interspeech 2008 incorporating SST 08

Factor Analysis Subspace Estimation for Speaker Verification with Short Utterances

Robbie Vogt, Brendan Baker and Sridha Sridharan

Speech and Audio Research Laboratory,
Queensland University of Technology, Brisbane, Australia.
{r.vogt, bj.baker, s.sridharan}@qut.edu.au

Abstract

Training the speaker and session subspaces is an integral problem in developing a joint factor analysis GMM speaker verification system. This work investigates and compares several alternative procedures for this task with a particular focus on training and testing with short utterances. Experiments show that better performance can be obtained when an independent rather than simultaneous optimisation of the two core variability subspaces is used. It is additionally shown that for verification trials on short utterances it is important for the session subspace to be trained with matched length utterances. Conversely, the speaker transform should always be trained with as much data as possible.

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

1 Introduction

Some of the most successful text-independent speaker verification systems in recent years have been based around the use of factor analysis modelling. Such techniques have gained prevalence due to their ability to deal with complex sources of intersession variation. Unlike techniques such as feature mapping, the factor analysis model treats the session (and speaker) components as a continuous variable rather than a discrete one. The explicit modelling of the session variation provides a more powerful mechanism to remove complex intersession effects.

This paper utilises a joint factor analysis model, similar to that described by Kenny, *et al.* [1], that combines both relevance MAP and subspace adaptation approaches to model both speaker and session variability.

For a factor analysis approach to be effective, appropriate models of both speaker and session variability need to be estimated. In this work, the models of both speaker and session variation take the form of low-rank transformation matrices. A core assumption is made that the majority of speaker and session variation can be modelled with a small number of variables. Each of these transformation matrices describes a subspace that encapsulates the main directions of variation. Ideally, the transformation matrices should accurately represent the types of inter- and intra-speaker variations expected within and between recording sessions.

In this paper, the problem of estimating the session and speaker subspaces is examined to determine if there is in fact value in the iterative refinement of the subspaces as has previously been assumed. Also investigated is the merit in independently optimising the speaker and session subspaces to better control the

Additionally, factor analysis methods have shown mixed results with reduced training and testing utterance lengths. While the incorporation of a speaker subspace has been shown to be beneficial [2, 3], accurately estimating session variables with shorter utterances has proven problematic [3, 4]. Training the speaker and session subspaces specifically for short utterances is also investigated to address this issue.

2 Factor Analysis for Speaker Verification

As in traditional approaches to speaker verification, factor analysis techniques are based around the use of GMM's to model a speaker. The factor analysis technique outlined by Kenny, *et al.* [1] is based on the decomposition of the GMM mean supervectors into speaker- and session-dependent parts. The motivation behind factor analysis techniques is to explicitly model and separate the speaker and session contributions.

Before describing the factor analysis model in detail, it is first useful to define some variables. Let F be the dimension of the acoustic feature vectors used. Also, define C as the total number of mixture components used to represent a GMM. 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 $c = 1, \ldots, C$.

In the commonly used GMM-UBM paradigm, during training only the GMM means are adapted. A GMM can therefore be conveniently expressed as $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^T \ldots \boldsymbol{\mu}_C^T\right]^T$, which is a $CF \times 1$ supervector.

The factor analysis model used for this study is a joint model of both speaker and session variability. It is useful to consider both of these sources of variation in isolation before defining the full joint factor model.

2.1 Speaker Variability

A combined relevance MAP [5] and subspace (eigenvoice) MAP [2] model formation is used to model the speaker-dependent variation. It is argued that the combined relevance and subspace form is able to provide accurate speaker models with limited available training, whilst also allowing for more detailed modelling when large amounts of training data are available.

