EXPERIMENTS IN SESSION VARIABILITY MODELLING FOR SPEAKER VERIFICATION

Robbie Vogt, Sridha Sridharan

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

ABSTRACT

Presented is an approach to modelling session variability for GMM-based text-independent speaker verification incorporating a constrained session variability component in both the training and testing procedures. The proposed technique reduces the data labelling requirements and removes discrete categorisation needed by previous techniques and provides superior performance. Experiments on Mixer conversational telephony data show improvements of as much as 46% in equal error rate over a baseline system. In this paper the algorithm used for the enrollment procedure is described in detail. Results are also presented investigating the response of the technique to short test utterances and varying session subspace dimension.

1. INTRODUCTION

While research in the field of speaker recognition and verification has been ongoing for many years, the greatest cause of errors still remains the same. The issue of mismatch caused by session variability. This term encompasses a number of phenomena including transmission channel effects, transducer characteristics, environment noise and variability introduced by the speaker.

A number of techniques have been proposed to compensate for various aspects of session variability at almost every stage in the verification process with some success; a state of the art verification system will often incorporate a number of these techniques. An example system [1] from the NIST Speaker Recognition Evaluation might include feature warping [2] and mapping [3, 4] to produce more robust features as well as score compensation techniques such as H- and T-Norm [5].

More recently a method for directly modelling the inter-session variability has been proposed and has provided impressive reductions in verification error rates [6, 7]. The motivation behind the proposed technique is to attempt to directly model session variability in the model space without discrete categories and with less restrictive data labelling requirements. The proposed technique incorporates session differences into the speaker modelling process in the form of session-dependent GMM mean offsets constrained to a low dimensional "session variability" subspace. This effects both the training and testing phases of the system.

This work draws heavily on the results of Kenny, *et al.* [6, 8] with some distinct differences, and builds on the work presented in [7]. In contrast to Kenny, *et al.* [6] the presented approach does not include a "speaker factor" subspace adaptation adopting a more traditional GMM-UBM structure and obviating the need to train a speaker subspace transform and reducing training complexity. A simplified verification score is also used that is more in line with the GMM-UBM approach. Finally, Z-Norm score normalisation is additionally applied to correct for the differing responses of trained speaker models.

This paper presents results on Mixer corpus data complementing the Switchboard-II results previously presented [7]. Section 2 describes the approach to modelling speakers in the presence of session variation, including the approach to representing a speaker during training, and how to exploit this method during testing. A more detailed explanation of the speaker enrollment algorithm is presented in Section 2.2 with results contrasting possible configurations in Section 3.2. Additional investigation is pursued on the effect of varying the dimension of the session subspace and the effect of using very short test utterances in Sections 3.1 and 3.3.

2. MODELLING SESSION VARIABILITY

The approach to modelling the session variability in telephony-based speaker verification adopted in this paper is to introduce a constrained offset of the speaker's Gaussian mixture model mean vectors to represent the effects introduce by the session conditions. In other words, the Gaussian mixture model that best represents the acoustic observations of a particular recording is the combination of a session-independent speaker model with an additional session-dependent offset of the model means. This can be represented in terms of the $R_y \times 1$ concatenated GMM component means supervectors as

$$\boldsymbol{m}_h(s) = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{y}(s) + \boldsymbol{U}\boldsymbol{z}_h(s),$$

where $R_y = MD$ where the GMM is of order M and dimension D.

Here, the speaker s is represented by the offset $\boldsymbol{y}(s)$ from the speaker independent (or UBM) concatenated mean supervector $\boldsymbol{m}$, scaled by the $R_y \times R_y$ diagonal matrix $\boldsymbol{D}$. To represent the conditions of the particular recording, designated with the subscript h, an additional offset of $\boldsymbol{U}\boldsymbol{z}_h(s)$ is introduced where $\boldsymbol{z}_h(s)$ is a low-dimensional representation of the conditions in the recording and $\boldsymbol{U}$ is the low-rank transformation matrix from the constrained session variability subspace of dimension R_z to the GMM mean supervector space of dimension R_y.

With this formulation both the speaker offset $\boldsymbol{y}(s)$ and the session factors $\boldsymbol{z}_h(s)$ are assumed to belong to a standard normal distribution, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Ideally, a training algorithm will be able to accurately discern the session-independent speaker model $\boldsymbol{y}(s)$ in the presence of session variability.

