THE THEORY OF SEGMENTAL HIDDEN MARKOV MODELS

M. J. F. Gales & S. J. Young

CUED/F-INFENG/TR 133

June 1993

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

Email: mjfg@eng.cam.ac.uk

Abstract

The most popular and successful acoustic model for speech recognition is the Hidden Markov Model (HMM). To use HMMs for speech recognition a series of assumptions are made about the waveform, some of which are known to be poor. In particular, the 'Independence Assumption' implies that all observations are only dependent on the state that generated them, not on neighbouring observations. In this paper, a new form of acoustic model is described called the Segmental Hidden Markov Model (SHMM) in which the effect of the 'Independence Assumption' on the observation likelihood is greatly reduced. In the SHMM all observations are assumed to be independent given the state that generated them but additionally they are conditional on the mean of the segment of speech to which they belong. Re-estimation formulae are presented for the training of both single and multiple Gaussian Inter Mixture models and a recognition algorithm is described. Additionally it is shown that the standard HMM, both in the single Gaussian mixture and multiple Gaussian mixtures cases, is just a subset of the SHMM. The new model is shown to provide better recognition performance on a wider set of synthetic data than the standard HMM.

Contents

1 Introduction **4**

2 The Hidden Markov Model **5**
2.1 Basic Theory 5
2.2 Viterbi algorithm 6
2.3 Forward Backward algorithm 6
2.4 Re-Estimation Formulae 7
2.5 Compensating for the Independence Assumption 7
2.5.1 Dynamic Coefficients 7
2.5.2 Explicit Time Correlation Modelling 7
2.5.3 Variable Frame Rate Analysis 8
2.5.4 Stochastic Segment Model 8

3 The Segmental Hidden Markov Model **9**
3.1 Relationship to MAP SHMM 11
3.2 Multiple Gaussian Inter Mixtures 11
3.3 Multiple Gaussian Intra Mixtures 12

4 Relationship to Standard HMMs **12**

5 Re-estimation Formulae **13**
5.1 Re-estimation for $\hat{\mu}_c$ 13
5.2 Re-estimation for $\hat{\Sigma}_c$ 14
5.3 Re-estimation for $\hat{\Sigma}$ 14
5.4 Approximate Solution 14
5.5 Comparison with Standard HMM re-estimation Formulae 15

6 Segmental HMM Estimation Stage **16**
6.1 Viterbi Re-Estimation 16
6.2 Baum-Welch Re-Estimation 17

7 Multiple Intra State SHMMs **17**
7.1 The New Complete Data Set 17
7.2 Multiple Intra States Probability 18
7.3 Estimating the Complete Data Set 18
7.4 Approximation to the Multiple Gaussian Intra Mixtures 20
7.5 Multiple Intra State Parameter Estimation 20

8 Relationship to Bayesian Speaker Adaptation **21**
8.1 Implementation within the SHMM framework 21
8.2 Maximum A-Posterior Estimate of the Mean 21
8.3 Bayesian Approach 23
8.4 Discussion 23

9 Maximisation Stage of the SHMM **23**
9.1 Approximate Solution 23
9.2 True Maximisation 23
9.3 Approximate True Maximisation 23
9.4 Parameter Constraint Implementation 24

10 Software Implementation **24**

11 Synthetic Data **25**
11.1 Training Procedure 25
11.2 Results on Synthetic Data 25

12 Preliminary Results on TIMIT **26**
12.1 Model Training 26
12.2 Recognition Performance 27

13 Conclusions **27**

A Segmental HMM Likelihood Derivation **30**

B Proof of Maximisation for Approximate Solution **31**
B.1 Proof for μ_c 31
B.2 Proof for Σ_c 31
B.3 Proof for Σ 31

1 Introduction

The problem of creating a machine to recognise or respond to speech has been actively studied since 1950. There are many reasons for studying the problem, due to the large number of direct practical applications. These include data retrieval, voice activated data entry, voice activated control, for instance of wheel chairs, and dictation machines. The problems associated with voice-operated systems are described below.

1. Between speaker variability. This may result from physical differences, such as length of vocal tract, sex and age. In addition there are variations in pronunciation due to regional dialects.

2. Within speaker variability. Speakers tend to alter their speech in situations of stress or excitement. There are variations in an individual speaker depending on whether the speech is being read, prompted or is free conversational speech. Also there is a certain amount of stress deliberately added to the speech, particularly when uttering a question.

3. Environmental variations. Speakers will naturally adapt their speech according to the environment, the well known Lombard effect [4]. The observed waveform may also be heavily distorted by high level interfering background noise such as occurs on a factory floor, or in a car. In addition there are problems associated with multipath. Speech sounds very different if spoken in an anechoic chamber compared to a normal room. Moreover, the medium used to record the speech tends to alter the speech.

4. There are also problems associated with fundamental confusability of speech. Given no context, it is hard to differentiate between *youth in Asia* and *Euthanasia*. So context and understanding must also be considered.

Ultimately any voice-operated system must be able to cope with all of the above problems.

This paper is concerned with the task of Automatic Speech Recognition (ASR), however it does not consider all of the issues detailed above, but concentrates on the problem of modelling the speech waveform in clean environments. Hence only the first two issues are addressed, those of between speaker variability and within speaker variability in speech recognition systems. The problem of ASR may be broken down into a series of stages. These are speech acquisition, speech parametrisation, modelling and recognition and finally the understanding stage. Each stage is not independent and tends to impact heavily on the other stages. Only the modelling and recognition stage of the task are examined here, the other aspects will be considered known and fixed. Various approaches to ASR have been and are still being studied. These may be split into four main approaches: template based, knowledge based, stochastic modelling and connectionist. The approach adopted here is a stochastic one. In particular it is related to the most popular and successful stochastic models in general use, the Hidden Markov Model (HMM).

This report is laid out as follows. The next section describes the standard HMM and the assumptions behind its use in speech recognition. In addition some of the methods used to overcome the problems associated with these assumptions are described. The third section introduces the Segmental Hidden Markov Model (SHMM) and describes the various forms it can take. The SHMM is then compared to the standard HMM and all forms of the standard HMM are shown to be subsets of the SHMM. Re-estimation formulae are described in section 5, where it is shown that closed forms for re-estimating SHMMs are only possible if additional assumptions about the model parameters are made. The re-estimation formulae are also compared to those of the standard HMM. In section 6 the two methods of completing the data set are given. An approximation to the multiple intra state model is described in section 7 and the alterations to the training and recognition algorithms are detailed. Bayesian speaker adaptation is described within the framework of the SHMM in section 8. The following section describes how, given the complete data set, the Maximum Likelihood parameters may be found. Section 10 briefly describes the software implementation and shows an example model. Results on synthetic data are given in section 11, and finally, in section 12 the recognition performance on various dialect regions of TIMIT are given.

2 The Hidden Markov Model

The HMM is the most popular and successful stochastic approach to speech recognition in general use. Its popularity and success are due to the existence of elegant and efficient algorithms for both training and recognition. However, the use of HMMs for speech recognition is dependent on certain assumptions. These are

1. Speech may be split into segments, states, in which the speech waveform may be assumed to be stationary. The transition between these states is assumed to be instantaneous.
2. The probability of a certain symbol being generated is only dependent on the current state, not on any previously generated symbols. This is a first order Markov assumption and is usually referred to as the 'Independence Assumption'.

In order that the first assumption is as valid as possible, a model having many states would be desirable, however this creates a problem in reliably estimating model parameters. Hence in any real system, a small number of states are used. The second assumption is not valid and is the major drawback to the use of HMMs for speech recognition.

2.1 Basic Theory

The basic theory for HMMs is presented here. A more detailed discussion of HMM theory and application for speech recognition is given by Rabiner [12].

HMMs are characterised by

1. N, the number of states in the model. The states will be denoted as $\mathbf{S} = \{S_1, S_2, ..., S_N\}$ and $q_i(t)$ indicates being in state S_i at time t. In addition the concept of segments will be used, a segment being defined as

$$s_i(t_i, t_j) = [q_i(t_i), q_i(t_i+1), .., q_i(t_j)] \tag{1}$$

   A particular segmentation will be defined as $\mathbf{s}_T = [s_1(1, t_1), s_i(t_1+1, t_2)..., s_N(t_{K-1}+1, T)]$, where K is the number of segments and states S_1 and S_N are the start and end states, respectively.

2. $\mathbf{A}$, the state probability transition matrix, where

$$a_{ij} = p(q_j(t+1)|q_i(t)) \tag{2}$$

3. $\mathbf{B}$, the output probability distribution, where

$$b_j(\mathbf{y}_t) = p(\mathbf{y}_t|q_j(t)) \tag{3}$$

   There are two general types of HMM split according to the form of their output distribution. If the output distribution is based on discrete elements then the models are called Discrete HMMs (DHMMs). Alternatively if the output distribution is continuous they are referred to as Continuous Density HMMs (CDHMMs). This report only considers CDHMMs. Usually in CDHMMs, the output distribution used is a multivariate Gaussian distribution or mixture of multivariate Gaussian distributions. For the multiple Gaussian mixture case

$$b_j(\mathbf{y}_t) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{y}_t; \mu_m, \mathbf{\Sigma}_m) \tag{4}$$

4. π, the initial state distribution, where

$$\pi_i = p(q_i(1)) \tag{5}$$

For convenience the following compact notation will be used

$$\mathcal{M} = (\mathbf{A}, \mathbf{B}, \pi) \tag{6}$$

There are many references dealing with the derivation of the re-estimation formulae, only the results are of interest here. An observation sequence is defined as $\mathbf{Y}_T = [\mathbf{y}_1, ..., \mathbf{y}_T]$, where each observation $\mathbf{y}_\tau$ is an n-dimensional vector $\mathbf{y}_\tau = [y_{\tau(1)}, y_{\tau(2)}, ..., y_{\tau(n)}]$. To be consistent with later work a segment is defined as being a group of consecutive parametrised frames, $\mathcal{Y}_{t_i,t_j} = [\mathbf{y}_{t_i}, \mathbf{y}_{t_i+1}, ..., \mathbf{y}_{t_j}]$. The probability, P, of a particular observation sequence is given by

$$\begin{aligned} P = p(\mathbf{Y}_T|\mathcal{M}) &= \sum_s p(\mathbf{Y}_T|\mathbf{Q}, \mathcal{M}) p(\mathbf{Q}|\mathcal{M}) \\ &= \sum_s p(\mathbf{Q}|\mathcal{M}) \prod_{\tau=1}^{T} p(\mathbf{y}_\tau|q_i(\tau), \mathcal{M}) \end{aligned} \tag{7}$$

where s is over all possible segmentations and $\mathbf{Q}$ is the frame state alignment associated with that segmentation. This is in fact a statement of assumption 2. Computationally equation 7 is very expensive if implemented directly. Luckily there exist efficient algorithms for this calculation.

2.2 Viterbi algorithm

The probability calculation shown above is a summation over all possible state sequences. However in HMMs it is usually assumed that for recognition and sometimes for training one sequence will dominate, so a Viterbi decoder may be used. A new variable $\phi_t(i)$ is introduced.

$$\phi_t(i) = \max_s [p(\mathbf{Y}_t, q_i(t)|\mathcal{M})] \tag{8}$$

where s is over all possible state sequences. Values of $\phi_t(i)$ may be efficiently calculated using the following recursive equation.

$$\phi_{t+1}(j) = \max_{1 \le i \le N} [\phi_t(i) a_{ij}]\, b_j(\mathbf{y}_{t+1}) \tag{9}$$

This algorithm is normally used for recognition and has a computational cost of $\mathcal{O}(T)$.

