NAP AND WCCN: COMPARISON OF APPROACHES USING MLLR-SVM SPEAKER VERIFICATION SYSTEM

Sachin S. Kajarekar[1] and Andreas Stolcke[1,2]

[1]SRI International, Menlo Park, CA, USA
[2]International Computer Science Institute, Berkeley, CA, USA

ABSTRACT

We compare two recently proposed techniques, within class covariance normalization (WCCN) [1] and nuisance attribute projection (NAP) [2], for intersession variability compensation in speaker verification. The comparison is performed using an MLLR-SVM speaker verification system. Both techniques model intersession variability using a within-speaker covariance matrix (WSCM). However, they manipulate eigenvectors of this matrix differently. We compare them on the 2005 and 2006 NIST speaker recognition evaluation (SRE) task. Results show that WCCN is more sensitive to the choice of background speakers and NAP is more sensitive to the choice of data for WSCM estimation. WCCN gives the best performance on 2005 SRE. On 2006 SRE, both techniques give similar performance under matched conditions. Further experiments with a simple combination of these techniques show slight improvements in the best performance of either technique. Overall results show that an MLLR-SVM system with either NAP or WCCN performs comparably to the best single systems in the 2006 NIST SRE.

Index Terms— Speaker recognition, Intersession variability, MLLR transforms, SVM

1. INTRODUCTION

We refer to speaker recognition as the task of recognizing speakers from their voices. We specifically look at the open-set speaker verification problem as formulated in NIST's annual speaker recognition evaluations (SREs). Commonly used modeling techniques in these evaluations include Gaussian mixture modeling (GMM) [3] and support vector machines (SVM) [4, 5]. Commonly used features include Mel frequency cepstral coefficients (MFCCs) for modeling spectral aspects and recently introduced nonuniform extraction region features (NERFs) [6] for modeling stylistic aspects of speech.

Channel variability is considered to be one major source of mismatch between training and testing in NIST SREs, and a variety of normalization techniques have been proposed [7, 8]. Recently, Kenny [9, 10] showed that another important source of mismatch is the intersession variability (ISV). Simply stated this is the (averaged) variation between different conversations by the same speaker. Obviously, ISV can be caused by channel variation; other causes may include variation in the phonetic content, emotional state, and so on. Kenny proposed a factor analysis model, shown to be very effective for GMM-based systems [11].

In parallel, researchers have been working on the issue of ISV in the SVM framework. One difference between GMM and SVM frameworks is in how features are constructed. A GMM uses a sequence of feature vectors sampled at a particular rate. An SVM uses some form of average statistics estimated from these features [4]. Two techniques have been proposed to address the effects of ISV in an SVM framework – nuisance attribute projection (NAP) [2, 12] and within-class covariance normalization (WCCN)[1].

We describe the experimental setup in Section 2. WCCN and NAP are described in Section 3. Section 4 compares results obtained with the two methods. In Section 5, we discuss combinations of the two techniques. We conclude with a summary in Section 6.

2. EXPERIMENTAL SETUP

For this study, we use maximum likelihood linear regression (MLLR) transform coefficients as features, which are obtained as a by-product of the ASR system used to transcribe the data. MLLR transforms are obtained for eight broad phonetic classes and for two genders. The features are rank-normalized to equate their dynamic ranges [13]. The normalized features are modeled using an SVM classifier with a linear inner-product kernel. This system is described in detail in [13] and has been the single best system in our NIST evaluation submission.

We use a variety of datasets in this paper because WCCN and NAP were proposed with different background sets and different datasets used to obtain the ISV statistics. We use the 2003-2006 SRE datasets and a set obtained from the Fisher database. From the 2003 SRE (SRE03) and 2004 SRE (SRE04) datasets, we use speakers with more than eight conversations to estimate the intersession variability (ISV dataset). The SRE03 data has about 625 unique speakers and the SRE04 data has about 310 unique speakers. The Fisher and SRE04 sets are used as negative training samples in SVM training (background dataset). We have used data from the 2005 SRE (SRE05) and 2006 SREs (SRE06) for evaluating the techniques. The former is used to tune parameters for each technique and the latter is used to test generalization.

For all the different SRE sets, the evaluation is performed with 2.5 minutes of training data and another 2.5 minutes of testing data. Each data point is obtained from one of the sides of a 5-minute-long of conversation. We present results in terms of both equal error rate (%EER) and decision cost function (DCF) with the cost specified by NIST [14]. Note that all the results presented in this paper are without any score normalization such as TNORM.

