Modelling Session Variability in Text-Independent Speaker Verification

Robbie Vogt, Brendan Baker, Sridha Sridharan

Speech Research Laboratory
Queensland University of Technology, Brisbane, Australia
{r.vogt,bj.baker,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 techniques such as feature mapping and H-Norm, while providing superior performance. Experiments on Switchboard-II conversational telephony data show improvements of as much as 48 % in detection cost with a single training utterance and 68 % with multiple training utterances over a baseline system.

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] to produce more robust features as well as score compensation techniques such as H- and T-Norm [4].

Feature mapping in particular addresses some of the known causes of session variability by attempting to map feature vectors extracted under identified conditions to a neutral feature space using a mapping based on adapted GMMs. It trains a discrete number of context models, such as for each handset transducer type, that are learnt from labelled training data. Due to this discrete identification of the context of a recording, feature mapping can only effectively be used for categorical data, such as transducer type, with ample labelled training data. Also, the number of contexts increases rapidly when combinations of factors are addressed.

The requirement for labelled training data can be overcome through blind clustering techniques as demonstrated for feature mapping in [5], however the discrete nature of the mapping is still problematic.

The techniques mentioned attempt to nullify session variability at the either the feature extraction or score normalisation stages but do not address the actual modelling of the speaker. Speaker model synthesis (SMS) [6] is a modelling based technique however it also suffers from issues with discrete decisions and data labelling requirements. It operates in a similar way to feature mapping but uses a transformation of GMM parameters to produce a model suited to the encountered context in which an utterance was recorded.

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. It is proposed to incorporate session differences into the way a speaker is modelled within a speaker verification system, in both the training and testing phases of the system. This work draws heavily on the results of Kenny *et al.* [7, 8] with some distinct differences.

In contrast to Kenny *et al.* [7] the presented approach does not perform speaker adaptation in a subspace adopting a more traditional GMM-UBM structure and obviating the need to train a speaker subspace transform and significantly reducing training complexity. The necessity to constrain session variability modelling to a low-dimensional space is also emphasised. Finally, a simplified verification score is used that is more in line with the GMM-UBM approach.

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. Training of the constrained session variability subspace is also addressed in 2.2. Experimental results are then presented on a modified NIST protocol using the Switchboard-II conversational telephony corpus in Section 3, and discussed in Section 4.

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 effect of session differences. 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 by,

$$\boldsymbol{m}_i(s) = \boldsymbol{m}(s) + \boldsymbol{U}\boldsymbol{x}_i(s). \tag{1}$$

Here, each speaker s has a supervector $\boldsymbol{m}(s)$ independent of session variations which is the concatenation of the GMM component mean vectors, $\boldsymbol{x}_i(s)$ is a low-dimensional representation of the variability in recording i and $\boldsymbol{U}$ is the low-rank transformation matrix from the constrained session variability subspace of dimension R_X to the GMM mean supervector space of dimension $R_S = MD$ (where the GMM is of order M and dimension D).

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

2.1. Speaker Model Training

Speaker models are trained through the simultaneous optimisation of the speaker model parameters themselves, which in this work corresponds to the supervector of GMM component means, and the session dependent subspace factors. In the case of the session variability vectors, there are as many of these as there are sessions available to train a particular speaker's model.

In this work the speaker mean supervector $\boldsymbol{m}(s)$ is optimised according to the *maximum a posteriori* (MAP) criterion often used in speaker verification systems [9]. The prior distribution in this case is derived from a universal background model (UBM).

The MAP criterion is also employed for optimising each of the session variablility vectors $\boldsymbol{x}_i(s)$. As described by Kenny *et al.* [7] the prior distribution in this case is 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 [7, 10].

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 model described in (1) as there is no sufficient statistics for mixtures of Gaussians due to the missing information of mixture component occupancy of each observation.