2.3 Forward Backward algorithm

An alternative to making the assumption that one state sequence dominates is to use the Forward Backward algorithm. For this, it is necessary to define new variables

$$\alpha_t(j) = p(\mathbf{y}_1, .., \mathbf{y}_t, q_j(t)|\mathcal{M}) \tag{10}$$

$$\beta_t(j) = p(\mathbf{y}_{t+1}, .., \mathbf{y}_T|q_j(t), \mathcal{M}) \tag{11}$$

Using these definitions it is possible to define iterative re-estimation formulae

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(\mathbf{y}_{t+1}) \tag{12}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{y}_{t+1}) \beta_{t+1}(j) \tag{13}$$

The probability of a particular observation sequence is now given by

$$P = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(\mathbf{y}_{t+1}) \beta_{t+1}(j) \tag{14}$$

This algorithm again has a computational cost of $\mathcal{O}(T)$.

2.4 Re-Estimation Formulae

For the standard single Gaussian mixture HMM the re-estimation formulae are

$$\hat{\mu}_j = \frac{\sum_{\tau=1}^{T} L_j(\tau)\mathbf{y}_\tau}{\sum_{\tau=1}^{T} L_j(\tau)} \tag{15}$$

and

$$\hat{\mathbf{\Sigma}}_j = \frac{\sum_{\tau=1}^{T} L_j(\tau)(\mathbf{y}_\tau - \hat{\mu}_j)(\mathbf{y}_\tau - \hat{\mu}_j)^T}{\sum_{\tau=1}^{T} L_j(\tau)} \tag{16}$$

where

$$L_j(t) = p(q_j(t)|\mathbf{Y}_T, \mathcal{M}) \tag{17}$$

There is a choice in the above expressions for the form of $L_j(t)$. If one state sequence is assumed to dominate then $L_j(t) \in \{0, 1\}$ and is given by the state sequence from equation 9. This will be referred to as Viterbi re-estimation. If this assumption is not used and the full summation, as defined in the Forward Backward algorithm, is used then

$$L_j(t) = \frac{1}{P}\alpha_j(t)\beta_j(t) \tag{18}$$

This will be referred to as Baum-Welch re-estimation.

2.5 Compensating for the Independence Assumption

As mentioned in the previous section, a major drawback in the use of HMMs is the invalidity of the 'Independence Assumption'. Various methods of overcoming this problem have been tried. Some methods are related to this work and are described below.

2.5.1 Dynamic Coefficients

The simplest method of handling correlation between observation vectors is the addition of dynamic coefficients [17]. These dynamic coefficients are added to the feature vector and are trained in the same way as the static coefficients. Many forms of dynamic coefficients are used, ranging from simple delta coefficients to acceleration coefficients to bandpass filtered coefficients.

The use of dynamic coefficients, or in a more general sense digitally filtered cepstral trajectories, has been shown to reduce recognition errors. This indicates that the assumptions behind the standard HMM using static coefficients are poor. However dynamic coefficients do not affect the static coefficients, so do not really compensate for the 'Independence Assumption', but just incorporate parameters which are less sensitive to it. In addition, the assumption of stationary segments of speech with instantaneous transitions between states is broken by the use of digital filtering of the cepstral parameters as the filters will run over the state boundaries.

2.5.2 Explicit Time Correlation Modelling

It is possible to modify the basic HMM assumption to allow improved modelling of speech. The standard 'Independence Assumption' results in

$$p(\mathbf{y}_t|\mathbf{Y}_{t-1}, \mathbf{Q}, \mathcal{M}) = p(\mathbf{y}_t|q_i(t), \mathcal{M}) \tag{19}$$

This may be modified to make the present observation conditional on the previous observation and state in addition to the present state. Hence

$$p(\mathbf{y}_t|\mathbf{Y}_{t-1}, \mathbf{Q}, \mathcal{M}) = p(\mathbf{y}_t|\mathbf{y}_{t-1}, q_i(t), q_j(t-1), \mathcal{M}) \tag{20}$$

Models of this form have been studied by Brown [15] and Wellekens [2]. Indeed, if desired, the probability may be made conditional on an arbitrary history of observations.

This method only partially overcomes the problem as the history over which the correlation is taken tends to be short in order to limit the increase in the number of parameters.

2.5.3 Variable Frame Rate Analysis

In Variable Frame Rate (VFR) analysis [7] the problem of compensating for the independence assumption is moved from a model adaptation process to a front end process. In this model, the first frame is retained and is assumed to be representative of all subsequent non-retained frames, until a frame is greater than some predefined distance from the previously retained frame. This new frame is then retained and the process repeated. The probability of a segment, a group of frames assumed to be represented by a single frame, is

$$p(\mathcal{Y}_{t_i,t_j}, s_i(t_i,t_j)|\mathcal{M}) \quad = \quad D_i(t_j - t_i + 1)\mathcal{N}(\mathbf{y}_{t_i}; \mu_i, \mathbf{\Sigma}_i) \tag{21}$$

where $D_i(\tau)$ is the probability of being in state S_i for duration τ, μ_i and $\mathbf{\Sigma}_i$ are the mean and variance associated with that state. By varying the distance threshold it is possible to vary the length of the segments and the effective frame rate.

This is a crude method of overcoming the correlation problem, as the first frame is assumed to be representative of all subsequent frames until a new frame is retained. This is not true for speech and results in the loss of acoustic information.

2.5.4 Stochastic Segment Model

One method of dealing with the independence assumption is to assume that the data is independent on the phone or model level. One implementation of this form is the Stochastic Segment Model (SSM) [13, 23].

Using the definition of the segment previously given, and stacking the observation vectors to form one large composite observation

$$p(\mathcal{Y}_{t_i,t_j}, s_i(t_i,t_j)|\mathcal{M}) \quad = \quad D_i(t_j - t_i + 1)\mathcal{N}(\mathcal{Y}_{\tau}; \mu_{\tau}, \mathbf{\Sigma}_{\tau}) \tag{22}$$

where $\mathcal{Y}_\tau$ is the mapped form of $\mathcal{Y}_{t_i,t_j}$, the mapping is described later, and τ is the number of 'mapped' frames. Here the observation probability parameters $\{\mu_\tau, \mathbf{\Sigma}_\tau\}$, are of dimension $k\tau$, where k is the length of each observation vector at each time instance. Durational modelling is not examined in this report, so $D_i(\tau)$ is not considered in the following work.

Two problems are associated with this style of model

Model	*Covariance Style*	*Number Parameters*
HMM	Diagonal	$NM(2k+1)$
HMM	Full	$NM(k((k+1)/2+1)+1)$
SSM	Temporal Full	$k(\tau((\tau+1)/2+1))$
SSM	Spatial Full	$\tau(k((k+1)/2+1))$
SSM	Full	$k\tau((k\tau+1)/2+1)$

Table 1: Number of Parameters per Model

1. Number of parameters. A comparison of the number of parameters in a full implementation of a SSM with those of a standard HMM is shown in table 1. In the table τ is the segment length, M the number of mixtures and k the length of the feature vector. A *Temporal Full* covariance matrix is one in which the elements of the feature vector are assumed to be spatially independent and correlated over time. The *Spatial Full* covariance case assumes that the parameters are independent over time, but spatially correlated with one another.

   From the table it can be seen that as τ increases, the number of parameters increases rapidly. In particular, this is a problem for the *Full* and *Temporal Full* covariance cases where the number of parameters rises as $\mathcal{O}(\tau^2)$. Unfortunately these cases are of most interest, since if the *Spatial Full* covariance model is used there is no obvious advantage over the standard HMM structure. Various other covariance styles may be implemented within the SSM framework, such as a block diagonal covariance matrix, but these appear to nullify the major advantages of the SSM over the previously mentioned schemes.

2. Varying length of Segments. In recognition and training there are a variable number of frames allocated to each state or phone, due to variations in speaker rate. In the SSM, it is assumed that the same number of frames are always allocated to each model. There are two basic methods of doing this.

   The first method takes the number of observed frames and maps it down to the required number, either using some temporal interpolation measure, or according to some distance measure [13]. This keeps τ low, but reduces the information content in a way similar to VFR analysis.

   Alternatively, the number of frames can be mapped up to the required number of frames, using an expected value for the unobserved frames given the observed frames [23]. This maintains all the information of the waveform but greatly increases the value of τ.

The SSM may be used to model the correlation well, however it dramatically increases the number of parameters to be estimated. This may result in problems in accurately estimating all the required parameters given a limited training data set.

3 The Segmental Hidden Markov Model

A model which minimises the effects of the 'Independence Assumption' without significantly increasing the number of parameters is required. To this end a new style of acoustic model is introduced, the Segmental Hidden Markov Model (SHMM). In the SHMM all observations are assumed to be independent given the state that generated them, but additionally they are conditional on the mean of the segment of speech to which they belong. The idea behind this assumption is that certain characteristics of the speech, such as speaker or stress condition, are fixed over the whole segment. Hence, when the first frame in the segment is observed, some characteristics are known and fixed. In standard HMMs this information is completely ignored. However, in making the observations conditionally dependent on the mean of the segment, the SHMM takes these effects into account, thereby making better use of the acoustic information.

With this new assumption it is necessary to calculate the probability of a segment given a particular model. For each state of the model $\mathcal{M}$ the output probability distribution will no longer be described by one distribution, but by two. One describing the distribution of the segment mean, the 'inter distribution', the other the observation probabilities given that mean, the 'intra distribution'. For this and subsequent sections a shorthand notation is used of representing $s_i(t_i, t_j)$ as s_i, where the segment boundaries are obvious. The required probability is

$$p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M}) \quad = \quad \int\limits_{\mathcal{R}^n} p(\mu|s_i,\mathcal{M})p(\mathcal{Y}_{t_i,t_j}|\mu,s_i,\mathcal{M})\,d\mu. \tag{23}$$

Again $D_i(\tau)$, the probability of staying in state S_i for duration τ is ignored. Using the assumption, that given the mean of the segment, all the observations within that segment are independent

$$p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M}) \quad = \quad \int_{\mathcal{R}^n} p(\mu|s_i,\mathcal{M}) \prod_{\tau=t_i}^{t_j} p(\mathbf{y}_\tau|\mu,s_i,\mathcal{M})\, d\mu \tag{24}$$

The above expression and assumptions have also been used by Russell [14]. However he uses an MAP approach to estimate the mean, this will be referred to as MAP SHMM. Here the full form of the above expression is used.

