Scatter Difference NAP for SVM Speaker Recognition

Brendan Baker, Robbie Vogt, Mitchell McLaren, and Sridha Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane, AUSTRALIA, 4001
{r.vogt,bj.baker,m.mclaren,s.sridharan}@qut.edu.au

Abstract. This paper presents Scatter Difference Nuisance Attribute Projection (SD-NAP) as an enhancement to NAP for SVM-based speaker verification. While standard NAP may inadvertently remove desirable speaker variability, SD-NAP explicitly de-emphasises this variability by incorporating a weighted version of the between-class scatter into the NAP optimisation criterion. Experimental evaluation of SD-NAP with a variety of SVM systems on the 2006 and 2008 NIST SRE corpora demonstrate that SD-NAP provides improved verification performance over standard NAP in most cases, particularly at the EER operating point.

1 Introduction

Automatic speaker verification technology has advanced considerably in recent years, and remains to be a highly active research area. This interest is evidenced by strong and continuing international participation by leading research groups in recent NIST Speaker Recognition Evaluations (SRE). These evaluations have fostered important developments in the technology, and have provided a public forum for participating institutions to publicise these advancements.

In recent NIST SRE's, there has been an increase in the popularity of support vector machine (SVM) approaches, with techniques based around SVM's achieving widespread success. A large variety of SVM speaker verification implementations have been proposed using differing feature representations. including cepstral polynomials [1], MLLR transform coefficients [2], recognised phonetic sequences [3] and adapted GMM mean supervectors [4].

An instrumental development that has attributed to the success of SVM techniques is that of nuisance attribute projection (NAP). Originally proposed by Solomonoff, *et al.* [5,6], NAP has been shown to be an effective method of reducing the performance degradation introduced by mismatch between the training and testing utterances of a speaker. Through a modification of the kernel function, NAP allows for the removal of dimensions of the feature space dominated by "nuisance variation." These removed dimensions 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

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 464–473, 2009.

can be observed through the differences between examples of the same speaker in the background population. A possible sub-optimality of this approach is that there is no mechanism to prevent desirable speaker information from also being removed along with the session variability. That is to say, 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.

Building upon previous work in this area [7], this paper addresses this potential sub-optimality in the NAP training by adopting an approach that explicitly avoids the incorporation of speaker information in the discarded dimensions. The alternate training method, termed *scatter difference nuisance attribute projection* (SD-NAP), is examined in detail. As described in [7], SD-NAP seeks to de-emphasise the speaker information in the NAP projection by incorporating between-class scatter information in the projection training. This paper provides further exploration of this alternate formulation, assessing its performance across different corpora and using varied feature sets and implementations.

The following sections describe both the standard NAP kernel function, and scatter difference analysis that underpins the modified training method. Following this, a series of experiments are performed to compare both the modified and standard NAP approaches. Initial experimentation is performed using a GMM mean supervector system on both 2006 and 2008 NIST SRE corpora. Following this, further results are presented on alternate SVM-based speaker verification systems, utilising different feature sets and kernel functions.

2 Nuisance Attribute Projection

NAP is used to combat errors introduced as a result of inter-session variation, or more simply, mismatch between training and testing utterances of a speaker. Through a modification of the kernel function, NAP allows for the removal of dimensions of the feature space dominated by this inter-session variation. These dimensions are generally determined through a data-driven approach using a large background database consisting of typically 100's or 1000's of speakers, each with numerous recordings or sessions. By examining the differences between examples of the same speaker in the background population, the within-class variation can be used as a model of the nuisance or inter-session variation.

More specifically, 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 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 dimensions of the GMM mean supervectors is $12,288 \times 1$) and the number of background data samples is relatively small (approximately 3,400 utterances from 430 speakers extracted from 2004 and 2005 NIST SRE data), the correlation matrix method [8] is used to determine the principal components. Determining the eigenvalues and eigenvectors of the $3,400 \times 3,400$ correlation matrix is evidently more practical and efficient than the direct eigen decomposition of the covariance matrix $\boldsymbol{S}_w$. This is essentially equivalent to kernel PCA.

