Speaker Verification using Sparse Representations on Total Variability I-Vectors

Ming Li[1], *Xiang Zhang*[2], *Yonghong Yan*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, USA
[2]Key Laboratory of Speech Acoustics and Content Understanding,
Chinese Academy of Sciences, Beijing, China

mingli@usc.edu

Abstract

In this paper, the sparse representation computed by l^1-minimization with quadratic constraints is employed to model the i-vectors in the low dimensional total variability space after performing the Within-Class Covariance Normalization and Linear Discriminate Analysis channel compensation. First, we propose the background normalized l^2 residual as a scoring criterion. Second, we demonstrate that the Tnorm can be efficiently achieved by using the Tnorm data as the non-target samples in the over-complete dictionary. Finally, by fusing with the conventional i-vector based support vector machine (SVM) and cosine distance scoring system, we demonstrate overall system performance improvement. Experimental results show that the proposed fusion system achieved 4.05% (male) and 5.25% (female) equal error rate (EER) after Tnorm on the single-single multi-language handheld telephone task of NIST SRE 2008 and outperformed the SVM baseline by yielding 7.1% and 4.9% relative EER reduction for the male and female tasks, respectively.

Index Terms: speaker verification, sparse representation i-vector modeling

1. Introduction

The use of joint factor analysis (JFA) [1, 2, 3] has contributed to state of the art performance in text independent speaker verification and hence is being widely used. It is a powerful technique for compensating the variability caused by different channels and sessions.

Recently, total variability i-vector modeling has gained significant attention due to its excellent performance, low complexity and small model size [4]. In this modeling, first, a single factor analysis is used as a front end to generate a low dimensional total variability space which models both the speaker and channel variabilities [4]. Then, within this total variability space, channel variability compensation methods, such as Within-Class Covariance Normalization (WCCN) [5], Linear Discriminative analysis (LDA) and Nuisance Attribute Projection (NAP) [6], are performed to reduce the channel variability. Finally, two classification approaches, namely support vector machine (SVM) and cosine distance scoring (CDS), are proposed for the verification task [4]. It is also shown in [4] that LDA followed by WCCN achieved the best performance. In this paper, we follow this framework and focus on further enhancing the performance of the total variability i-vector modeling, notably by exploring the sparse representations of the i-vectors.

More recently, a sparse representation computed by l^1-minimization (to approximate the l^0-minimization) with equality constraints was proposed to replace the SVM in the GMM mean supervector modeling and has been demonstrated to be effective in the closed set speaker identification task on the clean TIMIT database [7]. This approach was extended in our previous work [8] to handle the robust verification task (with multiple enrollment samples) against large session variabilities. In this approach [8], first, the sparse representation is computed by l^1-minimization with quadratic constraints rather than equality constraints. Second, by adding a redundant identity matrix at the end of the original over-complete dictionary, the sparse representation is made more robust to variability and noise [8, 9]. Third, the difference between the UBM and the MAP adapted model is mapped into the GMM mean shifted supervector, which not only preserves the distance of the associated GMM but also makes the supervector sparse, and therefore helps to achieve a robust sparse representation [8]. However, sparse representation on large dimension supervectors not only requires a large training data set (the number of samples must be greater than the supervector dimension [9]) but also consumes a large amount of memory space due to the over-complete dictionary which can limit the training sample numbers and slow down the recognition process. Thus, in this work, we adopt the i-vectors in the total variability space due to its excellent discriminative capability and small dimensionality.

Specifically, we first construct an over-complete dictionary using the background i-vector samples, and then calculate the sparsest linear representation via l^1-minimization for each test i-vector sample. The membership of the sparse representation in the over-complete dictionary itself captures the discriminative information given sufficient training samples [9, 8]. If the trial is true, the test sample should have a sparse representation whose nonzero entries concentrate mostly on the target samples whereas the test sample from a false trial should have sparse coefficients spread widely among multiple speakers [9]. In our speaker verification task, the number of non-target background speakers are naturally considerably larger than the number of target speakers. Thus the chance nonzero entries on the target training samples for a test sample from a false trial should be arbitrarily small and close to zero. Based on the overwhelming unbalanced non-target negative training samples and the very limited target positive training samples, in contrast to the SVM system which tunes the SVM cost values each time, the proposed framework utilizes the highly unbalanced nature of the training samples to form a sparse representation problem.