3. INTERSESSION VARIABILITY COMPENSATION

There are two issues in dealing with intersession variability: 1) how to compute variability and 2) how to compensate for it. Here

we compare the two techniques in terms of these issues. We describe WCCN first and interpret NAP as a simplified WCCN.

It is important to note some key properties of the eigenanalysis performed for WCCN and NAP. The dimensionality of MLLR features (T) is around 20k, and the number of conversations (M) available for computing the within-speaker covariance matrix is around 3-6k. It is computationally impossible to perform an eigenanalysis of a TxT covariance matrix, which itself is an ill-conditioned matrix and has only M-1 non-zero eigenvalues. Therefore, a kernel trick is used [15]. The eigenanalysis is performed with an MxM covariance matrix in the conversation space, and the eigenvectors are transformed back to the original feature space.

3.1. Within-class covariance normalization

Hatch et al. proposed the WCCN approach [1] and showed significant improvements on SRE04 and SRE05 data. Figure 1 shows how the ISV is computed ("analysis") and how the result is used ("application"). Note that this is only a brief overview of WCCN; for more a detailed explanation refer to [1].

In the analysis part, the MLLR features from the ISV dataset are normalized by within-speaker variance (WSV). This is done to ensure the proper conditioning of the within-speaker covariance matrix (WSCM) estimated in the next step. Eigenanalysis is performed on this covariance matrix and a set of eigenvectors (E_M) is computed using the kernel trick. WSV-normalized features are projected onto these eigenvectors and the within-speaker variance is again computed in the transformed space (WSV1).

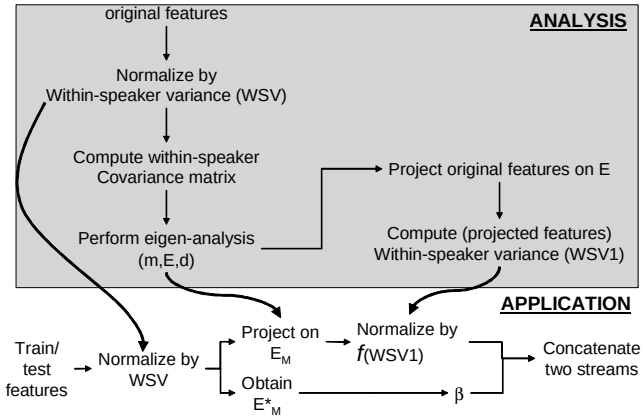

Figure 1. Block diagram for WCCN processing

During the application, the features are first normalized with WSV and projected onto E_M. The projections (V_1) have dimensionality M. Using the projections, a complement feature vector (V_2) is created by subtracting the reconstructed feature vector in E_M space from the original feature vector. This vector has dimensionality T. V_1 is normalized by a function of WSV_1, $f(WSV_1) = \alpha - (1-\alpha) \times \sqrt{WSV_1}$ and V_2 is weighted by a scalar β. Finally, the weighted V_1 and V_2 are concatenated to form a single $M+T$ dimensional feature vector. The scalars α and β are chosen on a development set.

3.2. Nuisance Attribute Projection

Solomonoff et al. proposed the NAP approach in [12]. The idea is based on principal component analysis and local linear embedding. The assumption is that unwanted variability can be sufficiently estimated in a high-dimensional feature space using second order statistics (the covariance matrix). Further, it is assumed that this variability lies in a lower-dimensional subspace spanned by the eigenvectors of the covariance matrix. Thus, one way to suppress the variability is to estimate this lower-dimensional subspace and remove it. Solomonoff et al. [2] have studied this approach extensively with many different ways of obtaining the covariance matrix. Recently, Burget et al. [16] used this technique in their 2006 SRE submission, where they applied this technique to their GMM-Supervector and MLLR features with SVM.

Figure 2 shows the analysis and application of NAP with a WCCN template to highlight the similarities between the two approaches. The analysis involves simply computing the WSCM and computing its eigenvectors. As mentioned earlier, the first N eigenvectors are ignored and the feature vector is reconstructed in the original feature space. The choice of N is made based on the performance of the system on the development set.

