Combining Continuous Progressive Model Adaptation and Factor Analysis for Speaker Verification

Mitchell McLaren[12], *Driss Matrouf*[1], *Robbie Vogt*[2], *Jean-Francois Bonastre*[1]

[1]University of Avignon, LIA, France
[2]Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
{m.mclaren, r.vogt}@qut.edu.au
{driss.matrouf, jean-francois.bonastre}@univ-avignon.fr

Abstract

This paper proposes a novel technique of incorporating factor-analysis-based inter-session variability (ISV) modelling in speaker verification systems that employ *continuous* progressive speaker model adaptation. Continuous model adaptation involves the use of all encountered trials in the adaptation process through the assignment of confidence measures. The proposed approach incorporates these confidence measures in the general statistics used in the ISV modelling process. Progressive SVM-based classification was integrated into the system through the utilisation of GMM mean supervectors. The proposed system demonstrated a gain of 50% over baseline results when trialled on the NIST 2005 SRE corpus. Adaptative score normalisation techniques were found to be beneficial to both GMM and SVM configurations alleviating the detrimental effects of score shift in progressive systems.

1. Introduction

Automatic speaker verification (ASV) systems suffer performance loss due to a number of factors. The two most dominant causes are limited training data and session variability [1, 2].

Although high performance can be obtained in an ASV system by enforcing large data requirements, it is often impractical for system users. One method of increasing the amount of speaker training data is to exploit the speech that is acquired during utilisation of the system. From this speech, client models can be updated to include statistics from trials that are determined to have originated from the appropriate target speaker. This is referred to as progressive or unsupervised speaker adaptation [3, 4].

The attainable performance of unsupervised ASV systems relies heavily on the decision criterion that is used to select additional training observations. In order to remove the challenge of selecting an appropriate threshold to make a binary decision, *continuous* progressive speaker adaptation can be employed in which all test data is weighted prior to use in the adaptation process. Preti et al. recently proposed one such system based on Gaussian mixture model (GMM) classification [4]. In the approach, a trial weight is determined using a world maximum a-posteriori (WMAP) function [5].

Several recent publications have demonstrated the necessity of session variability modelling in speaker verification systems [6, 7]. Session variability refers to the differences between channel and environmental conditions during the acquisition of training and testing utterances. A renowned approach of reducing the effect of session variation in ASV systems is factor analysis or inter-session variability (ISV) modelling [1, 6, 8].

Factor analysis attempts to model the effects of the session differences in the GMM modelling process as a mean offset constrained to a low-dimensional session subspace [6]. In this approach, the speaker model parameters and the session offset are simultaneously optimised according to maximum a-posteriori (MAP) criteria. Relevant speaker statistics are accumulated from a group of training utterances which are assumed to have originated from the same speaker. This assumption renders the current factor analysis approach unsuitable for continuous progressive speaker adaptation as all test utterances are used to adapt the speaker model.

Although similar to the threshold-based approach to model adaptation presented in [7], this paper focuses on the task of continuous adaptation through the integration of confidence measures into the statistics accumulated during inter-session variability modelling. Furthermore, support vector machine (SVM) classification is employed in order to observe the suitability of the discriminative classifier to the task of speaker model adaptation.

SVM classification can be integrated into a continuous progressive adaptation ASV system using the GMM mean supervector classifier [9]. As the GMM-based system combines all weighted training statistics for a speaker into a single model, the corresponding speaker SVM can be trained from the mean supervector extracted from this GMM.

A common problem observed in unsupervised ASV systems is score shift [3, 7]. Score shift occurs as a speaker model acquires additional training data and can be observed as a shift in both the target and impostor score distributions. As the score shifts become more apparent, the decision threshold determined for optimal classification of speaker models trained on a single utterance becomes unsuitable. Score shift is often removed using adaptative score normalisation techniques. The efficient nature of SVMs is utilised in this study to address the issue of score shift through adaptative Z-normalisation.

