Issues with Uncertainty Decoding for Noise Robust Speech Recognition

H. Liao and M. J. F. Gales

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

Abstract

Recently there has been interest in uncertainty decoding for robust speech recognition. Here the uncertainty associated with the observation in noise is propagated to the recogniser. By using appropriate approximations for this uncertainty, it is possible to obtain efficient implementations during decoding. The aim of these schemes is to obtain performance which is close to that of a model-based compensated system, without the computational cost. Unfortunately, in low SNR there is a fundamental issue with front-end uncertainty decoding where the model means and variances are updated according to the features. This is described in detail using the `Joint` and `SPLICE` with uncertainty forms, but is not limited to these two techniques. A solution for the `Joint` scheme is presented along with the implicit approach used in `SPLICE` with uncertainty. In addition, a model-based `Joint` uncertainty scheme is described, which is more efficient and powerful than the front-end schemes, and being model-based not affected by this problem. This issue is illustrated using the AURORA 2.0 database with these various systems.

Index Terms: model-based noise compensation, robust speech recognition, uncertainty decoding.

1. Introduction

Speech recognition in noise has been an area of active research for many years. Powerful model-based compensation schemes, such as Parallel Model Combination (PMC), Vector Taylor Series (VTS) and more recently `ALGONQUIN` [1], achieve good performance but are computationally expensive compared to feature compensation. Recently interest has grown in an elegant compromise between model-based and front-end schemes: *uncertainty decoding*, so called because a measure of the uncertainty introduced by the background acoustic noise is propagated into the recognition process [2, 3][1]. For front-end uncertainty schemes, this uncertainty is computed solely from the features.

Despite front-end uncertainty decoding achieving good performance for a range of acoustic environments [3], a fundamental problem arises. By passing a single uncertainty value to the decoder per frame, when the SNR is low large uncertainties can cause all the model variances to be rendered the same. When this occurs, the recogniser can no longer discriminate in these areas which can result in large numbers of insertion errors. This is especially the case when there is no other additional constraints such as a language model, e.g. in the AURORA digit recognition task [5]. This is not only an issue with specific implementations such as `SPLICE` with uncertainty [2] and front-end `Joint` [3], but all forms based on the uncertainty decoding framework presented here.

This paper examines the cause of this fundamental issue with front-end uncertainty decoding and how these specific implementations may be modified to address this problem. A model-based `Joint` uncertainty decoding scheme is discussed that is more effective, and actually more efficient, than standard front-end schemes without suffering from this limitation in low SNR conditions. The issues discussed are demonstrated on the standard AURORA 2.0 digit string recognition task.

2. The Uncertainty Decoding Framework

This section reviews the uncertainty decoding framework [2, 3]. A dynamic Bayesian network, as in figure 1, can represent the effects of environmental noise. Here, the noise corrupted speech observation $\boldsymbol{y}_t$ at time t is assumed to be conditionally independent of all other observations given the clean speech $\boldsymbol{x}_t$ and the noise $\boldsymbol{n}_t$ at that time. By also assuming the clean speech and noise are generated by HMMs with states θ_t^n for the noise[2] and θ_t for the clean speech, the corrupted speech likelihood may be expressed as

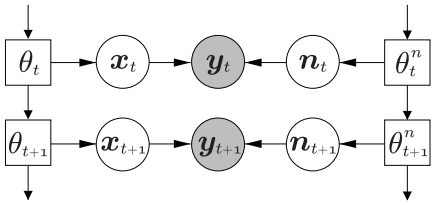

Figure 1: Uncertainty Decoding DBN.

$$p(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t) = \int p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})p(\boldsymbol{x}_t|\mathcal{M},\theta_t)d\boldsymbol{x}_t \quad (1)$$

where $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}}) = \int p(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{n}_t)p(\boldsymbol{n}_t|\check{\mathcal{M}},\theta_t^n)d\boldsymbol{n}_t$ and $\check{\mathcal{M}}$ the front-end compensation model. The acoustic model $\mathcal{M}$ consists of Gaussians each defined by a prior, c_m, mean, $\boldsymbol{\mu}^{(m)}$, and variance, $\boldsymbol{\Sigma}^{(m)}$. The likelihood calculation thus has two distinct parts. In equation 1, only the first part, $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})$, is a function of the noise. Front-end uncertainty decoding takes advantage of this factorisation, by making this conditional independent of the acoustic models; a wide variety approaches are then possible to model the conditional without adversely affecting decoding efficiency.

