Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 80 (2016) 60–70

www.elsevier.com/locate/specom

Feature sparsity analysis for i-vector based speaker verification

Wei Li[a,*], Tianfan Fu[b], Hanxu You[a], Jie Zhu[a], Ning Chen[c]

[a] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Department of Computer Science and Engineering (CSE), Shanghai Jiao Tong University, Shanghai 200240, China
[c] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200240, China

Received 30 October 2015; received in revised form 4 February 2016; accepted 4 February 2016
Available online 29 April 2016

Abstract

In recent years, the i-vector based framework has been proven to provide state-of-the-art performance in the speaker verification field. Each utterance is projected onto a total factor space and is represented by a low-dimensional i-vector. However, the degradation of performance in the i-vector space remains problematic and is commonly attributed to channel variability. Most techniques used for the channel compensation of the i-vectors, such as linear discriminant analysis (LDA) or probabilistic linear discriminant analysis (PLDA) aim to compensate for the variabilities caused by channel effects. However, in real-world applications, the duration of enrollment and test utterances by each user (speaker) are always very limited. In this paper, we demonstrate, from both analytical and experimental perspectives, that feature sparsity and imbalance widely exist in short utterances, in which case the conventional i-vector extraction algorithm, based on maximum likelihood estimation (MLE), may lead to over-fitting and decrease the performance of the speaker verification system, especially for short utterances. This prompted us to propose an improved i-vector extraction algorithm, which we term adaptive first-order Baum–Welch statistics analysis (AFSA). This new algorithm suppresses and compensates for the deviation from first-order Baum–Welch statistics caused by feature sparsity and imbalance.

We reported results on the male telephone portion of the core trial condition (short2-short3) and other short time trial conditions (short2-10sec and 10sec-10sec) on NIST 2008 Speaker Recognition Evaluations (SREs) dataset. As measured both by Equal Error Rate (EER) and the minimum values of the NIST Detection Cost Function (minDCF), 10%–15% relative improvement is obtained compared to the baseline of traditional i-vector based system.

Keywords: Speaker verification; i-vector; Total factor space; Feature variability; Adaptive first-order Baum–Welch statistics analysis (AFSA).

1. Introduction

Speaker verification technology is used to accept or reject a claimed identity by comparing two utterances. The first of these utterances is used for enrollment and is produced by the speaker with a target identity, whereas the second utterance is obtained from the speaker with a claimed identity and is used for testing purposes. In the last decade, the Gaussian mixture model based on the universal background model (GMM-UBM) framework has demonstrated strong performance and has become the most widely used method for speaker verification. In this framework, the mean vectors of each Gaussian component are commonly considered to represent most of a speaker's characteristics, whereas the other parameters, such as the weights and variances of each Gaussian component, are inherited from the UBM model (Reynolds et al., 2000). Traditional speaker models are obtained by employing a maximum a posteriori (MAP) adaptation. However, traditional MAP (or relevance MAP) treats each Gaussian component as a statistically independent distribution, which has many drawbacks in practical applications: only those components with a sufficient number of assigned speaker frames are well adapted, leaving the remaining components almost unchanged. Time-limited enrollment and test utterances or those that suffer from severe phonetic variability may lead to an obvious degradation in the performance of the verification system. Apart from

* Corresponding author. Tel.: +86 136 1163 6772; fax: +86 021 34205432.
E-mail address: liweisjtu@126.com (W. Li).

http://dx.doi.org/10.1016/j.specom.2016.02.008
0167-6393/

these disadvantages, traditional MAP does not have the ability to compensate for the effects of channel distortion, especially when the enrollment and test utterances use different channels.

An extension of the GMM-UBM framework, namely the factor analysis (FA) technique (Kenny et al., 2005, 2008), attempts to jointly model the speaker components. Each speaker is represented by a *mean supervector*, which is a linear combination of the set of *eigenvoices*. Generally, only a few hundred free parameters need to be estimated, which ensures that the speaker mean supervector converges quickly by using a training utterance with a relatively short duration. Based on the FA technique, *joint factor analysis (JFA)* (Kenny, 2005; Kenny et al., 2007) decomposes the GMM supervector into a speaker component **S** and a channel component **C**, and assumes these two components to be statistically independent. Although it is known by now that channel effects are not speaker-independent (for example, experimentally, gender-dependent eigenchannel modeling has been reported to be more effective than gender-independent modeling Kenny, 2010), compared to other methods JFA has still demonstrated good performance for text-independent speaker verification tasks in past NIST speaker recognition evaluations (SREs).

