Chapter 3

Speaker adaptation theory

This chapter discusses previous research in speaker adaptation, but places it within the context of our topic of cross-language adaptation. Reasonably detailed derivations of algorithms are given, especially when understanding of the algorithms are necessary for their proper use for cross-language adaptation versus for speaker adaptation as such.

3.1 Background on speaker adaptation

It has been established that if sufficient data is available, a speaker-dependent system outperforms a speaker independent system [50]. Research [51] has shown that differences between speakers is of much smaller magnitude than differences between phonemes, which is why speaker independent systems function reasonably well in spite of not modelling the exact characteristics of any given speaker. Speaker independent systems, however, perform poorly for speakers with different accents than those with which the system was trained. For most applications there is not enough speaker dependent data for the training of robust models and therefore speaker independent or speaker adaptive training is used. Speaker adaptive training uses large amounts of existing information from many speakers to improve the estimation of model parameters when faced with little data from a new speaker.

3.1.1 Speaker variation

The reason for performing adaptation is that there exists variation between the speech of different speakers. This variation can be classified into two main categories [52]:

- acoustic level differences, including
    - realisational,
    - physiological and
    - durational differences, and
- phonological level differences, including
    - lexical and
    - stress differences.

In this thesis we are mainly interested in the former category of speaker differences, or more accurately, in the correspondence between variation at this level across different languages. To the degree that the acoustic level speaker differences are not language specific, we expect direct cross-language re-use of acoustic information to be useful. In terms of a speech recognition system, the latter category of phonological differences between speakers is dealt with at the language (grammar) and pronunciation modelling level and is thus very language specific. However, since we deal with acoustic modelling, phonological speaker differences are not of direct importance.

Realisational factors comprise different methods of using the articulatory organs to produce wanted sounds. Physiological factors influence the generation of sounds by constraining the possible range of sounds that can be generated by an individual. For example the physical dimensions of the articulatory organs and notably the length of the vocal tract is known to influence the formant frequencies of voiced speech. Durational differences between speakers relate to the timing of the different aspects of generating specific sounds.

3.1.2 Speaker normalisation

Speaker normalisation groups together techniques that attempt to remove, or at least reduce, the differences between the speech of different speakers, while retaining the characteristics that distinguish the different phonetic categories. Vocal tract length normalisation (VTLN) is one such technique that estimates vocal tract length and computes a spectral shift accordingly [53]. An important aspect of normalisation is taking into account not only the characteristics of the particular speaker, but also being able to compensate for recording channel mismatch between training and testing conditions. Subtraction of the estimated mismatch between the training and testing conditions is often done in the cepstral domain, also called cepstral mean subtraction (CMS). CMS performs a cancellation of the effect of any linear operator in the frequency domain, such as is caused by using a microphone with a different frequency transfer function or by frequency filtering due to a transmission channel.

Normalisation is usually applied to the speech signal, or at least to the observation vector sequence as part of the pre-processing stage of the classifier. Normalisation can also be applied during the training phase to compensate for channel variations if applicable, or to reduce spectral differences between training speakers, resulting in more accurate models [54]. When considering the use of multiple databases for cross-language use of data it may be important to apply a normalisation technique such as CMS to take care of recording channel mismatch between the databases. Normalisation will, however, also remove overall spectral differences between the languages, influencing the distribution of feature vectors for all phones. The languages and databases concerned may differ significantly with respect to the phones and the relative quantities of these phones they contain, causing application of CMS to entire databases to be biased. A solution may be to weight the contribution of the data associated with each individual phone in computing the cepstral mean for a database. We discuss this topic in more detail in Section 5.1 where aspects regarding cross-database use of acoustic information are discussed.

Normalisation overlaps to a large degree with speaker adaptation, with normalisation usually seen as the application of adaptation techniques at the feature level, rather than at the model level. Some types of model adaptation, such as transformation-based adaptation, may implicitly perform normalisation, such as done by CMS, with an offset term and can approximate the spectral shift performed with VTLN in the cepstral mean transformation matrix [55], thus further blurring the distinction between adaptation and normalisation. Other adaptation techniques, such as Bayesian or discriminative training-based adaptation can not efficiently remove bias and thus the use of normalisation such as CMS in conjunction with adaptation may still be important to achieve good recognition performance. Zhao [56] performed experiments showing that acoustic normalisation (via CMS) followed by Bayesian adaptation achieved improved performance compared to performing only Bayesian adaptation when training on speech from one database and testing on speech from another database not exactly matched to the first.

3.1.3 Modes of applying speaker adaptation

Speaker adaptation can be applied in an on-line or an off-line mode. For dictation systems speaker adaptation can generally be performed in off-line or static mode, with adaptation occurring after initial enrolment and at intervals after collection of more data. For telephone-based systems, adaptation, if any, has to be applied on-line or dynamically on a per-call basis. The main difference between static and dynamic adaptation is in terms of the need for real-time implementation. Real-time constraints force dynamic adaptation to be performed on very little data, typically a single utterance, while static adaptation such as used for dictation systems, may use perhaps 30 minutes of speaker specific data. On-line methods use incremental techniques that typically only slightly change model parameters with each additional utterance used, while off-line methods perform batch-mode parameter updates that may completely re-estimate parameters. Cross-language adaptation is performed off-line since real-time constraints are not applicable. On-line adaptation may of course still be used after this to further increase performance when the system is applied to specific speakers.

Another important aspect to take into account is whether adaptation will be supervised or unsupervised. In supervised adaptation the adaptation speech has been labelled, or at least a transcription of the adaptation speech is available. In unsupervised adaptation, the speech to be used for adaptation is unknown and has to be recognised first before it can be used for adaptation. Chapter 2 discussed cross-language use of bootstrapping methods where transcriptions of the data in the target language were available, but the data was not labelled at phone level. Completely unsupervised cross-language adaptation is probably not feasible since the mismatch between the models and data would probably be too great for recognition in the target language to give acceptable results for further training or adaptation.

3.1.4 Categories of speaker adaptation

Speaker adaptation techniques have previously been classified into three categories [54] namely: (i) speaker classification, (ii) spectral transformation and (iii) speaker adaptive re-estimation of model parameters. We use a similar structure for our discussion of speaker adaptation techniques, but consider the transformation category to encompass newer techniques using transformations of model parameters and not only spectral or feature space transformations. Furthermore the third category of speaker adaptive re-estimation is quite wide and we limit ourselves in this chapter to the discussion of Bayesian adaptation techniques. A further field only recently applied to speaker adaptation, namely discriminative learning, is discussed in the next chapter. An overview of the three categories of

- speaker classification,
- transformation-based adaptation and
- Bayesian adaptation

is given next.

Speaker classification attempts to identify a specific set of models that best exhibit the characteristics of a new speaker and uses those models to perform recognition. The speaker classification category is of little interest to our research as it cannot change the characteristics of the acoustic space except to cluster it into segments. It is unlikely that significant overlap will occur between the clusters of speakers in different languages and even if there were significant overlap, the method would still only be useful in terms of handling speaker specific characteristics and not performing any adaptation to the new target language. The other two categories are more interesting to our research as they both can change source language model parameters in a structured way to better reflect the characteristics of the target language.

Transformation-based adaptation entails computing a transformation of pre-trained model parameters to better fit the speech of a new target speaker. This type of adaptation has ties with the technique of normalisation, discussed in Section 3.1.2, which operates on either frequency or cepstral domain features of the observation sequence to reduce spectral differences between speakers. Transformation-based adaptation, at its simplest, may entail only the subtraction of a global cepstral offset term, thereby improving the spectral match between the pre-trained model and the target speakers' speech in the same way that would be achieved with normalisation. However, when we refer to transformation-based adaptation, we usually imply that a matrix transformation of the model parameters is estimated - a more complex and powerful approach than frequency equalisation since (i) correlated noise can be removed and (ii) different transformations can be estimated for different phone groupings.

Transformation-based techniques assume a large degree of correlation between the feature distribution expressed by the current model and the feature distribution of the target speaker. This paradigm is well suited for the removal of correlated noise between source and target parameters. In contrast, Bayesian learning does not assume correlation with respect to changes from a current model, but assumes that prior knowledge exists about the distribution of the model parameters. Observations from a new speaker are treated as adding to the prior knowledge of the parameter distributions, thereby improving the estimate of the parameters. We expect Bayesian methods to work well in an environment where we have reasonably robust models in general, but which may need complex fine-tuning to achieve improved performance for a specific speaker or environment. The next two sections discuss in detail the implementation of Bayesian and transformation methods used in this thesis.

3.2 Bayesian adaptation

Bayesian estimation presents an alternative to maximum likelihood estimation and is preferred in particular when information about the distribution of unknown parameters is available. Bayesian learning also provides a framework for parameter smoothing and speaker adaptation when faced with a limited amount of data. In this section we are specifically interested in the use of Bayesian methods for adaptation. Bayesian estimation is well suited to the speaker adaptation paradigm, because information about (prior) distributions of model parameters can be estimated beforehand from a large set of speakers and then fine-tuned using measured observations from a new speaker.

Bayesian methods consider model parameters to be random variables with known *a priori* distributions. Observation of sample data from a new speaker converts the *a priori* density of a parameter into an *a posteriori* density, improving the estimate of the true value of the parameter and converging to the true value as the amount of observations increases. In Bayesian estimation, the unknown, but desired p.d.f. $p(\mathbf{x})$ is estimated by using the observed data $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and integrating over the parameter vector $\boldsymbol{\theta}$, which is considered a random variable taking values in the space Θ. The integral is expressed by [25, p. 51]

$$\begin{aligned} p(\mathbf{x}|\mathbf{X}) &= \int_{\Theta} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta}. \end{aligned} \tag{3.1}$$

Using Bayes rule, Equation 3.1 can be written, using the notation $g(\boldsymbol{\theta})$ for the prior distribution and $f(\mathrm{X}|\boldsymbol{\theta})$ for the likelihood function, by

$$p(\mathrm{x}|\mathrm{X}) = \int_{\Theta} p(\mathrm{x}|\boldsymbol{\theta}) \frac{f(\mathrm{X}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{p(\mathrm{X})} d\boldsymbol{\theta}, \tag{3.2}$$

where the observation probability, $p(\mathrm{X}) = \int_{\Theta} f(\mathrm{X}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$, is a constant that normalises the posterior density function. In practice Equation 3.2 does not offer a computationally feasible solution with current speech modelling techniques and computer technology due to the integration term. However, if $p(\boldsymbol{\theta}|\mathrm{X})$ peaks very sharply about some value $\hat{\boldsymbol{\theta}}$, Equation 3.2 may be approximated by

$$p(\mathrm{x}|\mathrm{X}) \approx p(\mathrm{x}|\hat{\boldsymbol{\theta}}). \tag{3.3}$$

This is especially applicable according to the Bayesian learning paradigm described by Duda & Hart [25, p. 54], which states in general that as the number of observations from a given distribution increases, the posterior distributions of the parameters peak more sharply around the true values of the parameters, ultimately approaching Dirac delta functions at the true values of the parameters as the number of observations approaches infinity. In this case the approximation is therefore entirely applicable.

However, even if the posterior parameter distribution is not sufficiently peaked, to reach a computationally feasible solution, it may still be necessary to estimate a single parameter value $\hat{\boldsymbol{\theta}}$ for use in place of the integration over the parameter space of Equation 3.2. The next section discusses a procedure to estimate such a parameter.

3.2.1 Bayes estimators

Because Bayesian methods consider parameters to be random variables, distributions of parameters are used, rather than fixed values. For efficiency, a single suitable value for the parameter may need to be estimated and for this purpose an estimator is used. The form of the estimator is not prescribed in Bayesian learning and remains to be decided by the statistician. The most important requirement of an estimator δ is that it delivers an estimate $\delta(\mathbf{X})$ (based on the observed data $\mathbf{X}$) that is close to the actual value $\mathbf{a}$ of the parameter $\boldsymbol{\theta}$ in an experiment. A sensible way of determining an estimator is by specifying a *loss function* $L(\mathbf{a}, \hat{\boldsymbol{\theta}})$ which measures the loss or cost when the true value of the parameter is $\boldsymbol{\theta} = \mathbf{a}$ and the estimate is $\hat{\boldsymbol{\theta}}$. The *Bayes estimator* [57, p. 275] is then given by the function $\delta^*(\mathbf{X})$ that, for every possible value $\mathbf{x}$ of $\mathbf{X}$, delivers the minimum expected loss, i.e.

$$\mathrm{E}[L(\boldsymbol{\theta}, \delta^*(\mathbf{X}))|\mathbf{X}] = \min_{\hat{\boldsymbol{\theta}} \in \Theta} \mathrm{E}[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})|\mathbf{X}], \tag{3.4}$$

where the unknown value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ takes values in the space Θ.

Minimum square error Bayes estimation

The loss function that is most commonly used is the *squared error loss function*, $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$. When the squared error loss function is used, the Bayes estimate is the value of $\hat{\boldsymbol{\theta}}$ for which $\mathrm{E}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})|\mathbf{X}]$ reaches a minimum value. The Bayes estimator for the squared error loss function is found by finding the root of the quadratic, i.e.

$$\begin{aligned}
\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \mathrm{E}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})|\mathbf{X}] &= 0 \\
\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \left[\mathrm{E}[\boldsymbol{\theta}^T \boldsymbol{\theta}|\mathbf{X}] - 2\hat{\boldsymbol{\theta}}^T \mathrm{E}[\boldsymbol{\theta}|\mathbf{X}] + \hat{\boldsymbol{\theta}}^T \hat{\boldsymbol{\theta}} \right] &= 0 \\
-2\mathrm{E}[\boldsymbol{\theta}|\mathbf{X}] + 2\hat{\boldsymbol{\theta}} &= 0
\end{aligned}$$

and thus the Bayes estimator is simply equal to the expectation value of the parameter $\boldsymbol{\theta}$,

$$\begin{aligned}
\delta^*(\mathbf{X}) = \hat{\boldsymbol{\theta}} &= \mathrm{E}[\boldsymbol{\theta}|\mathbf{X}] \\
&= \int_\Theta \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} \\
&= \int_\Theta \boldsymbol{\theta} \cdot \frac{f(\mathbf{X}|\boldsymbol{\theta}) g(\boldsymbol{\theta})}{p(\mathbf{X})} d\boldsymbol{\theta},
\end{aligned} \tag{3.5}$$

which in turn equals the first moment of the posterior density function $f(\mathbf{X}|\boldsymbol{\theta})g(\boldsymbol{\theta})/p(\mathbf{X})$. We refer to the Bayes estimator of Equation 3.5 as the MSE estimator in subsequent discussions since it produces the minimum squared error (MSE) solution to the Bayes loss function.

Other loss functions exist and may lead to different Bayes estimators, such as the absolute error loss function which leads to the Bayes estimate being equal to the median of the posterior distribution [57, p. 277]. An alternative to using a loss function in the Bayesian framework is to simply use the maximum value of the posterior distribution as the estimate, which in general will differ from the mean for asymmetric functions. This method is discussed next.

MAP Bayes estimation

Maximum *a posteriori* (MAP) estimation uses the parameter associated with the maximum *a posteriori* probability as the Bayes estimate. The MAP estimate for a parameter $\boldsymbol{\theta}$, given a prior distribution $g(\boldsymbol{\theta})$ and observation sequence $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_n\}$ is given by the mode of the posterior density function, i.e.