2.1. Speaker Model Training

Speaker models are trained through the simultaneous optimisation of all the model parameters over the set of training observations

This research was supported by the Australian Research Council (ARC) Discovery Grant Project ID: DP0453278.

$\boldsymbol{X}_h(s)$. In this work the model parameters are the component means offset supervector $\boldsymbol{y}$, and the session dependent subspace factors $\boldsymbol{z}_h, h = 1, ..., H$ (s has been dropped from this notation for clarity). The session variability vectors are not actually retained to model the speaker but their estimation is necessary to accurately estimate the true speaker means.

In this work the speaker mean offset supervector $\boldsymbol{y}$ is optimised according to the *maximum a posteriori* (MAP) criterion often used in speaker verification systems [9]. While the prior for this adaptation is the standard normal distribution, as stated above, with $\boldsymbol{D}$ satisfying $\boldsymbol{I} = \tau \boldsymbol{D}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{D}$ this is equivalent to Reynolds' prior with a "relevance factor" of τ; this form is referred to as *relevance MAP* [10].

The MAP criterion is also employed for optimising each of the session variability vectors $\boldsymbol{z}_h$. As described by Kenny, *et al.* [6] the prior distribution is also assumed to be a standard normal distribution in the subspace defined by the transformation matrix $\boldsymbol{U}$. The optimisation of such a criterion has previously been described for speaker recognition problems [10, 6].

The MAP criteria ensure that there is not a "race condition" between the simultaneous optimisation criteria as the prior information ensures a unique (local) optimum.

An EM algorithm is used to optimise the session variability model described above as there are no sufficient statistics for mixtures of Gaussians due to the missing information of mixture component occupancy of each observation. The expectation step is essentially identical to that used in iterative MAP adaptation [11] and maximum likelihood estimation of GMMs, while the maximisation step optimises both the speaker and session variables as described below.

2.2. Maximisation of the Session Variability Model

The direct solution to the simultaneous optimisation equations in the Maximisation step of this EM algorithm is possible, however it requires the decomposition of an $(R_y + HR_z) \times (R_y + HR_z)$ matrix for each iteration. This matrix is required to capture the relationships and cross correlations between the variables being optimised. As $R_y = 12288$ for the size of speaker model used in this work, this is a large matrix decomposition (many current SV systems use significantly larger models). Even with this matrix being positive-definite, this is difficult both in memory and processing requirements.

For this reason a procedure analogous to the Gauss-Seidel method for solving simultaneous equations is used as described in Algorithm 1. In this algorithm, $\boldsymbol{N}_h$ refers to the $MD \times MD$ diagonal component occupancy matrix with M diagonal blocks $N_{hc}\boldsymbol{I}_{D\times D}$ where N_{hc} is the observation count for mixture component c. $\boldsymbol{S}_{X,h}$ refers to the first cumulant vector of the observations $\boldsymbol{X}_h(s)$. These values are the required statistics to solve the MAP adaptation estimates at lines 9 and 11 which are the subspace MAP and relevance MAP estimates described above, respectively.

While this method converges more slowly than direct simultaneous optimisation, each iteration only requires the decomposition of one $R_z \times R_z$ matrix per training session and the trivial decomposition of an $R_y \times R_y$ diagonal matrix for the speaker supervector.

There are several interesting aspects to this algorithm. Firstly the question of the number of iterations that provide the best speaker verification performance. This algorithm adopts an iterative approach to finding the best estimate for the speaker and channel variables in the Gauss-Seidel algorithm and also for the estimation of the missing information of the mixture component occupancy in the EM algorithm. While increasing the iterations will provide models that better optimise the MAP training criterion, this may not necessarily translate to optimal recognition rates.