Assuming that $p(\mu|s_i,\mathcal{M})$ and $p(\mathbf{y}_\tau|\mu,s_i,\mathcal{M})$ are both Gaussian probability distributions, the output distribution is described by $\{\mathbf{\Sigma},\mu_c,\mathbf{\Sigma}_c\}$, the intra-state variance, the inter-state mean and the inter-state variance respectively. Hence

$$\begin{aligned} p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M}) \quad = \quad & \int_{\mathcal{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_c|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mu-\mu_c)\mathbf{\Sigma}_c^{-1}(\mu-\mu_c)^T) \\ & \prod_{\tau=t_i}^{t_j} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{y}_\tau-\mu)\mathbf{\Sigma}^{-1}(\mathbf{y}_\tau-\mu)^T)\, d\mu \end{aligned} \tag{25}$$

In appendix A the above equation is rewritten in terms of μ_n and $\mathbf{\Sigma}_n^{-1}$, which are independent of μ, to give the following form

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) = t\log(K) + \log(K_c) - \log(K_n) - \frac{1}{2}\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) \tag{26}$$

where

$$\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) = \mu_c\mathbf{\Sigma}_c^{-1}\mu_c^T + \sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau\mathbf{\Sigma}^{-1}\mathbf{y}_\tau^T - \mu_n\mathbf{\Sigma}_n^{-1}\mu_n^T \tag{27}$$

and

$$\mathbf{\Sigma}_n^{-1} \quad = \quad \mathbf{\Sigma}_c^{-1} + t\mathbf{\Sigma}^{-1} \tag{28}$$

$$\mu_n^T \quad = \quad \mathbf{\Sigma}_n\left(\mathbf{\Sigma}_c^{-1}\mu_c^T + \mathbf{\Sigma}^{-1}\mu_s^T\right) \tag{29}$$

$$\mu_s^T \quad = \quad \sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau^T \tag{30}$$

where $t = t_j - t_i + 1$, the segment duration, K, K_c and K_n are the standard normalising constants associated with $\mathbf{\Sigma}$, $\mathbf{\Sigma}_c$ and $\mathbf{\Sigma}_n$ respectively. The above equation has been derived for the full covariance matrix case. If it is now assumed that the covariance matrix is diagonal it is possible to simplify $\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M})$ in terms of the model parameters

$$\begin{aligned} & \mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) \\ = \quad & \sum_{i=1}^{n}\left[\frac{\mu_{c(i)}^2}{\mathbf{\Sigma}_{c(i,i)}} + \frac{1}{\mathbf{\Sigma}_{(i,i)}}\sum_{\tau=t_i}^{t_j}y_{\tau(i)}^2 - \frac{\mu_{c(i)}^2\mathbf{\Sigma}_{(i,i)}^2 + 2\mu_{c(i)}\mu_{s(i)}\mathbf{\Sigma}_{c(i,i)}\mathbf{\Sigma}_{(i,i)} + \mu_{s(i)}^2\mathbf{\Sigma}_{c(i,i)}^2}{\mathbf{\Sigma}_{c(i,i)}\mathbf{\Sigma}_{(i,i)}(t\mathbf{\Sigma}_{c(i,i)} + \mathbf{\Sigma}_{(i,i)})}\right] \\ = \quad & \sum_{i=1}^{n}\left[\frac{1}{\mathbf{\Sigma}_{(i,i)}}\left(\sum_{\tau=t_i}^{t_j}y_{\tau(i)}^2 - t\left(\frac{\mu_s}{t}\right)^2\right) + \mathcal{K}(s_i,\mathbf{\Sigma}_c,\mathbf{\Sigma})\left(\mu_{c(i)} - \frac{\mu_s}{t}\right)^2\right] \end{aligned} \tag{31}$$

where

$$\mathcal{K}(s_i,\mathbf{\Sigma}_c,\mathbf{\Sigma}) = \frac{t}{\mathbf{\Sigma}_{(i,i)} + t\mathbf{\Sigma}_{c(i,i)}} \tag{32}$$

From equation 31 it can be seen that $p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})$ is dependent on the vector sum, the vector squared sum and duration of the segment $\mathcal{Y}_{t_i,t_j}$. It is therefore sufficient to store these values for each possible segment to calculate the probability.

The above analysis has been performed for Gaussian inter and intra distributions. Any form of inter and intra distributions may be used. However, if only distributions where the inter distribution is a conjugate prior of the intra distribution are considered, closed forms for the probabilities and sufficient statistics given the complete data set exist.

3.1 Relationship to MAP SHMM

The SHMM may be related to the MAP SHMM. For this section it is assumed that the feature vector has dimensionality 1 so the index i is ignored. For the MAP SHMM

$$\begin{aligned} \log(p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) &= \log(\mathcal{N}(\hat{c};\mu_c,\Sigma_c)) + \sum_{\tau=t_i}^{t_j} \log(\mathcal{N}(y_\tau;\hat{c},\Sigma)) \\ &= \log(K_c) - \frac{1}{2\Sigma_c}(\hat{c}-\mu_c)^2 + t\log(K) - \frac{1}{2\Sigma}\sum_{\tau=t_i}^{t_j}(\hat{c}-y_\tau)^2 \end{aligned} \tag{33}$$

where the target mean, $\hat{c}$ is given by

$$\hat{c} = \frac{\mu_c\Sigma + \mu_s\Sigma_c}{\Sigma + t\Sigma_c} \tag{34}$$

Substituting 34 into 33 and simplifying gives

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) = t\log(K) + \log(K_c) - \frac{1}{2}\left(\frac{\mu_c^2}{\Sigma_c} + \frac{1}{\Sigma}\sum_{\tau=t_i}^{t_j} y_\tau^2 - \frac{(\mu_c\Sigma + \mu_s\Sigma_c)^2}{\Sigma\Sigma_c(\Sigma + t\Sigma_c)}\right) \tag{35}$$

This expression is identical to the probability equation for the SHMM proposed here, except for the term $\log(K_n)$. This normalisation constant may be written as

$$\log(K_n) = -\frac{1}{2}\left(\log(2\pi) + \log\left(\frac{\Sigma\Sigma_c}{t\Sigma_c + \Sigma}\right)\right) = \log(K) + \frac{1}{2}\left(\log(t) + \log\left(1 + \frac{\Sigma}{t\Sigma_c}\right)\right) \tag{36}$$

This term is solely dependent on the length of the segment, not the observations within that segment. Hence the use of the MAP estimate of the mean, instead of the true distribution for the mean, results in a model dependent bias on the length of the segmentation.

3.2 Multiple Gaussian Inter Mixtures

The above analysis has been performed assuming a single Gaussian inter mixture probability distribution for $p(\mu|s_i,\mathcal{M})$. If in fact this distribution is described by a multiple Gaussian mixture distribution the same style of analysis may be applied. Letting

$$p(\mu|s_i,\mathcal{M}) = \sum_{m=1}^{M} c_m \mathcal{N}(\mu;\mu_{c_m},\mathbf{\Sigma}_{c_m}) \tag{37}$$

and substituting this in equation 24

$$\begin{aligned} p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M}) &= \int_{\mathcal{R}^n} \sum_{m=1}^{M} c_m \mathcal{N}(\mu;\mu_{c_m},\mathbf{\Sigma}_{c_m}) \prod_{\tau=t_i}^{t_j} p(\mathbf{y}_\tau|\mu,s_i,\mathcal{M})\,d\mu \\ &= \sum_{m=1}^{M} c_m \int_{\mathcal{R}^n} \mathcal{N}(\mu;\mu_{c_m},\mathbf{\Sigma}_{c_m}) \prod_{\tau=t_i}^{t_j} p(\mathbf{y}_\tau|\mu,s_i,\mathcal{M})\,d\mu \end{aligned} \tag{38}$$

It can be seen that the analysis for the single mixture case may be directly applied to the multiple mixture case.

3.3 Multiple Gaussian Intra Mixtures

In the previous section multiple Gaussian inter mixture models are described. Multiple Gaussian intra mixtures models occur when

$$p(\mathbf{y}_\tau|\mu, s_i, \mathcal{M}) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{y}_\tau; \mu_m, \mathbf{\Sigma}_m) \tag{39}$$

where $\mu_m = \mu + \mathbf{\Delta}_m$. The delta intra means, $\mathbf{\Delta}_m$, are stored as model parameters. By substituting the above expression in equation 24 and assuming a single Gaussian inter mixture yields

$$p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{M}) = \int_{\mathcal{R}^n} \mathcal{N}(\mu; \mu_c, \mathbf{\Sigma}_c) \prod_{\tau=t_i}^{t_j} \left(\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{y}_\tau; \mu_m, \mathbf{\Sigma}_m) \right) d\mu \tag{40}$$

The analysis for the single intra mixture case can be seen to be inappropriate for the multiple Gaussian intra mixture case, due to the product of the weighted sum of Gaussians. In fact there are no sufficient statistics for the multiple intra mixture case [3].

4 Relationship to Standard HMMs

As previously stated the observation probability for a segment is given by

$$p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{M}) = \int_{\mathcal{R}^n} p(\mu|s_i, \mathcal{M}) \prod_{\tau=t_i}^{t_j} p(\mathbf{y}_\tau|\mu, s_i, \mathcal{M})\, d\mu \tag{41}$$

If the inter mixture variance, $\mathbf{\Sigma}_c$, is set to zero then $p(\mu|s_i, \mathcal{M}) = \delta(\mu_c - \mu)$, where $\delta(.)$ is the Dirac delta function, and the above equation may be simplified to

$$p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{M}) = \prod_{\tau=t_i}^{t_j} p(\mathbf{y}_\tau|\mu_c, s_i, \mathcal{M}) \tag{42}$$

This is the same expression as a standard single Gaussian mixture HMM with the mean set to the interstate mean, μ_c, and the variance set to the intra state variance, Σ. Hence, the standard single Gaussian mixture HMM may be viewed as a subset of the SHMM.

Equivalent standard models for the multiple Gaussian inter mixture model described in section 3.2 are also possible. By setting all the inter mixture variances to zero in equation 38 results in

$$p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{M}) = \sum_{m=1}^{M} c_m \left(\prod_{\tau=t_i}^{t_j} p(\mathbf{y}_\tau|\mu_{c_m}, s_i, \mathcal{M}) \right) \tag{43}$$

So the multiple Gaussian inter mixture case with zero inter mixture variances is the same as the standard multiple Gaussian mixture case with the added constraint that all observations in a segment associated with a given state are generated by the same Gaussian mixture.

One popular form of standard HMM used is the multiple Gaussian mixture model. An equivalent SHMM model would be of interest. Looking at equation 40 and setting $\mathbf{\Sigma}_c$ to zero

$$p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{M}) = \prod_{\tau=t_i}^{t_j} \left(\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{y}_\tau; \mu_m, \mathbf{\Sigma}_m) \right) \tag{44}$$

where $\mu_m = \mu_c + \mathbf{\Delta}_m$. This is identical to the standard multiple Gaussian mixture model and so the standard HMM multiple Gaussian mixture model may be viewed as a subset of the single Gaussian inter mixture, multiple Gaussian intramixture segment model.

So far only situations where the inter mixture variance has been set to zero have been considered. It is also possible to set the intra mixture variance to zero. The physical requirement of this zeroing is that all observations of a given segment are identical. A model of this form has already been proposed by Ponting [7], VFR. As previously described, this model assumes that the first observation of a segment is representative of the whole segment until an incoming frame is greater than some threshold from the first frame. An improvement described by Russell [14] uses the segment mean as representative of the whole segment. SHMMs and VFR trained models are not quite equivalent, due to the form of the distance measure used to decide on the segment lengths. In VFR a Euclidean distance measure is used. There is no easy mapping of this onto the segment lengths used for the SHMM.

5 Re-estimation Formulae

In order to re-estimate the parameters of the SHMM an auxiliary function $\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})$ is introduced

$$\begin{aligned}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) &= \sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\log(p(\mathbf{Y}_T,\mathbf{s}_T|\hat{\mathcal{M}}))\\ &= \sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\sum_{k=1}^{K}\log(p(\mathcal{Y}|s_k,\hat{\mathcal{M}}))+\log(p(\mathbf{s}_T|\hat{\mathcal{M}}))\right] \end{aligned}\tag{45}$$

where the summation on s is over every possible segmentation of $\mathbf{Y}_T$ and K is the number of segments in $\mathbf{s}_T$. It is shown by Baum [9] that if $\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})\geq\mathcal{Q}(\mathcal{M},\mathcal{M})$ then $p(\mathbf{Y}_T|\hat{\mathcal{M}})\geq p(\mathbf{Y}_T|\mathcal{M})$. For all the re-estimation formulae derived, diagonal covariance matrices are assumed and, for simplicity of notation, n is assumed to be 1 so that the vector notation has been dropped. In addition, the notation $\mathcal{Y}$ and s_i to represent $\mathcal{Y}_{t_i,t_j}$ and $s_i(t_i,t_j)$ respectively will be used and the means and variances will be assumed to relate to state S_i. For the following analysis left-to-right models are considered. Hence $K=N$ and only one segment from each utterance is associated with a particular state. This assumption is not necessary, but it further simplifies the notation.

