Temporal Information in a Binary Framework for Speaker Recognition

Gabriel Hernández-Sierra[1,2], José R. Calvo[1], and Jean-François Bonastre[2]

[1] Advanced Technologies Application Center, Havana, Cuba
[2] University of Avignon, LIA, France
{gsierra,jcalvo}@cenatav.co.cu,
jean-francois.bonastre@univ-avignon.fr

Abstract. In recent years a simple representation of a speech excerpt has been proposed, as a binary matrix 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 temporal information representation based on sequential changes present in the binary representation. A new temporal information is proposed in order to add it to speaker recognition systems. A new specificity selection approach using a mask in the cumulative vector space is also proposed. This aims to increase effectiveness in the speaker binary key paradigm. The experimental validation, done on the NIST-SRE framework, demonstrates the efficiency of the proposed solutions, which shows an EER improvement of 7%. The combination of i-vector and binary approaches, using the proposed methods, showed the complementarity of the discriminatory information exploited by each of them.

Keywords: speaker recognition, binary key representation, accumulative vector, temporal information.

1 Introduction

For the last years speaker identity information has been modeled using Gaussian Mixture Model (GMM)/Universal Background Model (UBM) paradigm [1]. 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 basis. In this framework, each speech excerpt is represented by a vector obtained from the concatenation of the means of the Gaussian components [2]. More recently, two major evolutions were proposed in the supervector framework: Joint Factor Analysis (JFA) [3], and i-vector [7].

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:

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 207–213, 2014.

- It is difficult to work with temporal speech information because each set of acoustic vectors is represented only by a point in the supervector or i-vector space. Which makes it impossible to take into account the continuous changes of the vocal tract, which are also discriminatory features of the speaker.
- The underlined paradigm is the statistical one, where the influence of specific information is mainly gathered by the frequency of this information. I.e.: if an event often occurs 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 speakers.

Alternatively, a simple representation of speech which shifts from a continuous probabilistic workspace to a binary discrete space was proposed in [4] and [5]. 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.

Due to this binary matrix representation of a speech excerpt, the speaker discriminant temporal information could be used as shown in [5], [6]. This work aims at adding new temporary information of the speech in the supervector space by using a very simple transformation process of the binary matrices. Each element of a new vector accumulates the state changes of the corresponding Gaussian component in their binary representation.

The main contributions of this work are: a method able to archive in a cumulative vector, frame by frame, the transformations that occur in a binary representation of the speech excerpt and a mask capable of selecting the most discriminatory specificities in the cumulative space. All results improve the speaker recognition performance.

Another result is the demonstration of the complementary information between binary and i-vector approaches by means 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, section 3 and 4 describe the proposed methods, temporal information representation and specificities selection; section 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 ([4], [5]), but all under the same philosophy.

We will use the GM proposed in [5]. 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).

GM allows the system, frame by frame, to perform a transformation $F : \mathbb{R}^d \rightarrow \mathbb{N}^E$ of d-dimensional acoustic frames to a high dimensional binary space ($E >> d$). 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 specificity.

The comparison criteria between two speakers A and B is defined as Intersection and Symmetric Difference Similarity (ISDS) as proposed in [5]. This similarity uses the CV of each speaker.

3 New Temporal Information Representation

In [5] we present a method for extracting the temporal information by blocks (Trajectory model), archiving the information for each block in a cumulative vector. This idea presented advantages for example: a cumulative vector in a segment of the utterance reflects a distribution of specificities related to the phonetic and prosodic contents of the segment, which provides suprasegmental information for speaker recognition. Moreover, an adequate segmentation and overlapping of speech could reduce the effect of noise and increase the robustness of the whole model. But the Trajectory model is not able to capture the temporal frame-level information.

The idea of this work arises from the following hypothesis, if we have a temporally ordered binary representation, containing global information of each frame of the utterance, it is possible to extract temporal frame-level information, which comes from continuous changes in vocal tract and enrich discriminatory characteristics of each speaker.

The new approach to obtain the temporal information uses the binary matrix resulting from the processing of a speech excerpt by the GM. The intention is to count the changes from one frame to another in the binary matrix, always sequentially throughout the utterance and weight each change by the importance of the frame, which is represented by the amount of changes archived on it.