For the model of speaker variation considered, the speaker-dependent GMM mean supervector can be represented by

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

In this model, $\boldsymbol{V}$ is a low-rank transformation matrix, and $\boldsymbol{D}$ is a $CF \times CF$ diagonal matrix. It is assumed that the majority of speaker variation is contained within the low-rank subspace defined by $\boldsymbol{V}\boldsymbol{V}^*$. The role of $\boldsymbol{D}\boldsymbol{z}(s)$ is to model the *residual* variability that is not captured by the speaker subspace. The vector $\boldsymbol{y}(s)$ is referred to as the speaker factors, and represent the parameters of the speaker in the specified subspace. The

speaker variability model is trained such that $\boldsymbol{y}(s)$ follows a standard normal distribution.

2.2 Session Variability

A similar decomposition is used to describe a model of inter-session variation. The GMM supervector representation of an utterance may be considered as the combination of a session-independent model with an additional offset of the model means representing the recording conditions of the session h. This can be expressed as

$$\boldsymbol{\mu_h}(s) = \boldsymbol{\mu}(s) + \boldsymbol{U}\boldsymbol{x}_h(s). \quad (2)$$

In this representation, $\boldsymbol{U}$ is a low-rank transformation matrix. The range of $\boldsymbol{U}\boldsymbol{U}^*$ can be thought of as defining a session effects space. We refrain from using the term *channel space* as the model also encaptures other forms of intra-speaker and session variation. The vector $\boldsymbol{x}_h(s)$ is an estimate of the session conditions (or latent session factors) within the session subspace, and follows a standard normal distribution.

2.3 Joint Factor Model

A joint factor representation can be obtained by combining the formulations in (1) and (2). The full joint factor model is then described by the set of speaker-independent *hyperparameters* $\boldsymbol{m}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{D}, \boldsymbol{\Sigma}$

The process of speaker model training involves the simultaneous estimation of the latent speaker, session and relevance factors. This is a non-trivial task requiring the decomposition of a very large matrix [1]. A direct solution to this optimisation problem is possible, however, this work employs an efficient, iterative algorithm based on the Gauss-Seidel approximation method [4].

The hyperparameters used in this joint factor analysis model require estimating through an offline training process. A standard background model GMM can be utilised as a source of estimates for the speaker-independent mean $\boldsymbol{m}$ and component covariance matrices $\boldsymbol{\Sigma}_c$. Also, to estimate the diagonal relevance MAP loading matrix $\boldsymbol{D}$, the empirical method outlined by Reynolds in [5] can be used. This method states that $\boldsymbol{D}$ is constrained to satisfy $\boldsymbol{I} = \tau \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{D}$ where τ is the relevance factor and $\boldsymbol{\Sigma}$ is a diagonal matrix consisting of the UBM component covariance matrices $\boldsymbol{\Sigma}_c$.

The remaining problem is to estimate the speaker and session variability transformation matrices $\boldsymbol{V}$ and $\boldsymbol{U}$.

3 Training the Speaker and Session Variability Subspaces

For the factor analysis model described in this paper to be effective, the speaker and session variability subspaces — described by the transformation matrices $\boldsymbol{V}$ and $\boldsymbol{U}$ — must be appropriately estimated. These matrices should represent the types of inter- and intra-speaker variations expected within and between recording sessions. To this end, the subspaces are trained on a database containing a large number of speakers each with several independently recorded sessions. This training database should include a variety of channels, handset types and environmental conditions that closely resembles the conditions on which the eventual system is to be used.

Estimates for the transformation matrices $\boldsymbol{V}$ and $\boldsymbol{U}$ can be obtained in different ways. Four different options for how to obtain these subspace transformation matrices are examined in this paper.

3.1 Principal Component Analysis

Each utterance in the training dataset is first converted into a single observation by training a relevance MAP adapted GMM.

The within- and between-class scatter matrices are then calculated from these observations to capture the intra-speaker and inter-speaker variability, respectively. The principal components of these scatter matrices are determined through eigen decomposition, with the factors corresponding to the R_x and R_y largest eigenvalues retained and used to form the transform matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively.

This PCA analysis forms a good starting point for further analysis but has some shortcomings. Firstly, through the relevance MAP adaptation process, each utterance is reduced to a single point estimate. This approach does not fully use all data available when calculating the transformation matrix. Secondly, this approach does not use the same optimisation criterion or training method as used in speaker model training and will therefore be suboptimal for this task.