5.1 Re-estimation for $\hat{\mu}_c$

Taking the partial derivative with respect to $\hat{\mu}_c$

$$\begin{aligned}\frac{\partial}{\partial\hat{\mu}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) &= \sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{\partial}{\partial\hat{\mu}_c}\left(\log(p(\mathcal{Y}|s_i,\hat{\mathcal{M}}))\right)\\ &= -\sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\left(\hat{\mu}_c-\frac{\mu_s}{t}\right)\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]\end{aligned}\tag{46}$$

where t is the length of the segment s_i. Equating the above equation to zero and solving for μ_c

$$\hat{\mu}_c=\frac{\sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\frac{\mu_s}{t}\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]}{\sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]}\tag{47}$$

It is necessary to find whether this is a maximum or minimum. Taking the second derivative

$$\frac{\partial^2}{\partial^2\hat{\mu}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = -\sum_{s} p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]\tag{48}$$

As $\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})>0$, and the second derivative is always negative, this is a maximum.

5.2 Re-estimation for $\hat{\Sigma}_c$

Taking the partial derivative with respect to $\hat{\Sigma}_c$

$$\begin{aligned}\frac{\partial}{\partial \hat{\Sigma}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) &= \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{\partial}{\partial \hat{\Sigma}_c}\left(\log(p(\mathcal{Y}|s_i,\hat{\mathcal{M}}))\right) \\ &= \frac{1}{2}\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})^2(\hat{\mu}_c-\frac{\mu_s}{t})^2-\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right] \qquad (49)\end{aligned}$$

Equating the above equation to zero yields no closed form for $\hat{\Sigma}_c$.

5.3 Re-estimation for $\hat{\Sigma}$

Taking the partial derivative with respect to $\hat{\Sigma}$

$$\begin{aligned}&\frac{\partial}{\partial \hat{\Sigma}}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) \\ &= \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{\partial}{\partial \hat{\Sigma}}\left(\log(p(\mathcal{Y}|s_i,\hat{\mathcal{M}}))\right) \qquad (50)\\ &= \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{1}{2\hat{\Sigma}^2}\left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - t\hat{\Sigma}+\hat{\Sigma}\hat{\Sigma}_c\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})-\frac{\mu_s^2}{t}+\frac{\hat{\Sigma}^2\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})^2(\hat{\mu}_c t-\mu_s)^2}{t}\right]\end{aligned}$$

Again equating the above expression to zero yields no closed expression for $\hat{\Sigma}$.

5.4 Approximate Solution

In the previous section closed form expressions for re-estimating all the parameters of the SHMM were shown not to exist. It was only possible to find such an expression for $\hat{\mu}_c$. Closed forms for estimating Σ and Σ_c would be desirable. Rewriting equation 32

$$\mathcal{K}(s_i,\Sigma_c,\Sigma)=\frac{1}{\Sigma_c}\left(\frac{1}{1+\frac{\Sigma}{t\Sigma_c}}\right) \qquad (51)$$

If it is assumed that $t\hat{\Sigma}_c \gg \hat{\Sigma}$, ie the between segment variability is far greater than the within segment variability, then

$$\mathcal{K}(s_i,\Sigma_c,\Sigma)=\frac{1}{\Sigma_c}\left(1-\frac{\Sigma}{t\Sigma_c}+\left(\frac{\Sigma}{t\Sigma_c}\right)^2-\ldots\right) \qquad (52)$$

and by ignoring terms in $\left(\frac{\Sigma}{t\Sigma_c}\right)$ and higher, $\mathcal{K}()$ is independent of the segmentation. Rewriting equation 47 yields

$$\hat{\mu}_c=\frac{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\frac{\mu_s}{t}\right]}{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})} \qquad (53)$$

Looking at equation 49

$$\frac{\partial}{\partial \hat{\Sigma}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\frac{\frac{1}{t^2}(\hat{\mu}_c t-\mu_s)^2-\hat{\Sigma}_c}{2\hat{\Sigma}_c^2}\right] \qquad (54)$$

Setting the above equation equal to zero and noting that $\hat{\Sigma}_c$ is independent of the segmentation then

$$\hat{\Sigma}_c = \frac{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{1}{t^2}(\hat{\mu}_c t-\mu_s)^2}{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})} \qquad (55)$$

Finally rewriting equation 51 and additionally ignoring terms in $\left(\frac{\Sigma}{\Sigma_c}\right)^2$

$$\frac{\partial}{\partial\hat{\Sigma}}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{1}{2\hat{\Sigma}^2}\left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - (t-1)\hat{\Sigma} - \frac{\mu_s^2}{t}\right] \tag{56}$$

Now equating the differential to zero and solving for $\hat{\Sigma}$

$$\hat{\Sigma} = \frac{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - \frac{\mu_s^2}{t}\right]}{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})(t-1)} \tag{57}$$

The proof that each of these estimates is a maximum is shown in appendix B.

5.5 Comparison with Standard HMM re-estimation Formulae

It was previously noted that the standard HMM is simply a subset of the SHMM. Having derived re-estimation formulae and optimisation criteria for the SHMM, it is interesting to see what models are generated if the HMM assumptions are true for a particular set of acoustic waveforms. The impact of the 'Independence Assumption' from the viewpoint of the SHMM is that for all segments $\mathcal{E}\{y_\tau\} = \mu_a$, where μ_a is some fixed value, the mean of the underlying acoustic waveform. Here it also assumed that the segments are long enough that

$$\frac{\mu_s}{t} \approx \mu_a \tag{58}$$

If this expression is substituted in equation 47

$$\begin{aligned}\hat{\mu}_c &= \frac{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mu_a\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]}{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]} \\ &= \mu_a \end{aligned} \tag{59}$$

Now substituting equation 58 in equation 49

$$\begin{aligned}\frac{\partial}{\partial\hat{\Sigma}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) &= \frac{1}{2}\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})^2(\hat{\mu}_c-\mu_a)^2 - \mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right] \\ &= -\frac{1}{2}\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma})\right]\end{aligned} \tag{60}$$

Finding a minimum or maximum to this expression is not possible as $\mathcal{K}(s_i,\hat{\Sigma}_c,\hat{\Sigma}) > 0$. Hence the derivative is always negative. However, there are bounds on the possible values of $\hat{\Sigma}_c$ as the variance must by definition be greater than or equal to zero. As the gradient is always negative the maximum log probability will occur when $\hat{\Sigma}_c$ is at a minimum, zero.

Finally rewriting equation 51 using equation 58 and the previous results

$$\frac{\partial}{\partial\hat{\Sigma}}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{1}{2\hat{\Sigma}^2}\left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - t\hat{\Sigma} - t\mu_a^2\right] \tag{61}$$

Setting the above expression to zero and solving for $\hat{\Sigma}$

$$\hat{\Sigma} = \frac{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - t\mu_a^2\right]}{\sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\,[t]}$$

$$= \frac{\sum_{s'} L_j(t_i,t_j) \left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - t\mu_a^2 \right]}{\sum_{s'} L_j(t_i,t_j)\,[t]} \tag{62}$$

where

$$L_j(t_i,t_j) = \frac{1}{P} p(\mathbf{Y}_T, s_i(t_i,t_j)|\mathcal{M}) = p(s_i|\mathbf{Y}_T,\mathcal{M}) \tag{63}$$

and the s' summation is over all possible s_i.

These expressions may be compared to the standard re-estimation formulae given in section 2. Comparing equation 15 with 59 it may be seen that they are identical provided a Viterbi re-estimation scheme is used. The inter variance has already been shown to tend to zero as required for a standard HMM. Comparing equation 16 with 62 and again by assuming Viterbi re-estimation and additionally noting that $\mathcal{E}\{(y_\tau - \mu)^2\} = \mathcal{E}\{y_\tau^2\} - \mu^2$ it may be seen that given long enough segments the two equations will yield the same result.

In the above analysis it was assumed that the mean for each segment was the same, μ_a. In reality, with finite length segments, this will not be true. There will be some variance on the means of the segments, this variance being $\frac{\Sigma_a}{t}$, where t is the length of the segment. Hence if the re-estimation formulae are applied to real data the values obtained will be close to, but not identical to the values of the standard HMM.

6 Segmental HMM Estimation Stage

As in the case of standard HMMs, the SHMM re-estimation formulae and recognition algorithms are in terms of the complete data set. It is therefore necessary to estimate the complete data set and iterate in a standard *EM* algorithm [1] style. Examining the number of possible segmentations assuming no bounds on segment length, yields 2^{T-1} to search over. It is not possible to search over all these solutions, so some efficient algorithm similar to the standard HMM algorithms are required. Again two options are available when estimating the complete data set. It may either be assumed that one path, segmentation, dominates the probability calculation, or that all paths must be considered. These will be referred to as Viterbi and Baum-Welch re-estimation respectively, in a similar way to the standard HMM re-estimation formulae.

6.1 Viterbi Re-Estimation

For Viterbi re-estimation it is necessary to find the most likely state sequence through the model. A new variable, $\phi_t(j,\tau)$, is introduced.

$$\phi_t(j,\tau) \quad = \quad \max_s [p(\mathbf{Y}_t, s_j(t',t), \overline{q}_j(t+1)|\mathcal{M})] \tag{64}$$

where $t' = t - \tau + 1$ and $\overline{q}_j(t+1)$ indicates that the model is not in state S_j at time $t+1$. Values of $\phi_t(j,\tau)$ may be calculated using the following recursive equation

$$\phi_t(j,\tau) \quad = \quad \max_{1\le i\le N, i\ne j} \left[\max_{1<\gamma<t-\tau} [\phi_{t-\tau}(i,\gamma)a_{ij}] \right] p(\mathcal{Y}_{t-\tau+1,t}|s_j(t',t),\mathcal{M})D_j(\tau) \tag{65}$$

This recursive expression is similar to that of a semi hidden Markov model [20]. Taking the maximum likelihood path through the model results in $L_j(t_i,t_j) \in \{0,1\}$. In a left to right model this will result in a maximum of N segments for each utterance.

Comparing the computational loads of the SHMM Viterbi re-estimation with that of the standard HMM, equation 9, shows a significant overhead for the SHMM. This overhead is due to the need to look at all possible previous segmentations, which results in a cost of $\mathcal{O}(T^2)$. This cost may be reduced by assuming a maximum duration in any state. If the maximum duration is t_{max} then the cost is $\mathcal{O}(Tt_{max})$.

Viterbi estimation can also be used for the recognition stage of an SHMM in an identical way to the standard HMM. For recognition equation 65 is used, so again there is a computational overhead at recognition time of $\mathcal{O}(Tt_{max})$.

