On the use of factor analysis with restricted target data in speaker verification

Javier Gonzalez-Dominguez[2], Brendan Baker[1], Robbie Vogt[1],
Joaquin Gonzalez-Rodriguez[2] and Sridha Sridharan[1]

[1]Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
[2]ATVS- Biometric Recognition Group, Escuela Politecnica Superior, Madrid, Spain

javier.gonzalez@uam.es, bj.baker@qut.edu.au, r.vogt@qut.edu.au

Abstract

Factor Analysis (FA) based techniques have become the state of the art in automatic speaker verification thanks to their great ability to model session variability. This ability, in turn, relies on accurately estimating a session variability subspace for the operating conditions of interest. In cases such as forensic speaker recognition, however, this requirement cannot always be satisfied due to the very limited quantity of appropriate development data. As a first step toward understanding the application of FA in these restricted data scenarios, this work analyzes the performance of FA with very limited development data and then explores several FA estimation methods that augment the target domain data with examples from a data-rich domain. Experiments on NIST SRE 2006 microphone data conditions demonstrate that telephone data can be effectively exploited to improve performance over a baseline system.

Index Terms: Session Variability, Factor Analysis, Forensic Speaker Recognition.

1. Introduction

The most successful text-independent Automatic Speaker Verification (ASV) systems in recent years have utilised some form of Factor Analysis (FA) as a technique for modelling both session and speaker variability. Systems using FA have gained prevalance due to their enhanced ability to deal with complex sources of intersession variation.

Factor Analysis, applied in the context of text-independent speaker verification, was first proposed by Kenny, *et al*. [1] integrated in a classical Gaussian Mixture Model (GMM). This initial work laid the foundations for later works which have successfully applied FA in both generative GMM [2] and discriminative SVM models [3], as well as at different levels of a ASV system [4, 1].

Although a number of variations in the configuration and implementation of the Factor Analysis model have been proposed, all of these techniques share the same basic principles: addressing the variability in a continuous manner, and making the assumption that the majority of speaker and session variation can be modelled with a small number of variables. Mathematically, these two concepts are embedded in the FA model, where the variability (speaker and session variability) is described through subspaces that encapsulate the corresponding main directions of variation. These subspaces are estimated from a background dataset and take in practice the form of low-rank transformation matrices.

From a statistical point of view, the variability subspaces act as a strong priors since target data (the data used in operational conditions) is constrained by the directions of variability described by these subspaces. As a consequence, an important issue in the successful application of the FA model is appropriate training of the subspace transform matrices. Ideally, these matrices should accurately represent the types of inter- and intra-speaker variations expected within and between recording sessions. In order to estimate the subspaces, appropriate training data (background data) which represents as much as possible the target conditions is required.

This paper considers the problem of data availability for training the FA subspaces, and the appropriate estimation of these subspaces. A particular focus is given to problems of data availability that may arise in the context of forensic applications.

In some forensic caseworks, the target conditions are well understood and significant amounts of data recorded under similiar conditions is available. Unfortunately, this is not always the case. In many forensic situations, the target conditions are highly variable, and the amount of example data available for "learning" the session conditions is extremely limited.

This work deals with the problem of applying FA techniques in these unfavourable cases where only a scant amount of data from the targetted operating conditions is available for estimation of the session variability subspace. The study explores the possibility of exploiting other available datasets (outside the target conditions) to improve the performance of the system in such situations.

The remainder of this paper is organised as follows. In Section 2, a summary of the FA framework is outlined. The problem of training the session variability subspace

is addressed in Section 3, where different strategies are presented. Experiments are presented that analyse the verification performance under simulated restricted training data scenarios. A comparison is made between the various alternatives for subspace estimation, showing the benefits that can be achieved through adjustments to the standard training approaches.

