Discriminant NAP for SVM Speaker Recognition

Robbie Vogt[1]*, Sachin Kajarekar*[2]*, Sridha Sridharan*[1]

[1]Speech Research Laboratory, Queensland University of Technology, Brisbane, Australia
[2]SRI International, Menlo Park, CA, USA

r.vogt@qut.edu.au, sachin@speech.sri.com, s.sridharan@qut.edu.au

Abstract

Nuisance Attribute Projection (NAP) provides an effective method of removing the unwanted session variability in a Support Vector Machine (SVM) based speaker recognition system by removing the principal components of this variability. There is no guarantee with the methods proposed, however, that desired speaker variability is retained.

This paper investigates the possibility of training NAP discriminatively to remove session variability while maintaining desirable speaker variability through an approach which is a variation on Scatter Difference Analysis (SDA). Experiments on NIST SRE tasks with a GMM mean supervector SVM system demonstrate a modest improvement by using SDA for NAP training by adding some speaker scatter.

Index Terms: NAP, session variability, SVM, Null-space linear discriminant analysis, scatter difference analysis

1. Introduction

The area of automatic speaker verification has seen a marked increase in research interest due to a number of factors both technological and from commercial and political perspectives, as evidenced by recent participation in NIST Speaker Recognition Evaluations (SRE). Noteworthy is the technological advances brought about by the widespread success and adoption of support vector machine (SVM) approaches based on a variety of features extracted from the audio signal including cepstral polynomials [1], MLLR transform coefficients [2], recognised phonetic sequences [3] and adapted GMM mean vectors [4].

The development of nuisance attribute projection (NAP) by Solomonoff, *et al.* [5, 6] has played an important role in the success of SVM approaches in the speaker verification domain particularly by introducing an effective method of reducing the performance degradation caused by the mismatch between the training and testing utterances of a speaker. NAP is a general approach that modifies the kernel function of an SVM classifier to remove the dimensions of the feature space that are dominated by nuisance variation through a reduced rank projection.

The dimensions to remove by NAP are generally determined through a data-driven approach over a large background population database. The most common form of NAP seeks to remove within-class variation which can be observed through the differences between examples of the same speaker in the background population. There is no mechanism in this training to prevent the desirable speaker information from also being removed along with the session variability.

In contrast, this work seeks to introduce a discriminative approach to training the NAP projection matrix that explicitly avoids incorporating speaker information in the discarded dimensions.

The following section describes the modified NAP kernel function proposed in [5] as well as the standard approach to training the projection, concluding with a brief comparison of the speaker and session variability actually captured by the projection on GMM mean supervector data. Section 3 proposes a discriminative approach to NAP projection training using a scatter difference criterion. Experimental results for a GMM mean supervector SVM system are presented and analysed for NIST SRE protocols in Section 4.

2. Nuisance Attribute Projection

NAP attempts to remove the unwanted within-class variation of the observed feature vectors [5, 6]. This is achieved by applying the transform

$$\boldsymbol{y}' = \boldsymbol{P}_n \boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{V}_n \boldsymbol{V}_n^T \right) \boldsymbol{y} \tag{1}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{V}_n$ is an $R_z \times R_y$ orthogonal projection matrix. $\boldsymbol{P}_n$ therefore introduces a null space of dimension R_z into the transformed features that corresponds to the range of $\boldsymbol{V}_n$.

As the purpose of NAP is to remove unwanted variability, $\boldsymbol{V}_n$ is trained to capture the principal directions of within-class variability of a training dataset, that is, it

This research was supported by SRI International as part of R. Vogt's 2006–2007 visit through a development contract with Sandia National Laboratories, and by the Australian Research Council Discovery Grant No DP0557387.

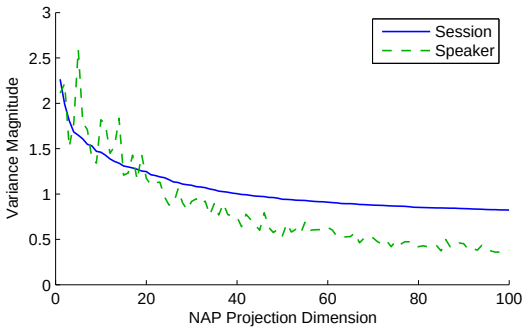

Figure 1: *Session and speaker variability magnitude of the SRE 2004 training data captured by the first 100 dimensions of the NAP projection.*

