Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 16–29

www.elsevier.com/locate/csl

Restricted Boltzmann machines for vector representation of speech in speaker recognition[☆]

Omid Ghahabi*, Javier Hernando

TALP Research Center, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona 08034, Spain

Received 26 October 2016; received in revised form 2 May 2017; accepted 19 June 2017
Available online 28 June 2017

Abstract

Over the last few years, i-vectors have been the state-of-the-art technique in speaker recognition. Recent advances in Deep Learning (DL) technology have improved the quality of i-vectors but the DL techniques in use are computationally expensive and need phonetically labeled background data. The aim of this work is to develop an efficient alternative vector representation of speech by keeping the computational cost as low as possible and avoiding phonetic labels, which are not always accessible. The proposed vectors will be based on both Gaussian Mixture Models (GMM) and Restricted Boltzmann Machines (RBM) and will be referred to as GMM−RBM vectors. The role of RBM is to learn the total speaker and session variability among background GMM supervectors. This RBM, which will be referred to as Universal RBM (URBM), will then be used to transform unseen supervectors to the proposed low dimensional vectors. The use of different activation functions for training the URBM and different transformation functions for extracting the proposed vectors are investigated. At the end, a variant of Rectified Linear Units (ReLU) which is referred to as variable ReLU (VReLU) is proposed. Experiments on the core test condition 5 of NIST SRE 2010 show that comparable results with conventional i-vectors are achieved with a clearly lower computational load in the vector extraction process.

Keywords: Restricted Boltzmann machine; Deep learning; Variable rectified linear unit; Speaker recognition; GMM−RBM vector; i-vector

1. Introduction

The low dimensional representation of a speech utterance based on the factor analysis technique is well-known as i-vector (Dehak et al., 2011a). Over the past few years, i-vectors have shown a great performance not only in speaker recognition but also in other applications (e.g., Dehak et al., 2011b; Bahari et al., 2012; Xia and Liu, 2012). Two commonly used scoring techniques for i-vectors are cosine distance (Dehak et al., 2010; 2011a) and Probabilistic Linear Discriminant Analysis (PLDA) (Prince and Elder, 2007; Kenny, 2010). PLDA scoring leads to a superior performance but needs speaker-labeled background data which is costly and not accessible easily.

[☆] This paper has been recommended for acceptance by Roger Moore.

* Corresponding author.
E-mail address: omid.ghahabi@upc.edu (O. Ghahabi).

http://dx.doi.org/10.1016/j.csl.2017.06.007

Motivated by the success use of Deep Learning (DL) in other speech processing applications (e.g., Mohamed et al., 2010; Dahl et al., 2012; Mohamed et al., 2012; Hinton et al., 2012; Senior et al., 2015), DL techniques have also been used in speaker recognition for different purposes. For example, DL techniques have been applied as a backend on i-vectors (Stafylakis et al., 2012b; Senoussaoui et al., 2012; Stafylakis et al., 2012a; Novoselov et al., 2014; Ghahabi and Hernando, 2014a; 2014b; 2017), used in the i-vector extraction algorithm (Lei et al., 2014; Kenny et al., 2014; Mclaren et al., 2015; Richardson et al., 2015; Liu et al., 2015; Campbell, 2014; Garcia-Romero et al., 2014), and also employed for compact representation of speech signals (Vasilakakis et al., 2013; Variani et al., 2014; Liu et al., 2015; Ghahabi and Hernando, 2015; Safari et al., 2016) and discriminative feature classification (Safari et al., 2015).

DL technology has been used in the i-vector extraction algorithm in two ways. First, a Deep Neural Network (DNN) has been used for acoustic modeling rather than the typical Gaussian Mixture Model (GMM) (Lei et al., 2014; Kenny et al., 2014; Campbell, 2014; Richardson et al., 2015; Garcia-Romero et al., 2014; Liu et al., 2015). Second, conventional spectral features have been replaced or appended by the so-called DNN bottleneck features and then a DNN or a GMM has been used as an acoustic model (Mclaren et al., 2015; Richardson et al., 2015; Liu et al., 2015). It has been shown that the best results are obtained when spectral features are appended by bottleneck features and a GMM is used as an acoustic model (Mclaren et al., 2015; Richardson et al., 2015; Lozano-Diez et al., 2016). However, the main problem is that the use of DNN as either an acoustic model or bottleneck feature extractor increases highly the computational cost of the i-vector extraction process. Moreover, in both cases phonetic labels are required for DNN training, which are not always accessible.

On the other hand, only a few works have tried to make use of DL techniques to build a compact representation of speech signals without using the conventional i-vector algorithm. In Vasilakakis et al. (2013), Variani et al. (2014), Liu et al. (2015), a deep architecture is trained using background feature vectors. Then the feature vectors of a given utterance are forward-propagated and the mean of the posterior probabilities of a particular hidden layer (Variani et al., 2014) or a PCA dimension reduced version of them (Liu et al., 2015), or a PCA dimension reduced version of the mean vectors (Vasilakakis et al., 2013) are considered as a new compact representation. In Safari et al. (2016), the parameters of the adapted networks are stacked to build a supervector. Then the dimension of the new supervectors are reduced by PCA. In Ghahabi and Hernando (2015), the authors used the GMM supervectors, rather than the feature vectors, as the inputs to a Restricted Boltzmann Machine (RBM). RBM has been used as a dimension reduction stage in that scenario. Although Liu et al. (2015) and Variani et al. (2014) have shown some success in text-dependent speaker recognition, still no significant improvement is reported for text-independent tasks. Moreover, working with DL techniques in feature vector domain is costly.

The aim of this work is to develop an efficient framework for vector representation of speech by keeping the computational cost as low as possible and avoiding phonetic labels. In order to achieve this goal, a global RBM referred to as Universal RBM (URBM) is trained given background GMM supervectors. The URBM tries to learn the total session and speaker variability among background supervectors. It will then be used to transform unseen supervectors to lower dimensional vectors which will be referred to as GMM−RBM vectors.

Compared to the preliminary work presented in Ghahabi and Hernando (2015), whitening in the supervector domain, which is computationally costly, is replaced by warping in the feature vector domain. This change makes possible to obtain higher speaker recognition accuracy, specially in lower dimensional vectors. Moreover, the effect of the type of the activation function for training the URBM and the type of the transformation function for GMM−RBM vector extraction are investigated. At the end, a variation of Linear Rectified Units (ReLU), which will be referred to as variable ReLU (VReLU), is proposed for training the URBM, and then a linear function is used for transformation in the vector extraction stage.

The core condition of NIST SRE 2006 (NIST, 2006) is used for the development and the core condition 5 of NIST SRE 2010 (NIST, 2010) with much bigger background data is used for the test and evaluation. The experiments on the evaluation set shows that the proposed GMM−RBM vectors achieve comparable performance with conventional i-vectors while lower computational cost is required for vector extraction. The conclusion is valid with both cosine and PLDA scoring. Moreover, the combination of GMM−RBM vectors and i-vectors at the score level improves the performance more.

The rest of the paper is organized as follows. Section 2 gives a brief background overview about conventional i-vectors and PLDA. Section 3 describes the proposed GMM−RBM vectors. Section 4 investigate the effect of