. The first derivative of the auxiliary in equation 11 with respect to the mean of component m and dimension i is

$$\frac{\partial\mathcal{Q}_J}{\partial\mu_{s,i}^{(m)}}=\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\gamma_t^{(m)}\left(\frac{\hat{\boldsymbol{a}}_i^{(rh)}\boldsymbol{o}_t+\hat{b}_i^{(rh)}-\mu_{s,i}^{(m)}}{\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2}}\right) \tag{13}$$

and with respect to the model variance

$$\frac{\partial\mathcal{Q}_J}{\partial\sigma_{s,i}^{(m)2}}=\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\frac{1}{2}\omega_{t,i}^{(m)}\left(\frac{(\hat{\boldsymbol{a}}_i^{(rh)}\boldsymbol{o}_t+\hat{b}_i^{(rh)}-\mu_{s,i}^{(m)})^2}{\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2}}-1\right) \tag{14}$$

where $\omega_{t,i}^{(m)}=\frac{\gamma_t^{(m)}}{\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2}}$. The Hessian matrix is composed of

$$\frac{\partial^2\mathcal{Q}_J}{\partial\mu_{s,i}^{(m)2}}=-\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\omega_{t,i}^{(m)} \tag{15}$$

$$\frac{\partial^2\mathcal{Q}_J}{\partial\left(\sigma_{s,i}^{(m)2}\right)^2}=\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\omega_{t,i}^{(m)}\left(\frac{1}{2}-\frac{\left(\hat{\boldsymbol{a}}_i^{(rh)}\boldsymbol{o}_t+\hat{b}_i^{(rh)}-\mu_{s,i}^{(m)}\right)^2}{\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2}}\right) \tag{16}$$

$$\frac{\partial^2\mathcal{Q}_J}{\partial\mu_{s,i}^{(m)}\partial\sigma_{s,i}^{(m)2}}=-\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\omega_{t,i}^{(m)}\left(\frac{\hat{\boldsymbol{a}}_i^{(rh)}\boldsymbol{o}_t+\hat{b}_i^{(rh)}-\mu_{s,i}^{(m)}}{\sigma_{s,i}^{(m)2}+\hat{\sigma}_{\mathsf{b},i}^{(rh)2}}\right) \tag{17}$$

From equations 13 and 14, it can be seen that contributions from observations when the SNR is low will be de-weighted by the uncertainty bias term $\hat{\sigma}_{\mathsf{b},i}^{(rh)2}$. When the noise completely subsumes the speech, the uncertainty bias will be infinite and these observations will not contribute to the model parameter update. If the SNR is high, the uncertainty bias will tend to zero, allowing these observations to fully contribute. This allows the canonical model to truly be a representation of clean, noise-free speech.

The estimation of the model variance is stabilised by limiting it to changing at most by a factor of υ

$$\hat{\sigma}_{s,i}^{(m)2}=\min\left(\max\left(\hat{\sigma}_{s,i}^{(m)2},\frac{1}{\upsilon}\sigma_{s,i}^{(m)2}\right),\upsilon\sigma_{s,i}^{(m)2}\right) \tag{18}$$

In practice, υ was set at 2. The Hessian matrix must also be negative definite for the optimisation to converge, however the 2nd derivative is not guaranteed to be. It may be re-expressed as

$$\frac{\partial^2\mathcal{Q}_J}{\partial(\sigma_{s,i}^{(m)2})^2}=w_{1,i}^{(m)}\left(-\hat{\vartheta}+\frac{1}{2}\right) \tag{19}$$

where $\hat{\vartheta}=\max\left(\vartheta,-\frac{w_{2,i}^{(m)}}{w_{1,i}^{(m)}}\right)$ and

$$w_{1,i}^{(m)}=\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\frac{\gamma_t^{(m)}}{\left(\sigma_{s,i}^{(m)2}+\sigma_{\mathsf{b},i}^{(rh)2}\right)^2} \tag{20}$$

$$w_{2,i}^{(m)}=-\sum_{h=1}^{H}\sum_{t=1}^{T^{(h)}}\gamma_t^{(m)}\left(\frac{(\boldsymbol{a}_i^{(rh)}\boldsymbol{o}_t+b_i^{(rh)}-\mu_{s,i}^{(m)})^2}{\left(\sigma_{s,i}^{(m)2}+\sigma_{\mathsf{b},i}^{(rh)2}\right)^3}\right) \tag{21}$$

This parameter ϑ should remain greater than a half to ensure stability of the optimisation. It may be observed that the ratio of $w_{2,i}^{(m)}$ to $w_{1,i}^{(m)}$ should converge to unity as the model parameters better approximate the training data, given the set of `Joint` transforms.

Lastly, instead of directly optimising the variance, the log of the variance is estimated to ensure that the converged value remains positive. Thus, the following change of variable is made

$$\boldsymbol{\varsigma}=\log\boldsymbol{\Sigma}_s^{(m)} \tag{22}$$

The derivatives may be easily recomputed to now optimise $\boldsymbol{\varsigma}$.