2. Factor Analysis for Speaker Verification

2.1. The JFA Model

The factor analysis model used for this study is a joint model of both speaker and session variability. As in traditional GMM-UBM approaches to speaker verification [5], factor analysis techniques are based around the use of mean-adapted GMM's to model a speaker. In the GMM-UBM paradigm, only the GMM means are adapted from the UBM during speaker enrollment. This GMM can therefore be conveniently expressed as $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^T \dots \boldsymbol{\mu}_C^T\right]^T$, which is a $CF \times 1$ mean supervector, where F is the dimension of the acoustic feature vectors and C is the number of GMM components.

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.

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

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

that is, as a linear offset from the UBM mean supervector $\boldsymbol{m}$. 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.

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

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

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.

A joint factor representation can be obtained by combining the formulations in (1) and (2). The process of speaker model training therefore involves the simultaneous estimation of the latent session, speaker and relevance factors, $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$, with the session factors subsequently discarded. This work employs an efficient, iterative algorithm based on the Gauss-Seidel approximation method for this task [6].

2.2. Subspaces Training Procedure

The full joint factor model is then characterised by the set of speaker-independent *hyperparameters* $\boldsymbol{m}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{D}$. These hyperparameters are estimated through an off-line training process as described in [7].

The UBM is used as a source of estimates for the speaker-independent mean $\boldsymbol{m}$. Also, to estimate the diagonal relevance MAP loading matrix $\boldsymbol{D}$, the empirical method outlined by Reynolds in [5] is be used: $\boldsymbol{D}$ 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 hyperparameters $\boldsymbol{V}$ and $\boldsymbol{U}$ describe the speaker and session variability subspaces, respectively. For the factor analysis model described in this paper to be effective these 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.

In this work, $\boldsymbol{U}$ and $\boldsymbol{V}$ were optimised independently in an attempt to explicitly capture the variability each subspace was intended to model. This method is described as the *disjoint* estimation approach in [7].

Under the disjoint approach, $\boldsymbol{U}$ is trained using the optimisation equations presented in [8] 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.

Similarly, $\boldsymbol{V}$ is trained by excluding $\boldsymbol{U}$ from the model, that is using the model in (1), and with no relevance MAP ($\boldsymbol{D} = \boldsymbol{0}$). This approach forces $\boldsymbol{V}$ to repre-

sent as much of the variability in the training dataset as possible with the constraint that all utterances from the same speaker have equivalent speaker factors, $\boldsymbol{y}$.

Both $\boldsymbol{U}$ and $\boldsymbol{V}$ were refined using an Expectation-Maximisation (EM) algorithm [8] after a principal components initialisation [7].

3. Training Session Variability Subspaces with Restricted Target data

As it was noted previously, the success of FA modelling is highly dependent on the proper estimation of the session variability subspace represented by $\boldsymbol{U}$. For this purpose, a suitable dataset that accurately represents the conditions of the target domain is essential.

Unfortunately, this requirement for suitable data cannot be satisfied in all situations. Forensic Speaker Recognition is an area that gives us a wide range of examples of this situation, mainly due to two factors. Firstly, despite the efforts made to collect new databases [9], the available data is still very limited. Secondly, real world forensic recordings tend to be extremely variable, making a case-by-case treatment necessary in most situations. In those cases where only a limited amount of data is available, the estimation procedure described above leads to poorly estimated variability subspaces since the real variability in target domain is not sufficiently represented.

The underlying idea of this work is to deal with this limited data problem by exploiting data from a data-rich domain in the session subspace estimation procedure in order to achieve a dual goal. First, to obtain a more robust estimation procedure by adding large amounts of data. Secondly, to incorporate certain session variability characteristics not present in the limited available target domain data but that could appear in the target domain. Three techniques for combining information from a data-rich domain and limited target domain data are presented in the remainder of this section.

3.1. Joining Matrices

A simple way to combine different session variability subspaces is to join session variability subspaces estimated on different datasets. This process is carried out by simply stacking the session variability directions estimated in each one of them in a bigger subspace.

This approach has the major advantage that subspaces can be treated and trained independently. From a practical point of view, this property is highly desirable because it allows us to keep a well-trained reference subspace trained on accumulated data that can be refined by simply appending new session variability information from new domains.

