Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 68 (2015) 11–22

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Low-complexity speaker verification with decimated supervector representations

B.C. Haris [*], Rohit Sinha

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

Received 16 June 2014; received in revised form 12 December 2014; accepted 17 December 2014
Available online 26 December 2014

Abstract

This work explores the use of a few low-complexity data-independent projections for reducing the dimensionality of GMM supervectors in context of speaker verification (SV). The projections derived using sparse random matrix and decimation are explored and are used as speaker representations. The reported study is done on the NIST 2012 SRE task using a state-of-the-art PLDA based SV system. Interestingly, the systems incorporating the proposed projections result in performances competitive to that of the commonly used i-vector representation based one. Both the sparse random matrix and the decimation based approaches are attributed to have very low computational requirements in finding the speaker representations. A novel SV system that exploits the diversity among the representations obtained by using different offsets in the decimation of supervector, is also proposed. The resulted system is found to achieve a relative improvement of 7% in terms of both detection cost and equal error rate over the default i-vector based system while still having lesser overall complexity.

Keywords: Speaker verification; GMM mean supervector; Data-independent projection

1. Introduction

The current text independent speaker verification (SV) systems predominantly use short-term cepstral feature extraction approaches to parameterize the speech signals and Gaussian mixture models (GMM) to model the distribution of the feature vectors (Kinnunen and Li, 2010). An efficient approach to train a GMM for a speaker is to adapt the parameters of a universal background model (UBM) using the speaker dependent data with the maximum-a posteriori (MAP) approach (Reynolds et al., 2000). The UBM is a GMM trained using large amount of speech data gathered from a number of speakers and thus it captures the speaker independent distribution of feature vectors. Adapting the means of the UBM only has been found to work well in practice. For verification purpose, the likelihood of the test data feature vectors over the claimed speaker's model is computed and is compared to a predefined threshold to obtain the decision.

Apart from the likelihood based approaches, the support vector machine (SVM) classifier is also been proposed for the SV task (Wan and Renals, 2005; Campbell et al., 2006a,b). Among the SVM based SV systems, the one using GMM mean super vector representations (Campbell et al., 2006b) happens to be the most popular. GMM mean supervectors are created by concatenating the mean vectors corresponding to a speaker adapted GMM-UBM. The interpretation of GMM mean supervectors as a fixed dimension representation of a speaker utterance has also led to the development of efficient session/channel compensation techniques for the SV task (Hatch et al., 2006; Yin et al., 2007; Kenny et al., 2007). The GMM mean supervec-

[*] Corresponding author.
E-mail addresses: haris@iitg.ernet.in (B.C. Haris), rsinha@iitg.ernet.in (R. Sinha).

http://dx.doi.org/10.1016/j.specom.2014.12.005
0167-6393/© 2015 Elsevier B.V. All rights reserved.

tors are found to be effective in representing speaker utterances, but are noted to be highly redundant in terms of speaker dependent information. To remove the redundancy of the supervectors, commonly, a factor analysis based approach is used in the front-end of SV systems. In that, a learned low-rank projection matrix called the *total variability matrix* (T-matrix) is used to derive the low-dimensional speaker representation which is commonly referred to as *i-vector* (Dehak et al., 2011). The systems using the i-vectors as speaker representations and a probabilistic linear discriminant analysis (PLDA) based classifier is considered to be the state-of-the-art for SV (Kenny, 2010; Garcia-Romero and Espy-Wilson, 2011). The major disadvantage of the i-vector based SV systems is the computational complexity and the memory requirements involved in deriving the i-vector representations. In addition to that a large amount of speech data is required to learn the T-matrix. Recently, some works simplifying the i-vector computation and reducing the memory requirements have been reported (Glembek et al., 2011; Cumani and Laface, 2013; Li and Narayanan, 2014).

In literature, data-independent projection approaches using random matrices are reported to provide a viable alternative to the data-dependent ones like principal component analysis (PCA) for reducing the dimensionality of very high dimensional vectors (Kaski, 1998). In a recent work (Haris and Sinha, 2014), we have explored some low-complexity data-independent projections using random matrices to reduce the dimensionality of GMM mean supervectors for a sparse representation classification based speaker identification system. Among the various projections explored, the sparse random matrix based one demands very little computational resources only and resulted in a performance competitive to that of the i-vector based approach. Such approaches are yet to be explored for an SV task and that forms the basic motivation of this work. The salient contributions reported in this paper are as follows. (1) The use of low-complexity sparse random projections for reducing the dimensionality of supervectors in context of the PLDA based SV system developed on a large multi-variability (NIST 2012 SRE) dataset, (2) A non-random data-independent projection with simple decimation of supervectors is proposed for dimensionality reduction and is shown to be as effective as the sparse random projection while having a much lower complexity. (3) A multi-offset decimation diversity based SV system is proposed which outperforms not only the individual offset decimation based systems but also the default i-vector based system while still having lesser computational requirements.

The rest of the paper is organized as follows. Section 2 presents the low-complexity data-independent projections used in this work. The descriptions on the SV system used and the experimental setup are given in Section 5. Section 6 contains the results and discussions followed by the presentation of the proposed multi-offset decimation diversity based SV system. The paper is concluded in Section 8.

2. Data-independent projections of GMM mean supervectors

In many pattern recognition applications, the use of data-independent random projections as an alternative to the data-dependent projections like PCA and factor analysis have been explored (Kaski, 1998; Bingham and Mannila, 2001). In order to reduce the dimensionality with random projections, the original d-dimensional supervector $\boldsymbol{x}$ is projected to a k-dimensional subspace using a random $k \times d$ matrix $\boldsymbol{R}$ as,

$$\hat{\boldsymbol{x}} = \boldsymbol{R}\boldsymbol{x} \tag{1}$$

where $\hat{\boldsymbol{x}}$ denotes the low dimensional representation of the data. The basis for the use of the random projection for classification tasks lies in the well known Johnson and Lindenstrauss (JL) lemma (Dasgupta and Gupta, 2003). It states that, a set of n points in high dimensional space can be mapped to a k-dimensional subspace such that the Euclidean distance between any two points changes by only a factor of $(1 \pm \epsilon)$, if $k \geq 12\frac{\ln n}{\epsilon^2}$.

The random projection matrix $\boldsymbol{R}$ is typically created using random numbers having a standard normal distribution. Such a projection matrix is much simpler to generate compared to the one derived using the data-dependent approaches like PCA or factor analysis. At the same time, the random matrix based approach does not provide any computational advantage in finding the projections compared to PCA. Also, this approach does not produce representations as compact as that in the data-dependent cases. As a result, the size of the projection matrix in case of the random projection approach would be much higher than that in case of the data-dependent ones. The generation and storage of such a large matrix is non-trivial. To address these issues, in Achlioptas (2001), the use of a non-Gaussian sparse random matrix is proposed. In the following we first describe the data-independent projection using sparse random matrix which is shown to have attractive computational advantages in our earlier work (Haris and Sinha, 2014) in context of sparse representation based speaker identification task. Additionally in this work we have explored the decimation of supervectors as a method of data-independent dimensionality reduction. Interestingly the decimation process can be interpreted as a projection using sparse matrix and the same is also described in the following.

2.1. Sparse random projection matrix

As proposed in Li et al. (2006), the elements of the sparse random projection matrix $\boldsymbol{R}$ are distributed as,

$$[\boldsymbol{R}]_{ij} = +\sqrt{s}\begin{cases} +1 & \text{with probability} \quad 1/2s \\ 0 & \text{with probability} \quad 1-1/s \\ -1 & \text{with probability} \quad 1/2s \end{cases} \tag{2}$$