Model-Based Approaches to Handling Uncertainty

M.J.F. Gales

Abstract A powerful approach for handling uncertainty in observations is to modify the statistical model of the data to appropriately reflect this uncertainty. For the task of noise robust speech recognition, this requires modifying an underlying ”clean” acoustic model to be representative of speech in a particular target acoustic environment. This chapter describes the underlying concepts of model-based noise compensation for robust speech recognition and how it can be applied to standard systems. The chapter will then consider important practical issues. These include: i) acoustic environment noise parameter estimation; ii) efficient acoustic model compensation and likelihood calculation; iii) and adaptive training to handle multi-style training data. The chapter will conclude by discussing the limitations of the current approaches and research options to address them.

1 Introduction

There are many sources of variability in the speech signal, such as inter-speaker variability, intra-speaker variability, background noise conditions, channel distortion and reverberant noise (longer term channel distortions). A range of approaches have been developed to try and reduce the level of variability: some are based on general linear transformations [38, 18]; others based on a model of how the variability impacts the acoustic models or features [37, 17]. This chapter will concentrate on one particular form of variability, background noise and convolutional distortion.

Handling background noise is still a fundamental issue in speech recognition. There are often high levels of mismatch between the training conditions of the acoustic models and test conditions in which they are required to operate. Even with no mismatch, background noise will impact the system performance. As the

M.J.F. Gales
Cambridge University Engineering Department, Trumpington Street, Cambridge e-mail: mjfg@eng.cam.ac.uk

level of noise increases the speech signal will become masked and the ability of the acoustic models to discriminate between words will decrease. Techniques for handling noise should be able to deal with this increase in uncertainty. This chapter examines approaches that handle background-noise and channel distortions by modifying the parameters of the underlying acoustic models, in this case Hidden Markov Models (HMMs) [52, 24]. This class of approaches is often referred to as *model-based noise compensation schemes* [1].

There is some debate as to whether model-based compensation schemes or feature-based compensation, where the "clean" speech is estimated, are the the most appropriate form for noise robust speech recognition. In practice the best scheme depends heavily on the computational resources available, whether the scheme needs to act causally, and the nature of the parametrisation being used. This chapter will briefly mention feature-based schemes, and how uncertainty is included. However as model-based approaches are the more natural approach to handle additional levels of uncertainty associated with noise robust speech recognition, this will be the focus of the discussion.

The next section will briefly discuss general forms of model adaptation to speakers or environment with a particular emphasis on how adaptation can be used to handle uncertainty. The impact of noise on speech and the forms of representation that are often used will then be described. This is followed by a brief discussion of feature-compensation. Model-based compensation is then described, along with a discussion of computational efficiency and estimation of all the model parameters. Finally conclusions are drawn along with possible future directions.

2 General Acoustic Model Adaptation

Given the range of variability (and related uncertainty) associated with speech there has been significant research devoted to handling this problem. Currently, one of the most popular approaches is to use linear transformations of the model parameters. This has been applied for rapid adaptation to speaker or environment changes.

Various configurations of linear transforms have been proposed. Note, the notation used in this section is consistent with the rest of the chapter. The clean speech parameters (the canonical model) will be indicated by an $\mathtt{x}$ in the subscript, and $\mathtt{y}$ for the corrupted speech (target condition) parameters. Thus the corrupted speech mean of component m will be indicated as $\boldsymbol{\mu}_{\mathtt{y}}^{(m)}$.

1. **Maximum Likelihood Linear Regression** (MLLR) [38, 23, 18]: one of the earliest and most popular forms of adaptation. Initially only adaptation of the means was considered [38]. This was extended to adapting the covariance matrices as

[1] There has been a large amount of work, and possible variants, for model-based noise compensation schemes. This chapter is not meant as a complete review of all such schemes. The presentation is (naturally) biased towards work performed at Cambridge University. However it is hoped that all sections are covered with appropriate background references.

well [57, 23, 18]. Here for component m

$$\boldsymbol{\mu}_{\mathtt{y}}^{(m)} = \mathbf{A}^{(r_m)}\boldsymbol{\mu}_{\mathtt{x}}^{(m)} + \mathbf{b}^{(r_m)} \tag{1}$$

$$\boldsymbol{\Sigma}_{\mathtt{y}}^{(m)} = \mathbf{H}^{(r_m)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\mathbf{H}^{(r_m)} \tag{2}$$

where r_m indicates the regression class to which component m belongs.

2. **Constrained MLLR** (CMLLR) [11, 18]: here the transformations of the means and covariance matrices, $\mathbf{A}^{(r_m)}$ and $\mathbf{H}^{(r_m)}$, are constrained to be the same, hence the name CMLLR. Originally used for diagonal transformation of the means and variances of the acoustic models [11], efficient estimation formulae and full transforms were investigated in [18].

$$\boldsymbol{\mu}_{\mathtt{y}}^{(m)} = \mathbf{H}^{(r_m)}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)} - \mathbf{b}^{(r_m)}) \tag{3}$$

$$\boldsymbol{\Sigma}_{\mathtt{y}}^{(m)} = \mathbf{H}^{(r_m)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\mathbf{H}^{(r_m)} \tag{4}$$

Rather than adapting the model parameters, for full transforms it is more efficient to implement this as a set of transformations of the features [18]. Thus the approach is sometimes referred to as Feature MLLR (FMLLR). Now the likelihood can be expressed as

$$p(\boldsymbol{y}_t|m) = |\mathbf{A}^{(r_m)}|\mathcal{N}(\mathbf{A}^{(r_m)}\boldsymbol{y}_t + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}) \tag{5}$$

where $\mathbf{A}^{(r_m)} = \mathbf{H}^{(r_m)\text{-}1}$, $\boldsymbol{y}_t$ is the corrupted speech observations at time t. This form of adaptation does not require the model parameters to be modified. For large vocabulary systems where there may be hundreds of thousands of components this is a very important attribute.

3. **Noisy CMLLR** (NCMLLR) [33]: this is an extension to CMLLR that is specifically aimed at handling situations with additional uncertainty. Here

$$p(\boldsymbol{y}_t|m) = |\mathbf{A}^{(r_m)}|\mathcal{N}(\mathbf{A}^{(r_m)}\boldsymbol{y}_t + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)} + \boldsymbol{\Sigma}_{\mathtt{b}}^{(r_m)}) \tag{6}$$

Thus NCMLLR may be viewed as a combination of CMLLR with a variance bias transform [57]. This form of transformation has the same structure as various noise model-compensation schemes [33].

All of the above approaches involve a transformation of the covariance matrix. Thus in all cases they can model changes in uncertainty in the target conditions by, for example, appropriately scaling the variances.

An interesting extension of these adaptation approaches is adaptive training [4]. Here the transforms are used during the training process. Rather than training a speaker (or noise) independent model-set to be adapted, a "neutral" canonical model is trained that is suitable for adaptation to each of the target conditions. Adaptive training schemes have been derived for all the above transforms [4, 18, 33]. For these adaptive training schemes changing levels of uncertainty in the training data should be reflected in the the contribution of those frames to the canonical model.

Frames with high levels of uncertainty should only make a small contribution to the model updates.

These general adaptation schemes do not rely on explicit models of speaker-differences or the impact of noise on the clean speech. Instead linear transforms, or sets of linear transforms, are estimated given adaptation data. Though advantageous in the sense that these transforms are able to model combinations of differences, they are only linear, or piecewise linear. Furthermore the number of parameters for each transform can be large, $\mathscr{O}(d^2)$ where d is the size of the feature vector, for full transforms. This makes them impractical for very rapid adaptation, though modifications to improve robustness are possible [7, 20].

To enable very rapid adaptation some low-dimensional representation of speaker differences or the impact of noise is needed. For speaker adaptation vocal tract length normalisation [37] is one such scheme. This requires a single parameter, the warping factor, to be estimated. The equivalent for noise robustness is the set of noise models associated with the particular acoustic environment and the *mismatch* function for how the noise alters the clean speech.

3 Impact of Noise on Speech

The first stage in any form of feature or model-based compensation scheme is to specify how the noise alters the clean speech for the parametrisation being used. In this section it is assumed that a "power-domain" MFCC-based feature vector is being used.

3.1 Static Parameter Mismatch Function

The standard, simplified model of the impact background additive noise, $\boldsymbol{n}_t$, and convolutional distortion, $\boldsymbol{h}_t$ on the clean $\boldsymbol{x}_t$, is [1]

$$\begin{aligned}\boldsymbol{y}_t &= \mathbf{C}\log\left(\exp(\mathbf{C}^{-1}(\boldsymbol{x}_t+\boldsymbol{h}_t)+\exp(\mathbf{C}^{-1}\boldsymbol{n}_t)\right)\\ &= \boldsymbol{f}(\boldsymbol{x}_t,\boldsymbol{h}_t,\boldsymbol{n}_t)\end{aligned} \tag{7}$$

where $\boldsymbol{y}_t$ is the corrupted speech observation at time t and $\mathbf{C}$ is the DCT. $\exp()$ and $\log()$ are element-wise exponential and logarithm respectively. It is simple to see that when the energy level of the noise is far greater than that of the (convolutionally distorted) clean speech then $\boldsymbol{y}_t \approx \boldsymbol{n}_t$. The clean speech is *masked* by the noise.

