Speaker modeling using local binary decisions

Jean-François Bonastre[1], *Xavier Anguera*[2], *Gabriel H. Sierra*[1,3],*Pierre-Michel Bousquet*[1]

[1]University of Avignon, LIA, Avignon, France
(jean-francois.bonastre,pierre-michel.bousquet)@univ-avignon.fr
[2]Telefonica Research, Barcelona, Spain
xanguera@tid.es
[3]Advanced Technologies Application Center, Havana, Cuba
gsierra@cenatav.co.cu

Abstract

Achieving an accurate speaker modeling is a crucial step in any speaker-related algorithm. Many statistical speaker modeling techniques that deviate from the classical GMM/UBM approach have been proposed for some time now that can accurately discriminate between speakers. Although many of them imply the evaluation of high dimensional feature vectors and represent a speaker with a single vector, therefore not using any temporal information. In addition, they place most emphasis on modeling the most recurrent acoustic events, instead of less occurring speaker discriminant information. In this paper we explain the main benefits of our recently proposed binary speaker modeling technique and show its benefits in two particular applications, namely for speaker recognition and speaker diarization. Both applications achieve near to state-of-the-art results while benefiting from performing most processing in the binary space.
Index Terms: speaker modeling, binary keys

1. Introduction

Accurate modeling of the speech uttered by a speaker is of great importance and has been the object of much research in the last few years [1, 2, 3, 4]. The range of possible applications include speaker recognition, speaker diarization or segmentation, and usually the same techniques can be applied to other areas like language ID or emotion recognition.

Traditionally, speaker information has been modeled by using the Gaussian Mixture Model/Universal Background model (GMM/UBM) paradigm [5] where a weighted sum of Gaussian distributions perform a direct acoustic modeling of the acoustic space. The usual approach is to first train a UBM model, through the estimation of a big number of free parameters, using as much data as possible from many different speakers. Then, each target GMM speaker model can be adapted from the UBM using much less data through Maximum a Posteriori (MAP) adaptation of the UBM means or, less commonly, also variances and weights.

More recently, there was a paradigm shift to model the speakers as a long, fixed-length vector. These include the supervectors [2]; the vectorized MLLR transformation matrix [1]; or the mean polynomial expansion of the feature vector in the GLDS SVM kernel [6]. An important evolution of these new models came with the direct modeling of session variability in the supervector space by using the Joint Factor Analysis (JFA) approach [7] (or equivalently the Nuisance Attribute Projection, NAP, in GMM/SVM systems [8]). Derived from the lessons learned from JFA, in [4] a dimensionality reduced version of the supervectors has been proposed, termed ivectors, with just a few hundred components but very good performance.

While the aforementioned techniques have shown extremely good performances, mostly on speaker recognition applications, they pose several limitations that we try to address with our proposed model. First, these modeling techniques rely on statistical distributions which require a large amount of acoustic data to be robustly estimated. Due to this constraint, usually a speaker (or a speech extract) is modeled using only a single vector and no temporal modeling is feasible. Also, both supervectors and ivectors put emphasis on the most frequently recurring acoustic information, taking less into account speaker intrinsically discriminant aspects, which are the core of speaker modeling. Indeed, an informative part of the acoustic space, where there is no frequent but significant speaker specific information, will be missed by the underlined model, the UBM. This negative aspect of UBM/GMM modelling is usually partially by-passed by using large model orders, which increases also the session mismatch and noise sensibility. Finally, most of these techniques involve high dimensional real-valued vectors or big models that are costly to compute and long to store, therefore not very appropriate for embedded applications.

In this paper we review a solution to these problems based on speaker modeling using binary vectors that was initially proposed in [3], and developed in [9, 10]. We focus here on a generic presentation of the underlying modeling technique applied to both speaker verification and speaker diarization tasks. We also detail the reasons why the proposed binary representation can solve the above mentioned problems and propose new extentions and scoring method using a segmental modelling for speaker recognition associated to new experimental results.

The rest of the paper is structured as follows: Next we review the proposed binary speaker modeling algorithm. Then, in section 3 we briefly describe the two different applications of this technology, emphasizing the common usage of the binary modeling technique and explaining the differences due to the particularities of each task. Next, in section 4 we show performance results for both applications and, in the last section, discuss why we believe the proposed technique is worth considering.

2. Overview of binary speaker modeling

The main goal of the proposed binary speaker modeling technique is to obtain, though an acoustic-to-binary transformation, a compact representation of a speaker within the binary space where the discriminability between speakers is maintained (or

enhanced) with respect to the acoustic (real-valued) space. Once in the binary space, several applications can be built to benefit from the ease of comparison, storage reduction and other properties of the binary vectors. Our binary speaker modeling systems are designed around three main blocks, which we describe next.

