# Analysis of Subspace Within-Class Covariance Normalization for SVM-based Speaker Verification

*Liang Lu*[1], *Yuan Dong*[1], *Xianyu Zhao*[2], *Jian Zhao*[1], *Chengyu Dong*[2], *Haila Wang*[2]

[1]Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]France Telecom Research & Development Center, Beijing, 100083, China
{luliang07, michaeljianzhao}@gmail.com
{xianyu.zhao, yuan.dong, chengyu.dong, haila.wang}@orange-ftgroup.com

## Abstract

Nuisance attribute projection (NAP) and within-class covariance normalization (WCCN) are two effective techniques for intersession variability compensation in SVM based speaker verification systems. However, by normalizing or removing the nuisance subspace containing the session variability can not guarantee to enlarge the distance between speakers. In this paper, we investigated the probability of using linear discriminant analysis (LDA) for discriminative training. To cope with the small sample size problem which prevents us from using LDA directly, we adapted the subspace LDA approach, which first projects the whole feature space into a relatively low dimensional subspace by PCA, and then performs LDA in the subspace. By some modification, the subspace LDA can be degenerated into a kind of WCCN approach, which we called subspace WCCN. Experiments on NIST SRE tasks showed that, the subspace WCCN outperformed the conventional direct WCCN, especially in low dimensional feature space.

**Index Terms**: Linear Discrimiant Analysis, Within-class Covariance Normalization, NAP, Support Vector Machine

## 1. Introduction

In recent years, the classifier of support vector machines (SVMs) has been successfully used in the task of speaker verification, and achieved state-of-the-art performance which can be comparable with conventional GMM-UBM systems, while the combination of the two are hopefully to further enhance the system performance [1]-[4]. To cope with the session variability, two techniques, namely, nuisance attribute projection (NAP) [5] and within-class covariance normalization (WCCN) [6, 7] are proposed for SVM-based systems and have achieve relatively promising result. Actually, NAP as well as its general version for nonlinear kernels [8] can be seen as to remove the nuisance subspace which mainly contains the session variability, while WCCN is mean to normalize the variability in the nuisance space in order to restrict the distance intra speakers.

Despite the success of NAP and WCCN, however, both of them still have their limitations. By removing or normalizing the nuisance subspace, the intra speaker distance is expected to be reduced, whereas the inter speaker distance, which is essential point for separation, is not expected to be enlarged accordingly. Actually, just as it is pointed in [9], the nuisance subspace selected by eigen-decomposition of the within-class covariance matrix also contains considerable amount of discriminative information of speakers, especially when the development set is relatively small when compared with the dimensionality of feature space. Both NAP and WCCN do not make well usage of this kind of information, and in addition, if the nuisance subspace is not properly selected, the performance of NAP will degrade apparently.

In this paper, we investigate the problem and try to enlarge the between class distance while restrict the within-class variability in feature space. For this target, linear discriminant analysis (LDA) is an appropriate approach, which is to select the directions to maximize the ratio of between-class scatter to that of within-class scatter. However, the insufficient training data can only offer a singular within-class scatter matrix which is impossible to solve the objective function of LDA. In addition, the low ranked between-class scatter matrix also makes LDA unable to find enough directions to describe the whole feature space. Hence, this difficulty, known as small sample size problem, makes it impossible to use LDA directly. For this, we borrowed the idea of subspace LDA [10] in face recognition filed which performs LDA in the subspace selected by PCA. By some modification, the subspace LDA approach can be degenerated into another equivalent form, which we named the subspace WCCN, and experiments shows that it outperforms the conventional direct WCCN approach, especially in low dimensional feature space.

The rest of paper is organized as follows. Section 2 presents a brief review of NAP and WCCN and section 3 describes the subspace LDA method and its equivalent form subspace WCCN in detail. Experimental results and discussion are given in section 4 and section 5 summarizes the paper as a conclusion.