The layout of this paper is as follows. Section 2 details the baseline continuous progressive speaker adaptation system. The novel approach to weight-based factor analysis is presented in Section 3 followed by a description of progressive GMM supervector SVM classification in Section 4. Score shift is addressed in Section 5. Experimental results are detailed in Section 6.

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

2. Continuous Progressive Speaker Adaptation in GMMs

Progressive speaker adaptation is most commonly performed using GMMs due to the availability of suitable algorithms such as maximum a-posteriori (MAP) adaptation [10]. MAP allows an updated speaker model to be efficiently trained from a universal background model (UBM) as new training data becomes available. This occurs whenever a trial is detected to have originated from the target speaker.

Previous studies have demonstrated the importance of excluding impostor trials from the data used to update a target model due to its potential to corrupt the model [3]. To prevent model degradation, an appropriate decision criteria must be employed. A lenient decision criteria, while being more accepting of target trials, will tend to accept a significantly greater portion of impostor data. On the other hand, a strict criteria will see few adaptations take place.

In order to alleviate the challenge of selecting an appropriate static threshold, several techniques have been proposed in which the adaptation of speaker models occurs with every encountered trial [3, 4]. Referred to as *continuous* progressive speaker adaptation, these approaches assign a weight or confidence measure to each trial as it is encountered. Updated speaker models are then adapted from either the UBM using all previous trial statistics or from the latest speaker model.

The fundamental GMM-UBM unsupervised system in this paper is based on the approach proposed by Preti et al. in which each trial confidence score is determined using a world MAP (WMAP) estimator [4]. 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, GMMs are used to model the TZ-normalised[1] score distributions. The WMAP function can formulated as

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

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.

Speaker model adaptation sees the new model means derived through MAP adaptation where the statistics corresponding to the adaptation data are obtained by combining all previously weighted utterance statistics.

3. The Weight-based Factor Analysis Model

3.1. Inter-session Variability Modelling

Recent techniques of directly modelling inter-session variation (ISV) through factor analysis have demonstrated significant performance improvements in GMM-UBM based ASV systems [1, 6]. The factor analysis model assumes that the conditions brought about by session variation can be represented in a low-dimensional subspace by a set of session-dependent mean offsets in the speaker model. In terms of GMM mean supervectors, the GMM that best represents a set of acoustic observations acquired over $h = 1, .., H$ sessions from speaker s is a combination of several factors: the session-speaker-independent model (UBM) means $\boldsymbol{m}$, the session-independent speaker means $\boldsymbol{y}_s$ and a session-dependent mean offset $\boldsymbol{x}_{(h,s)}$. This can be formulated as,

$$\boldsymbol{\mu}_{(h,s)} = \boldsymbol{m} + \boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x}_{(h,s)}. \qquad (2)$$

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

A GMM speaker model is trained through the simultaneous optimisation of the latent variables $\boldsymbol{y}_s$ and $\boldsymbol{x}_{(h,s)}$ over all speaker sessions. These latent variables are optimised according to the MAP criteria [1].

Prior to the estimation of these variables, a set of general data statistics is calculated. Specifically, the zero-order and first-order statistics of the speaker data with respect to the UBM model. The session-dependent and speaker-dependent zero-order statistics, $\boldsymbol{N}_{(h,s)}$ and $\boldsymbol{N}_s$ respectively, are estimated as

$$\boldsymbol{N}_{(h,s)}[g] = \sum_{t=1}^{T_{(h,s)}} \gamma_g(t); \boldsymbol{N}_s = \sum_{h=1}^{H_s} \boldsymbol{N}_{(h,s)} \qquad (3)$$

where each session h from speaker s contains $t = 1, .., T_{(h,s)}$ observations and the *a posteriori* probability of Gaussian g for the observation t is given by $\gamma_g(t)$. Similarly, the first-order session-dependent and speaker-dependent statistics $\boldsymbol{X}_{(h,s)}$ and $\boldsymbol{X}_s$ respectively are computed using as

$$\boldsymbol{X}_{(h,s)}[g] = \sum_{t=1}^{T_{(h,s)}} \gamma_g(t) \cdot \boldsymbol{v}_t; \boldsymbol{X}_s = \sum_{h=1}^{H_s} \boldsymbol{X}_{(h,s)} \qquad (4)$$

where $\boldsymbol{v}$ is the collection of training feature vectors. These general statistics are used to calculate the latent variables from which the true speaker means can be determined. Although not covered in this document, an efficient procedure for the optimisation of these latent variables is described in [1] or [6].