Inspired by the JFA approach, the authors in Dehak et al. (2011) proposed a combined speaker and channel space by defining a novel low-dimensional space named the *total factor space*. In this space, each utterance is represented by a low-dimensional feature vector termed an i-vector. The concept of an i-vector has opened the door for new ways in which to analyze speaker and session variability. As a result, various optimization and compensation techniques and scoring methods have been proposed (Bousquet et al., 2012, 2011; Dehak et al., 2011; Kenny, 2010), all of which have improved the results obtained with the JFA approach. Of late, i-vector extraction with length normalization and probabilistic linear discriminant analysis (PLDA) has become the state-of-the-art configuration for speaker verification (Bousquet et al., 2012; Kenny, 2010).

Although the i-vector based system has dominated the speaker verification field, recent research found the adaptive relevance factor, which replaces the traditional manually tuned relevance factor, capable of boosting the MAP-based Gaussian mixture model - support vector machine (GMM-SVM) framework to obtain a performance comparable to those of the JFA and i-vector frameworks (You et al., 2012, 2013).

Despite the success of the i-vector paradigm, it still has some shortcomings, one of which is that its applicability to *text-dependent speaker verification* continues to remain difficult. In cases in which the lexical contents of enrollment and test utterances are identical, we may assert that only those components with a sufficient number of speaker frames need to be adapted. However, FA-based techniques globally adapt all the Gaussian components, including those components with sparse or no speaker frames, which results in the performance of FA-based techniques not being comparable to that of traditional MAP adaptation in text-dependent speaker verification. Related work has also shown that results obtained with the traditional MAP approach can be better than those obtained with i-vector based methods (Aronowitz, 2012; Larcher et al., 2012; Stafylakis et al., 2013).

Considering that the i-vector based system is an extension of the text-dependent speaker verification field, in the context of text-independent verification, the principal challenge of this system in terms of achieving a low error rate is that the intra-speaker variability in the estimated parameters increases considerably as a result of variability in the lexicon and the training utterance duration (Hautamäki et al., 2013) (text-dependent speaker verification can be regarded as a special case of text-independent speaker verification of short utterances).

Two main streams have emerged to address the shortcomings of i-vector applicability to short utterances: normalizing those i-vectors derived from short utterances in the low-dimensional i-vector space (Kanagasundaram et al., 2012; Kenny et al., 2013; Larcher et al., 2013) and regularizing the Baum–Welch statistics of short utterances to obtain a more robust i-vector estimation (Hautamäki et al., 2013). In the analysis presented in this paper, we attempt to show that, although the FA-based i-vector approach is capable of successfully modeling speaker variability, on some occasions the traditional i-vector extraction algorithm may lead to overfitting. This problem typically occurs when the speech frames from the target speaker are sparse. In this paper, we continue to focus on the *text-independent* speaker verification field, and propose an improved i-vector extraction algorithm we have named *adaptive first-order Baum–Welch statistics analysis* (AFSA). AFSA attempts to suppress and compensate for the biased first-order Baum–Welch statistics caused by feature sparsity. Traditionally, zero-order and first-order Baum–Welch statistics are considered as determinate values extracted from a posteriori statistics of speaker frames based on the UBM model. Our approach is to provide first-order Baum–Welch statistics with a Bayesian explanation (although we continue to refer to it as "statistics"). Our proposed method treats phonetic variability and channel variability as mutually independent distributions, and as we show in the experimental section, various channel compensation techniques continue to work efficiently. Experiments were carried out on the core condition (short2-short3) and other conditions of a short duration (short2-10sec and 10sec-10sec) of NIST 2008 SREs. The experimental results show that, by applying the AFSA algorithm to the phase of i-vector extraction, a 10%–15% relative improvement is obtained compared with a baseline system in which a traditional i-vector algorithm with the same channel compensation techniques is adopted.

2. Supervector, total factor space, and i-vector extraction

The GMM-UBM framework is commonly considered to only require adaptation of the mean vectors of each Gaussian component for a given UBM model Ω and a training utterance. A *supervector* comprises the concatenation of each of the mean vectors. In the context of the i-vector framework,