Separate MAP Adaptation of GMM Parameters for Forensic Voice Comparison on Limited Data

Chee Cheun Huang[1,2]
[1]*School of Electrical Engineering and Telecommunications*
The University of New South Wales
Sydney, NSW 2052, Australia
chee.huang@student.unsw.edu.au

Julien Epps[1,2] and Ewald Enzinger[1,2]
[2]*National ICT Australia (NICTA)*
Australian Technology Park
Sydney, NSW 1430, Australia
j.epps@unsw.edu.au,
e.enzinger@student.unsw.edu.au

***Abstract*—Automatic forensic voice comparison (FVC) studies have often employed Gaussian Mixture Model – Universal Background Model (GMM-UBM) modeling based on mean-only maximum a posteriori (MAP) adaptation or full MAP adaptation with little consideration of other variants of MAP adapted configurations. Our study indicates that an FVC system improvement in log-likelihood ratio cost (C_{llr}) of up to 6.8% can be achieved via fusion of other MAP adapted configurations such as variance-only and weight-only adaptations. We also demonstrate a novel adaptation and fusion strategy named Separate MAP (SMAP) that yielded a substantial FVC performance improvement in C_{llr}, up to 24.2%, and which is more robust under limited data conditions compared with the conventional mean-only or full MAP adaptation. The fusion strategy involves fusing multiple MAP adapted GMM sub-configurations where in each of these GMM sub-configurations only a small subset of the total number of GMM parameters are MAP adapted separately based on the same speaker-specific training data.**

I. INTRODUCTION

The Gaussian Mixture Model – Universal Background Model (GMM-UBM) [4-6] is a modelling technique that has been used extensively in forensic voice comparison and speaker recognition. It is common to first train such a GMM-UBM covering the entire acoustic space for all speaker-independent acoustic classes using a large database via the expectation-maximization (EM) algorithm [7-8]. A subsequent step then involves tuning these acoustic classes, based on the speaker-dependent training data, via maximum a posteriori (MAP) adaptation, to produce speaker models. For a UBM with M mixture components comprising full covariance matrices and D-dimensional mean vectors, this involves the re-estimation of $M \times D$ means, $M \times D^2$ covariances, and M weight parameters.

Often only the mean of the Gaussians are adapted, while the weights and covariance matrices of the Gaussians are not adapted from the UBM [5]. An advantage of this 'mean-only' adaptation, particularly when relatively small amounts of speaker data are available, is that fewer parameters (only D per mixture) need to be estimated using the speaker-dependent data. In forensic casework, the offender and/or suspect speech recordings may be extremely short. Hence, it would be desirable to estimate even fewer parameters if possible.

A Gaussian mixture model is a weighted sum of M component densities.

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} w_i \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \quad (1)$$

Here $\mathbf{x}$ is a D-dimensional feature vector, w_i with $i = 1,..,M$ are the mixture weights, and the mean vector and covariance matrix for the ith mixture are denoted as $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ respectively.

For the purposes of definition, we repeat here details of the MAP adaption process [5] in Algorithm 1, in which K denotes the total number of MAP iterations. While full adaptation is perfectly valid, we consider only diagonal covariance matrices as it has been shown empirically that diagonal covariance GMMs can outperform full covariance GMMs [5].

Train $\lambda_{UBM} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}$, and use this to initialise the speaker-specific model λ_{sp}.

Given speaker-specific training vectors $\mathbf{x}_t$ with $t \in \{1,2,\ldots,T\}$

for k = 1 to K

Calculate:

$$\Pr(i|\mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j=1}^{M} w_j p_j(\mathbf{x}_t)} \quad (2)$$

$$n_i = \sum_{t=1}^{T} \Pr(i|\mathbf{x}_t) \quad (3)$$

$$E_i(\mathbf{x}) = \frac{1}{n_i}\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t)\,\mathbf{x}_t \quad (4)$$

$$E_i(\mathbf{x}^2) = \frac{1}{n_i}\sum_{t=1}^{T} \Pr(i|\mathbf{x}_t)\,\mathbf{x}_t^2 \quad (5)$$

Update λ_{sp} as follows:

$$\hat{\boldsymbol{\mu}}_i \leftarrow \alpha_i E_i(\mathbf{x}) + (1-\alpha_i)\boldsymbol{\mu}_i \quad (6)$$

End

Algorithm 1: Mean-only MAP adaptation [5].