6.2 Baum-Welch Re-Estimation

As mentioned in section 2 the standard HMM has an efficient algorithm for calculating the total probability summed over all possible paths, segmentations, the Forward Backward algorithm. A similar efficient algorithm for the SHMM is required. New versions of the α and β are defined for the segment based model.

$$\begin{aligned} \alpha_t(j,\tau) &= p(\mathbf{y}_1,..,\mathbf{y}_{t-\tau+1},..,\mathbf{y}_t,s_j(t-\tau+1,t),\overline{q}_j(t+1)|\mathcal{M}) & (66)\\ \beta_t(j) &= p(\mathbf{y}_{t+1},..,\mathbf{y}_T|s_j(t-\tau+1,t),\overline{q}_j(t+1),\mathcal{M}) \\ &= p(\mathbf{y}_{t+1},..,\mathbf{y}_T|q_j(t),\overline{q}_j(t+1),\mathcal{M}) & (67) \end{aligned}$$

where τ describes the length of the segment ending at time t. These segment model variables,α and β, are analogous to the standard re-estimation variables in section 2. The iterative re-estimation formulae for $\alpha_t(j,\tau)$ are

$$\begin{aligned} \alpha_t(j,0) &= \sum_{i=1,i\neq j}^{N}\left(\sum_{\tau=1}^{t-1}\alpha_{t-\tau}(i,\tau)\right)a_{ij} & (68)\\ \alpha_t(j,\tau) &= \alpha_{t-\tau+1}(j,0)p(\mathcal{Y}_{t-\tau+1,t}|s_j,\mathcal{M})D_j(\tau) & (69) \end{aligned}$$

Similar iterative formulae may be generated for $\beta_t(j)$.

$$\beta_t(j) = \sum_{i=1,i\neq j}^{N}\left(\sum_{\tau=1}^{T-t}\beta_{t+\tau}(i)p(\mathcal{Y}_{t,t+\tau}|s_i,\mathcal{M})D_i(\tau)\right)a_{ji} \quad (70)$$

Hence

$$p(\mathbf{Y}_T,s_j(t_i,t_j)|\mathcal{M}) = \alpha_{t_j}(j,(t_j-t_i+1))\beta_{t_j}(j) \quad (71)$$

As with Viterbi estimation for the SHMM the Baum-Welch estimation has a significant overhead compared to the standard HMM formulae of section 2. The SHMM algorithm has a cost of $\mathcal{O}(T^2)$, but again this may be reduced to $\mathcal{O}(Tt_{max})$ by assuming a maximum duration in any state.

7 Multiple Intra State SHMMs

Previously it was stated that there are no sufficient statistics for multiple intra mixture models. However, a computationally tractable approximation to the multiple intra mixture case, using multiple intra states, is possible. Multiple intra states occur when each segment, SHMM state, has multiple states associated with it. Each segment, S_i, is now defined by the parameter set $\{\mu_c,\mathbf{\Sigma}_c,\{\mathbf{\Delta}_1^i,...,\mathbf{\Delta}_L^i\},\{\mathbf{\Sigma}_1,...,\mathbf{\Sigma}_L\},\mathbf{A}^i\}$, where $\mathbf{\Delta}_l^i$ and $\mathbf{\Sigma}_l$ are the delta intra mean and intra variance associated with intra state l, and $\mathbf{A}^i$ is the intra state transition matrix.

7.1 The New Complete Data Set

In order to use multiple intra state models it is necessary to define a new complete data set. An additional layer is added, such that the definition of $\mathbf{Q}$ is extended, so $q_t(i,j)$ indicates being in intra state $I_j(i)$ of state S_i at time t, where $\mathbf{I}(i)=\{I_1(i),I_2(i),...,I_L(i)\}$, the set of intra states associated with state S_i, and L is the number of intra states associated with that state. A new segmentation is defined

$$s_i^c(t_i,t_j) = [q_{t_i}(i,j_1),...,q_{t_j}(i,j_l)] \quad (72)$$

This new segmentation, s_i^c, may be considered to have two components, a frame inter state/mixture allocation, s_i, and a frame intra state/mixture allocation, s^l.

7.2 Multiple Intra States Probability

For a given frame inter state/mixture allocation and frame intra state/mixture allocation it is necessary to calculate the probability of a particular segment.

$$\begin{aligned} p(\mathcal{Y}_{t_i,t_j}|s_i^c,\mathcal{M}) &= \int_{\mathcal{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_c|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mu-\mu_c)\mathbf{\Sigma}_c^{-1}(\mu-\mu_c)^T) \\ & \prod_{l=1}^{L}\left(\prod_{\tau=t_l}^{t_{l+1}-1} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{y}_\tau-\mu_l)\mathbf{\Sigma}_l^{-1}(\mathbf{y}_\tau-\mu_l)^T)\right) d\mu \end{aligned} \tag{73}$$

where $\mu_l = \mu + \mathbf{\Delta}_l$. By tying all the intra state variances together, letting this tied variance be $\mathbf{\Sigma}$, and additionally defining

$$\mathbf{y}_{\tau,l} = \mathbf{y}_\tau - \mathbf{\Delta}_l \tag{74}$$

it is possible to write

$$\begin{aligned} p(\mathcal{Y}_{t_i,t_j}|s_i^c,\mathcal{M}) &= \int_{\mathcal{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_c|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mu-\mu_c)\mathbf{\Sigma}_c^{-1}(\mu-\mu_c)^T) \\ & \prod_{\tau=t_i}^{t_j} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{y}_{\tau,l}-\mu)\mathbf{\Sigma}^{-1}(\mathbf{y}_{\tau,l}-\mu)^T)\, d\mu \end{aligned} \tag{75}$$

This is identical to the standard SHMM, so eliminating μ

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i^c,\mathcal{M})) = t\log(K) + \log(K_c) - \log(K_n) - \frac{1}{2}\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) \tag{76}$$

where

$$\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) = \mu_c\mathbf{\Sigma}_c^{-1}\mu_c^T + \sum_{\tau=t_i}^{t_j}\mathbf{y}_{\tau,l}\mathbf{\Sigma}^{-1}\mathbf{y}_{\tau,l}^T - \mu_n\mathbf{\Sigma}_n^{-1}\mu_n^T \tag{77}$$

and

$$\mathbf{\Sigma}_n^{-1} = \mathbf{\Sigma}_c^{-1} + t\mathbf{\Sigma}^{-1} \tag{78}$$

$$\mu_n^T = \mathbf{\Sigma}_n\left(\mathbf{\Sigma}_c^{-1}\mu_c^T + \mathbf{\Sigma}^{-1}\mu_s^T\right) \tag{79}$$

$$\mu_s^T = \sum_{\tau=t_i}^{t_j}\mathbf{y}_{\tau,l}^T \tag{80}$$

K, K_c and K_n are the standard normalising constants associated with $\mathbf{\Sigma}$, $\mathbf{\Sigma}_c$ and $\mathbf{\Sigma}_n$ respectively.

7.3 Estimating the Complete Data Set

The probability of the multiple intra state model has been obtained given the segmentation. Expressions to obtain the frame inter state/mixture allocation given the probability of a segment, $p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})$, in either a Viterbi or Baum-Welch style, have been defined in the previous section. Given this inter state/mixture allocation it is necessary to find the intra state/mixture allocation that maximises the probability. The probability of the segment may be expressed as

$$p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) = \sum_{s^l} p(\mathcal{Y}_{t_i,t_j},s^l|s_i,\mathcal{M}) \tag{81}$$

where s^l is represents a particular intra state segmentation. There are no efficient algorithms for obtaining the optimum frame intra state/mixture allocation over the complete summation.

However, by only considering

$$\begin{aligned}\hat{p}(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) &= \max_{s^l}\left[p(\mathcal{Y}_{t_i,t_j},s^l|s_i,\mathcal{M})\right] \\ &= \max_{s^l}\left[p(s^l|s_i,\mathcal{M})p(\mathcal{Y}_{t_i,t_j},s^l|s_i^c,\mathcal{M})\right] \end{aligned} \tag{82}$$

an iterative solution is possible. It has previously been shown that the SHMM is related to the MAP SHMM, the only difference being the term $\log(K_n)$. This term is independent of the frame intra state allocation, so for the purpose of optimising the intra frame state/mixture allocation may be ignored and the MAP SHMM may be considered. The MAP estimate of the mean, $\hat{c}$, is

$$\hat{c} = \frac{\mu_c\Sigma + \mu_s\Sigma_c}{\Sigma + t\Sigma_c} \tag{83}$$

Due to the new definition of μ_s, equation 80, $\hat{c}$ is dependent on the frame intra state allocation. Using this estimate of the mean, the probability of the observation sequence, given that particular complete data set, is maximised. Assuming that the covariance matrices are all diagonal and the dimensionality is one

$$\log(p(\mathcal{Y}_{t_i,t_j},s^l|s_i,\mathcal{M})) = \log(K_c) - \frac{1}{2\Sigma_c}(\hat{c}-\mu_c)^2 + \mathcal{F}(\hat{c},y_{\tau,l}) - \log(K_n) \tag{84}$$

where

$$\mathcal{F}(\hat{c},y_{\tau,l}) = t\log(K) - \frac{1}{2\Sigma}\sum_{\tau=t_i}^{t_j}(\hat{c}-y_{\tau,l})^2 + \log(p(s^i|s_i,\mathcal{M})) \tag{85}$$

Given the current complete data set, hence estimate of $\hat{c}$, the only term that is dependent on the observations is $\mathcal{F}(\hat{c},y_{\tau,i})$. This term may be rewritten as a standard HMM probability calculation

$$\mathcal{F}(\hat{c},y_{\tau,l}) = t\log(K) - \frac{1}{2\Sigma}\sum_{l=1}^{L}\left(\sum_{\tau=t_l}^{t_{l+1}-1}(\mu_l-y_\tau)^2\right) + \log(p(s^l|s_i,\mathcal{M})) \tag{86}$$

where $\mu_l = \hat{c} + \Delta_l$. If the standard HMM form of duration modelling is used then it is possible to find a new frame intra state allocation that maximises $\mathcal{F}(\hat{c},y_{\tau,l})$ given the current parameter set, by performing a standard Viterbi HMM estimation scheme. This yields $\mathcal{F}(\hat{c},\hat{y}_{\tau,l})$, where $\hat{y}_{\tau,i}$ are obtained using the new intra frame state allocation. So far nothing has been mentioned about what happens to the probability of the complete segment, only that

$$\mathcal{F}(\hat{c},\hat{y}_{\tau,l}) \geq \mathcal{F}(\hat{c},y_{\tau,l}) \tag{87}$$

However, Russell [14] has shown that for a given inter mixture segmentation the value of the estimated mean, $\hat{c}$, that maximises the value of the MAP SHMM segment probability is defined by equation 83. Hence defining

$$\mathcal{H}(\hat{c},y_{\tau,l},\mathcal{M}) = \log(K_c) - \frac{1}{2\Sigma_c}(\hat{c}-\mu_c)^2 + \mathcal{F}(\hat{c},y_{\tau,l}) - \log(K_n) \tag{88}$$

then

$$\mathcal{H}(\hat{c}_n,\hat{y}_{\tau,l},\mathcal{M}) \geq \mathcal{H}(\hat{c},\hat{y}_{\tau,l},\mathcal{M}) \geq \mathcal{H}(\hat{c},y_{\tau,l},\mathcal{M}) \tag{89}$$

where

$$\hat{c}_n = \frac{\mu_c\Sigma + \Sigma_c\left(\sum_{\tau=t_i}^{t_j}\hat{y}_{\tau,l}\right)}{\Sigma + t\Sigma_c} \tag{90}$$

A new complete data set has now been generated that is guaranteed not to decrease the probability of the segment. It is therefore possible to find a local maximum for $\hat{p}(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})$.

As with all iterative schemes it is necessary to make an initial estimate of the intra frame state/mixture allocation. With no prior knowledge, the frames may be assumed to be evenly distributed over all intra states. Given the additional computational overhead of estimating the optimal intra frame state allocation, this initial estimate may be used as the 'best estimate'. If the initial guess is used as the 'best estimate' then it is not necessary to tie all the intra state variances together.