The first block corresponds to the training of a generator model from acoustic data by structuring the speaker acoustic space into sub-areas thanks to a classical background model and representing each one using a set of Gaussian distributions. These Gaussian distributions are optimized to highlight speaker-dependent characteristics: the generator model embeds intrinsically speaker discriminant aspects. It differs from a classical background model which follows the Maximum Likelihood paradigm. These Gaussian components dedicated to speaker discriminant information are denoted "specificity models" to highlight their discriminant nature. The particular training method to be used for the generator model varies depending on the application and the information available. The generator model is a key element in our binary approach and it usually needs only to be trained once for every acoustic condition we want to apply the system to. While in speaker recognition it is usual to have separate training data for different speakers, in speaker diarization it is usual not to use any external training data, therefore the generator model needs to be trained from the same test data, where speaker boundaries are unknown a priori.

A second block describes how to obtain a binary representation from any input speaker data. It is generally done in two steps. In a first step, a transformation $T : \mathbb{R}^n \rightarrow \mathbb{N}^m$ between an n-dimensional feature vector and an m-dimensional binary vector is defined. The dimension m of the binary vector corresponds to the number of specificity models in the generator model, where each position in the binary vector is linked with one specificity. Such transformation projects each individual acoustic vector into the binary space. In general, the positions set to 1 in the binary vector indicate those specificity models are expected to be present in the given acoustic vector. The selection of the positions (set to 1) is done by a likelihood computation at the specificity model level associated to a relative selection (top-n highest value selection). This selection process could be applied globally on all the specificitiy models or independently for each acoustic sub-area (represented by a background model component). The final number of bits set to 1 per binary vector can be dynamically determined depending on the acoustic characteristics of the vector, or fixed according to a meta-parameter. Note that this parameter is not a threshold in the likelihood area and can be set fairly independently of the final application. The resulting stream of binary vectors obtained for a given input acoustic file represents an exhaustive time representation of the acoustic signal in the binary space. In many applications it is important to compress such representation into a more compact form, which is achieved in a second processing step by applying a segmental representation, with or without overlap. For any given segment of binary vectors, a single resulting binary vector is obtained by majority voting: a value 1 is set for the vector locations with most number of 1 in the sequence of binary vectors, while a value of 0 is given otherwise. Similarly to the first step, the number of 1 values can be set dynamically or set to a fix value. The unique resulting binary vector is used as the model for that particular speaker.

Finally, a third block defines a distance between any two speakers by comparing their binary speaker models. Given that these models reside in the binary domain, the computation of a similarity score gets reduced to selecting the appropriate distance between two binary vectors. As an example, Equation 1 is the similarity used for speaker verification and Equation 2 is used for speaker diarization.

$$S_v(x,y) = \frac{\sum_{i=1}^{N}(x[i] \wedge y[i])}{N} \quad (1)$$

$$S_d(x,y) = \frac{\sum_{i=1}^{N}(x[i] \wedge y[i])}{\sum_{i=1}^{N}(x[i] \vee y[i])} \quad (2)$$

where $\wedge$ indicates the boolean AND operator and $\vee$ indicates the boolean OR operator. By definition, both Equations 1 and 2 can take values in $[0, 1]$, where the bigger the value the more similar both speakers are.

3. Binary speaker modeling applications

Next, we explain the particular implementation of the binary speaker modeling approach for two different applications, namely for speaker verification and for speaker diarization. While the goals and particular challenges in both areas are different, they both need to model speakers to differentiate them from each other.

3.1. Speaker Verification

The task of speaker recognition involves a two-step process. First, a speaker model is learned from speech material pronounced by a given speaker. Then, a test acoustic utterance is compared to one speaker model in order to decide if the corresponding speaker has produced the utterance.

The first step consists in training the generator model, necessary to perform the transformation of acoustic data into the binary space. This step is done once using speakers data from as many speakers as possible. First, a GMM/UBM model is trained in a standard way using the EM/ML algorithm with all training data together. Then, a set of GMM speaker-dependent models is trained, taking independently the corresponding data for each available speaker (via one iteration of EM/ML, initialized from the general GMM/UBM model)[1]. The n Gaussian components of each speaker dependent GMM are grouped together to constitute the n specificity models set. Finally, the generator model is constructed by selecting, for each UBM component (n), a subset of specificity models that optimizes the coverage of the speaker acoustic space while individually discriminating among individual speaker characteristics. This is done by using a Maximum relevance, Minimum Redundancy algorithm. For a more detailed explanation refer to [10]. Note that all the generator model training is performed on training data (and speakers) others than those used for the performance evaluation (or the system real life).