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} f(\mathbf{X}|\boldsymbol{\theta})g(\boldsymbol{\theta}). \tag{3.6}$$

If $g(\boldsymbol{\theta})$ is considered fixed, but unknown, also known as a non-informative prior, then there is no knowledge about $\boldsymbol{\theta}$ and the MAP estimate is equal to the maximum likelihood (ML) estimate. We thus consider the selection of a suitable informative prior. The choice of a prior distribution is predicated as much by its suitability for expressing the prior distribution as by the possibility of deriving a solution for the Bayesian/MAP estimation problem. Similar to ML estimation, the computation of the MAP estimate is relatively easy when the family of p.d.f.'s $\{f(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ possesses a *sufficient statistic* of fixed dimension. For HMMs in the *incomplete data* modelling problem this is not true, but is addressed by iterative methods that solve the *complete data* modelling problem for which a sufficient statistic exists. Given that the family $\{f(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ possesses a sufficient statistic $t(\mathbf{X})$ of fixed dimension for the parameter $\boldsymbol{\theta}$, $f(\mathbf{X}|\boldsymbol{\theta})$ can be factored into two terms $f(\mathbf{X}|\boldsymbol{\theta}) = h(\mathbf{X})k(\boldsymbol{\theta}, t(\mathbf{X}))$ such that $h(\mathbf{X})$ is independent of $\boldsymbol{\theta}$ and $k(\boldsymbol{\theta}, t(\mathbf{X}))$ is the *kernel density*, which is a function of $\boldsymbol{\theta}$ and depends on $\mathbf{X}$ only through the sufficient statistic $t(\mathbf{X})$. If the prior density is thus chosen in a *conjugate family* $\{k(\cdot|\boldsymbol{\psi}), \boldsymbol{\psi} \in \Psi\}$ which includes the kernel density of the likelihood function $f(\cdot|\boldsymbol{\theta})$, the MAP estimate is greatly simplified since the posterior density is then of the same form as the prior, i.e. $k(\boldsymbol{\theta}|\boldsymbol{\psi}') \propto k(\boldsymbol{\theta}|\boldsymbol{\psi})k(\boldsymbol{\theta}, t(\mathbf{X}))$. With such a choice of prior, the procedure for finding the MAP estimate is similar to solving for the ML estimate - i.e. both find the mode of the kernel density.

Having a simple posterior density also eases implementation of other Bayesian estimators such as the MSE estimator which finds the mean of the posterior distribution. For symmetric distributions, such as the normal distribution, the mode and mean are equal and thus the MSE and MAP estimates are the same, while for asymmetric distributions, such as the Gamma distribution, the estimates will generally differ. We note at this point that it is only because a limited amount of adaptation data is used that the difference between the MAP and MSE estimators is considered. We do not expect the difference between the estimates produced by the methods to be large, but still wish to quantify the difference.

With some basic theory behind Bayesian estimation now covered, we proceed to discuss the implementation of Bayesian adaptation, and more specifically MSE and MAP adaptation for both the (single) Gaussian observation density case as well as for the more general Gaussian mixture distribution case. We assume that we are solving the *complete data* modelling problem as we shall discuss the implementation of the iterative estimation algorithm [24, 58] for the *incomplete data* modelling problem for HMMs in Section 3.2.4.

3.2.2 Gaussian density parameter distributions

In this section it is assumed that a sample from a Gaussian distribution is available and it is desired to derive the posterior distributions of the parameters of the Gaussian, i.e. the mean and variance of the Gaussian. The derivations closely follow DeGroot [59].

Mean-only adaptation

The simplest and also most used approach for Bayesian adaptation is to assume a normal distribution with mean m and precision τ (inverse of the variance) as the prior for the mean parameter μ (to be estimated) of the Gaussian observation distribution and a fixed, known value for the Gaussian precision parameter r. The prior distribution of μ

$$g(\mu) \propto \tau^{1/2} e^{-(\tau/2)(\mu-m)^2} \tag{3.7}$$

and the likelihood function $f(X|\mu)$ for observations $X = \{x_1, ..., x_n\}$

$$\begin{aligned} f(X|\mu) &\propto r^{n/2} e^{-(r/2)\sum_{i=1}^{n}(\mu-x_i)^2} \\ &\propto r^{n/2} e^{-(r/2)\,[nS+n(\mu-\bar{x})^2]} \end{aligned} \tag{3.8}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean and

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.9}$$

is the sample variance of the observations, can be combined to form the posterior p.d.f. $g(\mu|X)$ given by

$$g(\mu|X) \propto f(X|\mu)g(\mu) \propto \tau^{1/2} r^{n/2} e^{-(1/2)[\tau(\mu-m)^2+nr(\mu-\bar{x})^2+nrS]}. \tag{3.10}$$

By using the equality

$$\tau(\mu-m)^2 + nr(\mu-\bar{x})^2 = (\tau+nr)\Big(\mu - \frac{\tau m + nr\bar{x}}{\tau+nr}\Big)^2 + \frac{\tau nr}{\tau+nr} r(m-\bar{x})^2, \tag{3.11}$$

it is noted that the posterior p.d.f. $g(\mu|X)$ of μ is also a normal distribution (similar to the prior of μ in Equation 3.7), with mean $\frac{\tau m+nr\bar{x}}{\tau+nr}$ and precision $\tau + nr$ [59, p. 167] and is given by

$$g(\mu|X) \propto e^{-\frac{\tau+nr}{2}\left(\mu-\frac{\tau m+nr\bar{x}}{\tau+nr}\right)^2}. \tag{3.12}$$

Since the mode and the mean value of the normal distribution are equal, both the MAP estimate as well as the minimum squared error Bayesian estimate for μ are given by the mean of Equation 3.12, namely

$$\mu_{\mathrm{MAP}} = \mu_{\mathrm{MSE}} = \frac{\tau m + nr\bar{x}}{\tau + nr}. \tag{3.13}$$

Note that we refer to the mean value of a distribution as the expectation value of the parameter on which the distribution is conditioned. The estimate of μ is a linear combination of the prior mean m and the speaker dependent sample mean $\bar{x}$. When $n = 0$, no observations are available and the MAP estimate is simply equal to the prior mean value m. When a large number of observations are available, the MAP estimate converges to the ML estimate $\bar{x}$ asymptotically. When the prior precision τ is large, high confidence is associated with the prior mean m and a larger sample size will be necessary to significantly change the MAP estimate from the prior mean value than when the prior precision value is small. The difficulty with implementing Bayesian adaptation is in choosing suitable values for the prior distribution parameters. In the MAP approach suggested by Lee *et al.* [24], prior distribution parameters are estimated from speaker independent training data. Speaker independent Gaussian mixture distribution models are used to estimate weighted mean and precision values for the prior as well as the expected value of the precision through

$$m = \sum_{k=1}^{K} \tilde{c}_k \tilde{m}_k, \tag{3.14}$$

$$1/\tau = \sum_{k=1}^{K} \tilde{c}_k (\tilde{m}_k - m)^2 \tag{3.15}$$

and

$$1/r = \sum_{k=1}^{K} \tilde{c}_k \tilde{\sigma}_k^2 \tag{3.16}$$

where $\tilde{c}_k$ is the weight, $\tilde{m}_k$ is the mean and $\tilde{\sigma}_k^2$ is the variance of the kth mixture component of the speaker independent model. The weighted value of m is simply the sample mean of the speaker independent data when the weights are ML estimates while $1/\tau$ equals the variance of the mixture means around the global mean value and $1/r$ is the weighted average variance within a mixture. This choice of estimating the prior distribution and fixed variance makes especially good sense when we expect each mixture distribution to be representative of an individual speaker or type of speaker since Equation 3.15 then represents the expected between-speaker variance and Equation 3.16 the expected within-speaker variance.

Variance-only adaptation

Variance adaptation is proposed by Lee *et al.* [24] by assuming the value of the mean m to be fixed, but unknown and the variance a random variable with a prior distribution $g(\sigma^2)$ of the form

$$g(\sigma^2) = \begin{cases} \text{constant} & \text{if} \quad \sigma^2 \leq \sigma^2_{\text{min}} \\ 0 & \quad \text{otherwise,} \end{cases} \tag{3.17}$$

where σ^2_{min} is estimated from a large amount of speech data and should be a reasonable lower bound on the variance. We have arbitrarily chosen $\sigma^2_{\text{min}} = 10^{-4}$. The MAP estimate for the variance is then given by

$$\sigma^2_{\text{MAP}} = \begin{cases} S & \text{if} \quad S \geq \sigma^2_{\text{min}} \\ \sigma^2_{\text{min}} & \text{otherwise,} \end{cases} \tag{3.18}$$

where S is the sample variance as in Equation 3.9. While Equation 3.18 is not really useful by itself for speaker adaptation since the Gaussian variance plays a much less important role than the Gaussian mean value in speaker adaptation, it is of much use in any training situation when little data is available. The training procedure detailed in Chapter 2.1.4 and all adaptation methods detailed in this thesis also implement Equation 3.18 during parameter re-estimation in the form of a variance floor. This prevents variance values from reaching unrealistically low values when little data is used for estimation or adaptation purposes, thereby improving generalisation.

Mean and variance adaptation

Lee *et al.* [24] proposes a third approach where mean and precision parameters are adapted according to a joint mean and precision prior distribution derived from the set of speaker independent Gaussian mixtures. It has been shown [59, p. 169] that the choice of a normal-Gamma joint prior distribution forms a conjugate family for the mean and precision of a sample from a normal distribution. The joint prior distribution of the mean μ and precision r parameters is as follows: the conditional distribution of μ given r is a normal distribution with mean m and precision wr where $w > 0$, and the marginal distribution of r is a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, i.e.,

$$g(\mu, r) \propto r^{1/2} e^{-(wr/2)(\mu-m)^2} r^{\alpha-1} e^{-\beta r}. \tag{3.19}$$

The Gaussian likelihood function given by (similar to Equation 3.8)

$$f(X|\mu, r) \propto r^{n/2} e^{-(r/2)[nS+n(\mu-\bar{x})^2]} \tag{3.20}$$

can be combined with the prior $g(\mu, r)$ of Equation 3.19 to form the posterior p.d.f. $g(\mu, r|X)$

$$g(\mu, r|X) \propto f(X|\mu, r)g(\mu, r) \propto \{r^{1/2} e^{-(1/2)[wr(\mu-m)^2+nr(\mu-\bar{x})^2]}\} r^{\alpha+n/2-1} e^{-\beta r-(nr/2)S}. \tag{3.21}$$

By using the equality

$$wr(\mu - m)^2 + nr(\mu - \bar{x})^2 = (w+n)r\left(\mu - \frac{wm + n\bar{x}}{w+n}\right)^2 + \frac{wn}{w+n}r(m-\bar{x})^2, \tag{3.22}$$

it is noted that the posterior p.d.f. $g(\mu, r|X)$ (from Equation 3.21) of μ and r is also a joint normal-Gamma distribution (similar to the joint prior of μ and r in Equation 3.19) with the following form [59, p. 169]

$$g(\mu, r|X) \propto \{r^{1/2}e^{-(r/2)(w+n)(\mu-\hat{m})^2}\}r^{\hat{\alpha}-1}e^{-\hat{\beta}r}, \tag{3.23}$$

which is discussed in detail next. The part between braces on the right hand side of Equation 3.23 expresses the conditional distribution of μ for a given r and given the observations, which is a normal distribution with mean $\hat{m}$ given by

$$\hat{m} = \frac{wm + n\bar{x}}{w+n} \tag{3.24}$$

and precision $(w+n)r$. The second part on the right hand side of Equation 3.23 expresses the marginal distribution of r given the observations, which is a Gamma distribution with parameters $\hat{\alpha}$ and $\hat{\beta}$ given by

$$\hat{\alpha} = \alpha + n/2 \tag{3.25}$$

and

$$\hat{\beta} = \beta + \frac{n}{2}S + \frac{wn(m-\bar{x})^2}{2(w+n)}. \tag{3.26}$$

It is perhaps not immediately apparent from Equation 3.23 that the marginal distribution of r is simply the second part on the right hand side of the equation, until one considers that the integral over μ of the normalised first part on the right hand side of the equation (the normal distribution) is independent of r, rendering the remaining part the marginal distribution of r, i.e.

$$\begin{aligned} g(r|X) = \int g(\mu, r|X)d\mu \propto & \int \{[(w+n)r]^{1/2} e^{-(r/2)(w+n)(\mu-\hat{m})^2}\} r^{\hat{\alpha}-1} e^{-\hat{\beta}r} d\mu \\ & \propto r^{\hat{\alpha}-1} e^{-\hat{\beta}r} \int [(w+n)r]^{1/2} e^{-(r/2)(w+n)(\mu-\hat{m})^2} \, d\mu \qquad (3.27) \\ & \propto r^{\hat{\alpha}-1} e^{-\hat{\beta}r}. \end{aligned}$$

The posterior distribution of μ and r shows that they are dependent. The joint MAP estimate of μ and r is given by the mode of the 2-dimensional posterior distribution, while the MSE estimate is given by the mean of the distribution. Inspection reveals that the joint posterior distribution (Equation 3.23) has an axis of symmetry along $\mu = \hat{m}$ and thus the expectation value of μ, as well as the value of the mode of μ are independent of r and are equal to $\hat{m}$. Both the MAP and the MSE estimates for μ are given by the mean of the normal distribution

$$\mu_{\text{MAP}} = \mu_{\text{MSE}} = \hat{m} = \frac{wm + n\bar{x}}{w+n}. \qquad (3.28)$$

The MAP estimate of the Gaussian precision is calculated by differentiating the joint posterior distribution (Equation 3.23) with respect to r and finding the root of the equation. The calculation is greatly simplified since we know that the mode is located along $\mu = \hat{m}$ and thus we calculate

$$\frac{\partial}{\partial r} g(\hat{m}, r|X) = \frac{\partial}{\partial r} r^{\hat{\alpha}-1/2} e^{-\hat{\beta}r} = 0$$

$$(\hat{\alpha} - 1/2) r^{\hat{\alpha}-3/2} e^{-\hat{\beta}r} - r^{\hat{\alpha}-1/2}(-\hat{\beta}) e^{-\hat{\beta}r} = 0$$

and therefore the MAP estimate of the Gaussian precision is given by

$$r_{\text{MAP}} = \frac{\hat{\alpha} - 1/2}{\hat{\beta}} = \frac{2\alpha - 1 + n}{2\beta + nS + \frac{wn}{w+n}(m - \bar{x})^2}. \qquad (3.29)$$

The MSE estimate of the Gaussian precision is simply the mean of the marginal posterior Gamma distribution and is given by (from Equations 3.25 and 3.26)

$$r_{\text{MSE}} = \frac{\hat{\alpha}}{\hat{\beta}} = \frac{2\alpha + n}{2\beta + nS + \frac{wn}{w+n}(m - \bar{x})^2}. \tag{3.30}$$

We note at this point that Lee *et al.* [24] used the *mean* (not the mode as in Equation 3.29) of the marginal distribution of r, i.e. the mean of the Gamma p.d.f. $\hat{\alpha}/\hat{\beta}$, as the MAP estimate. This choice is inconsistent with the definition of MAP estimation, requiring use of the mode of the posterior distribution. In a later paper, Gauvain & Lee [58] refer to the correct MAP estimate $\hat{r} = (\hat{\alpha} - 1/2)/\hat{\beta}$. There is, however, a problem with using the mode of the posterior, since as Equation 3.29 shows, the precision is only valid (larger than zero) for $\hat{\alpha} = \alpha + n/2 > 1/2$. This may pose a problem when no observations are made ($n = 0$), depending on the value of α, for which case it is probably sensible to select to use the mean of the posterior, i.e. $\hat{\alpha}/\hat{\beta}$.

