Applying SVMs and Weight-based Factor Analysis to Unsupervised Adaptation for Speaker Verification

Mitchell McLaren[a,b,1,*], Driss Matrouf[a], Robbie Vogt[b], Jean-Francois Bonastre[a]

[a] *Laboratoire Informatique D'Avignon (LIA), Universite d'Avignon, Agroparc BP 1228, 84911 Avignon, Cedex 9, France*
[b] *Speech and Audio Research Laboratory, Queensland University of Technology (QUT), GPO Box 2434, Brisbane, Australia, 4001*

Abstract

This paper presents an extended study on the implementation of support vector machine (SVM) based speaker verification in systems that employ continuous progressive model adaptation using the weight-based factor analysis model. The weight-based factor analysis model compensates for session variations in unsupervised scenarios by incorporating trial confidence measures in the general statistics used in the inter-session variability modelling process. Employing weight-based factor analysis in Gaussian mixture models (GMM) was recently found to provide significant performance gains to unsupervised classification. Further improvements in performance were found through the integration of SVM-based classification in the system by means of GMM supervectors.

This study focuses particularly on the way in which a client is represented in the SVM kernel space using single and multiple target supervectors. Experimental results indicate that training client SVMs using a single target supervector maximises performance while exhibiting a certain robustness to the inclusion of impostor training data in the model. Furthermore, the inclusion of low-scoring target trials in the adaptation process is investigated where they were found to significantly aid performance.

Keywords: Speaker verification, factor analysis, Gaussian mixture model (GMM), support vector machine (SVM), unsupervised adaptation

1. Introduction

Classification performance of automatic speaker verification (ASV) systems relies heavily upon the availability of sufficient client training data and the use of session variation compensation [1, 2]. While enforcing large training data requirements ensures high performance, the system would be impractical for real-world use. One approach to maximising the amount of client training data while maintaining system practically is progressive speaker model adaptation [3, 4].

[*] Corresponding author. Tel.: +617 3138 9326
Email address: m.mclaren@qut.edu.au (Mitchell McLaren)

[1] Present Address: QUT, Brisbane, Australia. This work was conducted while the author was studying as part of an internship at LIA, Avignon, France.

Progressive speaker model adaptation exploits speech acquired through normal system use to progressively increase the amount of speaker model training data in an attempt to improve overall system performance. The collection of additional training data is often conducted in an *unsupervised* manner such that a decision criterion must be employed to determine whether a given speech segment originated from the target speaker and should, therefore, be used to update the speaker model. To alleviate the difficulties associate with selecting an appropriate hard decision threshold, Preti et al. proposed *continuous* progressive model adaptation [4]. Continuous model adaptation sees speaker models adapted from all encountered test data through the allocation of weights associated with the likelihood that the speech originated from the target speaker.

Recent publications have illustrated the necessity for session variability modelling in speaker verification systems [5, 6]. Session variability refers to the differences between channel and environmental conditions during the acquisition of training and testing utterances. Although the implementation of session compensation in progressive adaptation systems that employ a hard decision threshold is straightforward [6], the estimation of relevant session statistics when using continuous adaptation presents a number of difficulties due to the use of weighted training data. The weight-based factor analysis model [7] was recently proposed to address these issues by incorporating weights into the estimation of relevant session statistics to allow session variation compensation to be employed in a continuous adaptation system. The integration of SVM-based classification into this system provided promising results, provoking the further development of the unsupervised training of SVMs.

This study extends on the research in [7] by investigating the potential that SVM-based classification offers to unsupervised model adaptation with the weight-based factor analysis model for continuous model adaptation. In this paper, the use of multiple client target GMM supervectors [8] to train speaker SVMs is investigated and compared to the use of a single target supervector during unsupervised model training. The efficient nature of the SVM is also exploited in the reduction of the problematic score shift phenomena that adversely affects progressive model adaptation systems [3, 6]. Furthermore, the importance of including low-scoring target trials in the model adaptation process and how they aid SVM-based classification performance is investigated.

This paper details the baseline continuous progressive speaker adaptation system in Section 2. A description of score shift and a common technique used to counteract its effect is given in Section 3. Section 4 presents the novel weight-based factor analysis model followed by a the proposal of two SVM-based configurations for progressive model adaptation in Section 5. Experimental results are then detailed in Section 7 followed by a discussion on the effect of target-to-impostor trial ratios on unsupervised system performance in Section 8. Concluding statements are given in Section 9.

2. Continuous Progressive Speaker Adaptation in GMMs

Progressive model adaptation is often implemented using Gaussian mixture models (GMM) due to the availability of algorithms suitable for the task such as maximum a-posteriori (MAP) adaptation [9]. MAP allows an updated speaker model to be efficiently re-trained from a universal background model (UBM) as new training data becomes available. The standard approach to unsupervised progressive model adaptation uses a test speech segment in the adaptation of the a speaker model when it is decided to have originated from the target speaker. This decision is often made by applying a predefined hard decision threshold to the log-likelihood ratio (LLR) of the given speech segment [6]. While this approach can be effective, difficulties can arise when selecting a suitable threshold.

One of the most difficult challenges with selecting a hard decision threshold for unsupervised adaptation is preventing impostor data from being used in the adaptation process where it can severely degrade the quality of speaker models [3]. Although a strict threshold will accept minimal impostor data for adaptation, beneficial target speech will also be excluded thereby reducing the performance gains expected when employing unsupervised adaptation. In contrast, a lenient threshold will accept the bulk of target speech, however, the inclusion of impostor data will adversely affect performance.

The challenge of selecting a suitable hard decision threshold has been alleviated by the proposal of several techniques that adapt speaker models based on confidence measures [3, 4]. These model adaptation techniques assign confidence measures or weights to each trial segment prior to its use in the adaptation process such that greater weight is given to data that is more likely to have originated from the target speaker. Updated speaker models are then adapted from either the UBM using all previous trial statistics or from the latest speaker model. The use of all encountered trials in the adaptation process is often termed *continuous* model adaptation. The most promising of these approaches was proposed by Preti et al. [4] and serves as the fundamental GMM-UBM continuous adaptation system in this study.

