VARIABILITY COMPENSATED SUPPORT VECTOR MACHINES APPLIED TO SPEAKER VERIFICATION

Zahi N. Karam

MIT Lincoln Laboratory, Lexington MA
MIT, Cambridge MA

William M. Campbell

MIT Lincoln Laboratory, Lexington MA

ABSTRACT

Speaker verification using SVMs has proven successful, specifically using the GSV Kernel [1] with nuisance attribute projection (NAP) [2]. Also, the recent popularity and success of joint factor analysis [3] has led to promising attempts to use speaker factors directly as SVM features [4]. NAP projection and the use of speaker factors with SVMs are methods of handling variability in SVM speaker verification: NAP by removing undesirable nuisance variability, and using the speaker factors by forcing the discrimination to be performed based on inter-speaker variability. These successes have led us to propose a new method we call variability compensated SVM (VCSVM) to handle both inter and intra-speaker variability directly in the SVM optimization. This is done by adding a regularized penalty to the optimization that biases the normal to the hyperplane to be orthogonal to the nuisance subspace or alternatively to the complement of the subspace containing the inter-speaker variability. This bias will attempt to ensure that inter-speaker variability is used in the recognition while intra-speaker variability is ignored. In this paper we present the theory and promising results on nuisance compensation.

Index Terms— Support Vector Machines, Speaker Verification, Variability Compensation

1. INTRODUCTION

In a classification task there are two types of variability: the good which reflects the anticipated diversity needed for proper classification, and the bad which introduces undesirable information that confuses the classifier. An ideal classifier should, therefore, exploit the good and mitigate the bad. In the speaker verification task, inter-speaker variability is the desired variability and intra-speaker, e.g. channel and language, is the bad or nuisance variability. Techniques for handling nuisance, such as nuisance attribute projection (NAP) [2] and within class covariance normalization (WCCN) [5], are already used in SVM speaker verification. Recently, state of the art systems have revolved around joint factor analysis [3], which uses a Bayesian framework to incorporate estimates of subspaces containing nuisance and inter-speaker variability in the verification task. In this paper we introduce variability compensated SVM (VCSVM) which is a method to handle both good and bad variability by incorporating it directly into the SVM optimization. We will begin by motivating and describing our approach in a nuisance compensation framework. Modifications to the algorithm are then presented that allow for handling inter-speaker variability, as well as incorporating both the good and the bad variability simultaneously. We then discuss a probabilistic interpretation of the algorithm and finally present experimental results that demonstrate the algorithm's efficacy.

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions,and recommendations are those of the authors and are not necessarily endorsed by the United States Government. This work was also supported in part by the Texas Instruments Leadership University Program.

2. HANDLING NUISANCE VARIABILITY

Evidence of the importance of handling variability can be found in the discrepancy in verification performance between one, three and eight conversation enrollment tasks for the same SVM system. Specifically, for the SVM system in [1] performance improves from 5.0% EER for one conversation enrollment to 2.9% and 2.6% for three and eight, on all trials of the NIST SRE-Eval 06 core condition. One explanation for this is that when only one target conversation is available to enroll a speaker then the orientation of the separating hyperplane is set by the impostor utterances. As more target enrollment utterances are provided the orientation of the separating hyperplane can change drastically, as sketched in Figure 1. The additional information that the extra enrollment utterances provide is intra-speaker variability, due to channel, language, and other nuisance variables.

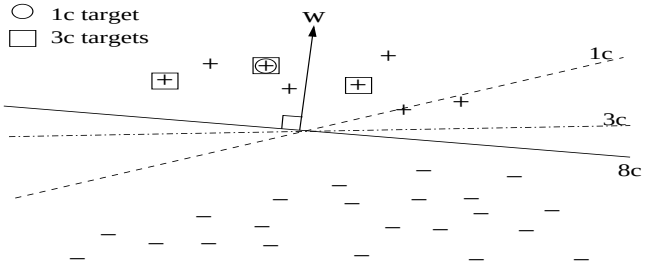


Fig. 1. *Different separating hyperplanes obtained with 1, 3, and 8 conversation enrollment.*

If an estimate of the principal components of intra-speaker variability for a given speaker were available then one could prevent the SVM from using that variability when choosing a separating hyperplane. However, since it is not possible in general to estimate intra-speaker variability for specific speakers, one could instead substitute a global estimate obtained from a large number of speakers. This is the approach taken by NAP, which handles nuisance variability by estimating a small subspace where the nuisance lives and removing it completely from the SVM features, i.e. not allowing any information from the nuisance subspace to affect the SVM decision. To handle this variability we propose VCSVM which allows for varying the degree to which the nuisance subspace is avoided by the classifier, rather than completely removing it.