In Algorithm 1, the data-dependent adaptation coefficient $\alpha_i = \frac{n_i}{n_i + r}$ is defined in terms of a relevance factor r (constant), and the probabilistic occupation count of the speaker dependent training data associated with the ith mixture component n_i .

The popularity of the GMM-UBM modeling in automatic speaker recognition has also disseminated to the area of forensic acoustics, where it has become the more commonly used method for modeling and likelihood ratio calculation in

automatic FVC systems, e.g. [1-3, 9-12]. In FVC, data from speakers from the relevant population (commonly known as the background database) are used to the train the UBM, which can be interpreted as a model representing the different-origin hypothesis. Similarly, speech data from the suspect are used in MAP adaptation to create GMMs, which can be interpreted as models representing the same-origin hypothesis. The speech data samples from an offender can subsequently be evaluated against these two models to output likelihood ratios [13-14].

In the case of GMM-UBM FVC, the likelihood ratio calculation is performed at the frame level for an offender recording parameterized as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. These likelihood ratios are then aggregated by computing a score s for the offender utterance

$$s = \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{p(\mathbf{x}_t|\lambda_{so})}{p(\mathbf{x}_t|\lambda_{do})}\right), \quad (7)$$

where λ_{so} denotes the probability density function modelling the same-origin hypothesis and λ_{do} denotes the probability density function modelling the different-origin hypothesis. The resulting value from this combination of likelihood ratio is referred to as a score, and a subsequent calibration step or score-to-likelihood–ratio transformation can be performed by using logistic regression calibration [15-16].

The general structure of the automatic FVC GMM-UBM system that is employed in the current paper is as illustrated in Fig. 1.

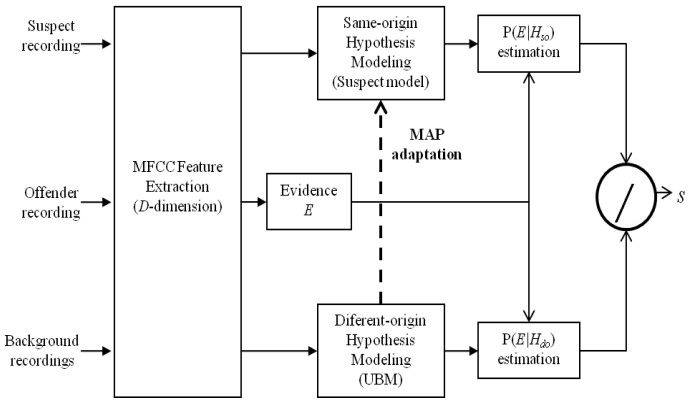

Figure 1. Automatic FVC system based on Gaussian Mixture Model – Universal Background Models (GMM-UBMs), after [1-3].

The database used by an automatic FVC system will typically be based on a relatively small homogeneous 'relevant' population (i.e. a population with a small range of variability) of speakers, where these speakers are selected on the basis of sounding sufficiently similar to a particular offender recording at a particular case a trial [13, 17].

In the context of forensic casework conditions, offender and/or suspect speech recordings are typically short in duration. GMM-UBM is currently one of the predominant FVC modelling techniques employed under such limited data conditions. For example, a GMM-UBM based experiment was investigated on the basis of a database of 68 male adult German speakers from the Pool 2010 corpus as in [9]. Similarly, a GMM-UBM based evaluation was performed on the basis of a database of 27 male speakers of Australian English as in [12].

In the context of automatic speaker verification or recognition studies, GMM-UBM has also been the core underlying fundamental technique employed under short utterance conditions. For example, a GMM-UBM based experiment reinforced with video information was proposed for short training and testing data in embedded speaker recognition for mobile devices in [18]. Another GMM-UBM based experiment combined with factor analysis was investigated for speaker verification with short utterances as in [19]. A GMM-UBM based experiment together with a phoneme class based multi-model method was investigated for short utterances of less than 2 seconds in [20].