3.2. Incorporating Weights in the General Statistics

Although similar to the approach detailed in [7], the novel technique presented here allows all test data to be used in the calculation of statistics as well as the model adaptation process by integrating confidence measures in the calculation of the statistics. The calculation of the zero-order and first-order statistics in (3) and (4) assume that all sessions $h = 1, .., H$ were acquired from the same speaker assigning an equal weighting to each. In the case of an ASV systems using continuous progressive speaker adaptation, all test data is used in the adaptation of the speaker model where few trials originate from the target speaker. In order to employ ISV modelling in such a task, an additional parameter is introduced into the calculation of statistics.

Equations (5) and (6) show the inclusion of $\alpha_{(h,s)}$ in the calculation of the session-dependent statistics,

$$\boldsymbol{N}_{(h,s)}[g] = \sum_{t=1}^{T_{(h,s)}} \gamma_g(t) \cdot \alpha_{(h,s)} \qquad (5)$$

$$\boldsymbol{X}_{(h,s)}[g] = \sum_{t=1}^{T_{(h,s)}} \gamma_g(t) \cdot \boldsymbol{v}_t \cdot \alpha_{(h,s)} \qquad (6)$$

where $\alpha_{(h,s)}$ is the confidence score determined by the WMAP estimator based on the LLR of session h scored against the

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

initial speaker model s. The calculation of the corresponding speaker-dependent statistics, $\boldsymbol{N}_s$ and $\boldsymbol{X}_s$ remains unchanged.

Operation of the unsupervised progressive system involves updating these statistics and the relevant latent variables after the acquisition of each trial $h = 1, .., H$. The updated speaker model is then trained using weighted data from the latest trial while also incorporating an appropriate weighting of statistics in the session variation compensation process.

4. Continuous Progressive SVM Classification using GMM Supervectors

The support vector machine (SVM) is a two-class discriminative classifier in which the maximum margin between classes is determined [9]. SVMs perform classification by projecting input vectors to a high-dimensional space in which the two classes are separated using a hyperplane. The discriminative nature of the SVM is particularly suited to the task of speaker verification in which each speaker is to be distinguished from others.

Continuous progressive speaker adaptation relies on the fact that the classifier can take into account utterance-dependent weights. Although a modified SVM kernel allowing observation-dependent weights could be used to train a speaker model, a more straightforward approach is adopted in which the weighted training data is fused into a single observation for SVM training. This can be accomplished using the GMM supervector SVM (GMM-SVM) configuration [9].

The GMM-SVM configuration combines the idea of representing acoustic observations in terms of adapted GMM mean vectors with discriminative SVM classification. The mean supervectors, formed through the concatenation of GMM component means, provide a convenient method of mapping a variable-length utterance to a fixed-dimension vector as required for use within an SVM classifier.

The progressive GMM-UBM system combines the statistics of all previously encountered trials into a single GMM using MAP adaptation (See equations (5) and (6)). As this model is updated, the corresponding mean supervector can be extracted to train an updated speaker SVM.

The testing phase involves the training of a GMM from the acquired test utterance. A supervector is then extracted from this model and tested against the client SVM. The resulting score is the distance that the test supervector lies from the separating hyperplane determined during SVM training.

5. Reducing Score Shift

Several studies have been presented in literature regarding the issue of score shift in progressive speaker adaptation systems [3, 7]. Score shift 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 results in a loss of performance as the static decision threshold used to evaluate system performance becomes unsuitable.