In addition, we propose three methods to enhance the robustness and the performance of our speaker verification task.

First, the background normalized (BNorm) L2 residual is proposed as a score measuring criterion. Second, by directly using the Tnorm i-vectors as the non-target background samples in the over-complete dictionary, the result of the sparse representation system with Tnorm is efficiently achieved by only one sparse representation computation. Finally, the results of these i-vector modeling systems are fused to further improve the overall verification performance.

The paper organization is as follows: Section 2 describes the proposed methods, Section 3 provides the experimental results and Section 4 summarizes the conclusions.

2. Methods

In this section, we first introduce the total variability i-vectors in section 2.1 and then present the details of our proposed sparse representation modeling in section 2.2. Finally, the description of our proposed sparse representation system with Tnorm score normalization is provided in section 2.3.

2.1. Total variability i-vectors and baseline modeling

In the total variability space, there is no distinction between the speaker effects and the channel effects. Rather than using the eigenvoice matrix V and the eigenchannel matrix U [1], the total variability space contains the speaker and channel variabilities simultaneously [4]. Given an utterance, the speaker and channel dependent GMM mean supervector can be written as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{m}$ is the UBM mean supervector, $\boldsymbol{T}$ is a rectangular total variability matrix of low rank and $\boldsymbol{w}$ is the so-called i-vector [4]. Considering a C-components GMM and F dimensional acoustic features, the total variability matrix $\boldsymbol{T}$ is a $CF \times L$ matrix which can be estimated the same way as learning the eigenvoice matrix V in [10] except that here we consider every utterance is produced by a new speaker [4].

In this total variability space, two channel compensation methods, namely Within Class Covariance Normalization (WCCN) [5] and Linear Discriminant Analysis (LDA) are applied to reduce the variabilities. WCCN uses the inverse of the within-class covariance to normalize the cosine kernel while LDA attempts to transform the axes to minimize the intra-class variance due to the channel effects and maximize the variance between speakers. After WCCN and LDA steps, SVMs with cosine kernel or cosine distance scoring is used for i-vector modeling. The cosine kernel between two i-vectors $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ is defined as follows:

$$k(\boldsymbol{w_1}, \boldsymbol{w_2}) = \frac{<\boldsymbol{w_1}, \boldsymbol{w_2}>}{\|\boldsymbol{w_1}\|_2 \|\boldsymbol{w_2}\|_2} \tag{2}$$

These two systems serve as our baseline systems.

2.2. Sparse representation for modeling

Given N_1 ($N_1 = 1$ in our case because only one recording for each target speaker and one target speaker per trial) target training samples $\boldsymbol{A_1}$ and N_2 non-target background training samples $\boldsymbol{A_2}$, we construct the over-complete dictionary $\boldsymbol{A}$:

$$\boldsymbol{A} = [\boldsymbol{A_1}\boldsymbol{A_2}] = [\boldsymbol{s_{11}}, \boldsymbol{s_{12}}, \cdots, \boldsymbol{s_{1N_1}}, \boldsymbol{s_{21}}, \boldsymbol{s_{22}}, \cdots, \boldsymbol{s_{2N_2}}]. \tag{3}$$