These past automatic speaker verification/recognition studies [18-20] and automatic FVC studies [1-3, 9-12] which employed GMM-UBM modeling are mainly based on mean-only MAP adaptation or full MAP adaptation without much exploration of other adaptation configurations such as variance-only and weight-only adaptations. This therefore motivates us to pose the question: How can adaptation be performed in a manner that operates on even fewer parameters than mean-only adaptation ?

II. METHODOLOGY

A. Background, Development and Test Databases

The database used in the present study [21] consists of voice recordings of 60 female speakers of Standard Chinese from Northeastern China (i.e. Mandarin/Putonghua), aged 23-45 [21-22]. The details of the data collection protocol can be found in [22]. In real forensic casework, recordings used during forensic evaluation should be restricted to those a lay person (such as a police officer) would conclude sound sufficiently similar to the offender voice recording that they would think it is appropriate to submit them for forensic analysis. The database in this study was therefore reasonably forensically realistic, in that the recordings were restricted at a minimum to be from speakers of the same gender and general age speaking the same language and dialect [13, 17]. For each speaker, two recordings were collected, separated by two to three weeks. The speaking task employed was an information-exchange task over the telephone with overall duration of about 10 minutes. A speaker at one end of the telephone received a "badly transmitted fax" with some illegible information, and therefore had to ask another speaker at the other end of the telephone for clarification of the fax material. All recordings were recorded using flat-frequency response lapel microphones (Sennheiser MKE 2 P-C) and an external soundcard (Roland® UA-25 EX) at high quality (44.1 kHz 16 bit) with each speaker recorded on a separate channel. Moreover, only the speech-active segments are retained in the database, with silence regions and any undesirable noises removed. Recordings were evenly allocated into background, development and test sets respectively based on recordings from the initial 20 speakers, the next 20 speakers and the final 20 speakers.

B. Baseline Forensic Voice Comparison System

The automatic FVC system employed in the current study was a M = 512 mixture component Gaussian Mixture Model –

Universal Background Model (GMM-UBM) as detailed in [2]. Information was extracted from the voice recordings in the form of D = 32-dimensional MFCCs (16 static coefficients and 16 delta coefficients) from each of the 20ms frames of the voice recordings, overlapped by 10ms with Hamming windows applied. In addition, feature normalization was performed to the extracted MFCCs via cumulative distribution mapping and no channel- or session-compensation technique was applied.

For conversion of test scores to interpretable likelihood ratios, logistic regression calibration or fusion was used [15-16, 23]. Calibration and fusion weights were calculated using the scores derived from the development data, and these weights were then used to calibrate and fuse scores derived from the test data. The pooled procedure for calculating the weights was adopted in this paper [1].

C. Proposed Separate MAP Parameter Adaptation

A novel Separate MAP (SMAP) adaptation process was proposed in our study which selectively and separately updates only a particular subset of MAP adapted parameters. A subsequent step involves fusion of multiple variants of SMAP adapted FVC systems as illustrated in Fig. 2, where N denotes the total number of fused SMAP adapted systems.

The reasoning behind the SMAP process is that adaptation of a small subset of parameters is less likely to overfit short speaker-specific utterances. Hence, in each of the SMAP adapted sub-configurations, a smaller subset of MAP adapted parameters were updated, each using the same speaker specific adaptation data, potentially allowing a more accurate estimation of this smaller subset of parameters in one particular SMAP adapted system. When this process is repeated for many parameter subsets, it produces a highly complementary set of adapted systems that are good candidates for fusion. The SMAP adaptation process is summarized below in Algorithm 2, where S_n denotes the nth mutually non-overlapping subset of GMM mean parameters. In this algorithm, at each iteration step, mean parameters outside of the subset S_n are 'reset' back to their UBM values at each step (equation (9)), so that effectively only the parameters from S_n are adapted during the nth separate MAP adaptation. SMAP algorithm was also empirically evaluated to have a similar convergence behavior to conventional mean-only MAP with increasing number of MAP iterations.

