DISCRIMINANT BINARY DATA REPRESENTATION FOR SPEAKER RECOGNITION

J.F. Bonastre, P.M. Bousquet, D. Matrouf

University of Avignon
LIA
Avignon, France

X. Anguera

Telefonica Research
Barcelona, Spain

ABSTRACT

In supervector UBM/GMM paradigm, each acoustic file is represented by the mean parameters of a GMM model. This supervector space is used as a data representation space, which has a high dimensionality. Moreover, this space is not intrinsically discriminant and a complete speech segment is represented by only one vector, withdrawing mainly the possibility to take into account temporal or sequential information. This work proposes a new approach where each acoustic frame is represented in a discriminant binary space. The proposed approach relies on a UBM to structure the acoustic space in regions. Each region is then populated with a set of Gaussian models, denoted as "specificities", able to emphasize speaker specific information. Each acoustic frame is mapped in the discriminant binary space, turning "on" or "off" all the specificities to create a large binary vector. All the following steps, speaker reference extraction, likelihood estimation or decision take place in this binary space. Even if this work is a first step in this avenue, the experiments based on NIST SRE 2008 framework demonstrate the potential of the proposed approach. Moreover, this approach opens the opportunity to rethink all the classical processes using a discrete, binary view.

Index Terms— Discrete, discriminant, binary, speaker recognition

1. INTRODUCTION

Speaker recognition main approaches are based on statistical modeling of the acoustic space. This modeling relies usually on a Gaussian Mixture Model (GMM), denoted as Universal Background Model (UBM), with a large number of components and trained using a large set of speech data gathered from hundreds of speakers. The UBM was originally seen as a seed to obtain the client speaker models. Each target model was derived from the UBM thanks to a MAP adaptation of the Gaussian mean parameters only [1]. An important evolution of the UBM/GMM paradigm was to consider the UBM as a definition of a new data representation space defined by the concatenation of the Gaussian mean parameters [2]. This space, denoted as "supervector space", allowed us to use Support Vector Machines (SVM). A second evolution step was crossed by the direct modeling of the session variability in the supervector space using the Joint Factor Analysis (JFA) approach (the Nuisance Attribute Projection is similar, for SVM/GMM systems[3]).

All the approaches derived from these evolutions follow the same scheme: a reparameterization step which takes as input a set of acoustic vectors and outputs a single vector representing this set.These approaches demonstrated their potential but show also two main limitations. First, as a set of acoustic vectors are represented by an unique point in the targeted space, it is difficult to exploit temporal or sequential information. It remains possible to work on sub-segments of a recording but it is difficult to obtain a strong estimate of the underlined statistics when the set of acoustic frames is small. Second, they rely on the concept of global and general information: an information is important because it appears frequently. The supervector space doesn't take into account intrinsically a speaker's discriminant aspects, which is the goal of speaker recognition. Building such a discriminant space remains difficult as a continuous probabilistic modeling needs a large amount of data to train the model parameters. In practice, the statistic estimation is done following the data missing case and a discriminant approach always increases this problem.

In this paper, we propose a solution able to answer the limitations of supervector based approaches, and more generally of UBM/GMM based approaches. We propose to move from a continuous probabilistic space to a discrete, binary space, able to handle directly the speaker discriminant information. A binary space offers several advantages. It allows to work with a large dimensionality but keeping a compact representation as one coefficient is coded using one bit. A binary representation allows to use specific arithmetic which are known for their computational efficiency. Using a binary representation allows also to go further in the direction of a (speaker) discriminative space. The data missing problem highlighted before is decreased as, for a given direction of the space, only the presence or non presence of a given specificity is evaluated.