The second step consists in the transformation of a speech file into the binary space using the generator model. It is performed efficiently by using the relationship (inside the generator model) between the GMM/UBM model and the sets of specificity models. For any given acoustic vector we first classically select the top GMM/UBM components, and then compute the likelihood values only for the corresponding specificity models. A final binary vector for a given input frame is constructed setting to 1 those locations associated to the specificity models with the highest likelihood values. This process is performed independently for each set of specificity models (i.e. for each selected GMM/UBM component). A binary key can be then constructed from a set of frame-binary vectors (corresponding

[1] In this case only means and variances are estimated, keeping the weights fixed to those in the initial model.

to a complete speech excerpt or to a temporal window) as described above.

The last step is the decision one. In the classical processing, the decision is directly computed as a distance between two binary vectors, each representing a speech excerpt (train and test) thanks to Equation 1. Additionally, some feature normalization techniques like NAP can be applied in the binary space to enhance the discriminability of the trained models, obtaining enhanced results, as will be shown in the experimental section. An alternative scoring is also proposed in order to explore the potential of our approach to take into account segmental aspects. A given speech file s is segmented into n_s temporal windows (possibly overlapped) and a binary vector, Bs^t is extracted (similarly than previously) for each window t. In addition, the global binary vector, Gs, representing the entire speech file is computed. The final segmental score is:

$$WS_v(x,y) = \frac{1}{n_x}\sum_{i=1}^{n_x} S_v(Bx_i, Gy) + \frac{1}{n_y}\sum_{i=1}^{n_y} S_v(By_i, Gx) \qquad (3)$$

3.2. Speaker diarization

Given an acoustic input signal containing speech from multiple speakers, the objective of a speaker diarization system is to split it into speakers. When performing such task, neither the number of speakers nor their identity is usually known a priory. In addition, most state-of-the-art systems avoid using any outside acoustic information and try to obtain speaker models from the test data itself through an iterative processing. Here we describe a bottom-up agglomerative speaker diarization system that follows [11] but uses the proposed binary approach.

Unlike in speaker recognition, here the generator model needs to be trained directly from the input data, which contains several intermixed speakers. Therefore, a different method is used to obtain a set of discriminative Gaussians. In a first step, a pool of Gaussians is computed by estimating a single Gaussian for every 2 seconds of data, with certain overlap between consecutive segments. Next, an iterative method is used to select the final set of m Gaussians in the generator model as follows: the first selected Gaussian is the one that most likely models the data used for training. Then, a single linkage clustering approach is followed to iteratively add to the selected set the Gaussians, among the remaining Gaussians in the pool, that have at every step the largest KL2 distance to the closest Gaussian among those already selected. A final step performs a standard EM/ML training iteration only on the Gaussian weights, by using all input data.

The speaker Diarization iterative agglomerative clustering of the input acoustic data is then performed as follows: the input data is first split into K initial clusters either homogeneously or using the obtained generator model. For each cluster the generator model is used to find its equivalent binary model. Then we iteratively merge the two clusters that are closest together according to Eq. 2 and reassign the input data into the closest cluster at the time. When reaching one cluster, the optimum number of speakers (and output clustering) is selected using a modified T-test metric inspired in the method proposed in [12]. For a more detailed explanation of the system refer to [9].

From the steps above, it is worth noting that the data reassignment step is performed sequentially on every 3 seconds of data by computing the segment's binary speaker model and comparing it to each of the current clusters using Eq. 2. Therefore in this application we show the capability of the proposed binary speaker modeling of using the temporal characteristics of the input data without losing in discriminatory power.

4. Performance evaluation

In this section we show experimental results both for the speaker recognition application and for speaker diarization using the proposed binary speaker models.

Table 1 reports the results of three configurations of the binary approach plus the baseline UBM/GMM (512 components), evaluated on the NIST SRE08 ([13]) short2-short3 condition, DET 7 (one session for enrollment and one session for testing, English language telephone speech in training and test, 470 target speakers and 638 tests segments are used to perform 6616 verification tests). The baseline system as well as the UBM/GMM functionalities are gathered from [14]. No score normalization or Factor Analysis (FA) based session variability modeling is used. All the binary configurations rely on a 128 components UBM and 256 specificity models per component (giving a binary vector dimension of 32768). For the frame binarization, top 3 UBM components and the highest 32 corresponding specificity models are set to 1. In configuration 1, each speech file is represented by one binary vector and the score is based on Eq. 1. In configuration 2, Eq. 3 is used for scoring, associated to temporal windows of 300 frames, with an overlap of 150 frames. In configuration 3, an adapted NAP procedure which works in our discrete space is applied on the configuration 1 system. For comparison, the performance of our best system during NIST 2010, an UBMGMM-SVM with a simplified version of JFA (eigenchannels) is also provided.