Train $\lambda_{UBM} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}$

Given speaker-specific training vectors $\mathbf{x}_t$ with $t \in \{1,2,\dots,T\}$, and $S_n \subset \{1,\dots,D\}$ such that

$$\bigcup_{n=1}^{N} S_n = \{1,\dots,D\},\ \bigcap_{n=1}^{N} S_n = \phi \tag{8}$$

for n = 1 to N

  for k = 1 to K

    Calculate (2)-(6)

    for each $j \in \{1,\dots,D\} \backslash S_n$

$$\hat{\mu}_i(j) = \mu_i(j) \quad \forall \quad i \tag{9}$$

    end

  end

end

Algorithm 2: Proposed SMAP adaptation process, in which MAP adaptation is performed separately on N mean parameter subsets.

D. Mean-, Variance- and Weight-Only Adaptation

In the first part of our study, numerous variants of the traditional MAP adapted configurations for automatic FVC system were investigated. Specifically, we investigated mean-only, variance-only and weight-only adaptations and the possible fusion combinations derived from these three adaptation configurations, as well as full (i.e. weight, mean and variance) adaptation (i.e. weight, mean and variance) on the performance of the automatic GMM-UBM based FVC system.

The number of MAP iterations K was empirically defined to be 3. The relevance factor r for this first part of study was

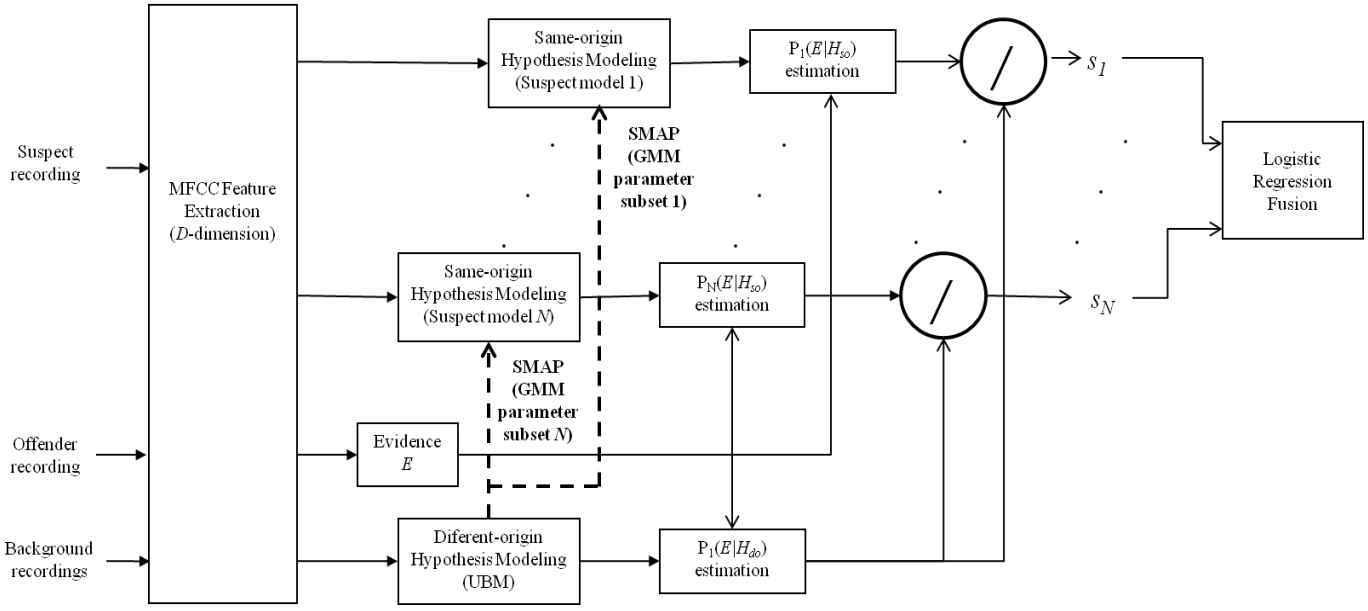

Figure 2. Automatic FVC system based on SMAP adapted sub-configurations.

necessarily defined to have a large value of 2000 to avoid some of the weights of the 512 mixture components being adapted to zero (i.e. zero-weighted-mixtures) due to the limited amount of speaker specific data, particularly for the configuration of weight-only adaptation.