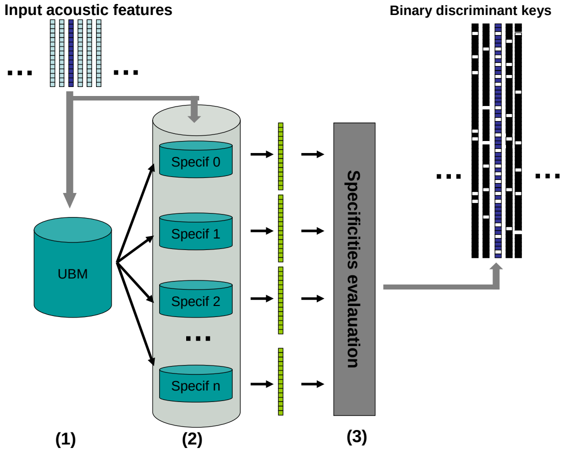

Fig. 1. Overview of the approach

2. METHOD OVERVIEW

Figure 1 presents an overview of the proposed approach, which is mainly composed of three elements. First, a classical UBM is used to structure the acoustic space in sub areas. The role of this UBM is only to tie each input frames with one or several Gaussian components, i.e. one or few acoustic regions. Second, each acoustic region is then populated with a set of Gaussian models, denoted as "speaker specificities". Each of these models emphasizes a given user specific information for this particular acoustic space region. These models are gathered from a training set comparable to the one used for the UBM training (a model is trained from data related to an unique speaker, in order to emphasize the specific information corresponding to this speaker). Finally, for an acoustic frame, each speaker specificity is evaluated and a corresponding binary value is set in the output vector.

Figure 2 shows how a region of the acoustic space defined by one of the UBM component is described by a large set of specificities (Gaussian components) organized in order to emphasize the discriminant information. It also shows how the input data is projected in the discrete, binary, space. The input data is associated with one binary value per specificity. Only the specificities close to the input data are associated with a value equal to 1 and the other with a value equal to 0. This process is expected to offer a large noise robustness as illustrated on Figure 2 by the brown (top) and green (down) parts. It presents the input data with and without noise and shows that the output binary information is very similar. This process is repeated for all selected subregions of the data space (i.e. all the in-interest areas corresponding to the selected UBM components, for the targeted input data). Over all, the main advantage of the proposed approach is its ability to finely describe the discriminant information present in each input data, with a large noise robustness. The proposed approach inherits a part of its concepts from the anchor model approach [4, 5]. A first implementation of the approach was proposed in [6]. Even if our approach is also close to [7, 8] which explored the posterior probabilities of mixtures as tokens. It differs from these works on two main aspects: the binary quantization and the intrinsic discriminant nature of the acoustic modeling.

The binary key generation is the core part of the proposed approach. It corresponds to a reparameterization step able to project an acoustic vector, as well as a set of acoustic vectors, into a (large) binary space. The proposed parameter space is designed in order to emphasize the individual speaker specificities (i.e. it is intrinsically a speaker discriminant space).

The binary key generation follows the UBM/GMM paradigm in order to take advantage of the statistical modeling power of this approach. It is composed by three steps: an acoustic space structuration based on a classical UBM, a speaker specificities model, denoted as generator model, and a binary key extraction module denoted as extractor. The following paragraphs describe the three parts of the binary key generation.

3. KEY GENERATOR

The first module of the binary key generator follows the classical UBM/GMM approach. It uses an UBM trained classically to structure the acoustic space. Using this UBM the *a posteriori* probability of all the UBM components are computed for each acoustic frame. These probabilities will be used by the generator and the extractor.

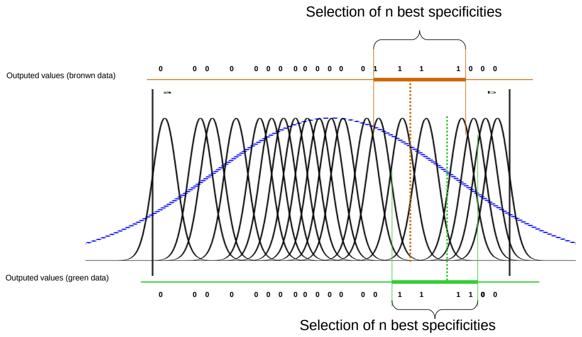

Fig. 2. Illustration of the input data projection in the binary space

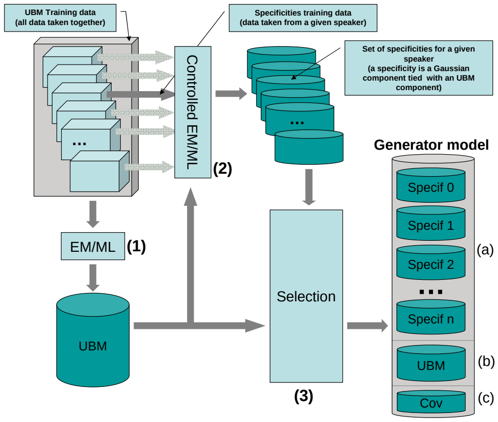
UBM Training data
(all data taken together)
Specificities training data
(data taken from a given speaker)
Set of specificities for a given speaker
(a specificity is a Gaussian component tied with an UBM component)
...
Controlled EM/ML
(2)
...
Generator model
Specif 0
Specif 1
Specif 2
...
Specif n
(a)
UBM
(b)
Cov
(c)
EM/ML
(1)
UBM
Selection
(3)

