Sparse Representation for Speaker Identification

Imran Naseem , Roberto Togneri, Mohammed Bennamoun
The University of Western Australia
imran.naseem@ee.uwa.edu.au, roberto@ee.uwa.edu.au, bennamou@csse.uwa.edu.au

Abstract

We address the closed-set problem of speaker identification by presenting a novel sparse representation classification algorithm. We propose to develop an overcomplete dictionary using the GMM mean supervector kernel for all the training utterances. A given test utterance corresponds to only a small fraction of the whole training database. We therefore propose to represent a given test utterance as a linear combination of all the training utterances, thereby generating a naturally sparse representation. Using this sparsity, the unknown vector of coefficients is computed via l_1-minimization which is also the sparsest solution [12]. Ideally, the vector of coefficients so obtained has nonzero entries representing the class index of the given test utterance. Experiments have been conducted on the standard TIMIT [14] database and a comparison with the state-of-art speaker identification algorithms yields a favorable performance index for the proposed algorithm.

1. Introduction

With increasing security concerns, automatic person identification has emerged as an active research area over last two decades. Human speech is a natural way of recognizing a person and therefore Automatic Speaker Recognition (ASR) systems have been widely deployed for secured authentication. Conventional speaker recognition algorithms make use of acoustic features to develop probabilistic speaker models and utilize an adequate statistical distance metric for the classification purpose. Gaussian Mixture Models (GMM) have been typically used to develop probabilistic model for each speaker in a given database [17]. The large scale acceptance of the GMMs as the standard in the ASR can be credited to a number of factors such as the high accuracy, the ability to scale training algorithms for large data sets and the probabilistic framework. A speech signal is naturally characterized by continuous changes in the spectral domain, consequently a number of Gaussian components (typically of the order of 64) are necessary to model the speaker-dependent features over the length of an utterance. The collection of these Gaussian components results in the complete Gaussian mixture model.

A more efficient approach is to develop a Universal Background Model (UBM) using utterances from a set of speakers and adapt this universal model with respect to a particular speaker using *Maximum-A-Posteriori* (MAP) adaptation [18]. This state-of-art approach is commonly referred to as the GMM-UBM. There are several benefits in using this approach that have accounted for significant performance improvements in the GMM-based classification. For instance, when training data is not available for the adaptation of components in the UBM, the speaker values revert to those in the UBM to provide a more robust speaker model. In contrast, when ample training data is available for a given GMM component, the values approach those of the ML estimate.

Recently an intriguing variation in the GMM-UBM approach has enabled representation of a speaker as a point in a high dimensional space i.e. the *speaker space* [4]. The main idea is to concatenate the means of the GMM components to form a so-called GMM mean supervector [4]. In this way, a variable-length utterance can be represented as a fixed-length feature vector in the feature space and therefore the problem of speaker recognition can be tackled as a general problem of pattern recognition. One technique that has received significant focus in pattern recognition literature is the Support Vector Machine (SVM). The discriminative nature of the SVM has been successfully applied to a variety of pattern recognition tasks. An SVM is basically a two-class classifier that fits a separating hyperplane between the two classes (assuming linear separability). In recent years, the SVM-based classification has become a major focus in the task of speaker identification and verification [4].

