SPEAKER VERIFICATION USING SPARSE REPRESENTATION CLASSIFICATION

Jia Min Karen Kua[1,2], *Eliathamby Ambikairajah*[1,2], *Julien Epps*[1,2], *Roberto Togneri*[3]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia
[3] School of Electrical, Electronic and Computer Engineering
The University of Western Australia, Western Australia, WA 6009, Australia
jmkua@student.unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au, Roberto.Togneri@uwa.edu.au

ABSTRACT

Sparse representations of signals have received a great deal of attention in recent years, and the sparse representation classifier has very lately appeared in a speaker recognition system. This approach represents the (sparse) GMM mean supervector of an unknown speaker as a linear combination of an over-complete dictionary of GMM supervectors of many speaker models, and ℓ_1-norm minimization results in a non-zero coefficient corresponding to the unknown speaker class index. Here this approach is tested on large databases, introducing channel-/session-variability compensation, and fused with a contemporary GMM-SVM system. Evaluations on the NIST 2006 SRE database show that when the outputs of the MFCC GMM-SVM-NAP based classifier are fused with the MFCC GMM-Sparse Representation Classifier-NAP (GMM-SRC-NAP) based classifier, a baseline EER of 6.56% can be reduced to 2.65%, significantly improving the performance of the speaker verification system.

Index Terms— sparse representation, compressive sensing, speaker verification

1. INTRODUCTION

An automatic speaker recognition system normally comprises three stages: feature extraction, speaker modelling and score computation. In current speaker recognition systems, speech is pooled from many speakers to train a single model, known as a universal background model (UBM). Individual speakers are then adapted from the UBM using the maximum a posterior (MAP) adaptation algorithm (Fig 1). Once the adaptation is complete, the mean *(μ_{spk})* will be different for varying speakers compared with the mean (μ_{UBM}) of the UBM. Normally the covariance and weight parameters Σ and ω are not adapted [1].

Campbell et al. [2] introduced the GMM based supervector concept for a Support Vector Machine (SVM) based classifier for a speaker verification task. The SVM tries to discriminate the target speaker from a set of impostor speakers by mapping the feature space (cepstral space) to a high dimensional SVM space (supervector space) to form a boundary between the two classes. The supervectors of target or impostor speakers are formed by concatenating the means of their corresponding GMMs (Fig 2). The GMM supervector for a particular speaker may be interpreted as a mapping between an utterance and a high dimensional vector. In addition, this transformation facilitates any length utterance to be represented by a fixed length supervector.

Compressive sensing/Sparse Representation is a recent development in digital signal processing. This technique samples a signal and simultaneously compresses it at a greatly reduced rate than the standard Shannon-Nyquist rate. The compressed signal can be reconstructed using an optimization process which employs non-adaptive linear projections that preserve the structure of the signal. In compressive sensing, the familiar least squares optimization is inadequate for signal reconstruction, and other types of convex optimization are used [3]. The sparse representation paradigm, when it was originally developed, was not intended for classification purposes; however sparse representation based classifiers are beginning to emerge for various applications [4-7].

Naseem et al. [8] were the first to introduce a classifier based on compressive sensing/sparse representation for speaker identification. Their experiments were conducted using the TIMIT database and they found that speaker identification using a sparse representation classifier showed good performance compared with GMM-SVM speaker identification algorithms.

In this paper, we propose the use of the sparse representation classifier (SRC) for a speaker verification task. In Section 2, we describe the sparse representation formulation. Section 3 discusses the sparse representation feature. Section 4 presents the experiments performed, followed by a discussion on the results obtained using the NIST 2001 and NIST 2006 speaker recognition databases.