2.1 Speaker Information Removed with NAP

As highlighted in [7], the original NAP formulation proposed by Solomonoff, *et al.* [5] does not explicitly avoid removing between-class variability, while it's assumed that it is this variability that is useful for discriminating between speakers. In Figure 1 a plot of the variability captured in the leading NAP dimensions is provided, calculated by measuring the variance of the supervector observations projected onto these dimensions for the 2,800 observations in the training database. It is this information that is discarded by the NAP kernel.

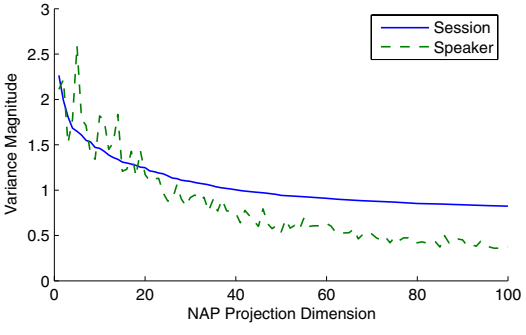

Fig. 1. Session and speaker variability magnitude of the SRE 2004 training data captured by the first 100 dimensions of the NAP projection

It can be seen from Figure 1 that there is 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 information it removes.

2.2 Scatter Difference Analysis

In [7], an alternate NAP formulation was proposed that seeked to reduce the speaker variability removed. This was achieved by incorporating the between class scatter information in the projection matrix optimisation criterion, and using the difference between both within- and between-class matrices such as in [9]. The criterion with this approach can be expressed as

$$J(\boldsymbol{v}) = \boldsymbol{v}^T \left(\boldsymbol{S}_w - m\boldsymbol{S}_b\right) \boldsymbol{v} \tag{4}$$

where $\boldsymbol{S}_b$ is the between-class scatter of the training data, and m controls the influence of the between-class scatter statistics. It should be noted that this approach introduces 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.[1]

To suppress the speaker information in the resulting transform, m is typically set to be in the range 0 to 1. The bounds of this range correspond to the special cases of standard NAP ($m = 0$), and equal weighting between session and speaker information ($m = 1$).

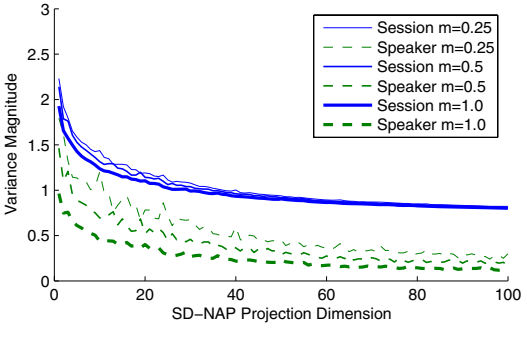

Fig. 2. Session and speaker variability magnitude of the SRE 2004 training data captured by the first 100 dimensions of the scatter difference NAP projection with different values of m

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 equally, as well as with $m = 0.5$ and $m = 0.25$, corresponding to a reduced influence of the between-class scatter statistic.

[1] The scatter difference: $\boldsymbol{S}_w - m\boldsymbol{S}_b$, is not necessarily positive definite, so generalised eigenvalue decomposition must be used rather than singular value decomposition.

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.

3 Database and Evaluation

The SD-NAP technique was evaluated using a number of different SVM speaker verification implementations, with varied feature sets, and across two separate evaluation corpora. A description of each of the system configuration and evaluation corpora follows.

3.1 Evaluation Corpora

The NIST 2006 and 2008 SRE corpora were used to evaluate and compare the performance of the proposed method. Results were derived from the "all trials" condition of the official evaluation protocol which includes trials spoken in all-languages.

3.2 GMM Mean Supervector SVM System

The mean of a MAP adapted GMM [10] in the form of a supervector provides a suitable representation of an utterance for modelling with an SVM classifier [4]. For the system used in this study, a GMM mean supervector is formed by concatenating 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 based around the system described in [11] with the resulting supervectors having dimension of $24 \times 512 = 12,288$. Each dimension of the SVM feature space is normalised by the mean and standard deviation of the corresponding observations in the background dataset. T-Norm score normalisation [12] was also used on the resulting verification scores using a T-Norm dataset constructed from the NIST 2004 SRE data.

