MAXIMUM MARGIN LINEAR KERNEL OPTIMIZATION FOR SPEAKER VERIFICATION

Mohamed Kamal Omar, Jason Pelecanos, Ganesh N. Ramaswamy

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
mkomar,jwpeleca,ganeshr@us.ibm.com

ABSTRACT

This paper describes a novel approach for discriminative modeling and its application to automatic text-independent speaker verification. This approach maximizes the margin between the model scores for pairs of utterances belonging to the same speaker and for pairs of utterances belonging to different speakers. A low-dimensional linear kernel is estimated which maximizes this margin. This approach emphasizes features which have a better ability to discriminate between scores belonging to pairs of utterances of the same target speakers and those of different speakers. In this paper, we apply this approach to the NIST 2005 speaker verification task. Compared to the Gaussian mixture model (GMM) baseline system, a 17.7% relative improvement in the minimum detection cost function (DCF) and a 11.7% relative improvement in equal error rate (EER) are obtained. We achieve also a 5.7% relative improvement in EER and 2.3% relative improvement in DCF by using our approach on top of a nuisance attribute projection (NAP) compensated GMM based kernel baseline system.

Index Terms— Speaker verification, maximum margin, discriminative training, GMM, nuisance attribute projection

1. INTRODUCTION

Maintaining data security and authenticity in speech-driven telephony applications can be performed effectively through speaker verification. Current automatic speaker verification systems face significant challenges caused by adverse acoustic conditions. Telephone band limitation, channel/transducer variability, as well as the natural speech variability all have a negative impact on the performance of speaker verification systems. Degradation in the performance of speaker verification and recognition systems due to inter-session variability has been one of the main challenges to actual deployment of speaker verification and recognition technologies.

A number of techniques have been proposed to solve these problems, including feature warping [1], feature mapping [2], and score normalization techniques like H-norm [9] and T-norm [3]. More recent approaches to compensate for channel effects and speech variability in the training and the testing utterances include using factor analysis [4], within-class covariance normalization (WCCN) [5], and nuisance attribute projection (NAP) [6]. Both WCCN and NAP modify a generalized linear kernel for a GMM based kernel to mitigate the effects of inter-session variability in the feature space.

Sequence kernel based methods have become one of the most important techniques for speaker verification. Since the amount of available training data for speaker verification is usually limited, and is affected by inter-session variability, the choice of the proper kernel was addressed many times in previous work. Most of previous work focused on generalized linear kernels with a high-dimensionality kernel feature space [7]. In [7], the linear kernel was set to a function of the enrollment and test utterances which included an inverse of the covariance matrix of the development data statistics. NAP is one of the most successful and widely used techniques for performing session compensation in speaker verification systems [6]. Our implementation of NAP is a variation of linear discriminant analysis (LDA) where the across-class covariance matrix is set to the identity matrix. However, the approach does not weight the retained directions in the feature space. In WCCN [5], the linear kernel is set to the inverse of the expected within-class covariance matrix over all speakers in the training data. This choice is motivated by showing that it minimizes a particular upper bound of the error of a binary classification problem which involves multiple classes.

In this paper, we propose training a low-dimensional kernel using semi-definite programming to maximize the margin between same-speaker and different-speaker inner product scores of the high-dimensional GMM mean supervector representation of utterances in the training data. The advantages of this approach compared to previous approaches include optimizing a discriminative criterion which can be made directly related to any weighted sum of the false alarm and the missing probabilities like the EER and the DCF objective functions. Also the estimation of the kernel is in a low-dimensional space to reduce the required amount of training data and the computational complexity associated with estimating the kernel. This makes our approach more suitable to applications with limited training data like speaker verification and speaker identification.

In the next section, we will formulate the problem and describe our objective criterion. In Section 3, the details of estimating the elements of the low-dimensional linear kernel to optimize our objective criterion are described. The experiments performed to evaluate the performance of our approach are described in Section 4. Finally, Section 5 contains a discussion of the results and future research.

In this paper, both vectors and matrices are in capital letters to be distinguished from scalars. Matrices are in boldface to be distinguished from vectors.

2. PROBLEM FORMULATION

In this section, we will discuss how the problem of estimating the linear kernel, which maximizes the margin between the model scores for pairs of utterances belonging to the same speaker and for pairs of utterances belonging to different speakers, can be reduced to an instance of semi-definite programming problems.

Before score normalization, the output scores of most GMM-based and generalized linear kernel SVM-based one-to-one match speaker verification systems can be represented by some kind of generalized inner product of two vectors representing the verification and the enrollment utterances.