2.1. Confidence Measure Estimation

The system proposed by Preti et al. [4] calculates a trials confidence measure using a world MAP (WMAP) estimator [10]. The WMAP estimator is a two-class Bayesian classifier based on two score models — target and impostor scores — learned from a development set. In this work, 12-component GMMs are used to model the TZ-normalised[2] [11] LLR distributions. The WMAP function used to equate confidence measures α in this work can be formulated as

$$\alpha = P(tar|s) = \frac{P(s|tar).P_{tar}}{P(s|tar).P_{tar} + P(s|imp).P_{imp}} \tag{1}$$

where $P(s|tar)$ and $P(s|imp)$ are the probabilities of the score given the target and impostor score distributions, respectively, and the prior probabilities of target and impostor trials are represented by P_{tar} and P_{imp}, respectively.

The plot in Figure 1 illustrates the range of WMAP confidence measures that result when using TZ-normalised LLRs as input. Based on this plot, it can be seen that a low LLR of around zero will be allocated a confidence measure of $\alpha \approx 0$. The highest confidence measure of $\alpha \approx 1$ occurs when a normalised LLR of approximately 10 is obtained. The extremities of the plot indicate that the WMAP classifier assumes the prior probabilities when the corresponding normalised LLRs were not encountered in the training data of the WMAP score GMMs. The effect that this assumption has on the model adaptation process is investigated in Section 7.4.

2.2. Speaker Model Adaptation

Speaker model adaptation sees the new model means derived using MAP adaptation [9]. Essentially, a single iteration of the expectation-maximisation (EM) process is employed in which the speaker model means $\boldsymbol{\mu}_s$ are adapted from the world model (UBM) means $\boldsymbol{m}$ via a set of statistics calculated from the training dataset.

[2] Application of T-norm prior to Z-norm

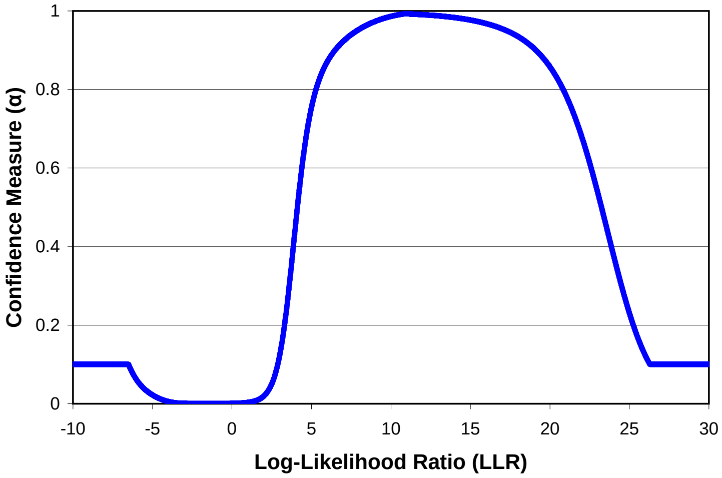

Figure 1: WMAP confidence measures calculated from TZ-normalised trial LLRs.

For this process, the zero- and first-order statistics of the training data are calculated with respect to the UBM model using all available speaker sessions $h = 1, \ldots, H$. As in the standard approach to MAP adaptation, the session-dependent zero- and first-order statistics, $\boldsymbol{N}_{(h,s)}$ and $\boldsymbol{X}_{(h,s)}$ respectively, are calculated as

$$\boldsymbol{N}_{(h,s)}[g] = \sum_{t=1}^{T_{(h,s)}} \gamma_g(t) \qquad \boldsymbol{X}_{(h,s)}[g] = \sum_{t=1}^{T_{(h,s)}} \gamma_g(t) \cdot \boldsymbol{v}_t \tag{2}$$

where each session h from speaker s contains $t = 1, \ldots, T_{(h,s)}$ observations and $\gamma_g(t)$ represents the *a posteriori* probability of Gaussian g for the observation at time t and $\boldsymbol{v}$ is the collection of training feature vectors. In this form, $\boldsymbol{N}_{(h,s)}$ represents the diagonal component occupancy matrix made up of the diagonal blocks $N_{(h,g)}\boldsymbol{I}$ where $N_{(h,g)}$ is the observation count for mixture g and $\boldsymbol{I}$ is an identity matrix having a size equal to the number of Gaussian mixture components.

The formulation of the zero- and first-order speaker-dependent statistics ($\boldsymbol{N}_s$ and $\boldsymbol{X}_s$ respectively) sees the integration of the confidence measure $\alpha_{(h,s)}$ such that,

$$\boldsymbol{N}_s = \sum_{h=1}^{H_s} \alpha_{(h,s)} \boldsymbol{N}_{(h,s)} \qquad \boldsymbol{X}_s = \sum_{h=1}^{H_s} \alpha_{(h,s)} \boldsymbol{X}_{(h,s)}. \tag{3}$$

Using these statistics and the UBM means $\boldsymbol{m}$, the MAP adaptation algorithm calculates the new speaker means $\boldsymbol{\mu}_s$ as

$$\boldsymbol{\mu}_s = (\boldsymbol{N}_s + r\boldsymbol{I})^{-1} (\boldsymbol{X}_s + r\boldsymbol{m}) \tag{4}$$

where r is the MAP relevance factor which is set to 14 in this study.

As new weighted training data is encountered during use of the progressive system, the corresponding session-dependent statistics (2) are calculated for the session and the speaker-dependent statistics in (3) are updated. The updated speaker means $\boldsymbol{\mu}_s$ are then estimated via (4) to incorporate the new training data in the speaker model.

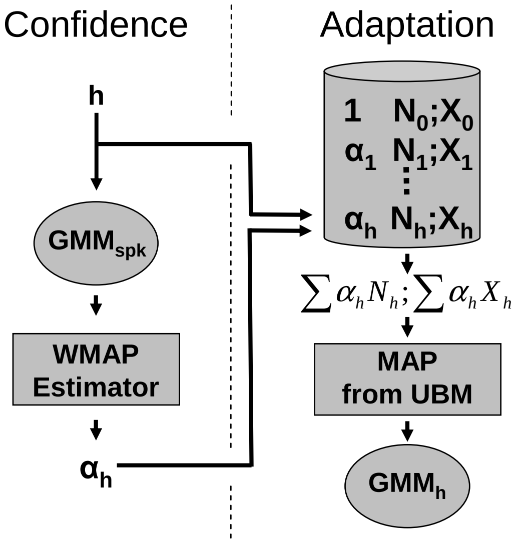

Figure 2: Illustration of GMM-based continuous progressive model adaptation system architecture when trial h is encountered.