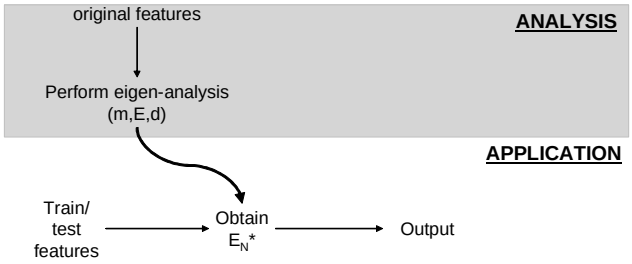

Figure 2. Block diagram for NAP processing with WCCN template for comparison

3.3. Comparison of WCCN and NAP

Apart from the WSV normalization for WCCN, these two techniques mainly differ in how the different eigenvectors are weighted. Here is a simple relationship between the two. As mentioned earlier, the total number of eigenvectors is T, We can partition the eigenbasis E_T such that

$$E = E_T = \left[E_M, E_{M \to T}\right]$$

Where E_M are the leading nonzero eigenvectors and $E_{M_\to T}$ are the eigenvectors corresponding to the zero eigenvalues. For practical purposes, $E_{M_\to T}$ are not computed explicitly but are computed as a complement of E_M :

$$xE_{M \to T} \Leftrightarrow x\left(I - E_M E_M^T\right) \Leftrightarrow xE_M^*,$$

where x is the MLLR feature vector.

A generic framework for generating new features can be defined as

$$x' = \left[f_M xE_M \ , \ CxE_M^*\right]$$

where f_M is a function that generates weightings for M eigenvectors and C is a constant used to weight the complement. With $f_M=f_{WSV}$, the features correspond to WCCN features. With $M=N$, $f_N=0$ and $C=1$, we obtain NAP features. With $f_M=C_1$ and $C=C_2$, as two constants, we obtain features proposed by [5], which we will come back to in Section 5.

4. RESULTS

As mentioned before, WCCN and NAP were originally proposed with different setups. WCCN used the SRE03 dataset for estimating the WSCM and Fisher corpus for the background data. The parameters α and β were trained on SRE04 and applied on SRE05. NAP was recently used with SRE04 data both for

estimating the WSCM and for obtaining the background data. The number of eigenvectors (N) to be ignored was decided using SRE05 data and the results were applied to SRE06.

In this paper, we adopt the NAP experimental setup. We estimate WCCN parameters (α and β) and NAP parameters (N) on SRE05 and apply the parameters to SRE06. For comparison with previous WCCN results, we use Fisher and SRE04 for background speakers. Again, note that the results presented here are without score normalization. The results are expected to be better with score normalization but we expect the trend to remain the same.

Table 1 shows the WCCN results for different conditions. Note that SRE05 is used to tune the parameters and thus shows the best possible improvement using this technique. The parameters are then applied to SRE06, showing how these techniques generalize to new data.

Table 1 shows that WCCN consistently gives 13-17% improvement over the baseline (row "N/A") on SRE05 data. The best performance is obtained using Fisher data for background speakers and SRE04 for estimating WSCM. However, the trend is different on SRE06 data. The configuration for the best performance on SRE05 actually gives the worst performance on SRE06. In addition, the results show a strong dependence on the choice of background data such that significant improvements are obtained with SRE04 data over Fisher data. The best performance on SRE06 is obtained using SRE04 data for background speakers and for the WSCM.

Table 1 WCCN Results (N/A=results without WCCN=baseline) Numbers in bold show the best performance

Back-ground data	Intersession variability estimated on	SRE05 (English) (DEV Set)		SRE06 (English) (EVAL set)	
		%EER	DCFx10	%EER	DCFx10
Fisher	N/A	5.872	0.190	4.639	0.224
	SRE03	5.066	0.154	4.314	0.198
	SRE04	**5.056**	**0.147**	4.477	0.216
SRE04	N/A	6.189	0.200	4.315	0.197
	SRE03	5.219	0.162	3.776	0.173
	SRE04	5.103	0.157	**3.603**	**0.166**

We analyze the results further by dividing the performance of WCCN features into those coming from normalized projections on E (referred to as V_1) and then appending these to the weighted reconstructed vector obtained from E^* (referred to as V_2). Table 2 shows the results corresponding to the results shown in the third and sixth column of Table 1 (best performance on SRE05 and SRE06). It also shows the best possible cheating performance obtained with parameters chosen on SRE06 data (last column, in parentheses). The results show that V_1 performs better than the baseline on SRE05 and the performance is further improved by adding V_2. However, the performance of V_1 does not generalize to SRE06, especially when Fisher data is used to model background speakers. We hypothesize that the lack of generalization is due to differences in the data collections for SRE04 and SRE06, e.g., the fact that SRE04 consists mostly of native speakers whereas SRE06 has a significant proportion of nonnative speakers. Further experimentation is needed to test and refine this hypothesis.