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 $(R_S+R_X)\times(R_S+R_X)$ matrix for each iteration. This matrix is required to capture the relationships between the variables being optimised. This translates to a $(12288+20)\times(12288+20)$ matrix decomposition for the size of speaker model and session variable subspace used in this work. Even with this matrix being positive-definite, this is impractical both in memory and processing requirements.

For this reason a procedure analogous to the Gauss-Seidel method for solving simultaneous equations is used:

1. Initialise all the session vectors and speaker model supervector estimates to $\mathbf{0}$.
2. Calculate the statistics and component occupancies of the observations in each training session based on the current variable estimates.
3. Re-estimate the session vector $\boldsymbol{x}_i(s)$ for each training session based on these statistics and the current estimate of the speaker supervector $\boldsymbol{m}(s)$.
4. Re-estimate the speaker model supervector $\boldsymbol{m}(s)$ based on these statistics and the new estimate of the session vectors $\boldsymbol{x}_i(s)$ obtained in step 3.
5. Repeat steps 2–4 until convergence.

While this method converges more slowly than direct simultaneous optimisation, each iteration only requires the decomposition of one $R_X \times R_X$ (20×20) matrix per training session and the trivial decomposition of an $R_S \times R_S$ diagonal matrix for the speaker supervector. In this work 5–10 iterations were found to provide sufficient convergence for the speaker verification task.

2.2. Training the Session Variability Subspace

For the session variation modelling described in the previous section to be effective, the constrained session variability subspace described by the transformation matrix $\boldsymbol{U}$ must represent the types of intra-speaker variations expected between sessions. To this end, the subspace is trained on a database containing a large number of speakers each with several independently recorded sessions.

Another EM algorithm is used to maximise the total (*a posteriori*) likelihood of all segments in the training database by training a speaker model for each speaker represented using the procedure in section 2.1. This procedure is described in detail in [8], with the caveat that a modified speaker model training procedure was used.

As stated in [7] this optimisation converges quite slowly and requires significant processing resources however our experiences with the process indicate that there is little improvement in verification performance to be gained with a fully converged algorithm; 10 iterations of the EM algorithm proved to be sufficient.

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 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 variation $\boldsymbol{x}_i(s)$ of the verification utterance for each speaker prior to performing standard Top-N ELLR scoring. This estimation is similar to that described in section 2.1 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 adaptation step 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

3.1. Baseline System and Evaluation Protocol

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 described in [2]. An adaptation relevance factor of $\tau = 8$ and 512-component models are used throughout and a session variation subspace of dimension $R_X = 20$ is used unless stated otherwise.

The proposed technique was evaluated using data from the NIST 2003 Speaker Recognition Evaluation Extended Data Task (EDT) [11]. The evaluation data is a subset of the Switchboard-II Phase 2 & 3 databases. To mirror the NIST 2004 evaluation conditions, the NIST EDT '03 evaluation procedure was restructured to include three training length conditions: one, three and eight conversation sides. The training and testing

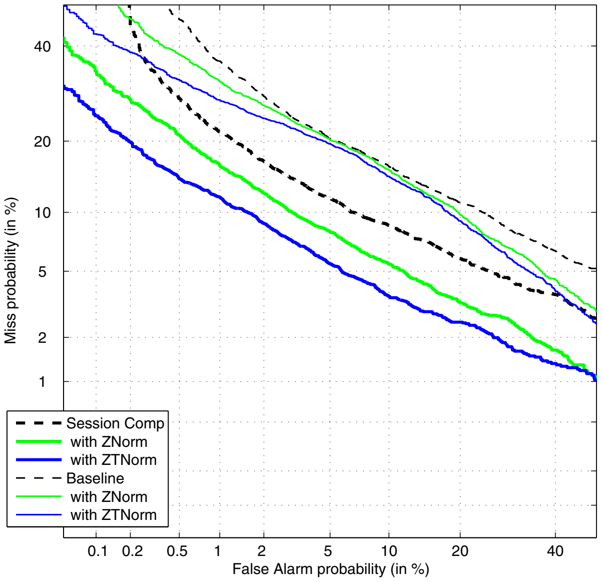

Figure 1: *DET plot for the 1-side training condition for the baseline system and one incorporating session variability compensation, with and without score normalisation.*