With the MAP estimates now derived, the selection of parameters for the normal-Gamma prior distributions remains to be addressed. Making exactly the same choices as in Equations 3.14-3.16 with respect to the prior normal distribution values (m and $\frac{1}{wr}$), as well as the prior expectation value of the variance (β/α), we get

$$m = \sum_{k=1}^{K} \tilde{c}_k \tilde{m}_k, \tag{3.31}$$

$$\frac{1}{wr} = \sum_{k=1}^{K} \tilde{c}_k (\tilde{m}_k - m)^2 \tag{3.32}$$

and

$$\beta/\alpha = \sum_{k=1}^{K} \tilde{c}_k \tilde{\sigma}_k^2. \tag{3.33}$$

By choosing somewhat arbitrarily the value of $\beta = 1$ we can solve for (Equation 3.33)

$$\alpha = \frac{1}{\sum_{k=1}^{K} \tilde{c}_k \tilde{\sigma}_k^2} \tag{3.34}$$

and using the prior mean value of the Gamma distribution $\alpha/\beta = \alpha$ in place of r in Equation 3.32 we solve for

$$w = \frac{\sum_{k=1}^{K} \tilde{c}_k \tilde{\sigma}_k^2}{\sum_{k=1}^{K} \tilde{c}_k (\tilde{m}_k - m)^2}. \tag{3.35}$$

Since β was chosen arbitrarily, the prior variance of the precision was not considered. We know, however, that for a sample from a Gamma distribution the expectation value of the variance is given by α/β^2, which in our case simply equals α. It is intuitively pleasing that the variance of the precision in the prior is equal to the chosen expectation value of the precision, meaning that large prior values of the precision are associated with larger variance and thus less certainty than for lower values of the precision.

The MAP equations we derived here are the same as those derived by Lee *et al.* [24], except for the offset in the variance estimate, but our derivation shows perhaps more clearly the meaning of the choices with respect to the prior parameters. The procedure outlined above is only for parameter estimation of univariate Gaussian distributions. This is not a problem if diagonal covariance matrices are used with multivariate Gaussian distributions, as they then simplify to independent univariate estimation problems. The next section discusses the implementation of Bayesian adaptation for the general multivariate case.

Multivariate normal distribution adaptation

The derivation of posterior distributions for a multivariate Gaussian distribution is a generalisation of the discussion in the previous section. We proceed to give the derivation of the Bayesian estimates for a joint mean and variance prior distribution. It has been shown [59, p. 177] that the choice of a normal-Wishart joint prior distribution forms a jointly conjugate family for the mean and precision of a sample from a multivariate normal distribution. The joint prior distribution of the mean $\boldsymbol{\mu}$ and precision $\mathbf{R}$ parameters is as follows: the conditional distribution of $\boldsymbol{\mu}$ given $\mathbf{R}$ is a normal distribution with mean vector $\mathbf{m}$ and precision matrix $w\mathbf{R}, w > 0$, and the marginal distribution of $\mathbf{R}$ is a Wishart distribution with $\alpha > D-1$ degrees of freedom and a symmetric positive definite precision matrix $\boldsymbol{\Upsilon}$. The joint prior normal-Wishart distribution is given by [59, p. 178]

$$g(\boldsymbol{\mu}, \mathbf{R}) \propto |\mathbf{R}|^{1/2} e^{-(w/2)(\boldsymbol{\mu}-\mathbf{m})^T \mathbf{R}(\boldsymbol{\mu}-\mathbf{m})} \, |\mathbf{R}|^{(\alpha-D-1)/2} e^{-(1/2)\,\mathrm{tr}[\boldsymbol{\Upsilon}\mathbf{R}]}. \tag{3.36}$$

With the multivariate Gaussian likelihood function for observations $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ given by

$$\begin{aligned} f(\mathbf{X}|\boldsymbol{\mu}, \mathbf{R}) &\propto |\mathbf{R}|^{n/2} e^{-(1/2)\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T \mathbf{R}(\mathbf{x}_i-\boldsymbol{\mu})} \\ &\propto |\mathbf{R}|^{n/2} e^{-(1/2)\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})^T \mathbf{R}(\mathbf{x}_i-\bar{\mathbf{x}})+n(\boldsymbol{\mu}-\bar{\mathbf{x}})^T \mathbf{R}(\boldsymbol{\mu}-\bar{\mathbf{x}})} \\ &\propto |\mathbf{R}|^{n/2} e^{-(n/2)\,[\mathrm{tr}(\mathbf{S}\mathbf{R})+(\boldsymbol{\mu}-\bar{\mathbf{x}})^T \mathbf{R}(\boldsymbol{\mu}-\bar{\mathbf{x}})]} \end{aligned} \tag{3.37}$$

and using the equality

$$\begin{aligned} &w(\boldsymbol{\mu}-\mathbf{m})^T \mathbf{R}(\boldsymbol{\mu}-\mathbf{m}) + n(\boldsymbol{\mu}-\bar{\mathbf{x}})^T \mathbf{R}(\boldsymbol{\mu}-\bar{\mathbf{x}}) = \\ &\qquad (w+n)\Big(\boldsymbol{\mu} - \frac{w\mathbf{m}+n\bar{\mathbf{x}}}{w+n}\Big)^T \mathbf{R}\Big(\boldsymbol{\mu} - \frac{w\mathbf{m}+n\bar{\mathbf{x}}}{w+n}\Big) + \frac{wn}{w+n}(\mathbf{m}-\bar{\mathbf{x}})^T \mathbf{R}(\mathbf{m}-\bar{\mathbf{x}}) \end{aligned} \tag{3.38}$$

the posterior p.d.f. $g(\boldsymbol{\mu}, \mathbf{R}|\mathbf{X}) \propto f(\mathbf{X}|\boldsymbol{\mu}, \mathbf{R}) g(\boldsymbol{\mu}, \mathbf{R})$ is also a normal-Wishart distribution with the following form [59, p. 178]

$$g(\boldsymbol{\mu}, \mathbf{R}|\mathbf{X}) \propto \{|\mathbf{R}|^{1/2} e^{-(1/2)(w+n)(\boldsymbol{\mu}-\hat{\mathbf{m}})^T \mathbf{R}(\boldsymbol{\mu}-\hat{\mathbf{m}})}\} \, \{|\mathbf{R}|^{(\alpha+n-D-1)/2} e^{-(1/2)\,\mathrm{tr}[\hat{\boldsymbol{\Upsilon}}\mathbf{R}]}\}, \tag{3.39}$$

which is discussed next. The posterior conditional distribution of $\boldsymbol{\mu}$ for a given $\mathbf{R}$ and given the observations is a normal distribution with mean $\hat{\mathbf{m}}$ given by

$$\hat{\mathbf{m}} = \frac{w\mathbf{m}+n\bar{\mathbf{x}}}{w+n} \tag{3.40}$$

and precision $(w+n)\mathbf{R}$. The marginal posterior distribution of $\mathbf{R}$ given the observations is a Wishart distribution with $\alpha + n$ degrees of freedom and precision matrix $\hat{\boldsymbol{\Upsilon}}$ given by

$$\hat{\boldsymbol{\Upsilon}} = \boldsymbol{\Upsilon} + n\mathbf{S} + \frac{wn}{w+n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T. \tag{3.41}$$

Since the posterior conditional normal distribution has axis of symmetry along $\boldsymbol{\mu} = \hat{\mathbf{m}}$, the MAP and MSE estimates of $\boldsymbol{\mu}$ are independent of $\mathbf{R}$ and are given by

$$\boldsymbol{\mu}_{\text{MAP}} = \boldsymbol{\mu}_{\text{MSE}} = \hat{\mathbf{m}} = \frac{w\mathbf{m} + n\bar{\mathbf{x}}}{w+n}. \tag{3.42}$$

The MSE estimate of the Gaussian covariance can be written in terms of the mean value of the posterior marginal Wishart p.d.f.

$$\mathbf{R}^{-1}_{\text{MSE}} = \frac{\hat{\boldsymbol{\Upsilon}}}{\alpha+n} = \frac{\boldsymbol{\Upsilon} + n\mathbf{S} + \frac{wn}{w+n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T}{\alpha+n} \tag{3.43}$$

while the MAP estimate can be derived by calculating the derivative of $g(\hat{\mathbf{m}}, \mathbf{R}|\mathbf{X})$ (from Equation 3.39) with respect to $\mathbf{R}$ and setting it equal to zero, which delivers

$$\mathbf{R}^{-1}_{\text{MAP}} = \frac{\boldsymbol{\Upsilon} + n\mathbf{S} + \frac{wn}{w+n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T}{\alpha+n-D}. \tag{3.44}$$

It can be attempted to estimate values for the parameters of the prior distributions from speaker independent mixture models in the same way as for the univariate case, using the criteria of Equations 3.31-3.33:

$$\mathbf{m} = \sum_{k=1}^{K} \tilde{c}_k \tilde{\mathbf{m}}_k, \tag{3.45}$$

$$(w\mathbf{R})^{-1} = \sum_{k=1}^{K} \tilde{c}_k (\tilde{\mathbf{m}}_k - \mathbf{m})(\tilde{\mathbf{m}}_k - \mathbf{m})^T \tag{3.46}$$

and

$$\boldsymbol{\Upsilon}/\alpha = \sum_{k=1}^{K} \tilde{c}_k \tilde{\boldsymbol{\Sigma}}_k, \tag{3.47}$$

where $\tilde{\Sigma}_k$ is the covariance matrix for mixture k of the speaker independent model. Setting $\mathbf{R}^{-1}$ equal to Υ/α in Equation 3.47 uses the expectation value of the prior covariance, but causes the equations to have no solution since $\mathbf{R}$ is over-determined. However, if diagonal dominance of the precision is assumed, use of the trace on both sides of Equation 3.46 allows a reasonable solution to be found for w. A choice with respect to either Υ or α still needs to be made. Without further information, it may be necessary to make an arbitrary assignment. A choice that will satisfy the constraints is e.g. selecting $\alpha = D + 1$. We do not discuss prior estimation for the multivariate case in more detail here, but return to the topic in Section 3.2.5 where a method for estimation of prior parameters for a multivariate mixture distribution is discussed.

The preceding procedures are applicable for the estimation of (single) Gaussian distributions, which we have found to be useful for speaker adaptation, even when cross-language prior models are used [33]. However, to estimate complex models commonly used for speaker independent recognition, we have to consider the problem of adaptation of mixture density models, which is addressed in the next section.

3.2.3 Mixture density HMM parameter distributions

This section expands on the previous sections that dealt with mean and variance adaptation in a Gaussian framework and places those derivations in the context of Gaussian mixture densities used as output distributions in an HMM with state transition probabilities. Gauvain & Lee [58, 60, 61] suggested applying Bayesian learning of Gaussian mixture components to speaker adaptation of CDHMMs. The method uses parameters of individual Gaussian components in a speaker independent HMM to compute prior distribution parameters for the adaptation of the Gaussian mean, variance and component weight, as well as for the adaptation of state transition parameters within a single framework. We proceed to discuss the prior distribution for a mixture density.

Mixture weight distributions

The Gaussian mixture density for a given state j can be considered a density associated with a statistical population consisting of a mixture of K component populations with mixing proportions $c_{j1}, ..., c_{jK}$. The sizes of the component populations can then be considered to be distributed according to a multinomial distribution, given by

$$f(n_{j1}, .., n_{jK} | c_{j1}, .., c_{jK}) \propto \prod_{k=1}^{K} c_{jk}^{n_{jk}} \tag{3.48}$$

where n_{jk} occurrences of each of the $1 \leq k \leq K$ mixture densities in state j are observed. It is known that the Dirichlet density [59, p. 174]

$$g(c_{j1}, .., c_{jK}) \propto \prod_{k=1}^{K} c_{jk}^{v_{jk}-1} \tag{3.49}$$

with prior parameters $v_{j1}, .., v_{jK}$ in this case, is a conjugate density for a sample from the multinomial distribution and is thus suitable for expressing prior information about the mixing proportions. The posterior Dirichlet p.d.f. of the mixing proportions, or mixture weights as we refer to them, is simply given by

$$\begin{aligned} g(c_{j1}, .., c_{jK} | n_{j1}, .., n_{jK}) &\propto f(n_{j1}, .., n_{jK} | c_{j1}, .., c_{jK}) g(c_{j1}, .., c_{jK}) \\ &\propto \prod_{k=1}^{K} c_{jk}^{n_{jk}} \prod_{k=1}^{K} c_{jk}^{v_{jk}-1} \\ &\propto \prod_{k=1}^{K} c_{jk}^{v_{jk}+n_{jk}-1}. \end{aligned} \tag{3.50}$$

The MAP estimate for the mixture weight is given by the mode of Equation 3.50 [61]

$$c_{jk\,\mathrm{MAP}} = \frac{\hat{v}_{jk} - 1}{\sum_{l=1}^{K} (\hat{v}_{jl} - 1)} \tag{3.51}$$

where $\hat{v}_{jk} = v_{jk} + n_{jk}$ is the parameter of the posterior Dirichlet distribution. The MSE estimate for the mixture weight is given by the mean of Equation 3.50 [59, p. 51]:

$$c_{jk\,\mathrm{MSE}} = \frac{\hat{v}_{jk}}{\sum_{l=1}^{K} \hat{v}_{jl}}. \tag{3.52}$$

Transition probability distributions

The HMM state transition probability parameters can be dealt with in much the same way as the mixture weight parameters. If the assumption is made that the transition probability parameters are independent of the other HMM parameters and that each row of the transition probability matrix **A** is independent, which is true for a first order Markov process, each row of the transition probability matrix can be considered to be the parameter of a multinomial distribution, characterising the number of transitions from state i to each state in the HMM, with likelihood function

$$f(n_{i1}, .., n_{iN} | a_{i1}, .., a_{iN}) \propto \prod_{j=1}^{N} a_{ij}^{n_{ij}} \tag{3.53}$$

where n_{ij} transitions from state i to each of the $1 \leq j \leq N$ states are observed. The prior Dirichlet density is expressed by

$$g(a_{i1}, .., a_{iN}) \propto \prod_{j=1}^{N} a_{ij}^{\eta_{ij}-1} \tag{3.54}$$

with prior parameters $\eta_{i1}, .., \eta_{iN}$ for the transition probabilities from state i. Similar to Equation 3.50, but calculating the joint p.d.f. of the transition probabilities from each state including dummy state 0, we derive the joint posterior distribution

$$\begin{aligned} g(\mathbf{A} | \{n_{ij}\}_{\, i=0,..,N;\, j=1,..,N}) &\propto \prod_{i=0}^{N} f(n_{i1}, .., n_{iN} | a_{i1}, .., a_{iN}) g(a_{i1}, .., a_{iN}) \\ &\propto \prod_{i=0}^{N} \left[\prod_{j=1}^{N} a_{ij}^{\eta_{ij}+n_{ij}-1} \right]. \end{aligned} \tag{3.55}$$

The MAP estimate for the transition probability parameters is given by the mode of Equation 3.55:

$$a_{ij\,\mathrm{MAP}} = \frac{\hat{\eta}_{ij} - 1}{\sum_{l=1}^{K} (\hat{\eta}_{il} - 1)} \tag{3.56}$$

where $\hat{\eta}_{ij} = \eta_{ij} + n_{ij}$ is the parameter of the posterior Dirichlet distribution. The MSE estimate for the mixture weight is given by the mean of Equation 3.55:

$$a_{ij\,\mathrm{MSE}} = \frac{\hat{\eta}_{ij}}{\sum_{l=1}^{K} \hat{\eta}_{il}}. \tag{3.57}$$

Now that we have prior distribution families for every parameter of the Gaussian mixture HMM in isolation, we combine these prior distributions to form a joint prior distribution.