2.3. System Architecture

In this study, operation of the continuous progressive GMM-based system consists of several distinct steps. As with non-adaptive systems, each client must first be enrolled. This involves adapting an initial speaker model GMM$_{spk}$ from the UBM using a speech sample from the client. Once a client is enrolled in the system, the testing and adaptation phases are looped with each encountered trial. The stages involved in this process include:

1. **Testing:** The classification score for trial h is obtained using the most recently updated speaker GMM$_{h-1}$.
2. **Confidence calculation:** The trials confidence measure α_h is determined by passing the LLR of h, when scored against the initial speaker GMM$_{spk}$, to the WMAP estimator.
3. **Model Adaptation:** An updated GMM is adapted from the UBM using all previously acquired zero- and first-order statistics and the newly acquired statistics from trial h weighted by α_h.

The last two stages are illustrated in Figure 2 and occur as each trial is encountered. Note that the most recently updated speaker GMM$_{h-1}$ is not used to calculate the confidence measure of trial h. Instead, the initially enrolled GMM$_{spk}$ is used to avoid adaptation problems due to *score shift*. This phenomena is described in the following section.

3. Reducing Score Shift

Score shift is a predominant issue occurring in progressive model adaptation systems that can significantly reduce classification performance if left uncorrected [3, 6]. This phenomena occurs as a target model accumulates additional training data and can be observed as a positive shift in both target and impostor score distributions. This can result in severe model corruption if a static threshold is used to select trial utterances for use in the model adaptation process as an increasing number of impostor trials will be incorrectly accepted for adaptation of the target model [6]. Likewise, the use of a static threshold to evaluate system performance quickly becomes unsuitable as the collection of scores are estimated with models trained on differing amounts of data.

A number of techniques have been proposed to counteract the effects of score shift. Mirghafori and Heck introduced client-specific classification thresholds which were adapted along with the client model to account for the shift in scores [12]. A more common approach, however, is to employ *adaptive* score normalisation techniques [6, 4]. Score normalisation techniques are commonly employed in ASV systems to compensate for many statistical variations in LLRs which in turn improves system performance [11]. In the adaptive scenario, score shift is counteracted by updating the score normalisation parameters to continually match the new characteristics of the speaker model.

A comprehensive study regarding adaptive T- and Z-score normalisation (denoted as Ta-norm and Za-norm) techniques in unsupervised conditions was recently presented by Yin et al. [6]. In the standard approach to score normalisation, T-norm statistics are calculated by scoring a given test utterance against a collection of pre-trained T-norm models. The adaptive extension of the technique, Ta-norm, dynamically selects a set of T-norm models that have been trained on a similar amount of data as the target model in an attempt to better imitate the target model characteristics in the calculation of the normalisation statistics. Similarly, the static approach to Z-norm scores a collection of impostor utterances against the enrolled client model from which normalisation statistics are calculated, while Za-norm recalculates these statistics after each model adaptation. Experiments in [6] found that Za-norm demonstrated greater performance stability over Ta-norm while the combination of both techniques (ZaTa-norm) provided the best results.

In this work, the confidence measure criterion is based on the *initial* speaker model GMM_{spk} and not the updated model GMM_{h-1} (see Section 2.3), thereby preventing score shift from adversely effecting the *adaptation* process. It should be noted that confidence estimation in the current system architecture does not account for the changes of a client's voice over time. While this is not considered an important factor for the purposes of this study involving data from the NIST SRE corpora, in practice, confidence estimation would ideally be based on the updated speaker model. As classification scores are obtained using the updated speaker model GMM_{h-1}, adaptive score normalisation must be applied to these scores in order to maximise performance.

The implementation of Ta-norm in a continuous progressive system is not straightforward. This is due to the use of weighted training data making it difficult to dynamically replicate equivalent training data in the T-norm models during the testing phase. Preti et al. previously demonstrated this by using a similar adaptive T-norm cohort approach to Yin et al., however a marginal loss of performance was observed [4]. In contrast, Za-norm uses a single impostor dataset irrespective of the amount of data used to train the target model. Therefore, static T-norm and Za-norm (TZa-norm) will be employed in this paper to counteract score shift.

4. The Weight-based Factor Analysis Model

This section describes the weight-based extension of the model that was recently proposed in [7]. The weight-based model was designed specifically to reduce session variation in systems that employ continuous progressive model adaptation by incorporating utterance-dependent weights in the modelling process.

Inter-session variation (ISV) is one of the most dominant factors that contribute to classification performance loss in ASV systems. Recent advances in techniques for modelling ISV through factor analysis have provided significant improvements in the performance of GMM-UBM based systems [1, 5]. The factor analysis model assumes that the session characteristics of a recording can be represented in a low-dimensional subspace. Represented in terms of GMM mean supervectors, a set of acoustic observations acquired over $h = 1, \ldots, H$ sessions from speaker s can be decomposed into three factors: a set of speaker-independent model (UBM) means $\boldsymbol{m}$, session-independent speaker means $\boldsymbol{y}_s$ and a session-dependent mean offset $\boldsymbol{x}_{(h,s)}$. This can be formulated as,

$$\boldsymbol{\mu}_{(h,s)} = \boldsymbol{m} + \boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x}_{(h,s)}. \tag{5}$$

where subscript (h, s) indicates session h from speaker s and $\boldsymbol{U}$ is the low-rank session variability transform matrix.

Speaker GMMs are trained through the simultaneous optimisation of the latent variables $\boldsymbol{y}_s$ and $\boldsymbol{x}_{(h,s)}$ over all speaker sessions according to the MAP criteria [1]. While this optimisation is only briefly covered in this document, a complete and efficient procedure for the optimisation of these latent variables is described in [1] or [5].