Though (7) is the most commonly used form, a range of alternative mismatch functions, or interaction functions, have also been proposed [17, 36, 10, 41, 21, 39]. These approaches can be split into two categories, domain-based and phase-based compensation.

1. **Domain-based** [17]: this is the simplest form of modified compensation where the domain of the speech and noise compensation is treated as a tunable parameter. Here

$$\boldsymbol{y}_t = \frac{1}{\gamma}\mathbf{Clog}\left(\mathbf{exp}(\mathbf{C}^{-1}\gamma(\boldsymbol{x}_t+\boldsymbol{h}_t)+\mathbf{exp}(\mathbf{C}^{-1}\gamma\boldsymbol{n}_t)\right) \tag{8}$$

   γ determines the domain n which the clean speech and noise are combined. $\gamma=1$ is the power-domain, $\gamma=1/2$ magnitude domain. Its value can be empirically tuned for a particular task.
2. **Phase-based** [10]: domain-based approaches are not motivated from the impact of noise on speech, they simply give a degree of flexibility enabling the mismatch function to be optimised. A more precise formulation is derived by taking into account the phase between the clean speech and noise vectors. Here

$$\boldsymbol{y}_t = \mathbf{Clog}\left(\mathbf{exp}(\mathbf{C}^{-1}(\boldsymbol{x}_t+\boldsymbol{h}_t))+\mathbf{exp}(\mathbf{C}^{-1}\boldsymbol{n}_t)+2\boldsymbol{\alpha}_t\circ\mathbf{exp}(\frac{\mathbf{C}^{-1}}{2}(\boldsymbol{x}_t+\boldsymbol{h}_t+\boldsymbol{n}_t))\right) \tag{9}$$

   where $\boldsymbol{\alpha}_t$ is the vector of phase factors (the cosine of the angle between the speech and the noise) at time instance t and $\circ$ is element-wise multiplication. There have been a range of approximations within this framework. In [41] a fixed value for all elements of the vector $\boldsymbol{\alpha}$ was empirically determined. This is the closest to the domain-based compensation schemes. Indeed the two approaches can be shown to equate for $\gamma=1$ and $\gamma=1/2$. The optimal value for the AURORA 2 task yielded similar mismatch functions for the two approaches [21].
   More precise forms of compensation treat $\boldsymbol{\alpha}$ as a random variable [10, 39, 65]. In [39] an analytic expression for the moments of $\boldsymbol{\alpha}$ were derived. Rather than using all the moments of the distribution of $\boldsymbol{\alpha}$, a simpler approach is to assume that it is Gaussian distributed and use the analysis in [39] to obtain the variance. However, since the phase factor has a physical interpretation, it should lie in the range -1 to +1. Thus an extension to this simple Gaussian approximation was used in [65] to compensate acoustic models using sample-based approaches. Here the variable is treated as

$$p(\alpha_i) \propto \begin{cases} \mathcal{N}(\alpha_i;0,\sigma^2_{\alpha i}) & \alpha\in[-1,+1] \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

   where $\sigma^2_{\alpha i}$ is the phase factor variance for element i.

The rest of this chapter will focus on the standard form of mismatch function given in (7). For some of the alternative mismatch functions model-based compensation has also been examined [41, 21, 39, 65].

3.2 Dynamic Parameter Mismatch Functions

The discussion so far has only considered the static parameters. The feature vector used for decoding usually consists of static and dynamic parameters. The standard form for the dynamic parameters is

$$\boldsymbol{\Delta y}_t = \frac{\sum_{\tau=1}^{w} \tau(\boldsymbol{y}_{t+\tau} - \boldsymbol{y}_{t-\tau})}{2\sum_{\tau=1}^{w} \tau^2} \tag{11}$$

where w is the window-width used to determine the delta parameters. Similar expressions are used for the delta-delta parameters, $\boldsymbol{\Delta}^2\boldsymbol{y}_t$. The form of (11) allows the dynamic parameters to be represented as a linear transform of the static parameters. This is the approach used in [8, 64]. The observation vector for decoding can be expressed as

$$\begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{\Delta y}_t \\ \boldsymbol{\Delta}^2\boldsymbol{y}_t \end{bmatrix} = \mathbf{D} \begin{bmatrix} \boldsymbol{y}_{t+w} \\ \vdots \\ \boldsymbol{y}_{t-w} \end{bmatrix} \tag{12}$$

and $\mathbf{D}$ is the linear transform determined from (11). Provided the appropriate correlations in the feature vector are modelled this allows the mismatch functions in the previous section to be used. Though yielding an accurate form of delta compensation this form is computationally expensive and requires non-standard clean-speech model statistics to be estimated. A similar style of formulation has been used for simple-difference delta and delta-delta parameters [17].

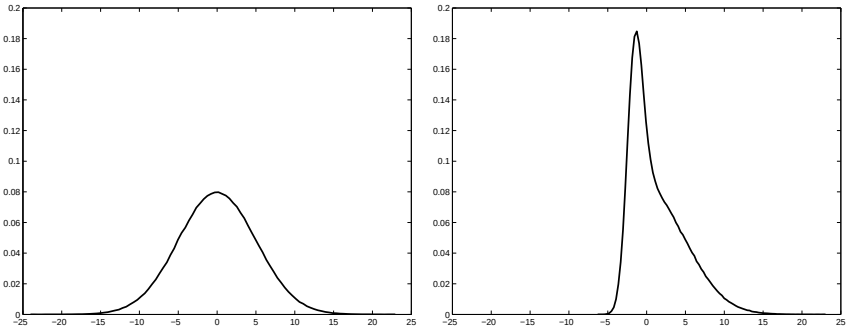
The above scheme is computationally expensive. Thus the most common form of mismatch function used is the continuous time approximation [25]. Here the following approximation is used

$$\boldsymbol{\Delta y}_t \approx \left.\frac{\partial \boldsymbol{y}}{\partial t}\right|_t = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\,\frac{\partial \boldsymbol{x}}{\partial t}\right|_t + \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}\,\frac{\partial \boldsymbol{n}}{\partial t}\right|_t \approx \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\boldsymbol{\Delta x}_t + \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}\boldsymbol{\Delta n}_t \tag{13}$$

This is the standard form used in, for example, VTS compensation [2]. For simplicity of presentation dynamic parameters are not discussed further in this chapter.

3.3 Corrupted Speech Distributions

Having derived a representation for the impact of noise on the clean speech it is useful to examine how it alters the form of the clean speech distribution. Under the mismatch function for the static parameters in (7) and the assumption that both the clean speech and the additive noise are Gaussian distributed the corrupted speech distribution will be non-Gaussian. This is illustrated for one dimension in Figure 1.



(a) Clean Speech (speech mean=0) (b) Corrupted Speech (noise mean=-2)

Fig. 1 Clean Speech (a) Corrupted Speech (b) distributions in the Log-Spectral Domain

As well as causing the distribution to be non-Gaussian, the "masking" property of noise on speech is clear. From Fig. 1 the low-energy speech is completely masked by the noise. This "masking" property has been used for some noise compensation approaches [66] and is also exploited in the missing feature noise robustness schemes [53, 59].

4 Feature Enhancement Approaches

The first forms of noise robustness were based on feature-enhancement approaches. Originally variants on spectral subtraction [6] were popular. These were then replaced by minimum mean square error estimation schemes (MMSE) [49, 9], either requiring stereo data [49, 48, 9, 3], or using noise model estimates [61]. This section will discuss MMSE style enhancement approaches and how uncertainty has been included into these schemes.

For MMSE-based approaches [49, 9] the estimated clean-speech at time t, $\hat{\boldsymbol{x}}_t$, is given by

$$\hat{\boldsymbol{x}}_t = \mathscr{E}\{\boldsymbol{x}|\boldsymbol{y}_t\} \tag{14}$$

The issue is what form the posterior distribution of the clean speech given the corrupted speech should have. For simplicity this is often assumed to be jointly Gaussian. However given the non-linear nature of the interaction of speech and noise in (7) a mixture of Gaussians is used to improve performance. Thus for *front-end* component n the joint distribution is modelled as

$$\left.\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix}\right| n \sim \mathscr{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{\mathrm{y}}^{(n)}\\ \boldsymbol{\mu}_{\mathrm{x}}^{(n)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathrm{y}}^{(n)} & \boldsymbol{\Sigma}_{\mathrm{yx}}^{(n)}\\ \boldsymbol{\Sigma}_{\mathrm{xy}}^{(n)} & \boldsymbol{\Sigma}_{\mathrm{x}}^{(n)}\end{bmatrix}\right) \tag{15}$$

