SPARSE REPRESENTATION OVER LEARNED AND DISCRIMINATIVELY LEARNED DICTIONARIES FOR SPEAKER VERIFICATION

Haris B. C. and Rohit Sinha

Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati -781039, India
{haris, rsinha}@iitg.ernet.in

ABSTRACT

In this work, a speaker verification (SV) method is proposed employing the sparse representation of GMM mean shifted supervectors over learned and discriminatively learned dictionaries. This work is motivated by recently proposed speaker verification methods employing the sparse representation classification (SRC) over exemplar dictionaries created from either GMM mean shifted supervectors or i-vectors. The proposed approach with discriminatively learned dictionary results in an equal error rate of 1.53 % which is found to be better than those of similar complexity SV systems developed using the i-vector based approach and the exemplar based SRC approaches with session/channel variability compensation on NIST 2003 SRE dataset.

Index Terms: speaker verification, learned dictionary, sparse representation, GMM mean supervector.

1. INTRODUCTION

Speaker verification (SV) task refers to the authentication of persons using their voice samples. The current state-of-the-art SV systems are based on the total variability i-vectors derived from the GMM mean supervectors for modeling speakers [1]. In last few years, there is a lot of interest generated about *sparse representation* and *compressive sensing* which provide a new directions to signal processing research. Recently the discriminative abilities of the sparse representation have also been exploited in various areas of the pattern recognition such as face recognition, texture classification, and speaker recognition. Following the work in [2] on face recognition by the sparse representation classification (SRC) with an exemplar dictionary, a similar approach for speaker identification task with an exemplar dictionary created using GMM mean supervectors was explored in [3]. In that work, the exemplar dictionary was created by arranging the supervectors corresponding to all speakers in training data as columns. The test data supervector was represented as the sparse linear combination of the atoms (columns) of the dictionary. The test supervector was assigned to the class associated to the atom having the highest non zero coefficient in the sparse vector.

Later the SRC with exemplar dictionary approach was extended to the speaker verification task in [4]. In that work, the dictionary for verifying a claim was constructed by arranging the GMM mean supervectors of the claimed speaker utterance and that of a set of imposter speaker utterances. The GMM mean supervector of the test utterance is represented as a sparse linear combination of the atoms of the dictionary. For verification purpose, the coefficient of the sparse vector corresponding to the target speaker vector is compared to that of the imposter speaker vectors with a suitable metric. In [5], the similar idea was explored with exemplar dictionary created using the total variability i-vectors. These SRC based approaches reportedly found to give competitive but lower performances in comparison to the existing high performing i-vector based approach. In the context of sparse representation, it is well known fact that the learned dictionaries not only outperform the exemplar ones but also are more data-independent [6]. Motivated by these facts, in this work, we propose a novel speaker verification approach employing sparse representation over learned and discriminatively learned dictionaries. The NIST 2003 SRE dataset is used for evaluating the performance.

The paper is organized as follows: In Section 2, the proposed SV system and the dictionary leaning algorithms are described. The different contrast SV systems and some session/channel variability compensation methods that are used in this work are briefly described in Section 3 and Section 4, respectively. Section 5 provides the details about database and experimental setup. The results are discussed in Section 6 and conclusions are given in Section 7.

2. PROPOSED SV SYSTEM USING SPARSE REPRESENTATION OVER LEARNED DICTIONARIES

We present a novel speaker verification system employing sparse representation over *learned dictionaries* and is referred to as SR-SV system in this work. In [7], the use of GMM mean shifted supervectors was explored for the SRC with exemplar dictionary on face video verification task. The mean shifting of the GMM supervectors was reported to enhance incoherence among atoms of the exemplar dictionary. Motivated by that we also use the GMM mean shifted supervectors for modeling. The GMM mean shifted supervector, $\boldsymbol{y}$ for a speaker utterance is defined as,

$$\boldsymbol{y} = \boldsymbol{s} - \boldsymbol{m} \tag{1}$$

where, $\boldsymbol{s}$ is the GMM mean supervector for the speaker utterance obtained from MAP adapted GMM-UBM and $\boldsymbol{m}$ is the speaker-independent UBM mean supervector. We model $\boldsymbol{y}$ using the sparse representation with a learned dictionary $\boldsymbol{D}$ as,