The factor analysis approach in [5] firstly estimates the session factors $\boldsymbol{x}_{(h,s)}$ using the zero- and first-order statistics defined by (2) and (3) as,

$$\boldsymbol{x}_{(h,s)} = \boldsymbol{L}^{-1}_{(h,s)}\boldsymbol{b}_{(h,s)} \tag{6}$$

where

$$\boldsymbol{L}_{(h,s)} = \boldsymbol{I} + \alpha_{(h,s)}\boldsymbol{N}_{(h,s)}\boldsymbol{U}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{U} \tag{7}$$

$$\boldsymbol{b}_{(h,s)} = \boldsymbol{U}^t\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}_{(h,s)} - \boldsymbol{m}\boldsymbol{N}_{(h,s)}\right) \tag{8}$$

where $\boldsymbol{\Sigma}$ is the UBM covariance matrix, $\boldsymbol{I}$ the identity matrix and $\alpha_{(h,s)}$ is the WMAP confidence measure assigned to trial h for client s.

The session compensated speaker component means $\boldsymbol{y}_s$ take into account the weighted session components using

$$\boldsymbol{y}_s = (\boldsymbol{N}_s + r\boldsymbol{I})^{-1}(\boldsymbol{X}_c + r\boldsymbol{m}) \tag{9}$$

where

$$\boldsymbol{X}_c = \boldsymbol{X}_s - \sum\nolimits_{h=1}^{H}\alpha_{(h,s)}\boldsymbol{N}_{(h,s)}\boldsymbol{U}\boldsymbol{x}_{(h,s)}. \tag{10}$$

Equation (9) is the same MAP adaptation formula given in (4) with a MAP adaptation factor or r, however, the *session-compensated* first-order statistics $\boldsymbol{X}_c$ are utilised in this case.

The continuous progressive system updates the general statistics and the relevant latent variables after the acquisition of each new trial. Session compensation is then performed when training an updated speaker model using these new weighted-statistics.

5. Continuous Progressive SVM Classification using GMM Supervectors

This section briefly describes SVM-based speaker verification along with the GMM supervector SVM configuration [8]. Based on this configuration, two approaches for the integration of SVM-based classification in a continuous model adaptation scenario are proposed.

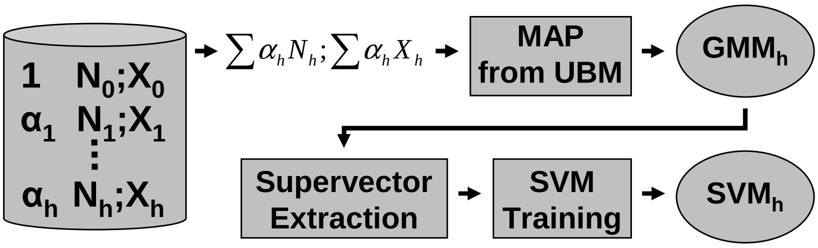

Figure 3: Adaptation data flow of the proposed Single Target Supervector (STS) SVM-based configuration for continuous progressive model adaptation.

5.1. SVMs and GMM Supervectors

In the context of speaker verification, a support vector machine (SVM) is a two-class discriminative classifier trained to separate client and impostor classes [13]. This separation occurs in a high-dimensional space via a kernel function where a hyperplane is positioned such that the margin between classes is maximised. The discriminative nature of the SVM is well suited to the task of speaker verification in which each speaker is to be distinguished from others.

Continuous progressive model adaptation can only be implemented using a classifier that can exploit utterance-dependent weights. As proposed in previous work [7], a straightforward approach to accomplishing this in SVMs involves fusing the weighted training data into the SVM features prior to model training. This is most readily accomplished using the GMM supervector SVM configuration [8].

The GMM mean supervector SVM provides a way of combining the generative modelling of adapted GMM mean vectors and discriminative SVM classification for speaker verification. GMM mean supervectors are formed through the concatenation of GMM component means, providing a convenient method of mapping a variable-length utterance to a fixed-dimension vector as required for use within an SVM classifier.

Proposed are two configurations for the incorporation of confidence weights in the GMM supervectors for SVM classification using *single* and *multiple* target supervectors in SVM training.

5.2. Single Target Supervector

The first approach involves integrating the statistics of all previously encountered trials into a single mean supervector, reflecting the SVM configuration in [7]. As the fundamental GMM-UBM system maintains a single GMM to represent each target (See Section 2), the single target supervector SVM (STS-SVM) configuration becomes a simple extension that utilises the mean supervector of the updated speaker model.

Following the same system architecture detailed in Section 2.3, enrolment sees a speaker SVM trained from the supervector extracted from the GMM trained using the clients enrolment data. The SVM-based testing phase involves training a GMM using the acquired test utterance from which the corresponding supervector is extracted and compared to the client SVM. The classification score is calculated as the distance between the test supervector and the hyperplane of the client SVM. The WMAP confidence estimation remains unchanged and based the GMM-based LLR from the initial speaker model. With each adaptation of the speaker GMM, a new SVM is trained using the supervector extracted from the updated GMM. Figure 3 depicts how the STS system incorporates the statistics and weights from the most recent test segment h and all previous trials $1, \ldots, h-1$ in the adaptation of the client SVM.

Table 1: Non-adaptive SVM-based trials on a selection of speakers from the 2005 SRE comparing modelling of speakers using single and multiple target GMM mean supervectors.

	Single Svec		**Multiple Svec**	
Training Sides	Min. DCF	EER	Min. DCF	EER
1 side	.0173	3.78%	.0173	3.78%
2 sides	.0109	1.97%	**.0096**	**1.81%**
3 sides	.0092	1.81%	**.0079**	**1.70%**

When observed in the SVM kernel space, the STS-SVM represents the target as a single positive point to be trained against a large impostor dataset. It is assumed that in using a single point to model the speaker, the high-weighted target data (including enrolment speech) will largely counteract the adverse effects of adapting using low-weight impostor data in the SVM kernel space. For example, when adapted with additional *target* data with a high confidence measure, the position of the positive speaker observation will move toward a point that better represents the characteristics of the speaker. In contrast, the use of low-weighted impostor data in the adaptation process will cause only a subtle shift in the target supervector in a sub-optimal direction.