2. SPARSE REPRESENTATION CLASSIFIER (SRC)

For a speaker identification task, [8] proposed the use of a GMM mean supervector to develop an over complete dictionary using training utterances from all the speakers. Following this, the GMM mean supervector of a test

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

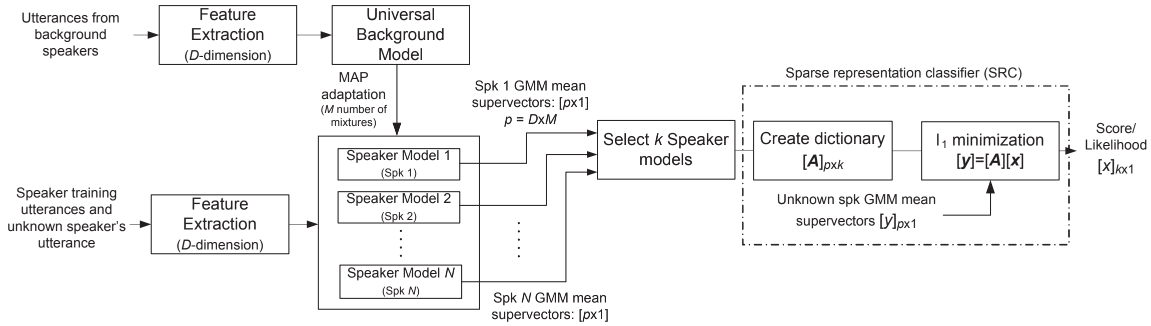
Utterances from background speakers
Feature Extraction (D-dimension)
Universal Background Model
MAP adaptation (M number of mixtures)
Spk 1 GMM mean supervectors: [px1] p = DxM
Speaker training utterances and unknown speaker's utterance
Feature Extraction (D-dimension)
Speaker Model 1 (Spk 1)
Speaker Model 2 (Spk 2)
Speaker Model N (Spk N)
Spk N GMM mean supervectors: [px1]
Select k Speaker models
Sparse representation classifier (SRC)
Create dictionary [A]pxk
l1 minimization [y]=[A][x]
Unknown spk GMM mean supervectors [y]px1
Score/ Likelihood [x]kx1

Figure 1 *A Sparse Representation Classifier (SRC) based speaker recognition system*

utterance from an unknown speaker is represented as a linear combination of this over-complete dictionary.

This representation is sparse since the test utterance corresponds to only a small fraction of the whole training database. As a result, the unknown vector of coefficients, obtained efficiently via ℓ_1- minimization, will have non-zero entries (ideally) corresponding to the class of the test utterance. Finding the sparsest representation implies that the various classes present in the over complete dictionary are automatically discriminated. In this paper we exploit this discriminative nature of sparse representation for a speaker verification task.

If we have *N* distinct speakers, then we have *N* supervectors (after MAP adaptation), each with a column vector of *D x M* rows; where *D* is the feature dimension and *M* is the number of Gaussian mixtures. According to Figure 1, we select *k* speaker models for a speaker verification task and a global dictionary matrix ***A*** can be constructed by concatenating all the supervectors of the *k* speaker models, thus creating a matrix of *D x M* rows and *k* columns. The speaker selection is random but includes one supervector from the claimed speaker. The test speaker supervector ***y*** can be represented as a linear combination of all supervectors of *k* speaker models. i.e $\boldsymbol{y} = \boldsymbol{Ax}$ (Fig 1):

$$\begin{array}{cccc} \text{Unknown speaker supervector} & & k\text{-1 imposters \& 1 true speaker supervectors} & k \text{ coefficients} \\ \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_p \end{bmatrix} & = & \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & a_{1k} \\ a_{21} & a_{22} & \cdot & \cdot & a_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & a_{p2} & \cdot & \cdot & a_{pk} \end{bmatrix} & \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \alpha_k \end{bmatrix} \end{array} \quad (1)$$

where $p = D\,x\,M$ and ***x*** is a *k*-dimensional vector coefficients that will be sparsely represented in ***A***. When equation (1) is solved, ideally all coefficients of α are zero, except one non-zero coefficient: its index corresponds to the class of the given unknown utterance ***y***.

