Pattern Recognition Letters 49 (2014) 17–23

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Session compensation using binary speech representation for speaker recognition ☆

Gabriel Hernandez-Sierra [a,b,*], Jose R. Calvo [a], Jean-François Bonastre [b], Pierre-Michel Bousquet [b]

[a] Advanced Technologies Application Center, Havana, Cuba
[b] University of Avignon, LIA, France

ARTICLE INFO

Article history:
Received 23 December 2013
Available online 14 June 2014

Keywords:
Speaker recognition
Session variability compensation
Nuisance attribute

ABSTRACT

Recently, a simple representation of a speech excerpt was proposed, as a binary matrix where each acoustic frame is represented by a binary vector. This new approach relies on the UBM paradigm but shifts the speaker recognition workspace from a continuous probabilistic to a discrete, binary discrete space, allowing easy access to the speaker discriminant information. In addition to the time-related abilities of this representation, it also allows the system to work with a more compact representation based on cumulative vectors. A cumulative vector is the sum of a set of frame-based binary vectors. In this space, global information can be exploited to compensate for the effects of session variability. This work is mainly dedicated to this aspect. A new variability compensation method in the cumulative vector space is proposed in order to remove not only the unwanted attributes of session variability but also the common attributes among speakers. This is done by incorporating in the projection matrix the common information to all classes. A specificity selection approach using a mask in the cumulative vector space is also proposed. This aims to reduce the non informative coefficients. The experimental validation, done on the NIST-SRE framework, demonstrates the efficiency of the proposed solutions, which shows an EER improvement from 42% to 61%. The combination of i-vector and binary approaches, using the proposed methods, showed the complementarity of the discriminatory information exploited by each of them.

1. Introduction

State-of-the-art speaker recognition methods are mainly based on the Gaussian Mixture Model (GMM)/Universal Background Model (UBM) paradigm [15,3]. In the GMM-UBM approach, a GMM — the UBM — represents the global acoustic space and a given speaker is defined by a GMM derived from the UBM, using the available speech data gathered for this speaker.

The supervector approach uses GMM-UBM as a root. In this framework, each speech excerpt is represented by a vector obtained by the concatenation of the means of the Gaussian components. These supervectors form a high dimensional representation space [8] where all the remaining processes are computed. This approach has been a major breakthrough in the evolution of speaker recognition systems. In particular, the supervector allows the use of Support Vector Machines (SVM) as discriminant classifiers, as well as direct modeling of session variabilities. More recently, two major evolutions were proposed in the supervector framework: Joint Factor Analysis (JFA) by Kenny et al. [13], and i-vector [9].

These algorithms showed a very good level of performance (for example in the NIST speaker recognition evaluations (SREs), all the best performing systems are based on JFA or i-vector). However, they are associated with two main drawbacks. First, it is difficult to work with sequential/temporal speech information because each set of acoustic vectors is represented only by a point in the supervector space. Second, the underlined paradigm is the statistical one, where the influence of specific information is mainly gathered by the frequency of this information. That is, if an event occurs often for a given speaker but very rarely for the other ones, it will scarcely be taken into account by these approaches, which could appear as paradoxal when the aim is to discriminate the speakers.

Alternatively, two binary approaches have emerged recently. First, a representation on the spectral levels called "Boosted Slice Classifiers" was proposed by Roy et al. [18]; for each speech excerpt a set of binary features carrying maximal discriminative information is estimated. To this end, a transformation

☆ This paper has been recommended for acceptance by Ajay Kumar.

* Corresponding author at: Advanced Technologies Application Center, Havana, Cuba. Tel.: +53 53479046.

E-mail addresses: gsierra@cenatav.co.cu (G. Hernandez-Sierra), jcalvo@cenatav.co.cu (J.R. Calvo), jean-francois.bonastre@univ-avignon.fr (J.-F. Bonastre), pierre-michel.bousquet@univ-avignon.fr (P.-M. Bousquet).

http://dx.doi.org/10.1016/j.patrec.2014.05.021
0167-8655/

$\phi : \mathbb{R}^d \rightarrow \{0, 1\}$ of the spectral space into a binary space is performed, using a simple operation of difference between each pair of spectral coefficients. This approach does not involve GMM-UBM, rendering it faster to obtain the binary representation. The work aims to address two problems concerning the computational cost and performance in noisy environments.

A second simple representation of speech which shifts from a continuous probabilistic workspace to a binary discrete space was proposed in [1,4,12]. It is based on local binary decisions, taken for each acoustic frame. Contrary to the previous statistical approaches, this binary-based framework is able to model infrequent and discriminant events. It also allows us to represent a speech excerpt as a binary matrix, since each acoustic frame is represented by a binary vector.

Thanks to this binary matrix representation of a speech excerpt, the speaker discriminant sequential/temporal information could be used as demonstrated in [5,12]. Moreover, using a very simple transformation process of the binary matrices, this representation also allows us to build a new supervector space. The transformation process is just an accumulation of the binary representation of the frames among a speech segment. It gives a cumulative vector of the same dimension as the frame binary vectors, thus yielding a representation of a speech excerpt through a compact yet informative vector. This specificity of the binary approach is of high interest since this new space allows the system to compensate for the effects of session variabilities, as in JFA or i-vector approaches, without losing its intrinsic qualities.