Fig. 3. Generator model

3.1. Generator model

The main role of the *generator* is to highlight the speaker specificities even if these speaker specificities are not present on average between the speakers. Of course, examples of these speaker specificities should be present in some training data. The basic idea to achieve this goal is to work on speaker dependent statistics when the UBM is working on speaker independent ones.

Figure 3 presents an overview of the *generator* model training process and of the model itself. In order to obtain the model, three phases are needed.

The first phase corresponds to a classical UBM training based on EM/ML algorithm. For the second training phase, the UBM training data set is decomposed in subsets. Each subset corresponds to one unique speaker. A speaker specific model is computed by EM/ML, initialized from the UBM. Unlike in the classical UBM/GMM speaker model estimation, mean and covariance parameters are computed, only the weights remain unchanged. For each UBM component i, a set of specificities E^i_j is obtained. A given specificity is a Gaussian component defined by the mean $\mu_{E^i_j}$ and the covariance matrix $\Sigma_{E^i_j}$. Then, an unique covariance matrix, Σ^i_{intra}, is computed for the set of specificities related to UBM component i as the average of $\Sigma_{E^i_j}$.

The covariance matrices are shared by all the specificities belonging to a given UBM component. All the available covariance matrices are averaged in order to represent intra-speaker variability. The third training process is a specificity selection. It takes as input the UBM, the set of speaker specificities and the averaged covariance matrices. The total number of speaker specificities is assumed to be very large. As this number drives directly the dimension of the discrete binary space, it is necessary to reduce it by selecting the most interesting specificities. This is done by a Maximum Relevance Minimum Redundancy algorithm. For each of the UBM component, the corresponding specificities are selected iteratively following the criterion:

$$arg_i max(\frac{\frac{1}{n}\cdot\sum_{j=1}^{j=n} d(S_i, E_j, \Sigma_{intra})}{d(S_i, ubm, \Sigma_{intra})}) \qquad (1)$$

where S_i is the next selected specificity, E is the set of selected specificities at the iteration i, $d(S_i, E_j, \Sigma_{intra})$ is a distance between two Gaussian components described respectively by the mean vector S_i and E_j and associated with the covariance matrix Σ_intra. The algorithm is initialized by selecting the closest specificity v.s. the corresponding UBM component.

Finally, the *generator* model is composed of: (a) a set of specificities for each of the UBM components, (b) the UBM itself and (c) the averaged covariance matrix.

3.2. Extractor

The role of the last part of the key generator, the extractor, is to project the acoustic data into the discriminant binary representation space. The extractor has to work at the frame level as well as at all the possible segmental levels. The extractor is composed by two modules. The first one takes as input the generator model and a unique acoustic frame and outputs a binary representation of this frame. The second one takes as input a set of binary vectors and outputs a unique binary vector. The two following paragraphs give details on the two modules.

In order to project an acoustic vector in the binary space, the a posteriori probabilities p^i_{ubm} of the input frame f_t for all the UBM components are classically computed by:

$$p^i_{ubm}(f_t) = \frac{w^i_{ubm}\cdot l(f_t|\mathcal{N}(\mu^i_{ubm}, \Sigma^i_{ubm}))}{\sum_{j=1}^{j=n} w^j_{ubm}\cdot l(f_t|\mathcal{N}(\mu^j_{ubm}, \Sigma^j_{ubm}))} \qquad (2)$$

where w^i_{ubm} is a priori probability of the UBM component i (the mixture weight) and $\mathcal{N}(\mu^i_{ubm}, \Sigma^i_{ubm})$ is the ubm component i.