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \tag{2}$$

The dictionary $\boldsymbol{D}$ is of $\boldsymbol{M} \times \boldsymbol{N}$ size where $\boldsymbol{M}$ corresponds to the dimension of supervector and $\boldsymbol{N}$ is the number of atoms. $\boldsymbol{D}$ can be learned on a suitable development data using algorithms described later. $\boldsymbol{x}$ is the sparse vector and is estimated using orthogonal matching pursuit (OMP) algorithm which minimizes l_0-norm with a constraint on the representation error as,

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x}\|_0 \ \text{ such that, } \ \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 < \epsilon \tag{3}$$

The sparse vector $\hat{x}$ obtained can be considered as a compact representation of the speaker.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

In case of learned dictionaries, the classification cannot be done simply by comparing the coefficients of the sparse representation of the test utterance as done in case of the exemplar dictionaries since the learned dictionaries do not have class labels associated with their atoms. To overcome this problem, we have devised a scoring method which compares the sparse representation of the test utterance with that of the claimed speaker's training utterance. We have used the cosine kernel metric for finding the similarity between the claimed and the test sparse vectors and that is compared with a threshold for the verification purpose as,

$$\frac{< \hat{\boldsymbol{x}}_{clm}. \, \hat{\boldsymbol{x}}_{tst} >}{\|\hat{\boldsymbol{x}}_{clm}\| \, \|\hat{\boldsymbol{x}}_{tst}\|} \lesseqgtr \gamma \quad \text{(Threshold)} \tag{4}$$

where $\hat{\boldsymbol{x}}_{clm}$ and $\hat{\boldsymbol{x}}_{tst}$ represent the sparse representations of the claimed and the test speakers, respectively.

In the following subsections, we describe two existing dictionary learning algorithms that are used for learning the dictionaries for the proposed SR-SV system.

2.1. The KSVD algorithm

The KSVD [6] is one of the most widely used algorithms for learning redundant dictionaries for sparse representations. It is a generalization of the well known K-means clustering algorithm. KSVD algorithm constructs a dictionary of K atoms that leads to the best possible representation for each member of the training examples with a minimum sparsity constraint. The dictionary learning problem is represented as,

$$\min_{\boldsymbol{D},\boldsymbol{X}} \left\{ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_2^2 \right\} \quad \text{subject to } \|\boldsymbol{x}_i\|_0 \leq T_0 \quad \forall i \tag{5}$$

where, $\boldsymbol{Y}$ is the set of dictionary training vectors, $\boldsymbol{D}$ is the dictionary, $\boldsymbol{X}$ is the set of sparse vectors corresponding to $\boldsymbol{Y}$ and T_0 is the constraint on sparsity. The learning is an iterative process and each iteration has two stages: the sparse coding stage and the dictionary update stage. In the sparse coding stage, any of the pursuit methods such as OMP can be used for finding the sparse representation of the given set of examples based on the current dictionary. The update of the dictionary atoms is done jointly with an update of the sparse representation coefficients related to it, thus resulting in accelerated convergence.