Joint prior distribution for HMM parameters

Assuming independence between the transition probability parameters, the mixture weight parameters and the parameters of the mixture distribution, the prior distributions of the parameters of the Gaussian mixture HMM $\boldsymbol{\lambda}$ can be combined in a joint prior distribution

$$g(\boldsymbol{\lambda}) \propto \prod_{i=1}^{N} \left[a_{0i}^{\eta_{0i}-1} \left[\prod_{j=1}^{N} a_{ij}^{\eta_{ij}-1} \right] \left[\prod_{k=1}^{K} c_{ik}^{\nu_{ik}-1} g(\boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}) \right] \right] \tag{3.58}$$

with the prior normal-Wishart mixture parameter distribution given by (see Equation 3.36)

$$g(\boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}) \propto |\mathbf{R}_{ik}|^{1/2} e^{-(w_{ik}/2)(\boldsymbol{\mu}_{ik} - \mathbf{m}_{ik})^T \mathbf{R}_{ik} (\boldsymbol{\mu}_{ik} - \mathbf{m}_{ik})} |\mathbf{R}_{ik}|^{(\alpha_{ik} - D - 1)/2} e^{-(1/2)\,\mathrm{tr}[\boldsymbol{\Upsilon}_{ik} \mathbf{R}_{ik}]}. \tag{3.59}$$

Under the complete data density assumption, which explicitly uses state and mixture alignment, posterior distributions for the parameters of an HMM can be derived. This is done next.

Complete data HMM likelihood function

The *complete data* likelihood for a mixture density HMM $\boldsymbol{\lambda}$ is the joint likelihood of the observations $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_T\}$, the state alignment given by $\mathbf{q} = \{q_1, ..., q_T\}$ and the mixture alignment given by $\mathbf{l} = \{l_1, ..., l_T\}$ (see Equation 2.6):

$$f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\boldsymbol{\lambda}) \propto \prod_{t=1}^{T} \left[a_{q_{t-1}q_t} c_{q_t l_t} |\mathbf{R}_{q_t l_t}|^{1/2} e^{-(1/2)(\boldsymbol{\mu}_{q_t l_t} - \mathbf{x}_t)^T \mathbf{R}_{q_t l_t} (\boldsymbol{\mu}_{q_t l_t} - \mathbf{x}_t)} \right], \tag{3.60}$$

From the state and mixture alignments, mixture occupancy $\gamma_{jk}(t)$ and transition occupancy $\xi_{ij}(t)$ (described by Equations 2.12 and 2.13 respectively) can be computed. From a decoding point of view, this correspond to Viterbi state alignment and choosing the most likely mixture at each state aligned observation frame. We note that the forward-backward algorithm can also be used to calculate values for the statistics $\gamma_{jk}(t)$ and $\xi_{ij}(t)$, but for the complete data likelihood we assume exact state and mixture alignment. In the following section (Section 3.2.4) this constraint will be eased when the estimation strategy is discussed. Further statistics can be defined:

$$\gamma_{ik} = \sum_{t=1}^{T} \gamma_{ik}(t), \tag{3.61}$$

$$\xi_{ij} = \sum_{t=1}^{T} \xi_{ij}(t), \tag{3.62}$$

$$\bar{\mathbf{x}}_{ik} = (1/\gamma_{ik}) \sum_{t=1}^{T} \gamma_{ik}(t) \mathbf{x}_t \tag{3.63}$$

and

$$\mathbf{S}_{ik} = (1/\gamma_{ik}) \sum_{t=1}^{T} \gamma_{ik}(t)(\mathbf{x}_t - \bar{\mathbf{x}}_{ik})(\mathbf{x}_t - \bar{\mathbf{x}}_{ik})^T \tag{3.64}$$

where γ_{jk} is the total occupancy of mixture k in state j, ξ_{ij} is number of transitions in the aligned data from state i to state j, and $\bar{\mathbf{x}}_{ik}$ is the sample mean and $\mathbf{S}_{ik}$ the sample variance of observations in mixture k of state i. Using the statistics of Equations 3.61-3.64 and the compact form of the Gaussian likelihood function (see Equation 3.37), the complete data likelihood function of Equation 3.60 can then be written as

$$f(\mathbf{X},\mathbf{q},\mathbf{l}|\boldsymbol{\lambda}) \propto \prod_{i=1}^{N}\left[a_{0i}^{\xi_{0i}}\left[\prod_{j=1}^{N} a_{ij}^{\xi_{ij}}\right]\left[\prod_{k=1}^{K} c_{ik}^{\gamma_{ik}}|\mathbf{R}_{ik}|^{\gamma_{ik}/2}e^{-(\gamma_{ik}/2)[\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik})+(\boldsymbol{\mu}_{ik}-\bar{\mathbf{x}}_{ik})^T\mathbf{R}_{ik}(\boldsymbol{\mu}_{ik}-\bar{\mathbf{x}}_{ik})]}\right]\right]. \tag{3.65}$$

Complete data posterior distribution

The prior $g(\boldsymbol{\lambda})$ (Equation 3.58) includes the kernel density of the complete data likelihood function $f(\mathbf{X},\mathbf{q},\mathbf{l}|\boldsymbol{\lambda})$ (Equation 3.65) and is thus a conjugate prior distribution for the complete data density. From Equations 3.58 and 3.65, the joint posterior distribution $g(\mathbf{q},\mathbf{l},\hat{\boldsymbol{\lambda}}|\mathbf{X})$ for the complete data density is therefore given by

$$\begin{aligned} g(\mathbf{q},\mathbf{l},\hat{\boldsymbol{\lambda}}|\mathbf{X}) \propto f(\mathbf{X},\mathbf{q},\mathbf{l}|\boldsymbol{\lambda})g(\boldsymbol{\lambda}) \propto \prod_{i=1}^{N}\Bigg[a_{0i}^{\eta_{0i}-1}\left[\prod_{j=1}^{N} a_{ij}^{\eta_{ij}-1}\right] \\ \left[\prod_{k=1}^{K} c_{ik}^{v_{ik}-1}|\mathbf{R}_{ik}|^{1/2}e^{-(w_{ik}/2)(\boldsymbol{\mu}_{ik}-\mathbf{m}_{ik})^T\mathbf{R}_{ik}(\boldsymbol{\mu}_{ik}-\mathbf{m}_{ik})}\,|\mathbf{R}_{ik}|^{(\alpha_{ik}-D-1)/2}e^{-(1/2)\,\mathrm{tr}[\boldsymbol{\Upsilon}_{ik}\mathbf{R}_{ik}]}\right]\Bigg] \\ \prod_{i=1}^{N}\left[a_{0i}^{\xi_{0i}}\left[\prod_{j=1}^{N} a_{ij}^{\xi_{ij}}\right]\left[\prod_{k=1}^{K} c_{ik}^{\gamma_{ik}}|\mathbf{R}_{ik}|^{\gamma_{ik}/2}e^{-(\gamma_{ik}/2)[\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik})+(\boldsymbol{\mu}_{ik}-\bar{\mathbf{x}}_{ik})^T\mathbf{R}_{ik}(\boldsymbol{\mu}_{ik}-\bar{\mathbf{x}}_{ik})]}\right]\right]. \end{aligned} \tag{3.66}$$

By re-arranging terms, $g(\mathbf{q},\mathbf{l},\hat{\boldsymbol{\lambda}}|\mathbf{X})$ can be written in the same form as the joint prior distribution $g(\boldsymbol{\lambda})$ (Equation 3.58) by:

$$\begin{aligned} g(\mathbf{q},\mathbf{l},\hat{\boldsymbol{\lambda}}|\mathbf{X}) \propto \prod_{i=1}^{N}\Bigg[a_{0i}^{\eta_{0i}+\xi_{0i}-1}\left[\prod_{j=1}^{N} a_{ij}^{\eta_{ij}+\xi_{ij}-1}\right] \\ \left[\prod_{k=1}^{K} c_{ik}^{v_{ik}+\gamma_{ik}-1}|\mathbf{R}_{ik}|^{1/2}e^{-(1/2)(w_{ik}+\gamma_{ik})(\boldsymbol{\mu}_{ik}-\hat{\mathbf{m}}_{ik})^T\mathbf{R}_{ik}(\boldsymbol{\mu}_{ik}-\hat{\mathbf{m}}_{ik})}\,|\mathbf{R}_{ik}|^{(\alpha_{ik}+\gamma_{ik}-D-1)/2}e^{-(1/2)\,\mathrm{tr}[\hat{\boldsymbol{\Upsilon}}_{ik}\mathbf{R}_{ik}]}\right]\Bigg], \end{aligned} \tag{3.67}$$

where the mean of the posterior Gaussian mean $\hat{\mathbf{m}}_{ik}$ is given by (see Equation 3.40)

$$\hat{\mathbf{m}}_{ik} = \frac{w_{ik}\mathbf{m}_{ik} + \gamma_{ik}\bar{\mathbf{x}}_{ik}}{w_{ik} + \gamma_{ik}}, \tag{3.68}$$

and the precision of the posterior Wishart precision $\hat{\mathbf{\Upsilon}}_{ik}$ is given by (see Equations 3.41 and 3.38)

$$\begin{aligned}\hat{\mathbf{\Upsilon}}_{ik} &= \mathbf{\Upsilon}_{ik} + \gamma_{ik}\mathbf{S}_{ik} + \frac{w_{ik}\gamma_{ik}}{w_{ik} + \gamma_{ik}}(\mathbf{m}_{ik} - \bar{\mathbf{x}}_{ik})(\mathbf{m}_{ik} - \bar{\mathbf{x}}_{ik})^T \\ &= \mathbf{\Upsilon}_{ik} + \sum_{t=1}^{T} \gamma_{ik}(t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)^T + w_{ik}(\hat{\mathbf{m}}_{ik} - \mathbf{m}_{ik})(\hat{\mathbf{m}}_{ik} - \mathbf{m}_{ik})^T. \end{aligned} \tag{3.69}$$

The solutions to the other posterior distribution parameters are also similar to those presented in Section 3.2.2 and this section, except that they are given in terms of the sufficient statistics of Equations 3.61-3.64. The parameters of the posterior Dirichlet transition probability and mixture weight densities ($\hat{\eta}_{ij}$ and $\hat{v}_{ik}$ respectively), the relative precision of the conditional posterior mean density, w_{ik} and the number of degrees of freedom of the posterior Wishart precision density are given by:

$$\hat{\eta}_{ij} = \eta_{ij} + \xi_{ij} \tag{3.70}$$

$$\hat{v}_{ik} = v_{ik} + \gamma_{ik} \tag{3.71}$$

$$\hat{w}_{ik} = w_{ik} + \gamma_{ik} \tag{3.72}$$

and

$$\hat{\alpha}_{ik} = \alpha_{ik} + \gamma_{ik}. \tag{3.73}$$

MAP and MSE parameter estimates

From the posterior distributions, MAP and MSE parameter estimates can be made. For the Gaussian mean distribution, the mean and the mode of the posterior distributions are the same and the MAP and MSE parameters are given by $\hat{\mathbf{m}}_{ik}$ (Equation 3.68)

$$\boldsymbol{\mu}_{ik\,\mathrm{MAP}} = \boldsymbol{\mu}_{ik\,\mathrm{MSE}} = \frac{w_{ik}\mathbf{m}_{ik} + \gamma_{ik}\bar{\mathbf{x}}_{ik}}{w_{ik} + \gamma_{ik}}. \tag{3.74}$$

For the Dirichlet and Wishart distributions the mean and mode differs. The MAP parameters are given by (see Equations 3.56, 3.51 and 3.44)

$$a_{ij\,\mathrm{MAP}} = \frac{\eta_{ij} + \xi_{ij} - 1}{\sum_{l=1}^{K}(\eta_{il} + \xi_{il} - 1)} \tag{3.75}$$

$$c_{ik\,\mathrm{MAP}} = \frac{v_{ik} + \gamma_{ik} - 1}{\sum_{l=1}^{K}(v_{il} + \gamma_{il} - 1)} \tag{3.76}$$

$$\mathbf{R}^{-1}_{ik\,\mathrm{MAP}} = \frac{\boldsymbol{\Upsilon}_{ik} + \sum_{t=1}^{T}\gamma_{ik}(t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)^T + w_{ik}(\hat{\mathbf{m}}_{ik} - \mathbf{m}_{ik})(\hat{\mathbf{m}}_{ik} - \mathbf{m}_{ik})^T}{\alpha_{ik} + \gamma_{ik} - D}, \tag{3.77}$$

and the MSE parameters are given by (see Equations 3.57, 3.52 and 3.43)

$$a_{ij\,\mathrm{MSE}} = \frac{\eta_{ij} + \xi_{ij}}{\sum_{l=1}^{K}(\eta_{il} + \xi_{il})} \tag{3.78}$$

$$c_{ik\,\mathrm{MSE}} = \frac{v_{ik} + \gamma_{ik}}{\sum_{l=1}^{K}(v_{il} + \gamma_{il})} \tag{3.79}$$

$$\mathbf{R}^{-1}_{ik\,\mathrm{MSE}} = \frac{\boldsymbol{\Upsilon}_{ik} + \sum_{t=1}^{T}\gamma_{ik}(t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)^T + w_{ik}(\hat{\mathbf{m}}_{ik} - \mathbf{m}_{ik})(\hat{\mathbf{m}}_{ik} - \mathbf{m}_{ik})^T}{\alpha_{ik} + \gamma_{ik}}. \tag{3.80}$$

It is apparent that the MAP estimates (Equations 3.75-3.77) are invalid under certain conditions ($\eta_{ij} + \xi_{ij} < 1$, $v_{ik} + \gamma_{ik} < 1$ and $\alpha_{ik} + \gamma_{ik} \leq D$). This is because the mode of the posterior distribution is undefined under these conditions. The MSE estimates do not suffer from this problem though.

The MAP and MSE estimates of Equations 3.74-3.80 have been derived based on the complete data assumption, i.e. that state and mixture alignment information is available. In practice, this information has to be computed from the adaptation data. The next section discusses an iterative estimation technique for the incomplete data scenario where state and mixture occupancy is not observed and also generalises the results of this section to include all possible state and mixture sequences.