On the other hand, it has several shortcomings. Firstly, it is necessary to restrict the size of each contributing subspace, loosing potentially useful directions of variability, in order to keep the overall size of the joined subspace relatively small as stipulated by the principles of FA. Second no particular emphasis is placed on the target domain data because all the directions play an equal role in the new subspace. Finally, even the main directions of session variability will tend to be poorly estimated for the target domain if there is severely limited data as the subspaces are estimated indepedendently.

3.2. Pooled Statistics

As an alternative to stacking two independently trained subspaces, the subspace estimation can also be supplemented with the data-rich telephone set simply by estimating a completely new session subspace. This time, estimation is performed by pooling all data. An obvious advantage of this method is that the estimation is performed using a substantial amount of data, making it potentially more robust. Unfortunately, there is no means of preventing the supplementary set dominating the estimation and having the biggest effect on the directions of variability.

3.3. Scaling Statistics

Based on the fact that we are usually most interested in the session variability present in a specific domain (the closest to the target domain conditions), it is reasonable to think that somehow these data should become more important in the subspace estimation procedure. Moreover, we should be able to get some advantage by using all the data available together rather than separately.

The approach presented here is based on giving a specific weight to each dataset in the training session variability subspace with a dual purpose. First, allow the estimation procedure to learn from a broader set of data leading us to more robust subspace estimation, and second to highlight the type of data which is considered most important. This second point is especially necessary when not enough data of this type is available and the variability presented could be *overshadowed* by the other types.

Specifically, first order statistics supervector extracted from each utterance is scaled by a previous fixed weight depending on the dataset to which it belongs. Thus, the matrix of first order statistics $\boldsymbol{S}$, input in the EM procedure for training the variability subspace take the following form:

$$\boldsymbol{S} = \alpha \boldsymbol{S}_{tgt} + (1-\alpha)\boldsymbol{S}_{bckg} \tag{3}$$

where $\boldsymbol{S}_{bkg}$ and $\boldsymbol{S}_{tgt}$ are the matrices whose columns are the first order statistics of utterances belonging to target data and other background data available respectively.

More generally, this could be extend to:

$$\boldsymbol{S} = \alpha_1 \boldsymbol{S}_1 \alpha_2 \boldsymbol{S}_2 ... + \alpha_n \boldsymbol{S}_n \tag{4}$$

with $\sum_{i=0}^{n} \alpha_i = 1$ and n different background sets.

In this way it is possible to balance the weight of each subset in the EM procedure such that the available data can be combined in an optimal way for the task at hand. A difficulty with this approach is the problem of finding the optimal selection of weights for each problem. Although this can be solved empirically, a reasonable option is choose the weights in a proportional way to the quantity of data in target domain, keeping at least a minimum weight for the rest of the sets.

4. Experiments

Evaluations are concentrated on examining scenarios in which the available background data similar to the acoustic conditions of the target domain is scarce. Comparisons are made with the FA baseline system described in Section 2.

4.1. Database and protocol

Data from the 2005 and 2006 NIST Speaker Recognition Evaluations (NIST SRE) was used to develop an experimental framework. These datasets were chosen for two main reasons. First, the datasets cover a wide range of acoustic (telephone and microphone) and environmental scenarios, allowing for vigourous testing under mismatched conditions. Secondly, following the well established NIST SRE protocols allows for future comparisions with other research groups.

Two development datasets, namely $dTel$ and $dMic$, were differentiated. The $dTel$ consists of SRE'04 and SRE'05 telephone data suplemented with data belonging to SWBII phase I and phase II databases. This collection was chosen to provide a broad coverage of telephone conditions, whilst also providing a high number of different speakers. The $dMic$ dataset was obtained from the microphone subset of the MIXER corpus and SRE'05 data.

In order to simulate the data scarcity problem, the $dMic$ set was divided into sets with differing amounts of data, obtaining different degrees of data scarcity. Specifically, three restricted sets were built: $dMic_{10}$, $dMic_5$ and $dMic_3$. These were formed with only 10, 5 and 3 utterances per speaker present in $dMic$. Table 1 shows a breakdown development dataset compositions.