An important consideration in uncertainty decoding is the form of representation for the conditional distribution,

Hank Liao is funded by Toshiba Research Europe Ltd.

[1]The uncertainty decoding framework, as described here, differs from the feature variance approach as discussed in [4].

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

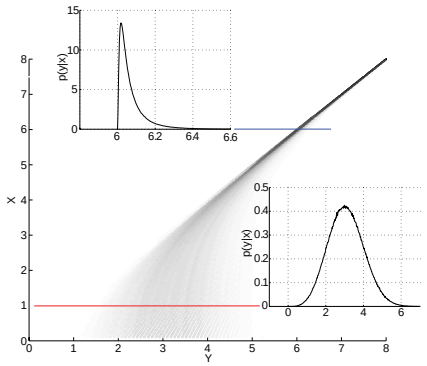

Figure 2: Joint distribution $p(x, y)$.

$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \check{\mathcal{M}})$. The joint distribution of the clean speech, x, and the noisy speech, y, in the log-spectral domain, where it is assumed that $y = \log(\exp(x) + \exp(n))$, shown in figure 2, illustrates this highly non-linear relationship and how the resultant conditional distributions may be non-Gaussian. Nevertheless, to model this complexity, uncertainty decoding schemes may represent the acoustic space with a GMM. From the observed noisy data, only the most likely component n is selected for efficiency and the form of distribution, as specified by that component, is passed to the recogniser. For both the `SPLICE` with uncertainty and the `Joint` schemes, the corrupted speech likelihood for state θ_t can be expressed as [3]

$$p(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t) \propto \sum_{m\in\theta_t} c_m \mathcal{N}\left(\boldsymbol{A}^{(n)}\boldsymbol{y}_t + \boldsymbol{b}^{(n)}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)} + \boldsymbol{\Sigma}_{\mathrm{b}}^{(n)}\right) \quad (2)$$

where $\boldsymbol{A}^{(n)}$, $\boldsymbol{b}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathrm{b}}^{(n)}$ are derived using different approximations in the `SPLICE` with uncertainty and `Joint` schemes.

`SPLICE` with uncertainty [2] makes use of Bayes' rule to express the conditional probability of the corrupted speech given the clean speech in terms of the conditional probability of the clean speech given the noisy. This requires an approximation for the clean speech distribution. A single global Gaussian is used, with mean, $\bar{\mu}_{x,i}$, and variance, $\bar{\sigma}^2_{x,i}$, for dimension i, and restricting $\boldsymbol{A}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathrm{b}}^{(n)}$ to be diagonal, gives

$$a_{ii}^{(n)} = \frac{\bar{\sigma}^2_{x,i}}{\bar{\sigma}^2_{x,i} - \check{\sigma}_i^{(n)2}}, \quad \sigma_{\mathrm{b}i}^{(n)2} = a_{ii}^{(n)}\check{\sigma}_i^{(n)2} \quad (3)$$

$$b_i^{(n)} = a_{ii}^{(n)}\left(\check{\mu}_i^{(n)} - \left(\check{\sigma}_i^{(n)2}/\bar{\sigma}^2_{x,i}\right)\bar{\mu}_{x,i}\right) \quad (4)$$

for dimension i. The parameters $\check{\mu}_i^{(n)}$ and $\check{\sigma}_i^{(n)2}$ are the means and variance respectively of $(x_{ti} - y_{ti})$ for the data associated with component n of the front-end GMM. In order to ensure that the uncertainty variance bias, $\boldsymbol{\Sigma}_{\mathrm{b}}^{(n)}$, is positive, the denominator in equation 3 is floored. In this work, the floor is set to a fraction α of the global clean variance, $\bar{\sigma}^2_{x,i}$, effectively ceiling $a_{ii}^{(n)}$ where

$$a_{ii}^{(n)} = \min\left(1/\alpha \ , \ \bar{\sigma}^2_{x,i}/\left(\bar{\sigma}^2_{x,i} - \check{\sigma}_i^{(n)2}\right)\right) \quad (5)$$

The next section discusses the effects of this in more detail.