3.2.4 Estimation algorithm

Gauvain & Lee [61] propose using an expectation maximisation (EM) [41] estimation strategy for MAP parameter estimation. The proposed strategy is based on the maximisation of the auxiliary function $R(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$, representing the *expectation* of the complete data posterior model log-likelihood ($\log[f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\hat{\boldsymbol{\lambda}})g(\hat{\boldsymbol{\lambda}})]$)

$$\begin{aligned} R(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) &= E\big[\log[f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\hat{\boldsymbol{\lambda}})g(\hat{\boldsymbol{\lambda}})|\mathbf{X}, \boldsymbol{\lambda}]\big] \\ &= E\big[\log[f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\hat{\boldsymbol{\lambda}})|\mathbf{X}, \boldsymbol{\lambda}]\big] + \log g(\hat{\boldsymbol{\lambda}}) \\ &= Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) + \log g(\hat{\boldsymbol{\lambda}}), \end{aligned} \tag{3.81}$$

given the observations $\mathbf{X}$, a current model $\boldsymbol{\lambda}$ and where $Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ is the auxiliary equation for conventional Gaussian mixture ML procedures and is given by

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \frac{1}{f(\mathbf{X}|\boldsymbol{\lambda})} \sum_{\mathbf{q}} \sum_{\mathbf{l}} f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\boldsymbol{\lambda}) \log f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\hat{\boldsymbol{\lambda}}). \tag{3.82}$$

Similar to maximising $Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ (see [62]), maximising $R(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ in each iteration, $R(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) > R(\boldsymbol{\lambda}, \boldsymbol{\lambda})$ implies a monotonic increase in posterior likelihood $f(\mathbf{X}|\hat{\boldsymbol{\lambda}})g(\hat{\boldsymbol{\lambda}}) > f(\mathbf{X}|\boldsymbol{\lambda})g(\boldsymbol{\lambda})$ until $\hat{\boldsymbol{\lambda}}$ reaches a critical point where $f(\mathbf{X}|\hat{\boldsymbol{\lambda}})$ attains a local maximum. Maximisation of $R(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ according to the procedure defined by [61] leads to exactly the re-estimation equations derived in the previous section (Equations 3.74-3.77), as we shall show shortly. The auxiliary function $Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ can be expanded (following [63, p. 9]):

$$\begin{aligned} Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) &= \frac{1}{f(\mathbf{X}|\boldsymbol{\lambda})} \sum_{\mathbf{q}} \sum_{\mathbf{l}} f(\mathbf{X}, \mathbf{q}, \mathbf{l}|\boldsymbol{\lambda}) \Big[\sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log c_{q_t l_t} + \sum_{t=1}^{T} \log \mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{q_t l_t}, \mathbf{R}_{q_t l_t}]\Big] \\ &= \sum_{i=0}^{N} Q_{a_i}[\boldsymbol{\lambda}, \{a_{ij}\}_{j=1}^{N}] + \sum_{i=1}^{N} Q_{c_i}[\boldsymbol{\lambda}, \{c_{ik}\}_{k=1}^{K}] + \sum_{i=1}^{N} \sum_{k=1}^{K} Q_{\mathcal{N}}[\boldsymbol{\lambda}, \boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}], \end{aligned} \tag{3.83}$$

where

$$\begin{aligned}
Q_{a_i}[\boldsymbol{\lambda}, \{a_{ij}\}_{j=1}^{N}] &= \frac{1}{f(\mathbf{X}|\boldsymbol{\lambda})} \sum_{\mathbf{q}} \sum_{\mathbf{l}} \sum_{t=1}^{T} \sum_{j=1}^{N} f(\mathbf{X}, q_{t-1} = i, q_t = j, \mathbf{l}|\boldsymbol{\lambda}) \log a_{ij} \\
&= \sum_{t=1}^{T} \sum_{j=1}^{N} \xi_{ij}(t) \log a_{ij} \quad \text{(reversing order of summation and using Equation 2.13)} \\
&= \sum_{j=1}^{N} \xi_{ij} \log a_{ij} \quad \text{(using Equation 3.62)},
\end{aligned} \tag{3.84}$$

$$\begin{aligned}
Q_{c_i}[\boldsymbol{\lambda}, \{c_{ik}\}_{k=1}^{K}] &= \frac{1}{f(\mathbf{X}|\boldsymbol{\lambda})} \sum_{\mathbf{q}} \sum_{\mathbf{l}} \sum_{k=1}^{K} \sum_{t=1}^{T} f(\mathbf{X}, q_t = i, l_t = k|\boldsymbol{\lambda}) \log c_{ik} \\
&= \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_{ik}(t) \log c_{ik} \quad \text{(reversing order of summation and using Equation 2.12)} \\
&= \sum_{k=1}^{K} \gamma_{ik} \log c_{ik} \quad \text{(using Equation 3.61)},
\end{aligned} \tag{3.85}$$

and

$$\begin{aligned}
Q_{\mathcal{N}}[\boldsymbol{\lambda}, \boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}] &= \frac{1}{f(\mathbf{X}|\boldsymbol{\lambda})} \sum_{\mathbf{q}} \sum_{\mathbf{l}} \sum_{t=1}^{T} f(\mathbf{X}, q_t = i, l_t = k|\boldsymbol{\lambda}) \log \mathcal{N}[\mathbf{x}_t, \boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}] \\
&\propto \sum_{t=1}^{T} \gamma_{ik}(t) \log \left[|\mathbf{R}_{ik}|^{1/2} e^{-(1/2)(\boldsymbol{\mu}_{ik} - \mathbf{x}_t)^T \mathbf{R}_{ik} (\boldsymbol{\mu}_{ik} - \mathbf{x}_t)} \right] \\
&\quad \text{(reversing order of summation and using Equation 2.12)} \\
&\propto \gamma_{ik} \log \left[|\mathbf{R}_{ik}|^{1/2} e^{-(1/2)[\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik}) + (\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})^T \mathbf{R}_{ik} (\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})]} \right] \\
&\quad \text{(following Equation 3.37).}
\end{aligned} \tag{3.86}$$

If we consider maximising $\Psi(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = e^{R(\boldsymbol{\lambda},\hat{\boldsymbol{\lambda}})}$, we get

$$\begin{aligned}
\Psi(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) &= e^{Q(\boldsymbol{\lambda},\hat{\boldsymbol{\lambda}})+\log g(\hat{\boldsymbol{\lambda}})} \\
&= g(\hat{\boldsymbol{\lambda}}) e^{\sum_{i=0}^{N} Q_a[\boldsymbol{\lambda},\{a_{ij}\}_{j=1}^{N}]+\sum_{i=1}^{N} Q_{c_i}[\boldsymbol{\lambda},\{c_{ik}\}_{k=1}^{K}]+\sum_{i=1}^{N}\sum_{k=1}^{K} Q_{\mathcal{N}}[\boldsymbol{\lambda},\boldsymbol{\mu}_{ik},\mathbf{R}_{ik}]} \\
&\propto g(\hat{\boldsymbol{\lambda}}) \left[\prod_{i=0}^{N} e^{Q_a[\boldsymbol{\lambda},\{a_{ij}\}_{j=1}^{N}]}\right] \left[\prod_{i=1}^{N} e^{Q_{c_i}[\boldsymbol{\lambda},\{c_{ik}\}_{k=1}^{K}]}\right] \left[\prod_{i=1}^{N}\prod_{k=1}^{K} e^{Q_{\mathcal{N}}[\boldsymbol{\lambda},\boldsymbol{\mu}_{ik},\mathbf{R}_{ik}]}\right] \\
&\propto g(\hat{\boldsymbol{\lambda}}) \prod_{i=1}^{N} \left[a_{0i}^{\xi_{0i}} \left[\prod_{j=1}^{N} a_{ij}^{\xi_{ij}}\right] \left[\prod_{k=1}^{K} c_{ik}^{\gamma_{ik}} |\mathbf{R}_{ik}|^{\gamma_{ik}/2} e^{-(\gamma_{ik}/2)[\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik})+(\boldsymbol{\mu}_{ik}-\bar{\mathbf{x}}_{ik})^T\mathbf{R}_{ik}(\boldsymbol{\mu}_{ik}-\bar{\mathbf{x}}_{ik})]}\right]\right],
\end{aligned} \tag{3.87}$$

which is of exactly the same form as the joint posterior distribution for the complete data density given in Equation 3.66 and therefore maximisation of Equation 3.87 leads to the MAP estimation equations derived in the previous section (Equations 3.74-3.77). Use of the auxiliary function in the derivation ensures that the likelihood of the MAP estimates monotonically increase in every iteration. Unfortunately this theoretical result does not apply for the MSE estimates. Since the maximisation of the auxiliary function is done for arbitrary unknown state and mixture alignment, either of the two main methods for iterative estimation of HMM parameters, namely the segmental and forward-backward methods of Chapter 2 can be used to calculate the sufficient statistics for the approximation of the posterior parameters. For computational efficiency we select to use the *segmental adaptation* method to locally maximise $f(\mathbf{X}, \mathbf{q}|\boldsymbol{\lambda})g(\boldsymbol{\lambda})$, but we could also have used the more general solution offered by the *forward-backward* adaptation algorithm to locally maximise $f(\mathbf{X}|\boldsymbol{\lambda})g(\boldsymbol{\lambda})$, as was assumed in the derivation of the maximisation of the auxiliary function $R(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$.

In the implementation of the segmental Bayesian adaptation algorithm, the Viterbi algorithm is used to compute the state alignment $(\bar{\mathbf{q}}(n))$ in iteration n of the observations with the current model estimate:

$$\bar{\mathbf{q}}(n) = \arg\max_{\mathbf{q}} f(\mathbf{X}, \mathbf{q}|\boldsymbol{\lambda}(n)). \tag{3.88}$$

The state alignment in iteration n is used, in turn, to estimate the statistics of Equations 3.61-3.64 and the MAP parameters of iteration $n+1$, as described by

$$\boldsymbol{\lambda}(n+1) = \arg\max_{\boldsymbol{\lambda}} f(\mathbf{X}, \bar{\mathbf{q}}(n)|\boldsymbol{\lambda})g(\boldsymbol{\lambda}), \tag{3.89}$$

where $\boldsymbol{\lambda}(0)$ is initialised to the model estimate when no data is observed, which is usually just the model that was used to seed the prior distribution. When applying the segmental Bayesian algorithm for speaker adaptation, use of only a single iteration may suffice, but we expect that for cross-language adaptation a relatively large number of iterations may be necessary, especially if there is a large mismatch between source and target data distributions. When a large number of iterations take place, unobserved model mixtures (mixtures with very low output probabilities) may converge to feature space regions where they contribute to the *a posteriori* probability function and are therefore adapted. We now turn our attention to the determination of the parameters of the prior distribution.

3.2.5 Prior density estimation

Section 3.2.2 discussed a method (from [24]) for prior density estimation for the mean and variance (or precision) parameters of a univariate Gaussian (Equations 3.31-3.33) and a multivariate Gaussian (Equations 3.45-3.47). The discussion centred around a way of using speaker independent Gaussian mixture models to estimate a normal-Gamma (univariate) and normal-Wishart (multivariate) prior distribution for the mean and variance of observations from the Gaussian observation distribution. One may apply this approach directly for Gaussian mixture observation distributions, but it would imply use of an identical prior distribution for every mixture. Another way of estimating parameters for the prior distribution is to set the prior mode equal to the parameters of a given HMM [61], typically an HMM trained on speaker independent data. The prior distribution, however, contains five parameters (v_{ik}, $\mathbf{m}_{ik}$, w_{ik}, α_{ik} and $\boldsymbol{\Upsilon}_{ik}$) for each mixture, while only three parameters ($\tilde{c}_{ik}$, $\tilde{\mathbf{m}}_{ik}$ and $\tilde{\mathbf{r}}_{ik}$) are associated with each mixture of the speaker independent HMM, essentially implying that we are unable to estimate the variance of the prior mean and, similar to the other approaches, that our estimate of the mean and the variance of the prior precision are dependent.

An elegant solution [61] can be found by limiting the family of the prior distribution to that of the kernel density of the complete-data likelihood. The prior family is expressed as a joint Dirichlet-normal-Wishart distribution (Equation 3.58) while the complete data likelihood function (Equation 3.65) is a *dependent* Dirichlet-normal-Wishart function. Element-wise comparison of the two equations delivers the following correspondence

$$\eta_{ij} - 1 \leftrightarrow \xi_{ij} \tag{3.90}$$

$$v_{ik} - 1 \leftrightarrow \gamma_{ik} \tag{3.91}$$

$$\alpha_{ik} - D \leftrightarrow \gamma_{ik} \tag{3.92}$$

$$w_{ik} \leftrightarrow \gamma_{ik}. \tag{3.93}$$

By selecting to retain η_{ij} and w_{ik}, the other two parameters of the prior distribution, namely v_{ik} and α_{ik}, can be written in terms of w_{ik} by

$$v_{ik} = w_{ik} + 1 \tag{3.94}$$

$$\alpha_{ik} = w_{ik} + D. \tag{3.95}$$

This reduction of the prior renders it of the same distribution family as the complete data likelihood function and the remaining parameters can then be estimated directly from the seed model parameters by using the prior transition probability

$$\eta_{ij} = \tilde{a}_{ij}, \tag{3.96}$$

the prior mixture weight value

$$w_{ik} = \tilde{c}_{ik}, \tag{3.97}$$

and similar to Equations 3.45 and 3.47:

$$\mathbf{m}_{ik} = \tilde{\mathbf{m}}_{ik}, \tag{3.98}$$

$$\frac{\boldsymbol{\Upsilon}_{ik}}{w_{ik} + D} = \tilde{\boldsymbol{\Sigma}}_{ik}. \tag{3.99}$$

To evaluate the meaningfulness of these choices we rewrite the affected posterior parameter estimates. The parameter reductions of Equations 3.94 and 3.95 are applied, as well as the choice of prior seed values (Equations 3.96-3.99) for the Gaussian mean estimates (from Equation 3.74)

$$\boldsymbol{\mu}_{ik\,\mathrm{MAP}} = \boldsymbol{\mu}_{ik\,\mathrm{MSE}} = \frac{\tilde{c}_{ik}\tilde{\mathbf{m}}_{ik} + \gamma_{ik}\bar{\mathbf{x}}_{ik}}{\tilde{c}_{ik} + \gamma_{ik}}, \tag{3.100}$$

for the MAP parameters (from Equations 3.75-3.77)

$$a_{ij\,\mathrm{MAP}} = \frac{\tilde{a}_{ij} + \xi_{ij} - 1}{\sum_{l=1}^{K}(\tilde{a}_{il} + \xi_{il} - 1)} \tag{3.101}$$

$$c_{ik\,\mathrm{MAP}} = \frac{\tilde{c}_{ik} + \gamma_{ik}}{\sum_{l=1}^{K}(\tilde{c}_{il} + \gamma_{il})} \tag{3.102}$$

$$\mathrm{R}^{-1}_{ik\,\mathrm{MAP}} = \frac{(\tilde{c}_{ik} + D)\tilde{\boldsymbol{\Sigma}}_{ik} + \sum_{t=1}^{T}\gamma_{ik}(t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)^T + \tilde{c}_{ik}(\hat{\mathbf{m}}_{ik} - \tilde{\mathbf{m}}_{ik})(\hat{\mathbf{m}}_{ik} - \tilde{\mathbf{m}}_{ik})^T}{\tilde{c}_{ik} + \gamma_{ik}}, \tag{3.103}$$

and for the MSE parameters (from Equations 3.78-3.80)

$$a_{ij\,\mathrm{MSE}} = \frac{\tilde{a}_{ij} + \xi_{ij}}{\sum_{l=1}^{K}(\tilde{a}_{il} + \xi_{il})} \tag{3.104}$$

$$c_{ik\,\mathrm{MSE}} = \frac{\tilde{c}_{ik} + \gamma_{ik} + 1}{\sum_{l=1}^{K}(\tilde{c}_{il} + \gamma_{il} + 1)} \tag{3.105}$$

$$\mathrm{R}^{-1}_{ik\,\mathrm{MSE}} = \frac{(\tilde{c}_{ik} + D)\tilde{\boldsymbol{\Sigma}}_{ik} + \sum_{t=1}^{T}\gamma_{ik}(t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)^T + \tilde{c}_{ik}(\hat{\mathbf{m}}_{ik} - \tilde{\mathbf{m}}_{ik})(\hat{\mathbf{m}}_{ik} - \tilde{\mathbf{m}}_{ik})^T}{\tilde{c}_{ik} + \gamma_{ik} + D}. \tag{3.106}$$

Although the parameter reduction has produced a MAP estimate that is defined for all valid prior parameter values, an artifact of the seeding is that the MAP variance estimate

(Equation 3.103) is not equal to the prior variance when no observations are available. We propose to remedy this by seeding the mode Υ_{ik}/w_{ik} (in place of the mean as in Equation 3.99) of the variance prior. This results in an elegant formula for the MAP variance estimate which is independent of the feature dimension D and is given by

$$\mathrm{R}^{-1}_{ik\,\mathrm{MAP}} = \frac{\tilde{c}_{ik}\tilde{\Sigma}_{ik} + \sum_{t=1}^{T}\gamma_{ik}(t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)(\hat{\mathbf{m}}_{ik} - \mathbf{x}_t)^T + \tilde{c}_{ik}(\hat{\mathbf{m}}_{ik} - \tilde{\mathbf{m}}_{ik})(\hat{\mathbf{m}}_{ik} - \tilde{\mathbf{m}}_{ik})^T}{\tilde{c}_{ik} + \gamma_{ik}}. \tag{3.107}$$

Examination of the posterior mean estimate (Equation 3.100) and the posterior variance estimates (Equations 3.103, 3.106 and 3.107) shows that $\tilde{c}_{ik}$ can be interpreted as a prior weighting factor associated with the kth mixture of state i. When $\tilde{c}_{ik}$ is large the mean and variance prior densities are sharply peaked around the values used for seeding the prior and less adaptation occurs than when $\tilde{c}_{ik}$ is small. This choice implies that we expect the weight associated with a mixture to express the confidence associated with the mixture, which makes intuitive sense. While the choice of seed value (Equation 3.97) makes sense, it leads to prior weight values in the range $[0, 1]$, which in Equations 3.100-3.107 implies that the weight associated with the prior distribution is less than that associated with a single observation frame. The prior weight $\tilde{c}_{ik}$ assigned to the prior distribution for each mixture is therefore multiplied by a global prior weight scaling factor ϖ. Unfortunately, the optimal value of ϖ cannot be determined easily from a small amount of training data, since it needs to be evaluated on independent data (target data not used for adaptation). We do not follow a cross-validation approach, but in experiments (Chapters 6 and 7) rather explicitly show the effect of the prior weight scaling factor on recognition performance. More detailed aspects of the application of Bayesian techniques for cross-language adaptation are covered in Chapter 5.