The SRE'06 data was utilised as the test dataset. Testing was performed using the test conditions specified in the NIST SRE'06 protocol, and using additional conditions specified and distributed by participating sites during the SRE'08 [1]. The test conditions examined were as follows: *1conv4w-1conv4w*, *1conv4w-1mic*, *1mic-1conv4w* and *1mic-1mic*.

[1] Additional conditions for auxiliary microphone training and testing were distributed on the SRE'08 Google Group list. Thanks to Doug Reynolds, David van Leeuwen, Albert Strasheim and Nicholas Scheffer for preparing and scrutinising these lists. Further details on these conditions can be obtained from the author or at http://groups.google.com/group/sre2008

	Databases	**# Speakers**	**# Utterances**
	SWB-II	325	1300
$dTel$	MIXER(SRE'04)	150	994
	MIXER(SRE'05-tel)	40	297
$dMic$	MIXER(SRE'05-mic)	45	1260
$dMic_{10}$	MIXER(SRE'05-mic)	45	450
$dMic_5$	MIXER(SRE'05-mic)	45	225
$dMic_3$	MIXER(SRE'05-mic)	45	135

Table 1: *Composition of development datasets.*

4.2. Results

As a starting point of this study, the effect of using restricted datasets in order to estimate a session variability subspace was analysed. For this purpose, the baseline JFA system presented in 2 was evaluated using the differing restricted microphone datasets described in Section 4.1 as training data for the low-rank session matrix $\boldsymbol{U}$. The results in Table 2 summarise the performance statistics of these restricted subspace training data experiments. Studying these results, it can be seen that when microphone data scarcity is simulated in the development stage (i.e. the amount of training data for $\boldsymbol{U}$ is reduced), system performace is degraded significantly. It is clear from these results alone that data availability for training the channel subspace has a large impact on overall performance.

For comparison purposes, results for a baseline system that does not include session compensation ($U = 0$) were also included in Table 2. It is obvious from the results that incorporating session compensation leads to significant improvements in performance across all train/test conditions. Interestingly, even when the data used to estimate the session subspace is mismatched to the conditions (channel type) of the evaluation trials, the inclusion of session compensation always results in an improvement. A session matrix estimated using purely telephone data reduces the error rates in the *1mic-1mic* condition. Similarly, a session matrix estimated using microphone data for telephone based trials provides some benefits over no session compensation at all. Expectedly, the best performance is achieved when the session subspace is trained using appropriate data (eg. $dMic$ used for *1mic-1mic*).

Experiments were then performed to examine whether the data rich sources - in this case the telephone data - could be used alongside the restricted data in the estimation of the session variability subspace $\boldsymbol{U}$, in order to improve the estimation and in turn, the overall performance. The first approach considered for this task was the joint subspace approach as outlined in Section 3.1. A new session variability subspace was generated simply by stacking two independently trained session subspaces, one estimated using the $dTel$ set and the other using the target domain data $dMic$. For this combination strategy, the top 50 and 20 eigenchannels from $\boldsymbol{U}_{\boldsymbol{dTel}}$ and $\boldsymbol{U}_{\boldsymbol{dMic}}$, respectively, were used to create a 70 eigenchannel joint

subspace [2]. The performance using both the full and restricted datasets are presented in Table 3.

Comparing the results in Table 3 with those in Table 2, it can be seen that supplementing the subspace training data with telephone data has a positive effect across nearly all evaluated tasks. While this effect seems obvious in those conditions where telephone data is involved, it is worth noting that even in the case condition *1mic-1mic*, including telephone data alongside the available microphone data in the subspace development stage is clearly beneficial. This suggests that it is possible to account for some session variability even in very apparently different acoustic subspaces. The biggest gains from supplementing the target domain microphone data with telephone data were observed when the target domain (microphone) data was restricted. For the most restricted training scenario $dMic_3$, a relative improvement of 28% resulted for the *1mic-1mic* condition when $dMic_3$ was supplemented using $dTel$.