4. EXPERIMENTS

A simplified Broadcast News system based on the 2003 CU-HTK system [11] was evaluated. MFCC parameters with the 0th cepstra, and associated 1st- and 2nd-order features for 39 dimensions were used with cross-word triphones and decision-tree clustered states.

There were 16 Gaussian components per state, yielding over 100K model components. The CU RT-03 diarisation system segmented and clustered the BN audio providing 143 hours of data for ML training. The dictionary contained 59K words. For decoding, a bigram LM generated lattices which were re-scored using a trigram LM. An initial decoding run provided the hypothesis for noise estimation. Figures are reported against the `dev03` test set encompassing 3 hours of shows from six different news sources aired in Jan 2001.

Compensation	Noise Est. Type	%WER
None	—	20.8
`Joint`	ML VTS	19.1
	ML `Joint`	18.8
VTS	ML VTS	18.8

Table 1. Broadcast News results using 256 `Joint` transforms or VTS with 2 full CMLLR transforms on `dev03` test set.

BN results are presented in table 1. There was a 2% absolute gain over the baseline system when applying either 256 `Joint` transforms or VTS compensation. When these compensation schemes are used with CMLLR, this improved to 18.0% and 17.7% respectively, although with `Joint` a small gain of 0.1% is maintained over solely using CMLLR. However, clearly the noise estimation type should match the compensation; there is a 0.3% gain for `Joint` compensation when `Joint` noise estimates are used rather than VTS. Preliminary experiments with JAT on BN showed no improvements. This was felt to be due to the training data being of relatively high SNR. Hence further experiments were conducted by artificially corrupting the RM task.

For experiments on the 1000-word Resource Management task, the same features and model topology as the BN system were used, except for 6 components per GMM. This gave 9492 system components. Data was corrupted with Car and Operations Room noise from the NOISEX-92 database to give 20 and 14 dB SNR tests; results are averaged across the `feb89`, `oct89` and `feb91` test sets totaling an hour. A multistyle 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 used as the initial model to begin JAT. The parameter ϑ was reduced from 2.5 to 1 in increments of 0.5 for 4 iterations of model re-estimation between each of 4 transform updates.

Acoustic Model Training	Compensation	Operations 20 dB	Operations 14 dB	Car 20 dB
Matched	—	7.4	14.3	—
Clean	—	38.0	83.7	49.7
	`Joint`	9.2	22.6	8.0
	VTS	8.4	23.6	7.4
Multistyle	—	7.0	15.5	43.5
	`Joint`	6.7	12.3	7.6
	VTS	6.5	12.0	6.9
JAT	`Joint`	6.2	11.4	6.2

Table 2. Baseline RM clean, multistyle, and JAT performance with 16 transform `Joint` and VTS compensation, %WER).

Table 2 provides RM results. As expected, clean performance was poor, while multistyle trained acoustic models gave results comparable to matched system. Applying VTS or 16 diagonal model-based `Joint` transforms to the clean models greatly improved results; but more interestingly, compensating multistyle models gave accuracies better than matched for either scheme. Using 16 diagonal constrained MLLR transforms was consistently poorer than `Joint` compensation for these conditions. The best training scheme was the JAT system, which exceeded matched and multistyle with VTS performance at both 20 and 14 dB SNR. The 20 dB Car test contains noise not seen in the training data. The results illustrate the weakness of multistyle training when the noise is not present in the training; the error rate only improves slightly to 43.5% from 49.7% on clean trained. However, the JAT system factors in the Car noise as easily as the Operations Room, yielding the same word error rate of 6.2%. Still, this is double the matched clean WER of 3.1%.

5. CONCLUSIONS

This paper has discussed various approaches to building robust automatic speech recognition systems. Multistyle training, where the data is used directly estimate the model parameters, may be also be compensated with VTS or `Joint` schemes to give additional robustness by reducing the mismatch between training and test conditions. Experiments on a multistyle, large vocabulary Broadcast News system show improvements with `Joint` and VTS compensation and demonstrate how matching the compensation used during noise estimation to that used during test improves accuracy. Moreover, a new form of adaptive training with `Joint` transforms gives the best results since the noise is factored out from the training data. The uncertainty due to noise will de-weight noisier segments of speech allowing JAT to accommodate a wide range of SNR in the training data. This results in acoustic models which truly represent the pure acoustic speech variability, rather than effects due to speaker differences or noise conditions. This was shown on experiments conducted on the RM database.

6. REFERENCES

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

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

[3] H. Liao and M. J. F. Gales, "Uncertainty decoding for noise robust speech recognition," Tech. Rep. CUED/F-INFENG/TR499, University of Cambridge, 2004.

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

[5] M. J. F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, Cambridge University, 1995.

[6] 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.

[7] H. Xu, L. Rigazio, and D. Kryze, "Vector Taylor series based joint uncertainty decoding," in *Proc. Interspeech*, 2006.

[8] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust speech recognition," in *Proc. ICSLP*, 2006.

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

[10] 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.

[11] D. Y. Kim, G. Evermann, T. Hain, D. Mrva, S. E. Tranter, L. Wang, and P. C. Woodland, "Recent advances in broadcast news transcription," in *Proc. ASRU*, 2003.