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1083

IEEE computer society

Typically, in pattern recognition problems, it is believed that high-dimensional data vectors are redundant measurements of an underlying source. The objective of manifold learning is therefore to uncover this "underlying source" by a suitable transformation of high-dimensional measurements to low-dimensional data vectors. The main objective is to find a basis function for this transformation, which could distinguishably represent patterns in the feature space. A number of approaches have been reported in the literature for dimensionality reduction. These approaches have been broadly classified in two categories namely *generative/reconstructive* and *discriminative* methods. Reconstructive approaches (such as Principal Component Analysis or the PCA [19]), are reported to be robust for noisy data, these methods essentially exploit the redundancy in the original data to produce representations with sufficient reconstruction property. Formally, given an input x and label y, the generative classifiers learn a model of the joint probability $p(x, y)$ and classify using $p(y|x)$, which is determined using the Bayes' rule. The discriminative approaches (such as Linear Discriminant Analysis or the LDA [3]), on the other hand, are known to yield better results in "clean" conditions [13] owing to the flexible decision boundaries. The optimal decision boundaries are determined using the posterior $p(y|x)$ directly from the data and are consequently more sensitive to outliers. In the speaker recognition community there is a growing interest for the exploration of these manifold learning methods. The PCA for instance, has shown some good results in this regard and is usually referred to as *Eignevoice* approach [15], [16]. Working on the same lines, the concept of *Fishervoice*, based on the LDA approach, has recently been proposed to address the problem of semi-supervised speaker clustering [10].

In this research we present a novel speaker identification algorithm in the context of sparse representation [1]. We propose to utilize the concept of the GMM mean supervector to develop an *overcomplete dictionary* using training utterances from all the speakers. The fixed-length GMM mean supervector of a given test utterance from an unknown speaker is represented as a linear combination of this overcomplete dictionary. This representation is naturally sparse since the test utterance corresponds to only a small fraction of the whole training database. Using this sparsity, we proposed to solve the inverse problem using the l_1-norm minimization as it is shown to be the sparsest solution [12]. The vector of coefficients thus obtained will have non-zero entries corresponding to the class of the test utterance. The proposed algorithm is evaluated on a subset of the widely available TIMIT speech corpus [14]. Comparative analysis with the state-of-art speaker recognition algorithms yields a fairly comparable performance index for the proposed algorithm. To the best of our knowledge, it is for the first time that sparse representation classification is used for the problem of speaker identification.

The rest of the paper is organized as follows: Section 2 presents the proposed algorithm, followed by close-set speaker identification experiments in Section 3. The paper is concluded in Section 4.

2. Sparse Representation for Speaker Identification

Sparse or parsimonious representation of signals is regarded as a major research area in the paradigm of statistical signal processing. Most of the signals of practical interest are compressible in nature. For example, audio signals are compressible in localized Fourier domain and digital images are compressible in Discrete Cosine Transform (DCT) and wavelet domains. Recent research in the area of *compressive sampling* has shown that if the optimal representation of a signal is sufficiently sparse when linearly represented with respect to an overcomplete dictionary (also referred to as *measurement matrix*), it can be efficiently computed using convex optimization [5, 6, 7, 8, 11, 12].

The main objective of the *compressive sensing* theory is to achieve computational efficiency for information processing using the parsimonious representation of signals. From this perspective, the compressive sensing theory basically tries to avoid the Shannon-Nyquist bound by sampling at a much lower rate and still safely recovering the original information [5]. Although the compressive sensing paradigm is not intended for classification purpose, the sparse representation of a signal with respect to a basis remains implicitly discriminative in nature. It selects only those basis vector which most compactly represents the signal and reject the others [20].

We exploit this discriminative nature of sparse representation to propose a novel speaker identification algorithm. The proposed algorithm incorporates the GMM mean supervector kernel approach [4] to represent the utterances as feature vectors of a fixed dimension.

We now present the basic framework of the proposed speaker identification algorithm. Let us assume that we have k distinct classes and n_i utterances are available for training from the ith class. Each variable-length training utterance is mapped to a fixed-dimension feature vector using the GMM mean supervector kernel [4]. Let the resultant feature vector be designated as $\mathbf{v}_{i,j}$ such that $\mathbf{v}_{i,j} \in \mathbb{R}^m$. Here i is the index of the class, $i = 1, 2, \ldots, k$ and j is the index of the training utterance, $j = 1, 2, \ldots, n_i$. All this training data from the ith class is placed in a matrix A_i such that $A_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \ldots\ldots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$. Let $\mathbf{y} \in \mathbb{R}^m$ be the GMM mean supervector for a test utterance from the ith speaker. A fundamental concept in pattern recognition indicates that patterns from the same class lie on a linear subspace [2], therefore if $\mathbf{y}$ belongs to the ith class and the training samples from the ith class are sufficient, $\mathbf{y}$ will approximately lie in the linear span of the columns of A_i:

$$\mathbf{y} = \alpha_{i,1}\mathbf{v}_{i,1} + \alpha_{i,2}\mathbf{v}_{i,2} + \cdots + \alpha_{i,n_i}\mathbf{v}_{i,n_i} \qquad (1)$$

where $\alpha_{i,j}$ are real scalar quantities. Since identity i of the test sample $\mathbf{y}$ is unknown we develop a *global* dictionary matrix A for all k classes by concatenating $A_i, i = 1, 2, \ldots, k$ as follows:

$$\mathbf{A} = [A_1, A_2, \ldots, A_k] \; \in \mathbb{R}^{m \times n_i k} \qquad (2)$$

The test pattern $\mathbf{y}$ can now be represented as a linear combination of all n training samples ($n = n_i \times k$):

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad (3)$$

where

$$\mathbf{x} = [0, \cdots, 0, \alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n_i}, 0, \cdots, 0]^T \in \mathbb{R}^n \qquad (4)$$

is an unknown vector of coefficients. Note that from equation 3 and our earlier discussion it is straight forward to note that only those entries of $\mathbf{x}$ that are nonzero, correspond to the class of $\mathbf{y}$ [20]. This means that if we are able to solve equation 3 for $\mathbf{x}$ we can actually find the class of the test pattern $\mathbf{y}$. Recent research in compressive sensing and sparse representation [6, 7, 8, 11, 12] has shown that the sparsity of the solution of equation 3, enables us to solve the problem using the l_1-norm minimization:

$$(l^1): \quad \hat{\mathbf{x}}_1 = argmin \left\|\mathbf{x}\right\|_1 \quad ; \mathbf{A}\mathbf{x} = \mathbf{y} \qquad (5)$$

Once we have estimated $\hat{\mathbf{x}}_1$, ideally it should have nonzero entries corresponding to the class of $\mathbf{y}$ and now deciding the class of $\mathbf{y}$ is a simple matter of locating indices of the non-zero entries in $\hat{\mathbf{x}}_1$. However due to noise and modeling limitations $\hat{\mathbf{x}}_1$ is commonly corrupted by some small nonzero entries belonging to different classes. To resolve this problem we define an operator δ_i for each class i so that $\delta_i(\hat{\mathbf{x}}_1)$ gives us a vector $\in \mathbb{R}^n$ where the only nonzero entries are from the ith class. This process is repeated k times for each class. Now for a given class i we can approximate $\hat{\mathbf{y}}_i = \mathbf{A}\delta_i(\hat{\mathbf{x}}_1)$ and assign the test pattern to the class with the minimum residual between $\mathbf{y}$ and $\hat{\mathbf{y}}_i$.

$$\underbrace{\min}_{i} r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_i(\hat{\mathbf{x}}_1)\|_2 \qquad (6)$$

3 Experimental Evaluation

The TIMIT corpus is a collection of phonetically balanced sentences sampled at 16 kHz (8 kHz bandwidth), consisting of 10 utterances from 630 speakers across 8 dialect regions in the USA [14]. Extensive experiments were conducted on a randomly selected subset of the TIMIT database consisting of 114 speakers. For our experiments we used 8 utterances per speaker for training (5 SX and 3 SI sentences) while 2 utterances (2 SA sentences) constituted the testing set. Refer to [14], [17] for further details.