lists for the new 1- and 3-side conditions were derived from the existing four conversation side lists. More impostor trials were also added to the evaluation to better reflect the minimum DCF operating region. Additionally, the protocol has three independent splits to allow for the development of background models and fusion training sets in an unbiased fashion. This modified protocol is referred to as the QUT EDT '03[1]. With 420 models under each training condition, there is a total of 35,130 trials, evenly divided between both genders.

3.2. Results

Figures 1 and 2 show detection error trade-off (DET) plots comparing systems with and without session variability modelling for the 1- and 3-side training conditions respectively. Tables 1 and 2 present the minimum detection cost function (DCF) and equal error rate (EER) performance corresponding to these DET plots.

With no score normalisation applied, the session modelling technique provided a 32 % reduction in DCF for the 1-side condition and a 54 % reduction in the 3-side condition with similar trends in EER. While the improvement in the 3-side training condition is very substantial, the 1-side result is at least as interesting and in many ways more surprising. In the 1-side condition, there was not multiple sessions from which to gain a good estimate of the true speaker characteristics by factoring out the session variations, however the technique successfully factored out the variations between the training and testing sessions even with the simplified verification approach described in this paper.

Also presented are results with normalisation applied to all systems. The normalisations applied were Z-Norm to characterise the response of each speaker model to a variety of (impostor) test segments followed by T-Norm to compensate for the variations of the testing segments, such as duration and linguistic content. Again the proposed technique outperforms the baseline system, but also in fact gains more from this normalisation process than the baseline system with the improvements in DCF growing to 48 % and 68 % respectively for the 1- and 3-side conditions.

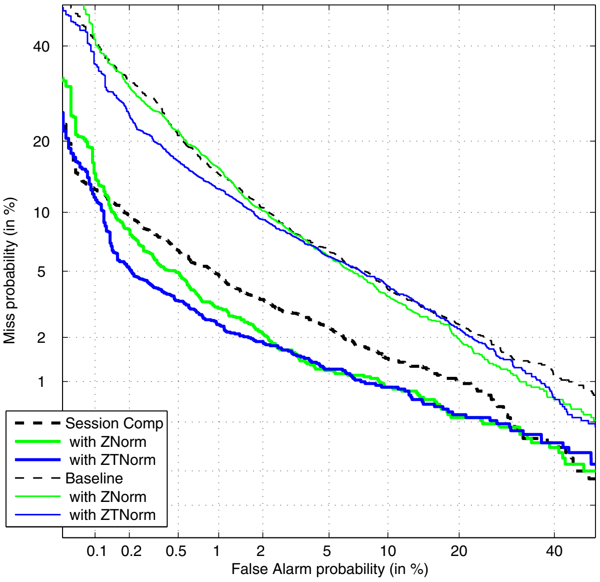

Figure 2: *DET plot for the 3-side training condition for the baseline system and one incorporating session variability compensation, with and without score normalisation.*

Table 1: *Minimum DCF and EER for the 1-side training condition for the baseline system and one incorporating session variability compensation, with and without score normalisation.*

	Baseline		Session Comp	
1-Side	DCF	EER	DCF	EER
No Norm	.0458	13.6	.0311	9.0
Z-Norm	.0415	13.0	.0251	6.8
ZT-Norm	.0367	12.7	.0191	5.3

Table 2: *Minimum DCF and EER for the 3-side training condition for the baseline system and one incorporating session variability compensation, with and without score normalisation.*

	Baseline		Session Comp	
3-Side	DCF	EER	DCF	EER
No Norm	.0243	5.9	.0110	2.8
Z-Norm	.0252	5.6	.0089	2.0
ZT-Norm	.0213	5.7	.0069	1.9

Presented in Table 3 are results obtained by varying the dimension of the session variability subspace for the female portion of the 1-side training condition. The necessity to constrain the amount of information in the utterance apportioned to session variation is highlighted by the degrading performance in the $R_X = 50$ case.