E. Parameter Subset Mean Adaptation

In the second part of our study, we explore a strategy based on the SMAP adaptation of a series of parameter subsets and their fusion. In this part of the study, mean-only SMAP adaptation was considered, the number of SMAP iteration *K* was empirically defined to be 3 and the relevance factor *r* was empirically defined to be at a more conventional value of 20.

Two scenarios were considered in the implementation of the SMAP adapted sub-configurations. Firstly, we considered only fusing two (N = 2) SMAP adapted sub-configurations in which S_1 comprised only the 16 static MFCC dimensions and S_2 comprised only the 16 delta MFCC dimensions. Secondly, we considered fusion of a set of N = 32 SMAP adapted sub-configurations S_n with $n \in \{1,2,\dots,N\}$. In this scenario, only the *n*th dimension of the MFCC mean vector was updated in each iteration of the SMAP adaptation process. This is an extreme example of SMAP parameter estimation in which every feature dimension is treated as a separate feature subset.

F. Effect of Limited Adaptation Data

In addition, we performed an experiment to validate the efficacy of this SMAP fusion strategy under limited adaptation and testing data. Specifically, we evaluated the robustness of the fusion strategy by using only a proportion of the full amount of speaker specific data from suspect recordings for SMAP adaptation as well as the offender recordings for testing. The proportions that considered were 25%, 50%, 75% and 100% (i.e. the full duration) of the respective training (suspect) and testing (offender) utterances.

We performed this robustness evaluation on the experimental setting with N = 2 described in Section IIE above, in which we comparatively evaluated both the traditional mean-only MAP adaptation and the fusion of the two SMAP adapted sub-configurations. Here, the number of SMAP iterations *K* was empirically chosen to be 3 and the relevance factor *r* was empirically chosen as 20.

Note that for the case of the 25% proportion, we can generate four different FVC system results based on the four possible quarters of the whole utterance we are using for the FVC system. Similarly, for a 50% proportion, we can generate two different FVC system results based on the possible half-utterances. In both these cases, an average of these different FVC results was taken as the final FVC result.

It is also worth noting that full utterances are used to train the UBM for all experiments. The assumption is that there is always a relatively large database for UBM training and that here the UBM should be independently fixed to evaluate the effect of SMAP adaptation from varying amounts of speaker specific training data.

G. Evaluation Metrics of Forensic Voice Comparison System

As a quantifiable metric for the validity or accuracy of FVC systems Log-likelihood ratio cost C_{llr}, proposed by Brümmer [23], is used in the current study and is defined as [13-14]:

$$C_{llr} = \frac{1}{2}\left[\frac{1}{N_{so}}\sum_{i=1}^{N_{so}} log_2\left(1+\frac{1}{LR_{so_i}}\right) + \frac{1}{N_{do}}\sum_{j=1}^{N_{do}} log_2\left(1+LR_{do_j}\right)\right], \quad (10)$$

where N_{so} and N_{do} are used to denote the number of same-origin and different-origin comparisons, whereas LR_{so} and LR_{do} are used to denote the likelihood ratios for the same-origin and different-origin comparisons. $C_{llr} \in [0,\infty]$ is a scalar quantity that measures of the quality of the whole forensic voice comparison system. The lower the value of C_{llr}, the better the performance of the system. Note that it is clear from (10) that C_{llr} severely penalizes misleading LRs based on their numerical values.