The Bayesian framework for estimation that we discussed in this section focussed heavily on the use of existing knowledge when facing the design of a new system, or when changing a system based on new observations. In the next section we discuss methods that attempt to exploit correlation between parameters when changing a current model to better reflect the characteristics of a new sample.

3.3 Transformation-based adaptation

Transformation-based techniques estimate a transformation of model parameters using a limited amount of observation data. A linear transformation of model parameters is usually computed and applied to an existing model for the model to better reflect the characteristics of the observations. Non-linear transformations, such as those implemented with multi-layer perceptrons (MLPs), have also been applied for the transformation of model parameters.

The motivation for using transformation-based adaptation, versus say Bayesian adaptation, is that if the changes in the observation characteristics can be approximated well enough by a simple parameter transformation, then only the parameters of the transformation have to be estimated which will typically be far fewer than those of the model being transformed. Parameters of unobserved distributions are adapted by implementing the same transformation for all the parameters or for groups of parameters and rapid adaptation can thus be achieved on little target data. When a reasonably large amount of adaptation data is available, such as for our application of cross-language adaptation, transformation-based adaptation does not automatically guarantee asymptotic behaviour with respect to a language dependent system.

The transformation approach can be applied at the feature or at the model level. When applied at the feature level, it is referred to as feature space adaptation or *spectral* transformation [54]. Feature space transformation can be implemented as part of the pre-processing stage of a system, transforming incoming speech from a new speaker to better match that of a reference speaker or speakers - thus normalising the speech of the new speaker with respect to the reference. Feature space transformation can also be used to perform compensation for spectral mismatch of recording conditions and channel effects between training and testing environments. When the transformation is implemented on cepstral features, as is usually done, a linear process in the frequency domain can be implemented (or counteracted) with a simple offset in the cepstral domain. Frequency warping or other non-linear frequency domain processes can be approximately implemented or counteracted with full transformations of the cepstral features. Feature space transformations have been used to perform phone-specific transformations to some degree by estimating several transformations across the entire feature space and implementing transformation of specific features using fuzzy class membership rules [64].

Model space transformations are generally accepted [65] to deliver better performance than feature space transformations since different transformations can be estimated for different phonetic groupings and also other parameters, such as Gaussian variance, can be transformed separately from the Gaussian mean parameters. Model space transformations can make better use of available data than feature space transformations by estimating few transformations when little adaptation data is available and estimating many transformations when a large amount of adaptation data is available.

An application of feature space transformation that is promising is the use of transformation to normalise speech from the training speakers with respect to some reference and then to retrain the models [66]. This approach is related to data augmentation, which transforms speech data from speakers close to the target speaker and subsequently performs retraining of models [67]. We discuss these methods in the context of using them for cross-language data augmentation, i.e. performing cross-language transformation and subsequent retraining. We now proceed to discuss the method most commonly used for transformation-based adaptation namely the linear transform.

3.3.1 Linear transformation of the Gaussian mean

Linear transformation of the Gaussian mean model parameters using target data attempts to improve the match between the model and target data through correlation between the distribution the model represents and the distribution of the target data. The Gaussian mean parameters are usually transformed since they specify positions in feature space that represent nuclei of the model distribution and can thus be directly compared with target data distributions. Transformation-based adaptation is usually performed with a linear transformation because it is well understood and leads to simple implementation. When a linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$ from parameters or observations $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ to parameters or observations $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_T\}$ is estimated, the squared error is given by

$$E = \sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{W}\mathbf{x}_t)^T (\mathbf{y}_t - \mathbf{W}\mathbf{x}_t) = \mathrm{tr}\big[(\mathbf{Y} - \mathbf{W}\mathbf{X})(\mathbf{Y} - \mathbf{W}\mathbf{X})^T\big] \tag{3.108}$$

and the minimum squared error (MSE) solution is found using the pseudo inverse form for the transformation matrix

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{3.109}$$

which is given in transpose form by

$$\mathbf{W}^T = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}^T \tag{3.110}$$

and for the transpose of row l of $\mathbf{W}$ by

$$\mathbf{w}_l^T = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}_l^T \tag{3.111}$$

for comparison with later equations, where $\mathbf{y}_l$ is the lth row of $\mathbf{Y}$ (not to be confused with a column $\mathbf{y}_t$ of $\mathbf{Y}$). Least squares linear regression has been used for estimation of feature space transformations [54], as well as for model adaptation by estimating transformations of parameters of CDHMM [68]. Cox [69] also used regression to estimate linear transformation of individual sound classes, exploiting correlation between classes. The most popular approach for estimating linear transformations is related to the least squares estimate and is discussed next.

Maximum likelihood linear regression

A maximum likelihood-based approach for linear transformation, termed maximum likelihood linear regression (MLLR), was proposed by Leggetter and Woodland [63, 27]. In the Gaussian mixture density HMM framework, MLLR estimates the linear transformation of the Gaussian means

$$\hat{\boldsymbol{\mu}}_{jk} = \mathbf{W}\boldsymbol{\mu}_{jk} \tag{3.112}$$

that maximises the likelihood $f(\mathbf{X}|\hat{\boldsymbol{\lambda}})$ of the observations given the transformed model

$$\hat{\boldsymbol{\lambda}} = \{A, (c_{jk}, \mathbf{W}\boldsymbol{\mu}_{jk}, \mathbf{R}_{jk})_{j=1;k=1}^{N\;\;\;K}\}. \tag{3.113}$$

The transformation matrices can be found by maximising the auxiliary function

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \sum_{\mathbf{q}} f(\mathbf{X}, \mathbf{q}|\boldsymbol{\lambda}) \log(f(\mathbf{X}, \mathbf{q}|\hat{\boldsymbol{\lambda}})) \tag{3.114}$$

with respect to $\mathbf{W}$ where $\hat{\boldsymbol{\lambda}}$ is the transformed model of Equation 3.113. Using the transformed model in the expansion of the auxiliary equation (Equation 3.86) delivers

$$\begin{aligned} Q_{\mathcal{N}}[\boldsymbol{\lambda}, \mathbf{W}\boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}] &\propto \gamma_{ik} \log\left[|\mathbf{R}_{ik}|^{1/2} e^{-(1/2)[\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik}) + (\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})^T \mathbf{R}_{ik}(\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})]}\right] \\ &\propto \gamma_{ik}\left[\frac{1}{2}\log|\mathbf{R}_{ik}| - \frac{1}{2}\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik}) - \frac{1}{2}(\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})^T \mathbf{R}_{ik}(\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})\right] \end{aligned} \tag{3.115}$$

To maximise $Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$, its derivative w.r.t. $\mathbf{W}$ is computed and equated to zero, i.e.

$$\begin{aligned} \frac{dQ(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})}{d\mathbf{W}} &= \frac{d}{d\mathbf{W}} \sum_{i=1}^{N}\sum_{k=1}^{K} Q_{\mathcal{N}}[\boldsymbol{\lambda}, \mathbf{W}\boldsymbol{\mu}_{ik}, \mathbf{R}_{ik}] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}\frac{d}{d\mathbf{W}}\left[\frac{1}{2}\log|\mathbf{R}_{ik}| - \frac{1}{2}\mathrm{tr}(\mathbf{S}_{ik}\mathbf{R}_{ik}) - \frac{1}{2}(\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})^T \mathbf{R}_{ik}(\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}\mathbf{R}_{ik}(\mathbf{W}\boldsymbol{\mu}_{ik} - \bar{\mathbf{x}}_{ik})\boldsymbol{\mu}_{ik}^T = 0, \end{aligned} \tag{3.116}$$

which delivers

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}\mathbf{R}_{ik}\mathbf{W}\boldsymbol{\mu}_{ik}\boldsymbol{\mu}_{ik}^{T}=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}\mathbf{R}_{ik}\bar{\mathbf{x}}_{ik}\boldsymbol{\mu}_{ik}^{T}. \tag{3.117}$$

For a diagonal covariance matrix (and thus diagonal precision also), the lth row on both sides of Equation 3.117 is given by

$$\mathbf{w}_l\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}r_{ikl}\boldsymbol{\mu}_{ik}\boldsymbol{\mu}_{ik}^{T}=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}r_{ikl}\bar{x}_{ikl}\boldsymbol{\mu}_{ik}^{T} \tag{3.118}$$

and we therefore find that the maximum likelihood estimate of the mean transformation matrix $\mathbf{W}$ can be expressed in a much simpler format than in the original publications [63, 27] by the expression

$$\mathbf{w}_l^T=\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}r_{ikl}\boldsymbol{\mu}_{ik}\boldsymbol{\mu}_{ik}^{T}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}r_{ikl}\bar{x}_{ikl}\boldsymbol{\mu}_{ik}\right] \tag{3.119}$$

for the lth row of $\mathbf{W}$. Equation 3.119 also clearly shows the relationship between the MLLR estimate and the MSE transformation estimate of Equation 3.111. The MLLR estimate is simply an MSE estimate that weights the contribution of each mixture component to the pseudo inverse with the amount of data associated with the mixture (γ_{ik}) multiplied by the precision of the mixture component separately for each feature dimension (r_{ikl}). The MLLR estimate can be written in the exact form of an MSE estimate (Equation 3.109) with $D \times KN$ dimensional matrices $\mathbf{X}$ and $\mathbf{Y}$, with the $(i \times k)$th column of $\mathbf{X}$ given by $(\gamma_{ik}r_{ikl})^{1/2}\boldsymbol{\mu}_{ik}$ and the $(i \times k)$th column of $\mathbf{Y}$ given by $(\gamma_{ik}r_{ikl})^{1/2}\bar{\mathbf{x}}_{ik}$.

If an offset term is added to the multiplicative term to make the transformation more general, the transformation of the kth Gaussian mean in the ith state can be expressed as the transformation

$$\hat{\boldsymbol{\mu}}_{ik}=\mathbf{W}\boldsymbol{v}_{ik} \tag{3.120}$$

of the extended mean vector $\boldsymbol{v}_{ik}=[\mu_{ik1},...,\mu_{ikD},1]^T$ by a $D \times (D+1)$ matrix $\mathbf{W}$. Closely following Equation 3.119, but sharing the transformation $\mathbf{W}_s$ across the K components of an arbitrary set of R states $\{s_1, ..., s_R\}$, the lth row of the transformation matrix $\mathbf{W}$ is given by

$$\mathbf{w}_{s_l}^T = \left[\sum_{r=1}^{R}\sum_{k=1}^{K} \gamma_{s_r k} r_{s_r k l} \boldsymbol{v}_{s_r k} \boldsymbol{v}_{s_r k}^T\right]^{-1} \left[\sum_{r=1}^{R}\sum_{k=1}^{K} \gamma_{s_r k} r_{s_r k l} \bar{x}_{s_r k l} \boldsymbol{v}_{s_r k}\right]. \tag{3.121}$$

Usually transformation-based adaptation is performed because there is too little data for re-estimation of parameters, thus necessitating the tieing of transformations across multiple states of multiple HMMs to obtain a robust estimates of the transformation. Equation 3.121 does not explicitly show tieing between states of different HMMs, but the group of states (called a regression class) tied in the transformation may be associated arbitrarily with different HMMs. The implementation of tieing used in this thesis groups together HMMs according to phonetic categories. A clustering algorithm may also be used to group together mixtures that are close to each other in feature space according to some metric [70]. When using phonetic groupings, the assumption is that sounds from the same categories undergo similar transforms, while the clustering approach assumes that mixtures that are closely located in feature space undergo similar transforms.

Inspection of Equation 3.121 reveals that the rank of the matrix that is inverted is less than or equal to the number of observed independent Gaussian mean vectors (at most RK). Since the matrix contains $D+1$ rows and columns, it follows that it is necessary that $RK \geq D+1$ for a non-singular matrix and thus for a unique solution to be found for $\mathbf{w}_{s_l}$. Writing Equation 3.121 in the familiar $\mathbf{Ax} = \mathbf{b}$ notation, it is apparent that $\mathbf{b}$ is within the column-space of $\mathbf{A}$ when the coefficients of the summation are not degenerate and therefore a solution exists, irrespective of the degree of mixture tieing. However, when $RK < D+1$, the solution is not unique and a range of values for $\mathbf{W}$ exist that exactly reproduce the maximum likelihood values for all the tied Gaussian means, i.e. the values obtained if mean-only training (re-estimation) is done on the adaptation data. The use of singular value decomposition is preferred in general for the solution of the least squares problem and may be used to determine a suitable transformation matrix $\mathbf{W}$, irrespective of whether the matrix **A** is singular or not. The transformation may, however, not be very useful if it merely implements re-estimation of the means.

When little target data is available, re-estimation is particularly troublesome as the re-estimated means are likely to be inaccurate. In this case a value of $RK >> D+1$ is desired to tie the transformation across a large number of mixtures for accuracy. However, if the amount of available target data increases, less tieing, i.e. more regression classes and thus a smaller RK is desired so that transformations may group together more closely related mixtures. In the event of a very large amount of data being available, $RK \leq D+1$ (effectively re-estimation) does not present a problem and may even be desirable because accurate estimates can be made on the target data alone. This is a very important point to make since it indicates that MLLR can exhibit asymptotic behaviour (in terms of mean estimation) with respect to a system trained on target data only, if the number of regression classes is allowed to increase in relation to the amount of target data available.