Algorithm 1 Speaker Model Estimation

1: $\boldsymbol{y} \leftarrow \boldsymbol{0}; \boldsymbol{z}_h \leftarrow \boldsymbol{0}; h = 1, \ldots, H$
2: **for** $i = 1$ to max. iterations **do**
3: **for** $h = 1$ to H **do**
4: Calculate $\boldsymbol{N}_h$ and $\boldsymbol{S}_{X,h}$ for session $\boldsymbol{X}_h$ where $\boldsymbol{\mu}_h = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{y} + \boldsymbol{U}\boldsymbol{z}_h$
5: **end for**
6: $\boldsymbol{N} \leftarrow \sum_{h=1}^{H} \boldsymbol{N}_h$
7: $\boldsymbol{S}_X \leftarrow \sum_{h=1}^{H} \boldsymbol{S}_{X,h}$
8: **for** $h = 1$ to H **do**
9: $\boldsymbol{z}_h \leftarrow \boldsymbol{A}_h^{-1}\boldsymbol{b}_h$
 where $\boldsymbol{A}_h = \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_h\boldsymbol{U}$
 and $\boldsymbol{b}_h = \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X,h|m} - \boldsymbol{N}_h\boldsymbol{D}\boldsymbol{y}\right)$
10: **end for**
11: $\boldsymbol{y} \leftarrow \boldsymbol{A}_y^{-1}\boldsymbol{b}_y$
 where $\boldsymbol{A}_y = \boldsymbol{I} + \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{D}$
 and $\boldsymbol{b}_y = \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_{X|m} - \sum_{h=1}^{H} \boldsymbol{N}_h\boldsymbol{U}\boldsymbol{z}_h\right)$
12: **end for**
13: **return** $\boldsymbol{y}$

Also of interest on line 11 is the term $\sum_{h=1}^{H} \boldsymbol{N}_h\boldsymbol{U}\boldsymbol{z}_h$ that links the results of the session factor estimation to the estimation of the speaker model parameters. Effectively, this calculation subtracts from $\boldsymbol{S}_{X,h}$ the portion explained by the estimated session variable $\boldsymbol{z}_h$. If, instead, the speaker parameters were optimised on the statistic $\boldsymbol{S}_X$ this would be an independent estimation; the equivalent of the Jacobi method.

These points will be investigated in Section 3.2.

2.3. Verification

The session variation introduced in the verification utterance must also be considered. There are a number of possible methods to achieve this that vary considerably in complexity and sophistication. This paper investigates only one possibility that is only marginally more complex than Top-N ELLR scoring [9] (the basis of most current text-independent speaker verification systems). Alternative approaches are discussed in Section 4.

The approach used in this paper is to estimate the session variables $\boldsymbol{z}_h$ of the verification utterance for each speaker prior to performing standard Top-N ELLR scoring. This estimation is similar to that described in Algorithm 1 with a few differences: It is a MAP estimation using the same standard normal prior distribution, however, the speaker supervector is considered known from previous training and not simultaneously estimated. Also, only a single EM iteration is used. To substantially reduce the processing required, a further simplification is made in that the mixture component occupancy statistics for the observations are calculated based on the UBM (rather than independently for each speaker). This allows for only one additional pass of the verification utterance than standard scoring and implies that only one matrix decomposition is necessary, regardless of the number of speakers being tested.

3. EXPERIMENTS

The baseline recognition system used in this study utilises fully coupled GMM-UBM modelling using iterative MAP adaptation and feature-warped MFCC features with appended delta coefficients, as

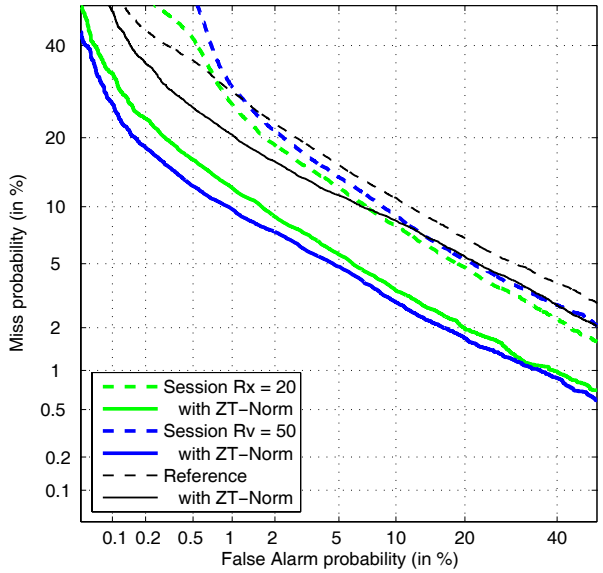

Fig. 1. DET plot for the 1-side condition when varying the number of session subspace dimensions, R_z with and without ZT-Norm score normalisation.