finds the vectors $\boldsymbol{v}$ that maximise the criterion

$$J(\boldsymbol{v}) = \boldsymbol{v}^T \boldsymbol{S}_w \boldsymbol{v} \tag{2}$$

where $\boldsymbol{S}_w$ is the within-class scatter of the training data. This is equivalent to finding the eigenvectors corresponding to the largest eigenvalues satisfying

$$\boldsymbol{S}_w \boldsymbol{v} = \lambda \boldsymbol{v}. \tag{3}$$

As the dimension of the input space is very large (the dimension of the GMM mean supervectors is approximately 100 000) and the number of background data samples is relatively small (approximately 2 800 utterances from 309 speakers extracted from 2004 NIST SRE data), the correlation matrix method [7] is used to determine the principal components. Determining the eigenvalues and eigenvectors of the $2\,800 \times 2\,800$ correlation matrix is evidently more practical and efficient than the direct eigen decomposition of the covariance matrix $\boldsymbol{S}_w$. This method is today more fashionably known as kernel PCA.

2.1. Speaker Information Removed with NAP

In the form proposed in [5], NAP does not explicitly avoid removing between-class variability, while it's assumed that it is this variability that is useful for discriminating between speakers. The amount of this variability captured in the NAP subspace was investigated for a GMM supervector SVM system. The variability captured in the leading NAP dimensions is plotted in Figure 1 by measuring the variance of the supervector observations projected onto these dimensions. This is the information *removed* by the NAP kernel.

From Figure 1, there is evidently a considerable amount of speaker variability removed along with the session variability using the NAP method and, in fact, for many of the first 20 dimensions the speaker variability is *greater* than the amount of session variability removed. This observation certainly motivates a NAP training algorithm that is more selective in the variability captured.

3. Discriminant NAP Training

The idea of the proposed discriminant NAP training approach is to minimise the speaker variability removed through the application of NAP by incorporating the between class scatter information in the projection matrix optmisation criterion.

The most obvious way to achieve this is by adopting the inverse of the traditional linear discriminant analysis (LDA) criterion, providing the greatest ratio of session to speaker variability in the training data. Thus we want the projection matrix $\boldsymbol{V}$ consisting of the vectors that maximise

$$J(\boldsymbol{v}) = \frac{\boldsymbol{v}^T \boldsymbol{S}_w \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{S}_b \boldsymbol{v}} \tag{4}$$

where $\boldsymbol{S}_w$ is the within-class scatter matrix and $\boldsymbol{S}_b$ is the between-class scatter matrix. This criterion is the inverse of the traditional LDA objective of maximising this ratio of between- to within-class variability. As with LDA, this criterion is optimised by solving the generalised eigenvalue problem

$$\boldsymbol{S}_b \boldsymbol{v} = \lambda \boldsymbol{S}_w \boldsymbol{v} \tag{5}$$

except the vectors with the smallest eigenvalues, as opposed to the largest, are retained. This approach does not produce a suitable NAP projection matrix, however, as it is not orthogonal.

It is possible to determine an orthonormal basis that spans the same space as the LDA projection. This meets the requirement of orthogonality for the projection matrix however initial experimentation encountered a number of further numerical issues with the LDA approach. Notably, due to the very large supervector space and limited observations, the ranges of $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are almost disjoint. This lead to extreme ranges of variability ratios (the eigenvalues) that tended to be dominated by accumulated rounding errors and offering very little useful information.

An alternative method of adding discrimination to NAP training is therefore considered in this work that is based on scatter difference analysis (SDA). This SDA approach swaps the variability ratio criterion of LDA for variability *differences* to avoid the issue of scaling in the whitening step of LDA. This is simply achieved by determining the principal directions of a weighted difference between the within and between scatter matrices.