Discussion of this aspect in the original MLLR paper [27] attributes poor performance in the extreme case of calculating a transformation of few tied mixtures using little data, to the accumulated matrices being close to singular and (matrix inversion) therefore causing computational errors. We feel that this is not the true reason for poor performance in the case of little data and few tied mixtures (many regression classes). Rather, as mentioned, the reason is that re-estimation on small amounts of target data is undesirable and therefore the inter-dependencies between more parameters should be shared in the transformation. Inspection of the condition of the matrices (from Equation 3.121) calculated in experiments in Chapters 6 and 7 also reveals that numerical accuracy is not of concern - also substantiated by the fact that the same results are obtained with Gauss-Jordan elimination (with full pivoting) than with a singular value decomposition-based approach.

Implementation of adaptation procedure

An iterative procedure is typically used to estimate the transformations, consisting of the following steps:

1. initialising current model estimates to trained source models,
2. computing sufficient statistics (Equations 2.12, 2.13 and 3.61-3.64) from target data using current model estimates and either Viterbi-alignment or forward-backward approaches,
3. computing the transformation for each regression class (Equation 3.121),
4. updating current model estimates (Equation 3.112) and
5. repeating the process from step 2 for a limited number of iterations or until convergence occurs.

The procedure usually converges within only a few iterations, but more iterations may be needed if the original source models match very poorly with the target data, which may be the case in particular for cross-language model transformation.

3.3.2 Variance transformation

A method for the transformation of both the Gaussian mean and variance parameters that is closely related to MLLR was suggested by Digalakis *et al.* [71]. The method computes the linear transformation of both Gaussian mean and variance parameters through the estimation of a transformation matrix $\mathbf{W}$ and an offset vector $\mathbf{b}$, yielding transformed Gaussian mean

$$\hat{\boldsymbol{\mu}}_{jk} = \mathbf{W}\boldsymbol{\mu}_{jk} + \mathbf{b} \tag{3.122}$$

and variance values

$$\hat{\Sigma}_{jk} = \mathbf{W}\Sigma_{jk}\mathbf{W}^T. \tag{3.123}$$

Unfortunately a closed form solution exists only for diagonal transformation matrices and therefore the transformation for each feature dimension is computed separately. The method has been found [65] not to perform as well as the standard MLLR approach, even though it also adapts the variance parameters, since it does not make use of dependencies between different feature dimensions. For this reason we did not pursue it further.

Maximum likelihood variance transformation

Another method for transforming both Gaussian mean and variance parameters, based on the extension of the MLLR adaptation framework, was proposed by Gales and Woodland [72]. Unlike the approach suggested by Digalakis *et al.* [71], the method optimises the mean and variance parameters in separate iterations, termed unconstrained transformation, thereby allowing a closed form solution for the ML variance transform estimate to be found. The Gaussian mean parameters are transformed in a first step using the standard MLLR approach discussed in the previous section (Equation 3.121). The Gaussian variance parameters are updated in a second step through

$$\hat{\Sigma}_{jk} = \mathbf{B}_{jk}^T\mathbf{H}\mathbf{B}_{jk}, \tag{3.124}$$

where $\mathbf{H}$ is the transformation to be estimated and $\mathbf{B}_{jk}$ is the inverse of the Choleski factor ($\mathbf{C}_{jk}$) of Σ_{jk}^{-1}, i.e.

$$\mathbf{B}_{jk} = \mathbf{C}_{jk}^{-1} \tag{3.125}$$

where

$$\boldsymbol{\Sigma}_{jk}^{-1} = \mathbf{C}_{jk}\mathbf{C}_{jk}^{T}. \tag{3.126}$$

The updated variance model $\bar{\boldsymbol{\lambda}}$ is given by

$$\bar{\boldsymbol{\lambda}} = \{\mathbf{A}, (c_{jk}, \hat{\boldsymbol{\mu}}_{jk}, \mathbf{B}_{jk}^{T}\mathbf{H}\mathbf{B}_{jk})_{j=1}^{N}{}_{;k=1}^{K}\} \tag{3.127}$$

where $\hat{\boldsymbol{\mu}}_{jk}$ is the MLLR updated Gaussian mean estimate.

Similar to the MLLR derivation, the transformation matrix $\mathbf{H}$ can be found by performing the derivative of the auxiliary function $Q(\hat{\boldsymbol{\lambda}}, \bar{\boldsymbol{\lambda}})$ (where $\hat{\boldsymbol{\lambda}}$ represents the MLLR updated mean model obtained using Equation 3.121 and $\bar{\boldsymbol{\lambda}}$ the MLLR updated mean and variance model) with respect to $\mathbf{H}$ and finding the root of the equation. For a transformation H_s, shared by the K components of a set of R states $\{s_1, ..., s_R\}$, each associated with observation sequences of length T_{s_r}, the estimation of the tied variance transformation can be represented by [72]

$$\mathbf{H}_s = \frac{\sum_{r=1}^{R}\sum_{k=1}^{K}\left\{\mathbf{C}_{s_r k}^{T}\left[\sum_{t=1}^{T_{s_r}}\gamma_{s_r k}(t)(\mathbf{x}_{s_r t} - \hat{\boldsymbol{\mu}}_{s_r k})(\mathbf{x}_{s_r t} - \hat{\boldsymbol{\mu}}_{s_r k})^{T}\right]\mathbf{C}_{s_r k}^{T}\right\}}{\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{t=1}^{T_{s_r}}\gamma_{s_r k}(t)} \tag{3.128}$$

where $\hat{\boldsymbol{\mu}}_{s_r k}$ is the MLLR updated Gaussian mean estimate and $\mathbf{C}_{s_r k}$ is given by Equation 3.126. The estimate of $\mathbf{H}_s$ in Equation 3.128 results in a full transformed covariance matrix. Full covariance matrices, however, are rarely used in speech recognition systems due to their greatly increased computational requirements. For diagonal covariance, which we also use, the diagonal entries of $\hat{\boldsymbol{\Sigma}}_{jk}$ are only affected by the diagonal entries of $\mathbf{H}$. The results is thus a diagonal transformation of variance - which does not take dependencies between the feature dimensions into account. In experiments, Gales and Woodland [72] reported an additional decrease in word error rate (WER) of 2% for speaker adaptation by using this mean and variance adaptation approach versus only MLLR mean adaptation, which by itself achieved 13% decrease in WER. Results [72] for noise and channel compensation produced greater increases due to variance adaptation (7% reduction in WER).

For cross-language adaptation, adaptation of the variance components may result in larger performance gains than for speaker adaptation, but may require a more complex approach than diagonal transformation. Recently, Gales [73] proposed a method for unconstrained full variance transformation which uses an iterative estimation algorithm to solve for the transformation. We, however, propose and evaluate an alternative approach.

Minimum squared error variance transformation

We propose a method for unconstrained full variance transformation that uses weighted least squares estimation to compute the variance transformation in a single iteration. The standard MLLR algorithm (Equation 3.121) is used to estimate transformed Gaussian mean parameters in a first stage, similar to the approach suggested by Gales & Woodland [72], followed by Gaussian variance transformation in the next stage. Since almost exclusive use is made of diagonal covariance matrices in speech recognition systems, we only consider the transformation of the variance parameter vector $\boldsymbol{\sigma}^2_{s_rk}$ on the diagonal of the covariance matrix Σ_{s_rk}. A full transformation of the variance parameters associated with the K component mixtures of a set of R states $\{s_1, ..., s_R\}$ can be expressed by

$$\hat{\boldsymbol{\sigma}}^2_{s_rk} = \mathbf{W}^*_s \boldsymbol{\sigma}^2_{s_rk} \tag{3.129}$$

where $\mathbf{W}^*_s$ is the (full) shared variance transformation matrix. We consider calculating the maximum likelihood estimate of the variance transformation of Equation 3.129, but find that the estimate can not be written in a closed-form, which reduces the attractiveness of the approach. We therefore consider using least squares estimation for the computation of $\mathbf{W}^*_s$. The squared error for the variance transformation of Equation 3.129 can be computed directly from the observation data and is then expressed by

$$E_1 = \sum_{r=1}^{R} \sum_{t=1}^{T_{s_r}} \sum_{k=1}^{K} \gamma_{s_rk}(t) \left[(\mathbf{x}_{s_rt} - \hat{\boldsymbol{\mu}}_{s_rk})^2 - \mathbf{W}^*_s \boldsymbol{\sigma}^2_{s_rk}\right]^T \left[(\mathbf{x}_{s_rt} - \hat{\boldsymbol{\mu}}_{s_rk})^2 - \mathbf{W}^*_s \boldsymbol{\sigma}^2_{s_rk}\right] \tag{3.130}$$

where $\hat{\boldsymbol{\mu}}_{s_r k}$ is an MLLR updated mean value (Equation 3.121) and assuming that the square of a vector implies computing the component-wise square of the vector (in the first term in brackets). Alternatively the squared error can also be expressed in terms of a statistic measuring the expected variance of the observation data by

$$E_2 = \sum_{r=1}^{R}\sum_{k=1}^{K} \gamma_{s_r k}\left[\mathbf{v}_{s_r k} - \mathbf{W}_s^* \boldsymbol{\sigma}^2_{s_r k}\right]^T \left[\mathbf{v}_{s_r k} - \mathbf{W}_s^* \boldsymbol{\sigma}^2_{s_r k}\right], \tag{3.131}$$

where $\mathbf{v}_{s_r k}$ is the target variance (vector) for mixture k of state s_r and is given by

$$\mathbf{v}_{s_r k} = \frac{\sum_{t=1}^{T_{s_r}} \gamma_{s_r k}(t)(\mathbf{x}_{s_r t} - \hat{\boldsymbol{\mu}}_{s_r k})^2}{\gamma_{s_r k}}. \tag{3.132}$$

We prefer to use Equation 3.131 because Equation 3.130 computes the fourth power of the distance between each observation and the transformed mean value, leading to very large estimates of the variance, while Equation 3.131 uses the average variance as computed in Equation 3.132. There are still, however, fundamental problems with the use of the variance transformation of Equation 3.129 as optimised using Equation 3.131 since:

- the constraint $\hat{\sigma}^2_{jkl} > 0$ is not guaranteed and
- the least squares error function measures an additive error and not a relative error, thereby biasing the transformation to decrease the error produced by large variance values and causing large relative errors for small variance values.

The transformed variance values can be forced to be valid by applying a variance floor, such as described in Section 3.2.2, but this does not really present a desirable solution. Also, if the magnitude of the variance values grouped together in a transformation have a large range, the relative error may be very large for small variance values, even if the relative error is small for large variance values. A better method for the MSE variance transform that overcomes both these problems is given next.

Minimum squared error log-variance transformation

We propose transforming variance parameters in log-space, thereby maintaining the constraint $\hat{\sigma}^2_{jkl} > 0$ and also minimising the relative error (in place of the absolute error) in the estimation of $\hat{\sigma}^2_{jkl}$. The transformation of the log-variance parameters by transformation matrix $\mathrm{W}^\dagger_s$ is given by

$$\log \hat{\boldsymbol{\sigma}}^2_{s_r k} = \mathrm{W}^\dagger_s \log \boldsymbol{\sigma}^2_{s_r k} \tag{3.133}$$

where $\log \boldsymbol{\sigma}^2_{s_r k}$ is the element-wise logarithm of $\boldsymbol{\sigma}^2_{s_r k}$. The squared error to be minimised can be written as

$$E = \sum_{r=1}^{R} \sum_{k=1}^{K} \gamma_{s_r k} \Big[\log \mathbf{v}_{s_r k} - \mathrm{W}^\dagger_s \log \boldsymbol{\sigma}^2_{s_r k} \Big]^T \Big[\log \mathbf{v}_{s_r k} - \mathrm{W}^\dagger_s \log \boldsymbol{\sigma}^2_{s_r k} \Big], \tag{3.134}$$

where the target variance $\mathbf{v}_{s_r k}$ is given by Equation 3.132. By writing the squared error in the following format

$$E = \sum_{r=1}^{R} \sum_{k=1}^{K} \gamma_{s_r k} \Big[\log \frac{\mathbf{v}_{s_r k}}{\hat{\boldsymbol{\sigma}}^2_{s_r k}} \Big]^T \Big[\log \frac{\mathbf{v}_{s_r k}}{\hat{\boldsymbol{\sigma}}^2_{s_r k}} \Big], \tag{3.135}$$

it is evident that the log-variance transform minimises the *relative* error between the transformed variance $\hat{\boldsymbol{\sigma}}^2_{s_r k}$ and the target variance $\mathbf{v}_{s_r k}$ and is therefore not as sensitive to the relative magnitudes of the variance components as the direct variance transformation.

Finally, the least squares estimate for the log-variance transformation matrix is given in pseudo inverse form solution (as in Equation 3.109):

$$\mathrm{W}^\dagger_s = \Big[\sum_{r=1}^{R} \sum_{k=1}^{K} \gamma_{s_r k} \log \mathbf{v}_{s_r k} \log \boldsymbol{\sigma}^{2\,T}_{s_r k} \Big] \Big[\sum_{r=1}^{R} \sum_{k=1}^{K} \gamma_{s_r k} \log \boldsymbol{\sigma}^2_{s_r k} \log \boldsymbol{\sigma}^{2\,T}_{s_r k} \Big]^{-1}. \tag{3.136}$$

We note that the same discussion that applied to the MLLR estimation equation (Equation 3.121 in Section 3.3.1) applies here with respect to the number of transformed mixtures and the dimension of the transformation. When equal or fewer mixtures than the dimension of the transformation are used, exact re-estimation of the variance values is the result. This, however, implies that inversion of the right-hand-side of Equation 3.136 is not attempted, but that the solution is found through e.g. a singular value decomposition-based approach. The more mixtures are grouped together in a transformation, the more robust, yet less accurate, the transformation becomes. When little data is available, few transformations should be calculated since direct estimation of the variance is problematic on little data.

This concludes our discussion of linear transformation-based adaptation. For speaker adaptation mean-only transformations are usually used, but we have covered variance adaptation in depth since it is important for cross-language adaptation. We have omitted discussion of the adaptation of mixture weight and transition probability parameters because it is inappropriate to apply transformation-based adaptation to them. For cross-language purposes, adaptation of mixture weight and transition probability parameters may be warranted. Other forms of adaptation as in Section 3.2 or even re-estimation may then be used on these parameters as they require far smaller amounts of data to estimate reliably than the Gaussian mean and variance parameters. We now proceed to discuss the application of *non-linear* transformation methods for adaptation.

3.3.3 Non-linear transformation adaptation

Non-linear transformation presents a more powerful paradigm than linear transformation, but present serious challenges in finding a suitable functional form for the transformation and also in optimising the parameters of the transform. As was noted in the previous section on linear transformation, only limited amounts of data are usually available. Relatively simple and well understood estimation techniques such as linear regression are able to use data relatively efficiently, while for the non-linear transformation approach gradient-based techniques must generally be used, which may not use limited data as efficiently.

Non-linear transformation of acoustic parameters has been performed for speaker adaptation using multi-layer perceptrons (MLPs) by Abrash *et al.* [74]. Gaussian mean components of a speaker independent model were adapted on speech from non-native American English speakers. A single non-linear (sigmoidal output function) hidden layer was used for the MLP. A linear transformation was used in parallel with the MLP, effectively adding direct connections from the inputs to the linear output neurons. The weights of the MLP were initialised to small random values, and the linear transformation was set to an identity matrix. Training both the linear transform and the MLP with gradient descent to maximise the observation data likelihood did not achieve the peak performance achieved with an MLLR-estimated linear transform when many transformation classes were allowed. However, when the linear transform was initialised with the MLLR estimate, a modest improvement on standard MLLR was achieved by applying gradient descent to both the linear transform and the MLP.

Choi and King [54] compared the performance of using an MLP with using linear transformations for speaker adaptation and found that the linear transformation delivered significantly better performance. The two studies thus indicate that, using current techniques, it may be difficult for non-linear transformations to improve on the performance of multiple linear transformations. For these above reasons, we restrict our further experimental investigations to linear transforms.