If the component that generated the distribution at time t is known (here $\hat{n}_t$), then the MMSE estimate of the clean speech will be a linear transform of the corrupted speech [29]

$$\hat{\boldsymbol{x}}_t = \mathscr{E}\{\boldsymbol{x}|\boldsymbol{y}_t, \hat{n}_t\} \tag{16}$$

$$= \boldsymbol{\mu}_{\mathtt{x}}^{(\hat{n}_t)} + \boldsymbol{\Sigma}_{\mathtt{xy}}^{(\hat{n}_t)} \boldsymbol{\Sigma}_{\mathtt{y}}^{(\hat{n}_t)\text{-}1} (\boldsymbol{y}_t - \boldsymbol{\mu}_{\mathtt{y}}^{(\hat{n}_t)}) \tag{17}$$

$$= \mathbf{A}^{(\hat{n}_t)} \boldsymbol{y}_t + \mathbf{b}^{(\hat{n}_t)} \tag{18}$$

In practice the component is not known, so needs to either be estimated or treated as a latent variable and marginalised over. For the latent variable case

$$\hat{\boldsymbol{x}}_t = \sum_n P(n|\boldsymbol{y}_t) \mathscr{E}\{\boldsymbol{x}|\boldsymbol{y}_t, n\} \tag{19}$$

There are a number of possible schemes that can be used both in terms of the treatment of the component and the estimation of the compensation parameters $\mathbf{A}^{(n)}$ and $\mathbf{b}^{(n)}$. If the joint distribution (and hence associated marginal distributions) is known then the posterior can be obtained from

$$P(n|\boldsymbol{y}_t) = \frac{P(n)\mathscr{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathtt{y}}^{(n)}, \boldsymbol{\Sigma}_{\mathtt{y}}^{(n)})}{\sum_m P(m)\mathscr{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathtt{y}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{y}}^{(m)})} \tag{20}$$

The estimate of a single component can also be found from the best posterior

$$\hat{n}_t = \operatorname*{argmax}_n \{P(n|\boldsymbol{y}_t)\} \tag{21}$$

An interesting alternative is to use an iterative EM-like process [3]. Either the joint distribution, or the transform, may be estimated from stereo data [49, 3] or using approaches based on model-based compensation [48, 61].

The estimate of the clean speech, $\hat{\boldsymbol{x}}_t$, is then passed to the clean recogniser. Thus the likelihood is approximated for a particular *recognition* component m by

$$p(\boldsymbol{y}_t|m) \approx \mathscr{N}(\hat{\boldsymbol{x}}_t; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}) \tag{22}$$

where $\boldsymbol{\mu}_{\mathtt{x}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}$ are the mean vector and covariance matrix of the clean speech-trained acoustic model for component m. Thus the underlying assumption behind this model is that the clean speech estimate is "perfect", irrespective of the level of background noise. However for low SNR conditions where $\boldsymbol{y}_t \approx \boldsymbol{n}_t$ it is difficult to get an accurate estimate of the clean speech.

One approach to address this problem is to add uncertainty to the estimate of the clean speech [5, 61]. Here the posterior is assumed to be Gaussian in nature

$$\boldsymbol{x}|\boldsymbol{y}_t \sim \mathscr{N}(\hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_t) \tag{23}$$

where $\boldsymbol{\Sigma}_t$ is the "uncertainty" associated with estimate at time instance t. The likelihood is then computed as

$$p(\boldsymbol{y}_t|m) \approx \int p(\boldsymbol{x}|\boldsymbol{y}_t)p(\boldsymbol{x}|m)d\boldsymbol{x} \tag{24}$$

$$\approx \int \mathcal{N}(\boldsymbol{x};\hat{\boldsymbol{x}}_t,\boldsymbol{\Sigma}_t)\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_{\mathrm{x}}^{(m)},\boldsymbol{\Sigma}_{\mathrm{x}}^{(m)})d\boldsymbol{x} \tag{25}$$

Though intuitively well motivated, from (24) it can be seen that the likelihood is not mathematically consistent.

An alternative more consistent scheme is to propagate the distribution of the corrupted speech given the clean speech [12, 43]. Here

$$p(\boldsymbol{y}_t|m) \approx \int p(\boldsymbol{y}_t|\boldsymbol{x})p(\boldsymbol{x}|m)d\boldsymbol{x} \tag{26}$$

Again the acoustic space is represented by a mixture model. Now the distribution (marginalizing over the components) is

$$p(\boldsymbol{y}_t|\boldsymbol{x}) = \sum_n P(n|\boldsymbol{x})p(\boldsymbol{y}_t|\boldsymbol{x},n) \tag{27}$$

Compared to (19) this is more complex as the component posterior, $P(n|\boldsymbol{x})$, is conditional on the clean speech latent variable $\boldsymbol{x}$ rather than the corrupted observation $\boldsymbol{y}_t$. Different approximations for this have been proposed [43, 12].

Though mathematically more consistent, this form of approach has an issue when using the (required) approximations for $P(n|\boldsymbol{x})$. This component posterior term should vary continuously as the "unseen" clean speech $\boldsymbol{x}$ changes. As this is highly computationally expensive to deal with, the approximations used produce a form of "average" component posterior term to use for enhancement. The posterior distribution $p(\boldsymbol{y}_t|\boldsymbol{x})$ is then the same for all "recognition" components m. At very low SNRs the averaged form of component posterior can often result in $p(\boldsymbol{y}_t|\boldsymbol{x}) = p(\boldsymbol{n}_t)$ as the vast majority, but not necessarily all, of clean speech and associated components will be completely masked. As the posterior is now independent of $\boldsymbol{x}$, all recognition components m have the same distribution, so the frame is ignored in terms of acoustic discrimination. The only form of discrimination will be associated with the language model. The information from any non-masked speech (and components) has been lost. Depending on the task this can have a large impact on recognition performance. This issue is discussed in detail in [46]. Given that the underlying attribute of feature-based approaches is that enhancement (with or without uncertainty) is decoupled from recognition components, this problem cannot be addressed within an enhancement framework[2]. As soon as there is a coupling be-

[2] Theoretically the exact value of $P(n|\boldsymbol{x})$ could be used. However as $\boldsymbol{x}$ is a function of the recognition component this effectively becomes model-based compensation. Interestingly if no uncertainty is used such as in SPLICE [9] this problem does not occur as only the means, not the variances, of the distributions can be altered.

tween the "enhancement" and the recognition component, the scheme becomes a model-based approach as discussed in the next section.

5 Model-Based Noise Compensation

The aim of model-based compensation schemes is to modify the acoustic model parameters so that they are representative of the HMM output distributions in the target domain. The advantages of model-based compensation schemes is that the additional uncertainty that results from the background noise is directly modelled. There is no need to estimate masks, or additional uncertainty.

From Fig. 1 it is clear that even if the clean speech and noise are Gaussian distributed, the resulting corrupted speech distribution is non-Gaussian. In practice when considering all elements in the feature vector the corrupted speech distribution may be highly complicated with a large number of modes. Some approaches attempt to model this complexity using for example GMMs [17]. Alternatively Gaussian approximations for the likelihood at the observation value $\boldsymbol{y}_t$ rather than for the whole distribution of $\boldsymbol{y}$ have been proposed [36]. Finally non-parametric schemes for the distribution of $\boldsymbol{y}_t$ have been used [65]. A common attribute of all these schemes is that they are computationally very expensive.

Rather than estimating the "true" distribution, a simple approximation is to assume that the distribution of the corrupted speech is Gaussian in nature [17, 2]. Thus

$$p(\boldsymbol{y}_t|m) \approx \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{y}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{y}}^{(m)}) \tag{28}$$

where $\boldsymbol{\mu}_{\mathrm{y}}^{(m)}$ and $\boldsymbol{\Sigma}_{\mathrm{y}}^{(m)}$ are the estimated mean vector and covariance matrix of the corrupted speech for the target environment. The task is now to obtain appropriate estimates of these corrupted model parameters. Using standard ML-estimation, these can be obtained using [17]

$$\boldsymbol{\mu}_{\mathrm{y}}^{(m)} = \mathcal{E}\{\boldsymbol{y}|m\}, \quad \boldsymbol{\Sigma}_{\mathrm{y}}^{(m)} = \mathcal{E}\left\{\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}|m\right\} - \boldsymbol{\mu}_{\mathrm{y}}^{(m)}\boldsymbol{\mu}_{\mathrm{y}}^{(m)\mathsf{T}} \tag{29}$$

There are a number of approximations for these expectations that sit within this *Parallel Model Combination* (PMC) framework. This chapter will only consider two such forms. The first, Vector Taylor Series (VTS) compensation [48, 2], is one of the most popular approaches. The second based on sampling schemes aims to improve the approximations underlying VTS. Other forms are possible for example the log-normal approximation [17], spline interpolations [58], and Jacobian compensation [56]. However, for all these schemes it is worth emphasising that however accurate the compensation scheme is, the final distribution is approximated by a single Gaussian.

For all these schemes the noise parameters are usually modelled using [17, 48]

$$\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}}), \quad \boldsymbol{h} = \boldsymbol{\mu}_{\mathrm{h}} \tag{30}$$

Thus the convolutional noise is assumed to be constant. Additionally the delta and delta-delta noise means are often assumed to be zero [44]. These parameters may be estimated [40], but the motivation for these estimates is not clear [3]. The estimation of these parameters will be described in Section 7.