Figure 3 compares the performance of the presented technique to a feature mapping system trained with a data-driven clustering [5] on equivalent development data (similar results can be achieved with standard feature mapping as described in [3]). Again, it can be seen that the session variation mod-

[1] The QUT EDT '03 protocol is available on request from the authors.

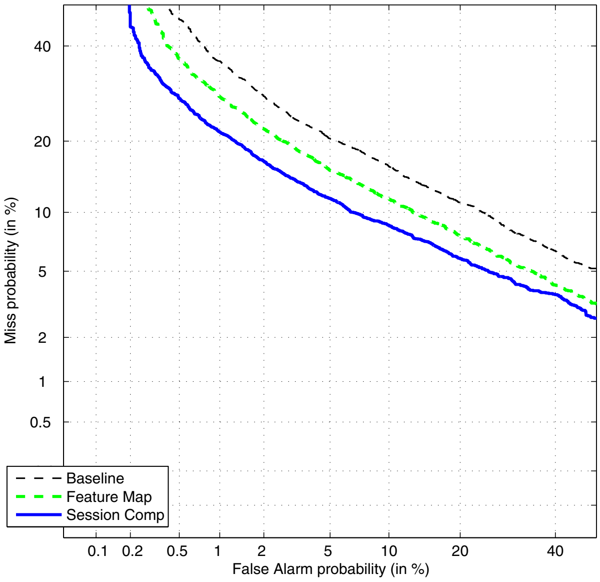

Figure 3: *Comparison of session variation modelling to blind feature mapping for the 1-side training condition.*

Table 3: *Minimum DCF and EER results for the female-only 1-side condition when varying the number of session subspace dimensions, R_X.*

1-Side	DCF	EER
Baseline	.0468	13.5
$R_X = 10$	.0335	8.9
$R_X = 20$	.0327	9.3
$R_X = 50$	.0353	10.7

elling technique has a clear advantage with a 19 % improvement at the minimum DCF operating point.

4. Discussion

One of the major advantages of the approach presented in this paper 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.

The benefits gained with score normalisation, particularly Z-Norm, seem to imply that a model produced with the proposed technique exhibits much more uniform response to a variety of test segments from different conditions. In contrast the baseline system improved little with Z-Norm while it is well known that H-Norm — utilising handset type labels — is more effective. This difference apparently indicates that the session modelling techniques are indeed successfully compensating for session differences such as handset type.

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 [12]. 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.

5. Conclusion

A technique was proposed to compensate for session variability in text-independent speaker verification by adding a session-dependent variable to the speaker modelling process that is constrained to lie in a session variation subspace. Experiments on conversational telephony data demonstrated the effectiveness of the technique for both single and multiple training session conditions with up to 68 % reduction in detection cost.

6. Acknowledgements

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

7. 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] D. Reynolds, “Channel robust speaker verification via feature mapping,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2003, pp. II–53–6.

[4] 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.

[5] M. Mason, R. Vogt, B. Baker, and S. Sridharan, “Data-driven clustering for blind feature mapping in speaker verification,” in *Eurospeech*, 2005, submitted.

[6] R. Teunen, B. Shahshahani, and L. Heck, “A model-based transformational approach to robust speaker recognition,” in *International Conference on Spoken Language Processing*, 2000, p. Paper 1642.

[7] 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.

[8] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. on Speech and Audio Processing*, in press.

[9] D. Reynolds, “Comparison of background normalization 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] National Institute of Standards and Technology, “NIST speech group website,” http://www.nist.gov/speech, 2004.

[12] 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.