Assume that the nuisance subspace is spanned by a set of N orthonormal eigenvectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N\}$, and let $\mathbf{U}$ be the matrix

whose columns are these eigenvectors. Let the vector normal to the separating hyperplane be $\mathbf{w}$. Ideally, if the nuisance was restricted to the subspace $\mathbf{U}$ then one would require the orthogonal projection of $\mathbf{w}$ in the nuisance subspace to be zero, i.e. $\left|\left|\mathbf{U}\mathbf{U}^T\mathbf{w}\right|\right|_2^2 = 0$. This requirement can be introduced directly into the primal formulation of the SVM optimization:

$$\min J(\mathbf{w},\epsilon) = ||\mathbf{w}||_2^2/2 + \xi\left|\left|\mathbf{U}\mathbf{U}^T\mathbf{w}\right|\right|_2^2/2 + C\sum_{i=1}^{m}\epsilon_i \quad (1)$$

$$\text{subject to } l_i(\mathbf{w}^T\mathbf{s}_i + b) \geq 1 - \epsilon_i \;\&\; \epsilon_i \geq 0,\;\; i = 0,\ldots,m$$

where $\xi \geq 0$, $\mathbf{s}_i$ denotes the utterance specific SVM features (supervectors) and l_i denotes the corresponding labels. Note that the only difference between (1) and the standard SVM formulation is the addition of the $\xi\left|\left|\mathbf{U}\mathbf{U}^T\mathbf{w}\right|\right|_2^2$ term, where ξ is a tunable (on some held out set) parameter that regulates the amount of bias desired. If $\xi = \infty$ then this formulation becomes similar to NAP compensation, and if $\xi = 0$ then we obtain the standard SVM formulation. Figure 2 sketches the separating hyperplane obtained for different values of ξ . We can rewrite the additional term in (1) as follows:

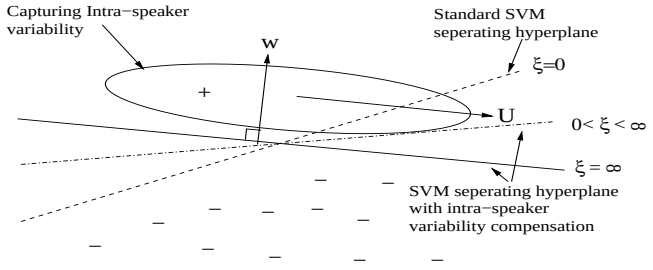


Fig. 2. *Sketch of the separating hyperplane for different values of ξ.*

$$\left|\left|\mathbf{U}\mathbf{U}^T\mathbf{w}\right|\right|_2^2 = (\mathbf{U}\mathbf{U}^T\mathbf{w})^T(\mathbf{U}\mathbf{U}^T\mathbf{w}) = \mathbf{w}^T\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T\mathbf{w} \quad (2)$$

$$= \mathbf{w}^T\mathbf{U}\mathbf{U}^T\mathbf{w}, \quad (3)$$

where the final equality follows from the eigenvectors being orthonormal ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$). Since $\mathbf{U}\mathbf{U}^T$ is a positive semi-definite matrix we can follow the recipe presented in [6] to re-interpret this reformulation as a standard SVM with the bias absorbed into the kernel. We begin by rewriting $J(\mathbf{w},\epsilon)$ in (1) as:

$$J(\mathbf{w},\epsilon) = \mathbf{w}^T(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)\mathbf{w}/2 + C\sum_{i=1}^{m}\epsilon_i, \quad (4)$$

and since $(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)$ is a positive definite symmetric matrix, then

