Issues with uncertainty decoding for noise robust automatic speech recognition

H. Liao *,[1], M.J.F. Gales *

Engineering Department, Cambridge University, Trumpington Street, Cambridge CB2 1PZ, United Kingdom

Abstract

Interest continues in a class of robustness algorithms for speech recognition that exploit the notion of uncertainty introduced by environmental noise. These techniques share the property that the uncertainty varies with the noise level and is propagated to the decoding stage, resulting in increased model variances. In observation uncertainty forms, the uncertainty variance is simply the variance of the error in enhancement that is added to the model variances. Another form, called uncertainty decoding, refers to a factorisation which results in a linear feature transform and model variance bias that increases with noise; using appropriate approximations, efficient implementations may be obtained, with the goal of achieving near model-based performance without the associated computational cost. Unfortunately, uncertainty decoding forms that compute the uncertainty in the front-end and pass this to the decoder may suffer from a theoretical problem in low signal-to-noise ratio conditions. This report discusses how this fundamental issue arises, and demonstrates it through two schemes: `SPLICE` with uncertainty and front-end joint uncertainty decoding (`FE-Joint`). A method to mitigate this for `FE-Joint` compensation is presented, as well as how `SPLICE` implicitly addresses it. However, it is shown that a model-based joint uncertainty decoding approach does not suffer from this limitation, like these front-end forms do, and is more computationally attractive. The issues described and performance of the various schemes are examined on two artificially corrupted corpora: the AURORA 2.0 digit string recognition and 1000-word Resource Management tasks.

Key words: Speech recognition, Uncertainty decoding, Noise robustness, AURORA2

1. Introduction

Improving the noise robustness of state-of-the-art automatic speech recognisers continues to be an important research area. Current continuous density HMM systems perform well on clean, uncorrupted speech, but in practice falter in noisy usage conditions. *Model-based compensation* techniques, which update the acoustic model means and variances, such as Parallel Model Combination (PMC) (Gales, 1995), vector Taylor series-based (VTS) compensation (Kim et al., 1998; Acero et al., 2000) and more recently `ALGONQUIN` (Kristjansson and Frey, 2002) have shown to be very effective forms of environmental noise compensation. Unfortunately, these are computationally expensive compared to less powerful front-end *feature enhancement* techniques like spectral subtraction and cepstral mean normalisation, that only compensate the noisy speech features.

Recently, research has focused on extending feature-based schemes by incorporating the uncertainty due to noise into the recognition search. *Observation uncertainty* forms, as termed in this paper, do so in an ad hoc fashion, by adding an un-

* Corresponding author.
Email addresses: `hl251@eng.cam.ac.uk` (H. Liao), `mjfg@eng.cam.ac.uk` (M.J.F. Gales).
URL: `http://mi.eng.cam.ac.uk/~hl251` (H. Liao).

[1] Supported by Toshiba Research Europe Limited.

certainty variance to HMM variances to represent the residual observation uncertainty after enhancement. The uncertainty, for example, may be based on the position of formants (Holmes et al., 1997), from a polynomial function of the signal-to-noise ratio (Arrowood and Clements, 2002), or the variance of the enhancement process (Deng et al., 2002; Benítez et al., 2004; Stouten et al., 2004; Deng et al., 2005; Wölfel and Faubel, 2007). In contrast, *uncertainty decoding* may be viewed as propagating the conditional probability of the corrupted speech, given the "clean" speech, into the decoding stage (Droppo et al., 2002; Liao and Gales, 2005). These two distinctly different approaches yield a similar form: the uncertainty is calculated efficiently in the front-end, and passed to the recogniser as a single, simple variance offset to the recognition model components. This can provide an elegant compromise of a fast feature-based compensation scheme with model-based accuracy.

Observation uncertainty techniques have demonstrated good results for a variety of tasks and environments, however they suffer from some inherent flaws. Observation uncertainty has been derived intuitively; some results show that the variances are ill-conditioned (Stouten et al., 2004; Deng et al., 2002) and gains disappear with adaptation (Wölfel and Faubel, 2007) . For front-end uncertainty decoding Droppo et al. (2002), there is a fundamental problem in low SNR when the noise masks the speech (Liao and Gales, 2006). Since a transform is selected in the front-end, which modifies all the model variances during decoding, in low SNR all the models may be altered to be the same. Hence no discrimination is possible, and if there are no other constraints such as a strong language model, then large numbers of insertions can take place in these areas of high uncertainty. Because model-based uncertainty decoding explicitly associates the corrupted speech conditional with the clean speech distribution, it does not suffer from this problem.

This paper discusses these issues in detail. Section 2 describes some feature enhancement schemes and how they can be extended to include uncertainty. This extension gives a similar decoding form to the front-end uncertainty decoding method reviewed in section 3. The fundamental problem with front-end uncertainty decoding is elaborated on in section 4 and demonstrated through the `SPLICE` with uncertainty and front-end joint uncertainty decoding techniques. We show that model-based uncertainty decoding does not suffer from this problem, is more effective, and at least as efficient as front-end forms. In chapter 5, these issues and techniques are evaluated on two artificially corrupted corpora: the often used small vocabulary AURORA 2.0 noisy digit string recognition task (Hirsch and Pearce, 2000) and the 1000-word Resource Management command and control database (Price et al., 1988). Overall conclusions and future work directions are presented in section 6.

2. Feature enhancement

Traditionally, fast front-end noise compensation techniques, such as spectral subtraction and more recently `SPLICE` (Deng et al., 2000), have removed the noise from the observed noisy, corrupted speech vector $\boldsymbol{y}_t$, and passed this estimate $\hat{\boldsymbol{x}}_t$ as if it were exactly the original clean speech vector $\boldsymbol{x}_t$ to the acoustic models as shown in Fig. 1. Hence, the decoding likelihood, at time frame t, is simply the evaluation of the enhanced clean speech against the uncompensated clean acoustic models

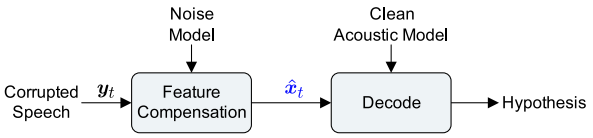

Fig. 1. The standard feature enhancement process.

$$p\left(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t\right) = p\left(\hat{\boldsymbol{x}}_t|\mathcal{M},\theta_t\right) \tag{1}$$

where $\mathcal{M}$ represents the set of clean acoustic model parameters, $\check{\mathcal{M}}$ represents some set of front-end parameters which may include noise or simplified speech models, and θ_t denotes the hidden clean speech state.

Simple enhancement schemes tend to be very efficient and take the following form

$$\hat{\boldsymbol{x}}_t = \mathcal{E}\left\{\boldsymbol{x}_t|\boldsymbol{y}_t,\check{\mathcal{M}}\right\} \tag{2}$$

The enhancement is based on the front-end parameters $\check{\mathcal{M}}$ and is usually independent of the actual acoustic model parameters $\mathcal{M}$; if the size of $\check{\mathcal{M}}$ is small, the enhancement may be fast. A recent algorithm, based on this framework, is called `SPLICE`. `SPLICE` partitions the noisy acoustic space using a front-end GMM with N components. The expected value of the clean speech given the corrupted speech is modeled by a piece-wise, linear function depending on the region in the acoustic space

$$\hat{\boldsymbol{x}}_t \approx \sum_n P(n|\boldsymbol{y}_t, \check{\mathcal{M}}) \mathcal{E}\left\{\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}}, n\right\} \quad (3)$$

with the posterior distribution of component n defined as

$$P\big(n|\boldsymbol{y}_t, \check{\mathcal{M}}\big) = \frac{\check{c}_n p\big(\boldsymbol{y}_t|\check{\mathcal{M}}, n\big)}{\sum_{i=1}^{N} \check{c}_i p\big(\boldsymbol{y}_t|\check{\mathcal{M}}, i\big)} \quad (4)$$

where $\check{c}_n$ is the component prior. The expected value of the clean speech posterior $\mathcal{E}\left\{\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}}, n\right\}$ is the mean of

$$p\big(\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}}, n\big) \approx \mathcal{N}\big(\boldsymbol{x}_t; \boldsymbol{y}_t + \check{\boldsymbol{\mu}}^{(n)}, \check{\boldsymbol{\Sigma}}^{(n)}\big) \quad (5)$$

which leads to

$$\mathcal{E}\left\{\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}}, n\right\} \approx \boldsymbol{y}_t + \check{\boldsymbol{\mu}}^{(n)} \quad (6)$$

