Factor Analysis Modelling for Speaker Verification with Short Utterances

Robbie Vogt, Chris Lustri, Sridha Sridharan

Speech Research Laboratory, Queensland University of Technology, Brisbane, Australia

{r.vogt, c.lustri, s.sridharan}@qut.edu.au

Abstract

This paper examines combining both relevance MAP and subspace speaker adaptation processes to train GMM speaker models for use in speaker verification systems with a particular focus on short utterance lengths. The subspace speaker adaptation method involves developing a speaker GMM mean supervector as the sum of a speaker-independent prior distribution and a speaker dependent offset constrained to lie within a low-rank subspace, and has been shown to provide improvements in accuracy over ordinary relevance MAP when the amount of training data is limited.

It is shown through testing on NIST SRE data that combining the two processes provides speaker models which lead to modest improvements in verification accuracy for limited data situations, in addition to improving the performance of the speaker verification system when a larger amount of available training data is available.

Index Terms: speaker verification, factor analysis, probabilistic PCA.

1. Introduction

The introduction of the GMM-UBM verification framework [1] and particularly *maximum a posteriori* (MAP) estimation of GMM speaker models was a significant step forward in text-independent speaker verification performance. Compared to maximum likelihood (ML) estimation, MAP incorporates prior knowledge in the model estimation criterion, in the form of a universal background model (UBM) trained on a large and diverse range of speech, to constrain the GMM parameter estimates [2]. The MAP approach allowed for significantly more complex speaker models to be trained with limited data that are not adversely affected by acoustic events that were unseen in the training data [1].

While the GMM-UBM approach has proven to be very successful at training speaker models with around 2 to 3 minutes of active speech, there are many applications of speaker verification technology which require the formation of accurate speaker models with far more limited durations of training and testing data. This problem has been addressed by introducing the concept of MAP adaptation in a low-dimensional speaker subspace, know as Probabilistic PCA, in order to train speaker models with a greatly reduced number of free parameters [3]. This method has been shown to produce more accurate speaker models than the full relevance MAP process when the amount of training data is very limited in a small vocabulary (digits), text-independent speaker verification task. Unfortunately, when there is a large amount of training data, the full relevance MAP process produces speaker models of a higher quality than those produced by the speaker subspace adaptation method due to the constrained nature of the subspace representation of the speaker.

Ideally, a speaker verification training system would produce accurate models when training data is limited, while also producing the best possible models when there is a large amount of training data and a finer detailed representation is possible. This paper outlines a method of combining both the relevance MAP and speaker subspace adaptation approaches in order to produce a more flexible system which provides accurate models with limited training data, while also producing high quality models as the amount of training data becomes large. The approach taken is to combine the relevance and subspace MAP methods with a simultaneous optimisation in a manner similar to the factor analysis approach of Kenny, *et al.* [4].

The following section briefly outlines the GMM-UBM framework and relevance MAP as the basis for this work with a description of the extended factor analysis model provided in Section 3. Experimental results on the NIST SRE 2005 data using a range of utterance lengths is then presented in Section 4.

2. GMM-UBM Verification System

The Gaussian mixture model (GMM) is a flexible probabilistic model commonly used for speaker modelling and consists of a weighted sum of normal distributions in the feature space [1]. A GMM is fully described by the set of mixture component distribution weights ω_c, means $\boldsymbol{\mu}_c$ and covariances $\boldsymbol{\Sigma}_c$ for each component $c = 1, \ldots, C$.

The GMM-UBM structure first proposed by Reynolds [1] has become the standard approach to text-independent speaker verification. The central advance introduced in the GMM-UBM approach is the extensive use of a universal background model (UBM) as both the basis of speaker model adaptation and to represent the null hypothesis in a likelihood ratio test. The UBM is a high-order GMM trained on a large quantity of speech obtained from a wide sample of the speaker population of interest and is desigend to capture the general form of a speaker model.