3.1. Scatter Difference Analysis

Another approach is to avoid the ratio of the LDA criterion by instead using the difference between the scatter matrices such as in [8]. The criterion with this approach is

$$J(\boldsymbol{v}) = \boldsymbol{v}^T \left(\boldsymbol{S}_w - m\boldsymbol{S}_b \right) \boldsymbol{v} \tag{6}$$

where m controls the influence of the between-class scatter statistics. Using such a criterion avoids many of

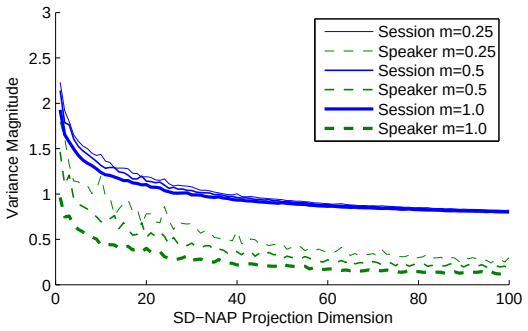

Figure 2: *Session and speaker variability magnitude of the SRE 2004 training data captured by the first 100 dimensions of the scatter difference NAP projection with different values of* m.

the issues associated with the LDA criterion such as dealing with singular scatter matrices, scaling by very small eigenvalues, and the resulting projection being non-orthogonal. This approach does, however, introduce a database-dependent tuning parameter to weight the relative importance of $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$.

The scatter difference criterion is optimised in the same manner as the standard NAP method, that is, by solving the eigenvalue problem. As with standard NAP, correlation matrices are used to avoid the issues caused by the very high dimensionality of the supervector features.

To suppress the speaker information in the resulting transform, m is assumed to be greater than 0. There are also some interesting special-case values of m with this approach. Specifically, in the case of $m = 0$, the criterion in (6) reverts to standard NAP training. Negative values of m will explicitly inculde more speaker variability and with $m = -1$ this approach is equivalent to PCA and finds the principal directions of variability in the total scatter matrix, $\boldsymbol{S}_t = \boldsymbol{S}_w + \boldsymbol{S}_b$.

Figure 2 shows the session and speaker variance in the leading dimensions of the projection trained with $m = 1$, that is weighting the within- and between-class scatter statistics evenly, as well as with $m = 0.5$ and $m = 0.25$, corresponding to a reduced influence of the between-class scatter statistic.

Comparing these results to Figure 1, it can be seen that the scatter difference criterion has significantly reduced the speaker variability captured by the NAP transform, as desired, with only a small drop in the session variance magnitude. Furthermore, as m increases the reduction in captured speaker variability becomes more pronounced, as expected.

NIST SRE 06	EER	Min. DCF
Baseline	5.07%	.0259
SD-NAP $m = 1.0$	4.53%	.0239
SD-NAP $m = 0.5$	4.48%	.0230
SD-NAP $m = 0.25$	4.48%	.0221
SD-NAP $m = 0.125$	4.42%	.0221
Standard NAP ($m = 0$)	4.37%	.0217
SD-NAP $m = -0.125$	4.26%	.0204
SD-NAP $m = -0.25$	**4.21%**	**.0198**
SD-NAP $m = -0.5$	4.26%	**.0198**
SD-NAP $m = -1.0$	4.32%	.0203

Table 1: *System performance on the common evaluation condition of the 2006 NIST SRE. For all NAP systems, 128 dimensions of session variability were removed.*

4. Results and Discussion

4.1. GMM Mean Supervector SVM System

The mean of a MAP adapted GMM [9] in the form of a supervector provides a suitable representation of an utterance for modelling with an SVM classifier [4]. A GMM mean supervector is formed by concatenated the component mean vectors of a MAP-adapted GMM that is $\boldsymbol{\mu}(s) = \left[\boldsymbol{\mu}_1(s)^T \dots \boldsymbol{\mu}_C(s)^T\right]^T$ where $\boldsymbol{\mu}_c(s)$ are the component means. The GMM-UBM system used in this work is described in [10] with the resulting supervectors have a dimension of $52 \times 2\,048 = 106\,496$. Similarly to Campbell, *et al.* [4], the supervectors are further normalised to be centred around the UBM mean and scaled by the UBM covariance such that the Euclidean norm of the resulting supervector is related to the Kullback-Leibler distance between the UBM and the adapted GMM. The resulting supervectors are modelled with a linear kernel SVM.

For these experiments, the background data (negative examples) for the SVM training consisted of an English-only subset of approximately 2 100 utterances drawn from a combination of the 2004 NIST SRE, Switchboard 2 and Fisher corpora. Another set of approximately 2 800 utterances from the 2004 SRE were used for training the NAP projection matrices, corresponding to 309 unique speakers.

4.2. Comparison of NAP and SD-NAP

Table 1 presents results of the proposed SD-NAP method in comparison to conventional NAP and a baseline system without compensation for session variability on the 2006 NIST SRE protocol. From these results it can be seen that NAP as proposed by Solomonoff, *et al.* imparts a significant performance improvement over the baseline system, as expected. Results for the scatter difference approach to NAP matrix training are also presented for a range of both positive and negative values of m.