As outlined in Section 3.2, a new subspace can also be estimated by pooling the statistics from both the data-rich set and target domain set. Results using this pooling method are presented in Table 4. An interesting point to highlight here is the case where the full microphone datset $dMic$ is available for subspace estimation. In this case, an improvement in performance over the joint matrix technique is observed for the case condition *1mic-1mic*. When less target domain (microphone) data is available for the subspace estimation, we see that the effectiveness of the session compensation is reduced when pooled statistics rather than stacked matrices are used. This suggests that for the pooled approach, the subspace estimation is being overwhelmed by the larger quantity of telephone data, and is not able to best utilise the available (but restricted) target domain data.

Finally, the method proposed in Section 3.3, where more emphasis can be placed on data from the target domain by performing a scaling of the statistics during subspace estimation, was evaluated. Results in Table 4 show the performance using various scaling weights, α. For these experiments, the closest simulation of real forensic applications, where only 3 utterances per speaker in $dMic$ was made available for subspace estimation was studied ($dMic_3$). It can be seen from these results, that in general, placing a larger weighting on the $dMic_3$ statistics results in an improvement in performance over straight pooling (unweighted). For the case condition *1mic-1mic*, a scaled statistics estimation results in a 6% relative improvement in EER over the straight pooling.

Figure 1 shows a final comparison of the considered estimation strategies for the session subspace, evaluated on the 1mic-1mic condition with only a limited amount of target domain data available ($dMic_3$). This chart clearly demonstrates the benefit of session compensation, but also the problems associated with a direct estimation of the subspace on a small dataset. Better results are achieved when subspace estimation is performed using the data-rich $dTel$ rather than $dMic_3$ alone. Importantly though, benefits result from supplementing the $dMic_3$ with other data. Each of the strategies for combining the two sets in estimation give improvements over either alone. The joint estimation approach using stacked subspaces achieves a better result than a straight pooling of the data, however, this trend can be reversed by introducing a simple scaling of the statistics during estimation. By weighting the target domain data more heavily during estimation, the best performance out of the considered approaches is achieved.

[2] An analysis of the eigenvalues for the microphone data showed a very rapid decline in values in comparison to the telephone data. For this reason, a reduced number (20) of dimensions were retained.

	Equal Error Rate (EER in %)			
$\boldsymbol{U}$ **Training**	1conv4w/ 1conv4w	1conv4w/ 1mic	1mic/ 1conv4w	1mic/ 1mic
$\boldsymbol{U} = 0$	5.97	8.20	7.81	11.03
$dTel$	3.49	4.31	3.95	6.79
$dMic$	5.80	5.19	5.30	6.64
$dMic_{10}$	5.99	5.69	5.50	7.51
$dMic_5$	5.93	6.06	5.72	8.07
$dMic_3$	5.99	6.13	5.72	8.33

Table 2: *Performance under restricted MIC data conditions in* $\boldsymbol{U}$ *training.*

	Equal Error Rate (EER in %)			
$\boldsymbol{U}$ **Training**	1conv4w/ 1conv4w	1conv4w/ 1mic	1mic/ 1conv4w	1mic/ 1mic
$dMic \vert dTel$	3.41	3.63	3.12	5.14
$dMic_{10} \vert dTel$	3.55	3.72	3.32	5.43
$dMic_5 \vert dTel$	3.55	4.15	3.63	5.74
$dMic_3 \vert dTel$	3.55	4.31	3.54	6.03

Table 3: *Performance using the joint matrices subspaces estimation approach.*

	Equal Error Rate (EER in %)			
$\boldsymbol{U}$ **Training**	1conv4w/ 1conv4w	1conv4w/ 1mic	1mic/ 1conv4w	1mic/ 1mic
$dMic + dTel$	3.73	3.54	3.43	4.97
$dMic_{10} + dTel$	3.61	3.72	3.43	5.47
$dMic_5 + dTel$	3.42	3.88	3.66	5.78
$dMic_3 + dTel$	3.49	4.12	3.76	6.19

Table 4: *Performance under restricted microphone data conditions when statistics are pooled with devTel.*