5.3. Multiple Target Supervector

The second approach produces an additional mean supervector for SVM training with each new trial. In this way, a collection of supervectors are used to train the updated client SVM. The motivation to adopt this approach in the unsupervised system comes from a number of experiments comparing the way in which multiple conversations can be utilised in the training of a client SVM.

Given a number of training sides, two options exist for the training of a client SVM based on GMM mean supervectors; (1) train a single target supervector to represent the combination of these utterances (as in the case of the STS-SVM described in Section 5.2) or (2) train a supervector from each available utterance to produce a multiple target supervector SVM (MTS-SVM).

A comparison of these two approaches to feature production for SVM training was conducted on a subset of male speakers and their corresponding trials from the NIST 2005 SRE using between 1 and 3 training sides. Unnormalised performance statistics in non-adaptive conditions when using FA-compensated supervectors are detailed in Table 1. The use of three training sides in the single supervector configuration, while offering good performance, is surpassed by that when using multiple supervectors. It is hypothesised that the use of multiple target observations allows the SVM to account for any residual intra-speaker variation remaining in the observations after being modelled by factor analysis. Contrary to this benefit, it was empirically found that the use of multiple target observations in an unsupervised scenario was likely to produce different score distributions between models trained on a different number of sides, such that the MTS-SVM may be more susceptible to the effects of score shift. The effective implementation of adaptive score normalisation is expected to counteract these score shifts to bring about the benefits of using multiple target supervectors for SVM-based classification in an unsupervised system.

The operation of the unsupervised multiple target supervector SVM (MTS-SVM) system follows that of the STS-SVM approach differing only in the model adaptation stage. Here, an additional target supervector is produced from the newly encountered trial data and enrolment

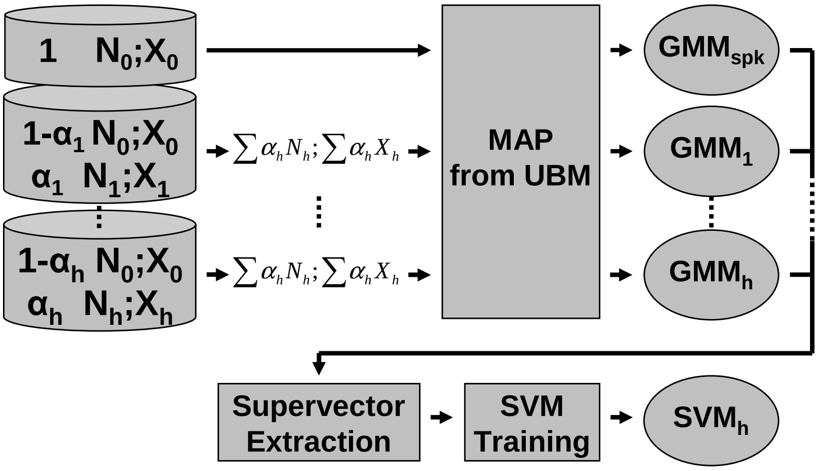

Figure 4: Adaptation data flow of the proposed Multiple Target Supervector (MTS) SVM-based configurations for continuous progressive model adaptation.

data. In this way, each positive target point in the SVM kernel space is independent of previously encountered trials. The process used to produce these additional supervectors is detailed in Figure 4. The trial h produces a corresponding supervector by weighting the test utterance statistics by α_h and the enrolment segment statistics by $1 - \alpha_h$, where α_h is determined using the WMAP estimator. This weighted enrolment data is used to reduce the influence of utterances that have a low WMAP confidence measure.

It should be noted that the repeated use of the enrolment data in the target supervectors is undesirable due to the increased emphasis given to this point in the SVM kernel space. However, the enrolment speech is the only training data that the system knows originated from the target speaker and must, therefore, be utilised in this counter-balance fashion in order to exploit weighted impostor and target training data in a multiple supervector configuration. Further, the MTS-SVM configuration detailed in this study could see, in practice, an infinite number of client supervectors being collected and utilised for model training. To better account for changes in client characteristics over time and to reduce system computation, the number of supervectors retained for this purpose could be limited to the most recently acquired or highest scoring N utterances.

In the SVM feature space, the collection of target supervectors can be viewed as a cloud centred around the enrolment training vector. When a low confidence score is assigned to a trial, the corresponding supervector will be more dependent on the speakers enrolment training data such that it resides close to the initial speaker supervector. A trial allocated a high confidence score will result in a supervector that is less dependent on the enrolment data, thereby providing additional intra-speaker variation information to the SVM.

6. Protocol

The GMM-UBM system in this study is the Mistral_SpkDet system based on the ALIZE platform[3] and distributed under an open source license. Speaker models were trained using mean-only MAP adaptation from a UBM with a relevance factor of 14. The 512-component UBM was trained on a selection of 1464 male utterances from the Fisher corpus. 512 components were utilised to match the previous work of Preti et al. [4] while providing a suitable performance-to-computation trade-off when dealing with high-dimensional GMM mean supervectors for SVM classification. Speaker utterances were represented by 19 linear frequency cepstral coefficients (LFCC) determined through filter-bank analysis, with their first derivatives, 11 of their second derivatives and the delta energy. Mean subtraction and variance normalisation were applied to features.

Evaluations were performed on the NIST 2005 corpus with the 1-sided training condition and restricted to male speakers only to give a total of 1231 target trials and 12317 impostor trials. The priors used in the WMAP estimator were 0.1 for target and 0.9 for impostors. All scores associated with WMAP were TZ-normalised. The 12-component WMAP score GMMs were trained on GMM-based scores from the 1conv4w male NIST 2006 SRE.

As a result of system developments in previous NIST SREs, separate normalisation datasets were used for the GMM- and SVM-based progressive systems. For GMM-based score normalisation, a single dataset was used for both T- and Z-norm consisting of a selection of 200 male utterances from the NIST 2004 corpus. Due to computational constraints, the statistics for Za-norm were updated only when a trial was assigned a confidence of greater than 0.1. The SVM-based configuration used a single impostor dataset for the background dataset, and T-norm and Z-norm. The use of a single impostor dataset for these roles has been shown to provide good performance compared to the use of separate impostor datasets [14]. This impostor dataset consisted of 180 male utterances selected from the Fisher corpus. T-norm SVMs were trained using the 'leave-one-out' approach and the Za-norm statistics were updated after every model adaptation. Performance statistics are given as the minimum decision cost function (DCF) and equal error rate (EER).