Often the SPLICE form in equation 3 is simplified by only applying the bias associated with the most likely component n^* of the corrupted speech Gaussian mixture model (GMM)

$$\hat{\boldsymbol{x}}_t = \mathcal{E}\left\{\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}}, n^*\right\} \quad (7)$$

$$= \boldsymbol{y}_t + \check{\boldsymbol{\mu}}^{(n^*)} \quad (8)$$

Here, the corrupted speech is updated simply by the bias vector $\check{\boldsymbol{\mu}}^{(n^*)}$, which is the expected value of difference between the clean and corrupted speech, associated with the most probable region n^* in the acoustic space. This is then used during decoding as representative of the clean speech feature vector. The corrupted acoustic space GMM with N components is given by

$$p\big(\boldsymbol{y}_t|\check{\mathcal{M}}\big) = \sum_{n=1}^{N} \check{c}_n p\big(\boldsymbol{y}_t|\check{\mathcal{M}}, n\big) \quad (9)$$

$$= \sum_{n=1}^{N} \check{c}_n \mathcal{N}\Big(\boldsymbol{y}_t; \boldsymbol{\mu}_y^{(n)}, \boldsymbol{\Sigma}_y^{(n)}\Big) \quad (10)$$

and the most likely component is determined by

$$n^* = \arg\max_n \Big[\check{c}_n P\big(\boldsymbol{y}_t|\check{\mathcal{M}}, n\big)\Big] \quad (11)$$

The correction vectors can be estimated using stereo data [2] with the following formula

$$\check{\boldsymbol{\mu}}^{(n)} = \mathcal{E}\left\{\boldsymbol{x}_t - \boldsymbol{y}_t|n\right\} \quad (12)$$

$$\check{\boldsymbol{\Sigma}}^{(n)} = \mathcal{E}\left\{(\boldsymbol{x}_t - \boldsymbol{y}_t)(\boldsymbol{x}_t - \boldsymbol{y}_t)^{\mathsf{T}}|n\right\} - \check{\boldsymbol{\mu}}^{(n)}\check{\boldsymbol{\mu}}^{(n)\mathsf{T}} \quad (13)$$

SPLICE has shown to be quite an effective compensation algorithm on the standard AURORA corpus (Droppo et al., 2001). It is efficient since the feature update is fast. The main cost is the N Gaussian evaluations in equation 11 to choose the acoustic region and associated correction bias.

2.1. *Observation uncertainty*

Traditional standard enhancement schemes that only update the feature vector and pass this on as if it were the true clean speech to the decoder, have recently been extended to reflect uncertainty of the de-noising process itself. Instead of assuming the feature cleaning process is exact, the posterior $p(\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}})$ is passed to the acoustic models, representing the uncertainty in the compensation. The mean of this distribution is the estimate $\hat{\boldsymbol{x}}$, but with an associated variance that may be the expected square error of the enhancement. This paper refers to this approach as *observation uncertainty*, although it has also been called uncertain observations or uncertain observation decoding in Arrowood and Clements (2002) and more confusingly referred to as uncertainty decoding, which is superficially similar but fundamentally different. This scheme is depicted in Fig. 2 which can be compared with Fig. 1.

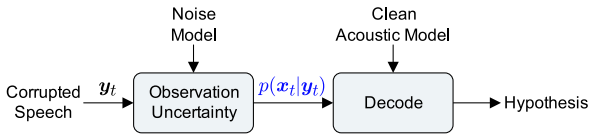

Fig. 2. The observation uncertainty form.

Thus if the clean speech feature vector is now considered a multivariate Gaussian distribution $\boldsymbol{x}_t \sim \mathcal{N}(\hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}})$, the decoding likelihood requires integration over all possible values

$$p\big(\boldsymbol{y}_t|\mathcal{M}, \hat{\mathcal{M}}, \theta_t, m\big)$$

$$\approx \int_{\mathcal{R}^d} p(\boldsymbol{x}_t|\boldsymbol{y}_t, \check{\mathcal{M}}) p(\boldsymbol{x}_t|\mathcal{M}, \theta_t) d\boldsymbol{x}_t \quad (14)$$

$$= \int_{\mathcal{R}^d} \mathcal{N}(\boldsymbol{x}_t; \hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}}) \mathcal{N}\Big(\boldsymbol{x}_t; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)}\Big) d\boldsymbol{x}_t \quad (15)$$

$$= \mathcal{N}\Big(\hat{\boldsymbol{x}}_t; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)} + \boldsymbol{\Sigma}_{\hat{\mathbf{x}}}\Big) \quad (16)$$

[2] In this paper, stereo data refers to two parallel channels of audio, with one providing a recording of clean speech data and the other a noisy version of the exact same acoustic speech.

Here $\hat{\boldsymbol{x}}_t$ is again the clean speech estimate, as is normally produced from standard enhancement schemes; the parameters $\boldsymbol{\mu}^{(m)}$ and $\boldsymbol{\Sigma}^{(m)}$ denote the mean and variance of Gaussian m in the clean speech acoustic model $\mathcal{M}$ used for decoding. The variance offset $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}$ is the expected square error of this enhancement process. In SPLICE this is

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}} = \check{\boldsymbol{\Sigma}}^{(n^*)} \tag{17}$$

and can be estimated from stereo data as in equation 13. Other enhancement schemes can be easily extended to provide this variance, for example considering formant frequencies as part of a heuristic measure (Holmes et al., 1997); using a weighted polynomial function of the SNR in the log spectral domain Arrowood and Clements (2002); or obtaining them from a parametric model of the clean speech (Deng et al., 2002), the classic Weiner filter (Benítez et al., 2004) or a particle filter (Wölfel and Faubel, 2007).

An interesting observation uncertainty form is the model-based feature enhancement (MBFE) technique extended to account for observation uncertainty (Stouten et al., 2004). It is notable because like the front-end joint uncertainty decoding form discussed in the next section, a GMM is embedded in the front-end and a joint distribution between the clean and corrupted speech is computed for each component. MBFE differs in that the joint distribution is used to compute the clean speech posterior, where the clean speech estimate for a particular component n is

$$\hat{\boldsymbol{x}}_t^{(n)} = \boldsymbol{\mu}_x^{(n)} + \boldsymbol{\Sigma}_{xy}^{(n)} \left(\boldsymbol{\Sigma}_y^{(n)}\right)^{-1} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(n)}\right) \tag{18}$$

and the associated variance or uncertainty

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}^{(n)} = \boldsymbol{\Sigma}_x^{(n)} - \boldsymbol{\Sigma}_{xy}^{(n)} \left(\boldsymbol{\Sigma}_y^{(n)}\right)^{-1} \boldsymbol{\Sigma}_{yx}^{(n)} \tag{19}$$

In this form of enhancement, as described in Stouten et al. (2004),the final clean speech estimate is formed by summing over all the estimates, weighted by the component posterior, rather than just choosing the most likely state as in the SPLICE form.

The computational cost of using observation uncertainty is similar to standard enhancement scheme. However the uncertainty that is propagated for the current frame must be added to all acoustic model components. This can total in the hundreds for a small task, such as in the reference AURORA recogniser (Hirsch and Pearce, 2000), to the hundred of thousands in state-of-the-art recognition systems such as the CU Broadcast News system (Kim et al., 2003). Moreover, this variance addition is not as simple to apply as, for example, a scaling of the variances—the Gaussian normalisation term that is usually cached must be re-computed. As stated in Arrowood and Clements (2002), assuming that Gaussian evaluations comprise 50% of the total computation cost of transcribing speech, the overhead of adding uncertainty is approximately 33%. Nevertheless, applying this variance update with a single global uncertainty is far cheaper than expensive model-based techniques such as VTS compensation, PMC or ALGONQUIN which separately update each acoustic model component individually depending on the effects of the noise on that Gaussian.

While in practice good results have been obtained using observation uncertainty (Benítez et al., 2004; Deng et al., 2002; Stouten et al., 2004; Wölfel and Faubel, 2007), there is a concern. Despite the presumption that enhanced observations may not be exact seems sensible, the resulting decoding form given in equation 14 do not appear to arise from any mathematical framework. Perhaps this is why the variances propagated seem ill-conditioned: they are deemed too large and imprecise (Deng et al., 2002), reduced by a factor of ten (Stouten et al., 2004), hurt performance compared to standard enhancement in higher SNR (Benítez et al., 2004), or give improvements which disappear with adaptation (Wölfel and Faubel, 2007). Hence, although gains have been obtained using this technique, it may be viewed as a heuristic, ad hoc approach.

3. Uncertainty decoding

In this section, the uncertainty decoding framework from (Droppo et al., 2002; Liao and Gales, 2005) is described along with two forms that exemplify it. A model of the corrupted speech as a function of the clean speech and noise can be expressed by a dynamic Bayesian network (DBN) as shown in Fig. 3. Here, the corrupted speech observation $\boldsymbol{y}_t$ is

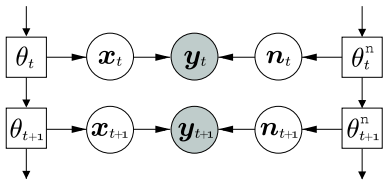

Fig. 3. Uncertainty decoding DBN.

assumed to be conditionally independent of all other observations given the clean speech $\boldsymbol{x}_t$ and the noise $\boldsymbol{n}_t$. The clean speech and noise are assumed to be generated by HMMs with states θ_t^{n} for the noise[3] and θ_t for the clean speech. Under these assumptions the corrupted speech likelihood is given by

$$p(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t)=\int_{\mathcal{R}^d} p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})p(\boldsymbol{x}_t|\mathcal{M},\theta_t)d\boldsymbol{x}_t \quad (20)$$

where

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}}) = \int_{\mathcal{R}^d} p(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{n}_t)p(\boldsymbol{n}_t|\check{\mathcal{M}},\theta_t^{\mathrm{n}})d\boldsymbol{n}_t \quad (21)$$

The likelihood calculation thus has two distinct parts. Only the first, $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})$, is a function of the noise, the other is the clean speech prior which is not dependent on the noise. Hence, this marginalisation is independent of the noise given $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})$. This uncertainty decoding framework can be depicted as shown in Fig. 4.