This can be described by the relation

$$s = E^T \mathbf{K} V, \tag{1}$$

where E is the supervector representing the enrollment utterance, V is the supervector representing the verification utterance, $\mathbf{K}$ is a positive semi-definite matrix, and s is the score corresponding to this pair of utterances. This representation can be more simplified to a standard inner-product relation by the substitutions

$$\Phi_e = \sqrt{\mathbf{K}} E, \tag{2}$$

$$\Phi_v = \sqrt{\mathbf{K}} V, \tag{3}$$

to be

$$s = \Phi_e^T \Phi_v. \tag{4}$$

For a GMM based kernel, typically both Φ_e and Φ_v are vectors in a high dimensional space of dimension equal to the product of the feature vector dimension and the number of Gaussian probability density functions in the universal background Gaussian mixture model. Estimating a linear transform in such a space will be hindered by such problems as data scarcity, training overfitting, and computational complexity. To avoid these problems, we choose to represent our vectors as $m \times n$ matrices, where m and n are arbitrary dimensions such that the product of the two dimensions of the matrix, $m \times n$, is equal to the dimension of the mean supervectors Φ_e and Φ_v. The score s is related to these two matrices using the relation

$$s = tr\left[\mathbf{X}^T \mathbf{Y}\right], \tag{5}$$

where $tr\,[.]$ is the trace of the matrix, $\mathbf{X}$ and $\mathbf{Y}$ are the $m \times n$ matrix representations of the supervectors Φ_e and Φ_v respectively.

Our goal is to introduce a positive semi-definite matrix $\mathbf{M}$ which is estimated by maximizing the margin between scores corresponding to pairs of utterances of the same speaker and those corresponding to pairs of utterances of different speakers. The scores generated by our system become

$$s = tr\left[\mathbf{X}^T \mathbf{M} \mathbf{Y}\right], \tag{6}$$

where $\mathbf{M}$ is an $m \times m$ matrix.

Therefore it can be shown by following the same steps for deriving the optimization problem for soft-margin SVM classifiers, [8], that our objective function to be minimized is

$$O(\mathbf{M}) = \alpha ||\mathbf{M}||_F + \lambda \sum_{i=1}^{p} \zeta_i + \gamma \sum_{d=1}^{q} \zeta_d \tag{7}$$

subject to the constraints

$$\mathbf{M} \geq \mathbf{0}, \tag{8}$$

$$tr\left[\mathbf{X}^{iT} \mathbf{M} \mathbf{Y}^i\right] \geq \beta - \zeta_i \quad \text{for } i = 1, 2, \ldots, p, \tag{9}$$

$$tr\left[\mathbf{X}^{iT} \mathbf{M} \mathbf{Y}^j\right] \leq -\beta + \zeta_d \quad \text{for } d = 1, 2, \ldots, q \text{ and } i \neq j, \tag{10}$$

$$\zeta_i \geq 0 \quad \text{for } i = 1, 2, \ldots, p, \tag{11}$$

$$\zeta_d \geq 0 \quad \text{for } d = 1, 2, \ldots, q, \tag{12}$$

where $||\mathbf{M}||_F$ is the Frobenius norm of the matrix $\mathbf{M}$ and is given by

$$||\mathbf{M}||_F = \sqrt{tr\left[\mathbf{M}\mathbf{M}^T\right]} \tag{13}$$

$$= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} m_{ij}^2}. \tag{14}$$

Here, m_{ij} is the element of matrix $\mathbf{M}$ in row i and column j, $\mathbf{X}^i$ is a matrix representation of an utterance from speaker i, $\mathbf{Y}^j$ is a matrix representation of an utterance from speaker j, $\mathbf{M} \geq \mathbf{0}$ means that $\mathbf{M}$ is a positive semi-definite matrix, α is the weight of the Frobenius norm of the matrix $\mathbf{M}$, ζ_i is the same-speaker slack variable fot the ith same-speaker utterance pair, ζ_d is the different-speaker slack variable fot the dth different-speaker utterance pair, p is the number of utterance pairs in the training data which belong to the same speaker, q is the number of utterance pairs in the training data which belong to different speakers, λ is the weight assigned to the term corresponding to the sum of the same-speaker slack variables, γ is the weight assigned to the term corresponding to the sum of the different-speaker slack variables, and β is a control variable to improve the numerical stability of the optimization problem.