Approach	Recognition Accuracy
GMM	92.98%
GMM-UBM	96.93%
GMM-SVM	97.80%
Sparse Representation	**98.24%**

Table 1. Experimental Results for the TIMIT database

At the feature extraction stage, GMM mean supervector approach [4] (consisting of 64 mixtures) is used to generate fixed-length feature vectors from variable length utterances. In all experiments a pre-emphasis filter with coefficient 0.97 was applied to the sampled waveform and features were extracted from each 25ms frame and generated every 10ms, all frames were windowed using the Hamming window function. Comparative analysis is performed using three state-of-the-art approaches i.e. the GMM [17], the GMM-UBM [18] and the GMM-SVM [4] speaker identification algorithms. For the implementation of the GMM and GMM-UBM systems the Hidden Markov Model ToolKit (HTK version 3.4.1) [21], was configured to model a single-state HMM with the standard MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum-A-Posteriori) adaptation scripts to adapt the UBM accordingly for the GMM-UBM models and GMM-SVM supervectors. For the GMM-SVM the SVM-KM toolbox [9] was used to implement the one-against-all SVM classifier.

Results are shown in Table 1. The proposed sparse representation identification algorithm achieves 98.24% recognition accuracy which is better than all the contesting approaches. The conventional GMM [17] approach for instance, attains 92.98% recognition which lags 5.26% as compared to the proposed approach. The state-of-art GMM UBM [18] approach yields a comparable identification success of 96.93%. Recently proposed GMM-SVM system [4] also attains a good performance with 97.80% recognition.

4 Conclusion

With the recent development in the paradigm of speaker recognition, variable-length utterances can be represented as fixed length features in a high dimensional feature space. The task of speaker identification can now therefore be viewed as a traditional pattern classification problem. Motivated with these studies, we propose a novel speaker identification algorithm based on sparse representation. Noting that a given test utterance from a particular speaker corresponds to only a fraction of the whole training database, we proposed to develop an *overcomplete dictionary* of all training utterances. A given test utterance is thus represented as a linear combination of all training utterances giving rise to a naturally sparse representation. The inverse problem is solved using the l_1-minimization (as it is the sparsest solution). Consequently the vector of coefficients is also sparse with non-zero entries corresponding to the class of the unknown speaker. The proposed algorithm is evaluated on the standard TIMIT database and comparative analysis is performed with the state-of-art speaker identification approaches. The proposed sparse representation classification algorithm has shown good performance index and is favorably comparable with all approaches.

Although the initial investigations for the proposed algorithm are quite good, the TIMIT database however characterizes ideal acquisition environment and does not depict key robustness issues (e.g. reverberant noise and session variability). Good performance index under clean conditions is encouraging enough to extend the proposed approach for robust speaker recognition addressing more challenging databases.

References

[1] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24, 2007.

[2] R. Barsi and D. Jacobs. Lambertian reflection and linear subspaces. *IEEE Trans. PAMI*, 25(3):218–233, 2003.

[3] V. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Tran. PAMI*, 17(7):711–720, July 1997.

[4] W. Campbell, D. Sturim, and D. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

[5] E. Candès. Compressive sampling. In *International Congress of Mathematicians*, 2006.

[6] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[7] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. on Pure and Applied Math*, 59(8):1207–1223, 2006.

[8] E. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Tran. Infm. Theory*, 52(12):5406–5425, 2006.

[9] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. SVM and kernel methods Matlab toolbox. *Perception Systemès et Information, INSA de Rouen, Rouen, France*, 2005.

[10] S. M. Chu, H. Tang, and T. S. Huang. Fishervoice and semi-supervised speaker clustering. *ICASSP*, pages 4089–4092, 2009.

[11] D. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, April 2006.

[12] D. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. on Pure and Applied Math*, 59(6):797–829, 2006.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2000.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. Darpa Timit: Acoustic-phonetic continuous speech corpus CD-ROM. LDC catalog number LDC93S1, 1993.

[15] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in Eigenvoice space. *IEEE Trans. on Speech and Audio Processing*, 8(6):695–706, Nov 2000.

[16] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. *ICSLP*, pages 1771–1774, 1998.

[17] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, August 1995.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3), 2000.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neurosicence*, 3(1):71–86, 1991.

[20] J. Wright, A. Yang, A. Ganesh, S. Sastri, S, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. PAMI*, 2008.

[21] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. Hidden Markov model toolkit (HTK) version 3.4 user guide. 2002.