We need to solve the system of linear equations given in (1) in order to obtain the sparse vector ***x***. Since the dimensionality (*p*) of the supervectors is much larger than the number of speakers (*k*), Equation (1) is over-determined and has no unique solution. Recent research in compressive sensing [3, 9] has shown that if ***x*** is sparse, it can be recovered with high probability by solving Equation (1) using ℓ_1-norm minimization as given in (2):

$$\hat{x}_1 = \arg\min \|x\|_1 \qquad subject\ to\ y = Ax \qquad (2)$$

Once we have estimated $\hat{x}_1$, due to modelling limitations, $\hat{x}_1$ would have small non-zero entries belonging to impostors in ***A***. Therefore, we should assign ***y*** to the class which has the largest support in $\boldsymbol{\alpha}$. i.e search for the largest value in the sparse vector ***x*** and locate the index which corresponds to the class of the unknown utterance ***y***.

In order to demonstrate Equation (1) using ℓ_1-norm minimization, we created a matrix ***A*** using a small number of 3-dimensional data, where the columns of ***A*** represent 5 impostors and a true speaker. These are represented by C1 to C6 in Figure 2. A test vector (representing the unknown speaker) was chosen such that it belongs to class C4. Solving Equation (2) produces the vector $\boldsymbol{x} \approx [0, 0, -0.2499, 0.8408, 0, 0.2136]^T$ with the largest ℓ_1-norm of 0.8408. This corresponds to class C4 and is identified as the correct class. This value is kept as the true speaker score and the remaining ℓ_1-norm values are denoted as impostor scores. In speaker verification experiments, all true speaker and impostor scores are collated to generate a Detection Error Tradeoff (DET) curve.

3. FEATURE EXTRACTION

In this paper, we use the conventional Mel Frequency Cepstral Coefficients (MFCC) and Spectral Centroid Frequency (SCF) in the feature extraction block of Figure 1 to evaluate the consistency of the new SRC approach across more than one feature type. The front-end of the recognition/verification system includes an energy-based speech detector which is applied to discard silence and noise frames. We used 20 ms Hamming-windowed frames, overlapped by 10 ms. The magnitude based MFCCs were extracted using a 26 Mel-scaled triangular filterbank and the frequency based SCF features were extracted as outlined below. Each feature vector consisted of 14 dimensions and the deltas were appended, following feature warping, providing a 28 dimensional feature vector.

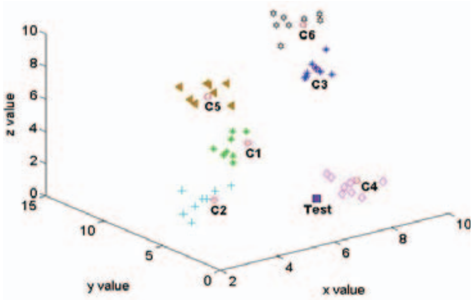

Figure 2: ***Plot of entries in matrix A [[5 5 5]^T, [3 3 3]^T, [7 4 9]^T, [8 3 2]^T, [4 6 8]^T, [9 10 9]^T] and the test vector y [7 4 1]^T***

In our previous work, the effectiveness of SCF features for speaker recognition was investigated [10]. We briefly outline the spectral centroid feature extraction process. Let $s[n]$ represent a frame of speech of N samples and $|S[f]|$ represent the magnitude spectrum of this frame. $|S[f]|$ is then divided into K sub-bands, where each sub-band is defined by a lower frequency edge (l_k) and an upper frequency edge (u_k).

We define *spectral moment* in a frequency sub-band as the spectral magnitude multiplied by its frequency. When the resultant spectral moment in a sub-band (numerator term in Equation (3)) is divided by the sum of the magnitude spectrum in that band, we obtain the spectral centroid frequency, f_{sc}:

$$f_{sc} = \sum_{f=l_k}^{f=u_k} f\,|S[f]| \Bigg/ \sum_{f=l_k}^{f=u_k} |S[f]| \qquad (3)$$

The spectral centroid frequency measure captures the center of gravity of each sub-band, detecting the approximate location of formants (in sub-bands containing multiple harmonics), but is also affected by changes in pitch and harmonic structure of the vocal source (particularly in sub-bands containing few harmonics). In our experiments, SCF are extracted with a 14 mel-scale Gabor filter-bank. We also increased the number of FFT points by an order of magnitude (from 160 to 2048) to better approximate the speech power spectrum, which was found to have a significant effect on the SCF performance [10].