3.2 Simultaneous Estimation of V and U

The simultaneous approach refines $\boldsymbol{U}$ and $\boldsymbol{V}$ through an EM algorithm with the speaker and session factors $\boldsymbol{y}(s)$ and $\boldsymbol{x}_h(s)$ as hidden variables. A maximum likelihood criterion over the entire dataset is optimised with each speaker s optimised as per the speaker model training described above. This method is described in [1] with the transformation matrix optimisation equations presented in [6].

This approach addresses the issues highlighted for the PCA approach, specifically using all data available in the matrix optimisation as well as optimising the same criterion as the speaker model enrollment. Compared to PCA, the simultaneous approach therefore provides a more refined and theoretically optimal solution for training both $\boldsymbol{U}$ and $\boldsymbol{V}$.

As the simultaneous method employs ML as the optimisation criterion it will fit the training data as well as it can. This result may actually not be desirable, considering the purpose of having separate subspaces. Specifically, these subspaces have been termed "speaker" and "session" subspaces but there is no means within an ML framework to constrain the $\boldsymbol{U}$ to capture only session variability and not capture speaker variability. If, for instance, $\boldsymbol{V}$ is not of high enough rank, there will be significant speaker variability captured by $\boldsymbol{U}$ as the next possible part of the FA model.

As the value and information contained in $\boldsymbol{x}_h(s)$ is effectively discarded in speaker model training, any speaker information captured in $\boldsymbol{U}$ will also be discarded.

It should be noted that the simultaneous optimisation is performed under the assumption that $\boldsymbol{D} = 0$, to ensure that as much of the observed variability is modelled by the low-rank speaker and session spaces.

3.3 Disjoint Estimation of V and U

Alternatively, $\boldsymbol{U}$ and $\boldsymbol{V}$ can be optimised independently in an attempt to explicitly capture the variability they are intended to model.

$\boldsymbol{U}$ is again trained using the optimisation equations presented in [6] but with $\boldsymbol{\mu}(s)$ estimated by a very loosely constrained relevance MAP, that is, by setting τ to be very small. As the relevance MAP adaptation will be preferred to model any common speaker characteristics found across sessions for a given speaker s in the training dataset, $\boldsymbol{U}$ will be preferred only to capture the *differences* between sessions of the same speaker, that is, the *inter-session* variability.

$\boldsymbol{V}$ is trained without $\boldsymbol{U}$ as part of the model and with no relevance MAP ($\boldsymbol{D} = 0$). This approach forces $\boldsymbol{V}$ to represent as much of the variability in the training dataset as possible.

This disjoint optimisation approach will generally produce a lower overall likelihood than the previously described simultaneous approach as it is not directly optimising the ML criterion, however, it can be argued that it is in fact superior as $\boldsymbol{U}$, particularly, is more likely to fulfil it's role in modelling *only* the session variability.

3.4 Coupled Estimation of $\boldsymbol{V}$ and $\boldsymbol{U}$

The coupled approach is very similar to the disjoint approach described above with the exception that an attempt is made to explicitly remove session variation during the optimisation of $\boldsymbol{V}$ by incorporating a pre-trained session variability component ($\boldsymbol{U}$). In the disjoint approach, session variability was effectively averaged across sessions of the same speaker. Under the coupled approach, variability likely to be caused by session conditions, as described by $\boldsymbol{U}$, is modelled explicitly.

$\boldsymbol{U}$ is first trained in an identical fashion to the disjoint estimation procedure.

Once optimisation of $\boldsymbol{U}$ is complete, $\boldsymbol{V}$ is optimised in the same fashion as the simultaneous approach by including $\boldsymbol{U}$ into the FA model for each speaker. This optimisation, however, is performed with the caveat that $\boldsymbol{U}$ is held constant rather than re-estimated. The optimisation of $\boldsymbol{V}$ is once again performed under the assumption of no relevance MAP component ($\boldsymbol{D} = 0$).

