Dependence of GMM Adaptation on Feature Post-Processing for Speaker Recognition

Robbie Vogt, Jason Pelecanos and Sridha Sridharan

Speech and Audio Research Laboratory,
Queensland University of Technology
GPO Box 2434, George St, Brisbane, AUSTRALIA, 4001.
{r.vogt, j.pelecanos, s.sridharan}@qut.edu.au

Abstract

This paper presents a study on the relationship between feature post-processing and speaker modelling techniques for robust text-independent speaker recognition. A fully coupled target and background Gaussian mixture speaker model structure is used for hypothesis testing in this speaker model based recognition system. Two formulations of the *Maximum a Posteriori* (MAP) adaptation algorithm for Gaussian mixture models are considered. We contrast the standard single iteration adaptation algorithm to adaptation using multiple iterations. Three post-processing techniques for cepstral features are considered; feature warping, cepstral mean subtraction (CMS) and RelAtive SpecTrA (RASTA) processing. It is shown that the advantage gained through iterative MAP adaptation is dependent on the parameterisation technique used. Reasons for this dependency are discussed.

1. Introduction

Effectively representing speaker specific information is crucial to the task of speaker recognition. To this end, the development of automatic, text-independent speaker recognition systems has seen research into improving both the extraction of robust speaker features form speech and the modelling of these characteristics.

Advancements in feature extraction have lead to cepstral-based methods such as mel-frequency cepstral coefficients (MFCC) with post-processing techniques to compensate for noise and channel distortion. In this work, we examine the use of MFCCs with RelAtive SpecTrA (RASTA) [1] processing, cepstral mean subtraction (CMS) [2] and feature warping [3] for channel and partial handset compensation. RASTA processing is designed to bandpass filter the time trajectories of each cepstral feature stream while cepstral mean subtraction removes the DC component of the cepstral feature streams. Feature warping performs a short-term marginal Gaussianisation of the cepstral stream.

For those speaker model representations using probability density functions, robust speaker model estimation has seen the trend of *Maximum Likelihood* (ML) estimation methods extend to *Maximum A Posteriori* (MAP) approaches. Estimation of probability density functions under the Maximum Likelihood criterion determines the parametric density through statistics obtained from training data only. MAP distribution estimation techniques incorporate prior knowledge of the distribution of the speaker model parameters in addition to the speech information provided by the speaker during enrollment.

A MAP adaptation approach [4] was successfully introduced to the speech recognition problem. Following this, a number of state-of-the-art text-independent speaker recognition systems [5, 6] used a form of adapted background Gaussian mixture modelling (GMM) using MAP adaptation. In [5], Reynolds derived an algorithm for a coupled target and background speaker model, and evaluated test utterances according to the expected frame-based log-likelihood ratio. For speaker modelling, the parameters of the Universal Background Model (UBM) are taken as a base model and are adjusted toward the speech of the target speaker. This has the advantages of prior information being tied into the modelling process and numeric stability of the parameters of the client speaker model. In this way, the robustness of the speaker models can be improved over the ML estimate.

In this MAP adaptation study, we refer to one of the more successful algorithms [5] and the assumptions that are placed on how the model is adjusted toward the target speaker. We examine an extended implementation of this algorithm that accounts for model parameter dependencies not considered with the standard implementation. The extended technique, which utilises the Expectation-Maximisation (E-M) algorithm [7] for model convergence, is shown to be a simple iterative extension to the standard algorithm. Thus, we discuss an implementation of the MAP algorithm that represents a fully coupled adaptation from the background speaker model with an iteratively determined target model.

We also investigate issues relating to interaction between the two MAP approaches and the different feature sets. This may identify which MAP adaptation algorithm is optimal for particular parameterisation schemes due to sparseness of model mixture components.

Section 2 discusses the general concept of MAP adaptation, MAP as it applies to Gaussian mixture models and the differences between the standard algorithm and its iterative form. Section 3 contrasts the feature post-processing techniques contrasted in this study. Section 4 presents experimental results and a discussion of the adaptation procedures and the parameterisation types with the conclusions in Section 5.