7.4 Approximation to the Multiple Gaussian Intra Mixtures

Using the multiple intra state model it is possible to obtain an approximation to the multiple Gaussian intra mixture model. If an ergodic intra state transition matrix is used, then the multiple intra state model may be mapped onto the multiple intra mixture model. As previously stated

$$p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) = \sum_{s^l} p(s^l|s_i,\mathcal{M})p(\mathcal{Y}_{t_i,t_j}|s_i^c,\mathcal{M}) \tag{91}$$

This expression has the same general form required for multiple Gaussian intra mixture SHMMs. However, it is necessary to tie various parameters in the transition matrix $\mathbf{A}^i$ together to obtain the exact form. Setting

$$a_{ij}^i = w_j \tag{92}$$

where w_j is the weight associated with j^{th} intra mixture [8], yields the correct form. In the previous section a closed expression for $p(\mathcal{Y}_{t_i,t_j}|s_i^c,\mathcal{M})$ have been derived, so it is possible to calculate the full probability for the multiple intra mixture case. The problem with this calculation is that there are no efficient algorithms to calculate the summation over every possible intra state segmentation. In a similar way to the multiple intra state case an approximation may be made where only $\hat{p}(\mathcal{Y}_{t_i,t_j}|s_i^c,\mathcal{M})$ is considered. Using this approximation the method used to complete the data set for the multiple intra state model may be used.

7.5 Multiple Intra State Parameter Estimation

It has already been shown that the multiple intra state case yields identical equations to the single intra state case if the observations, $\mathbf{y}_\tau$ are replaced by the transformed observations $\mathbf{y}_{\tau,l}$, this assumes that the complete data set is known. The re-estimation formulae derived for the standard parameter set $\{\mu_c, \mathbf{\Sigma}_c, \mathbf{\Sigma}\}$ may then be directly applied. It is only necessary to define re-estimation formulae for the intra state delta means, $\mathbf{\Delta}_l$ and intra state transition probabilities, $\mathbf{A}^i$.

Using the auxiliary function $\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})$ previously defined then

$$\begin{aligned}\frac{\partial}{\partial\hat{\Delta}_l}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) &= \sum_s p(\mathbf{Y}_T,\mathbf{s}_T^c|\mathcal{M})\frac{\partial}{\partial\hat{\Delta}_l}\left(\log(p(\mathcal{Y}|s_j^c,\hat{\mathcal{M}}))\right)\\ &= \sum_s p(\mathbf{Y}_T,\mathbf{s}_T^c|\mathcal{M})\left[\sum_{t_l}^{t_{l+1}-1}\frac{(y_\tau-\hat{\Delta}_l)}{\hat{\Sigma}} - \frac{(\mu_c\hat{\Sigma}+\hat{\mu}_s\hat{\Sigma}_c)}{\hat{\Sigma}(t\hat{\Sigma}_c+\hat{\Sigma})}(t_l-t_{l+1}+1)\right]\end{aligned} \tag{93}$$

where

$$\hat{\mu}_s = \sum_{l=1}^{L}\left(\sum_{\tau=t_l}^{t_{l+1}-1}(y_\tau-\hat{\Delta}_l)\right) \tag{94}$$

In re-estimating the state transition matrix $\mathbf{A}^i$, the standard HMM formulae may be used provided that the new complete data set is known. Hence

$$\hat{a}_{ij}^i = \frac{\sum_{\tau=1}^{T-1} p(q_\tau(k,i)|\mathcal{M})p(q_{\tau+1}(k,j)|\mathcal{M})}{\sum_{\tau=1}^{T} p(q_\tau(k,i)|\mathcal{M})} \tag{95}$$

where $\hat{a}^{i}_{ij}$ is the estimate of the element of the intra state transition matrix associated with state S_k of model $\hat{\mathcal{M}}$.

8 Relationship to Bayesian Speaker Adaptation

The SHMM has so far been described in terms of the assumptions behind the model. An alternative viewpoint is to consider the SHMM as an empirical Bayesian approach to speech recognition. The inter distribution may be viewed as the prior probability distribution for the segment mean and the intra distribution the observation distribution. Given this viewpoint, there are close analogies with the Bayesian speaker adaptation work of Gauvain and Lee [10, 5]. If, instead of talking about within segment and between segment variability, between speaker and inter speaker variability are considered, it is simple to see how the SHMM could be applied to the speaker adaptation problem.

8.1 Implementation within the SHMM framework

As previously mentioned in order to implement speaker adaptation within the SHMM framework it is necessary to talk about inter speaker and intra speaker variability. To estimate the output distribution parameters, $\{\mu_c, \Sigma_c, \Sigma\}$, for this model the auxiliary function, $\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})$, is redefined to be

$$\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) = \sum_{s} p(\mathbf{Y}_T, \mathbf{s}_T|\mathcal{M}) \left[\sum_{k=1}^{N} \log(p(\mathcal{Y}_{Sp}, s_k|\hat{\mathcal{M}})) + \log(p(\mathbf{s}_T|\hat{\mathcal{M}}))\right] \tag{96}$$

where $\mathcal{Y}_{Sp}$ is the composite segment of all frames belonging to a particular speaker. Hence the new 'segment', $\mathcal{Y}_{Sp}$, consists of all frames of speaker, Sp, allocated to state S_i. Again the auxiliary function is based on the complete data set, so there is a choice of Viterbi or Baum-Welch re-estimation. Using the above definition the only difference to the standard SHMM is the definition of the 'segment'. Therefore the previous parameter estimation schemes may be used.

8.2 Maximum A-Posterior Estimate of the Mean

The Maximum A-Posteriori (MAP) estimate of the mean is given in Duda and Hart [16]

$$\mu_{MAP} = \frac{\mu_c\Sigma + \mu_s\Sigma_c}{\Sigma + t\Sigma_c} \tag{97}$$

and

$$\Sigma_{MAP} = \frac{\Sigma\Sigma_c}{t\Sigma_c + \Sigma} \tag{98}$$

These are identical to μ_n and Σ_n as defined in appendix A. Given these MAP estimates of the mean, speaker adaptation may be implemented in a variety of forms.

1. Observation level. The recognition stage may be implemented as a standard HMM with parameters $\{\mu_{MAP}, \Sigma\}$. The a-posteriori probabilities are now given by

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{A}, \mathcal{M})) = \sum_{\tau=t_i}^{t_j} \log(\mathcal{N}(y_\tau; \mu_{MAP}, \Sigma)) \tag{99}$$

   where $\mathcal{A}$ are the adaptation data. This expression has been used by Lee [10] for speaker adaptation. However, the methodology for generating the prior distribution parameters $\{\mu_c, \Sigma_c, \Sigma\}$ has not previously been used. This scheme has the advantage of low computational overhead compared to the alternatives.

2. Segment level. The adapted SHMM may be used in the same way as Russell [14] implements segment based recognition. Here the model has observation parameters $\{\mu_{MAP}, \Sigma_{MAP}, \Sigma\}$ and the a-posterior probability is given by

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{A}, \mathcal{M})) = \log(\mathcal{N}(\mu_{MAP'}; \mu_{MAP}, \Sigma_{MAP})) + \sum_{\tau=t_i}^{t_j} \log(\mathcal{N}(y_\tau; \mu_{MAP'}, \Sigma)) \quad (100)$$

where

$$\mu_{MAP'} = \frac{\mu_{MAP}\Sigma + \Sigma_{MAP} \sum_{\tau=t_i}^{t_j} y_\tau}{\Sigma + t\Sigma_{MAP}} \quad (101)$$

This has a computational overhead associated with it compared with the standard HMM, see section 6.

3. Speaker level. The adaptation process is implemented on a speaker level. It should therefore be possible to implement the recognition on a speaker level. The probability should therefore be calculated on the basis of every segment from a particular utterance of a speaker allocated to a particular state. The model output probability parameters are again $\{\mu_c, \Sigma_c, \Sigma\}$ and the probability is given by equation 100. However now

$$\mu_{MAP'} = \frac{\mu_{MAP}\Sigma + \Sigma_{MAP} \sum_{\tau=1}^{T} y_\tau p(q_i(\tau))}{\Sigma + \Sigma_{MAP} \sum_{\tau=1}^{T} p(q_i(\tau))} \quad (102)$$

Implementation in this form requires that the complete segmentation for the whole utterance is known before any probabilities may be calculated. If implemented directly this causes a computational explosion. However, if used in a merge and split style scheme [22] or to reorder an N-Best scheme speaker level adaptation is implementable.

All the above schemes have assumed the complete data set for the adaptation data is known. There are a variety of ways of completing the data set on the adaptation data. The method proposed and implemented by Gauvain [5] uses an iterative *EM* style algorithm. An alternative to this approach, which is not iterative, is to use the decoder of Russell [14]. In this technique the new adapted models are automatically taken into account on a segment basis.

A single Gaussian inter mixture has so far been assumed. If a multiple inter mixture SHMM is to be used, then the new mixture weights, given the adaptation data, must be estimated

$$\begin{aligned} p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{A}, \mathcal{M}) &= \sum_{m=1}^{M} p(\mathcal{Y}_{t_i,t_j}, m|s_i, \mathcal{A}, \mathcal{M}) \\ &= \sum_{m=1}^{M} p(\mathcal{Y}_{t_i,t_j}|s_i, m, \mathcal{A}, \mathcal{M}) p(m|s_i, \mathcal{A}, \mathcal{M}) \end{aligned} \quad (103)$$

where M is the number of inter mixtures.

$$p(m|s_i, \mathcal{A}, \mathcal{M}) = \frac{p(\mathcal{A}|m, s_i, \mathcal{M}) p(m|s_i, \mathcal{M})}{p(\mathcal{A}|s_i, \mathcal{M})} \quad (104)$$

All the above may be calculated, since $p(m|s_i, \mathcal{M}) = c_m$ and the probabilities of the adaptation data may be calculated in one of the three ways detailed here. This 'adaptation' of the mixture weights is not performed in any Bayesian style, it is a Maximum Likelihood estimate of the mixture weights.

8.3 Bayesian Approach

The MAP approach detailed above assumes that the variance on the estimate of the mean is small. This may be in fact be false. The probability is now given by

$$p(\mathcal{Y}_{t_i,t_j}|s_i, \mathcal{A}, \mathcal{M}) \quad = \quad \int \mathcal{N}(\mu; \mu_{MAP}, \Sigma_{MAP}) \prod_{\tau=t_i}^{t_j} \mathcal{N}(y_\tau; \mu, \Sigma)\, d\mu \tag{105}$$

This equation is directly analogous to equation 24 and the analysis previously given may be applied. This probability calculation cannot be implemented on the observation level. For the segment level the above implementation is directly applicable. Again for the speaker level the definition of $\mathcal{Y}_{t_i,t_j}$ is altered to be a composite segment of all the frames within the utterance allocated to a particular model state.

8.4 Discussion

The previous subsections have briefly described how the SHMM may be modified and applied to the process of speaker adaptation. There are limits to its use in speaker adaptation. The adaptation so far described has purely concentrated on the means of the observation distribution. The SHMM as described is limited to this process. The variances and mixture weights may be compensated in a similar style as described by Lee [10]. However, the estimation of the prior parameters is far more complicated. Only single intra-mixture models have been looked at. The multiple intra state model may also be used in a similar way for speaker adaptation.

9 Maximisation Stage of the SHMM

In section 5 various forms for the re-estimation formulae for maximising the parameters of the SHMM given the complete data set were obtained. These are discussed in greater detail here.

9.1 Approximate Solution

If it is assumed that $t\hat{\Sigma}_c \gg \hat{\Sigma}$ then closed forms for the re-estimation of all the parameters of the SHMM have been derived. Hence the implementation of this method is straightforward for both Viterbi and Baum-Welch re-estimation formulae. Similar to standard HMMs, it is only necessary to keep running totals for the denominator and numerator on a per state basis.