$$J(\mathbf{w},\epsilon) = \mathbf{w}^T\mathbf{B}^T\mathbf{B}\mathbf{w}/2 + C\sum_{i=1}^{m}\epsilon_i, \quad (5)$$

where $\mathbf{B}$ can be chosen to be real and symmetric and is invertible. A change of variables $\tilde{\mathbf{w}} = \mathbf{B}\mathbf{w}$ and $\tilde{\mathbf{s}} = \mathbf{B}^{-T}\mathbf{s}$ allows us to rewrite the optimization in (1) as

$$\text{minimize} \quad J(\mathbf{w},\epsilon) = ||\tilde{\mathbf{w}}||_2^2/2 + C\textstyle\sum_{i=1}^{m}\epsilon_i \quad (6)$$

$$\text{subject to} \quad l_i(\tilde{\mathbf{w}}^T\tilde{\mathbf{s}}_i + b) \geq 1 - \epsilon_i \;\;\&\;\; \epsilon_i \geq 0, \quad i = 0,\ldots,m$$

which is then the standard SVM formulation with the following kernel:

$$K(\mathbf{s}_i,\mathbf{s}_j) = \mathbf{s}_i^T\mathbf{B}^{-1}\mathbf{B}^{-T}\mathbf{s}_j = \mathbf{s}_i^T(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)^{-1}\mathbf{s}_j. \quad (7)$$

Examining the kernel presented in (7) we realize that $(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)$ can be very large. This is of concern since the kernel requires its inverse. To circumvent this we use the Matrix Inversion Lemma [7] and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ to obtain:

$$\begin{aligned}(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)^{-1} &= \mathbf{I} - \sqrt{\xi}\mathbf{U}(\mathbf{I} + \xi\mathbf{U}^T\mathbf{U})^{-1}\sqrt{\xi}\mathbf{U}^T \\ &= \mathbf{I} - \xi\mathbf{U}[(1+\xi)\mathbf{I}]^{-1}\mathbf{U}^T \\ &= \mathbf{I} - \frac{\xi}{1+\xi}\mathbf{U}\mathbf{U}^T. \end{aligned} \quad (8)$$

The kernel can therefore be rewritten as:

$$K(\mathbf{s}_i,\mathbf{s}_j) = \mathbf{s}_i^T(\mathbf{I} - \frac{\xi}{1+\xi}\mathbf{U}\mathbf{U}^T)\mathbf{s}_j. \quad (9)$$

Examining (9) we notice that when $\xi = 0$ we recover the standard linear kernel, and more importantly when $\xi = \infty$ we recover exactly the kernel suggested in [2] for performing NAP channel compensation. An advantage of this formulation over NAP is that it does not make a hard decision to completely remove dimensions from the SVM features but instead leaves that decision to the SVM optimization.

It is of practical importance to note that (9) can be written as a linear combination of two kernels, and defining $\mathbf{x}_i = \mathbf{U}^T\mathbf{s}_i$ to be the channel factors:

$$K(\mathbf{s}_i,\mathbf{s}_j) = \mathbf{s}_i^T\mathbf{s}_j - \frac{\xi}{1+\xi}\mathbf{s}_i^T\mathbf{U}\mathbf{U}^T\mathbf{s}_j = \mathbf{s}_i^T\mathbf{s}_j - \frac{\xi}{1+\xi}\mathbf{x}_i^T\mathbf{x}_j. \quad (10)$$

This allows for a less costly implementation, because the two kernels need not be recomputed for each value of ξ and relatively little computation is required to obtain the second kernel, since the $\mathbf{x}_i$'s are typically low dimensional.

2.1. Should All Nuisance be Treated Equally?

As the choice of nuisance subspace gets larger one may find its more appropriate to handle directions within that subspace unequally, for example we might want to avoid using larger nuisance directions in discrimination more than we would smaller ones. One way to do this can be to use the eigenvalues corresponding to the different nuisance directions. Therefore, we allow the eigenvectors spanning the $\mathbf{U}$ matrix to be orthogonal but not orthonormal, specifically:

$$\mathbf{U}^T\mathbf{U} = \mathbf{\Lambda}, \quad (11)$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the eigenvalues corresponding to the columns of $\mathbf{U}$. We can now follow a formulation similar to that of the previous section, the difference will appear in the kernel when the matrix inversion lemma is applied:

$$K(\mathbf{s}_i,\mathbf{s}_j) = \mathbf{s}_i^T(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)^{-1}\mathbf{s}_j \quad (12)$$

$$= \mathbf{s}_i^T(\mathbf{I} - \sqrt{\xi}\mathbf{U}(\mathbf{I} + \xi\mathbf{U}^T\mathbf{U})^{-1}\sqrt{\xi}\mathbf{U}^T)\mathbf{s}_j$$

$$= \mathbf{s}_i^T(\mathbf{I} - \xi\mathbf{U}[\mathbf{I} + \xi\Lambda]^{-1}\mathbf{U}^T)\mathbf{s}_j. \quad (13)$$

An extreme example of this is where the whole SVM space is considered to contain nuisance information (i.e. $\mathbf{U}\mathbf{U}^T$ is full rank), which results in a formulation very similar to that of WCCN normalization [5]. WCCN proposes using inverse of the intra-speaker covariance matrix (i.e. full rank $\mathbf{U}\mathbf{U}^T$) as a kernel:

$$K(\mathbf{s}_i,\mathbf{s}_j) = \mathbf{s}_i^T(\mathbf{U}\mathbf{U}^T)^{-1}\mathbf{s}_j. \quad (14)$$

However, in practice $\mathbf{U}\mathbf{U}^T$ is ill-conditioned due to the noisy estimate and directions of very small nuisance variability, therefore

smoothing is applied to the intra-speaker covariance matrix to make inversion possible, and the WCCN suggested kernel becomes:

$$K(\mathbf{s}_i, \mathbf{s}_j) = \mathbf{s}_i^T((1-\alpha)\mathbf{I} + \alpha\mathbf{U}\mathbf{U}^T)^{-1}\mathbf{s}_j \qquad 0 \leq \alpha < 1. \quad (15)$$

Comparing (15) with (12) we see that they are similar. We should, however, mention that when $\mathbf{U}\mathbf{U}^T$ spans the full SVM space the ξ (in our implementation) and α (in the WCCN implementation) no longer set the amount of bias desired, instead they ensure that the kernel does not over-amplify directions with small amounts of nuisance variability.

3. USING INTER-SPEAKER VARIABILITY

Joint factor analysis [3] has been highly successful in the speaker verification task. Joint factor analysis estimates a "speaker" subspace, that captures good variability and is spanned by the columns of $\mathbf{V}$, and a "channel" subspace, that captures the nuisance and is spanned by the columns of $\mathbf{U}$. An utterance $\mathbf{s}_i$ is represented as a linear combination of a contribution from the speaker, $\mathbf{V}\mathbf{y}_i$, and one from the channel, $\mathbf{U}\mathbf{x}_i$, and a residual; where $\mathbf{y}_i$ are the speaker factors and $\mathbf{x}_i$ are the channel factors. Recently, promising results have been obtained by using just the speaker factors as features in a SVM speaker verification system. Based on this, we propose a VCSVM formulation similar to the one presented in the previous section to bias the SVM towards mostly using the data present in the inter-speaker variability space.

Assume that the inter-speaker subspace is spanned by a set of M orthonormal eigenvectors (eigenvoices) $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$, and let $\mathbf{V}$ be the matrix whose columns are these eigenvectors. Let the vector normal to the separating hyperplane be $\mathbf{w}$. Ideally if $\mathbf{V}$ captured all inter-speaker variability then we would want $\mathbf{w}$ to live in the $\mathbf{V}$ subspace and therefore be orthogonal to its complement, i.e. $\left|\left|(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{w}\right|\right|_2^2 = 0$. Similar to the previous section this requirement can be introduced directly into the primal formulation of the SVM optimization:

$$\min J(\mathbf{w}, \epsilon) = ||\mathbf{w}||_2^2/2 + \gamma\left|\left|(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{w}\right|\right|_2^2/2 + C\sum_{i=1}^{m}\epsilon_i \quad (16)$$

$$\text{subject to } l_i(\mathbf{w}^T\mathbf{s}_i + b) \geq 1 - \epsilon_i \quad \& \quad \epsilon_i \geq 0, \quad i = 0, \ldots, m$$