Then, mainly for computational efficiency, a component selection is classically performed by selecting only the components with the highest $p^i_{ubm}(f_t)$. For each selected UBM component, the likelihood of the acoustic frame f_t is computed for all the corresponding specificities (generator model, part (a)) thanks to:

$$l(f_t|E^i_j) = p^i(f_t)\cdot l(f_t|\mathcal{N}(\mu_{E^i_j}, \Sigma^i_{intra}) \qquad (3)$$

where $p^i(f_t)$ is the probability to have f_t tied to the UBM component i, E^i_j is the specificity j corresponding to the UBM component i. E^i_j is the specificity model defined by the mean $\mu_{E^i_j}$ and the covariance matrix Σ^i_{intra}. Several options are possible in order to estimate $p^i(f_t)$. In this work it is simply the a priori probability of the UBM component i. Finally, the output binary vector is computed by setting a 1 for the specificities with the highest $l(f_t|E^i_j)$ and a 0 for the others, including the non computed specificities (the specificities tied to a non selected UBM component are not computed). The number of coefficients set to 1 per vector could be dynamically determined or fixed as a meta parameter. The latter solution was retained for this work. It is important to notice that a fixed number of selected components is not like a threshold in the likelihood area.

After the previous step, a binary representation is available for each input acoustic frame. This binary vector could be used directly in order to propose a complete frame per frame representation of the input acoustic file. But it is also possible to compress the information in the time dimension, by applying a segmental representation with or without overlapping. The maximum compression is achieved when a complete input sequence of acoustic frames is represented by an unique binary vector. This solution was retained in this work. The time compression is simply obtained by a majority voting: a 1 value is set for the specificities with a maximum of votes (per-frame vector with a corresponding value set to 1) and a 0 value for the others. Similarly than for the previous module, a variable or a fixed number of coefficients set to 1 could be used, here a fixed one is proposed.

4. SPEAKER RECOGNITION SYSTEM

One of the main interest of the proposed approach is to open a new avenue in the design of speaker recognition system. With the binary discriminant representation data space, it is important to rethink all the steps currently used in speaker recognition systems, like session variability modeling, speaker models, score computation and score normalization or calibration. As it is possible to project each input frame, it is also possible to work on frame sequential information. Here, as a maximum compression is used at the key generator level, a speech excerpt is represented by a simple binary vector, able to emphasize the speaker specificities. This vector is itself a speaker model, denoted "binary key". Then, a simple distance between two binary vectors is enough for computing a similarity score. In the Information Theory area, several distances are available like Jaccard, Ghosh, Sokal & Michener, Sokal & Sneath and Yule criteria. In this work, we are using a criterion derived from the Sokal & Michener one:

$$S_{sm}(\mathbf{v}_{f1}, \mathbf{v}_{f2}) = \frac{P_c}{P} \qquad (4)$$

where P_c is the number of corresponding 1 in the two binary keys and P is the dimension of the binary keys.

5. EXPERIMENTS

All the results reported in this paper are evaluated on the NIST SRE08 ([9]) short2-short3 condition. This condition takes one session of the target speaker for enrollment and one session for testing. Short2-short3 is divided into several conditions and we are only interested in the male condition 7 with trials involving only English language telephone speech in training and test. In this condition, 470 target speakers and 638 tests segments are used to perform 6616 verification tests. Results are reported in terms of Equal Error Rate (%EER) and described by DET curves. The baseline system as well as the UBM/GMM functionalities are gathered from [10].

Figure 4 reports the performance of a system based on a 128 components UBM and a set of 256 specificities associated to each components. For the key generation, 3 UBM components are selected and 32 top specificities (for each selected UBM component) are set to 1 at the frame level when 2000 specificities are set to 1 at the file level. No score normalization or Factor Analysis (FA) based session variability modeling is used. The EER is about 12% which is far to the performance obtained by our baseline 512 components UBM/GMM system (about 8% of EER without score normalization and FA and about 4% of EER when session normalization is used). We tried also preliminary experiments with an adapted NAP procedure which works in our discrete space. Figure 5 shows the