Each sample $\boldsymbol{s_{ij}}$ is an L dimensional i-vector and is normalized to unit l^2 norm. This matches the length normalization in the SVM cosine kernel. Throughout the entire testing progress, the background samples $\boldsymbol{A_2}$ are fixed; and only the target samples $\boldsymbol{A_1}$ are replaced according to the claimed target identity in the test trial. Let us denote $N = N_1 + N_2$, then $N_1 \ll N_2$ and $L < N$ need to be satisfied for sparse representation. In our case, the dimensionality L of the i-vectors is significantly smaller than the number of training samples N. For any test sample $\boldsymbol{y} \in \mathbb{R}^L$ with unit l^2 norm, we want to use the over-complete dictionary $\boldsymbol{A}$ to linearly represent $\boldsymbol{y}$ in a sparse way. If $\boldsymbol{y}$ is from the target, then $\boldsymbol{y}$ will approximately lie in the linear span of training samples in $\boldsymbol{A_1}$ [9]. Since the equality constraint $\boldsymbol{Ax} = \boldsymbol{y}$ is not robust against large session variabilities [9], we constrain the Euclidian distance between the test sample and the linear combination of training samples to be smaller than ϵ which resulted in a standard convex optimization problem (l^1-minimization with quadratic constraints):

$$\text{Problem A :} \quad \min\|\boldsymbol{x}\|_1 \qquad \text{subject to} \|\boldsymbol{Ax} - \boldsymbol{y}\|_2 \le \epsilon \tag{4}$$

Since $N_1 = 1$ in our case, for each sample in the over-complete dictionary i, ($i = 1, \cdots, N$), let $\delta_i : \mathbb{R}^N \to \mathbb{R}^N$ be the characteristic function which selects the coefficient only associated with the i_{th} sample. For $\boldsymbol{x} \in \mathbb{R}^N$, $\delta_1(\boldsymbol{x}) \in \mathbb{R}^N$ is a new vector whose nonzero entries are the only entries in the first element of $\boldsymbol{x}$. Now based on the sparse representation $\boldsymbol{x}$, in addition to the l^1 norm ratio and l^2 residual ratio introduced in [8], we propose the new Background Normalized (BNorm) l^2 residual criterion for verification purposes. It uses the scores from the background data to perform a kind of Tnorm on the target score. Given a solved sparse representation, we can also consider every background sample as the target sample and calculate its minus l^2 residual as a similarity score. Without any additional sparse representation computation, just by rotating the role of each sample in this over-complete dictionary, we can instantly generate the similarity measure scores ($\boldsymbol{\phi}$) for all the samples.

$$\begin{aligned}
l^1\text{norm ratio} &= \|\delta_1(\boldsymbol{x})\|_1 / \|\boldsymbol{x}\|_1 & (5)\\
l^2\text{residual ratio} &= \frac{\|\boldsymbol{y} - \boldsymbol{A}(\Sigma_{i=2}^N \delta_i(\boldsymbol{x}))\|_2}{\|\boldsymbol{y} - \boldsymbol{A}\delta_1(\boldsymbol{x})\|_2} & (6)\\
\text{Bnorm } l^2\text{residual} &= \frac{-\|\boldsymbol{y} - \boldsymbol{A}\delta_1(\boldsymbol{x})\|_2 - \text{mean}(\boldsymbol{\phi})}{\text{std}(\boldsymbol{\phi})} & \\
\boldsymbol{\phi}_{j,j=2:N} &= -\|\boldsymbol{y} - \boldsymbol{A}\delta_j(\boldsymbol{x})\|_2 & (7)
\end{aligned}$$

A larger score represents a higher likelihood for the testing sample being from the target subject.

Due to large session variabilities, the test sample $\boldsymbol{y}$ can be partially corrupted. Thus an error vector $\boldsymbol{e}$ is introduced to explain the variability [9]:

$$\boldsymbol{y} = \boldsymbol{y_0} + \boldsymbol{e} = \boldsymbol{A}\boldsymbol{x_0} + \boldsymbol{e} \tag{8}$$

So the original optimization problem takes the following form:

$$\text{Problem B :} \quad \min\|\boldsymbol{z}\|_1 \qquad \text{subject to} \|\boldsymbol{Bz} - \boldsymbol{y}\|_2 \le \epsilon \tag{9}$$

$$\boldsymbol{B} = [\boldsymbol{A} \quad \boldsymbol{I}] \in \mathbb{R}^{L\times(N+L)}, \boldsymbol{z} = [\boldsymbol{x}^t \quad \boldsymbol{e}^t]^t \in \mathbb{R}^{(N+L)} \tag{10}$$

If the error vector $\boldsymbol{e}$ is sparse and has no more than $(L + N_1)/2$ nonzero entries, the new sparse solution $\boldsymbol{z}$ is the true generator according to (8) [9]. Finally, we redefine the three decision criteria based on the new sparse solution $\hat{\boldsymbol{z}} = [\hat{\boldsymbol{x}}^t \quad \hat{\boldsymbol{e}}^t]^t$.