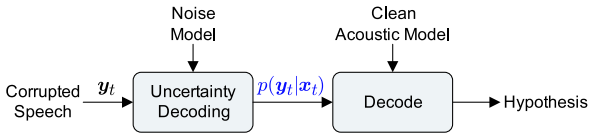

Fig. 4. Uncertainty decoding framework.

The term *uncertainty decoding* can be considered to encompass forms that exploit this factorisation by determining an efficient approximation for the conditional distribution $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})$ that easily completes the marginalisation and is cheap to compute. This distribution, conditioned on the clean speech, can be decoupled from the structure of the actual acoustic models. Thus there is significant freedom in choosing an appropriate form for this distribution that minimises the computational cost. Front-end uncertainty decoding forms decouple $\check{\mathcal{M}}$ and $\mathcal{M}$ completely and assure, through a series of approximations, that the information passed along to the decoding stage depends solely on the observed features. Model-based uncertainty maintains some coupling between the model components in $\check{\mathcal{M}}$ and $\mathcal{M}$, which leads to uncertainty information that is dependent on the "class" of the acoustic model component being passed to the decoder. In pure model-based approaches, such as ALGONQUIN, PMC or VTS compensation, the two distributions are fully tied by the clean speech variable. These explicitly compute the conditional for each model component, best reflecting the affect of noise on the model distributions, but at a significant computation cost.

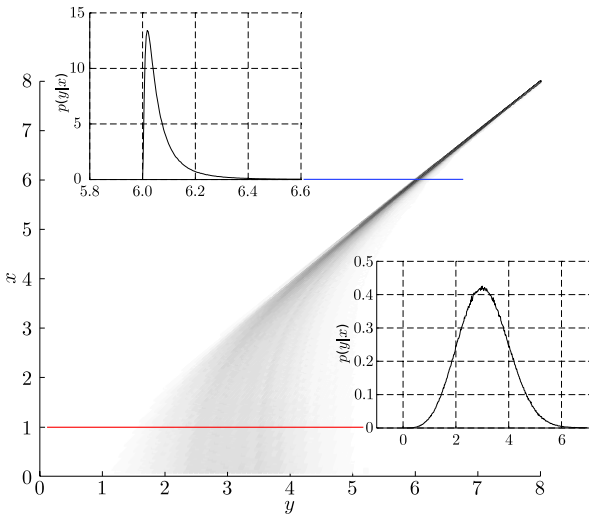

Fig. 5. Joint distribution $p(x,y)$. Additive noise is $\mathcal{N}(3,1)$.

3.1. *Front-end uncertainty decoding*

In front-end uncertainty decoding, a major focus is determining a form for the conditional distribution $p(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}})$ that can be efficiently computed, and is independent from the back-end acoustic models. The nature of the conditional can be explored by examining the joint clean-corrupted speech distribution as shown in Fig. 5. This simulation, which has also been detailed in Liao and Gales (2004) and Benítez et al. (2004), takes place in the log energy domain where x represents the clean speech and y the noise corrupted speech, where it is assumed that $y = \log(\exp(x) + \exp(n))$. The clean speech is a uniform distribution over $[0, 8]$ and the additive noise a constant Gaussian of mean 3 and variance 1. This relationship is highly non-linear especially in the low SNR region to the left and clearly non-Gaussian. Nevertheless, the approach taken in front-end uncertainty decoding is to represent the corrupted speech conditional given the clean speech with a GMM. By selecting a single, most probable component of the GMM given the observed noisy data, a single variance offset per frame is passed to the acoustic models during decoding. This is an approximation for efficiency, since not doing so would cause the complexity in the front-end process to multiply with the number of components in the back-end. The use of Gaussian distributions makes the marginalisation in equation 20 trivial. In this way,

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

an elegant compromise is achieved with fast front-end processing providing a simple acoustic model update.

Currently, two specific forms of front-end uncertainty decoding have been presented in the literature: SPLICE with uncertainty (Droppo et al., 2002) and the front-end joint uncertainty decoding (FE-Joint) method (Liao and Gales, 2005). For both, the resultant likelihood of the corrupted speech observation using the uncertainty parameters selected from component n of the front-end GMM can be expressed as

$$p(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t) \propto \sum_{m\in\theta_t} c_m \mathcal{N}\Big(\boldsymbol{A}^{(n)}\boldsymbol{y}_t+\boldsymbol{b}^{(n)};\boldsymbol{\mu}^{(m)},\boldsymbol{\Sigma}^{(m)}+\boldsymbol{\Sigma}_{\mathtt{b}}^{(n)}\Big) \quad (22)$$

where $\boldsymbol{A}^{(n)}$, $\boldsymbol{b}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathtt{b}}^{(n)}$ are the compensation parameters. In form, this is exactly the same as the observation uncertainty approach, as given in equation 16, with

$$\hat{\boldsymbol{x}} = \boldsymbol{A}^{(n)}\boldsymbol{y}_t+\boldsymbol{b}^{(n)} \quad (23)$$

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}} = \boldsymbol{\Sigma}_{\mathtt{b}}^{(n)} \quad (24)$$

but the parameters are derived from a fundamentally different perspective. SPLICE with uncertainty and the FE-Joint scheme have different methods of deriving these parameters. These two forms are discussed in further detail in the following sections.

3.1.1. *SPLICE with uncertainty*

SPLICE with uncertainty makes use of Bayes' rule to express the conditional probability of the corrupted speech given the clean speech in terms of the clean speech posterior distribution. For tractability, the denominator clean speech prior is modeled by a single Gaussian with the mean and variance for dimension i denoted by $\bar{\mu}_{x,i}$ and $\bar{\sigma}^2_{x,i}$. Using this approximation, with the restriction that matrices $\boldsymbol{A}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathtt{b}}^{(n)}$ are diagonal, gives

$$a_{ii}^{(n)} = \frac{\bar{\sigma}^2_{x,i}}{\bar{\sigma}^2_{x,i} - \check{\sigma}_i^{(n)2}} \quad (25)$$

$$b_i^{(n)} = a_{ii}^{(n)}\left(\check{\mu}_i^{(n)} - \frac{\check{\sigma}_i^{(n)2}}{\bar{\sigma}^2_{x,i}}\bar{\mu}_{x,i}\right) \quad (26)$$

$$\sigma_{\mathtt{b},i}^{(n)2} = a_{ii}^{(n)}\check{\sigma}_i^{(n)2} \quad (27)$$

The parameters $\check{\mu}_i^{(n)}$ and $\check{\sigma}_i^{(n)2}$ may estimated using equations 12 and 13. In order to ensure that the uncertainty variance bias $\boldsymbol{\Sigma}_{\mathtt{b}}^{(n)}$ is positive, the denominator in equation 25 is floored. In this work the floor is set to a fraction α of the global clean variance $\bar{\sigma}^2_{x,i}$. This floor effectively places a maximum value on $a_{ii}^{(n)}$ where

$$a_{ii}^{(n)} = \min\left(\frac{1}{\alpha}\ ,\ \frac{\bar{\sigma}^2_{x,i}}{\bar{\sigma}^2_{x,i} - \check{\sigma}_i^{(n)2}}\right) \quad (28)$$

The effects of this are discussed in section 4.