where $\gamma \geq 0$ is a tunable (on some held out set) parameter that enforces the amount of bias desired. If $\gamma = \infty$ then this formulation becomes similar to just using the speaker factors, and if $\gamma = 0$ then we obtain the standard SVM formulation. Note that since $\mathbf{I} - \mathbf{V}\mathbf{V}^T$ is a projection into the complement of $\mathbf{V}$ then we can replace it by $\mathbf{Q}\mathbf{Q}^T$, where $\mathbf{Q}$ is a matrix whose columns are the orthonormal eigenvectors that span the complement. With this substitution we obtain a formulation that is almost equivalent to that in (1), hence following the recipe in the previous section we see again can push the bias into the kernel of a standard SVM formulation. The kernel in this case is

$$K(\mathbf{s}_i, \mathbf{s}_j) = \mathbf{s}_i^T(\mathbf{I} - \frac{\gamma}{1+\gamma}\mathbf{Q}\mathbf{Q}^T)\mathbf{s}_j. \quad (17)$$

By substituting back $\mathbf{Q} = \mathbf{I} - \mathbf{V}\mathbf{V}^T$ we can rewrite (17) as:

$$K(\mathbf{s}_i, \mathbf{s}_j) = \mathbf{s}_i^T(\mathbf{I} - \frac{\gamma}{1+\gamma}(\mathbf{I} - \mathbf{V}\mathbf{V}^T))\mathbf{s}_j. \quad (18)$$

Note that we do not have to explicitly compute the orthonormal basis $\mathbf{Q}$, which can be rather large. When $\gamma = \infty$ the kernel becomes an inner-product between the speaker factors $\mathbf{y}_i = \mathbf{V}^T\mathbf{s}_i$:

$$K(\mathbf{s}_i, \mathbf{s}_j) = \quad \mathbf{s}_i^T\mathbf{V}\mathbf{V}^T\mathbf{s}_j = \mathbf{y}_i^T\mathbf{y}_j. \quad (19)$$

This kernel suggests that when one chooses to perform classification using only the inter-speaker subspace the resultant kernel is just an inner-product between the speaker factors.

4. INCORPORATING ALL VARIABILITY

A natural followup is to combine the previous sections into a single SVM formulation that attempts to handle all of the variation. In this section we will choose to treat all nuisance directions equally, however this can easily be extended to the setup in Section 2.1. This has to be done with some care since there is an overlap between the $\mathbf{U}$ subspace and the complement to the $\mathbf{V}$ subspace. Specifically, $\mathbf{U}$ lies in the complement of $\mathbf{V}$. With this in mind the resultant SVM formulation is as follows:

$$\begin{aligned} &\text{minimize } ||\mathbf{w}||_2^2/2 + \xi\left|\left|\mathbf{U}\mathbf{U}^T\mathbf{w}\right|\right|_2^2 \\ &\qquad +\gamma\left|\left|(\mathbf{I} - \mathbf{V}\mathbf{V}^T - \mathbf{U}\mathbf{U}^T)\mathbf{w}\right|\right|_2^2/2 + C\textstyle\sum_{i=1}^{m}\epsilon_i \\ &\text{subject to } l_i(\mathbf{w}^T\mathbf{s}_i + b) \geq 1 - \epsilon_i \quad \& \quad \epsilon_i \geq 0, \quad i = 0, \ldots, m. \end{aligned} \quad (20)$$

When $\xi = \gamma$ we obtain the inter-speaker result of Section 3 and if $\gamma = 0$ we obtain the intra-speaker result of Section 2. Recasting it as a standard SVM formulation yields the following kernel:

$$K(\mathbf{s}_i, \mathbf{s}_j) = \mathbf{s}_i^T(\mathbf{I} - \frac{\xi}{1+\xi}\mathbf{U}\mathbf{U}^T - \frac{\gamma}{1+\gamma}(\mathbf{I} - \mathbf{V}\mathbf{V}^T - \mathbf{U}\mathbf{U}^T))\mathbf{s}_j.$$

