Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 161–179

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Domain compensation based on phonetically discriminative features for speaker verification

Yanhua Long *, Hong Ye, Jifeng Ni

Department of Electronical and Information Engineering, Shanghai Normal University, Shanghai, 200234, China

Received 10 June 2015; received in revised form 13 January 2016; accepted 4 July 2016
Available online 7 July 2016

Abstract

This paper presents a new domain compensation framework by using phonetically discriminative features which are extracted from domain-dependent deep neural networks (DNNs). The domain compensation can be applied in both unsupervised and supervised manner, depending on whether the domain information of the development data is provided or not in advance. In supervised manner, the DNNs are trained on the development speech recordings of each given domain separately. While in the unsupervised manner, the development datasets are first automatically clustered into different domains, by using the Gaussian Mixture Model mean supervectors which are generated from each of the speech recordings, DNNs are then trained on the resulting clusters. Finally, we compensate the domain variabilities during the target speaker modeling step using support vector machines, by feeding in statistical vectors which are derived from the discriminative features extracted from the domain-dependent DNNs. The main strength of our proposed framework is that it does not need any speaker labels in the development dataset, which makes the proposed framework of great advantage over the state-of-the-art techniques that need speaker labels to train inter-speaker and/or intra-speaker variability models or channel compensation. Three speaker verification systems are investigated to examine the effectiveness of this new framework. Experimental results on the NIST SRE 2010 task demonstrate competitive performances to the state-of-the-art techniques in an initial implementation of the proposed framework.

Keywords: Discriminative features; I-vector; Domain compensation; Deep neural network; Speaker verification

1. Introduction

The session variability between training and test recordings of the same speaker can heavily degrade the system performances in speaker verification. This type of variability is usually attributed to the channel effects, although it also includes phonetic and intra-speaker variations such as the changes of speaker's emotion, health, etc. (Kenny et al., 2007; Kinnunen and Li, 2010). And nowadays, it remains to be the most challenging problem in the state-of-the-art speaker recognition.

In recent years, many techniques have been proposed to effectively eliminate these session variabilities. For instance, the Joint Factor Analysis (JFA) (Kenny et al., 2007, 2008) provides an extremely powerful ability to handle the channel compensation for Gaussian Mixture Models (GMM)-based speaker verification systems. It models a speaker conversation/recording as a linear combination of a speaker and channel/session components. Meanwhile, as a representative

* Corresponding author at: Department of Electronical and Information Engineering, Shanghai Normal University, Shanghai, 200234, China. Fax: +86-21-6432 2418.
E-mail address: yanhua@shnu.edu.cn (Y. Long).

http://dx.doi.org/10.1016/j.csl.2016.07.001
0885-2308/

of the discriminative speaker modeling techniques, the Support Vector Machines (SVM)-based speaker verification systems yield competitive results and provide a strong performance complementary to GMM-based systems (Campbell et al., 2006c; Guo et al., 2009; Sturim et al., 2011). The typical techniques to deal with channel effects for these systems are Nuisance Attribute Projection (NAP) (Campbell et al., 2006d; Sun et al., 2013) and Within-Class Covariance Normalization (WCCN) (Hatch et al., 2006). Moreover, the important recent advance in speaker recognition has been the development of i-vector representation technique, which can be regarded as an extension of JFA. Most state-of-the-art speaker verification systems are dominated by using i-vector to represent a given speech utterance (Dehak et al., 2011; Greenbery et al., 2014; Lei et al., 2013. Once i-vectors are extracted, the key to success of these systems mainly depends on the backend channel compensation techniques, such as the most recent studied variants of Probabilistic Linear Discriminant Analysis (PLDA) techniques (Cumani et al., 2013; Prince and Elder, 2007; Villalba and Brummer, 2011), and the i-vector based WCCN, NAP and LDA, etc (Dehak et al., 2011; Kanagasundaram et al., 2012; Vogt et al., 2008). Overall, these techniques have proven to be successful and brought a significant increase in accuracy on tasks like NIST speaker recognition evaluations (SREs). However, their success was mainly due to the sufficient data available in NIST SREs. To achieve optimal performances, not only the traditional GMM and SVM based channel compensation techniques need a large amount of speech data with accurate speaker labels, but also the backend channel compensation algorithms of recent i-vector based systems have the same data requirement. More importantly, they are typically trained on tens of thousands of speech cuts, which come from thousands of speakers with multiple cuts per speaker from different sessions, and these sessions always have a variety domain or condition coverage (Banse et al., 2014; Garcia-Romero et al., 2014). This indicates that a sufficient labeled (speaker and other kinds of metadata) development dataset is the key to guarantee a good speaker verification performance.

However, in real applications, it is typically extremely expensive to generate the required accurate labels, and in many cases, the speaker labels are unknown. Therefore, it is unrealistic to expect such a large set of labeled development data for every domain of interest. This motivates researchers start to explore approaches to alleviate the domain mismatch, or new channel compensation algorithms which require no or less labeled data. Such as the IDVC (Inter-dataset variability compensation approach proposed in Aronowitz, 2014a, 2014b), the whole development dataset was first split into several subsets according to the available metadata, then a low-dimensional inter dataset variability subspace was estimated by applying PCA (principal component analysis) on the averaged i-vectors which were derived from each subset. The dataset mismatches were then compensated by removing the low-subspace from all i-vectors as a pre-processing step before PLDA training and scoring. Meanwhile, works in Khoury et al. (2014) and Sun and Ma (2013) managed to obtain solid results without the use of any labeled data from unsupervised i-vector clustering point. In Garcia-Romero et al. (2014), an existing out-of-domain i-vector PLDA system trained on labeled data was used to cluster unlabeled in-domain data, and the resulting clusters were then used to adapt the parameters of PLDA system. Based on the variational Bayesian method, Villalba and Lleida (2014) applied the similar idea to adapt a well-trained PLDA model with unlabeled data. In Shum et al. (2014), active learning was investigated to obtain data labels from a noiseless oracle in the form of pairwise queries. However, most of these recent works still need multiple speech cuts which are collected from variety or diverse conditions for each speaker cluster. The actual deployment of a speaker recognition system into the real world is unlikely to warrant this requirement.

In this paper, we focus on dealing with the session variabilities which result from the domain mismatch between training and test recordings. The “domain” used here is not restricted to the typical domains as used in speech recognition, like “conversational” or “lecture”, it can also be one type of channel or condition. We refer to the variability/mismatch compensation as “domain” compensation in this paper. A new framework is proposed to eliminate these mismatches, and it can work well in both supervised and unsupervised manner. Motivated by the recent success of using deep neural network (DNN) as a front-end discriminative bottle-neck (BN) or phonetic posterior feature extraction in speaker verification (Diez et al., 2014; McLaren et al., 2015; Sarkar et al., 2014), in this framework, we also investigate these similar discriminative features, however, they are extracted from multiple domain dependent DNNs instead of a single one, and these features are utilized in a totally different way to model both the speaker traits and domain information. The session variabilities are compensated in the target speaker modeling step, which is based on the vector representation (i-vector, GMM mean supervector, etc) architecture. Three state-of-the-art speaker verification systems are built to examine the effectiveness of this new framework. And unlike recent works of McLaren et al. (2015) and Richardson et al. (2015), the discriminative features were concatenated to the acoustic features to form tandem features for speaker modeling, we expect the same performance improvements by examining the complementary information between discriminative and traditional acoustic features at the score-level system fusion stage.