3.5 Optimising for Short Utterance Conditions

A full factor analysis model incorporating session factors has not provided good results when short utterances (with 20 seconds or less) are used for testing and training [4, 3]. This poor performance can seemingly be attributed primarily to the inclusion of session factor modelling, as the inclusion of speaker factor modelling alone has provided improved performance in [3].

Two potential causes may explain the poor performance obtained when including the session variability model.

Firstly, the session factors may be poorly estimated with a short utterance. This is more damaging than poorly estimated speaker factors as the information modelled by the session factors is subsequently discarded whereas the speaker factors provide an additional offset to the speaker model.

Secondly, it is hypothesised that the $\boldsymbol{U}$ trained on longer utterances does not well characterise the differences observed between short utterances. Particularly, phonetic variation may be a much more dominant aspect of inter-session variability for short utterances than long utterances, as longer utterances will have a much higher phonetic coverage than only a few seconds of speech.

It is difficult to directly address the first of these points, but the second can be investigated by estimating subspace transforms on sessions of a matched length to the expected training and testing conditions. The effect of matched-length transform training, particularly for $\boldsymbol{U}$, will be explored in Section 4.2.

4 Experiments

The NIST 2005 Speaker Recognition Evaluation 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.

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 $\tau = 16$ and 512-component models and speaker and session variability subspaces of dimension $R_y = 200$ and $R_x = 50$ are used throughout. 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.

System	EER	Min. DCF
Baseline	9.17%	0.0439
FA PCA	5.87%	0.0241
FA Simultaneous	5.64%	0.0234
FA Disjoint	5.48%	0.0223
FA Coupled	5.56%	0.0217

Table 1: EER and minimum DCF on the 1-conv female subset of the 2005 NIST SRE common evaluation condition for alternate hyperparameter estimation methods.

4.1 Comparison of Subspace Estimation Techniques

Table 1 presents results comparing the four alternative algorithms described in Section 3 for estimating the speaker and session hyperparameter transformation matrices. Results are also provided for the baseline GMM-UBM recognition system where only standard relevance MAP training is performed.

Immediately obvious from the presented results is that all four training algorithms provide substantial improvements over the baseline system. Even the simple PCA-based factor analysis model provides a significant reduction in error rates. The results also demonstrate that further improvements in performance can be achieved through an iterative refinement of the PCA-based models. The EM-based training algorithms all obtain lower EER and minimum DCF values than the basic PCA-trained factor analysis.

Comparing the performance of the individual EM-based refinement algorithms, some interesting observations can also be made. Although the simultaneous solution improves over the seed PCA models, more appropriate refinements of the subspaces appear to be achieved through use of either the disjoint or coupled algorithms. As suggested in Section 3, it appears that a pure ML optimisation does not provide the best overall speaker verification performance as useful speaker information may be modelled in the the "session" subspace and subsequently discarded.

As stated, better solutions to estimating the session and speaker hyperparameters are provided by the disjoint and coupled algorithms. By independently optimising the session space, and allowing residual information to be accounted for by the relevance MAP component, one avoids discarding useful speaker information in the session adaptation process. The results suggest that this is indeed the case, with improvements in performance over both the simultaneous and PCA-based algorithms observed. Comparing these two algorithms, the disjoint algorithm achieves a better EER than the coupled, however, the reverse is true for the defined minimum DCF.

4.2 Reduced Training Lengths

The factor analysis speaker verification techniques have been tested and proven to be effective when large amounts of training and test data (eg. whole conversation sides) are available. Ideally, a speaker verification system should also produce accurate models when the training data is limited. Of particular interest is whether the speaker and session subspaces should be trained using as much data as a possible, or using segments matched to the length used for speaker model training and testing.

To investigate the performance and effect of reduced length scenarios, a range of shortened utterance lengths was examined.

	Subspace Training			
System	$\boldsymbol{V}$	$\boldsymbol{U}$	**EER**	**Min. DCF**
Full-length	1 conv	1 conv	13.47%	0.0544
Matched	20 sec	20 sec	12.04%	0.0498
Mixed	1 conv	20 sec	11.70%	0.0493

Table 2: EER and minimum DCF on a modified 20 second train/test condition for the female subset of the 2005 NIST SRE. Results are presented for systems using subspaces trained on different length segments.