In the front-end version of the `Joint` uncertainty decoding scheme, the joint distribution of the clean and corrupted speech is extracted. For component n of the front-end GMM the joint distribution is assumed to be Gaussian with parameters

$$\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y}_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x^{(n)} \\ \boldsymbol{\mu}_y^{(n)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x^{(n)} & \boldsymbol{\Sigma}_{xy}^{(n)} \\ \boldsymbol{\Sigma}_{yx}^{(n)} & \boldsymbol{\Sigma}_y^{(n)} \end{bmatrix}\right) \quad (6)$$

The compensation parameters are then given by

$$\boldsymbol{A}^{(n)} = \boldsymbol{\Sigma}_x^{(n)}\boldsymbol{\Sigma}_{yx}^{(n)\text{-}1}, \quad \boldsymbol{b}^{(n)} = \boldsymbol{\mu}_x^{(n)} - \boldsymbol{A}^{(n)}\boldsymbol{\mu}_y^{(n)} \quad (7)$$

$$\boldsymbol{\Sigma}_{\mathrm{b}}^{(n)} = \boldsymbol{A}^{(n)}\boldsymbol{\Sigma}_y^{(n)}\boldsymbol{A}^{(n)\mathsf{T}} - \boldsymbol{\Sigma}_x^{(n)} \quad (8)$$

Though the feature transform and variance bias may be full for the `Joint` scheme, they are typically made diagonal for efficiency.

When applying compensation schemes, it is important to assess the computational cost. Normally the computation of the front-end uncertainty parameters is dwarfed by the cost of applying the variance bias to the large number of model variances. This update is simply the addition of the variance bias, plus recomputing the cached determinant at a cost of $\mathcal{O}(d)$. While this compares favourable with model-based schemes, which are typically $\mathcal{O}(d^2)$ due to full matrix operations, unfortunately for front-end uncertainty schemes, the bias varies every time the front-end component changes, rather than when the acoustic environment changes.

3. Limitations of Front-End Schemes

There is a fundamental issue associated with front-end uncertainty decoding. Consider the joint distribution of the clean speech and noise shown in figure 2. Two conditional distributions, $p(y|x)$, are marked. The first is when there is a relatively high SNR ($x = 6$) yielding a highly skewed distribution that heavily peaks around $x = 6$. As the SNR increases this becomes more pronounced until it reaches a delta function, which does not affect the clean speech distribution when substituted in equation 1. In low SNR , e.g. $x = 1$ the conditional is very different. Here the distribution is exactly the same as the corrupting noise distribution, in this case a Gaussian distribution with mean 3 and variance of 1

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \check{\mathcal{M}}) \approx \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}) \quad (9)$$

where $\boldsymbol{\mu}_{\mathrm{n}}$ and $\boldsymbol{\Sigma}_{\mathrm{n}}$ are the noise mean and variance respectively. Although this result is not novel, the implication for uncertainty decoding has not been discussed in the literature. That is, substituted this into equation 1, the distribution of the corrupted speech is the same as the noise distribution

$$\begin{aligned} p(\boldsymbol{y}_t|\mathcal{M}, \check{\mathcal{M}}, \theta_t) &\approx \int \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}) p(\boldsymbol{x}_t|\mathcal{M}, \theta_t) d\boldsymbol{x}_t \\ &= \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}) \end{aligned} \quad (10)$$

This means irrespective of the recognition model component the distribution is always the same. For any front-end scheme, using this framework, where a single conditional is estimated and applied to all model components, in low SNR conditions a frame, or worse a sequence of frames, will have no discriminatory power between classes. With additional external restraints, such as a strong language model, these non-discriminatory regions are manageable; however for situations where language model constraints are weak, for example recognising digit strings in AURORA, the lack of discriminatory information can result in a large number of insertions.

This is clearly illustrated with the front-end `Joint` uncertainty decoding algorithm presented in [3]. Figure 3 shows the clean, corrupted and front-end `Joint` estimate, given by $\boldsymbol{A}^{(n)}\boldsymbol{y}_t + \boldsymbol{b}^{(n)}$, and the bias standard deviation, obtained from $\boldsymbol{\Sigma}_{\mathrm{b}}^{(n)}$, for a simple 16-component system. In regions of higher energy speech, e.g. frames 180 to 190, $\boldsymbol{\Sigma}_{\mathrm{b}}^{(n)}$ is small. But in low SNR, as in frames 225 to 230, the variance is off the scale, as is the `Joint` estimate of the value. These large variance values are