3.3 English Phonetic Lattice N-Gram SVM System

This system used phonetic transcripts produced by an English phone recogniser. The English phone recogniser was trained using data from the Fisher corpus [13]. Gender-dependent phone recognisers were used. 100 hours of speech per gender were extracted from the Fisher database to train three state, 16 mixture component, and gender-dependent HMMs using HTK. The system is capable of recognising a total of 43 phonetic labels. In a similar manner to the approach described in [14], rather than using 1-best transcriptions phone lattices were utilised. The expected frequencies of unigrams, bigrams and trigrams in each utterance were concatenated and used to form a feature vector. Only the 10,000

most frequently occurring trigrams (determined on the background set) were included. The n-gram frequencies were weighted according to their posterior probability of occurrence in the recognition lattice. T-Norm score normalisation was also applied to this system.

3.4 PPRLM Lattice N-Gram SVM System

The PPRLM (parallel phone recogniser with language modelling) system uses phone transcriptions obtained from multiple open-loop phone recognisers (OLPR) each trained on one of 6 languages; English, German, Hindi, Japanese, Mandarin and Spanish. This parallel stream architecture was first described for the speaker verification task in [15]. The multi-stream decoding was performed using QUTs HMM based OLPR - trained on the OGI multi-lingual database. The same support vector feature extraction process was used for all six language streams. The expected frequencies of unigrams, bigrams and trigrams in each utterance were concatenated and used to form a feature vector. Only the 10,000 most frequently occurring trigrams (determined on the background set) were included. Once again, the n-gram frequencies were weighted according to their posterior probability of occurrence in the recognition lattice. Scores from each stream were fused through a linear combination, with weightings calculated via logistic regression. T-Norm score normalisation was applied to each stream prior to combination.

3.5 MLLR SVM System

The MLLR system makes use of the same English phone recogniser as described in Section 3.3. The male and female English phone recogniser HMM models served as reference models for computing speaker-dependent MLLR transforms. Using the alignments produced by the phonetic decoder, a five class regression tree was used (4 data driven classes + silence) to obtain a set of MLLR transforms for each training segment. The acoustic features contain 39 components, resulting in a transform vector for each class of dimension 39×40 (39×39 transform matrix + 39×1 vector bias). The transform components for each class (excluding the silence class) were concatenated to form a single feature vector for each conversation side resulting in a total feature vector length of 6,240. Based on the suggestion in [2], transforms from both male and female models were generated for each conversation side and concatenated to form a final 12,480 length supervector. Once again, the SVM space was constructed with a linear kernel and was rank normalised. During development, T-score normalisation was found to provide little benefit for this system, and as such this process was excluded. This conclusion was also drawn in [2].

4 Results and Discussion

Initial experimentation was performed using the GMM mean supervector system (see Section 3.2). Table 1 presents results of the SD-NAP method in comparison

to conventional NAP and a baseline system without session variability compensation on both the 2006 and 2008 NIST SRE protocol. For the scatter difference approach to NAP matrix training, results are presented for a range of values for the matrix weighting term m.

It is clear from the results in Table 1 that significant performance improvements over the baseline can be achieved by adopting a form of nuisance attribute projection. As expected, the original NAP formulation, proposed by Solomonoff, *et al.* achieves a significant improvement over the baseline system.

The SD-NAP results indicate that further improvements can be achieved by minimising the amount of speaker variability removed. Almost all SD-NAP configurations trialled showed improvement over the standard NAP approach. The best results across 2006 and 2008 evaluations for both equal error rate and minimum detection cost criteria resulted when the SD-NAP technique was employed with $m = 0.2$.[2]

Table 1. Performance of the GMM Mean Supervector System evaluated on all trials (combined male and female) of the 2006 and 2008 NIST SRE's. 50 dimensions of session variability were removed using NAP/SD-NAP.