2.2. SKSVD algorithm

The SKSVD [8] is a *supervised* version of the KSVD algorithm for learning discriminative dictionary. It uses *class supervised simultaneous* OMP (CSSOMP) in the sparse coding stage of the dictionary learning process which differs from OMP in two aspects: (i) CSSOMP uses the same set of atoms from the dictionary to represent all examples from a given class and so attempts to extract the common internal structure of that class whereas OMP treats each example independently (ii) In addition to the original reconstruction criterion of minimum squared error used in OMP, CSSOMP also uses a discrimination measure which increases the separability among classes. The sparse discriminant dictionary learning problem is represented as,

$$\max_{\boldsymbol{D},\boldsymbol{X}} \left\{ \theta . J \left(\left\{ \left\{ \boldsymbol{x}_i^j \right\}_{i=1}^{n_j} \right\}_{j=1}^{c} \right) - \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left\| \boldsymbol{y}_i^j - \boldsymbol{D}\boldsymbol{x}_i^j \right\|_2^2 \right\}$$
$$\text{subject to} \quad \left\| \boldsymbol{x}_i^j \right\|_0 \leq T_0, \quad \forall i, j \tag{6}$$

The function $J(.)$ represents the discriminant measure defined as $:= \frac{trace(\boldsymbol{B})}{trace(\boldsymbol{W})}$ where $\boldsymbol{B}$ and $\boldsymbol{W}$ are the *between-class* and the *within-class* covariance matrices of the learning data, respectively. $\boldsymbol{D}$ is the learned dictionary, $\boldsymbol{y}_i^j$ is i^{th} example vector of j^{th} class from a set of dictionary training data having c classes with $n_j, \ 1 \leq j \leq c$ examples per class. $\boldsymbol{x}_i^j$ is the sparse coefficient vector corresponding to $\boldsymbol{y}_i^j$. θ is a parameter controlling the trade-off between discriminative and re-constructive terms in the learning criterion.

3. CONTRAST SPEAKER VERIFICATION SYSTEMS

In this work, for comparison purpose two different kinds of contrast SV systems are developed, one based on the cosine kernel scoring of either the GMM mean shifted supervectors or the i-vectors while the other based on the SRC with exemplar dictionaries created using either the GMM mean shifted supervectors or the i-vectors.

3.1. Total variability i-vector based SV system

The total variability i-vector based speaker verification system is the state-of-the-art method for speaker verification. In this, the GMM mean shifted supervector $\boldsymbol{y}$ for a speaker utterance is projected to a lower dimensional subspace as,

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{w} \tag{7}$$

where $\boldsymbol{T}$ is the low rank matrix referred to as 'total variability matrix' and the projection $\boldsymbol{w}$ is referred to as 'i-vector'. The total variability matrix $\boldsymbol{T}$ is learned on a large development data using probabilistic PCA method. The i-vector for a given GMM mean shifted supervector is found by using the pseudo-inverse of the matrix $\boldsymbol{T}$. Speaker verification is done by comparing the i-vectors corresponding to the test utterance and the claimed speaker's training utterance using the cosine kernel metric [1].

We have also implemented an SV system using the GMM mean shifted supervectors with cosine kernel scoring similar to the i-vector approach for better contrast purpose.

3.2. SRC with exemplar dictionary based SV system

Recently, two speaker verification systems employing SRC over exemplar dictionary using GMM mean supervectors and i-vectors are proposed in [4] and [5], respectively. In both methods, the exemplar dictionary is created for each of the claims by arranging the vectors representing the claimed speaker and a set of background speakers as,

$$\boldsymbol{D}_{clm} = [\boldsymbol{y}_{clm}, \ \boldsymbol{y}_{bg_1}, \boldsymbol{y}_{bg_2}, \ldots, \boldsymbol{y}_{bg_M}] \tag{8}$$

where $\boldsymbol{y}_{clm}$ denotes the appropriate vector representation of the claimed speaker's training utterance and $\{\boldsymbol{y}_{bg_i}\}_{i=1}^{M}$ denote those of M background speakers' utterances taken from the development data. The test vector $\boldsymbol{y}_{tst}$ is represented by a sparse vector $\boldsymbol{x}_{tst}$ over the dictionary $\boldsymbol{D}_{clm}$ as $\boldsymbol{y}_{tst} = \boldsymbol{D}_{clm}\boldsymbol{x}_{tst}$. For a given $\boldsymbol{y}_{tst}$ and $\boldsymbol{D}_{clm}$ the estimate of $\boldsymbol{x}_{tst}$ is obtained as,

$$\hat{\boldsymbol{x}_{tst}} = \underset{\boldsymbol{x}_{tst}}{\operatorname{argmin}} \|\boldsymbol{x}_{tst}\|_0 \quad \text{such that,} \quad \|\boldsymbol{y}_{tst} - \boldsymbol{D}_{clm}\boldsymbol{x}_{tst}\|_2^2 < \epsilon \tag{9}$$