Making use of the fact that the sum in Eq. 14 is usually dominated by the diagonal elements, we replace the Frobenius norm of the matrix $\mathbf{M}$ in Eq. 7 with the trace norm of the matrix $\mathbf{M}$ to simplify the optimization problem. Also taking into consideration that

$$tr\left[\mathbf{AB}\right] = tr\left[\mathbf{BA}\right], \tag{15}$$

the objective function becomes

$$O(\mathbf{M}) = \alpha tr\left[\mathbf{M}\right] + \lambda \sum_{i=1}^{p} \zeta_i + \gamma \sum_{d=1}^{q} \zeta_d \tag{16}$$

subject to the constraints

$$\mathbf{M} \geq \mathbf{0}, \tag{17}$$

$$tr\left[\mathbf{M}\mathbf{Y}^i\mathbf{X}^{iT}\right] \geq \beta - \zeta_i \quad \text{for } i = 1, 2, \ldots, p, \tag{18}$$

$$tr\left[\mathbf{M}\mathbf{Y}^j\mathbf{X}^{iT}\right] \leq -\beta + \zeta_d \quad \text{for } d = 1, 2, \ldots, q \text{ and } i \neq j, \tag{19}$$

$$\zeta_i \geq 0 \quad \text{for } i = 1, 2, \ldots, p, \tag{20}$$

$$\zeta_d \geq 0 \quad \text{for } d = 1, 2, \ldots, q. \tag{21}$$

The values of the matrices $\mathbf{Y}^j\mathbf{X}^{iT}$ and $\mathbf{Y}^j\mathbf{X}^{iT}$ in the margin constraints are mean normalized by removing the mean of their values over all the training utterance pairs.

3. IMPLEMENTATION

In this section, we present our implementation of the approach described in the previous section. Our goal is to calculate the elements of the weighting matrix $\mathbf{M}$ which minimize the objective function in Eq. 16 subject to the constraints in Equations 17- 21.

For all utterances in the development data, a mean based supervector is generated. For Gaussian mixture models (GMMs), a useful supervector is established by concatenating a vector function of the Gaussian means into a supervector. A GMM with c mixture components is used to construct the high-dimensional supervectors for the

enrollment utterance,Φ_e, and the verification utterance, Φ_v. These supervectors are constructed as follows

$$\Phi_i = \sqrt{w_i}\mathbf{\Sigma}_i^{-\frac{1}{2}}\left(\mu_i^{adapt} - \mu_i^{ubm}\right), \quad (22)$$

$$\Phi = \left[\Phi_1^T \Phi_2^T \ldots \Phi_c^T\right]^T, \quad (23)$$

where w_i is the weight of the ith Gaussian component in the GMM, μ_i^{adapt} is the MAP adapted mean for this component, μ_i^{ubm} is the universal background model (UBM) mean for this component, and $\mathbf{\Sigma}_i$ is the diagonal covariance matrix of the ith Gaussian component in the GMM. We use the single iteration MAP adaptation presented by Reynolds [9] to generate the utterance specific adapted means, $\left\{\mu_i^{adapt}\right\}$, from the UBM means, $\left\{\mu_i^{ubm}\right\}$.

For one set of experiments, the mean GMM based supervectors are NAP-compensated [6]. To determine the NAP subspace, the high-dimensional feature space directions with the greatest intra-class variability were used [5]. Let there be a series of nuisance directions described by a column-wise eigenvector matrix, $\mathbf{V}$. The nuisance directions may be removed from an utterance representation, Φ, using the following equation

$$\hat{\Phi} = (\mathbf{I} - \mathbf{V}\mathbf{V}^T)\Phi, \quad (24)$$

where $\hat{\Phi}$ is the NAP-compensated supervector, $\mathbf{I}$ is the identity matrix, and Φ is the supervector before the NAP compensation. After applying the subspace removal to both the enrollment and the verification supervectors, the NAP-compensated vectors were then used directly within an inner product scoring metric and the NAP-compensated score is given by

$$\hat{s} = \hat{\Phi}_e^T \hat{\Phi}_v^T, \quad (25)$$

$$= \Phi_e^T (\mathbf{I} - \mathbf{V}\mathbf{V}^T)\Phi_v^T, \quad (26)$$

since $(\mathbf{I} - \mathbf{V}\mathbf{V}^T)$ is a projection matrix and therefore $(\mathbf{I} - \mathbf{V}\mathbf{V}^T) = (\mathbf{I} - \mathbf{V}\mathbf{V}^T)^2$. These NAP-compensated scores are then normalized by ZT-norm [3, 9] to determine the final output scores.