## 2. SVM kernels of NAP and WCCN

NAP aims to remove the nuisance subspace that causes session variability in the kernel, and constructs a new one as

$$\begin{aligned} k(v_a, v_b) &= (Pb(v_a))^T (Pb(v_b)) \\ &= b(v_a)^T Pb(v_b) \\ &= b(v_a)^T (I - U_m U_m^T) b(v_b) \end{aligned} \tag{1}$$

where $U_m$ is a matrix whose columns are composed of eigenvectors corresponding to the top $m$ eigenvalues of the within-class covariance matrix, $S_W$,

$$S_W = \frac{1}{N}\sum_{i=1}^{c}\sum_{x_k \in X_i}(x_k - \mu_i)(x_k - \mu_i)^T \tag{2}$$

where $N$ is the total number samples and $c$ is the number of class. Different with NAP, for WCCN, the expected within class covariance matrix is used in the generalized linear kernel to suppress the session variability, i.e.

$$k(v_a, v_b) = v_a^T S_W^{-1} v_b \tag{3}$$

### 2.1. Overcome the limitations

Although NAP and WCCN are effective for session variability compensation, however, they are not mean to enlarge the between-class distance. In addition, the authors in [9] pointed out that, the nuisance space removed by NAP actually has considerable amount of speaker variability which is useful for discrimination. To avoid these problems, they proposed a method named discriminant NAP to increase its discriminative ability by using an inverse version of LDA criteria, which aims to find the nuisance space with high intra-speaker variability while low inter-speaker variability and then remove it. Different with their idea, in this paper, we try to use the subspace version of LDA criteria to find the most discriminative directions. More detailed description about this approach is presented in the follow-up section.

## 3. LDA and WCCN in Subspace

LDA attempts to find the optimal projection $W_{opt}$ as the matrix with orthonormal columns which maximizes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples, i.e.:

$$\begin{aligned} W_{opt} &= \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \\ &= \left[w_1 w_2 \cdots w_m\right] \end{aligned} \tag{4}$$

where $S_B$ and $S_W$ denote the between-class scatter matrix and within-class scatter matrix, respectively, and $S_W$ is defined as equation (2) while $S_B$ is as follows:

$$S_B = \frac{1}{N}\sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

If $S_w$ is a non-singular matrix, then the optimal projection $W_{opt}$ in (4) can be obtained by solving the eigenvalue problem of the following equation:

$$S_W^{-1} S_B w_i = \lambda_i w_i, i = 1, 2, \ldots, m \tag{6}$$

Unfortunately, in practice, $S_W$ is always singular since the number of training samples is far smaller than the dimensionality of the feature space. This difficulty, known as the small sample size problem, prevents us form using LDA directly for supervised discriminative training. To overcome it, different methods have been proposed and applied successfully to image retrieval, object and face recognition tasks. Here, we introduce a widely used method in face recognition filed, named subspace LDA or fisherfaces [10], which first projects the whole feature space into a low dimensional subspace by PCA, in which $S_W$ is full rank and do LDA in the relatively small subspace. More formally, the optimal matrix $W_{opt}$ in (4) would be:

$$W_{opt}^T = W_{lda}^T W_{pca}^T \tag{7}$$

where

$$W_{pca} = \arg\max_W \left|W^T S_T W\right|$$

$$W_{lda} = \arg\max_W \frac{\left|W^T W_{pca}^T S_B W_{pca} W\right|}{\left|W^T W_{pca}^T S_W W_{pca} W\right|}$$

$$S_T = S_B + S_W = \frac{1}{N}\sum_i^N (x_i - \bar{\mu})(x_i - \bar{\mu})^T$$

However, it is very interesting to find out that, the subspace LDA method can degenerate into subspace WCCN, when performed a little modification. As it is known that

$$W_{opt} = \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} = \arg\max_W \frac{\left|W^T S_T W\right|}{\left|W^T S_W W\right|} \tag{8}$$

And if the PCA basis are normalized by its corresponding eigenvalues, namely,

$$\begin{aligned} S_T &= W_{pca} \Lambda W_{pca}^T \\ \overline{W}_{pca} &= W_{pca} \Lambda^{-1/2} \end{aligned}$$

Using the normalized PCA basis, then subspace LDA would be as follows:

$$\begin{aligned} W_{lda} &= \arg\max_W \frac{\left|W^T \overline{W}_{pca}^T S_B \overline{W}_{pca} W\right|}{\left|W^T \overline{W}_{pca}^T S_W \overline{W}_{pca} W\right|} \\ &= \arg\max_W \frac{\left|W^T \overline{W}_{pca}^T S_T \overline{W}_{pca} W\right|}{\left|W^T \overline{W}_{pca}^T S_W \overline{W}_{pca} W\right|} \\ &= \arg\max_W \frac{\left|W^T I W\right|}{\left|W^T \overline{W}_{pca}^T S_W \overline{W}_{pca} W\right|} \end{aligned} \tag{9}$$