5.1 Vector Taylor Series Compensation

A currently popular form of model-based compensation is VTS. Here a first-order Taylor series approximation to the non-linearity of (7) is used. Thus for component m the random variable for the corrupted speech $\boldsymbol{y}$ is related to the clean speech $\boldsymbol{x}$ and noise $\boldsymbol{n}$ random variables by [48]

$$\boldsymbol{y}|m \approx \boldsymbol{f}(\boldsymbol{\mu}_{\mathrm{x}}^{(m)}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}) + \mathbf{J}^{(m)} \left((\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{x}}^{(m)}) + (\boldsymbol{h} - \boldsymbol{\mu}_{\mathrm{h}}) \right) + (\mathbf{I} - \mathbf{J}^{(m)})(\boldsymbol{n} - \boldsymbol{\mu}_{\mathrm{n}}) \quad (31)$$

where the Jacobian $\mathbf{J}^{(m)}$ is defined as

$$\mathbf{J}^{(m)} = \left. \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right|_{\boldsymbol{\mu}_{\mathrm{x}}^{(m)}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}} \quad (32)$$

Using this approximation yields the following estimates for the corrupted speech distribution

$$\boldsymbol{\mu}_{\mathrm{y}}^{(m)} = \boldsymbol{f}(\boldsymbol{\mu}_{\mathrm{x}}^{(m)}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{\mu}_{\mathrm{n}}) \quad (33)$$

$$\boldsymbol{\Sigma}_{\mathrm{y}}^{(m)} = \mathbf{J}^{(m)} \boldsymbol{\Sigma}_{\mathrm{x}}^{(m)} \mathbf{J}^{(m)\mathsf{T}} + (\mathbf{I} - \mathbf{J}^{(m)}) \boldsymbol{\Sigma}_{\mathrm{n}} (\mathbf{I} - \mathbf{J}^{(m)})^{\mathsf{T}} \quad (34)$$

As the Jacobian will be full, this results in a full covariance matrix for the corrupted speech distribution $\boldsymbol{\Sigma}_{\mathrm{y}}^{(m)}$. It is common to diagonalise this covariance matrix to maintain efficient likelihood calculation and control the number of model parameters. Thus in practice the likelihood is computed as

$$p(\boldsymbol{y}_t|m) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{\mu}_{\mathrm{y}}^{(m)}, \mathrm{diag}(\boldsymbol{\Sigma}_{\mathrm{y}}^{(m)}) \right) \quad (35)$$

For a discussion of the impact of this approximation see [64].

A nice aspect of VTS, and one of the reasons for its popularity, is that the linearisation simplifies the estimation of noise and clean speech model parameters [48]. This is discussed in Section 7. However this linearisation may be expected to impact performance, thus alternative schemes are of interest.

[3] These may be interpreted as a general mismatch function rather than motivating from the physical impact of noise on speech.

5.2 Sampling-Based Approximations

VTS relies on a first-order (or possibly higher) Taylor series expansion. To improve this form of approximation it is possible to use sampling-style approaches. This section briefly describes two of these schemes. Both aim to directly estimate the integrals of the form, taking the mean of component m as an example,

$$\boldsymbol{\mu}_{\mathrm{y}}^{(m)} = \int\int \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{n}) p(\boldsymbol{x}|m) p(\boldsymbol{n}) d\boldsymbol{n} d\boldsymbol{x} \tag{36}$$

where $\boldsymbol{f}(.)$ is given in (7).

The simplest approximation is based on Monte-Carlo sampling. As both the clean speech and the noise are Gaussian distributed there are no problems generating samples from them. This approximation, Data-Driven PMC (DPMC) [17], then uses, for example, the following update formula for the mean

$$\boldsymbol{\mu}_{\mathrm{y}}^{(m)} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{f}(\boldsymbol{x}^{(k)}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{n}^{(k)}) \tag{37}$$

where $\boldsymbol{x}^{(k)}$ is a sample drawn from $\boldsymbol{x}|m \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathrm{x}}^{(m)})$ and $\boldsymbol{n}^{(k)}$ is a sample drawn from $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}})$. Note in this case the noise and speech samples are drawn independently.

The advantage of this form of compensation is that in the limit as $K \rightarrow \infty$ the compensation will be "exact" (given the assumptions that the corrupted speech distribution is Gaussian in nature). However a major disadvantage of this straightforward scheme is that as the number of dimensions being sampled from increases, the number of samples needs to be increased in order to get robust estimates.

One approach to address these limitations is to use unscented transforms [31]. Rather than drawing independent samples from the speech and noise, a set of samples are jointly drawn given the means and variances of the clean speech and noise. Here the approximation, again for the mean, has the form

$$\boldsymbol{\mu}_{\mathrm{y}}^{(m)} = \frac{1}{\sum_{k=0}^{K} w^{(k)}} \sum_{k=0}^{K} w^{(k)} \boldsymbol{f}(\boldsymbol{x}^{(k)}, \boldsymbol{\mu}_{\mathrm{h}}, \boldsymbol{n}^{(k)}) \tag{38}$$

The samples are drawn in a deterministic fashion. If the overall dimension of the combined vector $\boldsymbol{z}^{(k)}$ has dimensionality $2d$ (the feature vector is d-dimensional)

$$\boldsymbol{z}^{(k)} = \begin{bmatrix} \boldsymbol{x}^{(k)} \\ \boldsymbol{n}^{(k)} \end{bmatrix} \tag{39}$$

$2d+1$ samples are then drawn in a symmetric fashion based on (noting the dependence of the combined vector on the clean speech component m)

$$\boldsymbol{z}^{(0)} = \boldsymbol{\mu}_{\mathbf{z}}^{(m)}; \;\; w^{(0)} = \frac{\kappa}{2d+\kappa} \tag{40}$$

$$\boldsymbol{z}^{(k)} = \boldsymbol{\mu}_{\mathbf{z}}^{(m)} + \left[\sqrt{(2d+\kappa)\boldsymbol{\Sigma}_{\mathbf{z}}^{(m)}}\right]_k^{\mathsf{T}}; \;\; w^{(k)} = \frac{1}{2(2d+\kappa)} \tag{41}$$

$$\boldsymbol{z}^{(k+2d)} = \boldsymbol{\mu}_{\mathbf{z}}^{(m)} - \left[\sqrt{(2d+\kappa)\boldsymbol{\Sigma}_{\mathbf{z}}^{(m)}}\right]_k^{\mathsf{T}}; \;\; w^{(k+2d)} = \frac{1}{2(2d+\kappa)} \tag{42}$$

where $\left[\sqrt{\mathbf{A}}\right]_k^{\mathsf{T}}$ indicates the transpose of kth row of the Choleski factorisation of $\mathbf{A}$ and κ is a tunable parameter. The number of samples increases linearly as the number of dimensions increases.

Unscented transform compensation has been applied, with gains over VTS and simpler forms of PMC, for both model compensation and feature-based enhancements [60, 28].

6 Efficient Model-Compensation and Likelihood Calculation

One of the issues with model-based compensation schemes is that they are computationally expensive. Applying schemes such as VTS to large vocabulary speech recognition systems is currently impractical for real-time compensation. The costs associated with model-based compensation schemes can be split into three parts: i) estimation of the noise parameters; ii) estimation of the compensation parameters; iii) applying the compensation parameters to the acoustic models. The estimation of the noise parameters is not discussed here, but in the next section. This section will briefly describe approaches for reducing the computational load of the remaining two stages.

One approach to address the problem of computational cost is to express model-based compensation in a factored form [16]. To improve the efficiency this can be rewritten in the following approximate form

$$p(\boldsymbol{y}_t|m) = \int p(\boldsymbol{y}_t|\boldsymbol{x},m)p(\boldsymbol{x}|m)d\boldsymbol{x} \tag{43}$$

$$\approx \int p(\boldsymbol{y}_t|\boldsymbol{x},r_m)p(\boldsymbol{x}|m)d\boldsymbol{x} \tag{44}$$

where r_m indicates the regression class that component m belongs to. The distribution of the clean speech is known, it's given by the clean speech HMM. Thus the problem is to find the conditional distribution, $p(\boldsymbol{y}_t|\boldsymbol{x},r_m)$. It is interesting to compare this form with the enhancement schemes in Section 4. Here the posterior is dependent on either the component or regression class, whereas for feature enhancement it is not. This means that the approximate averaging over the complete acoustic space discussed in [46] and Section 4 will not occur for model-based compensation (unless very few regression classes are used).

6.1 Compensation Parameter Estimation

For schemes such as VTS the compensation parameters required are the Jacobians associated with each component m, $\mathbf{J}^{(m)}$. This is needed to compensate the covariance matrices. This form of Jacobian can be computed as [2]

$$\mathbf{J}^{(m)} = \mathbf{C}\mathbf{F}^{(m)}\mathbf{C}^{-1} \tag{45}$$

where $\mathbf{C}$ is the DCT matrix and $\mathbf{F}^{(m)}$ is a diagonal covariance matrix where the elements on the leading diagonal are given by

$$f_{ii}^{(m)} = \frac{1}{1+\exp([\mathbf{C}^{-1}]_i(\boldsymbol{\mu}_{\mathtt{n}} - \boldsymbol{\mu}_{\mathtt{x}} - \boldsymbol{\mu}_{\mathtt{h}}))} \tag{46}$$

and $[\mathbf{C}^{-1}]_i$ is the ith *row* of $\mathbf{C}^{-1}$. This calculation is dominated by a matrix-matrix multiplication (in the dimensionality of the static parameters) per recognition Gaussian component. For large vocabulary speech recognition this rapidly becomes impractical.