To estimate the elements of the matrix $\mathbf{M}$ which minimize the objective function in Eq. 16 subject to the constraints in Eq. 17 to Eq 21, we used the C library for semi definite programming (CSDP) [10]. The CSDP library is designed to handle constraint matrices with general sparse structure. CSDP can handle inequality constraints by converting them to equality constraints with additional non-negative auxiliary variables. The positive semi-definite matrix which the library tries to estimate to solve the optimization problem in Eq. 16 to Eq 21 consists of three blocks: the $\mathbf{M}$ matrix and two diagonal matrices which have on the diagonal the values of the slack variables and the auxiliary variables to convert inequality constraints to equality constraints. Therefore the dimensions of this three-block matrix are $(2p + 2q + m) \times (2p + 2q + m)$.

4. EXPERIMENTS

The performance of the maximum margin linear kernel (MMLK) system was evaluated on the common and the core conditions of the NIST 2005 Speaker Recognition Evaluation (SRE) [11]. The utterances consist of one side from approximately 5 minutes of a two channel telephone conversation. This provides, on average, approximately two and a half minutes of usable speech.

The development data set consists of a combination of audio from the NIST 2004 speaker recognition database and the Switchboard II Phase III corpora. The collection contains 4862 utterances: 2105 utterances of male speakers and 2757 of female speakers. The total number of speakers in the development data is 978 speakers: 536 female speakers and 442 male speakers. Thus, on average, there are almost 5 utterances per speaker to estimate the expected within-class covariance matrix over all speakers in the development data for our implementation of the NAP compensation.

The front-end features consist of 38 dimensional features forged from 19 cepstral coefficients and their corresponding deltas. There are 24 filters in the filter bank, over a frequency range of 125-3800 HZ, used to generate these cepstral coefficients. Feature warping is applied to the resulting feature vectors [1]. Each utterance in both the training and the testing data is represented by a GMM mean based supervector of dimension 9728. This representation was generated using a UBM of 256 Gaussian components using MAP adaptation. The system performance was measured at two operating points, namely in terms of the Equal-Error Rate (EER) and the minimum Detection Cost Function (DCF) as defined in the evaluation plan [11].

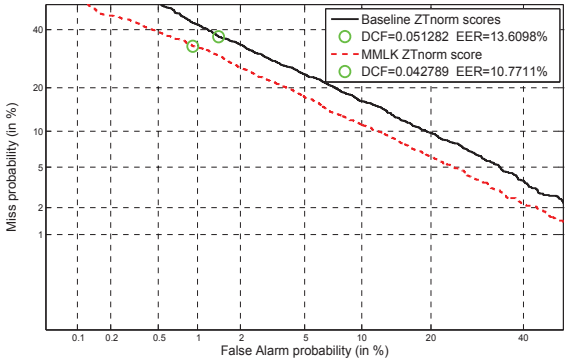

Fig. 1. Baseline and MMLK Results on the NIST05 Common Condition.

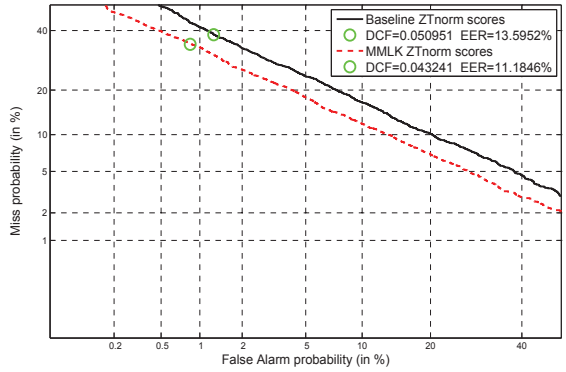

Fig. 2. Baseline and MMLK Results on the NIST05 Core Condition.