7. Results

The following experiments were designed with several objectives in mind: (1) to illustrate the effectiveness of the weight-based FA model to counteract session variation in a continuous model adaptation scenario, (2) to observe how the GMM-based configuration compares to the proposed SVM-based systems in terms of performance and robustness to score shift, and (3) to investigate the effect of low-weighted impostor and target trials in the model adaptation process.

7.1. Evaluation of Weight-based Factor Analysis Model

Initial experiments in this section reflect the findings of [7] and were aimed at illustrating the robustness of the recently proposed weight-based factor analysis model to the inclusion of weighted impostor data when used in unsupervised model adaptation. For this task, performance from the fundamental GMM-based system without session compensation was compared to that of the proposed session-compensated configuration.

[3] http://mistral.univ-avignon.fr

Table 2: GMM-based progressive NIST 2005 trials (male only) with and without the weight-based factor analysis model.

	non-FA		**weight-based FA**	
System	Min. DCF	EER	Min. DCF	EER
Base TZ	.0358	9.18%	.0162	4.71%
Adapt TZa	.0277	6.17%	.0089	2.36%
Oracle TZa	.0176	4.06%	.0050	1.62%

Table 2 details the results of the 1-sided male SRE'05 trials from both the standard and the session-compensated GMM system in three adaptation scenarios: no adaptation (Base), continuous unsupervised adaptation (Adapt) and supervised adaptation (Oracle) which involved adapting speaker models using only target trials with a confidence measure of one. TZa-norm was applied to all scores to counteract the score shift phenomena as described in Section 3.

The non-FA results in Table 2 indicate that continuous model adaptation provided relative gains of 23% in min. DCF and 33% in EER over baseline results. The oracle results indicate that the best possible gain one could expect from the system was 51% and 56%. This means that the baseline system acheived around half the gains possible from unsupervised adaptation suggesting that the system does not fully exploit the adaptation process.

Table 2 also details results when employing the weight-based FA model. In this system, relative gains of 38% in min. DCF and 50% in EER were observed when using the continuous model adaptation over the baseline system. The oracle results demonstrate that a relative gain of 69% and 66% in min. DCF and EER, respectively, would occur if confidence weights were correctly assigned. This indicates that continuous adaptation can better exploit weighted trial data when session compensation is employed.

A comparison of results from the non-FA and weight-based FA configurations demonstrates that significant improvements in performance were found when using session-compensation in both baseline and adaptive scenarios. This was most evident in the oracle trials where session compensation provided a significant relative gain of 72% in min. DCF and 60% in EER over non-FA oracle results with very similar gains being found in the unsupervised scenario. These results indicate, firstly, that the weight-based FA model is well suited to the task of session compensation in continuous model adaptation scenario, and secondly, that it is robust to the inclusion of impostor data in the calculation of the session statistics.

7.2. SVM-based Classification

This section compares the performance and characteristics of the proposed progressive SVM-based configurations detailed in Section 5 to the GMM-based system. All systems adopt the weight-based factor analysis model proposed in Section 4 to counteract session variation.

Evaluations were conducted on SRE'05 using the GMM-UBM, STS-SVM and MTS-SVM configurations for continuous progressive model adaptation. For each system, performance was evaluated for the Base, Adapt and Oracle conditions with unnormalised, TZ-normalised and TZa-normalised results being stated for each applicable condition in Table 3. This section focuses only on the TZa-normalised results (note, this is non-adaptive in the Base scenario) while the remaining results are discussed in Section 7.3.

When considering the TZa-normed results in Table 3, similar Base performance was achieved in both the GMM- and SVM-based configurations when no model adaptation occurred. In the unsupervised scenario, the STS-SVM provided the best performance holding a 9% gain in min.

Table 3: Comparison of GMM- and SVM-based configurations on progressive 1conv4w, male trials from the NIST 2005 SRE using the weight-based factor analysis model.

	GMM-UBM		**STS-SVM**		**MTS-SVM**	
System	Min. DCF	EER	Min. DCF	EER	Min. DCF	EER
Base	**.0190**	**4.00%**	.0191	4.95%	.0191	4.95%
Base TZ	**.0162**	4.71%	**.0162**	**4.48%**	**.0162**	**4.48%**
Adapt	.0118	3.34%	**.0116**	**3.15%**	.0203	5.20%
Adapt TZ	.0131	3.74%	**.0099**	**2.68%**	.0197	4.72%
Adapt TZa	.0089	2.36%	**.0081**	**2.27%**	.0092	3.25%
Oracle	.0104	2.92%	**.0088**	**2.44%**	.0196	4.55%
Oracle TZ	.0104	3.25%	**.0076**	**1.87%**	.0211	4.14%
Oracle TZa	**.0050**	**1.62%**	.0056	1.71%	.0068	1.87%

DCF over the GMM-based configuration. Although in the same scenario the MTS-SVM provided a similar min. DCF to the GMM-UBM, its EER suffered a relative reduction of 38%. In contrast to the unsupervised results, performance in the Oracle conditions was maximised with the GMM-based configuration followed closely by the STS-SVM. While the MTS-SVM offered reasonable performance in the oracle trials, the relatively poor unsupervised results demonstrate that it was heavily affected by the inclusion of weighted-impostor data in the speaker model and, therefore, not well suited to the task of continuous model adaptation. When comparing the two alternate configurations, the GMM-based system provided superior results in the Oracle condition while the STS-SVM held the best performance in the unsupervised scenario. This indicates that the STS-SVM is more robust to the inclusion of impostor data than the GMM-UBM system and is, therefore, the best configuration for maximising overall performance in the unsupervised conditions of the NIST 2005 SRE.

7.3. Comparison of Score-shift in Classifiers

When designing a progressive model adaptation system, it is desired that the employed classifier is robust to the detrimental effects of score shift while providing good classification performance. Following is an analysis of the effects of score shift in the GMM and SVM classifiers along with how well adaptive score normalisation counteracts these shifts. The results presented in Table 3 provide insight into these system characteristics.