Under the GMM-UBM approach, a speaker model is estimated through maximum *a posteriori* (MAP) estimation allowing for prior knowledge in the form of a prior distribution to be incorporated into the estimation process. In the form proposed by Reynolds, known as *relevance* MAP [3], the prior distribution for this estimation is determined by the UBM parameters and a factor τ governing the influence or relevance of the UBM on the final speaker model.

The relevance MAP solution can be conveniently expressed in terms of the concatenated GMM component mean vectors. In this way, a speaker dependent model for speaker s, is fully defined as $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^T \ldots \boldsymbol{\mu}_C^T\right]^T$ which is a $CF \times 1$ supervector containing the means of each mixture component in the speaker GMM, where F represents the dimensionality of the feature vectors used in the model and C denotes the total number of

This research was supported by the Australian Research Council Discovery Grant No DP0557387.

mixture components used to represent the GMM. The relevance MAP model then takes the form

$$\boldsymbol{\mu}(s) = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}(s) \quad (1)$$

where, $\boldsymbol{m}$ represents the prior distribution in mean supervector space, while $\boldsymbol{D}$ is set to be a $CF \times CF$ diagonal matrix, and $\boldsymbol{z}(s)$ takes the form of a speaker-dependent offset vector from the UBM mean. $\boldsymbol{z}(s)$ is estimated by optimising a MAP criterion [5] with the standard normal distribution, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, as the prior.

To match Reynolds' formulation, $\boldsymbol{D}$ is constrained to satisfy $\boldsymbol{I} = \tau \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{D}$. Here τ is the relevance factor and $\boldsymbol{\Sigma}$ is a diagonal matrix consisting of the UBM component covariance matrices $\boldsymbol{\Sigma}_c$.

Due to the large number of parameters that need to be estimated in the relevance MAP process, the speaker model requires a large amount of training data in order to take full advantage of the technique. When limited training data is available, the model is unable to saturate, and the ability of the speaker model produced by the process to accurately model the speaker is limited.

3. Factor Analysis for Speaker Verification

The factor analysis techniques outlined by Kenny, *et al.* [4] are based on the decomposition of the GMM mean supervectors into speaker- and session-dependent parts. As such, a GMM supervector representation of a given utterance may be expressed as the sum of a speaker-dependent contribution and a speaker-independent session contribution. The motivation behind factor analysis is to explicitly model each of these contributions in a low-dimensional subspace of the GMM mean supervector space in order to form a more accurate speaker GMM for speaker verification purposes.

3.1. Speaker Variability

Speaker subspace adaptation, first proposed for speaker recognition by Lucey and Chen [3], is a process in which it is assumed that the majority of speaker variation is contained within a low-rank subspace of the full supervector space, as opposed to the full space. In this situation, the speaker model may be expressed as

$$\boldsymbol{\mu}(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s). \quad (2)$$

In this model, $\boldsymbol{V}$ is a low-rank transformation matrix which has been trained on a variety of background speakers in order to capture the main directions of speaker variation [3, 4] and vector $\boldsymbol{y}(s)$ represents the parameters of the speaker in the specified subspace. $\boldsymbol{V}$ is trained so that $\boldsymbol{y}(s)$ follows a standard normal distribution. To train a speaker model, $\boldsymbol{y}(s)$ is again optimised according to a MAP criterion. This model enables a speaker GMM to be estimated within the subspace that captures the most salient characteristics with a low number of parameters.

A significant advantage of using speaker subspace adaptation is that it requires far fewer free parameters to be estimated in order to train a GMM. Furthermore, the parameters which are estimated give information about the speaker in the subspace that contains the greatest variation between speakers, therefore forming a model that captures the most important information about the speaker while requiring less data than the full relevance MAP. Unlike relevance MAP, adaptation constrained to lie within the subspace also infers information about acoustic events that may not have occurred in the training utterance as the individual mixture component mean offsets are related through the definition of the subspace transform $\boldsymbol{V}$. This implies that the speaker subspace adaptation process can produce more accurate models with limited speech data than a full relevance MAP training process, as demonstrated by Lucey and Chen [3].