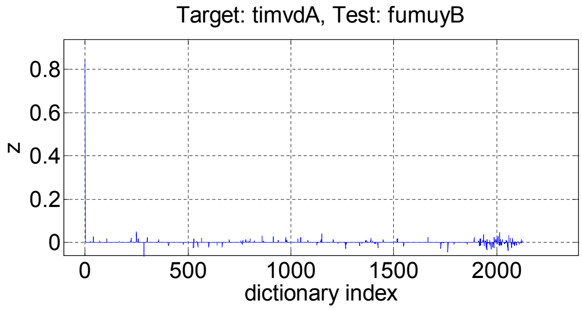

Figure 1: The sparse solution of a true trial with problem B (9)

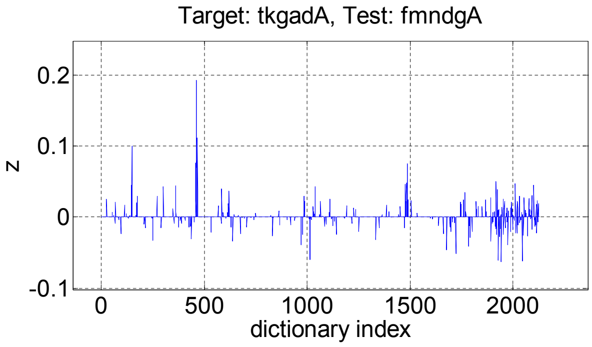

Figure 2: The sparse solution of a false trial with problem B (9)

Table 1: The two configurations (S1 and S2) and the corresponding complexity of sparse representation with Tnorm.

	Trail Score		Tnorm Score (N_3 samples)		SR Times
	$\boldsymbol{A}_1$	$\boldsymbol{A}_2$	$\boldsymbol{A}_1$	$\boldsymbol{A}_2$	
S1	Target	Background	i^{th} Tnorm	Background	1+N_3
S2	Target	Tnorm	None	None	1

$$
\begin{aligned}
l^1\text{norm ratio} &= \|\delta_1(\hat{\boldsymbol{x}})\|_1/\|\hat{\boldsymbol{x}}\|_1 && (11)\\
l^2\text{residual ratio} &= \frac{\|\boldsymbol{y}-\hat{\boldsymbol{e}}-\boldsymbol{A}(\Sigma_{i=2}^{N}\delta_i(\hat{\boldsymbol{x}}))\|_2}{\|\boldsymbol{y}-\hat{\boldsymbol{e}}-\boldsymbol{A}\delta_1(\hat{\boldsymbol{x}})\|_2} && (12)\\
\text{Bnorm}l^2\text{residual} &= \frac{-\|\boldsymbol{y}-\hat{\boldsymbol{e}}-\boldsymbol{A}\delta_1(\hat{\boldsymbol{x}})\|_2-\text{mean}(\boldsymbol{\phi})}{\text{std}(\boldsymbol{\phi})}\\
\boldsymbol{\phi}_{j,j=2:N} &= -\|\boldsymbol{y}-\hat{\boldsymbol{e}}-\boldsymbol{A}\delta_j(\hat{\boldsymbol{x}})\|_2 && (13)
\end{aligned}
$$

Fig.1 and 2 demonstrate the sparse solutions of two trials in the evaluation using problem B (9) before Tnorm.