NIST SRE '05	EER	Min. DCF
Baseline	5.66%	.0243
SD-NAP $m = 1.0$	5.21%	.0210
SD-NAP $m = 0.5$	5.13%	.0201
SD-NAP $m = 0.25$	4.92%	.0193
SD-NAP $m = 0.125$	4.92%	.0187
Standard NAP ($m = 0$)	**4.80%**	.0182
SD-NAP $m = -0.125$	4.92%	**.0179**
SD-NAP $m = -0.25$	5.05%	**.0179**
SD-NAP $m = -0.5$	5.29%	.0184
SD-NAP $m = -1.0$	5.49%	.0190

Table 2: *System performance on the common evaluation condition of the 2005 NIST SRE. For all NAP systems, 128 dimensions of session variability were removed.*

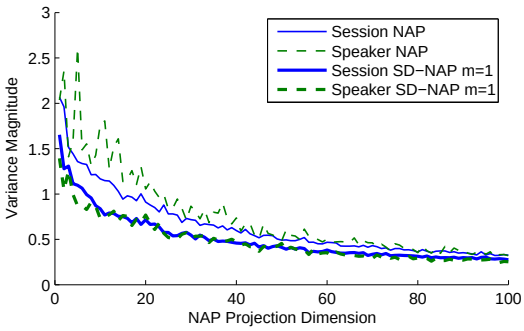

Figure 3: *Session and speaker variability magnitude of the SRE 2006 evaluation corpus captured by the first 100 dimensions of the proposed SD-NAP projection compared to standard NAP training.*

These results indicate that positive values of m, corresponding to reducing the the speaker information removed by NAP, produce results that are inferior to the standard NAP transform but ahead of the baseline system excluding NAP. Furthermore, an obvious trend of degrading performance can be observed as the value of m increases.

The results with $m < 0$ are more interesting: Contrary to expectations, all the negative values of m that were tested produced improved performance over standard NAP ($m = 0$) according to both the equal error rate and minimum detection cost criteria with the best results given by $m = -0.25$. These systems correspond to explicitly *adding* speaker variability to the NAP projection.

Results on the 2005 NIST SRE protocol are presented in Table 2. Results similar to the 2006 SRE are observed for the positive range of values for m, that is, no positive value of m provides results equivalent to the standard NAP approach but do improve substantially on the baseline system. The negative values m on the other hand are not consistent with the 2006 SRE results: While the DCF is marginally ahead for small negative values of m ($m = -0.125$ and $m = -0.25$), the rest of the results are inferior to standard NAP. This is particularly evident for the EER with $m = -1.0$, which falls close to the performance of the baseline system.

The improved performance with $m < 0$ for the 2006 SRE is an interesting outcome, however, the poor performance with positive values — the intended approach — warrants further analysis. Under the hypothesis that the degraded performance for these conditions was due to worse generalisation across databases of the SD-NAP approach compared to NAP, the amount of variability captured by the respective transform was measured for the 2006 SRE test data. A plot of between- and within-class variance captured by the leading dimensions of both the standard NAP projection and the proposed SD-NAP projection for this data is depicted in Figure 3.

It can be seen by comparing Figure 3 to Figure 2 that the speaker variability captured by the SD-NAP projection has been significantly reduced for the projection training data, but the same cannot be said for the 2006 SRE data.

Interestingly, comparing the variability captured with the standard NAP training in Figures 3 and 1, the speaker variability captured by the projection appears very similar but the session variability is reduced. In fact, for most of the dimensions the speaker variability is greater than the session variability.

Furthermore, the introduction of the discriminative criterion for SD-NAP has not produced the same results as for the 2004 SRE data used for training the projections. There is a reduction in the speaker information captured by the SD-NAP transform as desired, however, this reduction is not as pronounced as for the training data. Unlike the results for 2004 SRE data, the corresponding reduction in captured session variability is also significant. These results indicate that the SD-NAP projection has not generalised particularly well to the 2006 SRE data, but, does in fact appear to improve the ratio of session to speaker variability captured by the transform.

Finally, an analysis of the variability captured by the SD-NAP transform with $m = -0.25$, which is the best configuration for the 2006 SRE, reveals almost identical results as for $m = 1$.