Rather than using VTS the approximation in (44) can be used. The aim is to obtain an efficient form for the regression-class specific conditional distribution, $p(\boldsymbol{y}_t|\boldsymbol{x},r)$. One approach is Joint Uncertainty Decoding (JUD) [42]. Here the joint distribution is assumed to be Gaussian at the regression class level. Thus

$$\left[\begin{array}{c}\boldsymbol{y}\\ \boldsymbol{x}\end{array}\right] \Bigg| r \sim \mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{\mu}_{\mathtt{y}}^{(r)}\\ \boldsymbol{\mu}_{\mathtt{x}}^{(r)}\end{array}\right], \left[\begin{array}{cc}\boldsymbol{\Sigma}_{\mathtt{y}}^{(r)} & \boldsymbol{\Sigma}_{\mathtt{yx}}^{(r)}\\ \boldsymbol{\Sigma}_{\mathtt{xy}}^{(r)} & \boldsymbol{\Sigma}_{\mathtt{x}}^{(r)}\end{array}\right]\right) \tag{47}$$

The conditional distribution is also Gaussian where

$$\boldsymbol{\mu}_{\mathtt{y}|\mathtt{x}}^{(r)} = \boldsymbol{\mu}_{\mathtt{y}}^{(r)} + \boldsymbol{\Sigma}_{\mathtt{yx}}^{(r)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(r)\text{-}1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathtt{x}}^{(r)}) \tag{48}$$

$$\boldsymbol{\Sigma}_{\mathtt{y}|\mathtt{x}}^{(r)} = \boldsymbol{\Sigma}_{\mathtt{y}}^{(r)} - \boldsymbol{\Sigma}_{\mathtt{yx}}^{(r)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(r)\text{-}1}\boldsymbol{\Sigma}_{\mathtt{xy}}^{(r)} \tag{49}$$

As all distributions are Gaussian, the marginal will also be Gaussian. The likelihood in the joint framework can be computed as

$$p(\boldsymbol{y}_t|m) = \mathcal{N}(\boldsymbol{y}_t; \mathbf{H}^{(r_m)}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)} - \mathbf{b}^{(r_m)}), \mathbf{H}^{(r_m)}(\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)} + \boldsymbol{\Sigma}_{\mathtt{b}}^{(r_m)})\mathbf{H}^{(r_m)\mathsf{T}}) \tag{50}$$

where the compensation transform parameters are obtained using

$$\begin{aligned}\mathbf{H}^{(r)} &= \boldsymbol{\Sigma}_{\mathtt{yx}}^{(r)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(r)\text{-}1},\\ \mathbf{b}^{(r)} &= \boldsymbol{\mu}_{\mathtt{x}}^{(r)} - \mathbf{H}^{(r)\text{-}1}\boldsymbol{\mu}_{\mathtt{y}}^{(r)}\\ \boldsymbol{\Sigma}_{\mathtt{b}}^{(r)} &= \mathbf{H}^{(r)\text{-}1}\boldsymbol{\Sigma}_{\mathtt{y}}^{(r)}\mathbf{A}^{(r)\text{-}\mathsf{T}} - \boldsymbol{\Sigma}_{\mathtt{x}}^{(r)}\end{aligned} \tag{51}$$

These compensation parameters only need to be computed for each of the R regression classes, rather than all recognition components. All parameters of the joint

distribution, other than the cross term $\boldsymbol{\Sigma}_{\mathsf{xy}}^{(r)}$, can be obtained using for example VTS, or from the clean speech training data. For VTS the cross term can be found using [62, 43]

$$\boldsymbol{\Sigma}_{\mathsf{xy}}^{(r)} = \boldsymbol{\Sigma}_{\mathsf{x}}^{(r)} \mathbf{J}^{(r)\mathsf{T}} \tag{52}$$

The cost of computing the compensation parameters per regression class is more expensive than computing them for a single component, but the number of regression classes can be made orders of magnitude smaller than the number of components. It is also flexible as the number of regression classes can be controlled depending on the available compute resources.

6.2 Compensating the Model Parameters

Having derived the compensation parameters the model parameters must then be modified. In a similar fashion to (35), whatever form of compensation is used it should require only diagonal covariance matrix likelihood calculations. Directly applying the VTS compensation parameters requires calculating the means and covariance matrices for every component. For large systems this rapidly becomes impractical. Three alternative options for model compensation are described below.

1. **VTS-JUD** [67]: this form is the most closely related to VTS. Equation (50) is used with diagonal covariance matrices. Thus the likelihood is computed as

$$p(\boldsymbol{y}_t|m) = \mathcal{N}\left(\boldsymbol{y}_t; \mathbf{H}^{(r_m)}(\boldsymbol{\mu}_{\mathsf{x}}^{(m)} - \mathbf{b}^{(r_m)}), \mathrm{diag}\left(\mathbf{H}^{(r_m)}(\boldsymbol{\Sigma}_{\mathsf{x}}^{(m)} + \boldsymbol{\Sigma}_{\mathsf{b}}^{(r_m)})\mathbf{H}^{(r_m)\mathsf{T}}\right)\right) \tag{53}$$

   This scheme requires the all recognition parameters to be transformed. Thus the cost of applying the compensation parameters is comparable to standard VTS. However there is the advantage of only computing the compensation parameters at the regression class level.
2. **JUD** [43]: here the likelihood is computed as

$$p(\boldsymbol{y}_t|m) = |\mathbf{A}^{(r_m)}|\mathcal{N}(\mathbf{A}^{(r_m)}\boldsymbol{y}_t + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}_{\mathsf{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathsf{x}}^{(m)} + \boldsymbol{\Sigma}_{\mathsf{b}}^{(r_m)}) \tag{54}$$

   where the compensation transform parameters are obtained using $\mathbf{A}^{(r)} = \mathbf{H}^{(r)\text{-}1}$. This form of compensation only requires a bias to be applied to the clean covariance matrix. However to limit the computational cost this covariance bias term, $\boldsymbol{\Sigma}_{\mathsf{b}}^{(r)}$, needs to be diagonal. Using a full joint distribution and diagonalising the covariance matrix in (54) yields poor performance [43]. To address this problem, the form of the joint distribution can be modified. Here

$$\left[\begin{array}{c}\boldsymbol{y}\\ \boldsymbol{x}\end{array}\right] \Bigg| r \sim \mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{\mu}_{\mathsf{y}}^{(r)}\\ \boldsymbol{\mu}_{\mathsf{x}}^{(r)}\end{array}\right], \left[\begin{array}{cc}\mathrm{diag}(\boldsymbol{\Sigma}_{\mathsf{y}}^{(r)}) & \mathrm{diag}(\boldsymbol{\Sigma}_{\mathsf{yx}}^{(r)})\\ \mathrm{diag}(\boldsymbol{\Sigma}_{\mathsf{xy}}^{(r)}) & \mathrm{diag}(\boldsymbol{\Sigma}_{\mathsf{x}}^{(r)})\end{array}\right]\right) \tag{55}$$

This yields diagonal forms for the compensation parameters in (51). Note it will also be more efficient to compute the compensation parameters. This form only requires compensation parameters at the regression class level, and only a variance bias to be applied at the recognition component level.
This form of compensation has exactly the same form as NCMLLR (6) but derived from a noise compensation perspective. For a discussion of the attributes and comparison of the two approaches see [34].

3. **Predictive CMLLR** (PCMLLR) [22]: this uses the same form of transformation as CMLLR [18]. However rather than estimating the transform parameters from adaptation data, they are estimated from the model-based corrupted speech distributions. The form of likelihood calculation is

$$p(\boldsymbol{y}_t|m) = |\mathbf{A}^{(r_m)}|\mathcal{N}(\mathbf{A}^{(r_m)}\boldsymbol{y}_t + \mathbf{b}^{(r_m)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}) \tag{56}$$

Here the model parameters are not altered, but there is additional costs in estimating $\mathbf{A}^{(r)}$ and $\mathbf{b}^{(r)}$ from the compensation form. For a discussion of the computational costs of this see [67]. Though PCMLLR is an approximation to the corrupted distribution, it has additional flexibility. By using full or block-diagonal transformations, correlation changes in the feature-vector can be efficiently modelled. This is not possible with standard VTS where diagonal covariance matrices are used. This flexibility has been found to yield improved performance [67]. Another advantage of this approach is that adaptive training is very simple, as the standard CMLLR adaptive training approach can be used [67].
Interestingly PCMLLR has exactly the same form as the MMSE estimate in (18). However there are two important differences. First PCMLLR is dependent on the regression class. Second the compensation parameters are derived from minimising the KL-divergence to the estimate of the corrupted speech distribution rather than from a MMSE perspective [63]. As the KL divergence looks at complete distributions (rather than just the first-order moments in MMSE) changes in the uncertainty can be modelled with PCMLLR.

It is simple to show that when the number of regression classes is the same as the number of components, both VTS-JUD, JUD and PCMLLR become identical to the standard model compensation scheme being used to derive the joint distribution.