	Equal Error Rate (EER in %)			
Scaling (α)	1conv4w/ 1conv4w	1conv4w/ 1mic	1mic/ 1conv4w	1mic/ 1mic
$Unweighted$	3.49	4.12	3.76	6.19
0.6	3.46	4.15	3.74	5.95
0.7	3.55	4.15	3.67	**5.82**
0.8	3.80	4.49	3.32	**5.82**
0.9	4.30	4.64	3.78	6.50

Table 5: *Performance using scaled statistics during ML estimation. Results using 3 mic utterances per speaker*

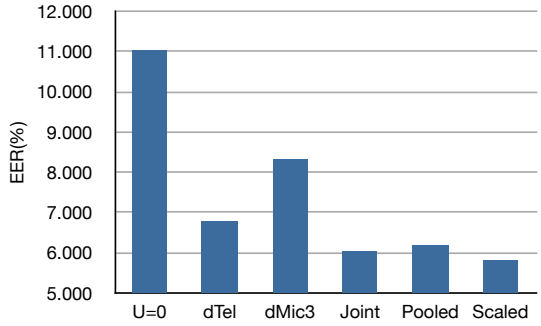

Figure 1: *Comparison of performance for 1mic-1mic condition using different U training strategies*

5. Conclusions

The successful application of FA techniques is highly dependent on the proper estimation of session variability as represented by the subspace transformation matrices. This work has analysed the problem of applying FA in situations where a scant amount of data similar to the expected operating conditions is available — a common situation in forensic speaker recognition work.

A range of experiments using the microphone condition of the well-known NIST SRE 2006 database and protocol were initially conducted exploring the effect of reducing the quantity of available development data. These experiments clearly demonstrated the importance of a well-estimated session variability subspace as using poorly matched telephone data or heavily restricting the available microphone development data resulted in significantly increased error rates. In these situations, current estimation procedures lead to poorly estimated subspaces and consequently far from optimal FA performance.

To deal with this problem, several methods were explored to combine different variability information obtained from different sources of data, including joining subspace matrices, and pooling estimation statistics. These techniques are based on the idea that variability present in different databases can be exploited in order to provide more robust subspace estimates. Experiments with these methods show that a suitable method of combining information from both the target domain and a data-rich development domain can be very useful in the restricted data scenarios, particularly if emphasis can be placed on the limited available target domain data. Specifically, introducing a scaling or weighting on the statistics in the EM algorithm used for training the session variability subspace has demonstrated good gains.

These results open the door to the possibility of applying FA in unfavourable scenarios such as those encountered in most of forensic cases. The lack of suitable development data for training session subspaces in these cases has typically rendered state-of-the-art session compensation approaches unusable.

6. Future Work

The work presented in this paper aims at being an initial step in the application of state-of-the-art session variability compensation in scenarios with very limited target domain data, as is common in forensic casework. Significant future work is planned to further develop an understanding of this important issue as well as develop new techniques. This work will include investigations with authentic forensic case data obtained through the AHUMADA III data collection [9]. Further work will also investigate a maximum a posteriori criterion for estimating session variability subspaces to further enhance the robustness of the techniques presented in this paper.

7. Acknowledgements

Research on this project was carried out at QUT during a research visit by J. G.-D. QUT thanks the Australian Research Council for supporting the facilities used to undertake this collaborative research under the Discovery Grant No DP0557387. J. G.-D. thanks the Spanish Ministry of Education for supporting his doctoral research under the project TEC2006-13170-C02-01. Thanks also to the Spanish Ministry of Education for supporting forensic speaker recognition under the project TEC2009-14719.

8. 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] P. Kenny, P. Oullet, V. Dehak, N. Gupta, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2006, pp. I–97–I–100.

[4] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

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

[7] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, 2008, pp. 853–856.

[8] 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.

[9] D. Ramos, J. Gonzalez-Rodriguez, and J. Gonzalez-Dominguez, J. Lucena, "Adressing database mismatch in forensic speaker recognition with ahumada iii: a public real-casework database in spanish," in *Interspeech*, 2008.