In the Adapt and Oracle results of Table 3, consistent performance gains were observed when using TZa-norm over TZ-norm (ie. Z-norm statistics gathered only using the initial speaker model). Not only does this demonstrate that each of the classifiers are subject to some degree of score shift, but also that the effects of score shift can be adequately counteracted in each configuration using adaptive score normalisation. The effects of score shift were most prominent in the Oracle results of the MTS-SVM configuration where significant relative improvements of 68% in min. DCF and 55% in EER were observed when using adaptive normalisation over the standard score normalisation approach. The GMM system showed similar gains while the STS-SVM showed improvements of only 26% and 9% in min. DCF and EER, respectively, despite having similar overall performance to the other classifiers.

Of particular importance in this study is the reduction score shift in the unsupervised scenario. In this case, Table 3 indicates that the application of adaptive score normalisation over the static approach provided improvements of 53% in min. DCF and 31% in EER for the MTS-SVM,

Table 4: Lower-thresholding the WMAP confidence measurement in the male SRE'05 trials using the progressive STS-SVM system with the weight-based factor analysis model.

			Thresholded Trials	
WMAP Configuration	**Min. DCF**	**EER**	**targets**	**impostors**
Standard system	.0080	2.27%	0	0
Minimum/maximum threshold	**.0079**	2.27%	2	1099
Floored at α < .002	.0081	2.27%	15	8277
Floored at α < .005	.0081	**2.25%**	26	10058
Floored at α < .010	.0087	2.68%	37	10898
Floored at α < .020	.0086	3.09%	47	11399
Floored at α < .050	.0088	3.17%	62	11853

over 32% in both min. DCF and EER for the GMM-based configuration while the STS-SVM held a mere 18% and 15% gain in min. DCF and EER, respectively, along with the best overall performance. These statistics show that the STS-SVM configuration is considerably less susceptible to the effects of score shift than the alternate models, thus further supporting the use of the STS-SVM for continuous unsupervised progressive model adaptation.

7.4. Robust Confidence Score Estimation

This section aims to investigate some of the potential drawbacks of the WMAP-based approach (see Section 2) to the estimation of a trials corresponding confidence score.

The first aspect to be studied is the confidence measures returned by WMAP function when the trial LLRs are at the extremities of the WMAP curve. From the plot in Figure 1 of Section 2, it can be seen that the WMAP curve has a minimum and maximum. On either side of these points, the curve returns to a value of 0.1. This is because the scores used to train the WMAP GMMs do not contain examples in these outer regions and, therefore, the prior probabilities are assumed as detailed in Section 6. This poses a potential problem in that both very high scoring target trials and very low scoring impostor trials in the evaluation corpus are allocated a confidence of 0.1.

Experiments were conducted to find out whether or not these outlying confidence measures adversely effect the adaptation process. For this task, confidence measures were floored to zero when the LLR was less than the minimum point that was determined to be an LLR of -1.45, set to 1 when the LLR was above the maximum of 10.87 and allowed to remain unchanged within these boundaries. The corresponding WMAP *minimum/maximum threshold* results from the unsupervised STS-SVM evaluation of the male SRE'05 are compared to the standard WMAP configuration in to the top of Table 4. It can be observed that minimal performance improvements were found by thresholding the confidence measures. Also listed in Table 4 are the number of target and impostor trials that produced a confidence measure beyond the upper or lower threshold; in this case only two target trials were affected while more than 1000 impostors were removed from the adaptation process.

The second aspect studied in this section was the affect of low confidence measures on model adaptation from both impostor and target trials. While it is anticipated that the majority of impostor trials will be allocated a low confidence measure, their use in the adaptation process naturally contributes to some degree of model corruption. In contrast to this apparent disadvantage, the target trials that are difficult to classify are also included in the adaptation process even though that reside in this low-confidence region. As the number of low-weighted impostor examples

Table 5: TZ-normalised results for the STS-SVM configuration employing the weight-based FA model on the unsupervised adaptation mode of the NIST 2008 SRE.

System	Min. DCF	EER
Base TZ	**.0131**	2.92%
Adapt TZa	.0144	**2.51%**

are expected to far outweigh the number of low-weighted target trials, the question arises as to whether the benefits of using the current approach surpass those that would come about if the majority of low-weighted impostor data was prevented from potentially corrupting speaker models. The approach used to address this question involves combining the WMAP function with a strict lower-threshold. In particular, the minimum/maximum threshold approach, described at the beginning of this section, is modified to have a variable minimum threshold whereby a desired LLR threshold can be selected with which all lower values are floored to zero.

The results in Table 4 detail the performance obtained when using this variable threshold on the WMAP confidence measure. It can be seen that, initially, the increased minimum threshold provided few benefits to performance. However, once this threshold reached a value of .010, a notable loss in performance occurred. Performance then continued to degrade by a further 15% in EER as this threshold of .010 was increased to .020. Surprisingly, this occured despite the removal of 501 additional impostor trial from the model adaptation process and a mere 10 target trials. These results cleary demonstrate that the benefits of using low-weighted target trials far outweighs the detrimental effects that occur from the accumulation of low-weighted impostor data in a speaker model. In light of these observations, it can be stated that target trials that are difficult to classify provide a significant amount of information in the model adaptation process and their potential can only be realised through *continuous* progressive model adaptation.

7.5. Unsupervised Adaptation in the NIST 2008 SRE

The unsupervised adaptation mode of the recent NIST 2008 SRE [15] was evaluated using the unsupervised STS-SVM system employing the weight-based FA model described in this paper. The male English-only results from these trials are presented in Table 5.

Notable in these results is the 14% relative improvement in EER achieved through the use of the adaptive configuration over the baseline system. However, a performance loss of 9% in minimum DCF was also observed from the unsupervised adaptation system relative to the baseline configuration.

The reason that significant gains were not observed in both performance statistics through unsupervised adaptation in the NIST 2008 SRE compared to the NIST 2005 SRE (Table 3) can be found by analysing the ratio of target to impostor trials of the evaluation set. In the NIST 2005 SRE, there was one target trial for every 10 impostor trials — that is, 9.1% target trials. With this small proportion, the use of unsupervised adaptation provided substantial performance improvements over the baseline system. In the NIST 2008 SRE, however, only 6.7% of trials were from target speakers. Significantly less gains were achieved under these circumstances. The following section elaborates on this aspect further.