The score for verification is found using the l_1-norm ratio metric given by $\|\delta_1(\hat{\boldsymbol{x}_{tst}})\|_1 / \|\hat{\boldsymbol{x}_{tst}}\|_1$ where, $\delta_1(\hat{\boldsymbol{x}_{tst}})$ is a vector whose nonzero entries are the only entries in the first element of $\hat{\boldsymbol{x}_{tst}}$. In our implementation we have used GMM mean shifted supervectors instead of GMM mean supervectors for being consistent with other methods.

4. SESSION/CHANNEL VARIABILITY COMPENSATION

The session/channel variability compensation methods form an integral part of all current SV systems. In the following, we describe in brief the different session/channel variability compensation methods that are applied to different SV systems considered in this work.

4.1. Joint factor analysis

In joint factor analysis (JFA) [9], the GMM mean shifted supervector $\boldsymbol{y}$ for a speaker is represented as the sum of three factors as,

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{u} + \boldsymbol{V}\boldsymbol{v} + \boldsymbol{D}\boldsymbol{d} \tag{10}$$

where $\boldsymbol{U}$ is the session/channel subspace matrix, $\boldsymbol{V}$ is the speaker subspace matrix, and $\boldsymbol{D}$ represents the diagonal residual matrix. The vectors $\boldsymbol{u}$, $\boldsymbol{v}$ and $\boldsymbol{d}$ are the projections of $\boldsymbol{y}$ in their respective subspaces. The session/channel compensated GMM mean shifted supervector is given by $\boldsymbol{y}' = \boldsymbol{V}\boldsymbol{v} + \boldsymbol{D}\boldsymbol{d}$. In our implementation, we have used $\boldsymbol{V}\boldsymbol{v}$ factor only ignoring the residual factor and the compensated supervectors of the training and testing utterances are compared using cosine kernel method as suggested in [10]

4.2. Linear discriminant analysis

Linear discriminant analysis (LDA) is a commonly used method for dimensionality reduction and is widely used in pattern recognition applications. In LDA, the feature vectors are projected down to a set of new orthogonal axises where the discrimination between different classes is maximum. The projection matrix is composed by the eigen vectors corresponding to the best eigen values of the eigen analysis equation, $(\boldsymbol{W}^{-1}\boldsymbol{B})\boldsymbol{v} = \boldsymbol{\lambda}\boldsymbol{v}$, where $\boldsymbol{W}$ is the within-class covariance matrix, $\boldsymbol{B}$ is the between-class covariance matrix, $\boldsymbol{v}$ is an arbitrary vector, and $\boldsymbol{\lambda}$ is the diagonal matrix of eigen values [1].

4.3. Within class covariance normalization

In within class covariance normalization (WCCN) method, the feature vectors are transformed using a matrix which minimizes the upper bounds on the classification error metric and hence minimizes the classification error [11]. The transformation matrix $\boldsymbol{B}$ is obtained by Cholesky decomposition of the inverse of the within-class covariance matrix $\boldsymbol{W}$ as, $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t$.