7 Adaptive Training and Noise Estimation

So far the discussion has assumed that all the model parameters required for compensation are known. In practice this is rarely the case. Originally the background noise was simply estimated from periods of "silence" in the test conditions. This required the use of a voice activity detection scheme, and removed any link between the clean model parameters and the estimates of the noise. Furthermore there is no way to estimate the convolutional noise. For the clean speech parameters it was assumed that clean (high SNR) training data was always available to estimate

the clean models. However this did not allow application domain, or found, data to be used in the training process. Thus recently there has been growing interest in training both the acoustic models [45, 30, 32] and noise model [48, 35, 44, 40] in a full ML framework. This research area has parallels with developments in speaker adaptation where the speaker transform parameters are often estimated in an ML fashion [38] and the canonical model parameters are estimated using adaptive training [4].

The standard approach to estimate the parameters is to maximise the likelihood of the data. Thus the aim is to find the model parameters, $\hat{\mathscr{M}}$, that maximise

$$\mathscr{F}(\hat{\mathscr{M}}) = \sum_{\theta} P(\boldsymbol{\theta}) \prod_{t} \sum_{m \in \theta_t} \mathscr{N}\left(\boldsymbol{y}_t; \hat{\boldsymbol{\mu}}_{\mathrm{y}}^{(m)}, \mathrm{diag}(\hat{\boldsymbol{\Sigma}}_{\mathrm{y}}^{(m)})\right) \tag{57}$$

where the summation over $\boldsymbol{\theta}$ includes all possible state sequences for the observation sequence. In common with standard HMM parameter training, EM is used. Thus the following auxiliary function is maximised (ignoring all terms independent of the model to be estimated)

$$\mathscr{Q}(\hat{\mathscr{M}}; \mathscr{M}) = \sum_{m,t} \gamma_t^{(m)} \log\left(\mathscr{N}\left(\boldsymbol{y}_t; \hat{\boldsymbol{\mu}}_{\mathrm{y}}^{(m)}, \mathrm{diag}(\hat{\boldsymbol{\Sigma}}_{\mathrm{y}}^{(m)})\right)\right) \tag{58}$$

where the posterior of the observation at time t being generated by component m, $\gamma_t^{(m)}$, is determined using the "current" model parameters, $\mathscr{M}$. The task is now to estimate the clean speech model parameters for each of the components, $\hat{\boldsymbol{\mu}}_{\mathrm{x}}^{(m)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{x}}^{(m)}$, and noise model parameters, $\hat{\boldsymbol{\mu}}_{\mathrm{n}}$, $\hat{\boldsymbol{\mu}}_{\mathrm{h}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{n}}$, that maximise (58).

Two approaches have been described in the literature. The first is to introduce a second level of EM, where the clean speech, or noise, at time t are considered as continuous latent variables. This will be referred to as the EM approach. The second is a direct approach based on second-order optimisation schemes. This section gives a summary of some of the attributes of these schemes. Neither of the forms used is exact, a series of approximations is made in each case. The best scheme needs to be determined empirically for the task (and approximations) of interest. For a more detailed analysis and contrast of the two approaches see [15].

7.1 EM-based Approaches

From the VTS approximation (31) it can be seen that the corrupted observation can be written in the form of a generative model where

$$\boldsymbol{y}|m \approx \mathbf{J}^{(m)}\boldsymbol{x}^{(m)} + (\mathbf{I} - \mathbf{J}^{(m)})\boldsymbol{n} + \mathbf{J}^{(m)}\hat{\boldsymbol{\mu}}_{\mathrm{h}} + \mathbf{g}^{(m)} \tag{59}$$

and

$$\boldsymbol{x}^{(m)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}, \hat{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)}) \tag{60}$$

$$\boldsymbol{n} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathtt{n}}, \hat{\boldsymbol{\Sigma}}_{\mathtt{n}}) \tag{61}$$

$$\mathbf{g}^{(m)} = \boldsymbol{f}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\mu}_{\mathtt{h}}, \boldsymbol{\mu}_{\mathtt{n}}) - \mathbf{J}^{(m)}(\boldsymbol{\mu}_{\mathtt{x}}^{(m)} + \boldsymbol{\mu}_{\mathtt{h}}) - (\mathbf{I} - \mathbf{J}^{(m)})\boldsymbol{\mu}_{\mathtt{n}} \tag{62}$$

This now has the form of a general factor analysis style model, for which EM-based update formulae can be applied [55, 26, 35, 30, 33]. This allows the clean speech parameters and the noise parameters to be found in an iterative fashion. Note the convolutional noise bias is not estimated within an EM-style framework (as it has no variance) but is estimated in an EM-style approach and is related to the bias transform estimation [57] (and also to the estimation scheme in [48]).

The estimates of the clean speech means and covariances can then be expressed as[4]

$$\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)} = \frac{\sum_t \gamma_t^{(m)} \mathcal{E}\{\boldsymbol{x}|\boldsymbol{y}_t, m\}}{\sum_t \gamma_t^{(m)}} \tag{63}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)} = \mathrm{diag}\left(\frac{\sum_t \gamma_t^{(m)} \mathcal{E}\left\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\boldsymbol{y}_t, m\right\}}{\sum_t \gamma_t^{(m)}} - \hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)\mathsf{T}}\right) \tag{64}$$

and the noise parameters as

$$\hat{\boldsymbol{\mu}}_{\mathtt{n}} = \frac{\sum_{m,t} \gamma_t^{(m)} \mathcal{E}\{\boldsymbol{n}|\boldsymbol{y}_t, m\}}{\sum_{m,t} \gamma_t^{(m)}} \tag{65}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathtt{n}} = \mathrm{diag}\left(\frac{\sum_{m,t} \gamma_t^{(m)} \mathcal{E}\left\{\boldsymbol{n}\boldsymbol{n}^{\mathsf{T}}|\boldsymbol{y}_t, m\right\}}{\sum_{m,t} \gamma_t^{(m)}} - \hat{\boldsymbol{\mu}}_{\mathtt{n}}\hat{\boldsymbol{\mu}}_{\mathtt{n}}^{\mathsf{T}}\right) \tag{66}$$

where the expectations are over the distribution determined by the current model parameters.

However compared to the standard general FA-style EM-estimation approaches, which are guaranteed not to decrease the likelihood at each iteration, there are two important additional approximations being made.

1. Fixed 'loading matrix' and bias. For this form of FA-style estimation $\mathbf{J}^{(m)}$ and $\mathbf{g}^{(m)}$ are assumed not to be functions of the clean speech and noise parameters to be estimated. This is not the case as the Jacobian and bias will change as the model parameters change.
2. Diagonal covariance matrices. Using the form of generative model in (59) means that the corrupted speech distribution will be a full covariance matrix (the loading matrix $\mathbf{J}^{(m)}$ is full). However this covariance matrix is diagonalised for efficient decoding (35). The joint distribution between the clean speech and corrupted speech (this is the basis for the FA-style estimation) thus has the form

[4] For simplicity of notation the multiple noise conditions that would normally be present for adaptive training have been ignored.

$$\left[\begin{array}{c}\boldsymbol{y}\\ \boldsymbol{x}\end{array}\right]\Bigg| m \sim \mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{\mu}_{\mathtt{y}}^{(m)}\\ \boldsymbol{\mu}_{\mathtt{x}}^{(m)}\end{array}\right], \left[\begin{array}{cc}\mathrm{diag}(\boldsymbol{\Sigma}_{\mathtt{y}}^{(m)}) & \mathbf{J}^{(m)}\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\\ \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\mathbf{J}^{(m)\mathsf{T}} & \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\end{array}\right]\right) \tag{67}$$

However from the generative model in (59) the corrupted speech covariance matrix can be expressed as

$$\hat{\boldsymbol{\Sigma}}_{\mathtt{y}}^{(m)} = \mathbf{J}^{(m)}\hat{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)}\mathbf{J}^{(m)\mathsf{T}} + (\mathbf{I}-\mathbf{J}^{(m)})\hat{\boldsymbol{\Sigma}}_{\mathtt{n}}(\mathbf{I}-\mathbf{J}^{(m)})^{\mathsf{T}} \tag{68}$$

From (67) this should be diagonal. For these two expressions to be consistent, the off-diagonal terms that results from $\mathbf{J}^{(m)}$ being full and $\hat{\boldsymbol{\Sigma}}_{\mathtt{x}}^{(m)}$ being diagonal must be cancelled out by elements from the noise terms. This is not possible for all components as $\mathbf{J}^{(m)}$ is component specific whereas the noise is common for all components. Hence the generative model is not consistent with the joint distribution [5] so the EM-style approach is not guaranteed to increase the auxiliary function. Similar issues arise for the noise estimation case. An alternative, though approximate solution, is to diagonalise the Jacobian. this is the approach adopted in [13]. However this introduces additional approximations in the form of the generative model.

Both of these approximations mean that the update is not guaranteed to increase the auxiliary function. To overcome this problem it is possible to "back-off" estimates by explicitly evaluating the auxiliary function. This becomes important if multiple iterations are performed. Thus though mathematically elegant it is important to be aware of the approximations being used with this approach.