5. PROBABILISTIC INTERPRETATION

In [6], the author makes a connection between the suggested kernel and the probabilistic interpretation of SVMs proposed in [8]. The SVM problem can be thought of as one of maximization of the likelihood of $\mathbf{w}$ given the training data ($\{\mathbf{s}_i, l_i\}$ pairs) by writing it as

$$\max \; l(\mathbf{w}|\{\mathbf{s}_i, l_i\}) = -\mathbf{w}^T\mathbf{w}/2 - C\sum_{i=1}^{m}h(l_i(\mathbf{w}^T\mathbf{s}_i + b)), \quad (21)$$

where $h()$ is the hinge loss. In this formulation the SVM can be though of as just computing the MAP estimate of $\mathbf{w}$ given the training data, where the $\mathbf{w}^T\mathbf{w}$ term is essentially a Gaussian ($N(0, \mathbf{I})$) prior and the second term is the log-likelihood of the training data given $\mathbf{w}$. This Gaussian prior on $\mathbf{w}$ in the standard SVM does not bias the angle of $\mathbf{w}$ in any direction since the components of $\mathbf{w}$ in the prior are independent. In VCSVM, when we introduce the bias to handle the variability this only affects the first term in (21) and therefore changes the prior on $\mathbf{w}$ in the MAP estimation interpretation (we will focus on nuisance variability):

$$\begin{aligned} \max \; l(\mathbf{w}|\{\mathbf{s}_i, l_i\}) \quad = \quad & -\mathbf{w}^T(\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)\mathbf{w}/2 \\ & -C\sum_{i=1}^{m}h(l_i(\mathbf{w}^T\mathbf{s}_i + b)). \end{aligned} \quad (22)$$

The prior on the MAP estimate of $\mathbf{w}$ is still a Gaussian $N(0, (\mathbf{I} + \xi\mathbf{U}\mathbf{U}^T)^{-1})$ but with its principal components orthogonal to the nuisance subspace and the variance along the principle components set by ξ. Hence, the prior is biasing $\mathbf{w}$ to be orthogonal to the nuisance subspace. A similar connection can be made for the full setup proposed in Section 4.

6. EXPERIMENTAL RESULTS

We have chosen to demonstrate VCSVM in two scenarios, the first is as an alternative to NAP to handle nuisance in the GSV system presented in [1], and the second to handle nuisance in a system presented in [4] where SVM speaker verification is performed using low-dimensional speaker factors. The goal of this section is not to compare the performance of these two systems, but rather to show that VCSVM is applicable to both.

We begin with the speaker verification system proposed in [4],

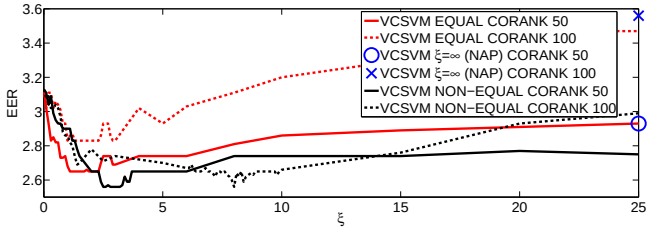


Fig. 3. *Results on English trials of the NIST SRE-Eval 06 core task with speaker factor SVM system: EER vs ξ for equal and non-equal weighting of nuisance subspace, and various subspace sizes.*

which represents each utterance using a vector of 300 speaker factors from the joint factor analysis system in [9]. The speaker factor vectors are normalized to have unit L_2-norm and used as features in a SVM speaker verification system. Figure 3 shows how the equal error rate (EER) changes as a function of ξ on our development set, the English trials of the NIST SRE-Eval 06 core task, for 50 and 100 dimensional nuisance subspaces when equal and non-equal weighting of the nuisance dimensions are used. The figure shows that non-equal weighting of the nuisance directions yields more favorable results than equal weighting. It also shows that VCSVM allows for nuisance compensation in such a small space, while NAP performs poorly since it completely removes the estimated nuisance dimensions which are a large percentage of the total dimensionality. Based on the development results we choose $\xi = 3$ and a corank of 50 for the VCSVM system and present results on all trials of the Eval 08 core task in Figure 5 (a).