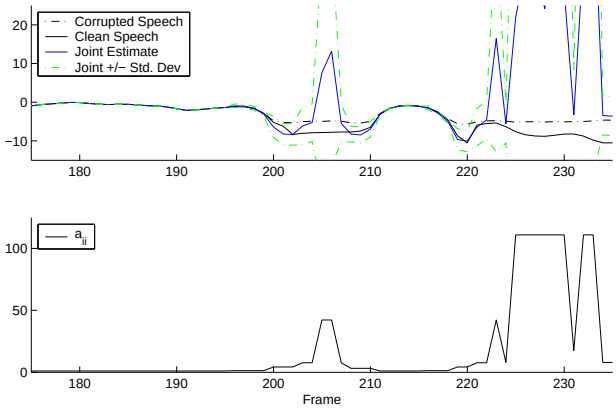

Figure 3: Plot of log energy for snippet from AURORA digit string 8-6-Zero-1-1-6-2, showing joint estimate, a_{ii} and variance bias.

reflected in the large values of a_{ii} associated with those regions. In frames 225 to 230 the value is around 100. With greater numbers of front-end components, these effects are amplified.

The reason for these very large values of a_{ii}, and associated variance biases, becomes clear when examining what happens in low energy speech regions to the joint distribution in equation 6. In these SNR regions, the corrupted speech distribution will be dominated by the noise (the standard masking effect), and the cross-variance term $\mathbf{\Sigma}_{xy}^{(n)}$ becomes

$$\mathbf{\Sigma}_{xy}^{(n)} = \mathcal{E}\left\{(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(n)})(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(n)})^{\mathsf{T}}\right\} \approx \mathbf{0} \tag{11}$$

The clean speech and the corrupted speech will be uncorrelated. From equation 7, this lack of correlation drives $\boldsymbol{A}^{(n)}$ to infinity. Given equation 11, the relationship to equation 10 becomes clearer by re-expressing equation 2, for component m, as

$$\begin{aligned} p(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t,m) = &\mathcal{N}\big(\boldsymbol{y}_t; \mathbf{\Sigma}_{yx}^{(n)}\mathbf{\Sigma}_x^{(n)\text{-}1}\big(\boldsymbol{\mu}^{(m)} - \boldsymbol{\mu}_x^{(n)}\big) + \boldsymbol{\mu}_y^{(n)}, \\ &\mathbf{\Sigma}_{yx}^{(n)}\mathbf{\Sigma}_x^{(n)\text{-}1}\big(\mathbf{\Sigma}^{(m)} - \mathbf{\Sigma}_x^{(n)}\big)\mathbf{\Sigma}_{yx}^{(n)}\mathbf{\Sigma}_x^{(n)\text{-}1} + \mathbf{\Sigma}_y^{(n)}\big) \\ = &\mathcal{N}\big(\boldsymbol{y}_t; \boldsymbol{\mu}_y^{(n)}, \mathbf{\Sigma}_y^{(n)}\big) \end{aligned} \tag{12}$$

which in low SNR is simply the noise distribution. Hence, leaving $\boldsymbol{A}^{(n)}$ unconstrained may result in large numbers of insertions.

Thus, in the `Joint` scheme, it is inevitable that in low SNR the correlation matrix, and hence the covariance matrix, will tend to zero yielding extreme compensation parameters in equation 7. This is the correct behaviour given the simple assumptions made for efficiency, however to prevent these extremes it would be sensible to limit the possible values for the compensation parameters. The obvious approach is to examine the correlation coefficients

$$\rho_{xy,i}^{(n)} = \sigma_{xy,i}^{(n)}\left(\sqrt{\sigma_{x,i}^{(n)2}\sigma_{y,i}^{(n)2}}\right)^{\text{-}1} \tag{13}$$

for dimension i. The compensation parameters estimates in equation 7 can then be re-expressed in terms of this coefficient as

$$a_{ii}^{(n)} = \frac{\sigma_{x,i}^{(n)}}{\rho_{xy,i}^{(n)}\sigma_{y,i}^{(n)}}, \qquad \sigma_{\mathsf{b},i}^{(n)2} = \frac{\sigma_{x,i}^{(n)2}}{\rho_{xy,i}^{(n)2}} - \sigma_{x,i}^{(n)2} \tag{14}$$

for the diagonal form of front-end `Joint` uncertainty decoding. The compensation parameters can then be restricted by enforcing a minimum on the correlation coefficient used in 14 as follows

$$\hat{\rho}_{xy,i}^{(n)} = \max(\rho_{xy,i}^{(n)}, \rho) \tag{15}$$

where ρ is an empirically set constant. Increasing the value of ρ raises the minimum acceptable correlation, decreasing the maximum variance bias. In the limit, it is possible to set $\rho = 1$, resulting in a zero variance bias in equation 14. The effects of this flooring on the same snippet as figure 3 is shown in figure 4. As anticipated, the extremes previously observed have disappeared.