And the solution of equation (9) is to solve the eigenvalue problem of:

$$\left(\overline{W}_{pca}^T S_W \overline{W}_{pca}\right)^{-1} W_{opt} = W_{opt} \vec{\lambda} \tag{10}$$

Thus, in this case, the eigenvectors of $\overline{W}_{pca}^T S_W \overline{W}_{pca}$ is the most discriminative directions in normalized PCA subspace. And by normalizing those directions using their corresponding eigenvalues to suppress the variance, we will get the subspace WCCN, i.e.:

$$\begin{aligned} \overline{W}_{pca}^T S_W \overline{W}_{pca} &= U \tilde{\Lambda} U^T \\ \overline{W}_{wccn} &= U \tilde{\Lambda}^{-1/2} \end{aligned} \tag{11}$$

Hence, the subspace WCCN approach is equivalent with LDA criteria in the subspace, and compared with the conventional direct WCCN as equation (3) showed, it is more capable to explore the discriminative information in the training examples.

### 3.1. Some practical rules and discussion

To achieve satisfying results, the following rules and principles should be noticed when performing subspace WCCN.

1) When PCA is used for dimension reduction, the kernel trick can be used which allows doing PCA in the relatively small sample space, and then transforms to the original feature space.

2) Only used the PCA subspace will lose large amount of information in its orthogonal complement space. To avoid this problem, just as [7], the PCA-complement is also reserved. Thus the final feature space is actually as:

$$\Phi(x) = \begin{bmatrix} (1-\sigma)\cdot A^T \Phi_{PCA}(x) \\ \sigma \cdot \Phi_{\overline{PCA}}(x) \end{bmatrix}, \qquad \sigma \in [0,1] \tag{12}$$

where $A = \overline{W}_{pca}\overline{W}_{wccn}$ and $\Phi_{\overline{pca}}(x) = \left(I - W_{pca}W_{pca}^T\right)x$. $\sigma$ is set to be 0.5 in our experiments.

3) It should be noted that, because $S_T = S_W + S_B$, and compared with $S_W$, $S_B$ is much sparser (because there are only hundreds of speakers in the development set but the sessions can be thousands). Hence it is very likely that the total scatter matrix $S_T$ will be dominant by $S_W$ and the eigenvectors of $S_T$ will be nearly the same as that of $S_W$. In that case, the PCA subspace will mainly contain the within-class variability and the direct WCCN approach will have the similar performance with subspace WCCN, especially when the dimensionality of the original feature space is huge. This fact will be illustrated in the following up experiments.

4) Another interesting fact lies in that, the subspace WCCN approach is very similar with the method described in [7], which performs WCCN in no-normalized PCA subspace. Actually, there is no much difference between the two when $S_W$ is dominant in $S_T$. However, if the factor of $S_B$ can not be ignored, then the basis selected by the two are different. Limited by the length, however, we do not present the detailed comparison of the two approaches in this paper.

## 4. Experiments and Discussion

In this section, we report some experimental results of SVM-based speaker verification systems using NAP and WCCN for session variability compensation. The purpose of these experiments is to compare the performance of direct and subspace WCCN in low and high dimensional feature space, respectively. For this purpose, two systems are used in our experiments, one of which is MLLR based SVM system and the other is the widely used GMM supervector kernel based SVM system. All experiments were conducted on the 2006 NIST SRE corpus, and we focused on male part of the 1conv4w-1conv4w task, which contains 1570 true trails and 20561 false trails.

Before reporting the results of experiments, it is necessary to specify the direct WCCN approach in this paper. In practice, the within-class covariance matrix $S_W$ is always not full rank or ill-conditioning. Hence, it's not appropriate to use equation (3) directly. In this paper, we modified the direct WCCN approach into two modes as follows:

**Mode 1**: Only use the subspace of $S_W$ with high eigenvalues, i.e. the feature space would be

$$\Phi(x) = \overline{U}_n^T x \tag{13}$$

**Mode 2**: Concatenate two subspaces, just as equation (12) did, the feature space would be

$$\Phi(x) = \begin{bmatrix} (1-\sigma)\cdot \overline{U}_n^T x \\ \sigma \cdot \left(I - U_n U_n^T\right)x \end{bmatrix}, \qquad \sigma \in [0,1] \tag{14}$$

where in equation (13) and (14), $S_W = U\Lambda U^T$ and $\overline{U}_n = U_n \Lambda_n^{-1/2}$. $U_n$ is composed by the columns of $U$ corresponding to the top $n$ eigenvalues. $\sigma$ in equation (14) is set to be 0.5 in our experiments.