3.1.2. *Front-end joint uncertainty decoding*

In the front-end version of JUD, FE-Joint, the corrupted speech conditional distribution given the clean speech is directly modeled by a GMM. To derive this conditional, a joint distribution of the clean and corrupted speech is estimated for each region of the acoustic space. For component n of the front-end corrupted speech GMM the joint distribution is assumed to be Gaussian with parameters

$$\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{y}_t\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x^{(n)}\\ \boldsymbol{\mu}_y^{(n)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_x^{(n)} & \boldsymbol{\Sigma}_{xy}^{(n)}\\ \boldsymbol{\Sigma}_{yx}^{(n)} & \boldsymbol{\Sigma}_y^{(n)}\end{bmatrix}\right) \quad (29)$$

The joint distribution parameters may be estimated from stereo data or predicted using a model-based technique such as VTS compensation (Xu et al., 2006). From the joint distribution, the conditional distribution can be derived as follows

$$p\big(\boldsymbol{y}_t|\boldsymbol{x}_t,\check{\mathcal{M}},n\big) \quad (30)$$

$$\approx \mathcal{N}\Big(\boldsymbol{y}_t;\boldsymbol{\mu}_y^{(n)} + \boldsymbol{\Sigma}_{yx}^{(n)}\boldsymbol{\Sigma}_x^{(n)\text{-}1}\big(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(n)}\big), \boldsymbol{\Sigma}_y^{(n)} - \boldsymbol{\Sigma}_{yx}^{(n)}\boldsymbol{\Sigma}_x^{(n)\text{-}1}\boldsymbol{\Sigma}_{xy}^{(n)}\Big) \quad (31)$$

$$= \big|\boldsymbol{A}^{(n)}\big|\mathcal{N}\Big(\boldsymbol{\Sigma}_x^{(n)}\boldsymbol{\Sigma}_{yx}^{(n)\text{-}1}\big(\boldsymbol{y}_t-\boldsymbol{\mu}_y^{(n)}\big)+\boldsymbol{\mu}_x^{(n)};\boldsymbol{x}_t, \boldsymbol{\Sigma}_x^{(n)}\boldsymbol{\Sigma}_{yx}^{(n)\text{-}1}\boldsymbol{\Sigma}_y^{(n)}\boldsymbol{\Sigma}_x^{(n)}\boldsymbol{\Sigma}_{yx}^{(n)\text{-}1}-\boldsymbol{\Sigma}_x^{(n)}\Big) \quad (32)$$

$$= \big|\boldsymbol{A}^{(n)}\big|\mathcal{N}\Big(\boldsymbol{A}^{(n)}\boldsymbol{y}_t + \boldsymbol{b}^{(n)};\boldsymbol{x}_t,\boldsymbol{\Sigma}_{\mathtt{b}}^{(n)}\Big) \quad (33)$$

If instead of using the full GMM to represent the conditional, only the one associated with the most probably corrupted speech component n is used, then the compensation parameters in equation 22 are given by

$$\boldsymbol{A}^{(n)} = \boldsymbol{\Sigma}_x^{(n)}\boldsymbol{\Sigma}_{yx}^{(n)\text{-}1} \quad (34)$$

$$\boldsymbol{b}^{(n)} = \boldsymbol{\mu}_x^{(n)} - \boldsymbol{A}^{(n)}\boldsymbol{\mu}_y^{(n)} \quad (35)$$

$$\boldsymbol{\Sigma}_{\mathtt{b}}^{(n)} = \boldsymbol{A}^{(n)}\boldsymbol{\Sigma}_y^{(n)}\boldsymbol{A}^{(n)\mathsf{T}} - \boldsymbol{\Sigma}_x^{(n)} \quad (36)$$

Though the joint covariance matrix terms in equation 29 may be full, they may be diagonalised to

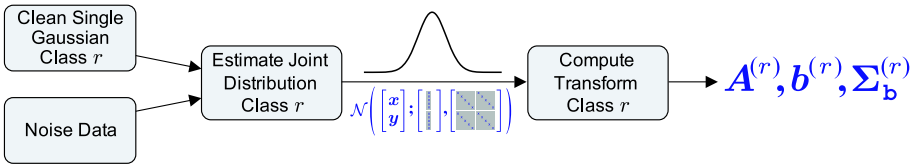

Fig. 6. Model-based joint uncertainty decoding.

make $\boldsymbol{A}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathbf{b}}^{(n)}$ diagonal for efficiency. Also, the normalisation term $\left|\boldsymbol{A}^{(n)}\right|$ in equation 33 is not necessary since it is the same for all likelihood calculations for each time frame. The selection of the appropriate conditional distribution based on the observed corrupted speech, rather than the hidden clean speech, is a significant approximation and is discussed in more detail in Liao and Gales (2004).

It is interesting to compare these parameters with the ones derived using the clean speech posterior in the MBFE observation uncertainty form given by equations 18 and 19. Although both estimate joint distributions for front-end components representing acoustic regions of the corrupted speech space, and have similar decoding likelihood forms, the actual compensation parameters are completely different. For MBFE with uncertainty these are

$$\boldsymbol{A}^{(n)} = \boldsymbol{\Sigma}_{xy}^{(n)} \boldsymbol{\Sigma}_{y}^{(n)\text{-}1} \tag{37}$$

$$\boldsymbol{b}^{(n)} = \boldsymbol{\mu}_{x}^{(n)} - \boldsymbol{A}^{(n)} \boldsymbol{\mu}_{y}^{(n)} \tag{38}$$

$$\boldsymbol{\Sigma}_{\mathbf{b}}^{(n)} = \boldsymbol{\Sigma}_{\hat{\mathbf{x}}}^{(n)} = \boldsymbol{\Sigma}_{x}^{(n)} - \boldsymbol{A}^{(n)} \boldsymbol{\Sigma}_{yx}^{(n)} \tag{39}$$

Note in the original form (Stouten et al., 2004), the clean speech estimate is formed from a weighted contribution from all the components in the front-end GMM, not just the most likely.

3.2. *Model-based uncertainty decoding*

In the previous section describing front-end uncertainty decoding, the conditional distribution in equation 21 is completely decoupled from the acoustic models. However, more precise forms can arise from maintaining this link. In the ALGONQUIN scheme, an *interaction likelihood* $\boldsymbol{\Psi}$ (Kristjansson and Frey, 2002) captures the residual error in the mismatch function $f(\boldsymbol{x}_t, \boldsymbol{n}_t)$. This is propagated to the recognition search as the conditional in the uncertainty decoding framework

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{y}_t; f(\boldsymbol{x}_t, \boldsymbol{n}_t), \boldsymbol{\Psi}) \tag{40}$$

However, these model-based schemes are known to be computationally expensive. For instance, ALGONQUIN uses a variational Bayes algorithm to iteratively approximate the non-Gaussian corrupted speech distribution; this is conducted for every recognition component. Therefore, ALGONQUIN is comparable in form to pure model-based schemes such as PMC or VTS compensation as the effect of the noise is considered independently for each acoustic model component.

Model-based joint uncertainty decoding (M-Joint) (Liao and Gales, 2005) sits in a middle ground between front-end JUD and pure model-based forms. The uncertainty parameters are estimated for a group of similar acoustic model components rather than globally in the front-end or for each model component separately. Compared to front-end uncertainty decoding, instead of having each component in the front-end associated with a region of acoustic space n, link it to a set of similar recognition model components r. For example, one may choose to have two recognition classes, one for silence and another for speech; more classes may be derived by using a regression tree depending on the amount of data available (Shinoda and Watanabe, 1995). The joint distribution can then be computed over this class of recognition components r where the mean vectors and covariance matrices of the clean and corrupted speech are given by

$$\boldsymbol{\mu}_{x}^{(r)} = \frac{\sum_{m \in r} \gamma_t^{(m)} \boldsymbol{x}_t}{\sum_{m \in r} \gamma_t^{(m)}} \tag{41}$$

$$\boldsymbol{\mu}_{y}^{(r)} = \frac{\sum_{m \in r} \gamma_t^{(m)} \boldsymbol{y}_t}{\sum_{m \in r} \gamma_t^{(m)}} \tag{42}$$

$$\boldsymbol{\Sigma}_{x}^{(r)} = \frac{\sum_{m \in r} \gamma_t^{(m)} \boldsymbol{x}_t \boldsymbol{x}_t^{\mathsf{T}}}{\sum_{m \in r} \gamma_t^{(m)}} - \boldsymbol{\mu}_{x}^{(r)} \boldsymbol{\mu}_{x}^{(r)\mathsf{T}} \tag{43}$$

$$\boldsymbol{\Sigma}_{y}^{(r)} = \frac{\sum_{m \in r} \gamma_t^{(m)} \boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{T}}}{\sum_{m \in r} \gamma_t^{(m)}} - \boldsymbol{\mu}_{y}^{(r)} \boldsymbol{\mu}_{y}^{(r)\mathsf{T}} \tag{44}$$

where $\gamma_t^{(m)}$ is the component posterior at time instance t given the observation sequence. The notation $\sum_{m \in r}$ denotes summation over the recognition components m in model class r. The cross-

covariance terms between the clean and corrupted speech are then given by

$$\boldsymbol{\Sigma}_{xy}^{(r)} = \frac{\sum_{m \in r} \gamma_t^{(m)} \boldsymbol{x}_t \boldsymbol{y}_t^{\mathsf{T}}}{\sum_{m \in r} \gamma_t^{(m)}} - \boldsymbol{\mu}_x^{(r)} \boldsymbol{\mu}_y^{(r)\mathsf{T}} \quad (45)$$

Having obtained the joint distribution parameters for a class r, the compensation parameters can be derived using equations 34 to 36. Fig. 6 depicts how these parameters are estimated for a class of recognition components r. Each class r is associated with a single clean speech Gaussian, $\mathcal{N}(\boldsymbol{\mu}_x^{(r)}, \boldsymbol{\Sigma}_x^{(r)})$, computed from the clean speech acoustic model and state occupancy statistics. The rest of the joint parameters may be predicted using model-based compensation approaches such as PMC or VTS compensation (Liao and Gales, 2007), given some noise model. In contrast to the `FE-Joint` scheme, during decoding all the front-end components representing different groups of recognition components are active and pass their measure of uncertainty to the recogniser. This operation is similar to constrained MLLR (Gales, 1998a), but with the addition of a variance bias.