9.2 True Maximisation

It is not necessary to make the assumption used in section 9.1, since it is possible to maximise the auxiliary function using standard optimisation techniques, gradient descent, conjugate gradient descent etc. For these optimisation techniques it is generally necessary to store sufficient statistics to calculate the value and gradient, or higher derivative, at any point, if the true local maximum for a given complete data set is to be found. Hence the duration, sum and sum of squares of the feature vector for every possible segment need to be stored, so no sufficient statistics of a fixed dimension are available. However, if Viterbi re-estimation is used, with left to right models, there are at most N segments generated for each utterance. Thus for small training data sets using Viterbi re-estimation it is possible to use this true maximisation procedure.

9.3 Approximate True Maximisation

If there is not enough memory, or Baum-Welch re-estimation is to be used, an alternative form for the maximisation, which does not require sufficient statistics of arbitrary dimension, is needed. If knowledge is restricted to knowing the value, gradient and possibly higher derivatives, of the function at the present estimate of model parameters, then sufficient statistics of a fixed dimension are available. Standard optimisation procedures, such as gradient descent or quick-prop [19], exist for the task where only this limited knowledge is available. These techniques have been used previously for Maximum Mutual Information training of HMMs [15]. In general they work by making a 'best guess' at a new parameter set given the present point in space, the derivatives at that point and possibly previous step information. These statistics may be calculated on a per state basis, giving sufficient statistics of a fixed dimension. Optimisation of this form will converge more slowly than the true maximisation scheme as, after each estimation step, it is not guaranteed to increase the probability. Secondly, depending on the probability surface of the function being optimised, the 'best guess' may or may not be close to the local maximum.

9.4 Parameter Constraint Implementation

It is necessary to enforce constraints on the variances, as these are, by definition, positive. For the approximate solution this is not required as the variances are positive from the re-estimation formulae. However for the other maximisation options the variances must be constrained to be positive. This can be achieved in two ways. The first is to optimise the standard deviation and square the resultant value, guaranteeing that the variance is positive. Alternatively the log of the variance may be maximised. Since the log function increases monotonically, maximising the log of the variance maximises the variance, with the added constraint that the variance is positive, as the exponent is positive when the value of the log is real.

10 Software Implementation

All software was implemented within the framework of the Hidden Markov Model Toolkit (HTK) [21]. The model syntax was modified to incorporate the SHMM features. An example model is given below.

```
<BeginHMM>
 <NumStates> 3 <VecSize> 10 <MFCC_E_D> <nullD> <diagC>
 <StreamInfo> 1 10
 <State> 2 <NumMixes> 2
 <Mixture> 1 0.5
 <NumIntraMixes> 1
    <Mean> 10
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 10
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
    <IntraVariance> 10
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
    <IntraGConst> 1.0
 <Mixture> 2 0.5
 <NumIntraMixes> 2
    <Mean> 10
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
    <Variance> 10
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
    <IntraMixture> 1 0.5
        <IntraMean> 10
          0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
        <IntraVariance> 10
          1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
    <IntraMixture> 2 0.5
        <IntraMean> 10
          0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
```

```
        <IntraVariance> 10
          1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
 <TransP> 3
 0.0 1.0 0.0
 0.0 0.9 0.1
 0.0 0.0 0.0
<EndHMM>
```

Standard HMMs may still be described by the syntax given above.

11 Synthetic Data

In order to test the training algorithms and the Viterbi decoder, artificial data was generated using a set of three SHMMs. A total of 1000 utterances were generated, approximately evenly distributed over the three models, using the underlying assumptions of the SHMM. These were then split into a training set of 500 utterances and a test set of 500.

11.1 Training Procedure

The following training procedure was used to generate the models. The number of emitting states in all cases was set to 3, the same as all the source models.

1. Generate a standard HMM, having the same number of states as the SHMM to be trained. For the modelling process a single Gaussian mixture model, was used initialised using uniform segmentation.
2. Using the segmentation generated by the standard HMM as the initial estimate for the complete data set, the models were updated using a Viterbi style estimation scheme, as described in section 6. The model parameters were estimated using the true maximisation scheme, detailed in section 9.3. Hence it was necessary to store information on a per segment basis. The maximisation was performed using conjugate gradient descent optimisation. The initial start point for the maximisation was obtained from the approximate estimate. For these experiments the transition matrix, $\mathbf{A}$, was not updated and was set to the correct value.

11.2 Results on Synthetic Data

Model	*Number Mixtures (Intra)*	*% Recognition Rate*
Source Model	1 (1)	77.2
HMM	1	51.6
HMM	2	57.4
SHMM	1 (1)	78.2

Table 2: Synthetic Data Performance from SHMM data

From table 2 it can be seen that if the source model is in the form of a SHMM it is not possible to dramatically increase performance by adding additional mixtures to a standard HMM. In fact even if the three mixture standard HMMs are trained on all the data, including the test data, the performance is still only 58.6%. The models for this test were optimised with both Viterbi and Baum-Welch re-estimation. It can also be seen that the training routine performs well, giving a performance only marginally worse than that of the source model.

As a comparison, a set of standard HMMs were then used as the sources. The above procedure was repeated and a set of HMMs and SHMMs were generated.

The performances of all three models are approximately the same. When trained, the SHMM set all the inter variances to low values, the largest inter to intra ratio being 0.07. This agrees with

Model	*Number Mixtures (Intra)*	*% Recognition Rate*
Source Model	1	95.9
HMM	1	96.0
SHMM	1 (1)	95.8

Table 3: Synthetic Data Performance from HMM data

the discussion in section 4 where the HMM is stated to be equivalent to the SHMM when the inter variance is zero.

12 Preliminary Results on TIMIT

12.1 Model Training

The training procedure used for the synthetic data was repeated on various dialect regions of TIMIT [6]. For these experiments the 48 KFL phone set [11] , using a standard folding from the TIMIT labels, was used. In addition to the optimum training, the suboptimal, approximate, training was also implemented, where it was assumed that $t\Sigma_c \gg \Sigma$. The transition matrix $\mathbf{A}$ was not updated in either the standard HMMs or SHMMs, the value being set identically over all models. Three emitting state models were used for all phones. The speech was parametrised using Mel-Frequency Cepstral Coefficient (MFCC) [18] including the energy and delta coefficients. This resulted in a 26 element feature vector.

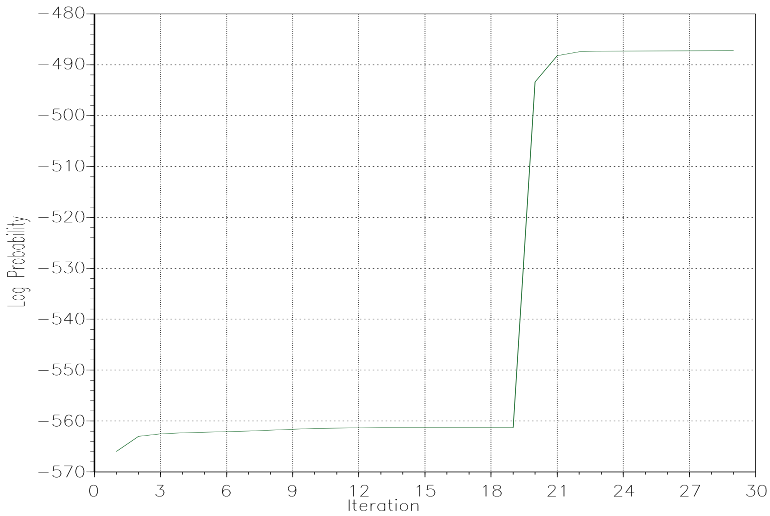

Figure 1: Maximum Likelihood Training of *ih* on Dialect Region 1

Figure 1 shows the log likelihood against training iteration for a single mixture standard HMM from iteration 1 to iteration 19. The segmentation of this standard HMM is used to initialise the training of a single inter, single intra Gaussian distribution SHMM from iteration 20 to 29.

The standard HMM has a maximum log probability of -561.3. This may be compared with the maximum for the SHMM of -487.2. It is worth comparing this with the maximum log probability of a 3 Gaussian mixture standard HMM, which has twice the number of parameters of the SHMM, of -545.2. The SHMM converges to a far higher log probability for the training data than the standard HMM. This is true for all the phones trained and given the fact that the standard HMM is a subset of the SHMM it should always be true provided that there is not a local minima problem with the SHMM training.

12.2 Recognition Performance

The segment based models were tested on subsets of TIMIT. The 48 KFL phone set was folded down to the standard 39 phone set for scoring. In order to reduce the computation time, a maximum possible duration,t_{max}, in any one state was set. For these experiments a maximum duration of 40 frames was used. This was assumed to be sufficient for all models other than the silence model. Hence it was necessary to map multiple observations of silence onto a single observation. All the results given include the SA sentences in both training and testing.

Model	*No Mixtures (Intra)*	*%Correct*	%Accuracy
HMM	1	56.87	50.94
SHMM_{sub}	1 (1)	53.98	47.61
SHMM	1 (1)	55.99	49.87

Table 4: Recognition performance on Dialect Region 1 of TIMIT

The first dialect region examined was dialect region 1. The *sub* subscript indicates that the models were generated using the suboptimal training routine. From the results in table 4, it can be seen that the SHMM does not improve the performance over a standard single mixture HMM. In addition the optimal training was superior to the suboptimal training. Though the recognition results are worse for the SHMM, the log likelihoods for the test sentences were higher than for the standard HMM.

Model	*No Mixtures (Intra)*	*%Correct*	%Accuracy
HMM	1	61.85	56.25
SHMM	1 (1)	58.44	52.12
HMM	2	63.10	57.54
SHMM	2 (1)	61.14	53.51

Table 5: Recognition performance on Dialect Region 2 of TIMIT

Recognition experiments were then performed on dialect region 2, which is approximately twice the size of dialect region 1. The results are shown in table 5. Again the SHMM does not perform as well as the standard HMM.

Model	*No Mixtures (Intra)*	*%Correct*	%Accuracy
HMM	1	60.74	56.09
SHMM	1 (1)	57.08	52.10

Table 6: Recognition performance on the training data for Dialect Region 2 of TIMIT

In addition the models were tested on the training data to see if the SHMM was over training. The results in table 6 indicate that the SHMMs are not over trained.

13 Conclusions

A new acoustic model for speech has been proposed, the SHMM. Both re-estimation formulae and recognition algorithms have been derived for this new model. It has been shown that the standard

HMM, both single and multiple gaussian mixtures, are a subset of the SHMM. For synthetic data the new model has been shown to perform better on a wider set of acoustic waveforms than standard HMMs, even when the standard HMM has over twice the number of parameters. However the performance on real data is disappointing. The most likely reason for the poor performance is the lack of inter segment mean correlation modelling. If independence is assumed on a phone or model level, such as for the SSM, good performance may be achieved. In the SHMM independence is assumed on a segment level, where the typical segment length is less than a phone. The same assumption is used for standard HMMs, however in the SHMM, by making the segment probabilities conditional on the segment mean, the information from intra state variations are weighted far more. In order to overcome this problem, an extension to the standard SHMM has been proposed in the form of the multi intra state SHMMs. Using the multiple intra state model, it is possible to assume independence on a phone or model level whilst having multiple states within the model. Work on this style of model is currently under way.

Acknowledgement

The authors are grateful to Dr. M.J. Russell for many helpful discussions about the work presented here.

References

[1] Dempster A.P., Laird N.M., and Rubin D.B. Maximum liklihood from incomplete data via the EM algorithm. In *J. Roy. Stat. Soc.*, volume 39, pages 1–38, 1977.

[2] Wellekens CJ. Explicit time correlation in hidden Markov models for speech recognition. In *Proceedings ICASSP*, pages 384–386, 1987.

[3] Gauvain J and Lee C. Improved acoustic modelling with bayesian learning. In *Proceedings ICASSP*, pages 481–484, 1992.