In this paper, we found in low dimensional feature space, where there are relatively sufficient training samples to estimate $S_W$, the direct WCCN with mode 1 will get slightly better performance while in large dimensional feature space, mode 2 is more appropriate. The detailed description of the experiments is in the follow-up sections.

### 4.1. Experiment on MLLR Kernels

In maximum likelihood linear regression (MLLR) [11], an affine transform (A, b) is applied to the Gaussian mean vectors to map from speaker-independent to speaker-dependent means by $\mu_s = A\mu + b$, where $A$ is a full matrix and $b$ is a vector. Using MLLR kernels for SVM-based speaker verification was first proposed in [12], which stacked the MLLR transformation coefficients together as features for SVM, and achieved very promising performance.

To achieve satisfying results, the set of phone models are always partitioned or clustered by similarity, and a separate transform is applied to each cluster. In our system, however, only one MLLR transform matrix was used to all the clusters, because our purpose of using MLLR based SVM is to examine the performance of subspace WCCN in the condition of low dimensional feature space. Thus, it may result in suboptimal baseline system performance.

As for the features, 18 MFCC coefficients plus C0 are computed, and then cepstral mean subtraction (CMS) and RASTA filtering are followed to alleviate the channel effects. The First order derivatives computed over 5 frames are appended to each frame vector, which results in dimensionality 38. Thus, the feature dimensionality of MLLR based SVM is $38 \times 39 = 1482$. The NIST SRE 2004 corpus was used as the development set for session variability compensation, in which we used 1790 sessions of 310 speakers. Hence, compared with the dimension of feature space extended by SVM kernel, the small sample size problem was not very serious.

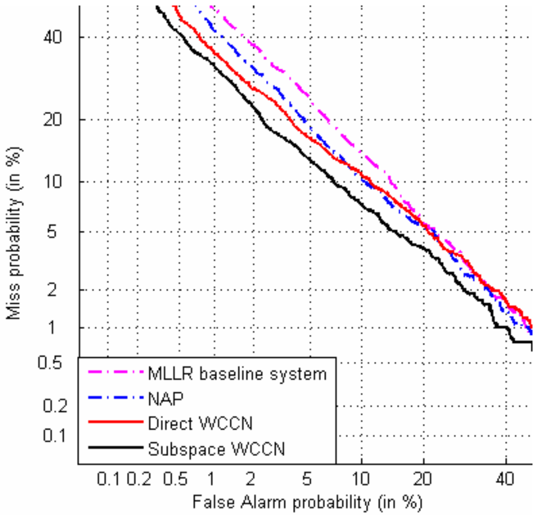

Figure 1: *Comparison of subspace WCCN, direct WCCN and NAP on MLLR based SVM system.*

#### *4.1.1.* Results of MLLR Kernels

In this experiment, The direct WCCN approach is performed as mode 1 described, where the rank of $\overline{U}_n$ is 800. For subspace WCCN, the rank of $\overline{W}_{pca}$ is 300 and the number of

NAP eigenvectors is 10. Figure 1 shows the results of subspace WCCN in MLLR based SVM system in the form of DET curves. It is obvious that, from figure 1, the subspace WCCN approach outperforms both direct WCCN and NAP significantly, about more than 15% relative improvement in both EER and min DCF. Hence, when the background data is relatively sufficient for the feature space, subspace WCCN can achieve really satisfying results, because it is a kind of discriminative training. Unfortunately, most kernels used in SVM-based speaker verification which can achieve state-of-the-art performance are huge in dimensionality of feature space and suffer from the small sample size problem. The following set of experiments is performed to examine the performance of subspace WCCN on one of such kernels, namely, GMM supervector kernels.

## 4.2. Experiment on GMM Supervector Kernels

GMM mean supervector kernels have been wildly used since it was proposed in [1]. A GMM mean supervector is formed by concatenated the component mean vectors of a MAP-adapted GMM that is $\mu(s)=\left[\mu_1(s)^T \ldots \mu_c(s)^T\right]^T$ , where $\mu_i(s)$ is the $i^{th}$ component mean.

For the features in the experiment, 12 MFCC coefficients plus C0 are computed and cepstral mean subtraction (CMS) and feature warping over 300 frames are applied. RASTA filtering of the features follows. First, second and third order derivatives computed over 5 frames are appended to each feature vector, which results in dimensionality 52. HLDA is used to reduce the feature dimension from 52 to 51. The number of Gaussian mixture components is 512 and the final dimension of supervector is 26112.