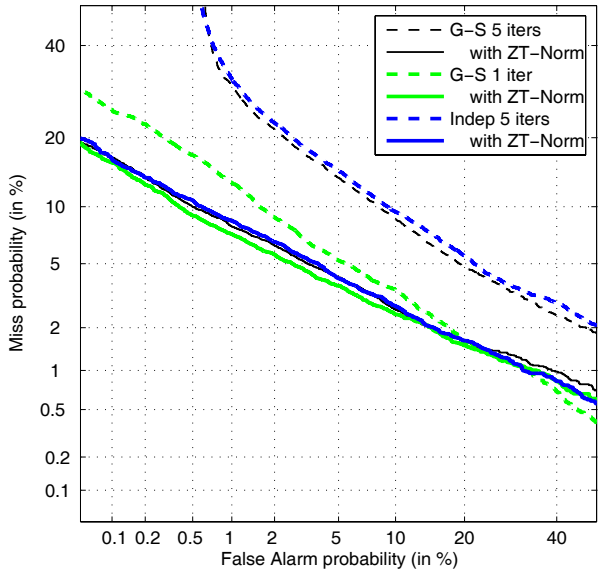

Fig. 2. DET plot for the 1-side training condition comparing optimisation algorithms in speaker enrolment, with and without score normalisation.

described in [2]. An adaptation relevance factor of $\tau = 8$ and 512-component models are used throughout.

The proposed technique was evaluated using data from the NIST 2004 Speaker Recognition Evaluation [12] with a modified testing and training protocol. 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 cellular handsets and channels. The testing and training data from the 2004 evaluation was pooled and then separated into three splits (in a similar fashion to previous NIST EDT evaluations). This separation was in order to facilitate cross validation of individual systems, tuning and evaluation of fusion techniques and selecting of thresholds. A total of 314 speakers with 925 models were tested in 139,084 trials.

3.1. Session Subspace Size

In [7] we noted the importance of severely constraining the dimension of the session variability subspace citing degrading performance comparing results for the $R_z = 50$ case to $R_z = 20$. Further experiments revealed this to not necessarily be the case. As Figure 1 shows, increasing R_z from 20 to 50 results in worse performance based on the raw output scores but after normalisation is applied the situation has reversed, with $R_z = 50$ giving both superior minimum DCF and EER (Table 1). Both configurations show significant advantages over the reference system with as much as a 46% reduction of EER and a 42% reduction of minimum detection cost after normalisation is applied.

	Raw Scores		ZT-Norm	
System	DCF	EER	DCF	EER
Reference	.0389	10.6%	.0300	9.0%
$R_z = 20$	**.0358**	**8.7%**	.0211	5.4%
$R_z = 50$	.0391	9.4%	**.0174**	**4.8%**

Table 1. Minimum DCF and EER results for the 1-side condition when varying the number of session subspace dimensions, R_z.

The implications of this result are that increasing the power of the system's ability to model session variability can provide improved performance but score normalisation is required to realise these benefits. This leads to the conclusion that the session variability modelling method produces inherently less calibrated raw scores than standard GMM-UBM methods with ELLR scoring, particularly as R_z is increased.

It is also apparent that it is not possible to make accurate conclusions about the comparative performance of different configurations after normalised based on raw system scores. In this work this point has proven true numerous times.

3.2. Comparison of Training Methods

As stated in Section 2.2 there are several possibilities for the algorithm used to simultaneously optimise the set of variables $\{\boldsymbol{y}, \boldsymbol{z}_h; h = 1, \ldots, H\}$ during speaker enrolment. Results using several configurations are presented in Figure 2.

Simultaneously optimising using the Gauss-Seidel method (labelled "G-S 5 iters") works better than independently optimising (labelled "Indep 5 iters") with multiple iterations however this advantage is surprisingly small.

Interestingly dropping back to only 1 iteration of the EM procedure gives much better performance than using more iterations based on the raw output scores without normalisation; a 40% reduction in both minimum DCF and EER was observed in this case. While this advantage was reduced with score normalisation, the 1-iteration version was consistently ahead. The 1-iteration result may indicate that it is better to fully optimise the session variables independently of the speaker variable $\boldsymbol{y}$ and then determine the speaker parameters on what is effectively the *residual* variability after removing the channel effects and other forms of session variability.