2. MAP Adaptation for Speaker Recognition

2.1. MAP Adaptation Concept

In many direct probability density estimation schemes, the set of speaker features are solely used to determine the speaker model parameters based on the Maximum Likelihood criterion. Thus,

given an utterance $\boldsymbol{X}$ from a speaker and that the parameters of a speaker model are represented by λ, the aim is to determine the parameters of the speaker model λ, that directly maximise the likelihood.

$$\lambda_{ML} = \arg\max_{\lambda} p(\boldsymbol{X}|\lambda) \quad (1)$$

Alternatively, the Bayesian inference approach is to select the speaker model parameters to maximise the posterior likelihood of the model given the speaker features and the model parameter distribution. The model parameters are optimised according to *Maximum A Posteriori* estimation.

$$\lambda_{MAP} = \arg\max_{\lambda} p(\lambda|\boldsymbol{X}) = \arg\max_{\lambda} \frac{p(\boldsymbol{X}|\lambda)p(\lambda)}{p(\boldsymbol{X})} \quad (2)$$

Since the likelihood of $p(\boldsymbol{X})$ is independent of the speaker model parameters, the denominator in Equation 2 can be ignored. Given a non-informative prior distribution of the model parameters, specified by $p(\lambda) = constant$, the MAP solution is equivalent to the Maximum Likelihood estimation mentioned earlier. The process of MAP estimation allows for prior information about the model parameters to be tied into the target speaker model.

2.2. MAP Estimation of Gaussian Mixture Models

For this result, let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ be a sample of T independent and identically distributed vector observations from an N-component multivariate Gaussian mixture density of dimension D. The joint density is given in Equation 3.

$$p(\boldsymbol{X}|\lambda) = \prod_{t=1}^{T} \sum_{i=1}^{N} w_i g(\boldsymbol{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (3)$$

The speaker model is described by the mixture component weights w_i, means $\boldsymbol{\mu}_i$ and diagonal covariances $\boldsymbol{\Sigma}_i$. ie. $\lambda = \{\{w_1, w_2, \ldots, w_N\}, \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_N\}, \{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_N\}\}$. The density of a sample from a Gaussian distribution is given in Equation 4. (Here and throughout the paper, $(')$ represents the vector or matrix transpose.)

$$g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right\} \quad (4)$$

For MAP adaptation, no *sufficient statistic* of a fixed dimension exists for the parameter set λ. For the purpose of a simplified presentation of this work only the mixture component means will be adapted using prior information. Work by Reynolds, et al. [5] also supports the notion that the more useful parameters to adjust are the mixture component means. Given that the mixture component means will be adapted, the prior density is assumed Gaussian and is given by Equation 5.

$$g(\boldsymbol{\mu}_i|\Theta_i) \propto \exp\left\{-\frac{\tau_i}{2}(\boldsymbol{\mu}_i - \boldsymbol{m}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{m}_i)\right\} \quad (5)$$

where $\Theta_i = \{\tau_i, \boldsymbol{m}_i\}$ are the set of hyperparameters with $\tau_i > 0$ and $\boldsymbol{m}_i$ is a D-dimensional vector. In this work, the relevance factor, τ_i, for each mixture component is set to τ. For the prior information, assuming independence between the parameters of the individual Gaussian mixture components, the joint prior density of the speaker model mean vector parameters λ is given.

$$p(\lambda) = \prod_{i=1}^{N} g(\boldsymbol{\mu}_i|\Theta_i) \quad (6)$$

The MAP solution is then solved by maximising $p(\lambda)p(\boldsymbol{X}|\lambda)$. The Expectation-Maximisation algorithm [7] is used to maximise the joint likelihood of this function. It is achieved by maximising an auxilliary function whereby given an old model parameter vector estimate $\hat{\lambda} = \{\hat{\boldsymbol{\mu}}_i\}$, the likelihood of the new parameter estimate $\lambda = \{\boldsymbol{\mu}_i\}$ is equal to or better than the old estimate. The E-M result for the mean adaptation process is given.

$$\boldsymbol{\mu}_i = \frac{\tau_i \boldsymbol{m}_i + \sum_{t=1}^{T} c_{it} \boldsymbol{x}_t}{\tau_i + \sum_{t=1}^{T} c_{it}} \quad (7)$$