Unfortunately, as the amount of available speaker data increases a speaker model formed by the relevance MAP will become more accurate than a speaker model constrained to lie in the speaker subspace due to the fact that the speaker subspace becomes saturated [3]. As such, speaker models produced by the speaker subspace adaptation method will not converge asymptotically to the maximum likelihood estimate as the amount of training data increases as the relevance MAP approach will.

3.2. Session Variability

It is possible using a similar method to explicitly model mismatch between different sessions of the same speaker. It is assumed in this formulation of the speaker model that the most significant session variability effects may also be described in a low-dimensional subspace of the full mean supervector space. This allows for a channel compensation supervector to be introduced into the speaker model, in order to minimise the effect of this inter-session variability. To achieve this, a speaker GMM may be considered as the combination of a session-independent speaker model with an additional offset of the model means representing the recording conditions of the session h. This can be expressed as

$$\boldsymbol{\mu_h}(s) = \boldsymbol{\mu}(s) + \boldsymbol{U}\boldsymbol{x}_h(s). \quad (3)$$

In this representation, $\boldsymbol{U}$ is a low-rank transformation matrix representing the main directions of session variation. The matrix $\boldsymbol{U}$ is determined using a wide range of speakers, and comparing different sessions of the same speakers in order to determine the subspace which contains the most significant session variability effects [4, 6]. Similarly to the speaker factor adaptation, the vector $\boldsymbol{x}_h(s)$ is an estimate of the session conditions with the session subspace, and follows a standard normal distribution.

By explicitly modelling the session conditions in this way it is possible to remove the most significant linear effects of session variability and generate a more reliable estimate of the true characteristics of the speaker mean, $\boldsymbol{\mu}(s)$. This approach has been demonstrated to provide significant improvements in the verification accuracy of speaker verification systems [6, 4] when sufficient data is available to estimate the session conditions.

3.3. Combining Relevance and Subspace MAP

A significant limitation of the speaker subspace adaptation approach is the assumption that all speaker variation lies in a low-dimensional subspace of the full speaker space. While this subspace is enough to provide a speaker model of reasonable quality when the amount of available training data is limited, it becomes insufficient to provide the most accurate model when the amount of data increases.

It is therefore useful to assume that the speaker model takes a form which combines both relevance MAP and speaker subspace adaptation [4]. This form is able to provide accurate speaker models with limited available training data, while also converging asymptotically to the maximum likelihood estimate as the amount of training data increases. In this case, the

System	1 conv	60 sec	20 sec	10 sec
Baseline	.0442	.0456	.0608	.0752
$R_y = 50$	.0437	.0451	.0598	.0732
$R_y = 100$	.0434	.0452	.0592	.0736
$R_y = 200$	**.0422**	**.0434**	**.0571**	**.0727**

Table 1: *DCF on the female subset of the 2005 NIST SRE common evaluation condition.*

System	1 conv	60 sec	20 sec	10 sec
Baseline	9.51%	9.93%	14.73%	20.88%
$R_y = 50$	9.34%	9.76%	14.31%	20.79%
$R_y = 100$	9.26%	9.60%	14.14%	20.62%
$R_y = 200$	**9.09%**	**9.51%**	**13.80%**	**19.87%**

Table 2: *EER on the female subset of the 2005 NIST SRE common evaluation condition.*

speaker GMM takes the form expressed as

$$\boldsymbol{\mu}(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s) + \boldsymbol{D}\boldsymbol{z}(s). \quad (4)$$

In addition to this, a joint speaker model may be formed by combining the relevance MAP method and speaker subspace training with session variability training by using the formulation (4) in (3). This enables the combined model to compensate for channel effects in addition to modelling the speaker variation and retaining the property of asymptotic convergence to the maximum likelihood estimate with large amounts of training data.

To optimise the full model described in (4) it is necessary to simultaneously optimise the variables $\boldsymbol{y}(s)$ and $\boldsymbol{z}(s)$ as well as the set $\boldsymbol{x}_h(s)$ in the case of session variability modelling. This is a non-trivial task requiring the decomposition of a very large matrix [4]. A direct solution to this optimisation problem is possible, however, this work employs an efficient, iterative algorithm based on the Gauss-Seidel approximation method [6].