2.3. Sparse representation with Tnorm

Test normalization (Tnorm) is an important technique for normalizing the variance of the testing score based on a set of cohort models and hence is widely adopted in verification tasks. Calculating the similarity scores between the testing sample and all the cohort models can be computationally expensive for the sparse representation system. Thus, as shown in Table 1, we propose a new setup for the sparse representation system to efficiently perform Tnorm score normalization. Compared to the straightforward configuration S1, the new setting S2 only requires a single sparse representation calculation which reduces the computational complexity significantly. In the configuration S2, we directly employ the Tnorm data as the non-target background samples in the over-complete dictionary and use the score distribution of the Tnorm data to normalize the target sample's score using eq (13). Note that in problem B setting, the number of samples in the over-complete dictionary $(L+1+N_3)$ is always bigger than the i-vector dimensionality L. Therefore, the condition of sparse representation is still satisfied.

Table 2: Corpora used to estimate the UBM, total variability matrix (T), WCCN, LDA ,SVM imposter and the Tnorm data.

	Switchboard	NIST04	NIST05	NIST06
UBM		$\surd$		
T	$\surd$	$\surd$	$\surd$	$\surd$
WCCN		$\surd$	$\surd$	$\surd$
LDA		$\surd$	$\surd$	$\surd$
SVM-Imposter		$\surd$	$\surd$	
Tnorm				$\surd$

3. Experimental results

3.1. Corpus and i-vector generation

We performed experiments on the NIST 2008 speaker recognition evaluation (SRE) corpus [11]. Our focus is the single-side 1 conversation train, single-side 1 conversation test, and the multi-language handheld telephone task, which is one part of the core test condition. This setup resulted in 3832 true trials and 33218 false trials. We used equal error rate (EER) and the minimum decision cost value (minDCF) as the metrics for evaluation [11].

For cepstral feature extraction, a 25ms Hamming window with 10ms shifts was adopted. Each utterance was converted into a sequence of 36-dimensional feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. An energy-based speech detector was applied to discard low-energy frames. Feature warping is applied to mitigate channel effects.

The training data included Switchboard II part2 and part3, Switchboard Cellular, NIST SRE 2004, 2005, and 2006 corpora. The description of the dataset used in each step is provided in Table 2. The gender-dependent GMM UBMs consist of 1024 mixture components, which were trained using EM with the data from NIST SRE 04 corpus. The background data was the same as UBM. We used all of the training data for estimating the total variability space. The NIST SRE 2004, 2005 and 2006 datasets were used for training WCCN and the the LDA matrix, and a data set chosen from NIST SRE 2006 corpus was used for Tnorm score normalization, including 367 male utterances and 340 female utterances. The SVMLight toolkit [12] was used for SVM modeling.

3.2. Results and discussion

The performance of sparse representation using S1 configuration and problem B setting with different score measuring criteria is shown in Table 3 and Table 4. It is demonstrated in [8] that l^1 norm ratio is better than l^2 residual ratio for verification tasks. This matches with our experimental results here. Furthermore, the proposed BNorm l^2 residual criterion achieved the best performance among all three score measurement with 0.0226 and 0.0293 minDCF value after Tnorm for the male and female tasks, respectively. It is shown in both Table 5 and Table 6 that S1 configuration based sparse representation system performed better than the S2 configuration in terms of minDCF value. This might be because in S2 setting, each Tnorm target sample was not scored on test sample independently and the

Table 3: Performance of the sparse representation system on the Male part of the NIST 08 test with configuration S1.

System	Without Tnorm		With Tnorm	
	EER	minDCF	EER	minDCF
l^1 norm ratio	5.68%	0.0243	5.13%	0.0235
l^2 residual ratio	6.25%	0.0247	5.13%	0.0240
Bnorm -(l^2 residual)	**5.35%**	**0.0236**	**4.94%**	**0.0226**

Table 4: Performance of the sparse representation system on the Female part of the NIST 08 test with configuration S1.

System	Without Tnorm		With Tnorm	
	EER	minDCF	EER	minDCF
l^1 norm ratio	7.21%	0.0327	6.95%	0.0317
l^2 residual ratio	7.26%	0.0325	6.86%	0.0307
Bnorm -(l^2 residual)	**6.76%**	**0.0310**	**6.19%**	**0.0293**

Tnorm set is smaller than the background data set. However, the S2 setting is significantly more efficient. Moreover, by fusing the sparse representation systems with the SVM and CDS baselines, the S2 based system demonstrates superior performance over S1 setting in terms of EER for both male and female tasks.