This work aims to address the compensation of session variability in the cumulative vector space. Each of the coefficients of a cumulative vector reports the level of activation, in terms of number of frames, of a given specificity model (a Gaussian model).

The main contributions of this work are: a mask capable of selecting the most discriminatory specificities in the cumulative space and a new variant to represent its within-class scatter for compensation techniques able to capture more variability information, both results improve the speaker recognition performance in front to session variability.

Another results are: the application of two variability compensation techniques on the cumulative space, the evaluation of the behavior of these techniques on i-vector framework and the demonstration of the complementary information between binary and i-vector approaches by mean of fusion of their speaker recognition scores.

The rest of the article is structured as follows: Section 2 brings an overview of the Speaker Binary Key, Sections 3 and 4 describe the proposed methods, specificities selection and session compensation respectively; Sections 5 and 6 are dedicated to an experimental validation based on NIST SRE 2008 protocol; finally, Section 7 brings some conclusions.

2. Overview of Speaker Binary Key

The Speaker Binary Key relies mainly on the "Generator Model" (GM). The GM contains all the acoustic descriptions of the "specificities," which are speaker discriminant information on which local binary decisions will be made. This acoustic model is built *a priori* during the development phase. Several methods have been proposed to create the GM [1,4,12], but all under the same philosophy.

We will use the GM proposed in [12], illustrated in Fig. 1. This model is composed of a classic UBM associated with a bag of (mono) Gaussian models. The UBM plays a structural role. It defines a partition of the acoustic space into particular acoustic regions; each one of which is associated with one of the UBM components. The bag of Gaussian components contains the specificity models and it is divided into several sets. Each set is linked to a particular acoustic region, as determined by the UBM. The specificity models are selected from a set of GMM, trained with matrices that are composed of the centers of the components, which belong to the adapted models (the same speakers to create the UBM were used). Refer to [12] for details.

The logic behind this GM model is to associate the well known power of the UBM in terms of acoustic space structuration to a fine modeling of the discriminant aspects, separately for each of the regions. This GM-based modeling increases the discriminative power compared to a classical GMM-UBM approach, thanks to both a higher number of parameters per acoustic region and the ability to capture infrequent speaker characteristics.

A sparse binary matrix that represents each speech utterance is obtained using the GM. This binary matrix represents the best relationships between each acoustic frame and the bag of specificity models, organized by UBM components. For each frame, the top UBM components are identified first. Within each of these UBM component the specificities with greater likelihoods, above a given threshold, are considered "activated", following a local binary decision framework. The binary coefficients previously mentioned before, this process gives a fine spatial description of the speaker specific information for each acoustic feature, contrarily to classical JFA or i-vector algorithms.

GM allows the system, frame by frame, to perform a transformation $F : \mathbb{R}^d \rightarrow \mathbb{N}^m$ of d-dimensional acoustic frames to a high dimensional binary space ($m \gg d$). Then, the cumulative vector (CV), which is a compact form of the matrix, is simply obtained by adding the rows of the binary matrix. The CV highlights the level of activation of each GM specificities. Finally, a third representation, also the shortest one, is a binary vector (BV) obtained from the CV by changing the non zero values in the CV into 1. The BV represents the active specificities for a given speech excerpt, independently of the level of activation of the various specificities.

Notice that cumulative vectors belong the to natural domain ($CV \in \mathbb{N}$, including zero), and differs from the i-vectors domain, which requires a specific treatment.

The comparison criteria between two speakers A and B is defined as Intersection and Symmetric Difference Similarity (ISDS) as proposed in [12]. This similarity uses the CV and BV for each speaker.

$$ISDS(A, B) = \frac{\sum_{i=1}^{|A \cap B|} a_i + b_i}{\left(\sum_{j=1}^{A-B} a_j + \sum_{j=1}^{B-A} b_j\right) * \sum_{i=1}^{|A \cap B|} |a_i - b_i|} \quad (1)$$

where $\{\forall a \in A,\ \forall b \in B | A - B \neq \emptyset \text{ and } \exists\ a \neq b | (a, b) \in A \cap B\}$.[1]

The similarity measure is driven by the BVs of two speakers but is applied on the corresponding CV. Elements that belong to the intersection and the symmetric difference between the BVs are used as indexes in CV.

3. Information selection (mask)

The mask for cumulative vectors, described below, is focused on reducing the dimensionality by discarding uninformative coefficients inside the CV space. The process to obtain the mask consists in a selection algorithm, based on the specificities with little or no variance within the population that do not have interesting information. Therefore, these specificities are not necessary to compare two cumulative vectors.

[1] Note: Given the nature of our sets that ensures that all sets have the same number of elements, the cases $A \supset B$ or $B \supset A$ do not exist, therefore $A - B \neq \emptyset \iff B - A \neq \emptyset$.