System	1 conv	60 sec	20 sec	10 sec
Baseline	.0442	.0456	.0608	.0752
$R_y = 200$	.0422	.0434	**.0571**	**.0727**
Channel	.0305	.0373	.0702	.0857
Chan $R_y = 200$	**.0295**	**.0350**	.0671	.0880

Table 3: *DCF on the female subset of the 2005 NIST SRE common evaluation condition for systems with and without channel compensation.*

System	1 conv	60 sec	20 sec	10 sec
Baseline	9.51%	9.93%	14.73%	20.88%
$R_y = 200$	9.09%	9.51%	**13.80%**	**19.87%**
Channel	7.24%	8.67%	19.11%	28.87%
Chan $R_y = 200$	**6.90%**	**8.33%**	18.01%	26.35%

Table 4: *EER on the female subset of the 2005 NIST SRE common evaluation condition for systems with and without channel compensation.*

4. Experiments

The baseline recognition system used in this study utilises fully coupled GMM-UBM modelling using MAP adaptation and feature-warped MFCC features with appended delta coefficients [7]. An adaptation relevance factor of $\tau = 8$ and 512-component models are used throughout and a session variability subspace of dimension $R_x = 50$ is used when session variability modelling is applied. The transforms for both the speaker and session subspaces were trained on a combination of Switchboard-2 and Mixer data drawn from earlier NIST SRE's.

A modified version of the NIST 2005 Speaker Recognition Evaluation [8] corpus and protocol was used for the presented experiments. This data is drawn from the recent Mixer conversational telephony corpus which includes a wide variety of mismatched conditions with speakers using both landline and mobile handsets and channels. To investigate the trends of the evaluated techniques, a range of shortened utterance lengths was tested. The shortened utterances were obtained by truncating the utterances of the 1conv4w-1conv4w condition to the specified length of active speech data for both training and testing. Utterance lengths of 10, 20 and 60 seconds were examined, as well as the full available conversation side for comparison purposes (typically with 100–120 seconds of active speech).

4.1. Results

The results presented in Tables 1 and 2 demonstrate a consistent but small advantage in combining relevance MAP adaptation with speaker adaptation when training speaker models for the purpose of speaker verification. It can be seen that the accuracy of the verification system is an improvement on the ordinary relevance MAP process over all tested utterance lengths.

The improvements shown in the results of the shorter tests demonstrate that this approach has merit when limited utterance data is available for training speaker models. There is an improvement in the resultant DCF and EER scores of the tests using both 10 seconds and 20 seconds of training data when both speaker adaptation and relevance MAP are incorporated into the speaker training process, with the best relative improvement observed for the 20 sec condition showing 6% relative improvements in both DCF and EER.

It is noted that using 200 speaker factors produced the lowest DCF and EER scores for the entire range of utterance durations tested. These results, particularly when the 10 sec condition in considered, indicate that an increase in speaker factors could well be supported even with these short utterance lengths. Of course, increasing the number of speaker factors will also require an increase in the size of the corpus used to train the transform $\boldsymbol{V}$ to a ensure sufficiently representative subspace.

It is also notable that the combined speaker subspace adaptation and relevance MAP process produces improved results for longer training utterances in addition to short-length training utterances. This shows that combining speaker subspace adaptation and relevance MAP techniques to train speaker models provides improvements in verification accuracy over a range of training durations, therefore providing a more robust and flexible speaker verification system.

Tables 3 and 4 present the comparative performance of systems additionally incorporating session variability modelling into the factor analysis model. Again, the best performance is given by a system with combined speaker subspace and relevance MAP for all utterance lengths, however, incorporating channel factors appears to be effective only for longer utterances. These results agree with previously published results in [6] that demonstrated a testing utterance length of 10 to 20 seconds were needed to observe improved performance from session variability modelling when the speaker models were trained on a full conversation side.