In this formulation, the *a posteriori* probability for the Gaussian mixture component i given the observation $\boldsymbol{x}_t$ and the current model estimate $\hat{\lambda}$ is specified.

$$\begin{aligned} c_{it} &= Pr(i|\boldsymbol{x}_t, \hat{\lambda}) \\ &= \frac{w_i g(\boldsymbol{x}_t|\hat{\boldsymbol{\mu}}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{N} w_j g(\boldsymbol{x}_t|\hat{\boldsymbol{\mu}}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \quad (8)$$

Thus, the MAP estimation procedure for GMMs is iterative. The process is initialised by setting the old estimate of the model parameters $\hat{\lambda}$ to λ_0. The initial model λ_0 may be determined by using a close speaker model to seed the procedure or by the use of a general speaker model derived from a GMM trained on a large quantity of diverse speech. The MAP algorithm is then used to determine a better estimate λ, from the old estimate $\hat{\lambda}$. At the end of each iteration, $\hat{\lambda}$ is set to λ. The process is iterated until a required convergence criteria is achieved.

In many established speaker recognition systems, the hyperparameters may be derived from a universal speaker model of generally higher mixture component order trained from a vast quantity of diverse speech. One procedure, proposed by Reynolds [8, 5], utilises a universal background speaker model to derive the relevant adapted speaker models. Thus, the prior distribution means are set to the UBM component means (ie. $\boldsymbol{m}_i = \boldsymbol{\mu}_i^{ubm}$). In previous work, adaptation was proposed as a *single adaptation step* whereas this work uses the *iterative solution*.

2.3. Iterative and Standard MAP Adaptation

There are a number of factors to consider in contrasting one of the standard MAP approaches to its iterative form. The standard MAP technique is simply a single iteration of Equation 7 while the E-M based result is iterative. The iterative version of this result allows for the variation of mixture component means to become dependent not only on previous iterations but also on other components to further refine the MAP estimate. Alternatively, the single-iteration approach assumes that the mixture component means vary in a completely indepenedent manner, thus only a single iteration is required to find the MAP solution. This assumption is not always beneficial and the sparsity of the features used may determine the appropriateness of either MAP technique.

2.4. Classification with Fully Coupled Speaker Modelling

For classification we decide between two speaker classes identified as the target speaker and the non-target (UBM) speaker

set specified by models, λ_{target} and λ_{ubm}. Given a test utterance $\boldsymbol{X}$, the joint likelihood ratio may be determined. However, this is typically not a robust estimate for a target speaker model closeness measure, since the observations are neither independent or identically distributed. A more robust measure for speaker verification is the expected frame-based log-likelihood ratio measure.

$$E[LLR(\boldsymbol{x})] = E\left[\log \frac{p(\boldsymbol{x}|\lambda_{target})}{p(\boldsymbol{x}|\lambda_{ubm})}\right] = \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{p(\boldsymbol{x}_t|\lambda_{target})}{p(\boldsymbol{x}_t|\lambda_{ubm})}\right) \quad (9)$$

This type of testing structure in concerto with MAP adaptation, with coupled target and background speaker model components, is an effective method of performing speaker recognition. A significant advantage of a fully coupled system is that the coupling enables discrimination between regions of space that the GMM has learned from the training speech. Consequently, if there are no adaptation observations in the regions nearby a mixture component, the mixture component will remain unadapted. Conversely, mixture components near training observations will be adjusted toward the speech data. With this type of scoring structure, test observations that are scored against unadapted mixture components will contribute toward a zero expected log-likelihood ratio, whilst adapted regions will be more discriminative.

3. Feature Post-Processing Techniques

Cepstral features, and in particular MFCCs, provide an effective method for extracting speaker specific information from speech. However, cepstral features are susceptible to degradation under noisy and mismatched conditions, such as those experienced in telephony environments. To compensate for this, post-processing techniques are commonly applied. We will highlight three distinct post-processing approaches in this section and will subsequently examine the effect each has on MAP adaptation.