Table 3 shows the results using NAP on different datasets for background speakers and for estimating WSCM. Note that the results for the row "N/A" are slightly different from Table 1 due to implementation differences. In Table 1, the results are obtained by running the baseline system without WCCN. In Table 3, the results are obtained by removing zero eigenvectors (default case for NAP). As a part of the NAP procedure, the global mean is subtracted from the original features for this default case, which leads to small numerical differences in features and scores. However, the difference between the two "N/A" results is not significant (α=0.05). The purpose of showing different result for NAP is to validate the experimental procedure by verifying that the default case gives results similar to the baseline.

Table 2 WCCN results with session variability computed on SRE04 data showing breakdown of results ($E=E_M$)

Back-ground data	Experiment	SRE05 (English) (DEV Set)		SRE06 (English) (EVAL set)	
		%EER	DCFx10	%EER	DCFx10
Fisher	baseline	5.872	0.190	4.639	0.225
	E x $f(\mathrm{WSV}_1)$	5.542	0.182	5.556	0.266
	+ E^*x β	5.056	0.147	4.477	0.216
SRE04	baseline	6.189	0.200	4.315	0.197
	E x $f(\mathrm{WSV}_1)$	5.907	0.203	4.956	0.217
	+ E^* x β	5.103	0.157	3.603 (3.452)	0.166 (0.162)

The numbers in parentheses in the last row of each table show the cheating performance on SRE05, using optimal parameters for that dataset. These results show a different trend. NAP gives more improvement on SRE06 than SRE05. In addition, NAP seems to be more sensitive to the choice of data for WSCM than to the choice of data for background speakers. The best performance for NAP on SRE05 is with the Fisher-SRE04 configuration for background and ISV estimation, respectively, but the best performance on SRE06 is with SRE04-SRE04.

Table 3 NAP Results (N/A=results without NAP=baseline). Numbers in bold show the best performance

Back-ground Data	Intersession variability estimated on	SRE05 (English) (Dev Set)		SRE06 (English) (Eval Set)	
		%EER	DCF x10	%EER	DCF x10
Fisher	N/A	5.899	0.190	4.641	0.225
	SRE03	5.653	0.166	4.423	0.206
	SRE04	**5.470**	**0.158**	3.999	0.196
SRE04	N/A	6.189	0.202	4.312	0.197
	SRE03	5.744	0.172	3.831	0.180
	SRE04	5.664	0.163	**3.614** (3.567)	**0.170** (0.167)

Comparison of WCCN and NAP results shows a difference in the best configurations for SRE05 and SRE06. It also shows the importance of matched setups and that the worst-case mismatch in the configuration gives only a small improvement in performance. The comparison also shows the dependence of these techniques on choice of background corpus and data used for WSCM. Comparison of the cheating performance also shows that WCCN suffers from over-training more than NAP. This is not surprising because the former uses more parameters. However, the best performance for both methods is obtained with the SRE04-SRE04 configuration, where both methods give comparable results.

5. COMBINATION OF NAP AND WCCN

Next, we explore simple combinations of NAP and WCCN. The

idiosyncrasies of these approaches are as follows: NAP uses a very simple, binary weighting for the eigenvectors. WCCN models the subspace spanned by the eigenvectors corresponding to nonzero eigenvalues separately, and uses a complex weighting for the eigenvectors. Separating the subspaces is based on previous work [5], where it was shown that, for cepstral features, it is advantageous to model these subspaces separately. However, the weighting proposed in [5] was simpler than the one used in WCCN.

Two combinations of NAP and WCCN were devised, as follows:

1. NAP→WCCN – Obtain the best NAP result, separate the subspaces as WCCN (ignore leading N eigenvectors) and apply simple weights as suggested in [5].
2. WCCN→NAP – Obtain the best WCCN result and modify the weighting so the first few eigenvectors are set to zero.

