Computers and Electrical Engineering 48 (2015) 226–238

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Non-speaker information reduction from Cosine Similarity Scoring in i-vector based speaker verification ☆

Hossein Zeinali [a,*], Alireza Mirian [a], Hossein Sameti [a], Bagher BabaAli [b]

[a] *Department of Computer Engineering, Sharif University of Technology, Iran*
[b] *School of Mathematics, Statistics, and Computer Science, University of Tehran, Iran*

ARTICLE INFO

Article history:
Received 11 March 2015
Received in revised form 30 August 2015
Accepted 2 September 2015
Available online 17 October 2015

Keywords:
Cosine similarity
I-vector
Speaker verification
Non-speaker information

ABSTRACT

Cosine similarity and Probabilistic Linear Discriminant Analysis (PLDA) in i-vector space are two state-of-the-art scoring methods in speaker verification field. While PLDA usually gives better accuracy, Cosine Similarity Scoring (CSS) remains a widely used method due to simplicity and acceptable performance. In this domain, several channel compensation and score normalization methods have been proposed to improve the performance. We investigate non-speaker information in cosine similarity metric and propose a new approach to remove it from the decision making process. I-vectors hold a large amount of non-speaker information such as channel effects, language, and phonetic content. This type of information increases the verification error rate and hence it should be removed from the scoring method. To this end we propose a method that estimates non-speaker information between two i-vectors using the development set and subtracts it from cosine similarity. The results indicate that the proposed method performed better than other implemented methods based on the cosine similarity. Furthermore, in certain cases the performance of this method was better than the PLDA method and when combined with PLDA performance was improved in most cases.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Low dimensional representation of speech in the so called Total Variability Space (TVS) which is obtained by Factor Analysis (FA) in the Gaussian Mixture Model (GMM) mean supervector space has been popular in the state-of-the-art speaker verification systems. This representation, which is commonly known as i-vector, maps utterances with arbitrary duration into a low and fixed dimensional space [1].

I-vector holds a substantial amount of information and therefore it has been found to be useful in various applications. Language and particularly phonetic details are among the most important information found in this vector. Thus, several methods have exploited this vector for Language Identification (LID) purposes [2–6]. Using this vector for age [7,8] and emotion estimation [9] along with accent recognition [10,11] are among the other applications of this kind of representation. All these applications as well as its main application which is speaker recognition, show the great amount of information it contains, hence it seems reasonable to reduce the unrelated information of this vector for better speaker recognition.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. A. Isazadeh.

[*] Corresponding author.

E-mail addresses: zeinali@ce.sharif.edu (H. Zeinali), alirezamirian@ce.sharif.edu (A. Mirian), sameti@sharif.edu (H. Sameti), babaali@ut.ac.ir (B. BabaAli).

http://dx.doi.org/10.1016/j.compeleceng.2015.09.003
0045-7906/© 2015 Elsevier Ltd. All rights reserved.

Different methods have been proposed for speaker modeling in i-vector space [12,13]. Gaussian PLDA is the most common one which ignores the process by which i-vectors are extracted (i.e., the point estimate of hidden variables in FA model) and instead pretends that they are random vectors generated by using PLDA method. Although in most cases PLDA results in better accuracy, CSS is its close competitor with less logical and computational complexity. Usually, CSS performance gets close to PLDA after a couple of channel compensation methods like LDA and Within-Class Covariance Normalization (WCCN) and also followed by normalization techniques like ZT-norm. Simplicity of CSS in comparison with PLDA and its good performance, particularly in applications where the speech duration is short (e.g. text-prompted speaker verification) encourages the idea of extending or modifying it in order to improve its performance. A number of channel compensation techniques such as SN-LDA, SN-WLDA and WLDA have been suggested to improve performance of CSS based i-vector systems [14]. Additionally, certain studies have been carried out in the score domain such as the normalized cosine kernel introduced in [15].

In general, the best modeling technique for speaker recognition is the one where only the speaker-related information is taken into account. Regarding the mentioned applications for i-vector, it is apparent that this vector is not the ideal option for the speaker recognition. In addition, the best similarity measurement for scoring in the speaker recognition is the measure where only speaker-related information is considered. Experiments show that the cosine similarity measure does not have this quality and measures speaker-related information as well as non-speaker information such as channel effects, phonetic content, and perhaps other kinds of information.

In this paper, we address the issue of these kinds of information in cosine similarity and try to reduce the effects of them from the decision making process. The rest of the paper is organized as follows: In Section 2 some of the related works are described briefly and then in the following section the parts of an i-vector based speaker verification system are explained. In Section 4, three observations which suggest the existence of non-speaker information in cosine similarity are described. Section 5 is dedicated to the proposed method based on these observations. Experiments and results are reported in Section 6 and finally conclusions and a discussion of future works are presented in Section 7.

2. Related works

In general, two types of error occur in speaker verification systems. The first is the false rejection error. The main reason for this type of error is the mismatch between the train and test times, which can be due to various causes such as change of channel or handset[16–18], change in the state of the speaker (e.g. stress [19], hurry, etc.) or even intentional changes in the utterance such as whispering. All these variations in conditions of speech at test and train times increase the intra-class variances. This leads to higher false rejection errors.

Different compensation methods for reducing the effects of such variations have been proposed:

1. NAP: Nuisance Attribute Projection (NAP) removes the nuisance subspaces. An orthogonal projection depending only on the speaker is performed in the complementary space of the channel by the projection matrix. This method has been employed in both mean supervector and i-vector spaces [1,20–22].
2. WCCN: The principal goal of this method is to minimize the false rejection and false acceptance errors in the training stage of Support Vector Machine (SVM). In other words, this method scales the space to remove the dimensions with high intra-class variance. Therefore, this method reduces intra-class variance [1,23,24].
3. JFA: The objective of Joint Factor Analysis (JFA) method is to model the channel effects and the speaker separately in two different subspaces. In this method the mean supervector is decomposed to two separate supervectors where one is related to the speaker and the other is related to the channel. After this decomposition, the supervector representing the channel effects is discarded and hence the channel effects are excluded from the decision [16,25,26].
4. LDA: This method is used in pattern recognition for reducing intra-class variances and to increase discrimination between classes. This method is also used for dimensionality reduction in classification. Since this method decreases intra-class variations, it can be used for reducing channel effects [1].
5. PLDA: This is a probabilistic method which aims at decomposing the i-vector space to two separate subspaces for the speaker and channel. This method is similar to JFA in some respects [12,13].

The second type of error in speaker verification systems is the false acceptance error. The main cause for this error is the presence of similarity between vectors of different speakers. These similarities have various reasons and can be grouped in two categories: the first is the set of speaker-related similarities. This type of similarity enables us to identify a speaker using their vectors. In the case of two speakers with very similar voices (whether naturally or by voice emulation/conversion), the speaker-related similarity between their respective vectors increases and this in turn leads to increase in false acceptance error rate. The second category of similarities is the set of non-speaker similarities of the vectors which may be due to different sources such as similarity in channel, microphone or similarity in the uttered text. As far as we know no method has been proposed to reduce the effects of this kind of similarity. This research aims at investigating non-speaker information in cosine similarity and proposing a method to reduce their effects.

Download English Version:

https://daneshyari.com/en/article/455162

Download Persian Version:

https://daneshyari.com/article/455162

Daneshyari.com