3.3. *Computational cost*

The additional costs for different noise compensation schemes in the front-end processing, and during decoding are summarised in Table 1. `M-Joint` compensation with diagonal variances is surprisingly efficient in comparison to front-end uncertainty decoding, eg `SPLICE` with uncertainty or `FE-Joint`. With the same number of transforms, $R = N$, they have a similar front-end computational cost. The main difference is that the variance bias applied to the recognition model-set is fixed, and therefore maybe cached, given a particular acoustic environment. In contrast, the uncertainty bias in front-end schemes will vary if either the acoustic environment or the front-end component changes. Thus in general, model-based uncertainty approaches are at least as computationally efficient as front-end uncertainty forms.

Comparing this uncertainty decoding cost to, for example, the simplest form of model compensation, the log-add approximation (Gales, 1995), shows that the uncertainty decoding has the potential for large reductions in computational cost. For the log-add approximation, the dominant cost is a D-dimensional matrix vector multiplication for each recognition component, a cost of $\mathcal{O}(D^2)$.

Table 1
Summarising computational cost for different noise compensation schemes

Compensation Scheme	Front-end Cost	Decoding Cost
Feature Enhancement	$\mathcal{O}(DTN)$	None
Front-end Uncertainty	$\mathcal{O}(DTN)$	$\mathcal{O}(DTM)$
Model-based Uncertainty	$\mathcal{O}(DTR)$	$\mathcal{O}(DM)$
Model-based Forms	None	$\mathcal{O}(D^2M)$

D–number of feature dimensions
T–number of frames
N–number of front-end GMM components
R–number of acoustic model classes
M–number of acoustic model components

If the variances are compensated as well, the cost increases dramatically (Gales, 1998b). Using a truncated first-order VTS compensation scheme (Acero et al., 2000) requires the computation of two $D \times D$ matrices per Gaussian in the acoustic model and then several matrix multiplies to compensate. The model-based uncertainty form shares parameters for similar components, so that there is a great saving in their estimation. Compensation is also cheaper since the variance bias addition and re-computation of the normalisation term are both $\mathcal{O}(D)$. Therefore `M-Joint` compensation may be considered a fast form of model-based compensation.

4. Issues with front-end uncertainty decoding

The previous chapter discussed one serious drawback with front-end uncertainty based schemes: that the model variances must be updated every time the variance bias changes. Although, the computation is simple compared to a technique such as model-based VTS compensation, it still involves an expensive recomputation of the typically cached Gaussian normalisation term. However, there is an even larger concern for front-end uncertainty decoding forms.

4.1. *A fundamental problem*

Consider the joint distribution of the clean and noisy speech shown previously in Fig. 5 where the Gaussian noise source is constant. Two corrupted speech conditional distributions, $p(y|x)$, are marked. The first results when the SNR is relatively high, with the clean speech $x = 6$. This yields a highly skewed distribution that peaks sharply at

$x = 6$ that is highly non-Gaussian; still, in JUD it is modelled with a normal distribution without serious degradation (Liao and Gales, 2004). As the SNR increases the skewing becomes more pronounced until the distribution becomes a delta function, yielding the clean speech distribution when substituted in equation 20. This is expected, since when the SNR is high the noise should have no influence on compensating the acoustic models.

The corrupted speech conditional distribution looks very different when the SNR is low, with $x = 1$ while $n = 3$. At this point, the distribution is Gaussian, matching the corrupting noise distribution $\mathcal{N}(3, 1)$. Thus in low SNR, the conditional distribution approaches the noise distribution

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \check{\mathcal{M}}) \approx \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}) \tag{46}$$

where $\boldsymbol{\mu}_{\mathrm{n}}$ and $\boldsymbol{\Sigma}_{\mathrm{n}}$ are the noise mean and variance respectively. This intuitively makes sense, since the noise masks the speech. This singularity has also been documented in Benítez et al. (2004) however the consequences for uncertainty decoding have not been previously examined. It will be shown that front-end uncertainty decoding forms can exhibit problems because of this, while model-based forms do not. If equation 46 is substituted into equation 20, the distribution of the corrupted observation is the same as the noise distribution

$$\begin{aligned} p(\boldsymbol{y}_t|\mathcal{M}, \check{\mathcal{M}}, \theta_t) &\approx \int \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}) p(\boldsymbol{x}_t|\mathcal{M}, \theta_t) d\boldsymbol{x}_t \\ &= \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}) \end{aligned} \tag{47}$$

since the conditional distribution is no longer a function of the clean speech.

Thus regardless of the original recognition model component, the compensated distribution used during decoding will always be identical to the noise distribution. When a single conditional distribution is estimated and used for all components, as is the case with front-end uncertainty decoding schemes, in low SNR no acoustic discrimination is possible since every model distribution has been transformed to the noise. If the recognition task has additional constraints beyond the acoustic models, such as a language model, then it may be possible to distinguish between different models in these regions. However, when there is no language model or other restrictions, for example with a continuous digit recognition task such as AURORA, these areas where no discriminatory acoustic information is available will be prone to errors. These errors will likely be insertions since these are probably background non-speech regions, although low-energy speech may be substituted by other models if the noise is significant enough to mask the speech.

A clear illustration of this issue with `FE-Joint` compensation (Liao and Gales, 2004, 2005) is presented in Fig. 7. This figure shows the clean speech, corrupted speech, `FE-Joint` speech estimate, given by $a^{(n)} y_t + b^{(n)}$, and the uncertainty standard deviation, $\sigma_{\mathtt{b}}^{(n)}$, for a simple system with a 16-component front-end GMM. For those regions of higher energy speech, for example frames 210 to 220 where the vowel 'i' is articulated, the variance bias is small. On the other hand, in the lower energy regions around this vowel, for example frames 225 to 230, the variance becomes too large to be measured on this scale, as is the `FE-Joint` estimate of the value. These large variances are associated with large values of the scale factor $a^{(n)}$ as shown in Fig. 7. In this example, from frames 225 to 230 the value of $a^{(n)}$ is around 100. With greater numbers of front-end components, these effects are amplified as extremes are no longer smoothed over fewer components.

The reason that the magnitude of the scaling factor $a^{(n)}$, and hence the variance biases, both become very large, can be ascertained by examining the nature of the joint distribution, as given in equation 29, in low energy speech regions. For regions with low SNR, the corrupted speech distribution is dominated by the noise; in other words, the noise masks the speech. Consider the cross-variance term $\boldsymbol{\Sigma}_{xy}^{(n)}$ for a front-end component associated with these regions of low speech energy

$$\boldsymbol{\Sigma}_{xy}^{(n)} = \boldsymbol{\Sigma}_{yx}^{(n)\mathsf{T}} = \mathcal{E}\left\{(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(n)})(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(n)})^{\mathsf{T}}\right\} = \mathbf{0} \tag{48}$$

that is, the clean speech and the corrupted speech are uncorrelated since the clean speech and noise processes are independent. This lack of correlation drives $\boldsymbol{A}^{(n)}$, from equation 34, to infinity along with the model variance offsets. In front-end uncertainty decoding, this is expected behaviour because the front-end has determined that in these areas, the uncertainty is high, since the SNR is low. Given equation 48, the relationship to equation 47 becomes clearer by re-expressing equation 22, for component m, as

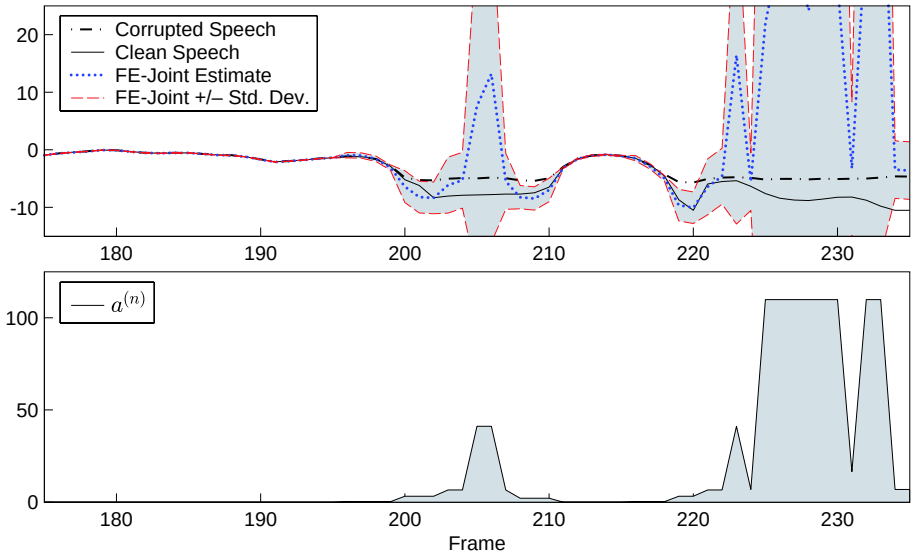

Fig. 7. Plot of log energy dimension for snippet from AURORA digit string 8-6-Zero-1-1-6-2, showing 16-component GMM `FE-Joint` estimate $a^{(n)}y_t + b^{(n)}$, uncertainty bias $\sigma^{(n)}_{\mathbf{b}}$, and $a^{(n)}$.