The results of this analysis are inconclusive: While the SD-NAP approach appears to have achieved the objective of reducing the speaker variability relative to the session variability for the test data, it did not produce improved performance. Furthermore, explicitly adding speaker variability with $m < 0$ did produce improved performance but only for the 2006 SRE. It is possible that there is some characteristic of the 2006 SRE data collection that is causing these results; SVM systems using other features such as MLLR coefficients will be used to explore this phenomenon further.

System	EER	Min. DCF
Speaker LDA	5.82%	.0307
Standard NAP	4.37%	.0217
Fused	**3.88%**	**.0204**

Table 3: *System performance on the common evaluation condition of the 2006 NIST SRE utilising the speaker-dominant LDA projection.*

4.3. Using Speaker-Discriminating Features Directly

An alternative approach was investigated for exploiting a discriminative approach to preparing features for SVM modelling in a similar vein to previous work in [11]. The main idea of this approach was to combine a system using standard LDA with low dimension with a system capable of representing the entire supervector space.

Firstly, a traditional LDA transform was trained on GMM supervector features to produce a low-dimensional space maximising the ratio of speaker to session variability. The LDA analysis was achieved through a simultaneous diagonalisation approach [7] where $\boldsymbol{S}_t$ — the total scatter matrix — is first whitened and the transformed version of $\boldsymbol{S}_b$ is then diagonalised. As with the previous techniques, the size of the supervector space requires this analysis to be performed on correlation matrices due to practical computational limits. As the purpose of this transform was to performing modelling and recognition in this space, rather than remove it, the orthogonality constraint of NAP was not a concern.

The 2006 SRE results for a system using 300 most discriminative LDA dimensions are presented in the first row of Table 3. Comparing to Table 1, this system using only 300-dimensional features shows a relatively small drop in performance compared to the Baseline system that uses a feature space of over 100 000 dimensions.

A simple linear combination of this system with the standard NAP system (repeated from the fifth row of Table 1) provided both a slight improvement in DCF and a substantial improvement in the EER operating point (as indicated in the final row of Table 3). This result seems to indicate that there is some value to investigating more discriminatory methods of preparing features for modelling with an SVM approach.

4.4. Discussion

Although it appears that there is merit in preparing more discriminatory features for SVM modelling as evidenced by the limited performance gains demonstrated above, further work is required to realise these potential gains. The reliance of the proposed techniques on traditional variance analysis and F-ratio-maximising approaches, such as LDA, may not be appropriate for optimising SVM classification; unlike traditional generative approaches that model the full distribution, SVM's are only interested in training observations on or within the separating margin. It is hypothesised that optimising the feature space to maximise the resulting margin will provide more reliable gains in performance.

5. Conclusions

This paper investigates incorporating a discriminative training approach to NAP transform training to avoid removing important speaker information while suppressing the session information. A criterion based on scatter difference analysis was proposed for this purpose where the between-class scatter, as well as the within-class scatter, statistic is exploited.

Results for a GMM mean supervector SVM system demonstrated generally inferior performance for both the 2005 and 2006 SRE's, however, adding speaker variation showed modest gains for the 2006 SRE. Further experiments with a more traditional LDA approach also demonstrated an improved EER when fused with a standard NAP system. These results lend some support to investigating more discriminatory approaches to SVM feature preparation.

6. References

[1] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. 161–164.

[2] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in mllr-transform-based speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2006.

[3] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, "Phonetic speaker recognition with support vector machines," *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[4] 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.

[5] A. Solomonoff, C. Quillen, and W. Campbell, "Channel compensation for SVM speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 57–62.

[6] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *ICASSP*, vol. I, 2005, pp. 629–632.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, California, USA: Academic Press, 1990.

[8] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face recognition using kernel scatter-difference-based discriminant analysis," *Neural Networks, IEEE Transactions on*, vol. 17, no. 4, pp. 1081–1085, 2006.

[9] 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.

[10] S. Kajarekar, L. Ferrer, M. Graciarena, E. Shriberg, K. Sönmez, A. Stolcke, G. Tur, and Y. Solewicz, "2006 NIST speaker recognition evaluation: SRI system description," in *NIST Speaker Recognition Workshop Booklet*, 2006.

[11] S. Kajarekar, "Four weightings and a fusion: a cepstral-SVM system for speaker recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 17–22.