The NIST SRE 2004 corpus was used as the development set for session variability compensation, in which we used 4603 sessions of 310 speakers. In this experiment, the direct WCCN was performed as mode 2 described, and the rank of $\overline{U}_n$ is 400. For the subspace WCCN, the rank of $\overline{W}_{pca}$ is also 400. The number of NAP eigenvectors is 60.

### *4.2.1.* Results of GMM Supervector Kernels

Table 1 shows the results of NAP, direct and subspace WCCN on GMM supervector kernels in terms of EER and min DCF. Compared with the MLLR based system, it is obvious that subspace WCCN did not achieve so notable improvement, although it still slightly outperforms direct WCCN. The reason is just as it is discussed subsection 3.1, in more than 20 thousands dimensional feature space, the between class covariance matrix $S_B$ estimated by several hundred speakers is really spare and noisy. Hence, we can take for $S_T \approx S_W$ , and the directions found by PCA in $S_T$ are very similar with that of $S_W$ . Thus, that will make little difference between direct and subspace WCCN approach.

However, we can expect the subspace WCCN approach will achieve better performance when added more speakers and sessions in background training examples. But such training data is always difficult or expensive to collect, and even pooled in several hundreds more speakers, the training samples is still insufficient when compared to the feature space. Hence, the classical LDA criteria may not be appropriate for SVM classification in large feature space. Similar discussion is also presented in [9]. To prepare more discriminative features in SVM modeling, more works on this direction is still needed.

Table 1. *Results of GMM supervector kernels experiment*

| Φ | Baseline system | NAP | Direct WCCN | **Subspace WCCN** |
|---|---|---|---|---|
| **EER** | .0630 | .0554 | .0510 | **.0503** |
| **Min DCF** | .0300 | .0263 | .0250 | **.0245** |

# 5. Conclusion

In this paper, we investigated the subspace WCCN approach for discriminative training in SVM-based speaker verification systems. In this approach, WCCN is performed in the normalized PCA subspace and we demonstrated that it will be equivalent with the LDA criteria. Experiment on NIST sre06 corpus showed that, subspace WCCN can achieve significant performance gains over direct WCCN or NAP when the training samples are relatively sufficient for the feature space. However, in the large feature space where the training data is insufficient, the performance gains are limited, because the PCA subspace will mainly only contain the within-class variability. Our future work will attempt some modified LDA criteria for more efficient and robust discriminative training approaches, and examine the performance of subspace WCCN on some other tasks, e.g. short duration and cross channel tasks.

# 6. References

[1] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff, “SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *Proc. ICASSP*, 2006, pp. 97-100.

[2] H. Yang, Y. Dong, X. Zhao, L. Lu, and H. Wang, “Cluster adaptive training weights as features in SVM-based speaker verification,” in *Interspeech*, 2007, pp. 2013-2016.

[3] X. Zhao, Y. Dong, H. Yang, J. Zhao and H. Wang, “SVM-based speaker verification by location in the space of reference speakers,” in *Proc. ICASSP*, 2007, pp. 281-284.

[4] V. Wan, S. Renals, “Speaker verification using sequence discriminant support vector machines,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 203-210, 2005.

[5] A. Solomonoff, W. M. Campbell and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, 2005, pp. 629-632.

[6] A. O. Hatch, A. Stolcke, “Generalized linear kernels for one-versus-all classification: application to speaker recognition,” in *Proc. ICASSP*, 2006, pp. 585-588.

[7] A. O. Hatch, S. Kajarekar and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. ICSLP*, 2006, pp. 1471-1474.

[8] X. Zhao, Y. Dong, H. Yang, J. Zhao, L. Lu and H. Wang, "Nonlinear kernel nuisance attribute projection for speaker verification," in *proc. ICASSP*, 2008, pp. 4125-4128.

[9] R. Vogt, S. Kajarekar and S. Sridharan, "Discriminant NAP for SVM Speaker Recognition," in *proc. Odyssey Speaker and Language Recognition Workshop,* 2008.

[10] P.N. Belhumeur, J.p. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997.

[11] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[12] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg and A. Venkataraman, “MLLR transforms as features in speaker recognition,” in *Proc. Eurospeech*, 2005, pp. 2425-2528.