$$p\left(\boldsymbol{y}_t|\mathcal{M},\check{\mathcal{M}},\theta_t,m\right) \tag{49}$$

$$= \mathcal{N}\Big(\boldsymbol{y}_t;\boldsymbol{\Sigma}^{(n)}_{yx}\boldsymbol{\Sigma}^{(n)\text{-}1}_{x}\Big(\boldsymbol{\mu}^{(m)}-\boldsymbol{\mu}^{(n)}_{x}\Big)+\boldsymbol{\mu}^{(n)}_{y}, \tag{50}$$

$$\boldsymbol{\Sigma}^{(n)}_{yx}\boldsymbol{\Sigma}^{(n)\text{-}1}_{x}\Big(\boldsymbol{\Sigma}^{(m)}-\boldsymbol{\Sigma}^{(n)}_{x}\Big)\boldsymbol{\Sigma}^{(n)\text{-}1}_{x}\boldsymbol{\Sigma}^{(n)\mathsf{T}}_{yx}+\boldsymbol{\Sigma}^{(n)}_{y}\Big)$$

$$= \mathcal{N}\left(\boldsymbol{y}_t;\boldsymbol{\mu}^{(n)}_{y},\boldsymbol{\Sigma}^{(n)}_{y}\right) \tag{51}$$

which is simply the noise distribution in a low energy region. Therefore, allowing an unconstrained estimate of $\boldsymbol{A}^{(n)}$ may result in large numbers of errors, mainly insertions, depending on the task. `FE-Joint`may be viewed as operating in an "uncertainty space" since the features are transformed by $\boldsymbol{A}^{(n)}$; this is why in noisy regions the features are highly scaled rather than the models being transformed to the noise distribution.

Though this is correct in the sense that given the assumptions, this provides the compensation parameters to use, the assumptions are only simple approximations chosen to make `FE-Joint` compensation efficient. Consequently, it may be prudent to mitigate the extreme symptoms that result by restraining the possible values for the compensation parameters. The obvious approach is to examine the correlation coefficients discussed earlier for each of the dimensions, defined as

$$\rho^{(n)}_{xy,i} = \frac{\sigma^{(n)}_{xy,i}}{\sqrt{\sigma^{(n)2}_{x,i}\sigma^{(n)2}_{y,i}}} \tag{52}$$

The compensation parameter estimates given in equations 34 to 36 can then be re-expressed in terms of the correlation coefficient as

$$a^{(n)}_{ii} = \frac{\sigma^{(n)}_{x,i}}{\rho^{(n)}_{xy,i}\sigma^{(n)}_{y,i}} \tag{53}$$

$$b^{(n)}_{i} = \mu^{(n)}_{x,i} - \frac{\sigma^{(n)}_{x,i}}{\rho^{(n)}_{xy,i}\sigma^{(n)}_{y,i}}\mu^{(n)}_{y,i} \tag{54}$$

$$\sigma^{(n)2}_{\mathbf{b},i} = \frac{\sigma^{(n)2}_{x,i}}{\rho^{(n)2}_{xy,i}} - \sigma^{(n)2}_{x,i} \tag{55}$$

for the diagonal form of `FE-Joint`. To restrict the extreme values of $a^{(n)}_{ii}$ and $\sigma^{(n)2}_{\mathbf{b},i}$ that can be obtained, a minimum value on the correlation coefficient can be enforced. Accordingly, the correlation $\rho^{(n)}_{xy,i}$ in equations 53-55 is set to

$$\hat{\rho}^{(n)}_{xy,i} = \max(\rho^{(n)}_{xy,i}, \rho) \tag{56}$$

where ρ is an empirically determined constant. Increasing the value of ρ raises the minimum acceptable correlation, decreasing the maximum variance bias. This can be viewed as enforcing a SNR floor since SNR is highly related to correlation Borga (2001). The effects of this flooring on the same snippet of artificially corrupted speech from Fig. 7 is shown in Fig. 8. As anticipated, the extremes in the variance bias observed before have been tempered.

Since this fundamental issue of all distributions becoming the same in low SNR theoretically affects

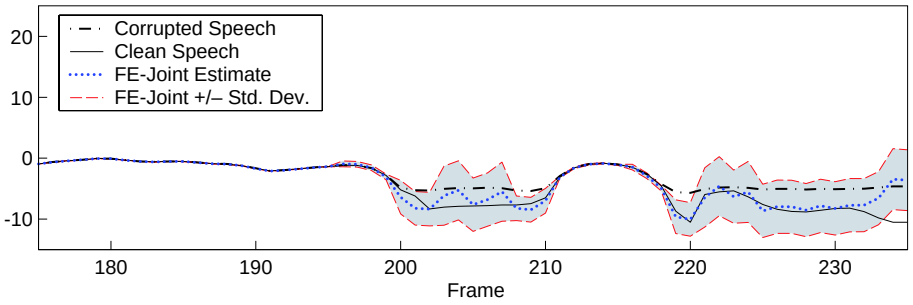

Fig. 8. Plot of log energy dimension for snippet from AURORA digit string 8-6-Zero-1-1-6-2, showing 16-component GMM `FE-Joint` estimate, $a^{(n)}y_t + b^{(n)}$, and uncertainty bias $\sigma_{\mathtt{b}}^{(n)}$, with correlation flooring $\rho = 0.1$.

all front-end uncertainty forms, `SPLICE` with uncertainty should also suffer from it. However this has not been observed, for example, on the AURORA results presented in Droppo et al. (2002). This is because a limit is applied on the maximum value of the variance bias scaling factor $a_{ii}^{(n)}$ to $1/\alpha$ in equation 28. Here α is also an empirically determined parameter. In addition to this explicit flooring, there is also an under-estimate of the value of $a_{ii}^{(n)}$. In order to make the `SPLICE` with uncertainty form tractable, a single clean speech Gaussian for the denominator of equation 25 is used. Since its variance will be larger than any of the individual front-end components that should be used, the scaling estimate will be lower than expected as can be discerned from equation 25. This under-estimation will become larger as the number of front-end components increases because the variance of the individual model components will become smaller and smaller compared to the global variance. This is exactly the situation when a component might be expected to be associated only with a low-energy noise region.

4.2. *Comparison with other uncertainty-based schemes*

The problem with front-end uncertainty decoding is that the conditional distribution $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \check{\mathcal{M}})$ is well suited for acoustic model components that model the region n^* in the acoustic space, but not others; in low SNR this leads to all components being transformed to the equivalent of the noise model. In contrast, `M-Joint` has multiple different $p(\boldsymbol{y}_t|\boldsymbol{x}_t, r)$ dependent on the model class r. Pure model-based schemes estimate this conditional distribution for every model component m. Thus model-based compensation schemes, such as `M-Joint`, do not suffer from this problem since all the models are not globally affected in the same manner. Each model, or group of components, is compensated individually and hence the relative affect of the corrupting noise is taken into account. For example high-energy speech is less influenced by noise than the original background models. This is a result of maintaining the tie between the conditional distribution and the clean speech distribution in equation 20. With recognition components being compensated differently relative to the noise, they may be distinguishable from others, until the interfering noise subsumes all possible speech. Therefore, this theoretical issue with all front-end uncertainty based techniques is not present for model-based forms.

With observation uncertainty, this is also not an issue. While one view may be that the observation uncertainty should be infinite when noise subsumes speech, with a different clean speech prior the uncertainty can be bounded. In MBFE with uncertainty, the clean speech estimate in 18 and variance in equation 19 applied to the acoustic models, where the cross correlation again is zero, becomes

$$\hat{\boldsymbol{x}}_t^{(n)} = \boldsymbol{\mu}_x^{(n)} \tag{57}$$

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}^{(n)} = \boldsymbol{\Sigma}_x^{(n)} - \boldsymbol{\Sigma}_{xy}^{(n)} \left(\boldsymbol{\Sigma}_y^{(n)}\right)^{-1} \boldsymbol{\Sigma}_{yx}^{(n)} \tag{58}$$

$$= \boldsymbol{\Sigma}_x^{(n)} \tag{59}$$

which in this case is the simplified clean speech model for this noise region n, which in low SNR *is* the noise variance of the clean speech background model. Obviously, substituting this clean speech posterior into equation 16 is not problematic. In contrast, the variance of the Weiner filtering process yields the noise variance $\boldsymbol{\Sigma}_{\mathbf{n}}$ of the noisy test condition when the SNR approaches $-\infty$ (Benítez et al., 2004). This inconsistency is due primarily to the different forms of the clean speech posterior,

but may be more generally symptomatic of lacking a formal framework.