The preliminary results of these combinations do not show a significant improvement over the best NAP and WCCN results, but they do show interesting trends. In the NAP→WCCN combination, the results show that separating the spaces does not give any advantage over combining them. It also shows the same trend as WCCN results whereby the performance of the features that are projections onto the eigenvectors (V_1) does not generalize from SRE05 to SRE06. In the WCCN→NAP combination, the results do not change significantly if the weights of the leading eigenvectors are set to zero. This shows that the proposed weighting scheme is optimal in the given setup. However, there is a potential for pursuing different functional forms (e.g., sigmoid) for more compact and generic weightings.

6. SUMMARY AND CONCLUSIONS

We compared two techniques – NAP and WCCN – for compensating for intersession variability in an MLLR-SVM speaker verification system. Both techniques model intersession variability as a within-speaker covariance matrix and weight the resulting eigenvectors to minimize the variability. We performed experiments with different sets of background data and with different databases for estimating the variability. The results show that NAP is more sensitive to the choice of data for the WSCM and WCCN is more sensitive to the choice of background set. In general, Fisher-SRE04 is the best combination for SRE05, and SRE04-SRE04 is the best combination for SRE06. WCCN gives the best performance on SRE05 but does not generalize to SRE06. We attribute this to the fact that projections onto normalized eigenvectors do not generalize from SRE05 to SRE06. Although WCCN requires more parameters and has generalization issues, it still performs comparably to NAP under the best configuration. Furthermore, we explored combinations of these two techniques. Our preliminary results show limited gains but there is a potential for using a functional form for the weightings of eigenvectors that will be more compact and more general. In summary, the application of NAP and WCCN techniques improves the performance of MLLR-SVM and makes it comparable to the best single systems in NIST 2006 SRE.

7. ACKNOWLEDGMENTS

The authors thank Andrew Hatch from ICSI for his help with WCCN experiments and Pavel Matejka from Brno University for his help with up NAP experiments. This work is funded under NMA401-02-9-2001 and NSF IIS-0544682. The views herein are those of the authors and do not necessarily reflect the views of the funding agencies.

8. REFERENCES

[1] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class Covariance Normalization for SVM-based Speaker Recognition," Proc. of ICSLP, Pittsburgh, PA, 2006.

[2] A. Solomonoff, C. Quillen, and I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," Proc. of ICASSP, Philadelphia, USA, 2005.

[3] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Mixture Models," *Digital Signal Processing*, vol. 10, pp. 181-202, 2000.

[4] W. M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," Proc. of ICASSP, Orlando, 2002.

[5] S. Kajarekar, "Four Weightings and a Fusion: A Cepstral-SVM system for speaker recognition," Proc. of ASRU, San Juan, 2005.

[6] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, vol. 46, pp. 455-472, 2005.

[7] D. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," Proc. of ICASSP, Hong Kong, China, 2003.

[8] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," Proc. of 2001: A Speaker Odyssey: The Speaker Recognition Workshop, Crete, Greece, 1996.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in Factor Analysis Based Speaker Verification," Proc. of ICASSP, Toulouse, France, 2006.

[10] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," Proc. of ICASSP, Philadelphia, PA, 2005.

[11] R. Vogt, B. Baker, and S. Shridharan, "Modeling Session Variability in Text-independent Speaker Verification," Proc. of Eurospeech, Lisbon, Portugal, 2005.

[12] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel Compensation for SVM Speaker Recognition," Proc. of Odyssey: The Speaker and Language Recognition Workshop, Toledo, Spain, 2004.

[13] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-Transform-based Speaker Recognition," Proc. of IEEE Odyssey 2006 Speaker and Language Recognition Workshop, San Juan, Puerto Rico, 2006.

[14] NIST, "http://www.nist.gov/speech/tests/spk/index.htm."

[15] B. Schoelkopf, A. J. Smola, and K.-R. Mueller, "Kernel Principal Component Analysis," *Lecture Notes in Computer Science*, pp. 583-588, 1997.

[16] L. Burget, P. Matejka, P. Shwarz, O. Glembek, M. Karafiat, J. Cernocky, and F. Grezl, "NIST Speaker Recognition Evaluation 2006," Proc. of NIST 2006 Speaker Recognition Workshop, San Juan, Puerto Rico, 2006.