4. SPEAKER VERIFICATION EXPERIMENTS

4.1. Database

Speaker verification experiments were conducted using the NIST 2001 SRE database, which is a convenient size for initial validation without SVM-NAP. The NIST 2001 SRE development database consists of 38 male speakers and 22 female speakers. The evaluation database comprises 74 male speakers and 100 female speakers for training, 850 male speakers and 1188 female speakers for testing. The training utterance for each speaker was 2 minutes and the testing segment duration was less than 60 seconds.

The final evaluation data is the core test condition of the NIST 2006 SRE (1conv4w-1conv4w), containing 51 448 trials of which 3612 are true and 47 836 are false. The background data consists of 3079 speech utterances from the NIST 2004 SRE, which cover a number of speakers (female and male). The Nuisance Attribute Projection (NAP) [11] training data includes approximately 10000 speech utterances from the NIST 2004 and 2005 SRE corpus.

4.2. Speaker verification systems

The baseline back-end of the verification system for the NIST 2001 SRE database was based on Universal Background Model-Gaussian Mixture Models (UBM-GMMs) for simplicity. The baseline back-end of the NIST 2006 SRE database was a GMM-SVM-NAP classifier (NAP rank 40). We also created an SRC back-end using linear programming to implement the sparse representation classifier as per Section 2, with k = 11 and k = 15-16 for 2001 and 2006 respectively, and applied NAP.

4.3. Results and discussion

Results for the speaker recognition experiments based on the NIST 2001 and NIST 2006 SRE database with different features and classifiers are given in Table 1. It can be observed that SRCs are able to achieve almost comparable performance to the UBM-GMM back-end on the NIST 2001 database. On NIST 2001, the fusion of two systems with a different classifier improved on an 8.28% EER baseline to 7.01% for MFCC and for SCF, from 8.83% to 7.43% as shown in Table 2. Furthermore, the fusion of three subsystems (not shown in the table), MFCC (UBM-GMM), MFCC (GMM-SRC) and SCF (GMM-SRC) reduced the EER to 6.32% which demonstrates the complementary nature of the classifiers and features.

Interestingly, SCF which was introduced as a complementary feature to MFCC in [10] outperforms MFCC on NIST 2006 database for both GMM-SVM and GMM-SRC back-end (without NAP compensation). In addition, by incorporating NAP compensation to SRC, we could improve the EER from 15.06% to 11.19% and from 14.18% to 12.82% for MFCC and SCF respectively. Although GMM-SVM-NAP outperforms GMM-SRC-NAP for both features on NIST 2006 (Table 1), we observed from the fused results, given in Table 2, that the fusion of systems with different classifiers significantly improves on the individual subsystem performance. The MFCC features with the standard back-end, produced the lowest single-feature system EER of 6.56%, while the sparse representation classifier with NAP produced an EER of 11.19%. However, the fusion of these two systems improved on a 6.56% EER baseline to 2.65% as shown in Figure 3. This result provides strong encouragement that sparse representation classifier (GMM-SRC-NAP) carries complementary information to the standard GMM-SVM-NAP classifier. Furthermore, the fusion of SCF-based GMM-SVM-NAP (EER 7.67%) and SCF-based GMM-SRC-NAP (EER 12.82%) showed an improvement and the

EER was reduced to 3.51%. as shown in Figure 4. Similar improvements after fusion were also observed for this database without NAP. Subsequently it was observed that the *A* matrix composition shouldn't be drawn from the same database. Therefore we repeated the experiments on MFCC GMM-SRC-NAP with *A* matrix composed using NIST 2005 SRE database (372 Female and 274 Male), with an EER of 6.66%. Similarly the fusion of MFCC GMM-SVM-NAP and MFCC GMM-SRC-NAP improved on a 6.56% EER baseline to 1.83%.