Although it is possible to account for score shift using client-specific thresholds [3], a more common approach is to employ *adaptative* score normalisation techniques. Score normalisation techniques are commonly employed in ASV systems due to the performance gains they offer by compensating for many statistical variations in LLR's [11]. Adaptative score normalisation reduces score shift by continually matching the normalisation statistics to the updated speaker model.

A comprehensive study regarding adaptative T and Z score normalisation (referred to as Ta-norm and Za-norm) techniques in unsupervised conditions was recently conducted by Yin et al. [7]. When compared to Ta-norm, it was found that Za-norm demonstrated greater stability in terms of performance while the best results were obtained when using both techniques (ZaTa-norm).

In this paper, the decision criteria is based on the initial speaker model removing the possibly of score shift adversely effecting the adaptation process. Nonetheless, adaptative normalisation is required as a post-score step in order to maximise performance.

The efficient testing procedure in SVM classification motivates the use of adaptive Z-normalisation as relevant statistics can be updated in a fraction of the time required by the GMM counterpart. The implementation of Ta-norm in a continuous progressive system is not straightforward due to the use of weighted training data. Preti et al. previously demonstrated this by using a similar adaptative T-norm cohort approach to Yin et al., however a marginal loss of performance was observed [4]. Therefore, only TZa-norm will be employed in this paper.

6. Experiments

6.1. Protocol

Evaluations are performed on the NIST 2005 corpus with the 1-sided training condition and restricted to male speakers only. The GMM-UBM system in this study is the LIA_SpkDet system [2] based on the ALIZE platform [3] and distributed under an open source license. This system produces speaker models using MAP adaptation by adapting only the means 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. Speaker utterances were represented by 19 linear frequency cepstral coefficients (LFCC) determined through filterbank analysis, with their first derivatives, 11 of their second derivatives and the delta energy. Mean subtraction and variance normalisation were applied to features.

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 WMAP score GMMs were trained on the scores from the 1-sided male 2006 evaluation.

Two progressive systems were evaluated: One GMM-based and the other SVM-based. For the GMM-based scores, a single dataset was used for both T- and Z-normalisation 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 weight of greater than 0.1.

The GMM-SVM configuration employed the GMM supervector kernel proposed by Campbell [9]. The background dataset consisted of 180 male utterances selected from the Fisher corpus. This dataset was also used to gather statistics for both T- and Z-normalisation where the T-norm SVMs were trained using the 'leave-one-out' approach. The Za-norm statistics were updated after every model adaptation.

6.2. System Evaluation on NIST Corpora

The proposed weight-based FA model was trialled on the NIST 2005 SRE corpora with results detailed in Figure 1 and Table 1.

The results demonstrate that a significant gain can be achieved through the use of speaker model adaptation. In the

[2] http://alize.univ-avignon.fr/heberges/ALIZE/LIA_RAL/

[3] http://alize.univ-avignon.fr/heberges/ALIZE/

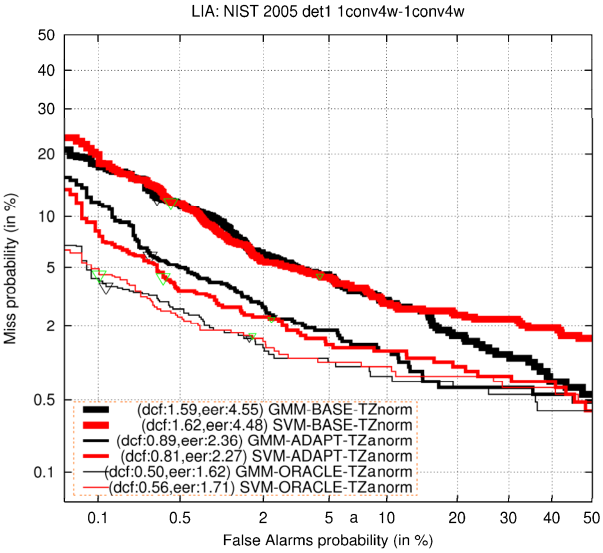

Figure 1: DET plot of the proposed progressive ASV system using the weight-based factor analysis model trialled on the NIST 2005 SRE corpus with TZa-normalisation.