8. Discussion

The differences observed between the unsupervised evaluation of the NIST 2005 and 2008 SRE are largely due to the difference in protocol between the two datasets. This adds weight to the study presented by Preti et al. in [2] where it was found that the proportion and the distribution of positive tests can have a significant impact on the potential gains offered through unsupervised adaptation. Similar to the findings in Section 7.5, Preti et al. found that the higher rate of impostors found in the NIST 2006 SRE compared to the NIST 2005 SRE saw little or no improvement through unsupervised adaptation in the SRE'06.

The order that target examples appear in the evaluation protocol can also have a dramatic effect on how rapidly a model is adapted to better characterise the target speaker. For instance, observing all target utterances prior to the impostors segments will produce better classification scores than a protocol that encounters the impostor segments prior to the target segments.

Given these variabilities in evaluation conditions, it would appear somewhat difficult to produce a suitable protocol for the evaluation of unsupervised systems. This becomes increasingly challenging when the expected target-to-impostor trial ratio is unknown in the intended scenario for the system's application (ie., telephony banking). The likelihood of encountering an impostor trial shortly after speaker enrolment for a given application would also need to be taken into account when designing an evaluation protocol for appropriate system development.

Future studies are expected to focus on determining an appropriate evaluation protocol for unsupervised model adaptation in speaker verification systems. Such studies are likely to investigate the difference in system performance between the best and worse case scenarios. For instance, allowing the system to encounter all the target trials prior to the impostors could be considered 'best case' as it provides the greatest opportunity for the system to adapt highly-robust speaker models. In contrast, the 'worst case' would see all the impostor trials attempting to degrade the enrolled model prior to the availability of target data.

While the expected case, such as the randomly encountered, 10% target proportion of the NIST evaluations, are important for the purpose of demonstrating how the unsupervised technology operates, the applicability of the system should be recognised in a worst case scenario. Unsupervised systems that minimise the difference between the best and worst cases while providing improvements in overall classification performance could be viewed as robust to potential challenges faced through real-world use of the system. The worst conditions that the system may face in real-world situations should, therefore, be incorporated into unsupervised system development such that robustness to these conditions is addressed.

9. Conclusions

This paper investigated the integration of SVM-based classification alongside the weight-based factor analysis model in ASV systems that employ *continuous* progressive speaker adaptation. Two SVM-based configurations were proposed for integration into the system using GMM mean supervectors.

Evaluation of the weight-based factor analysis model on the NIST 2005 SRE corpus in an unsupervised scenario showed that it was robust to the inclusion of impostor data and provided relative improvements of 68% and 62% in min. DCF and EER respectively over the GMM-based configuration without session compensation. The proposed single target supervector (STS) SVM configuration provided the best unsupervised performance with a min. DCF of .0081 and an EER of 2.27%. This system demonstrated a relative gain of 9% in min. DCF over the session-compensated GMM-based system in the unsupervised condition despite the GMM-based configuration holding a marginal gain in the oracle condition. The STS-SVM was shown to be more robust to the inclusion of impostor training data than the GMM-based system and as well as less prone to the effects of score shift.

The use of low-weighted trials in the model adaptation process was investigated to determine whether they adversely effect classification performance when employing the WMAP approach to estimate confidence measures. Despite the number of low-weighted impostor trials far outweighing those of target trials, the target trials were shown to provide beneficial information to speaker model adaptation even when allocated a low confidence measure.

Acknowledgements

This research was supported by the Australian Research Council (ARC) Discovery Grant Project ID: DP0877835.

[1] R. Vogt, S. Sridharan, Experiments in Session Variability Modelling for Speaker Verification, in: IEEE International Conference on Acoustics, Speech and Language Processing, Vol. 1, 2006, pp. 897–900.

[2] A. Preti, J. F. Bonastre, D. Matrouf, F. Capman, B. Ravera, Confidence measure based unsupervised target model adaptation for speaker verification, in: Proc. Interspeech, 2007, pp. 754–757.

[3] L. Heck, N. Mirghafori, On-line unsupervised adaptation in speaker verification, in: Proc. International Conference on Spoken Language Processing, Vol. 2, 2000, pp. 454–457.

[4] A. Preti, J. F. Bonastre, F. Capman, A continuous unsupervised adaptation method for speaker verification, in: Proc. International Joint Conferences on Computer, Information and System Sciences, and Engineering (CISSE), 2006, pp. 461–465.

[5] D. Matrouf, N. Scheffer, B. Fauve, J. F. Bonastre, A straightforward and efficient implementation of the factor analysis model for speaker verification, in: Interspeech, 2007, pp. 1242–1245.

[6] S. Yin, R. Rose, P. Kenny, A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification, IEEE Transactions on Audio, Speech and Language Processing 15 (7) (2007) 1999–2010.

[7] M. McLaren, D. Matrouf, R. Vogt, J. F. Bonastre, Combining continuous progressive model adaptation and factor analysis for speaker verification, in: Proc. Interspeech, 2008, pp. 857–860.

[8] W. Campbell, D. Sturim, D. Reynolds, Support vector machines using GMM supervectors for speaker verification, Signal Processing Letters 13 (5) (2006) 308–311.

[9] D. Reynolds, T. Quatieri, R. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10 (1) (2000) 19–41.

[10] C. Fredouille, J. F. Bonastre, T. Merlin, Bayesian approach based-decision in speaker verification, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, 2001, pp. 77–81.

[11] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, Digital Signal Processing 10 (1) (2000) 42–54.

[12] N. Mirghafori, L. Heck, An adaptive speaker verification system with speaker dependent a priori decision thresholds, Proc. of the International Conference on Spoken Language Processing 2 (2002) 589–592.

[13] C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.

[14] M. McLaren, R. Vogt, B. Baker, S. Sridharan, Data-driven background dataset selection for SVM-based speaker verification, In print, IEEE Trans. Audio, Speech and Language Processing.

[15] National Institute of Standards and Technology, The NIST year 2008 speaker recognition evaluation plan, available from: http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf (2008).