3.3. Reduced Test Utterance Length

An important part of the session modelling method is estimating the session vector $\boldsymbol{z}_h$ for the test utterance. While this is a low-

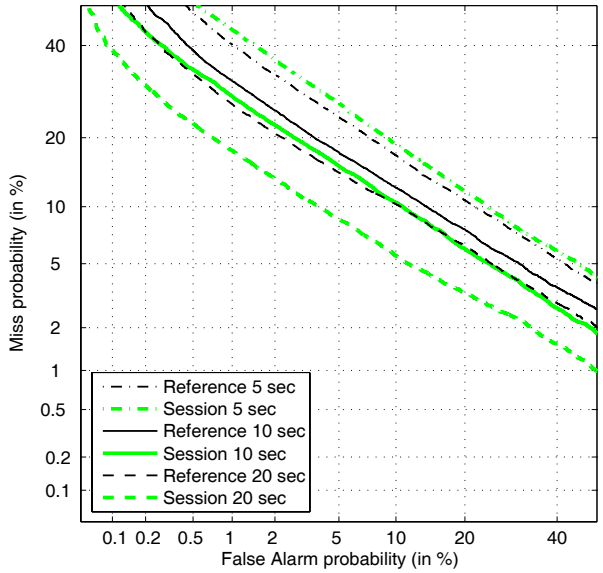

Fig. 3. DET plot for the 1-side training condition comparing baseline and session compensation results for short test utterance lengths.

dimensional variable estimating it accurately will require a sufficient quantity of speech. This experiment aims to determine the minimum requirements for extracting improved results from the session compensation method.

Figure 3 shows the impact of reducing the test utterance length for both the session variability modelling method and standard GMM-UBM modelling with test utterance lengths of 5, 10 and 20 seconds of active speech. These results indicate that approximately 10 seconds of speech is required to estimate the session factors sufficiently accurately to produce results that improve on standard modelling and scoring practice, while 20-second tests produce advances in performance starting to approach those experienced with full-length testing utterances, approximately 20% relative for both DCF and EER.

4. DISCUSSION

The results presented in this paper confirm the effectiveness of modelling session variability in a constrained subspace. The techniques presented have translated well from Switchboard conditions to the more challenging Mixer corpus which exhibits a greater variety of channel conditions and handset types.

One of the major advantages of the approach presented in this paper and related work is the more relaxed requirements for training corpus labelling. This technique removes the necessity of labelling databases for channel, handset type and other forms of session variability, which is often difficult, error prone and expensive if not impossible.

More sophisticated verification techniques are also possible. Future research will investigate the effectiveness of Bayes factor techniques in conjunction with modelling session variability in a similar approach to [13]. Under this approach the speaker model parameters are not assumed to be known at testing time, but rather to have posterior distributions refined by the training procedure. A similar formulation of the test utterance likelihood is also used as Kenny's verification criterion [6].

It may also be the case that the choice of scoring method also invalidates some of the training method conclusions drawn from the presented experiments; generally it is advantageous for the verification and enrolment criteria to match.

5. CONCLUSION

This paper confirmed the effectiveness of the session modelling technique for speaker verification on Mixer corpus of conversational telephony data by demonstrating a 46% reduction in EER and a 42% reduction in minimum detection cost.

Experimental results also indicate the effectiveness of the Gauss-Seidel training method and that approximately 10 seconds of speech is the minimum test utterance length required for sufficiently accurate estimates of the session variables to gain from the session modelling approach.

6. REFERENCES

[1] M. Mason, R. Vogt, B. Baker, and S. Sridharan, "The QUT NIST 2004 speaker verification system: A fused acoustic and high-level approach," in *Australian International Conference on Speech Science and Technology*, 2004, pp. 398–403.

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

[3] M. Mason, R. Vogt, B. Baker, and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification," in *Eurospeech*, 2005, pp. 3109–3112.

[4] D. Reynolds, "Channel robust speaker verification via feature mapping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2003, pp. 53–56.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[6] 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.

[7] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *Interspeech*, 2005, pp. 3117–3120.

[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. Reynolds, "Comparison of background normalizaftion methods for text-independent speaker verification," in *Eurospeech*, vol. 2, 1997, pp. 963–966.

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

[11] J. Pelecanos, R. Vogt, and S. Sridharan, "A study on standard and iterative map adaptation for speaker recognition," in *International Conference on Speech Science and Technology*, 2002, pp. 190–195.

[12] National Institute of Standards and Technology, "NIST speech group website," http://www.nist.gov/speech, 2004.

[13] R. Vogt and S. Sridharan, "Bayes factor scoring of GMMs for speaker verification," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 173–178.