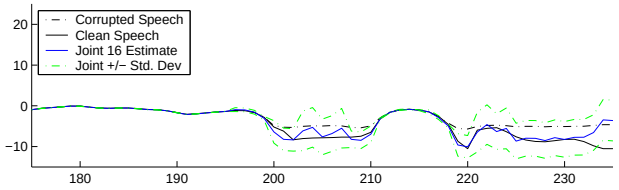

Figure 4: Plot of log energy for AURORA digit string 8-6-Zero-1-1-6-2, with correlation flooring, $\rho = 0.1$.

As this issue of all distributions becoming the same theoretically affects all front-end uncertainty decoding schemes, the `SPLICE` with uncertainty decoding should also suffer from it. However the expected issues have not been observed, for example on the AURORA results presented in [2]. This is because `SPLICE` with uncertainty limits the maximum value of $a_{ii}^{(n)}$ to $1/\alpha$ in equation 5. There is also an under-estimate of the value of $a_{ii}^{(n)}$. In order to make the calculation of the `SPLICE` uncertainty efficient, a global variance is used in the denominator of equation 3. Since this will be larger than any individual component that should be used, the scaling estimate will be lower than expected. This under-estimation becomes larger as the number of front-end GMM components increases—exactly the when a component might expected to only be associated with a low-energy noise region. These two limiting factors on $a_{ii}^{(n)}$, keep the uncertainty from becoming too large in `SPLICE` with uncertainty and causing this issue.

4. Model-Based `Joint` Uncertainty Decoding

All front-end uncertainty decoding schemes may result in regions of no discrimination because only a single set of compensation parameters is propagated from the front-end model to the recogniser. Model-based schemes do not suffer from this problem as the effective set of compensation parameters propagated to the recogniser is explicitly linked to the recognition component. With model-based `Joint` [3] transforms, instead of linking front-end of components to regions of the feature space, each is associated with a set of recognition model components. Similar to the front-end `Joint` scheme, the joint distribution of the clean and corrupted speech features are required. For example, the cross-covariance terms between the clean and corrupted speech are given by

$$\mathbf{\Sigma}_{xy}^{(r)} = \frac{\sum_{m\in r_m}\gamma_m(t)\boldsymbol{x}_t\boldsymbol{y}_t^{\mathsf{T}}}{\sum_{m\in r_m}\gamma_m(t)} - \boldsymbol{\mu}_x^{(r)}\boldsymbol{\mu}_y^{(r)\mathsf{T}} \tag{16}$$

where $\gamma_m(t)$ is the component posterior at time instance t and r_m is the set of recognition components associated with component r.

Having obtained the component parameters, the compensation parameters can be derived using equations 7 and 8. During recognition, in contrast to the front-end `Joint` scheme, all the front-end components are active and pass their measure of uncertainty to the recogniser. This operation is similar to using a multiple-transform constrained MLLR scheme [6], but with the addition of a variance bias. The model-based scheme is actually more efficient than the front-end, since the variance bias applied to the recognition model-set is fixed given a particular acoustic environment, in

contrast to the front-end scheme where it will vary if either the acoustic environment or the front-end component changes. As all the front-end components are active, if one of them is associated with a low energy region, such that $\mathbf{\Sigma}_{xy}^{(r)}$ is very small, then this will only affect the recognition components in class r_m, not all the recognition components. Thus there is no problem with regions lacking discrimination between classes.

5. Experiments

Experiments were conducted on the standard AURORA 2.0 small vocabulary digit string recognition task [5]. The reference acoustic model setup was used with an internal version of HTK 3.3 and its native front-end processing; this resulted in slight differences from HTK 2.2. Compensation parameters were estimated using stereo data. This allows the techniques to be assessed without considering inaccuracies arising from noise estimation, or approximations in the mismatch function. In practice, the compensation parameters can be estimated using PMC or VTS style schemes. The front-end uncertainty schemes used diagonal transformations.