3.3.4 Transformation for normalisation before training

The use of transformations as a pre-processing stage for the normalisation of speech from different speakers before commencing with HMM training has shown promising results. A procedure for *data augmentation* was suggested by Bellegarda *et al.* [75] that performs a least squares linear mapping from the acoustic space of a reference speaker to that of a new speaker. A large amount of data from a reference speaker is transformed to augment the little data from a new speaker to serve for the training of speaker dependent models for the new speaker. Separate linear transformations are estimated for the data associated with groups of elementary speech models. A problem that was reported with the procedure was that too much transformed data from a single reference speaker overwhelmed the small amount of speaker specific data. This situation was improved in subsequent research [76] by

- implementing transformations from multiple reference speakers - thereby reducing the amount of data per reference speaker to approximately the amount of data available for the new speaker,
- implementing a selection procedure to choose reference speakers that are "close" in some sense to the new speaker and
- tieing all the models for a reference speaker in estimating the transformation.

Further improvements to the approach are detailed in [67] and include using MLLR to estimate the transformations and using gender dependent models to estimate alignments instead of using reference speaker specific models.

Procedures related to the previous approach have been used for speaker normalisation before training. Ishii and Tonomura [77] implemented a procedure for speaker normalisation through transformation. The method estimates MLLR mappings from each speaker to the SI model trained on speech from all the speakers, subtracts the MLLR offsets from the speech data and retrains the SI models. This procedure is repeated iteratively and delivers speaker independent models that do not model speaker variation offset and may thus have narrower distributions. For recognition purposes MLLR is used to estimate the transformation (including offset) from the normalised SI models for a new speaker. A closely related approach was also proposed by Nagesha and Gillick [66]. MLLR mappings are also estimated from SI models to each of a set of speakers, but speaker specific data is then transformed using the inverse of the MLLR estimated transformation for each speaker. The reverse transformed data is then used to retrain SI models and the procedure is repeated. Speaker independent models are thus produced that are invariant to linear transformations of the speech from speakers used to train them. Obviously, to accurately recognise speech from a new speaker, a transformation from the normalised models to the new speaker must first be estimated.

The procedures discussed in this section are of interest for cross-language adaptation, because they may be applied to the normalisation of data from multiple databases containing multiple languages to a single target language. Further detail regarding application of the methods is given in Chapter 5. The next section discusses how Bayesian and transformation-based techniques can be combined to improve adaptation performance.

3.4 Combined Bayesian and transformation-based adaptation

Both the Bayesian adaptation approach detailed in Section 3.2 and the transformation-based adaptation approach detailed in Section 3.3 have their respective strengths and weaknesses. Bayesian methods have in particular two perceived advantages over the transformation-based approach namely that with Bayesian methods

- expected performance is asymptotic with respect to a target system - i.e. the performance converges to that of a target dependent system when a large amount of data is available, and
- the degree to which adaptation takes place is automatically controlled by the amount of adaptation data available - i.e. when little data is available little adaptation takes place and as more data is available, more adaptation takes place.

We note that the asymptotic performance property of Bayesian techniques is not true for a transformation-based adaptation approach in general, but as we discussed in Section 3.3.1, may be achieved for transformed values if the number of transformation classes is allowed to increase with the amount of adaptation data. The Gaussian mean values then eventually agree with the target dependent mean values when there are fewer independent Gaussian mixtures per transformation than the dimension of the transformation itself. This argument may be extended to the Gaussian variance values if they are transformed separately from the means. Transformation-based adaptation, on the other hand, has the advantage over Bayesian adaptation that by sharing transformations across groups of phonemes, unobserved parameters can be adapted, leading to more rapid adaptation than is possible with Bayesian adaptation.

Methods to combine Bayesian and transformation-based adaptation are researched in an attempt to retain desired properties from both strategies. We discuss two main techniques that combine Bayesian and transformation-based methods, the first technique focusing on combining rapid transformation-based adaptation with the asymptotic performance property of Bayesian adaptation and the second technique focusing on using Bayesian techniques to control transformation-based adaptation when little data is available.

3.4.1 Linear transformation-MAP

Digalakis and Neumeyer [78] proposed combining Bayesian and transformation-based adaptation in two stages. Constrained transformation-based adaptation [71] is performed in the first stage, using a diagonal transformation to adapt both mean and variance (Equations 3.122 and 3.123) parameters with the adaptation data for a new speaker. This has the advantage of rapidly and accurately compensating for significant bias between source models and target data, such as is exhibited by channel effects. The resulting (speaker adapted) models are used as the starting point for the second adaptation stage, implementing an approximate MAP (AMAP) adaptation algorithm for the Gaussian mean and variance parameters. The Gaussian mean parameters are estimated using an interesting variation to the MAP mean estimate of Equation 3.74 given for mixture k of state i by [78]

$$\boldsymbol{\mu}_{ik\,\mathrm{AMAP}} = \frac{\varpi\gamma_{ik}^{\mathrm{SI}}\boldsymbol{\mu}_{ik}^{\mathrm{SA}} + (1-\varpi)\gamma_{ik}^{\mathrm{SD}}\bar{\mathbf{x}}_{ik}^{\mathrm{SD}}}{\varpi\gamma_{ik}^{\mathrm{SI}} + (1-\varpi)\gamma_{ik}^{\mathrm{SD}}}, \tag{3.137}$$

where $\gamma_{ik}^{\mathrm{SI}}$ and $\gamma_{ik}^{\mathrm{SD}}$ are the mixture occupancy statistics of the speaker independent and speaker dependent data respectively, $\boldsymbol{\mu}_{ik}^{\mathrm{SA}}$ is the (speaker adaptive) transformed mean value, $\bar{\mathbf{x}}_{ik}^{\mathrm{SD}}$ is the sample mean of the speaker dependent data and ϖ is a global adaptation rate factor, in this case taking on values between zero and one. Gaussian variance parameter estimation is computed in a similar fashion to the mean estimation, calculating a linear combination of transformed variance statistics and speaker dependent variance statistics. Digalakis and Neumeyer report [79] that their technique approximately halves the recognition error rate for non-native speakers of American English with only a small amount of adaptation data, approaching the speaker independent accuracy achieved for native speakers.

Comparing the method to MAP mean estimation as derived in Section 3.2.5 (Equation 3.100 in particular), the mixture weight prior seed $\tilde{c}_{ik}$ associated with a mixture in the prior has been replaced by the occupancy statistics for that mixture and the learning factor ϖ is incorporated in a different way. Using occupancy statistics for weighting causes mixtures with high occupancy in the prior to be adapted more slowly than mixtures for which little data was observed when the prior was trained. This may be useful for speaker adaptation, but not necessarily for cross-language adaptation, as the frequency of occurrence of a phoneme in a source language may not give an accurate indication as to its suitability for seeding a prior distribution for target language model estimation. In fact, we found that use of source language occupancy statistics (as in Equation 3.137) delivered poorer performance than use of the mixture weight prior seed $\tilde{c}_{ik}$ (as in Equation 3.100) and therefore in experiments in Chapters 6 and 7 we used the MAP estimates of Section 3.2.5 in implementing MLLR-MAP adaptation.

Thelen *et al.* [80] also implemented a combination of linear transformation and MAP adaptation, similar to that of Digalakis and Neumeyer [78], but using least squares to estimate a full transformation matrix for the Gaussian mean parameters and then used the standard MAP algorithm to derive the final mean values. Better results were obtained with phonetically derived regression classes than with clustering procedures. Interestingly, they reported that their linear regression-MAP algorithm did not achieve asymptotic performance with a speaker dependent system as was planned. They give as a reason the fact that, even with a large amount (several hours) of adaptation data from a single speaker, less than 70% of the transformed densities are observed during MAP adaptation and are thus unadapted. Most parameters of the adapted system are therefore not optimised for the target speaker beyond the initial transformation. The percentage of unobserved densities may have been even higher if the initial transformation had not been performed. This points to a deficiency in the Bayesian estimation framework, namely that when the distribution of the adaptation data differs significantly from the distribution of the data that was trained on, only a fraction of the total parameter set that corresponds to the adaptation data is adapted. For cross-language and cross-database acoustic adaptation we expect that the overlap between source and target feature distributions may be relatively poor, which may negatively influence recognition performance. We therefore evaluate the performance of using MLLR-MAP, showing in Chapter 7 that it leads to improved performance for cross-language, cross-database adaptation.

3.4.2 MAP-MLLR

Chou [81] recently proposed an alternative combination of Bayesian and transformation-based adaptation termed maximum *a posteriori* linear regression (MAPLR). The goal of the method is not to ensure asymptotic performance, but to control the amount of adaptation when little data is available by using prior distributions. It incorporates prior knowledge by biasing the MLLR transformation to more closely match a unity transformation when little adaptation data is available and to more closely match the MLLR estimate when a large amount of adaptation data is available.

MAPLR

MAPLR assumes an elliptic symmetric *a priori* distribution for the transformation matrix. The solution to MAPLR entails diagonalising the matrix inversion of the MLLR estimate

(Equation 3.121) through the addition of a diagonal matrix, i.e.

$$\begin{aligned}\hat{\mathbf{w}}_{s_l}^T &= (\hat{\mathbf{G}}_{sl})^{-1}\hat{\mathbf{z}}_{sl} \\ &= (\mathbf{G}_{sl} + \mathbf{D}_{sl})^{-1}(\mathbf{z}_{sl} + \mathbf{D}_{sl}\tilde{\mathbf{w}}_l^T)\end{aligned} \tag{3.138}$$

where $\mathbf{G}_{sl}$ is equal to the first term in brackets and $\mathbf{z}_{sl}$ the second term in brackets on the right hand side of Equation 3.121, $\mathbf{D}_{sl}$ is the scale factor (acting as a diagonalising term) and $\tilde{\mathbf{w}}_l^T$ is the lth row of the location parameter of the transformation. Choosing the location parameter ($\tilde{\mathbf{W}}$) to be the identity matrix backs off the transformation to an identity transform when there are no observations and is the implementation approach that we use. When a large number of observations are available, the occupancy statistics of $\mathbf{G}_{sl}$ and $\mathbf{z}_{sl}$ dominate the equation, ensuring convergence to the MLLR estimate.

Chou [81] uses a global MLLR transformation to estimate the prior location parameters ($\tilde{\mathbf{w}}_l$), but does not describe how to estimate the scale parameters ($\mathbf{D}_{sl}$). We propose using as diagonalising term the diagonal of $\mathbf{G}_{sl}$ (from Equation 3.121), normalised with respect to the amount of data and multiplied by an overall prior weight scaling factor ϖ, i.e.

$$d_{sli} = \varpi \frac{\sum_{r=1}^{R}\sum_{k=1}^{K}\gamma_{s_r k} r_{s_r kl} v_{s_r ki}^2}{\sum_{r=1}^{R}\sum_{k=1}^{K}\gamma_{s_r k}}, \tag{3.139}$$

where d_{sli} is the ith term on the diagonal of $\mathbf{D}_{sl}$ and $v_{s_r ki}$ is the ith term of the extended mean vector $\boldsymbol{v}_{s_r k}$. The value of ϖ depends on the suitability of the prior distribution and should be determined empirically.

MAP-like log variance transformation

We propose using a similar approach to MAPLR for the diagonalisation of the MSE log-variance transformation. Since our attempts at obtaining an ML estimate for the variance and log-variance transformations did not produce a closed-form solution, MAP estimation is not attempted. We propose simply adding a diagonalising term (scaling parameter)

$\mathbf{D}_s^\dagger$, similar to the scaling parameter in Equation 3.138, to the pseudo inverse solution of Equation 3.136, producing the MAP-like estimate

$$\hat{\mathbf{W}}_s^\dagger = \Big[\sum_{r=1}^{R}\sum_{k=1}^{K} \gamma_{s_r k} \log \mathbf{v}_{s_r k} \log \boldsymbol{\sigma}_{s_r k}^{2}{}^{T} + \mathbf{D}_s^\dagger\Big]\Big[\sum_{r=1}^{R}\sum_{k=1}^{K} \gamma_{s_r k} \log \boldsymbol{\sigma}_{s_r k}^{2} \log \boldsymbol{\sigma}_{s_r k}^{2}{}^{T} + \mathbf{D}_s^\dagger\Big]^{-1}. \tag{3.140}$$

When no data is observed, $\hat{\mathbf{W}}_s^\dagger$ backs off to a unity transformation and when a large amount of data is observed, $\hat{\mathbf{W}}_s^\dagger$ converges to the MSE estimate. We propose calculating the diagonal term $\mathbf{D}_s^\dagger$ in a similar fashion to the MAPLR diagonal term (Equation 3.139), producing the equation

$$d_{si}^\dagger = \varpi \frac{\sum_{r=1}^{R} \sum_{k=1}^{K} \gamma_{s_r k} (\log \sigma_{s_r k i}^2)^2}{\sum_{r=1}^{R} \sum_{k=1}^{K} \gamma_{s_r k}}, \tag{3.141}$$

where $d_{si}^\dagger$ is the ith term on the diagonal of $\mathbf{D}_s^\dagger$ and the overall prior weight scaling factor ϖ is shared with Equation 3.139.

3.4.3 Comparison of MLLR-MAP and MAP-MLLR

Figure 3.1 shows conceptually the difference between the MLLR-MAP and MAP-MLLR approaches. While MAP-MLLR controls the amount of adaptation the transformation can effect, MLLR-MAP uses the MLLR transformed models to seed prior distributions for MAP estimation.

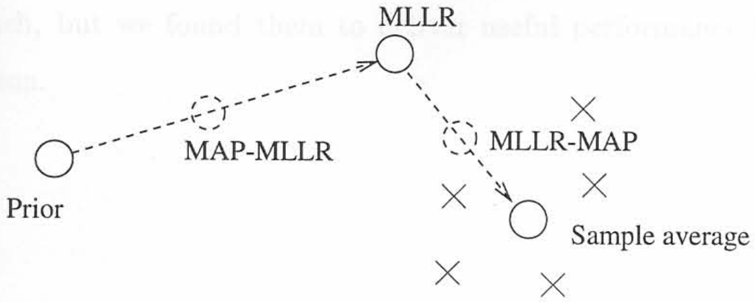

Figure 3.1: Graphical comparison of the working of the MAP-MLLR and MLLR-MAP techniques, showing adaptation of the Gaussian mean

Both techniques may be useful for cross-language adaptation in different ways. MLLR-MAP should provide better asymptotic performance than MLLR adaptation alone and should also improve performance if source and target language models are poorly matched by removing channel effects before commencing with MAP adaptation. MAP-MLLR may in particular improve the robustness of estimates for transformation classes with little data by decreasing the over-fitting effect when complex transformations are estimated from limited data. Both methods are experimented with in Chapter 6 and 7 and are shown to significantly improve performance.

3.5 Discussion

We have discussed the major classes of methods used for speaker adaptation, namely Bayesian and transformation-based methods, as well as combinations of these techniques. A new technique for full transformation of variance parameters in log-space and utilising MAP-like control over adaptation was proposed, specifically with cross-language model adaptation in mind. Some aspects regarding the application of the techniques for cross-language adaptation were mentioned, but will only be discussed in detail in Chapter 5. Experimental comparisons of Bayesian, transformation-based and combined techniques are given in Chapters 6 and 7).

In the next chapter we discuss a third class of methods applicable for acoustic adaptation, namely discriminative training methods. These methods are not generally used for speaker adaptation as such, but we found them to deliver useful performance for cross-language acoustic adaptation.