One of the advantages of these FA-style training approaches is that it is simple to incorporate discriminative training criteria such as MPE [50] into the adaptive training framework [13].

7.2 Second-Order Approaches

Rather than using an EM-style approach it is possible to use standard gradient descent style schemes to directly maximise (58) [45, 32]. For a general second-order approach the update has the form

$$\left[\begin{array}{c}\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}\\ \hat{\boldsymbol{\sigma}}_{\mathtt{x}}^{(m)2}\end{array}\right] = \left[\begin{array}{c}\boldsymbol{\mu}_{\mathtt{x}}^{(m)}\\ \boldsymbol{\sigma}_{\mathtt{x}}^{(m)2}\end{array}\right] + \zeta \left[\begin{array}{cc}\frac{\partial^2\mathscr{Q}()}{\partial\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)2}} & \frac{\partial^2\mathscr{Q}()}{\partial\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}\partial\hat{\boldsymbol{\sigma}}_{\mathtt{x}}^{(m)2}}\\ \frac{\partial^2\mathscr{Q}()}{\partial\hat{\boldsymbol{\sigma}}_{\mathtt{x}}^{(m)2}\partial\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}} & \frac{\partial^2\mathscr{Q}()}{\partial\left(\hat{\boldsymbol{\sigma}}_{\mathtt{x}}^{(m)2}\right)^2}\end{array}\right]^{-1} \left[\begin{array}{c}\frac{\partial\mathscr{Q}()}{\partial\hat{\boldsymbol{\mu}}_{\mathtt{x}}^{(m)}}\\ \frac{\partial\mathscr{Q}()}{\partial\hat{\boldsymbol{\sigma}}_{\mathtt{x}}^{(m)2}}\end{array}\right] \tag{69}$$

where $\mathscr{Q}(\hat{\mathscr{M}};\mathscr{M})$ is written as $\mathscr{Q}()$ to save space, $\boldsymbol{\sigma}_{\mathtt{x}}^{(m)2}$ is the vector of leading diagonal elements of $\boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}$, and ζ is the learning rate. Considering the estimation of

[5] It is not clear that the joint covariance matrix in (67) is related to any generative model of the form given in (59) where the noise model is shared over multiple components.

the clean speech mean, $\boldsymbol{\mu}_{\mathrm{x}}^{(m)}$, the derivative can be written as (the fixed variables are explicitly expressed to make the form of the derivative clear)

$$\left.\frac{\partial \mathcal{Q}()}{\partial \hat{\boldsymbol{\mu}}_{\mathrm{x}}^{(m)}}\right|_{\hat{\boldsymbol{\sigma}}_{\mathrm{x}}^{(m)}} = \left.\frac{\partial \hat{\boldsymbol{\mu}}_{\mathrm{y}}^{(m)}}{\partial \hat{\boldsymbol{\mu}}_{\mathrm{x}}^{(m)}}\right|_{\hat{\boldsymbol{\sigma}}_{\mathrm{x}}^{(m)}} \left.\frac{\partial \mathcal{Q}()}{\partial \hat{\boldsymbol{\mu}}_{\mathrm{y}}^{(m)}}\right|_{\hat{\boldsymbol{\sigma}}_{\mathrm{y}}^{(m)}} + \left.\frac{\partial \hat{\boldsymbol{\sigma}}_{\mathrm{y}}^{(m)2}}{\partial \hat{\boldsymbol{\mu}}_{\mathrm{x}}^{(m)}}\right|_{\hat{\boldsymbol{\sigma}}_{\mathrm{x}}^{(m)}} \left.\frac{\partial \mathcal{Q}()}{\partial \hat{\boldsymbol{\sigma}}_{\mathrm{y}}^{(m)2}}\right|_{\hat{\boldsymbol{\mu}}_{\mathrm{y}}^{(m)}} \tag{70}$$

In common with the FA-style approaches these second-order approaches make a number of approximations.

1. Second-order approximation. In common with all second-order schemes there is the assumption that the "error-surface" is quadratic in nature. In practice this is not the case. Additionally the form of the Hessian is often modified, for example diagonalised, and approximated to simplify optimisation.
2. Approximate derivatives. The mean derivative given in (70) is often not used. For example in [32] the second term in (70) is assumed to be zero. Thus the gradient is approximated by

$$\left.\frac{\partial \mathcal{Q}()}{\partial \hat{\boldsymbol{\mu}}_{\mathrm{x}}^{(m)}}\right|_{\hat{\boldsymbol{\sigma}}_{\mathrm{x}}^{(m)}} \approx \mathbf{J}^{(m)} \left.\frac{\partial \mathcal{Q}()}{\partial \hat{\boldsymbol{\mu}}_{\mathrm{y}}^{(m)}}\right|_{\hat{\boldsymbol{\sigma}}_{\mathrm{y}}^{(m)}} \tag{71}$$

   Though simplifying the derivative, it shifts the stationary points of the function.

As there is no guarantee of increasing the likelihood, for noise estimation backing-off approaches can be used [44]. For the model parameter estimation additional smoothing can also be added [44].

8 Conclusions and Future Research Directions

This chapter has reviewed a number of schemes associated with model-based approaches for handling uncertainty. The discussion concentrates on techniques for handling high levels of background noise and channel distortion as this is one of the most important forms of varying uncertainty in the speech signal. A number of approaches, as well as issues, are highlighted. These include: the model compensation process itself; the computational costs associated with this process; and how the parameters of all elements of the process can be estimated from data. Though no performance figures have been given in this chapter, the references given allow a comparison of a number of approaches to be made. In particular the AURORA 2 test set [27] has been used to evaluate a number of systems within a consistent framework.

One of the interesting aspects of model-based compensation research is that techniques originally developed for general linear transform adaptation schemes (whether speaker or environment) are being increasingly used. Thus schemes based

on ML-estimation of the model parameters [38], adaptive training schemes [4] are becoming popular. Additionally discriminative training is also being used [13].

Though there has been improvements in the level of noise robustness for speech recognition, there are still a number of issues that need to be addressed. The author feels these will become increasingly important as the complexity of the task and range of conditions under which ASR systems are required to operated increases.

1. **Impact of noise on speech**: it is not possible to derive representations for the impact of noise on the speech for all forms of parametrisation. This chapter has assumed that MFCC parameters are being used. Even the introduction of basic front-end schemes such as CMN mean that the mismatch function cannot be derived, though approached geared to handling this have been derived [47]. Due to this reason, and the added problem of delta and delta-delta parameters, feature-enhancement schemes based on stereo training [3] are used to combine noise robustness with state-of-the-art front-end processing such as semi-tied transforms [19] and fMPE [51]. Generalising model-based compensation techniques to handle state-of-the-art front-ends will be an important research area.
2. **Handling changes in correlation**: though the Jacobian associated with schemes such as VTS are block-diagonal in structure the resulting covariance matrices are diagonalised for speed of decoding. This is known to degrade recognition performance [43, 64, 67]. Predictive linear transform schemes are one framework for addressing this [22]. However to date research in addressing this problem has been limited. As performance requirements for robust ASR in low SNR conditions increases this topic will become increasingly important.
3. **Improved distribution modelling**: the majority of model-based compensation schemes assume that each speech and noise component pairing will yield a Gaussian distributed corrupted speech component. As previously discussed this is not true. Obtaining more efficient non-Gaussian schemes than the current versions may yield improved performance over the Gaussian approximations.
4. **Speed of compensation/parameter estimation**: one of the main objections to model-based approaches is that they are slow. For large vocabulary systems there may be hundreds of thousands of Gaussian components. Improving the speed of all aspects of model-compensation is essential for it to be broadly applied. For example using incremental forms of noise estimation/compensation [14] is one approach to handling this.
5. **Reverberant noise**: extending the range of environments for which model compensation schemes can be used. For example to handle long-term reverberant noise as well as additive noise, a model-based approach is described in [54].
6. **Improved acoustic modelling**: as the level of background noise increases, and the associated uncertainty of the speech increases, it may become increasingly important to improve the form of the acoustic models being used for the clean speech, noise and corrupted speech. One approach in this direction is to use HMM generative models to obtain scores for use in a discriminative classifier [21].

In summary model-based compensation schemes are a very natural way of handling uncertainty in speech recognition. However there is still significant research required to enable these techniques to achieve the levels of performance, both speed and accuracy, to allow their general deployment in a wide-range of speech applications.