Table 1: *Performance of speaker recognition systems*

System	DCF(*100)	EER %
(1) one binary vect. per file	5.73	12.30
(2) one binary vect. per 300 frames window	5.69	11.16
(1)+NAP	2.69	5.47
UBM/GMM	3.54	7.74
UBM/GMM-SVM with FA	1.55	2.73

In speaker diarization the most commonly used evaluation metric is called diarization error rate (DER) and computes the percentage of evaluated time that has been incorrectly labeled, either attributing it to the wrong speaker or classifying it incorrectly between speech and silence. In the experiments shown here we used all datasets available from the NIST Rich Transcription meeting evaluation campaigns, which consists of 4 sets of meeting excerpts, totaling 32 recordings that contain from 4 to 10 (unknown a priori) speakers and are recorded in various unknown acoustic conditions. As a baseline algorithm we use a GMM-based agglomerative bottom-up diarization system as described in [11]. Table 2 shows the DER values and the real-time running factor of both systems. The speaker binary system was executed using $m = 896$ Gaussians in the generator model.

Table 2: *Comparison speaker diarization results on DER and real-time factor*

System	RT05	RT06	RT07	RT09	xRT
Acoustic Baseline	24.96	24.32	17.39	26.80	1.19
Binary speaker models	27.75	27.48	20.02	29.05	0.175

5. Discussion

For both applications, the novel binary approach proposes an interesting level of performance even if the error rates are equivalent or higher than the corresponding baseline. This result

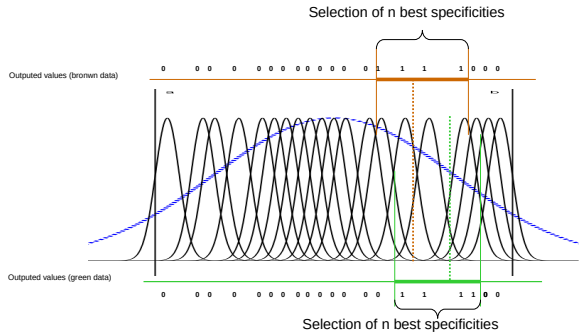

Figure 1: Illustration of the input data projection in the binary space

could be seen as disappointing but it comes from a first implementation of a new paradigm, with very little parameter optimization. Moreover, the generator model, which is the core of the approach, is trained using a straightforward criterion which should be revisited in the focus of binary or discrete statistical approaches.

However, the NAP-based configuration for speaker recognition obtains about 5% of EER to be compared to about 8% for the baseline and about 4% for our best NIST 2010 system when a Factor Analysis is used. The NAP-based result shows clearly that, despite the current limitations, the binary space has at least the same discrimination power than the supervector space. The binary approach demonstrates also its ability to work at a lower level than a file. For speaker recognition slightly better results are obtained when the score are based on the average of 300 frame-based binary vectors, compared to the file-file distance. Our diarization approach obtains similar results to the equivalent acoustic GMM-based system while running around 10 times faster.

One of the expected advantages of the proposed approach, as mentioned in the introduction, is to allow to model precisely the speaker specificities even if the related acoustic events are not frequently observed. With an usual UBM/GMM system, the only way to achieve this goal is to increase the number of components. Moreover, it is well known that a limit (about few thousands of components) is reached in UBM size, as both the quality of the likelihood estimation and the robustness to noises and mismatches decreases. Figure 1 illustrates how a region of the acoustic space defined by one of the UBM component can be described in the binary approach 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 1 by the brown (top) and green (down) parts, which present the input data with and without noise and show that the output binary information is very similar. This characteristic of our approach is expected to allow a fine and robust description of the discriminant information present in each input data.

6. Acknowledgements

This work was done in collaboration between Telefonica and University of Avignon. The proposed approach is the object of the filled patent proposal FR 10/57732.

7. References

[1] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "Mllr transforms as features in speaker recognition," in *in Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 2425–2428.

[2] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308 – 311, May 2006.

[3] X. Anguera and J.-F. Bonastre, "A novel speaker binary key derived from anchor models," in *Proc. Interspeech*, 2010.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788 –798, May 2011.

[5] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *Journal on Applied Signal Processing, Special issue on biometric signal processing*, 2004.

[6] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 1, 2002, p. I.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435 –1447, May 2007.

[8] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, 2005, pp. 629 – 632.

[9] X. Anguera and J.-F. Bonastre, "Fast speaker diarization based on binary keys," in *Proc. ICASSP*, 2011.

[10] J. Bonastre, P. M. Bousquet, D. Matrouf, and X. Anguera, "Discriminant binary data representation for speaker recognition," in *Proc. ICASSP*, 2011.

[11] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, US Virgin Islands, USA, Dec. 2003.

[12] T. H. Nguyen, E. S. Chng, and H. Li, "T-test distance and clustering criterion for speaker diarization," in *Proc. Interspeech*, 2008.

[13] A. F. Martin and C. S. Greenberg, "NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels," 2009, pp. 2579–2582.

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