	NIST SRE 2006		**NIST SRE 2008**	
System	**EER**	**Min. DCF**	**EER**	**Min. DCF**
Baseline	7.30%	.0363	10.26%	.0542
Standard NAP ($m = 0$)	4.82%	.0232	7.14%	.0375
SD-NAP $m = 0.05$	4.73%	.0230	7.10%	.0373
SD-NAP $m = 0.1$	4.67%	.0225	6.96%	.0372
SD-NAP $m = 0.2$	**4.54**%	**.0221**	**6.80**%	**.0366**
SD-NAP $m = 0.5$	4.73%	.0235	7.05%	.0372
SD-NAP $m = 1$	5.34%	.0261	7.46%	.0388

The SD-NAP technique was evaluated for the remaining verification systems, with comparisons made again to both standard NAP and baseline systems. Table 2 provides a performance summary of the trialled configurations. For the SD-NAP configurations, the weighting factor m was optimised on the 2006 data. Once again, it is immediately clear from these results that incorporating nuisance attribute projection of some form allows for significant improvements in performance to be achieved. Both NAP and SD-NAP configurations provide marked improvements over the baseline systems for all systems and evaluations.

Comparing the NAP and SD-NAP configurations, for most cases, the SD-NAP approach achieves superior error rates to that of standard NAP. This is with the exception of the multi-stream PPRLM n-gram system. In this case, adopting the SD-NAP approach in substitution of standard NAP resulted in a slight degradation across both evaluations.

[2] While negative values for the weighting m were investigated in [7] and gave improved performance, these same negative values gave only degraded performance in this study. It is hypothesized that the results in the previous study were anomalis and possibly related to differences or details in the system implementatations.

Table 2. Performance of the SVM speaker verification systems evaluated on all trials of the 2006 and 2008 NIST SRE's. 50 dimensions of session variability were removed using NAP/SD-NAP.

	NIST SRE 2006		NIST SRE 2008	
System	**EER**	**Min. DCF**	**EER**	**Min. DCF**
GMM SVM				
Baseline	7.30%	.0363	10.26%	.0542
Standard NAP	4.82%	.0232	7.14%	.0375
SD-NAP $m = 0.2$	**4.54%**	**.0221**	**6.80%**	**.0366**
English N-gram				
Baseline	13.82%	.0630	18.90%	.0910
Standard NAP	11.42%	.0496	14.49%	**.0794**
SD-NAP $m = 0.2$	**10.04%**	**.0480**	**13.42%**	**.0794**
PPRLM N-gram				
Baseline	15.38%	.0631	17.24%	.0764
Standard NAP	**8.24%**	**.0426**	**11.85%**	**.0719**
SD-NAP $m = 0.2$	8.30%	.0442	11.96%	**.0719**
MLLR				
Baseline	10.34%	.0442	14.34%	.0626
Standard NAP	9.43%	.0418	12.30%	.0542
SD-NAP $m = 0.2$	**9.26%**	**.0409**	**11.85%**	**.0535**

Table 3. Comparison of NAP and SD-NAP for individual streams of PPRLM system on all trials (combined male and female) of the 2006 NIST SRE

	Standard NAP		SD-NAP	
System	**EER**	**Min. DCF**	**EER**	**Min. DCF**
English	11.75%	.0521	11.01%	.0522
German	12.36%	.0549	11.48%	.0542
Hindi	11.64%	.0521	10.87%	.0522
Japanese	12.64%	.0547	12.00%	.0543
Mandarin	11.89%	.0521	11.23%	.0520
Spanish	12.61%	.0539	11.86%	.0536
Combined	8.24%	.0426	8.30%	.0442

It is also interesting to note that the optimal value for the SD-NAP weighting m remained fairly stable across the different systems and evaluations. Using a constant value $m = 0.2$ for all systems gave the best performance.

4.1 Analysis of Individual PPRLM Streams

Further analysis was performed to gain a better indication as to why the SD-NAP approach did not outperform the standard NAP method when used for the PPRLM system. Table 3 provides a performance summary breakdown for the six individual language streams when tested on the 2006 SRE, along with the combined result.