System	SNR(dB) 20	15	10	5
Clean	4.62	12.20	31.13	59.16
Matched	1.85	2.81	5.01	11.41
SPLICE	1.95	3.07	6.13	16.47
+Uncertainty, $\alpha = 0.1$	2.15	3.22	5.95	14.50
+Uncertainty, $\alpha = 0.95$	2.00	3.20	5.58	12.29
FE-Joint	22.67	25.82	28.38	34.37
FE-Joint, $\rho = 0.9$	1.81	2.88	5.71	14.62

Table 1: Clean, matched and SPLICE on AURORA 2.0 test set A, averaged across N1-N4, WER(%).

Table 1 shows baseline, SPLICE and Joint systems' performance. The 256-component SPLICE systems approach matched performance, significantly improving the poor clean system. To investigate the effects of the flooring α, from equation 5, on SPLICE with uncertainty, a range of values of were tried. Performance at the cited value of $\alpha = 0.1$ in [2], can be improved slightly by increasing it to 0.95. This can be compared against the front-end Joint scheme. As expected, without flooring ρ, a vast number of insertions occur. In contrast, this behaviour was not seen on Resource Management [3] because of the constraining language model. With ρ set to 0.9, the performance is now comparable to the various SPLICE systems. Both these optimal flooring values significantly reduce the uncertainty passed to the decoder.

Table 2 summarises the results of the model-based Joint approach. The number and form of transforms were explored, where the diagonal transforms are similar to the front-end schemes and contrasted with full matrix forms of $\boldsymbol{A}^{(r)}$ and $\mathbf{\Sigma}_{\mathrm{b}}^{(r)}$. A 16 transform model-based Joint scheme performed slightly worse than appropriately floored 256-component front-end schemes, but at considerably less computational cost; with the same number of diagonal transforms, the model-based system is superior to all of the front-end systems examined. Moreover, using a full transform gave substantial gains. In low SNR, the 16 full transform model-based system is better than matched. However as the variance bias is a full matrix, there is the impractical cost of performing a full covariance matrix decode, compared to the diagonal covariance matched system. This does indicate an opportunity to obtain excellent results using this model-based Joint approach.

System	Number of Transforms	SNR(dB) 20	15	10	5
Diagonal Transformations					
M-Joint	1	3.33	5.92	13.35	31.96
	16	2.47	3.82	7.25	16.63
	256	1.90	2.73	5.19	12.00
Full Transformations					
M-Joint	1	2.43	3.82	6.97	17.14
	16	1.95	2.80	4.23	9.89

Table 2: Model-based Joint systems' performance on AURORA 2.0 test set A, averaged across N1-N4, WER(%).

6. Conclusions

This paper has discussed important differences within the uncertainty decoding framework between front-end and model-based approaches. In the former, by only propagating a single vector of features and probabilities, during high noise the ability to effectively discriminate can be lost. This causes insertion errors in the search if all models are rendered acoustically equivalent. With another source for discrimination, such as a language model, this can be less of an issue as it guides the search when the SNR is low and uncertainty is high. This issue was explored on the AURORA task, which practically has no language constraints, using the Joint form of uncertainty decoding, where it was found that flooring the correlation was beneficial, and the SPLICE with uncertainty form, which implicitly floors uncertainty parameters. However, model-based schemes are not affected by this problem, hence better results were obtained than in front-end systems with equivalent numbers of parameters. The best system was the model-based Joint scheme with full matrix parameters; though this increases the decoding computational cost, it does indicate the possible benefits of this framework. Major limitations of this paper are that experiments are conducted on artificially corrupted data and assume noise stationarity; however, recent work has explored using this Joint form on found data such as Broadcast News [7].

7. References

[1] T. T. Kristjansson and B. J. Frey, "Accounting for uncertainty in observations: A new paradigm for robust speech recognition," in *Proc. ICASSP*, Orlando, Florida, May 2002.

[2] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. ICASSP*, Orlando, Florida, May 2002.

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

[4] L. Deng, J. Droppo, and A. Acero, "Exploiting variances in robust feature extraction based on a parametric model of speech distortion," in *Proc. ICSLP*, 2002.

[5] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the evaluation of speech recognition systems under noisy conditions," in *Proc. ASR-2000*, 2000.

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

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