[4] Junqua J and Anglade Y. Acoustic and perceptual studies of lombard speech: Application to isolated-word automatic speech recognition. In *Proceedings ICASSP*, pages 841–844, 1990.

[5] Gauvain J.L. and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. In *IEEE Transactions SAP*, 1993. To be published.

[6] Garofolo J.S. *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database.* National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

[7] Ponting K.M. and Peeling S.M. The use of variable frame rate analysis in speech recognition. In *Computer Speech and Language*, volume 5, pages 169–179, 1991.

[8] Liporace LA. Maximum likelihood estimation for multivariate observation of markov sources. In *IEEE Transactions on Information Theory*, pages 729–734, 1982.

[9] Baum L.E., Petrie T., Soules G., and Weiss N. A maximisation technique occuring in the statistical analysis of probabilistic functions of Markov chains. In *Ann. of Math. Stat.*, volume 41, pages 164–171, 1970.

[10] Chin-Hui Lee and Juang B.H. A study on speaker adaptation of the parameters of continuous density hidden Markov models. In *IEEE Transactions ASSP*, volume 39, pages 806–814, 1991.

[11] Kai-Fu Lee. *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers, Boston, 1989.

[12] Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, February 1989.

[13] Ostendorf M. and Roukos S. A stochastic segment model for phoneme-based continuous speech recognition. In *IEEE Transactions ASSP*, volume 37, pages 1857–1869, 1989.

[14] Russell MJ. A segmental statistical model for speech pattern recognition. In *Proceedings IOA*, volume 14, pages 503–510, 1992.

[15] Brown P.F. The acoustic-modelling problem in automatic speech recognition. Technical report, IBM Thomas J. Watson Research Centre, August 1987.

[16] Duda R.O. and Hart P.E. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 1974.

[17] Furui S. Speaker independent isolated word recognition using dynamic features of speech spectrum. In *IEEE Transactions ASSP*, pages 52–59, 1986.

[18] Davis S.B. and Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions ASSP*, volume 28, pages 357–366, 1980.

[19] Fahlman S.E. An empirical study of learning speed in back-propagation networks. Technical report, School of Computer Science, Carnegie Mellon University, 1988.

[20] Levinson SE. Continuous speech recognition by means of acoustic/phonetic classification obtained from a hidden Markov model. In *Proceedings ICASSP*, pages 93–96, 1987.

[21] Young SJ. *HTK: Hidden Markov Model Toolkit V1.2 Reference Manual.* Cambridge University Engineering Department Speech Group, 1990.

[22] Digalakis V., Ostendorf M., and Rohlicek J. Fast algorithms for phone classification and recognition using segment-based models. *IEEE Transactions ASSP*, 40(12):2885–2896, December 1992.

[23] Digalakis V., Ostendorf M., and Rohlicek J.R. Improvements in the stochastic segment model for phoneme recognition. In *Proceedings of the DARPA Workshop*, October 1989.

A Segmental HMM Likelihood Derivation

The aim is to make the probability independent of μ.

$$p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M}) = \int\limits_{\mathcal{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_c|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mu-\mu_c)\mathbf{\Sigma}_c^{-1}(\mu-\mu_c)^T) \prod_{\tau=t_i}^{t_j}\frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{y}_\tau-\mu)\mathbf{\Sigma}^{-1}(\mathbf{y}_\tau-\mu)^T)\,d\mu \tag{106}$$

Taking the log probability

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) = t\log(K)+\log(K_c)+\log\left[\int\limits_{\mathcal{R}^n}\exp(-\frac{1}{2}\mathcal{F}(\mathcal{Y}_t,\mathcal{M},\mu))\,d\mu\right] \tag{107}$$

where

$$\begin{aligned}
\mathcal{F}(\mathcal{Y}_{t_i,t_j},\mathcal{M},\mu) &= (\mu-\mu_c)\mathbf{\Sigma}_c^{-1}(\mu-\mu_c)^T+\sum_{\tau=t_i}^{t_j}(\mathbf{y}_\tau-\mu)\mathbf{\Sigma}^{-1}(\mathbf{y}_\tau-\mu)^T && (108)\\
K &= \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} && (109)\\
K_c &= \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_c|^{\frac{1}{2}}} && (110)
\end{aligned}$$

and $t = t_j - T_i + 1$. Expanding the terms in $\mathcal{F}(\mathcal{Y}_{t_i,t_j},\mathcal{M},\mu)$

$$\begin{aligned}
\mathcal{F}(\mathcal{Y}_{t_i,t_j},\mathcal{M},\mu) &= \mu\mathbf{\Sigma}_c^{-1}\mu^T-2\mu\mathbf{\Sigma}_c^{-1}\mu_c^T+\mu_c\mathbf{\Sigma}_c^{-1}\mu_c^T+\sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau\mathbf{\Sigma}^{-1}\mathbf{y}_\tau^T-2\mu\mathbf{\Sigma}^{-1}\sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau^T+t\mu\mathbf{\Sigma}^{-1}\mu^T\\
&= \mu\left(\mathbf{\Sigma}_c^{-1}+t\mathbf{\Sigma}^{-1}\right)\mu^T-2\mu\left(\mathbf{\Sigma}_c^{-1}\mu_c^T+\mathbf{\Sigma}^{-1}\mu_s^T\right)+\mu_c\mathbf{\Sigma}_c^{-1}\mu_c^T+\sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau\mathbf{\Sigma}^{-1}\mathbf{y}_\tau^T \quad (111)
\end{aligned}$$

where μ_s is defined as $\mu_s^T = \sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau^T$. Let

$$\begin{aligned}
\mathbf{\Sigma}_n^{-1} &= \mathbf{\Sigma}_c^{-1}+t\mathbf{\Sigma}^{-1} && (112)\\
\mu_n^T &= \mathbf{\Sigma}_n\left(\mathbf{\Sigma}_c^{-1}\mu_c^T+\mathbf{\Sigma}^{-1}\mu_s^T\right) && (113)
\end{aligned}$$

and expanding

$$\begin{aligned}
(\mu-\mu_n)\mathbf{\Sigma}_n^{-1}(\mu-\mu_n)^T &= \mu\mathbf{\Sigma}_n^{-1}\mu^T-2\mu\mathbf{\Sigma}_n^{-1}\mu_n^T+\mu_n\mathbf{\Sigma}_n^{-1}\mu_n^T && (114)\\
&= \mathcal{F}(\mathcal{Y}_{t_i,t_j},\mathcal{M},\mu)-\left(\mu_c\mathbf{\Sigma}_c^{-1}\mu_c^T+\sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau\mathbf{\Sigma}^{-1}\mathbf{y}_\tau^T\right)+\mu_n\mathbf{\Sigma}_n^{-1}\mu_n^T
\end{aligned}$$

Noting that

$$\int\limits_{\mathcal{R}^n}\exp(-\frac{1}{2}(\mu-\mu_n)\mathbf{\Sigma}_n^{-1}(\mu-\mu_n))^T\,d\mu = (2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}_n|^{\frac{1}{2}} = \frac{1}{K_n} \tag{115}$$

Hence

$$\log(p(\mathcal{Y}_{t_i,t_j}|s_i,\mathcal{M})) = t\log(K)+\log(K_c)-\log(K_n)-\frac{1}{2}\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) \tag{116}$$

where

$$\mathcal{G}(\mathcal{Y}_{t_i,t_j},\mathcal{M}) = \mu_c\mathbf{\Sigma}_c^{-1}\mu_c^T+\sum_{\tau=t_i}^{t_j}\mathbf{y}_\tau\mathbf{\Sigma}^{-1}\mathbf{y}_\tau^T-\mu_n\mathbf{\Sigma}_n^{-1}\mu_n^T \tag{117}$$

B Proof of Maximisation for Approximate Solution

It is necessary to prove that, given the assumption $t\Sigma_c \gg \Sigma$, there is only one turning point of $\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})$ and that turning point is a maximum.

To prove that only one maximum exists it is sufficient to show that all the coefficients of the differential are positive. Hence there may be only one point at which the equation equals zero if the equation is linear in the variable of interest.

B.1 Proof for μ_c

It has already been shown in section 5 that for $\hat{\mu}_c$ the function is strictly concave.

B.2 Proof for Σ_c

$$\frac{\partial}{\partial\hat{\Sigma}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) \quad = \quad \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[\frac{\frac{1}{t}^2(\hat{\mu}_c t-\mu_s)^2-\hat{\Sigma}_c}{2\hat{\Sigma}_c^2}\right] \tag{118}$$

It is necessary to show that there is only one turning point for the above equation. The denominator, $2\hat{\Sigma}_c^2$, may be ignored as this is independent of the segmentation, so the expression is linear in Σ_c. Examining the coefficients

$$\frac{1}{t^2}(\hat{\mu}_c t-\mu_s)^2 \geq 0 \tag{119}$$

and the probability by definition is positive. Hence provided that for some segmentation $p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M}) > 0$ for the state of interest, there is only one turning point as defined in equation 55.

To show this is a maximum equation 118 is differentiated

$$\frac{\partial^2}{\partial^2\hat{\Sigma}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) \quad = \quad \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[-\frac{1}{\hat{\Sigma}_c^3}\left[\frac{1}{t^2}(\hat{\mu}_c t-\mu_s)^2-\hat{\Sigma}_c\right]-\frac{1}{2\hat{\Sigma}^2}\right] \tag{120}$$

At the turning point described in equation 55 the above expression may be reduced to

$$\frac{\partial^2}{\partial^2\hat{\Sigma}_c}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}}) \quad = \quad \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\left[-\frac{1}{2\hat{\Sigma}^2}\right] \tag{121}$$

As the variance is positive,by definition, this expression is negative.

B.3 Proof for Σ

$$\frac{\partial}{\partial\hat{\Sigma}}\mathcal{Q}(\mathcal{M},\hat{\mathcal{M}})$$
$$= \sum_s p(\mathbf{Y}_T,\mathbf{s}_T|\mathcal{M})\frac{1}{2\hat{\Sigma}^2}\left[\sum_{\tau=t_i}^{t_j} y_\tau^2-(t-1)\hat{\Sigma}-\frac{\mu_s^2}{t}\right] \tag{122}$$

Again the denominator, $2\hat{\Sigma}_c^2$ is ignored and examining the coefficients

$$\sum_{\tau=t_i}^{t_j} y_\tau^2 \quad \geq \quad \frac{\mu_s^2}{t} \tag{123}$$

and $t \geq 1$. The probability is also positive. Hence provided that some segment is longer than one sample, and not all elements in that are identical, then there is only one turning point.

Differentiating equation 122

$$\begin{aligned}
\frac{\partial^2}{\partial^2 \hat{\Sigma}} \mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) &= \sum_s p(\mathbf{Y}_T, \mathbf{s}_T | \mathcal{M}) \left[-\frac{1}{\hat{\Sigma}^3} \left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - (t-1)\hat{\Sigma} - \frac{\mu_s^2}{t} \right] - \frac{(t-1)}{2\hat{\Sigma}^2} \right] \\
&= \sum_s p(\mathbf{Y}_T, \mathbf{s}_T | \mathcal{M}) \left[-\frac{1}{\hat{\Sigma}^3} \left[\sum_{\tau=t_i}^{t_j} y_\tau^2 - \frac{1}{2}(t-1)\hat{\Sigma} - \frac{\mu_s^2}{t} \right] \right] \qquad (124)
\end{aligned}$$

Again examining the turning point described in equation 57

$$\frac{\partial^2}{\partial^2 \hat{\Sigma}} \mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = \sum_s p(\mathbf{Y}_T, \mathbf{s}_T | \mathcal{M}) \left[-\frac{(t-1)}{2\hat{\Sigma}^2} \right] \qquad (125)$$

As $t \geq 1$ and the variance is positive, the above expression is negative.