6. CONCLUSION

In this paper, we investigated the discriminative nature of a sparse representation classifier (SRC) for a speaker verification task. Sub-systems were developed using MFCCs and SCFs as features and various back ends were used in the system. We demonstrated that SRC classifiers benefit from NAP compensation as expected and provide good performance particularly when two sub-systems are fused. Evaluation on the NIST 2006 database using a fusion of MFCC GMM-SVM-NAP and MFCC GMM-SRC-NAP subsystems, demonstrated relative improvements of 60% over the performance of MFCC GMM-SVM-NAP only. Similarly when SCF was used as features instead of MFCC, we find a relative improvement of 77%. This strongly supports the hypothesis that these two classifiers carry complementary information.

7. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, pp. 19-41.

[2] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters,* vol. 13, pp. 308-11, 2006.

[3] E. J. Candès, "Compressive sampling," in *Proc. Int'l Congress of Mathematicians*, 2006.

[4] K. Huang and S. Aviyente, "Sparse representation for signal classification," *Advances in Neural Information Processing Systems,* vol. 19, p. 609, 2007.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 210-227, 2008.

[6] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. of ICASSP*, 2010, pp. 4370-4373.

[7] S. F. Cotter, "Sparse Representation for accurate classification of corrupted and occluded facial expressions," in *Proc of ICASSP*, 2010, pp. 838-841.

[8] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse Representation for Speaker Identification," in *Proc. of ICPR*, 2010, pp. 4460-4463.

[9] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory,* vol. 52, pp. 1289-1306, 2006.

[10] J. M. K. Kua, T. Tharmarajah, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010, pp. 34-39.

[11] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. of ICASSP*, Philadelphia, PA, USA, 2005, pp. 629-32.

Table 1 ***The speaker verification results on the NIST 2001 SRE and NIST 2006 database***

FEATURES	NIST 2001		NIST 2006			
	UBM-GMM	SRC	GMM-SVM	GMM-SVM-NAP	SRC	SRC - NAP
MFCC	8.28	8.65	12.90	6.56	15.06	11.19
SCF	8.83	9.18	10.93	7.67	14.18	12.82

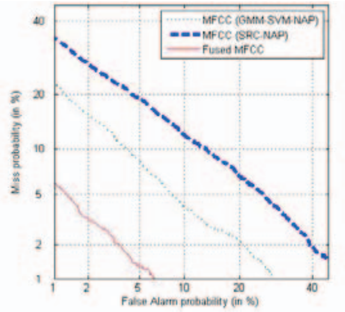

Figure 3 ***DET curve showing the MFCC GMM-SVM-NAP, MFCC SRC-NAP and fused speaker recognition system, tested on NIST 2006***

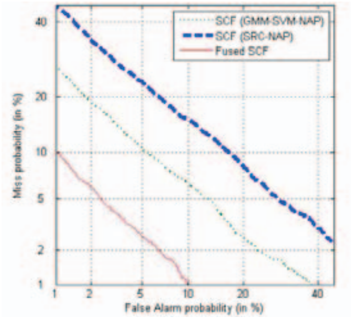

Figure 4 ***DET curve showing the SCF GMM-SVM-NAP, SCF SRC-NAP and fused speaker recognition system, tested on NIST 2006***

Table 2 ***Fused EER for speaker verification on NIST 2001 and NIST 2006 SRE database***

FEATURES (BACKEND)	NIST 2001			FEATURES (BACKEND)	NIST 2006		
	MFCC (SRC)	SCF (UBM-GMM)	SCF (SRC)		MFCC (SRC-NAP)	SCF (GMM-SVM-NAP)	SCF (SRC-NAP)
MFCC (UBM-GMM)	7.01	7.64	6.60	MFCC (GMM-SVM-NAP)	**2.65**	5.96	**2.97**
MFCC (SRC)	-	6.37	7.89	MFCC (SRC-NAP)	-	3.27	10.99
SCF (UBM-GMM)	-	-	7.43	SCF (GMM-SVM-NAP)	-	-	3.51