Interestingly, a number of the individual language streams display equal or better performance when the SD-NAP approach is used rather than the standard NAP formulation. Particularly, improvement is apparent for the EER performance statistic, where using SD-NAP lowers the error rate across all languages. Unfortunately, as already presented, once the streams are combined using logistic regression, this improvement is negated, with the standard NAP approach surpassing SD-NAP. These results warrant further investigation in future studies. A number of factors, including the order in which T-Norm score normalisation and fusion is performed, and the callibration of input scores prior to logistic regression combination may need to be re-considered when using SD-NAP.

5 Conclusions and Future Work

This paper examined an alternate NAP training approach that explicitly avoids the incorporation of speaker information in the discarded dimensions, The method proposed uses an alternate training criterion based on scatter difference analysis. The SD-NAP method seeks to de-emphasise important speaker information in the NAP projection by incorporating a weighted version of the between-class scatter.

The SD-NAP technique was evaluated and compared to both baseline and standard NAP approaches using a number of different SVM speaker verification implementations and evaluation corpora. Results demonstrated that consistent improvements over the standard NAP approach could be achieved. Although the SD-NAP method introduces an additional database-dependent tuning parameter m, experiments revealed that the optimal value for this weighting term remained fairly stable across the different systems and evaluations.

The only exception to the observed improvement achieved through use of SD-NAP was for the multi-stream PPRLM-based speaker verification system. Further analysis of this system revealed that for the individual component streams, equal or better performance was generally achieved over NAP. This trend, however, was reversed after the application of T-Norm and stream combination. The effect that SD-NAP has on output score distributions along with its interaction with score normalisation schemes such as T-Norm warrants further investigation.

Acknowledgements. This research was supported by the Australian Research Council (ARC) Discovery Grant Project ID: DP0877835.

References

1. Campbell, W.: Generalized linear discriminant sequence kernels for speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 161–164 (2002)
2. Stolcke, A., Ferrer, L., Kajarekar, S.: Improvements in MLLR-transform-based speaker recognition. In: Odyssey: The Speaker and Language Recognition Workshop (2006)

3. Campbell, W., Campbell, J., Reynolds, D., Jones, D., Leek, T.: Phonetic speaker recognition with support vector machines. In: Advances in Neural Information Processing Systems, vol. 16 (2004)
4. Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A.: 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, pp. I–97–I–100 (2006)
5. Solomonoff, A., Quillen, C., Campbell, W.: Channel compensation for SVM speaker recognition. In: Odyssey: The Speaker and Language Recognition Workshop, pp. 57–62 (2004)
6. Solomonoff, A., Campbell, W., Boardman, I.: Advances in channel compensation for SVM speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. I, pp. 629–632 (2005)
7. Vogt, R., Kajarekar, S., Sridharan, S.: Discriminant NAP for SVM speaker recognition. In: Odyssey: The Speaker and Language Recognition Workshop (2008)
8. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, San Diego (1990)
9. Liu, Q., Tang, X., Lu, H., Ma, S.: Face recognition using kernel scatter-difference-based discriminant analysis. IEEE Transactions on Neural Networks 17(4), 1081–1085 (2006)
10. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10(1/2/3), 19–41 (2000)
11. McLaren, M., Vogt, R., Baker, B., Sridharan, S.: A comparison of session variability compensation techniques for SVM-based speaker recognition. In: Interspeech 2007, pp. 790–793 (2007)
12. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing 10(1/2/3), 42–54 (2000)
13. Cieri, C., Miller, D., Walker, K.: The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In: International Conference on Language Resources and Evaluation, pp. 69–71 (2004)
14. Hatch, A., Peskin, B., Stolcke, A.: Improved phonetic speaker recognition using lattice decoding. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2005)
15. Andrews, W., Kohler, M., Campbell, J., Godfrey, J., Hernández-Cordero, J.: Gender-dependent phonetic refraction for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 149–152 (2002)