Two sets of experiments were conducted to test our technique. In the first set of experiments, we used as a baseline the GMM system which generates the scores for each pair of utterances using the inner product of the corresponding GMM based mean supervectors. ZT-Norm is applied to these scores to generate the final scores. For the MMLK system, we estimated a kernel of dimensions 38×38, (i.e. it has the same dimension as the original feature vector). A held-out set of 1550 utterances from the development data was used to tune the parameters in Equations 16 to 21; namely α, λ, β, and γ. The final values were found by updating one parameter at a time in relatively small steps until very small changes in performance measured in DCF or EER on the held-out set were found. The final value for α was 0.01, for λ was 1.05, for γ was 9.98, and for β was 0.05. To balance the number of pairs of different-speaker utterances and same-speaker utterances, we restricted the different-speaker utterances to speakers of the same gender and within 20% of the value of the average score. We removed outlier values from the same-speaker utterances which had a score more than 50% larger or smaller than the average score. As shown in Figure 1, the maximum margin kernel method reduces the DCF objective function by 17.7% relative and reduces the EER objective function by 11.7% relative on the common condition subset of the NIST05 evaluation set. As shown in Figure 2, the maximum margin kernel method reduces the DCF objective function by 15.2% relative and reduces the EER objective function by 17.8% relative on the core condition subset of the NIST05 evaluation test set.

After verifying the efficiency of our approach in the first setup, we performed the experiments of the second setup with a baseline that has the nuisance attribute projection applied to the GMM based mean supervectors before generating the scores using an inner product of the compensated supervectors and then applying ZT-Norm to these scores to generate the final scores. In this setup, a maximum margin linear kernel of dimensions 38×38 is estimated using the NAP-compensated GMM-based mean supervectors. We generated two sets of tuning parameters: one optimized on the DCF of the held out data set, and the other is optimized on the EER of the held out data set. The final values of the former are $\alpha = 0.01$, $\lambda = 1$, $\gamma = 9.9$, and $\beta = 0.02$. The final values for the latter are $\alpha = 0.01$, $\lambda = 1$, $\gamma = 10.3$, and $\beta = 0.02$. As shown in Table 1, the best MMLK system reduces the DCF compared to the NAP-GMM baseline system by 2.3% relative and reduces the EER compared to the NAP-GMM baseline system by 5.7% on the common condition subset of the NIST05 evaluation set. The results in Table 2 show that the best MMLK system reduces the DCF compared to the NAP-GMM baseline system by 2.1% relative and reduce the EER compared to the NAP-GMM baseline system by 3.7% for the core condition subset of the NIST05 evaluation set.

System	min. DCF	EER
NAP Baseline	0.01857	5.4136%
$\text{MMLK}_{\text{DCF-tuned}}$	**0.01814**	5.40227%
$\text{MMLK}_{\text{EER-tuned}}$	0.01851	**5.1086%**

Table 1. Comparison of the NAP-compensated Baseline and the MMLK systems on the NIST05 Common Condition

5. CONCLUSIONS

In this paper, we examined an approach for optimizing a low-dimensional linear kernel to maximize the margin between the inner product scores corresponding to pairs of utterances of the same speaker and those corresponding to pairs of utterances of different speakers. We applied this approach to the NIST 2005 automatic speaker verification task. This approach decreased the minimum DCF by 17.7% and the EER by 11.7% compared to the ZT-Norm GMM baseline system. We achieved also small gains in DCF and EER compared to a NAP-compensated baseline system. This improvement may be attributed to emphasizing elements in the feature vector which better discriminate between different speakers.

Further investigation of the performance of our approach on other evaluation tasks will be our main goal.

System	min. DCF	EER
NAP Baseline	0.019473	6.004%
$\text{MMLK}_{\text{DCF-tuned}}$	**0.019065**	**5.7819%**
$\text{MMLK}_{\text{EER-tuned}}$	0.019343	5.8174%

Table 2. Comparison of the NAP-compensated Baseline and the MMLK systems on the NIST05 Core Condition

6. REFERENCES

[1] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, Crete,Greece, 2001, pp. 213–218.

[2] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. of ICASSP*, Orlando, Florida, 2002, pp. 53–56.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, January 2000.

[4] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," The Centre de Recherche Informatique de Montreal, Technical Report CRIM-06/08-13, 2005.

[5] A. Hatch, S. Kajarekar, and A. Stolke, "Within-Class covariance normalization for SVM-based speaker recognition" in *Proc. of Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, 2006.

[6] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition" in *Proc. of ICASSP*, Philadelphia, PA, 2005.

[7] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation" in *Proc. of ICASSP*, Toulouse, France, 2006.

[8] V. N. Vapnik, *Statistical Learning Theory,* Wiley, New York, NY, 1998.

[9] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[10] B. Borchers, "CSDP, A C Library for Semidefinite Programming," *Optimization Methods and Software*, vol. 11, no. 1, pp. 613–623, 1999.

[11] National Institute of Standards and Technology, "The NIST Year 2005 Speaker Recognition," http://www.nist.gov/speech, 2008.