4.3. *Summary*

In this section, a major issue for front-end uncertainty decoding forms has been discussed. There can be low SNR regions where models are rendered acoustically indistinguishable, which can result in spurious insertion errors if no other search constraints are available. This was demonstrated with two such front-end uncertainty forms: `SPLICE` with uncertainty and `FE-Joint`. A solution for the `FE-Joint` form was discussed and an analysis of why `SPLICE` with uncertainty does not explicitly display these problems was presented. In comparison to these forms, the observation uncertainty approach does not have this inherent issue, but does not appear to be based on any formal mathematical foundation. Also pure model-based techniques like PMC or `ALGONQUIN` do not suffer from this problem, as they compensate each component differently.

5. Experiments

This section reports quantitative results on the standard small-vocabulary AURORA task and the medium-sized Resource Management corpus—both artificially corrupted databases.

5.1. *The AURORA system*

AURORA 2.0 is a small vocabulary digit string recognition task (Hirsch and Pearce, 2000). Utterances are one to seven digits long based on the TI-DIGITS database with noise artificially added. The clean training data is comprised of 8440 utterances with 55 male and 55 female speakers. For matched training, 422 sentences are provided for each of 16 conditions: 4 different SNRs ranging from 20 to 5 dB, and with the 4 different additive noise sources N1 to N4: subway, babble, car and exhibition hall. Each of the 16 conditions also has a test set of a 1001 sentences with 52 male and 52 female speakers.

The reference recogniser uses a 39 dimensional feature vector consisting of 12 MFCCs appended with unnormalised log energy, delta and delta-delta coefficients. The acoustic models are whole word digit models, each with 16 emitting states, 3 mixtures per state and silence and inter-word pause models to give 546 components. For this work, an internal HTK 3.3 alpha system was used, as opposed to the reference 2.2 version used to report the standard results. This resulted in very minor differences in the baseline performance.

5.2. *The Resource Management system*

The 1000-word naval ARPA Resource Management (RM) database (Price et al., 1988) was corrupted with noise at the waveform level from the NOISEX-92 database (Varga et al., 1992). The clean data was recorded in a sound-isolated room using a head mounted Sennheiser HMD414 noise-canceling microphone yielding a high signal-to-noise ratio of 49 dB[4]. The speech was recorded with 16 bit resolution at 20 kHz and down-sampled subsequently to 16kHz. The speaker independent training data for this task consists of 109 speakers reading 3990 sentences of prompted script; the utterances vary in length from about 3 to 5 seconds totalling 3.8 hours of data.

The NATO NOISEX-92 database provides recording samples of various artificial, pedestrian and military noise environments recorded at 20 kHz with 16 bit resolution. The Destroyer Operations Room noise was sampled at random intervals and added to the clean speech data at the waveform level prior to parameterisation. The noise itself has a dominant low frequency background hum, an unknown repetitive 6 Hz broadband noise of a machine, and other intermittent speech and spontaneous noises.

The baseline recogniser was built using the RM recipe distributed with HTK (Young et al., 2004). The 39 dimensional feature vector consists of 12 MFCCs appended with the log energy, velocity and acceleration coefficients. The cross-word, state-clustered triphone acoustic models with six components per state, giving 9492 recognition components, were used along with a simple word pair grammar. All results are quoted as an average of three of the four available test sets, `Feb'89`, `Oct'89` and `Feb'91`, unless otherwise stated; the `Sep'92` test data was not used. This gave a total of 30 test speakers and 900 utterances. All decoding experiments were run using this system as the standard RM configuration unless otherwise stated.

[4] The `wavmd` tool from the NIST Speech Quality Assurance Package v2.3 was used to determine the SNR.

5.3. *Estimation of compensation parameters*

The compensation parameters for all schemes were estimated using stereo data for the specific noise condition. This allows the techniques to be assessed without having to consider inaccuracies that result from the noise estimation process, or approximations in the mismatch function. In practical situations where stereo data is not available, the compensation parameters can be estimated using PMC or VTS style schemes (Liao and Gales, 2007). For the front-end uncertainty schemes only diagonal transformations were used.

For the front-end uncertainty decoding schemes, `SPLICE` with uncertainty and `FE-Joint`, two sets of front-end GMMs were trained using iterative mixture splitting; these are *clean speech GMMs* and *corrupted speech GMMs* respectively. The first used clean speech data to train clean speech GMMs from which corrupted speech GMMs were derived, for each condition, using stereo data. For the front-end uncertainty decoding schemes described here, this is the preferred way of building models, since provided a noise model is available or can be estimated, the compensation parameters can be simply estimated using VTS or PMC style approaches (Liao and Gales, 2004; Xu et al., 2006). The second set of models were directly estimated from corrupted speech data. This should better represent the corrupted speech acoustic space.

For `M-Joint`, the GMM was not trained explicitly, but rather each Gaussian is linked with a class or cluster of recognition model parameters. Model component classes were derived in a top-down fashion as they are when using constrained MLLR. When using these model-based schemes the parameters in, for example, equation 41 were obtained from the clean data.

5.4. *Results*

Table 2 shows the baseline word error rates along with `SPLICE` system performance on the AURORA task. As usual, the addition of noise seriously degrades the performance of the system unless the clean models are compensated. The matched, approximate "best", target performance is also shown; this matched system was built using stereo data and single-pass re-training (Gales, 1995) to maintain the clean speech transition probabilities, but update the output distributions to reflect the corrupted speech. `SPLICE` was evaluated with both clean and corrupted speech GMMs. The results presented in the table are with a 256-component GMM, but the same general trends were observed for both more and less components. Not surprisingly the use of the corrupted speech trained GMM, as presented in Droppo et al. (2002), outperformed the clean speech trained GMM. It is curious that the `SPLICE` with uncertainty schemes were so sensitive to the choice of GMM. However this may be an attribute of the limitations of the front-end schemes discussed in section 4. To investigate the effects of the flooring on `SPLICE` with uncertainty a range of values of α (see equation 28) were tried. Table 2 shows the performance for 0.1 (as recommended in Droppo et al. (2002)) and the best observed over the range of SNRs 0.95. By increasing the value of α from 0.1 to 0.95 slight performance gains were obtained, especially on the lower SNR conditions. The best configuration was `SPLICE` with uncertainty and $\alpha = 0.95$.

Table 2
Clean, matched and `SPLICE` with 256 components systems' performance on AURORA 2.0 test set A, averaged across N1-N4, WER(%)

	SNR(dB)			
System	20	15	10	5
Clean	4.62	12.20	31.13	59.16
Matched	1.85	2.81	5.01	11.41
Clean Speech GMM				
`SPLICE`	1.97	2.96	6.24	15.74
+Uncertainty, $\alpha = 0.1$	2.49	4.13	8.88	23.06
+Uncertainty, $\alpha = 0.95$	2.30	3.88	8.30	21.38
Corrupted Speech GMM				
`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

Table 3 shows the performance of `FE-Joint` compensation. Two configurations were run. The first used no flooring value for ρ. In contrast to the RM system in Liao and Gales (2004) where significant performance gains were obtained with no ρ flooring, the performance was significantly worse than the baseline `SPLICE` system. While slightly fewer deletions and substitutions occurred overall, a vast number of insertions appeared in regions where there were a series of frames with low-correlation coeffi-

cients. For example, on the car noise at 20 dB, using FE-Joint with 256 components and the clean speech GMM, the number of insertion errors was 421, a magnitude-fold increase from 31 when no noise is present. If ρ is set to 0.9, this drops to a reasonable 19 insertion errors, compared to a total of 13 for the matched system. The correlation floor at this level gives significantly improved performance for both the clean and corrupted speech trained GMM systems. The modified FE-Joint scheme is comparable to SPLICE with uncertainty for both clean and corrupted GMM systems.

Table 3
256-component FE-Joint system performance on AURORA 2.0 test set A, averaged across N1-N4, WER(%)

	SNR(dB)			
System	20	15	10	5
Clean Speech GMM				
FE-Joint	16.99	20.50	25.95	42.78
FE-Joint, $\rho = 0.9$	1.93	2.98	6.09	16.36
Corrupted Speech GMM				
FE-Joint	22.67	25.82	28.38	34.37
FE-Joint, $\rho = 0.9$	1.81	2.88	5.71	14.62

To present a more detailed view of how individual frames and elements in each of those frames are affected by this flooring, results of a 16-component simplified system are presented in Table 4. When ρ is in greater than 0.9 all the low-energy and background related coefficients are affected, severely restraining the magnitudes of the mean and variance biases. Nevertheless, this appears to be an effective strategy.