Cepstral mean subtraction (CMS) [2] is one of the more widely used methods of compensating for stationary linear channels. It is applied to a speech segment by subtracting the mean value of each cepstral feature stream from all features in that stream. This method arises from a signal processing approach, as CMS is equivalent to performing convolution of the time-signal by an estimate of the inverse of the linear channel. While CMS is an effective method of removing channel distortion it also removes some speaker specific information, which leads to degraded performance for clean speech. Also, CMS does not account for the distortion introduced by additive noise.

RelAtive SpecTrA (RASTA) [1] essentially applies a bandpass filter to each stream of log-filterbank or cepstral coefficients. This filter is designed to suppress spectral components that are detrimental to speech and speaker recognition. Hence, the approach adopted by RASTA processing is to utilise knowledge of physical constraints of the speech production system to emphasise the speech-like content of the modulation spectrum.

In the presence of additive noise and channel distortion the distribution of log-energy based cepstral features over time undergoes a nonlinear distortion. Feature warping [3] was designed to compensate for this nonlinearity by remapping the distribution of a feature stream to a target distribution through cumulative distribution function matching. In the typical case of a standard normal target distribution, this can be interpreted as short-term marginal Gaussianisation of each cepstral feature stream. Even with a suboptimal normal target distribution, a significant performance improvement is attained through this mapping. This result indicates that there is more speaker specific information in the relative positions of components in a mixture model than their absolute positions.

While only CMS explicitly attempts to compensate for linear channel effects, it is interesting to note that all three techniques effectively compensate for these effects through removing the DC component of cepstral features.

4. Experiments

The recognition system used in this study utilises fully coupled GMM-UBM modelling using MFCC features with delta coefficients, as described in [3]. We evaluate the use of multiple MAP iterations for three different parameterisation enhancement techniques. For this evaluation, the NIST 1999 Speaker Recognition Evaluation database was used. (For further information see [6]). This database includes a collection of 230 male and 309 female target speakers, each providing approximately two minutes of enrollment speech. There are 1448 male and 1972 female test segments of up to one minute in length.

Figure 1 presents the effect of the number of mixture components, parameterisation method and the type of MAP adaptation on speaker recognition performance according to the minimum Detection Cost Function (DCF) criterion [6] used in the 1999 NIST evaluation. The minimum DCF is defined as the associated cost at which the weighted sum of the miss and false alarm probabilities, for an ensemble of test speech segments, is a minimum.

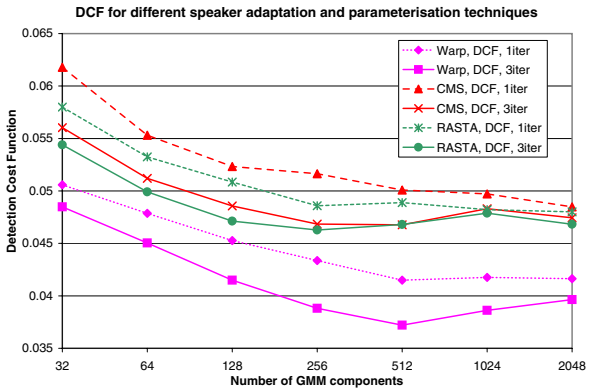

Figure 1: *Plot of Detection Cost versus the GMM order for different parameterisations and adaptation approaches (1 and 3-iteration) using all NIST 1999 male tests.*

With reference to the results in Figure 1, the DCF error rates significantly improve when the multiple iteration MAP is performed instead of the basic algorithm for the NIST 1999 speech corpus. In addition, the extended MAP procedure tends to reach an optimal error rate using slightly fewer Gaussian mixture components. In this evaluation, feature warping is an improvement on both RASTA and CMS channel compensation techniques. However, under normalised handset and test segment conditions, the improvement for warping was not as pronounced.

Following from this discussion, Figure 2 shows the same systems evaluated at the Equal Error Rate (EER) operating region. Interestingly, the performance of the multi-iteration

MAP approach is sub-optimal to the standard algorithm for RASTA and CMS processing for 128/256 mixture components and above. In contrast to this result, the feature warping technique introduces an improvement across the range of model orders for multiple MAP iterations. This presents the issue of why, for multiple iterations, feature warping improves in performance at the EER operating point while RASTA and CMS degrade. Possible reasons for this result may be model overtraining, the coupled target and background model nature of the GMM-UBM system and possibly the sparseness of the speaker feature space attributed to the type of parameterisation.