As a quantifiable metric for the reliability or precision of FVC systems, Morrison et. al. [2, 13, 24-25] proposed empirical procedures for estimating the 95% Credible Interval (CI) based on the likelihood-ratio estimates from the FVC systems. In this study, the parametric procedure is employed in all calculations associated with reliability of forensic voice comparison systems. The mathematical formulation of the parametric procedure is summarised here for ease of reference. In particular, CI can be estimated based on the *t* distribution of the pooled within-group posterior standard deviation of the likelihood ratio estimates (σ') with degrees of freedom (*df*) defined as the total number of likelihood ratio estimates minus the total number of speaker-comparisons [2, 26]

$$CI = \pm t_{1-\frac{\alpha}{2},df}\sigma' \quad (11)$$

$$df = \sum_i (n_i - 1) \quad (12)$$

Here n_i denotes the number of log-likelihood ratio estimates calculated for speaker-comparison i and α is set to 0.05 for 95% CI calculation. The posterior standard deviation (σ') is defined to be equal to the sample standard deviation ($\hat{\sigma}$) with flat prior standard deviation (σ). The 95% CI is based solely on the sample variance [2]

$$\hat{\sigma}^2 = \frac{1}{df}\sum_i\left(\sum_{j=1}^{n_i}(\bar{x}_i - x_{ij})^2\right) \quad (13)$$

where $\bar{x}_i$ is the within-comparison mean based on the individual log likelihood ratio estimates x_{ij} which is defined as the *j*th likelihood ratio estimate of the speaker comparison *i*. The within-comparison mean is defined as [2]

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad (14)$$

Two versions of C_{llr} are reported, namely C_{llr}-mean and C_{llr}-pooled. C_{llr}-mean is calculated on the basis of the within-group means of log likelihood ratios where each log likelihood ratio in a group is an independent measure comparing the same pair of speakers [2, 24-25]. C_{llr}-mean together with 95% CI, form a pair of metrics where one is a measure of validity (accuracy) and the other a measure of reliability (precision). C_{llr}-pooled ignores the group membership and is calculated using the pooled likelihood ratio values instead of using the within-group means; as such it does not differentiate between accuracy and precision. It is a goodness metric that combines

both accuracy and precision, with the relative contribution of each being unknown [11, 23, 27]. On all of the above metrics, smaller values indicate better performance.

In the context of FVC analysis, the usage of log-likelihood ratio cost as a metric for measuring validity or goodness of the FVC system is consistent with the likelihood-ratio framework as it is based on gradient strength of evidence rather than hard thresholding with binary decision, i.e. correct-classification rates or classification-error rates typically used in automatic speaker recognition/verification studies, see in particular [24] and Appendix B of [12] for discussion.

III. RESULTS AND DISCUSSIONS

Results for the mean-only, variance-only and weight-only MAP adaptation and the possible fusion combinations derived from these three adaptation configurations, as well as the full adaptation (i.e. adapting based on weight, mean and variance parameters concurrently) on the performance of the automatic GMM-UBM based FVC system are tabulated in Table I. Among the four individual MAP adaptations (i.e. mean-only, variance-only, weight-only and full), the best MAP adapted configuration belongs to the mean-only MAP adaptation which is not surprising considering that this has been demonstrated by previous experimental results, e.g. [5], in the context of speaker verification. Fusion of the different MAP adapted configurations however revealed that a further improvement over the mean-only MAP adaptation is possible. The best fused FVC performance was achieved by fusing the mean-only, variance-only and weight-only FVC results which gave an improvement in pooled C_{llr} of 6.8% over the mean-only MAP adapted FVC result.

TABLE I. PERFORMANCE OF AUTOMATIC GMM-UBM FVC SYSTEM BASED ON PARTIAL ADAPTATION AND FUSED COMBINATIONS

	Cllr-mean	***95% CI***	***Cllr-pooled***
Mean-only adaptation r = 2000	0.168	0.965	0.196
Variance-only adaptation (r = 2000)	0.185	1.066	0.221
Weight-only adaptation (r = 2000)	0.834	0.439	0.848
Full adaptation (r = 2000)	0.274	0.902	0.302
Fusion of mean-only + variance-only adaptation	0.156	0.962	0.183
Fusion of mean-only + weight-only adaptation	0.156	1.027	0.187
Fusion of variance-only + weight-only adaptation	>1	>1	>1
Fusion of mean-only + variance-only + weight-only adaptation	0.151	1.048	0.182