Table 4
Flooring $\rho_{xy,i}$ on 16-component FE-Joint system on AURORA 20dB SNR, WER(%)

	ρ floor						
Method	0.99	0.95	0.9	0.5	0.1	0.01	-1.0
%Frames affected	100	100	100	58	33	27	0
%Elements affected	99	97	90	46	28	12	0
%WER	2.79	2.52	2.52	3.75	24.5	20.8	20.2

M-Joint compensation was tested on this task with results reported in Table 5. Five systems were built. The first three used diagonal transformations, similar to the front-end schemes. The performance of the 16-transform M-Joint scheme was slightly

Table 5
M-Joint system performance on AURORA 2.0 test set A, averaged across N1-N4, WER(%)

	Number of	SNR(dB)			
System	Transforms	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

worse than the appropriately floored 256 component front-end schemes, but with a considerably reduced computational cost. With an equal number of diagonal transforms, 256, the model-based system is far superior to the front-end version. Moreover, using full transformations gave considerable gains. The 16-transform full variance model-based approach yielded better performance at low SNR than the matched system. However as the variance bias is a full matrix, there is the high cost of performing a full covariance matrix decode, compared to the diagonal covariance matched system. Nevertheless, this does indicate an opportunity to obtain excellent performance using this M-Joint approach.

On the RM task, as shown in Table 6, the incorporation of this correlation flooring does not affect performance, until it is severely set to 0.9, where on this task, at this level, it degrades performance. The presence of a language model to guide the recognition during low SNR regions makes the flooring unnecessary. The corrupted speech GMM in this FE-Joint system was derived from a clean speech GMM and single-pass re-training rather than directly from the corrupted speech data.

Table 6
Flooring $\rho_{xy,ii}$ on FE-Joint system on RM 20dB SNR, WER(%)

	# of	ρ_{const} Floor				
System	Comps.	0.9	0.5	0.1	0.01	-1.0
Clean	—	33.2				
FE-Joint	16	10.8	9.3	9.8	9.7	9.8
	256	10.3	8.2	8.2	8.4	8.4
Matched	—	7.2				

Lastly, results of M-Joint compensation on RM are presented in Table 7. As in the AURORA results, greater transform specificity improves results; however, increasing the number of transforms beyond 16 did not affect performance much. Also similar to AURORA, the most powerful full M-Joint systems give results comparable to matched system performance. This essentially incorporates the correlations between dimensions while using diagonal acoustic model variances.

Table 7
M-Joint system performance on RM 20dB SNR, WER(%)

	# of	Transform Kind	
System	Transforms	Diagonal	Full
Clean	—	33.2	
M-Joint	16	8.2	7.4
	256	8.0	7.4
Matched	—	7.2	

6. Conclusions

This report has presented a fundamental problem with front-end uncertainty decoding methods: by only propagating a single vector of features and probabilities, during periods where the noise is dominant, the ability to effectively discriminate acoustically can be lost. When all the models become identical in these situations, this causes insertion errors in the search. 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 the uncertainty is high. For the FE-Joint compensation scheme, a correlation floor can be used to enforce a bound on the uncertainty decoding scaling, ensuring that all models are not updated to be the same. In SPLICE with uncertainty, the flooring of the variance of the clean speech posterior and the use of a global clean speech prior, both aid in preventing this issue from occurring. M-Joint does not suffer from this problem since the corrupted speech conditional is tied to the clean speech acoustic model. This ensures each recognition component, or group of components, is compensated differently depending much the noise affects them. If the uncertainty parameters can be shared across classes of similar recognition components, such as with M-Joint, efficiency is similar to the front-end versions, without this fundamental problem.

These factors were explored on the small vocabulary AURORA and 1000-word RM corpora. The need to floor the correlations was demonstrated for front-end uncertainty decoding such as SPLICE and FE-Joint forms on the AURORA task; these two algorithms perform comparably. M-Joint compensation gave better results than either on this small task. Similar trends were observed on the medium vocabulary RM database, however the correlation flooring was not necessary due to the presence of a language model to guide the search in low SNR areas. Overall, M-Joint is a superior uncertainty decoding form, since it achieves the best noise robustness and efficiency compared to the other uncertainty forms evaluated.

A major limitation of this paper is that experiments are all conducted on artificially corrupted data and assume stationarity of the noise. This limitation is addressed in Liao and Gales (2007) where a process for estimating noise models for M-Joint compensation is presented along with preliminary results using found data such as Broadcast News.

Acknowledgments

The authors would like to thank the two reviewers for their detailed feedback and constructive comments. This work was sponsored by Toshiba Research Europe Ltd.

References

Acero, A., Deng, L., Kristjansson, T. T., Zhang, J., Oct. 2000. HMM adaptation using vector Taylor series for noisy speech recognition. In: Proc. ICSLP. Beijing, China.

Arrowood, J. A., Clements, M. A., Sep. 2002. Using Observation Uncertainty In HMM Decoding. In: Proc. ICSLP. Denver, Colorado.

Benítez, C., Segura, J. C., A. de la Torre, , Ramírez, J., Rubio, A. J., Oct. 2004. Including uncertainty of speech observations in robust speech recognition. In: Proc. ICSLP. Jeju Island, Korea.

Borga, M., Jan. 2001. Canonical correlation, a tutorial. Available from: http://people.imt.liu.se/~magnus/cca/.

Deng, L., Acero, A., Plumpe, M., Huang, X. D., Oct. 2000. Large vocabulary speech recognition under adverse acoustic environments. In: Proc. ICSLP. Beijing, China, pp. 806–809.

Deng, L., Droppo, J., Acero, A., 2002. Exploiting variances in robust feature extraction based on a parametric model of speech distortion. In: Proc. ICSLP.

Deng, L., Droppo, J., Acero, A., May 2005. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Transactions on Speech and Audio Processing 12 (3).

Droppo, J., Acero, A., Deng, L., May 2002. Uncertainty decoding with splice for noise robust speech recognition. In: Proc. ICASSP. Orlando, Florida.

Droppo, J., Deng, L., Acero, A., Sep. 2001. Evaluation of the SPLICE algorithm on the Aurora 2 database. In: Proc. of Eurospeech 2001. Aalborg, Denmark, pp. 217–220.

Gales, M. J. F., 1995. Model-based techniques for noise robust speech recognition. Ph.D. thesis, Cambridge University.

Gales, M. J. F., Jan. 1998a. Maximum Likelihood Linear Transformations For HMM-Based Speech Recognition. Computer Speech and Language 12.

Gales, M. J. F., 1998b. Predicative model based compensation schemes for robust speech recognition. Speech Communication 25.

Hirsch, H.-G., Pearce, D., Sep. 2000. The AURORA experimental framework for the evaluation of speech recognition systems under noisy conditions. In: Proc. ASR-2000. pp. 181–188.

Holmes, J. N., Holmes, W. J., Garner, P. N., Sep. 1997. Using formant frequencies in speech recognition. In: Proc. Eurospeech. Rhodes, Greece.

Kim, D. Y., Evermann, G., Hain, T., Mrva, D., Tranter, S. E., Wang, L., Woodland, P. C., 2003. Recent advances in broadcast news transcription. In: Proc. ASRU.

Kim, D. Y., Un, C. K., Kim, N. S., Jun. 1998. Speech recognition in noisy environments using first-order vector Taylor series. Speech Communication 24 (1), 39–49.

Kristjansson, T. T., Frey, B. J., May 2002. Accounting for uncertainty in observations: A new paradigm for robust speech recognition. In: Proc. ICASSP. Orlando, Florida.

Liao, H., Gales, M. J. F., 2004. Uncertainty decoding for noise robust speech recognition. Tech. Rep. CUED/F-INFENG/TR499, University of Cambridge, available from: mi.eng.cam.ac.uk/~hl251.

Liao, H., Gales, M. J. F., 2005. Joint uncertainty decoding for noise robust speech recognition. In: Proc. Interspeech.

Liao, H., Gales, M. J. F., 2006. Issues with uncertainty decoding for noise robust speech recognition. In: Proc. Interspeech.

Liao, H., Gales, M. J. F., 2007. Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In: Proc. ICASSP.

Price, P., Fisher, W. M., Bernstein, J., Pallett, D. S., May 1988. The DARPA 1000-word resource management database for continuous speech recognition. In: Proc. ICASSP. Seattle, Washington, USA.

Shinoda, K., Watanabe, T., Sep. 1995. Speaker adaptation with autonomous control using tree structure. In: Proc. Eurospeech. Madrid, Spain.

Stouten, V., van Hamme, H., Wambacq, P., Oct. 2004. Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement. In: Proc. ICSLP. Vol. I. Jeju Island, Korea, pp. 105–108.

Varga, A. P., Steeneken, H. J. M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Tech. rep., Speech Research Unit, Defence Research Agency, Malvern, U.K., available from NOISEX-92 CD-ROMS.

Wölfel, M., Faubel, F., 2007. Considering uncertainty by particle filter enhanced speech features in large vocabulary continuous speech recognition. In: Proc. ICASSP.

Xu, H., Rigazio, L., Kryze, D., 2006. Vector Taylor series based joint uncertainty decoding. In: Proc. Interspeech.

Young, S. J., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. C., Mar. 2004. The HTK Book (for HTK Version 3.3). University of Cambridge.