The shortened utterances were obtained by truncating the utterances of the NIST2005 1conv4w-1conv4w condition to the specified length of active speech data for both training and testing. Utterance lengths of 10, 20, 40 and 80 seconds were examined, as well as the full available conversation side for comparison purposes (typically with 100–120 seconds of active speech). For all factor analysis configurations considered, the coupled training algorithm described in Section 3.4 was used.

The first set of experiments examined the performance of the factor analysis model on the 20 second training test. Table 2 shows the performance of three different factor analysis systems for this condition. The first system utilised subspaces trained using the full-length conversation sides. The second system matched the subspace training to the lengths expected in speaker model training/testing, optimising both speaker and session spaces on reduced 20s segments. Finally, a third system (Mixed) consisted of a speaker subspace trained on full-length recordings and a session space trained on the 20s segments.

The results in Table 2 show that better performance is achieved when the session transform is trained using 20s segments as opposed to full-length utterances. Matching the session subspace training to use the shorter segments results in relative improvements of least 8% for both minimum DCF and EER.

Interestingly, the mixed transform training-length system outperforms the matched system. This result suggest that the speaker subspace training benefits from increased training data, however, in estimating the session subspace it is important to use data that matches the conditions and length expected in speaker model training and scoring. The need for a matched session subspace adds weight to the discussion in Section 3.5, suggesting that the inter-session variability has substantially different characteristics when short utterances are used. Table 4 adds more support to this argument as the trace of the session subspace increases significantly as the utterance length is reduced. This additional variability may be attributed to an increase in the significance of phonetic variation between shorter utterances.

In Table 3 results are presented for the other shortened train/test conditions. A comparison is made between a GMM-UBM baseline, a full-length trained factor analysis and finally a factor analysis system using the mixed-length training approach. The results show the same trend as observed for the 20s condition. For all training conditions, the best performance is obtained when the session transformation matrix is optimised on utterances matched to the evaluation condition. It can be seen, however, that the FA model still has a diminishing advantage as the utterance length is reduced as previously observed [4, 3]. This result adds support to the first hypothesis in Section 3.5.

A useful implications of these results is that the speaker subspace does not appear to need retraining to match the training/testing segment lengths.

System	**80 sec**	**40 sec**	**20 sec**	**10 sec**
Baseline	0.0442	0.0501	0.0617	0.0753
FA Full-length	0.0238	0.0346	0.0544	0.0797
FA Mixed	0.0234	0.0337	0.0493	0.0708

Table 3: Minimum DCF on the female subset of the 2005 NIST SRE common evaluation for reduced utterance length conditions.

Utt. Length	**1 conv**	**80 sec**	**40 sec**	**20 sec**	**10 sec**
$tr(\boldsymbol{U}\boldsymbol{U}^*)$	105.7	116.9	148.8	213.0	329.8

Table 4: Trace of the session subspace covariance with $\boldsymbol{U}$ trained with different length utterances.

5 Conclusions

A key step to ensuring the success of a factor analysis approach is the appropriate estimation and optimisation of the speaker and session variability subspace transformation matrices. The manner in which the estimation is performed, as well as the way the data is prepared and presented to the optimisation algorithms must be considered.

A number of alternate algorithms for this estimation of the subspace transformation matrices were evaluated. Experiments showed that better performance can be obtained when independent rather than simultaneous optimisation of the two core variability subspaces is used. The outlined disjoint and coupled algorithms achieved the best performance.

The effect of shortening the utterances used in subspace training and speaker model training and scoring was also examined. Results suggest that in training the speaker subspace, as much data possible should be used. For the session variability subspace however, it appeared important to optimise on data matched to the conditions and utterance lengths expected to be encountered in training and testing.

6 Acknowledgements

This research was supported by the Australian Research Council Discovery Grant No DP0877835.

7 References

[1] 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.

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

[3] R. Vogt, C. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[4] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

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

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

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