Next, we present the performance of VCSVM using a GSV system

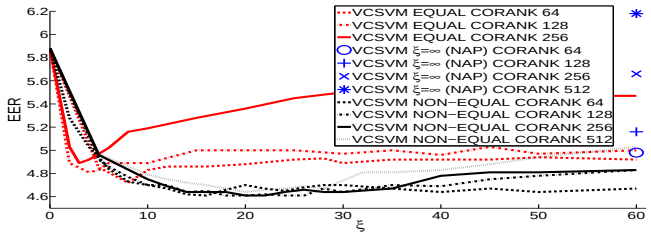
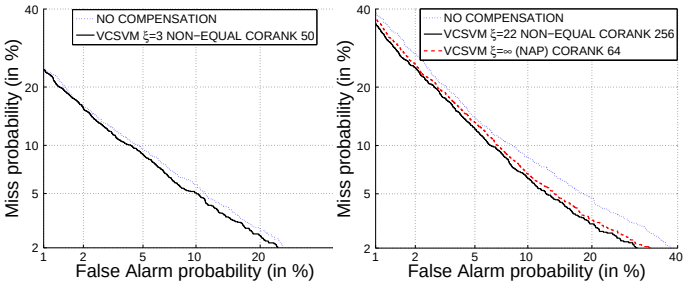


Fig. 4. *Results on all trials of the NIST SRE-Eval 06 core task with GSV system: EER vs ξ for equal and non-equal weighting of nuisance subspace, and various subspace sizes.*

[1] with 512 mixture GMMs and 38 dimensional, 19 cepstral and deltas, RASTA compensated feature vectors. Figure 4 presents results on the development set, all trials of the NIST SRE-Eval 06 core condition, of how the EER changes as a function of ξ, corank, and whether equal or non-equal weighting was used. Again this shows that non-equal weighting of the nuisance directions is preferable over equal weighting. It also shows that non-equally weighted VCSVM is fairly stable with regards to varying ξ and the corank, which is not the case with NAP. Based on these development results we compare, in Figure 5 (b), no nuisance compensation to the best-performing NAP system, with a corank of 64, and the best VCSVM system,



(a) Speaker factor SVM system (b) GSV System

Fig. 5. *Detection error plots on all trials of NIST Eval 08 core task.*

with $\xi = 22$ and corank of 256. We see that even in a large dimensional space such as this, it is preferable to not completely remove the nuisance subspace.

7. CONCLUSION

This paper presents variability compensated SVM (VCSVM), a method for handling both good and bad variability directly in the SVM optimization. This is accomplished by introducing into the minimization a regularized penalty, which biases the classifier to avoid nuisance directions and use directions of inter-speaker variability. With regard to nuisance compensation, an advantage of our proposed method is that it does not make a hard decision on removing nuisance directions, rather it decides according to performance on a held out set. Another benefit is that it allows for unequal weighting of the estimated nuisance directions, e.g. according to their associated eigenvalues. This flexibility allows for improved performance over NAP, increased robustness with regards to the size of the estimated nuisance subspace, and successful nuisance compensation in small SVM spaces. Future work will focus on using this method for handling inter-speaker variability and all variability simultaneously.

8. ACKNOWLEDGMENTS

The authors would like to thank the members of the "Robust Speaker Recognition Over Varying Channels" team at the JHU Summer Workshop 08, for the invaluable discussions and stimulating environment.

9. REFERENCES

[1] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, 2005.

[2] Alex Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proceedings of ICASSP*, 2005.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[4] Najim Dehak, Patrick Kenny, Reda Dehak, Ondrej Glember, Pierre Dumouchel, Lukas Burget, Valiantsina Hubeika, and Fabio Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *submitted to ICASSP*, 2009.

[5] Andrew O. Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proceedings of Interspeech*, 2006.

[6] Luciana Ferrer, Kemal Sonmez, and Elizabeth Shriberg, "A smoothing kernel for spatially related features and its application to speaker verification," in *Proceedings of Interspeech*, 2007.

[7] Mike Brookes, "The matrix reference manual," http://www.ee.ic.ac.uk/hp/staff/www/matrix/intro.html.

[8] P. Sollich, "Probabilistic interpretation and bayesian methods for support vector machines," in *Proceedings of ICANN*, 1999.

[9] P. Matejka et al., "BUT system for the NIST 2008 speaker recognition evaluation," *submitted to ICASSP*, 2009.