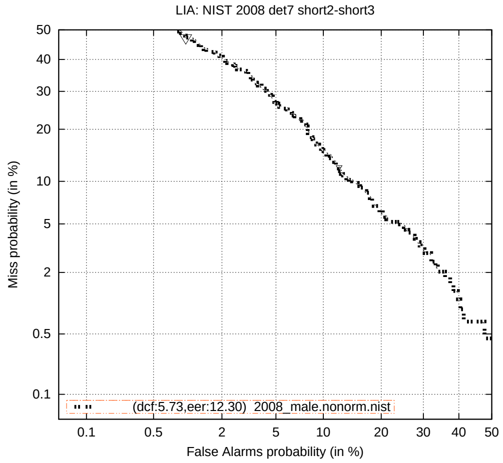

Fig. 4. DET of the binary system (using 128 components UBM and 256 specificities per component), NIST 2008 male, det 7 condition

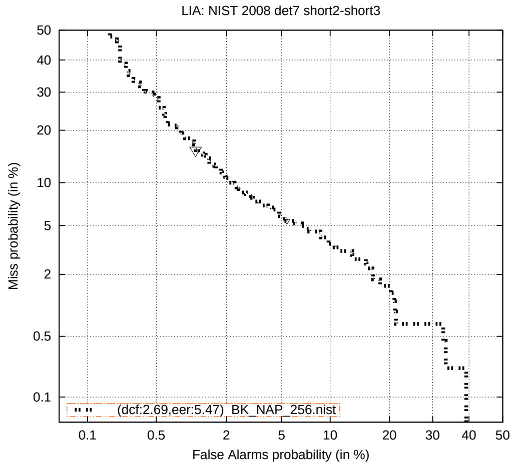

Fig. 5. DET of the binary system (using 128 components UBM and 256 specificities per component) when NAP is applied in the binary space, NIST 2008 male, det 7 condition

corresponding DET curve, still without score normalization. Our approach obtains about 5% of EER to be compared with about 4% for our best UBM/GMM system when a Factor Analysis is used to compensate for session variability.

6. CONCLUSIONS AND FUTURE WORKS

In this work, we proposed a new paradigm for speaker recognition which projects the acoustic frames in a binary space built in order to emphasize the speaker specificities. The strong points of the proposed approach are that it is able to deal with the frame sequence, to concentrate all the acoustic related processes in one module which is applied on train and test speech data, and that it works in a discriminant binary space. The latter point opens also the possibility to rethink the session variability modeling as well as the decision or score normalization modules in the view of such a binary data representation space.

The proposed approach obtained about 12% of EER. This level of performance remains acceptable compared to our baseline systems (respectively, 8% and 4% for the UBM-GMM without and with Factor Analysis) as it is only a first version for our approach, with few tuning or optimization. When a session variability modeling is applied, in the binary space, the error rates decreased significantly, until about 5% of EER, which is close to our best performance using a classical UBM/GMM system with a JFA-based session variability modeling (4% of EER).

Several points of the system should be optimized, like the UBM, the way to train the specificities and to select the best ones. The links between the session variability modeling and the binary space building should also be explored in the next future. Finally, such an approach opens a large room for research in several directions. We expect to come back to the temporal/sequential information which is known to be important for speaker characterization. We will also try to use the specific nature of our system to look not only at the performance but in order to explain how and why given results are obtained, for example, by coming back to the phonetic information.

7. REFERENCES

[1] F. Bimbot, J-F Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, April 2004.

[2] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, vol. 1, pp. 637–640.

[4] T. Merlin, J-F. Bonastre, and C. Fredouille, "Non directly acoustic process for costless speaker recognition and indexation," in *Workshop on Intelligent Communication Technologies and Applications*, 1999.

[5] Y. Mami and D. Charlet, "Speaker identification by location in an optimal space of anchor models," in *ICSLP-2002*, 2002.

[6] X. Anguera and J.F. Bonastre, "Novel binary key representation for biometric speaker recognition," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (interspeech'2010)*, 2010.

[7] N. Scheffer and J.-F. Bonastre, "Ubm-gmm driven discriminative approach for speaker verification," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 2006, pp. 1 –7.

[8] P.A. Torres-Carrasquillo, D.A. Reynolds, and Jr. Deller, J.R., "Language identification using gaussian mixture model tokenization," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 2002, vol. 1, p. I.

[9] A. F. Martin and C. S. Greenberg, "NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels," in *International Conference on Speech Communication and Technology*, 2009, pp. 2579–2582.

[10] J. F Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, Pouchoulin G, and Evans N, "ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition," in *Speaker Odyssey Workshop*, 2008.