In Table 5, we see a significant improvement was achieved by fusing the proposed S2 sparse representation system with either the SVM baseline or the CDS baseline in terms of minDCF value. The fusion system (ID 8) achieved the best result on male task with 4.05% EER and 0.0204 minDCF value. Similar results are demonstrated in Table 6 for female data task. The minDCF value of the cosine distance baseline system was improved from 0.0302 to 0.0272 by the fusion with the sparse representation system (fusion system ID 9). The overall system performance was improved to 5.25% EER and 0.0262 minDCF value by the fusion system ID 8.

It is shown in Table 5 and 6 that, after T-norm, sparse representation did not achieve superior performance compared to SVM or CDS baseline in terms of single system performance in this task. This might be because only one enrollment (positive) sample in the over-complete dictionary. Furthermore, we can see that the improvements of Tnorm score normalization on sparse representing systems are less significant than the SVM and CDS baselines. It might be due to the fact that score distribution being not gaussian (majority of the l^1 norm ratio scores concentrate on 0 value), suggesting that we need to investigate other distribution based score normalization. Future work also includes investigating the usage of sparse representation on the language identification task and the potential way to represent the speaker/language/channel information in the sparse manner.

4. Conclusions

A robust speaker verification approach using a sparse representation on the total variability i-vectors is proposed. The main contributions are as follows. First, we propose the background normalized l^2 residual as a score measuring criterion. Second, we demonstrate that the Tnorm can be efficiently achieved by using the Tnorm data as the non-target samples in the over-complete dictionary. Finally, by fusing with the conventional i-vector based SVM system and cosine similarity system, we show that the overall system performance is improved, and achieves state of the art results. Future work includes investigating the non-gaussian distribution based score normalization and the usage of sparse representation for the language identification task and information representation in the sparse manner.

Table 5: Performance on the Male part of the NIST 08 test

ID	System	Without Tnorm		With Tnorm	
		EER	minDCF	EER	minDCF
1	SVM-base	4.75%	0.0231	4.36%	0.0216
2	CDS-base	4.76%	0.0256	4.43%	0.0221
3	SR S1	5.35%	0.0236	4.94%	0.0226
4	SR S2			4.82%	0.0235
6	Fusion1+3			4.18%	0.0202
7	Fusion2+3			4.30%	0.0205
8	Fusion1+4			**4.05%**	**0.0204**
9	Fusion2+4			4.22%	0.0204

Table 6: Performance on the Female part of the NIST 08 test

ID	System	Without Tnorm		With Tnorm	
		EER	minDCF	EER	minDCF
1	SVM-base	5.86%	0.0278	5.52%	0.0268
2	CDS-base	6.87%	0.0326	5.93%	0.0302
3	SR S1	6.76%	0.0310	6.19%	0.0293
4	SR S2			6.40%	0.0314
6	Fusion1+3			5.40%	0.0263
7	Fusion2+3			5.55%	0.0272
8	Fusion1+4			**5.25%**	**0.0262**
9	Fusion2+4			5.53%	0.0272

5. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[2] P. Kenny, G. Boulianne, P. Dumouchel, and P. Ouellet, “Speaker and Session Variability in GMM-Based Speaker Verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

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

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, no. 99, pp. 1–1, 2010.

[5] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” in *Proc. Interspeech*, vol. 4, no. 2.2, 2006.

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

[7] I. Naseem, R. Togneri, and M. Bennamoun, “Sparse Representation for Speaker Identification,” in *Proc. ICPR*, 2010, p. 4460.

[8] 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*, 2011, paper available at http://www-scf.usc.edu/ mingli/publication.htm.

[9] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal Mach Intell*, vol. 31, no. 2, pp. 210–227, 2008.

[10] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[11] “The NIST Year 2008 Speaker Recognition Evaluation Plan,” *http://www.nist.gov/speech/tests/spk/2008/index.html*, 2008.

[12] T. Joachims, “SVMLight: Support Vector Machine,” *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 1999.