Let $X_s = [x_1, x_2, ..., x_T]$ be a binary matrix obtained from a transformation of a speech excerpt of the s-speaker using the GM. Each column represents a binary vector and the rows represent their E specificities. Then, the temporal information is obtained by:

$$TV[i] = \sum_{j=1}^{T-1} (D_{i,j} * W_j), \tag{1}$$

where $D_{i,j}$ with $i = 1, 2, ..., E$ and $j = 1, 2, ..., T-1$, is a binary matrix that contains the changes between frames of the corresponding specificity defined by $D_{i,j} = |X_{i,j} - X_{i,j+1}|$ and the weight $W_j = \sum_{z=1}^{E} D_{z,j}$.

Then we obtain a *temporal vector TV* with E-dimensions, able to contain the temporal information of speaker utterance frames. Finally, as shown in the eq. 1, the weight of the frame obtained by the changes between a binary vector and next vector, is incorporated.

4 Information Selection (Mask)

The mask for cumulative vectors 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 does not have interesting information.

Let $X = [x_1, x_2, ..., x_S]$ be a matrix of CVs obtained from a population of S impostor speakers. Each column represents a CV and the rows represent their E specificities. Then, the mask is obtained by:

$$mask_j = \begin{cases} 1 \text{ if } \frac{1}{S} \sum_{i=1}^{S} (X_{j,i} - \overline{x}_j)^2 \geqslant \theta \\ \\ 0 \qquad \text{otherwise} \end{cases}$$

where $\bar{x_j}$ is the mean value of the corresponding specificity j with $j = 1, 2, ..., E$, θ is a specificity selection threshold, and $mask$ is the result: a boolean vector where the coefficients related to specificities with variances greater than θ are set to 1 and the others are set to 0.

The variance threshold θ was calculated in the following manner:

1. Specificities are sorted in descending order in terms of variance values.
2. Sets of specificities of greater variance than a value are selected.
3. Using the specificities selected, some speaker recognition experiments with the measure proposed by [5] are performed.
4. The threshold is selected where the effectiveness reflects a minimal value.

In order to obtain the mask, more discriminating specificities were selected with variance greater than θ. The mask has E elements and k selected specificities, $k < E$. Then, the use of the mask consists in selecting the CV or TV values in the indices where the mask is set to 1, resulting in a vector with less dimensions and more discriminative, as we will see in experimental section.

5 Experiments

In all the experiments presented in this section, the front end uses 19 linear frequency cepstral coefficients (LFCCs) together with energy, delta and delta-delta coefficients, giving 60-dimensional feature vectors. A unique GM is used for

the binary-based systems. It is based on a UBM with 512 Gaussian components trained on the NIST SRE 2005 database. The specificity models are derived from the UBM via a selection of the mean vectors, which are obtained through a MAP adaptation of the NIST SRE 2005 speech utterances. The dimension of the CV and TV generated by GM is given by the amount of specificities $m = 36872$.

For the frame binarization process, the top 3 UBM components are selected and the corresponding specificity coefficients with a posterior probability greater than 0.001 are set to 1.

In order to fix the variance threshold in the mask, $\theta = 0.11$, we used CV obtained from 2450 multilingual signals of 124 speakers taken from the NIST SRE 2004. Speaker recognition experiments were evaluated with the NIST SRE 2008 short2-short3, condition det 7, to validate the threshold. The mask has E elements ($E = 36872$), then the specificities of greater variance of cumulative vectors are selected using the mask, resulting in $k = 24000$ dimensions.

Text-independent speaker detection experiments are performed based on the NIST SRE 2008 protocol, male speakers, condition det 7 (telephone-telephone English conversations). This condition uses 1270 speakers (short2) and 671 test utterances (short3); 6615 verifications are performed, including 439 target tests; the rest are non-target tests.

For evaluation of the methods presented cosine and variability compensation technique Probabilistic Linear Discriminant Analysis (PLDA) are used as similarity measures. To train the matrices of LDA projections and PLDA, are used NIST SRE 2004 and 2005.

6 Results

All performance results presented in this section are expressed in terms of Equal Error Rate (EER) and classical Minimum Detection Cost Function (MinDCF).

We first evaluated the impact of the variance-based specificities selection approach (mask), proposed in section 4. Table 1 presents the results of the binary-key framework without and with the mask respectively. ISDS method, proposed by [5] is used as a similarity measure reference on cumulative vector.