References

1. A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. PhD thesis, Carnegie Mellon University, 1990.
2. A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang. HMM adaptation using vector Taylor series for noisy speech recognition. In *Proc. ICSLP*, pages 869–872, Beijing, China, October 2000.
3. M. Afify, X. Cui, and Y. Gao. Stereo-based stochastic mapping for robust speech recognition. In *Proc. ICASSP*, 2007.
4. T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proc. ICSLP*, 1996.
5. J. A. Arrowood and M. A. Clements. Using Observation Uncertainty In HMM Decoding. In *Proc. ICSLP*, Denver, Colorado, September 2002.
6. S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions Audio Speech and Signal Processing*, 27:113–120, 1979.
7. W. Chou. Maximum a posterior linear regression with elliptically symmetric matrix variate priors. In *Proc. Eurospeech*, 1999.
8. A de la Torre, D Fohr, and J-P Haton. Statistical adaptation of acoustic models to noise conditions for robust speech recognition. In *Proc. ICSLP*, pages 1437–1440, 2002.
9. L. Deng, A. Acero, M. Plumpe, and X. D. Huang. Large vocabulary speech recognition under adverse acoustic environments. In *Proc. ICSLP*, pages 806–809, Beijing, China, October 2000.
10. L. Deng, J. Droppo, and A. Acero. Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. *IEEE Transactions on Speech and Audio Processing*, 12:133–143, 2004.
11. V. V. Digalakis, D. Rtischev, and L. G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions Speech and Audio Processing*, 3:357–366, 1995.
12. J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. ICASSP*, Orlando, Florida, May 2002.
13. F. Flego and M. J. F. Gales. Discriminative adaptive training with VTS and JUD. In *Proc. ASRU*, 2009.
14. F. Flego and M. J. F. Gales. Incremental predictive and adaptive noise compensation. In *Proc. ICASSP*, Taipei, Taiwan, 2009.
15. F. Flego and M. J. F. Gales. Adaptive Training and Noise Estimation for Model-Based Noise Compensation for ASR. Technical Report CUED/F-INFENG/TR653, University of Cambridge, 2010.
16. B. Frey, L. Deng, A. Acero, and T. T. Kristjansson. ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Proc. Eurospeech*, Aalbork, Denmark, September 2001.
17. M. J. F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, Cambridge University, 1995.
18. M. J. F. Gales. Maximum Likelihood Linear Transformations For HMM-Based Speech Recognition. *Computer Speech and Language*, 12, January 1998.

19. M. J. F. Gales. Semi-Tied Covariance Matrices For Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 7:272–281, 1999.
20. M. J. F. Gales. Cluster Adaptive Training of Hidden Markov Models. *IEEE Transactions Speech and Audio Processing*, 8:417–428, 2000.
21. M. J. F. Gales and F. Flego. Discriminative classifiers with adaptive kernels for noise robust speech recognition. *Computer Speech and Language*, 2010.
22. M. J. F. Gales and R. C. van Dalen. Predictive linear transforms for noise robust speech recognition. In *Proc. ASRU*, pages 59–64, 2007.
23. M. J. F. Gales and P. C. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264, 1996.
24. M. J. F. Gales and S. J. Young. The application of hidden Markov models in speech recognition. *Foundation and Trends in Signal Processing*, 1(3):195–304, 2008.
25. R. A. Gopinath, M. J. F. Gales, P. S. Gopalakrishnan, S. Balakrishnan-Aiyer, and M. A. Picheny. Robust speech recognition in noise - performance of the IBM continuous speech recognizer on the ARPA noise spoke task. In *Proc. ARPA Workshop on Spoken Language System Technology*, pages 127–130, Austin, Texas, 1995.
26. R. A. Gopinath, B. Ramabhadran, and S. Dharanipragada. Factor analysis invariant to linear transformations of data. In *Proc. ICSLP*, pages 397–400, 1998.
27. H.-G. Hirsch and D. Pearce. The AURORA experimental framework for the evaluation of speech recognition systems under noisy conditions. In *Proc. ASR*, pages 181–188, September 2000.
28. Y. Hu and Q. Huo. *Chinese Spoken Language Processing*, chapter An HMM Compensation Approach Using Unscented Transformation for Noisy Speech Recognition. Springer Berlin / Heidelberg, 2006.
29. X. D. Huang, A. Acero, and H. W. Hon. *Spoken Language Processing*. Prentice Hall, 2001.
30. Q. Huo and Y. Hu. Irrelevant variability normalization based hmm training using VTS approximation of an explicit model of environmental distortions. In *Proc. Interspeech*, pages 1042–1045, Antwerp, Belgium, 2007.
31. S. J. Julier and J. K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.
32. O. Kalinli, M.L. Seltzer, and A. Acero. Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition. In *Proc. ICASSP*, pages 3825–3828, Taipei, Taiwan, April 2009.
33. D. Kim and M. J. F. Gales. Adaptive training with noisy constrained maximum likelihood linear regression for noise robust speech recognition. In *Proc. Interspeech*, Brighton, UK, 2009.
34. D. Kim and M. J. F. Gales. Noisy constrained maximum likelihood linear regression for noise robust speech recognition. *IEEE Transactions Audio Speech and Language Processing*, 2010.
35. D. Y. Kim, C. K. Un, and N. S. Kim. Speech recognition in noisy environments using first-order vector Taylor series. *Speech Communication*, 24(1):39–49, June 1998.
36. T. T. Kristjansson. *Speech Recognition in Adverse Environments: a Probabilistic Approach*. PhD thesis, Waterloo University, Waterloo, Canada, 2002.
37. L. Lee and R. C. Rose. Speaker Normalisation Using Efficient Frequency Warping Procedures. In *ICASSP'96*, Atlanta, 1996.
38. C. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9, 1995.
39. V. Leutnant and R. Haeb-Umbach. An analytic derivation of a phase-sensitive observation model for noise robust speech recognition. In *Proc. Interspeech*, pages 2395–2398, 2009.
40. J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero. High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. In *Proc. ASRU*, pages 65–70, Kyoto, Japan, December 2007.
41. J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero. HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition. In *Proc. ICASSP*, pages 4069–4072, April 2008.

42. H. Liao. *Uncertainty Decoding for Noise Robust Speech Recognition*. PhD thesis, Cambridge University, Cambridge, UK, sep 2007.
43. H. Liao and M. J. F. Gales. Joint uncertainty decoding for noise robust speech recognition. In *Proc. Interspeech*, 2005.
44. H. Liao and M. J. F. Gales. Joint uncertainty decoding for robust large vocabulary speech recognition. Technical Report CUED/F-INFENG/TR552, University of Cambridge, 2006. Available from: `mi.eng.cam.ac.uk/~mjfg`.
45. H. Liao and M. J. F. Gales. Adaptive Training with Joint Uncertainty Decoding for Robust Recognition of Noisy Data. In *Proc. ICASSP*, volume 4, pages 389–392, Honolulu, USA, April 2007.
46. H. Liao and M. J. F. Gales. Issues with uncertainty decoding for noise robust speech recognition. *Speech Communication*, 2008.
47. Y. Minami and S. Furui. A maximum likelihood procedure for a universal adaptation method based on HMM composition. In *Proc. ICASSP*, pages 129–132, 1995.
48. P. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon University, 1996.
49. L. Neumeyer and M. Weintraub. Probabilistic optimum filtering for robust speech recognition. In *Proc. ICASSP*, volume 1, pages 417–420, 1994.
50. D. Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, 2003.
51. D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig. fMPE: Discriminatively trained features for speech recognition. In *Proc. ICASSP*, Philadelphia, 2005.
52. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
53. B. Raj and R. Stern. Missing Feature Approaches in Speech Recognition. *IEEE Signal Processing Magazine*, 22(5):101–116, 2005.
54. C. K. Raut, T. Nishimoto, and S. Sagayama. Maximum likelihood based HMM state filtering approach to model adaptation for long reverberation. In *Proc. ASRU*, 2005.
55. D. Rubin and D. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, March 1982.
56. S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi. Jacobian approach to fast acoustic model adaptation. In *Proc. ICASSP*, 1997.
57. A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4:190–202, May 1996.
58. M. Seltzer, K. Kalgaonkar, and A. Acero. Acoustic model adaptation via linear spline interpolation for robust speech recognition. In *Proc. ICASSP*, 2010.
59. M. Seltzer, B. Raj, and R. Stern. A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition. *Speech Communication*, 43(4):379393, 2004.
60. Y. Shinohara and M. Akamine. Bayesian feature enhancement using a mixture of unscented transformations for uncertainty decoding of noisy speech. In *Proc. ICASSP*, pages 4569–4572, 2009.
61. V. Stouten, H. van Hamme, and P. Wambacq. Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement. In *Proc. ICSLP*, volume I, pages 105–108, Jeju Island, Korea, October 2004.
62. V. Stouten, H. van Hamme, and P. Wambacq. Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement. In *Proc. ICASSP*, volume I, pages 433–436, Philadelphia, USA, March 2005.
63. R. C. van Dalen, F. Flego, and M. J. F. Gales. Transforming features to compensate speech recogniser models for noise. In *Proc. InterSpeech*, 2009.
64. R. C. van Dalen and M. J. F. Gales. Extended VTS for noise-robust speech recognition. In *Proc. ICASSP*, Taipei, Taiwan, 2009.
65. R. C. van Dalen and M. J. F. Gales. Asymptotically exact noise-corrupted speech likelihoods. In *Proc. InterSpeech*, 2010.

66. A. P. Varga, R. K. Moore, J. Bridle, K. Ponting, and M. Russel. Noise compensation algorithms for use with hidden Markov model based speech recognition. In *Proc. ICASSP*, 1988.
67. H. Xu, M. J. F. Gales, and K. K. Chin. Improving joint uncertainty decoding performance by predictive methods for noise robust speech recognition. In *Proc. ASRU*, 2009.