System	1 conv	60 sec	20 sec	10 sec
Baseline	.0429	.0449	.0589	.0748
$R_y = 200$	.0417	.0423	**.0562**	**.0719**
Channel	.0280	.0348	.0668	.0854
Chan $R_y = 200$	**.0274**	**.0337**	.0648	.0862

Table 5: *DCF on the female subset of the 2005 NIST SRE common evaluation condition for systems with Z-Norm applied.*

System	1 conv	60 sec	20 sec	10 sec
Baseline	8.92%	9.93%	14.73%	20.79%
$R_y = 200$	8.59%	9.18%	**13.89%**	**20.12%**
Channel	**6.48%**	8.00%	18.43%	29.12%
Chan $R_y = 200$	**6.48%**	**7.74%**	17.59%	26.60%

Table 6: *EER on the female subset of the 2005 NIST SRE common evaluation condition for systems with Z-Norm applied.*

It is notable that the verification system trained on the 10 second condition was able to gain advantages through estimating 100 or 200 speaker factors, but was not able to estimate 50 channel factors effectively. It is hypothesised that this inconsistency is due to the fact that information represented by the channel factors is effectively removed, therefore poorly estimated channel factors may cause useful information to be discarded. On the other hand, poorly estimated speaker factors may still offer a useful representation of the data and correspondingly some improvement in the verification system.

It is alos noted that these results were obtained using a relevance factor of $\tau = 8$ throughout, which may be suboptimal for combining both relevance MAP and subspace adaptation. Future work will investigate methods for determining the most appropriate relevance factor for the combined model. It is not clear at this stage whether the relevance factor should be obtained through a tuning process using the combined model or via a direct optimisation of the matrix $\boldsymbol{D}$. Determining the most appropriate relevance factor is expected to result in further improvements for the combined relevance MAP and subspace adaptation process.

Finally, results are presented in Tables 5 and 6 for the evaluated systems incorporating Z-Norm score normalisation. Comparing to the results in Tables 3 and 4, the introduction of score normalisation has not made significant changes to the observed trends with a performance gain observed with Z-Norm in almost all cases. It is worth noting, however, that Z-Norm has reduced the advantage of incorporating the speaker subspace adaptation for the systems with channel compensation and that Z-Norm is not helpful for the EER in the 10-second condition.

5. Conclusions

It was proposed in this work that the factor analysis modelling approach to GMM speaker verification is an ideal solution for combining the benefits of speaker subspace MAP adaptation for short utterances and standard relevance MAP adaptation for longer utterances to provide a speaker modelling approach that is optimal over a wide range of utterance lengths. MAP adaptation within a low-dimensional subspace has been demonstrated to capture the dominant characteristics of a speaker with very short training but is limited by the subspace as the training length increases. Conversely, relevance MAP asymptotically approaches the optimal ML solution for large amounts of training but is unable to infer the characteristics of a speaker for unseen acoustic events. These characteristics were combined in the FA approach by simultaneously optimising a combined subspace and relevance MAP criterion.

Experiments conducted on 2005 NIST SRE data using a range of training and testing utterance lengths from 10 seconds up to a conversation side of typically 2 to 3 minutes of active speech demonstrate a consistent advantage to the combined subspace and relevance MAP criterion over relevance MAP alone. Further, additionally including session factors in the model gave improved performance for longer trials but degraded performance as the utterance lengths dropped to 20 seconds and below.

6. References

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

[2] J.-L. Gauvain and C.-H. Lee, "Bayesian adaptive learning and MAP estimation of HMM," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. Soong, and K. Paliwal, Eds. Boston, Mass: Kluwer Academic, 1996, pp. 83–107.

[3] S. Lucey and T. Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Eurospeech*, 2003, pp. 2021–2024.

[4] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 219–226.

[5] J. Gauvain, L. Lamel, and B. Prouts, "Experiments with speaker verification over the telephone," *Eurospeech*, 1995.

[6] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2006, pp. 897–900.

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

[8] National Institute of Standards and Technology, "The NIST year 2005 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf, 2005.