Results for fusion of the different SMAP adapted sub-configurations on the performance of the automatic GMM-UBM based FVC system are tabulated in Table II. Fusion of the two SMAP adapted sub-configurations with N = 2 (i.e. with feature vectors split into static MFCCs and delta MFCCs) resulted in a moderate FVC system improvement in C_{llr}-pooled of 5.4% over the traditional mean-only MAP adaptation. A more substantial FVC system improvement in C_{llr}-pooled of 24.2% over the traditional mean-only MAP adaptation was observed from fusion of a set of 32 SMAP adapted systems (N = 32), where in this case the SMAP adapted parameters corresponding to only one particular MFCC dimension were updated in each sub-configuration. Other intermediate N values such as N = 4, 8 and 16 were also investigated for experimental completeness, results generally exhibited a trend of increasing FVC system improvement over the traditional mean-only MAP adaptation with increasing value of N. These results tend to imply that the more separately the GMM parameters can be estimated, the better the FVC system performance can be achieved through a collective fusion of FVC system results based on these systems.

TABLE II. PERFORMANCE OF AUTOMATIC GMM-UBM FVC SYSTEM BASED ON FUSING DIFFERENT GMM SMAP ADAPTED SUB-CONFIGURATIONS.

	Cllr-mean	***95% CI***	***Cllr-pooled***
Mean-only adaptation (r = 20)	0.040	1.317	0.056
Fusion of 2 MFCC systems (2 individually SMAP adapted configurations based on 16 static and 16 delta MFCCs respectively)	0.039	1.211	0.053
Fusion of 32 MFCC systems (32 individually SMAP adapted configurations based on each of the 32 MFCC dimensions)	0.027	2.811	0.042

Given the performance improvements of SMAP relative to conventional mean-only or full adaptation, it is reasonable to ask under what conditions these will hold. As the amount of offender data tends towards infinity, there should be no incremental improvement from SMAP relative to the conventional mean-only MAP adaptation approach [5]. This intuition seems to be upheld by the results of Fig. 3, which show that as the amount of speaker-specific data increases, SMAP seems to converge to the conventional mean only result. On the other hand, for increasingly small amounts of speaker-specific data, there seems to be an increasingly large advantage to performing separate adaptation. These results imply that the SMAP fusion strategy is indeed more robust comparatively with the traditional mean-only MAP adaptation under limited amounts of speaker training or testing data.

IV. CONCLUSION

Traditionally, most automatic FVC studies have employed GMM-UBM modeling based on mean-only MAP adaptation or full MAP adaptation with little exploration into other variants of MAP adapted configurations. An evaluation of a fusion strategy that involves fusing multiple MAP adapted GMM sub-configurations, where in each of these sub-configurations only a small subset of the GMM parameters are updated based on the same amount of limited speaker specific data, demonstrated that this fusion strategy can achieve a substantial FVC system improvement and more robust under the conditions of limited speaker training data in comparison with the traditional mean-only MAP adaptation. Therefore, this fusion strategy can be recommended for automatic FVC systems where the limited speaker data conditions hold. As future work, these experimental evaluations can be repeated on other forensic databases to further validate the findings.

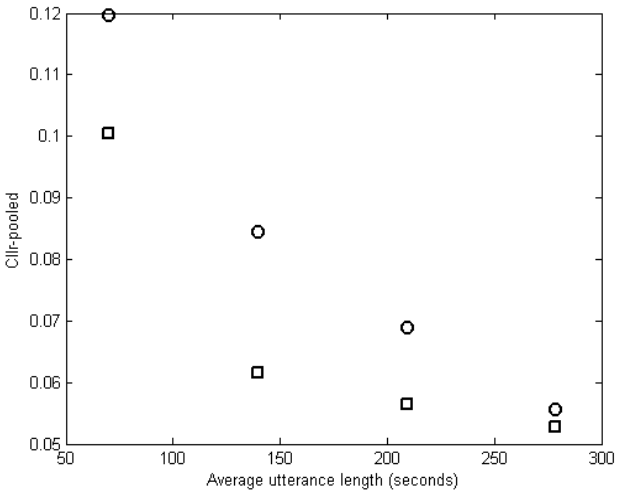

Figure 3. Evaluation of the robustness of the SMAP fusion strategy by using only a fraction of the full amount of speaker specific data from suspect recordings for MAP adaptation as well as the offender recordings for testing. The proportions that were considered are 25%, 50%, 75% and 100% (i.e. the full duration with an average of 278.7 seconds per utterance in the test database) of the respective training (suspect) and testing (offender) utterances. Solid circles: mean-only traditional MAP adaptation FVC results, Solid squares: mean-only SMAP adaptation FVC fusion results.