supervised mode (Oracle), GMM classification provides a relative gain of 69% and 64% in DCF and EER respectively over baseline results while SVM classification held marginally lower gains. Of particular interest however is how the classifiers respond in the unsupervised mode (Adapt) where impostor data is able to deteriorate the speaker model. It can be seen that SVM classification is superior with a realtive gain of approximately 50% in both DCF and EER while the GMM system achieves a gain of 44% in DCF and 48% in EER. This suggests that when compared to the GMM, the SVM is more robust to the inclusion of detrimental training data.

A comparison was made between the application of static TZ-norm (ie. Z-norm statistics gathered from initial speaker model only) to TZa-norm in the GMM and SVM classifiers. Results are detailed in Table 1. It can be seen that both classifiers greatly benefit from the application of TZa-norm with an average performance gain of around 26% in the unsupervised mode. The use of TZ-normalisation however results in a performance loss of approximately 9% in the GMM system. On the contrary, unsupervised SVM classification provides a performance gain of 11% in DCF and 14% in EER when using TZ-normed scores. These findings suggest (1) that score normalisation must be *adaptative* when employed in a progressive GMM system, and (2) that SVMs appear to be less prone to score shift than GMMs.

7. Conclusions

This paper detailed the implementation of a weight-based factor analysis model for inter-session variability modelling. Such a technique is particularly suited to ASV systems employing *continuous* progressive speaker adaptation.

Evaluated on the NIST 2005 SRE corpus, results indicate firstly that similar performance is achievable in both GMM and SVM configurations when used in a non-adaptative or supervised conditions. In the unsupervised mode however, SVM classification was found to be marginally more robust to the inclusion of impostor training data. The use of adaptative Z-score normalisation was highly beneficial in both configurations while requiring less computation in the SVM-based system. A comparison between TZ and TZa-norm in the progressive GMM and SVM configurations suggested that SVMs may be less prone to the detrimental effects of score shift.

Table 1: Comparison of normalisation methods on GMM and SVM configurations on progressive male trials from the NIST 2005 SRE corpora.

System	GMM		SVM	
	Min. DCF	EER	Min. DCF	EER
Base	.0190	4.00%	.0191	4.95%
Base TZ	.0159	4.55%	.0162	4.48%
Adapt	.0116	3.35%	.0111	3.01%
Adapt TZ	.0126	3.65%	.0099	2.59%
Adapt TZa	.0089	2.36%	.0081	2.27%
Oracle	.0104	2.92%	.0088	2.44%
Oracle TZ	.0104	3.25%	.0076	1.87%
Oracle TZa	.0050	1.62%	.0056	1.71%

Future work will investigate methods of combining weighted training data in the SVM rather than the GMM using methods such as kernel fusion. A thorough study into the degree that score shift exists in progressive SVM classification would also be beneficial.

8. References

[1] R. Vogt and 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, and B. Ravera, “Confidence measure based unsupervised target model adaptation for speaker verification,” in *Interspeech*, 2007, pp. 754–757.

[3] L. Heck and N. Mirghafori, “On-line unsupervised adaptation in speaker verification,” in *International Conference on Spoken Language Processing*, vol. 2, 2000, pp. 454–457.

[4] A. Preti, J. F. Bonastre, and F. Capman, “A continuous unsupervised adaptation method for speaker verification,” in *International Joint Conferences on Computer, Information and System Sciences, and Engineering (CISSE)*, 2006.

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

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

[7] S. Yin, R. Rose, and P. Kenny, “A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification,” in *IEEE International Conference on Acoustics, Speech and Language Processing*, vol. 15, no. 7, 2007, pp. 1999–2010.

[8] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor Analysis Simplified,” in *IEEE International Conference on Acoustics, Speech and Language Processing*, vol. 1, 2005, pp. 637–640.

[9] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[10] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[11] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.