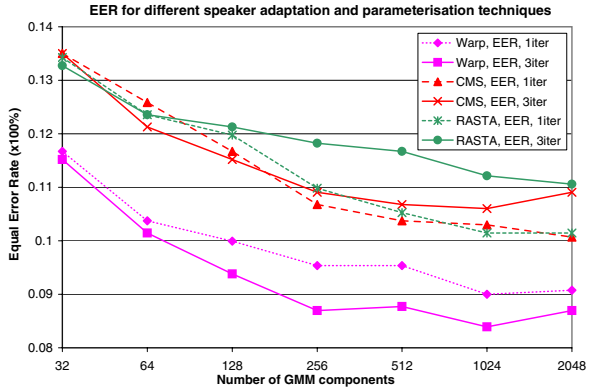

Figure 2: *Plot of EER versus the GMM order for different parameterisations and adaptation approaches (1 and 3-iteration) using all NIST 1999 male tests.*

In addressing the issue of the sparseness of the feature space, the average inter-component distance of each background model was measured using the Bhattacharyya [9] and Kullback-Leibler [10] distances (Figure 3). It was observed that the Bhattacharyya distance for feature warping background models was significantly smaller than either the RASTA or CMS feature processing models. Consequently, for feature warping, if the UBM mixture component distributions are more overlapped, the use of multiple MAP iterations becomes essential to accommodate for the mixture component interactions. The Kullback-Leibler distance indicated a similar trend with feature warping producing more overlapping distributions than the other techniques.

5. Conclusions

Experiments on the NIST 1999 Speaker Recognition Evaluation indicate that iterative MAP adaptation can be an effective method for improving speaker recognition performance. In particular, DCF error rates improved for feature warping, RASTA processing and cepstral mean subtraction. In contrast, the equal error rates improved for feature warping and degraded for RASTA and CMS at higher mixture orders. It was proposed that feature warping, using multiple MAP iterations, improved in a consistent manner because of the tightly clustered nature of the Gaussian modes represented in the background model. Iterative MAP adaptation, within the E-M algorithm theory, accounts for the mixture component interactions when attempting to find the MAP solution. Single step adaptation assumes sparse, independent, Gaussian clusters. In conclusion, the theory and results indicate that the performance of the adaptation procedures can be affected by the sparseness of the speech feature space inherent to the type of parameterisation.

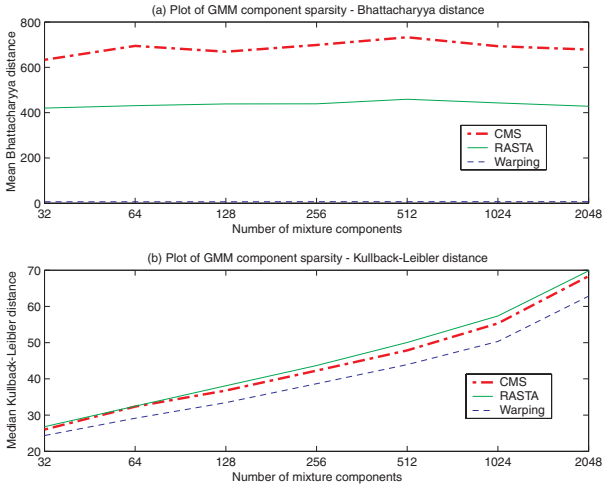

Figure 3: *Plot of UBM average inter-component (a) Bhattacharyya and (b) Kullback-Leibler distances versus GMM order for different parameterisations.*

6. Acknowledgements

This work was supported by the Australian Defence Science and Technology Organisation.

7. References

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254–272, 1981.

[3] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 213–218, 2001.

[4] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[5] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[6] National Institute of Standards and Technology, "NIST speech group website." http://www.nist.gov/speech, 2003.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, vol. 2, pp. 963–966, 1997.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, California, USA: Academic Press, 1990.

[10] K. Shinoda and C. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 276–287, 2001.