ACKNOWLEDGMENT

The writing of this paper was supported financially by NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

REFERENCES

[1] G.S. Morrison, T. Thiruvaran, and J. Epps, "An issue in the calculation of logistic-regression calibration and fusion weights for forensic voice comparison", in Proc. of the 13th Australasian International Conference on Speech Science and Technology, 2010, pp. 74-77.

[2] G.S. Morrison, T. Thiruvaran, and J. Epps, "Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system", in Proc. of Odyssey 2010: The Language and Speaker Recognition Workshop, Brno, Czech Republic, 2010, pp. 63-70.

[3] C. Zhang, G.S. Morrison, and T. Thiruvaran, "Forensic voice comparison using Chinese /iau/", in Proc. of the 17th International Congress of Phonetic Sciences, Hong Kong, China, 2011, pp. 2280-2283.

[4] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, vol. 17, no. 1-2, pp. 91-108, 1995.

[5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.

[6] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[7] C.M. Bishop, Pattern recognition and machine learning. Vol. 4, New York: springer, 2006.

[8] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern classification, New York: John Wiley, 2001.

[9] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models", in Proc. Interspeech, 2008, pp. 1505-1508.

[10] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition", Computer Speech & Language, vol. 20, no. 2-3, pp. 331-355, 2006.

[11] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2104-2115, 2007.

[12] G.S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM)", Speech Communication, vol. 53, no. 2, pp. 242-256, 2011.

[13] G.S. Morrison, Forensic Voice Comparison, in Expert evidence (Ch. 99), I. Freckelton and H. Selby, Editors, Thomson Reuters: Sydney, Australia, 2010.

[14] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood Ratio Calibration in a Transparent and Testable Forensic Speaker Recognition Framework", in IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006, pp. 1-8.

[15] N. Brümmer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2072-2084, 2007.

[16] G. Morrison, Robust version of train_llr_fusion. m from Niko Brümmer's FoCal Toolbox. 2009, release 2009-07-02. http://geoff-morrison.net.

[17] G.S. Morrison, F. Ochoa, and T. Thiruvaran, "Database selection for forensic voice comparison", in Proc. of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore, 2012, pp. 62-77.

[18] A. Larcher, J.-F. Bonastre, and J.S. Mason, "Short utterance-based video aided speaker recognition", in IEEE 10th Workshop on Multimedia Signal Processing, 2008, pp. 897-901.

[19] R.J. Vogt, C.J. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances", in Proc. of IEEE 10th Workshop on Multimedia Signal Processing, 2008, pp. 897-901.

[20] C. Zhang, X. Wu, T.F. Zheng, L. Wang, and C. Yin, "A K-phoneme-class based multi-model method for short utterance speaker recognition", in Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012, pp. 1-4.

[21] C. Zhang and G.S. Morrison (2011) Forensic database of audio recordings of 68 female speakers of Standard Chinese. http://databases.forensic-voice-comparison.net/.

[22] G.S. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice", Australian Journal of Forensic Sciences, vol. 44, no. 2, pp. 155-167, 2012.

[23] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection", Computer Speech & Language, vol. 20, no. 2-3, pp. 230-275, 2006.

[24] G.S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems", Science & Justice, vol. 51, no. 3, pp. 91-98, 2011.

[25] G.S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system", Forensic Science International, vol. 208, no. 1–3, pp. 59-65, 2011.

[26] W.M. Bolstad and J. Wiley, Introduction to Bayesian statistics. Vol. 2: Wiley-Interscience New York, 2007.

[27] D. van Leeuwen and N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in Speaker Classification I, Springer Berlin Heidelberg, pp. 330-353, 2007.