5. EXPERIMENTAL SETUP

The experiments are performed using the NIST 2003 SRE database. It contains speech data of 356 target speakers collected over cellular phone network. The evaluation of the system is done as per the NIST 2003 SRE evaluation plan for primary task [12]. This experimental setup contains 24981 trials for verification task including true and false trials. The standard MFCC feature vectors of 39-dimensions with cepstral mean and variance normalization are used. An energy based VAD is used for selecting the speech frames. The Switchboard Cellular Part 2 corpus is used as the development data for all the systems. A gender-independent UBM model of 1024 Gaussian mixtures created using approximately 10 hours of the development speech data is used for all the systems. The GMM supervectors are created by adapting only the mean parameters of the UBM using maximum *a posteriori* (MAP) approach with the speaker specific data. The total variability matrix of 400 columns for the i-vector based system and the dictionary of 400 atoms for the proposed SR-SV systems are created using 1872 speech utterances taken from the development database. θ of value 0.7 is used for learning the discriminative dictionary. 400 imposter speaker utterances from the development database are used for creating the dictionary for the SRC system with exemplar dictionary. The JFA is made up of 300 speaker factors and 100 channel factors without the residual factor. The LDA and WCCN matrices are created using the same development data which is used for learning the total variability matrix and the dictionaries. The LDA for the i-vector system uses 250 top dimensions where as the proposed SRC based system uses LDA of 375 top dimensions. All the above mentioned parameters are chosen out of experimentation. The performance of the SV systems are evaluated using the equal error rate (EER) and the minimum detection cost function (minDCF).

Table 1. *Performances of proposed and various contrast speaker verification systems on NIST 2003 SRE dataset*

System		EER (%)	minDCF
Cosine kernel	super vector	8.42	0.161
	i-vector	4.21	0.072
SRC, xmplr dict	super vector	6.50	0.117
	i-vector	6.78	0.121
Proposed SR-SV	lrnd dict	5.23	0.097
	discr. lrnd dict	**2.89**	**0.051**

6. RESULTS AND DISCUSSION

The performances of the contrast systems and the proposed SR-SV system on NIST 2003 SRE dataset are given in Table 1. The cosine kernel scoring based systems with the GMM mean shifted supervectors and the i-vectors have resulted in an EER of 8.42 % and 4.21 %, respectively. The reduced dimension i-vector based system has already reported to significantly outperform the much larger dimension GMM mean shifted supervector based system. The SRC over exemplar dictionary based systems result in an EER of 6.50 % and 6.78 % for the GMM mean shifted supervector and the i-vector cases, respectively. As reported in the literature [4, 5], the performances of both SRC with exemplar dictionary based systems turn out to be lower than that of the i-vector based system. In the proposed SR-SV system, the simple learned and discriminatively learned dictionaries have been tried and have resulted in an EER of 5.23 % and 2.89 %, respectively. It is to note that among the four sparse representation based systems tried, the learned dictionary ones have significantly outperformed the exemplar ones and thus emphasizing the effectiveness of learned dictionaries for classification task.

There are some obvious similarities between the proposed SR-SV and the i-vector based systems those can be noted by comparing Eq. 2 with Eq. 7. The matrices $\boldsymbol{D}$ and $\boldsymbol{T}$ have the same size and the projections $\boldsymbol{x}$ and $\boldsymbol{w}$ of the GMM mean shifted supervector derived from those matrices are used for classification with the same scoring metric. The main differences between the two lie in the different criteria used for learning those matrices and the nature of the projections derived from them. The projection in case of the SR-SV system is sparse while the one in case of the i-vector based system is full. Further, we hypothesize that training of the reduced rank total variability matrix $\boldsymbol{T}$ and learning of the redundant dictionary $\boldsymbol{D}$ have somewhat similar goals i.e., to develop a more compact model for classification. The better performance of the i-vector based system compared to that of the SR-SV system with simple learned dic-

Table 2. *Performances of proposed and various contrast speaker verification systems with session/channel compensation on NIST 2003 SRE dataset*