Table 1. Speaker recognition using cumulative vectors CV, without and with the mask

Mask	score	DCF*100	EER%
no	$ISDS$	5.5	12.5
yes	$ISDS$	4.8	11.6

The results show an improvement with fewer dimensions ($24000 < 36872$). This represents a reduction of 35% of the specificities. If we focus on Table 1, we can see that the mask improves the effectiveness of the algorithm, because the specificities that were removed do not contain discriminatory information but contain information detrimental to the classifier performance. The rest of the experiments will be performed using this mask.

The main results of the experiments are presented in Table 2, where the performance of the new representation containing temporal information (TV) and global information (CV) is evaluated. The LDA to compensate the session variability is used and two scoring methods cosine and PLDA are evaluated.

Table 2. Evaluating representations CV and TV

repr.	extract.	compensation and scoring	DCF*100	EER%
CV	GM	LDA+cos	5.64	10.50
CV	GM	LDA+PLDA	**2.76**	4.92
TV	GM	LDA+cos	5.36	10.45
TV	GM	LDA+PLDA	2.78	**4.77**
Score fusion of CV and TV				
(CV:LDA+PLDA) + (TV:LDA+PLDA)			2.7	4.50

In Table 2, the results show an improvement (3% regarding the best results of both representations) in terms of EER is obtained using the proposed TV approach compared to CV. We can also see a significant improvement in the results obtained by the PLDA regarding the cosine. Finally, the fusion between scores of both representations is presented, obtaining a better result (4.50% EER) which shows that there is complementary information between CV and TV.

Table 3 shows the performance of the current methods on the i-vector approach, looking for a comparison with our approaches and the fusion between the two approaches presented in Table 4.

Table 3. State-of-the-art speaker recognition system

repr.	extract.	compensation and scoring	DCF*100	EER%
i-vector	FA	LDA+cos	2.24	3.80
i-vector	FA	LDA+PLDA	1.89	2.96

Again, the PLDA classifier robustness in front of the cosine is appreciated, although it is seen that the PLDA has a higher computational cost than the cosine.

In Table 4, a score fusion of both paradigms is obtained. An important result is shown (***2.73*** EER), which allows us to conclude that there is complementary information between the two approaches, highlighting even more the importance of the binary representation.

Table 4. Score fusion of i-vector and CV or TV

fusion	DCF*100	EER%
(i-vector:LDA+PLDA) + (CV:LDA+PLDA)	1.85	2.95
(i-vector:LDA+PLDA) + (TV:LDA+PLDA)	***1.80***	***2.73***

7 Conclusions

In this article, we aimed at associating the power of a new speech representation, the Speaker Binary Key, with new temporal information, being a new step in the development of the binary approach, which is impossible to obtain in the supervectors or i-vectors space. In addition, compared to cumulative vectors it has a 3% EER average improvement and its fusion shows the best result obtained (4.50% EER) into the binary framework.

We proposed a new specificities selection method for cumulative and temporal vector, capable of removing one third (12872 of 36872) of specificities of the Generator Model with little or no information, with an increase in effectiveness.

It is important to notice that both approaches, i-vectors and the temporal information representation, contain complementary information, responsible for the improvement (7% of EER) in the score fusion, which reaffirms the importance of including the temporal information in the binary framework as a new speaker discriminative characteristic.

In future studies, we will examine ways to combine Speaker Binary Key and another binary framework, Boosted Slice Classifiers [8], mainly focusing on the selection of the most discriminative binary features.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
2. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-carrasquillo, P.A.: Support Vector Machines for speaker and language recognition. Computer Speech and Language 20, 210–229 (2006)
3. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech & Language Processing 15, 1448–1460 (2007)
4. Anguera, X., Bonastre, J.F.: A novel speaker binary key derived from anchor models. In: INTERSPEECH, pp. 2118–2121 (2010)
5. Hernández-Sierra, G., Bonastre, J.-F., Calvo de Lara, J.R.: Speaker recognition using a binary representation and specificities models. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 732–739. Springer, Heidelberg (2012)
6. Bonastre, J.F., Miro, X.A., Sierra, G.H., Bousquet, P.M.: Speaker modeling using local binary decisions. In: INTERSPEECH, pp. 13–16 (2011)
7. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech & Language Processing 19, 788–798 (2011)
8. Roy, A., Magimai-Doss, M., Marcel, S.: A fast parts-based approach to speaker verification using boosted slice classifiers. IEEE Transactions on Information Forensics and Security 7, 241–254 (2012)