System + session/channel comp.		EER (%)	minDCF
Cosine kernel	super vector + JFA	3.61	0.066
	i-vector + LDA + WCCN	2.24	0.037
SRC, xmplr dict	super vector + JFA	4.01	0.069
	i-vector + LDA + WCCN	5.42	0.102
Proposed sprs rep. lrnd dict.	discrm. dict + LDA	2.75	0.048
	discrm. dict + WCCN	2.71	0.049
	discrm. dict + JFA	**1.53**	**0.028**

tionary is hypothesized to be the attribute of the probabilistic PCA method used for the creation of $\boldsymbol{T}$ matrix which could have imparted some discriminative ability to it. On the other hand when an explicit discriminative criterion is employed in dictionary learning it significantly boosts the classification ability. As a result, proposed SR-SV system with discriminatively learned dictionary has shown significantly improved performance in comparison to the state-of-the-art system based on the i-vectors.

To explore the effectiveness of the proposed SV system in presence of the session/channel variability compensation, we have applied suitable methods among JFA, LDA and WCCN to different SV systems considered. In SR-SV systems only the one with discriminatively learned dictionary is considered as it's performance is much better than that of the simple learned dictionary one. For the cosine kernel scoring of GMM mean shifted supervector based system, the supervectors are JFA compensated prior to scoring. But for the i-vector based system, LDA and WCCN are applied to the i-vector as suggested in [1]. Similarly, for the SRC over exemplar dictionary based systems, the appropriate compensation methods are applied consistent with the kind of vectors used. For the considered SR-SV system, the compensation is applied in two ways: one as LDA/WCCN applied to the sparse projection and other as JFA applied to the supervectors prior to the dictionary learning. The performance of different SV systems on NIST 2003 SRE dataset with appropriate kind(s) of session/channel compensation applied are given in Table 2. On comparing Table 1 and Table 2, we note that the relative ordering of the performances of different systems considered remains the same with and without application of session/channel variability compensation. The two best performing systems after session/channel variability compensation are the SR-SV with discriminatively learned dictionary and the i-vector based system having EER of 1.53 % and 2.24 %, respectively. Note that both systems have undergone similar relative improvement over their uncompensated performances with the application of suitable session/channel compensation. Further we note that for sparse representation based systems, preprocessing of the the vector with JFA has been found to be more effective compared to postprocessing with LDA/WCCN.

7. CONCLUSIONS

A novel SR-SV system has been proposed employing the sparse representation of the GMM mean shifted supervectors over the learned dictionaries. For dictionary learning, both simple as well as discriminative criteria have been explored. The proposed system was compared to two recently suggested SRC over exemplar dictionary based SV systems as well as the existing i-vector based SV system. On NIST 2003 SRE dataset, the proposed SR-SV system with discriminatively learned dictionary is found to outperform all other SV systems considered both with and without the session/channel variability compensation. As a future work, we would like to explore the reasons behind such enhanced performance exhibited by the proposed SV system in detail and also to evaluate it on a more up-to-date publicly available datasets such as NIST 2005 SRE.

8. ACKNOWLEDGEMENT

This work has been supported by the ongoing project grant No. 12(4)/2009-ESD sponsored by the Department of Information Technology, Government of India.

9. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788 –798, May 2011.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[3] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for speaker identification," in *Proc. ICPR*, Aug 2010, pp. 4460–4463.

[4] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," in *Proc. ICASSP*, May 2011, pp. 4548–4551.

[5] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Proc. Interspeech*, Aug 2011, pp. 2729–2732.

[6] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[7] M. Li and S. Narayanan, "Robust talking face video verification using joint factor analysis and sparse representation on gmm mean shifted supervectors," in *Proc. ICASSP*, May 2011, pp. 4835–4838.

[8] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," IMA Preprint 2213, University of Minnesota, Tech. Rep., Jun 2008.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1448 –1460, May 2007.

[10] D. Garcia-Romero and C. Y. Espy-Wilson, "Joint factor analysis for speaker recognition reinterpreted as signal coding using overcomplete dictionaries," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010.

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

[12] NIST 2003 Speaker Recognition Evaluation Plan, www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.pdf.