A Novel Speaker Binary Key Derived from Anchor Models

Xavier Anguera[1], *Jean-François Bonastre*[2]

[1]Multimedia Research Group, Telefonica Research, Barcelona, Spain
[2]University of Avignon, LIA, Avignon, France.
xanguera@tid.es, jean-francois.bonastre@univ-avignon.fr

Abstract

The approach presented in this paper represents voice recordings by a novel acoustic key composed only of binary values. Except for the process being used to extract such keys, there is no need for acoustic modeling and processing in the approach proposed, as all the other elements in the system are based on the binary vectors. We show that this binary key is able to effectively model a speaker's voice and to distinguish it from other speakers. Its main properties are its small size compared to current speaker modeling techniques and its low computational cost when comparing different speakers as it is limited to obtaining a similarity metric between two binary vectors. Furthermore, the binary key vector extraction process does not need any hard threshold and offers the opportunity to set the decision steps in a well defined binary domain where scores and decisions are easy to interpret and implement.

Index Terms: binary key, speaker modeling, biometrics

1. Introduction

Voice-based biometric systems face several kinds of variabilities. Such systems take advantage of inter-subject variability and try to minimize intra-subject variability, which includes the effect of pathological state of the subjects, speaking styles, the linguistic content of the messages, etc. These two variabilities are taken into account by any speaker recognition engine. Unfortunately, several other "negative" variabilities are present in this area and are usually grouped together and labelled as "session variability". The main aspects of session variability are environmental noises, recording equipment and transmission channel effects.

Statistical modeling has proven for quite some time its efficiency in dealing with these different variabilities. A classical speaker recognition system is usually based on the GMMUBM paradigm [1] associated to a session variability modeling like the Factor Analysis approach [2]. "Negative" variability problems are also taken into account by adding several normalization processes to the former approaches, applied at different levels, such as the acoustic vectors or the scores.

This statistical framework implies large and complex models. These models require large amounts of training data and are generally dependent on the targeted scenario. The different normalization levels further increase the global complexity of the systems. The scores issued from these processes are also difficult to interpret and sometimes require a new normalization in order to fit, for example, a Bayesian framework [3]. This is important since easy to understand scores allow easy to fit thresholds, which are a key point for real case applications.

The approach presented in this paper aims to solve this problem by representing a voice recording via a simple binary vector. In this approach, like in [4] or in the *anchor* model approach [5], no acoustic modeling of a given speaker is needed even if an unique general statistical acoustic model is still present, which is only used in order to extract the binary key. Unlike the latter approaches, our proposal is based on a novel acoustic key composed only of binary values. Its main properties are its small size compared to current speaker modeling techniques like GMM and GMM-SVM systems [6, 7] and its speed, as voice comparisons correspond to a simple similarity metric between two binary vectors. Furthermore, the binary key vector extraction process —the only acoustic process in the system— does not need any hard threshold, making the system easy to adapt to different environmental conditions. It also offers the opportunity to base the decision step on a well defined binary domain where scores and decisions are easy to interpret and implement.

Note that, although in this paper the experimental section shows how the binary keys are able to discriminate between speakers, its formulation is general enough to be applied to other acoustic classification problems, which will be addressed in future work.

The main objective of the presented work is to assess whether the Speaker Binary Key is able to effectively model a speaker's voice and to distinguish it from other speakers. This paper is organized as follows. In section 2, the key element of the system, the binary key extraction algorithm, is presented, with a focus set on the underlined acoustic model, the Binary Key Background Model, which constitutes the core of the approach. Section 3 is dedicated to the experimental validation of the proposed idea, and the last section draws some conclusions and proposes possible future work.

2. Binary key extraction algorithm

The key extraction algorithm is composed of two main blocks. On the one hand, a binary key background model (KBM) needs to be computed only once at the beginning of the process, and is later used to convert speaker utterance(s) into the desired binary key. On the other hand, a method is described for transforming an acoustic utterance into a binary-valued multidimensional key using the KBM. Such background model takes advantage on the modeling abilities of the Gaussian Mixture Model framework, known to be able to capture the general form of the acoustic space: it consists of a big GMM model, close to the classical Universal Background Model (UBM) usually used in speaker recognition. However, the process used to build the KBM tries to emphasize the discriminant aspects of the information gathered by the model by distributing the model components within the acoustic space where speaker specific data is later expected to be modeled.

During the development of this work X. Anguera was partially funded by the Torres Quevedo Spanish program.

2.1. Obtaining the Binary Key Background Model

The KBM addresses the global acoustic space, as the classical Universal Background Model (UBM), but it also highlights speaker specificities. In order to follow these two goals, we start with a classical UBM and we extend it with a set of N_s *anchor* speakers used to put the focus on speaker specificities. We borrow the concept of anchor speaker from [4, 5] as it is also used as a base of speaker characteristics in which to project the data of a test segment. Selecting the set of anchor speakers as well as the size of this set are two important tasks. Even if more sophisticated approaches borrowed from information retrieval could be used, in this preliminary work we just randomly select the set of anchor speakers from the available ones and we propose an analysis of results concerning the effect of the size of the set considered.

In order to obtain the UBM model we use standard EM training, starting with a single Gaussian and by iteratively splitting all Gaussians until we reach the power of two closest to the desired number of Gaussians, then iteratively splitting the single Gaussian with highest posterior probability until reaching the desired complexity.

The KBM model is then obtained by concatenation of the Gaussian Mixture Models (GMM) of each of the anchor speakers, obtained in turn by EM or MAP adaptation (means only) of the UBM, with N_g Gaussian Mixtures, to each anchor speaker's data. If enough data is available for each anchor speaker a single full EM training iteration can be applied by taking the UBM as the initial model, else we found Reynold's Gaussian-independent adaptation coefficient method [8] to work better than using MAP with a global α adaptation ratio. In all cases we apply a lower cap of 0.5 to the Gaussians variance in order to prevent the model from overfitting to the data. The KBM model will finally have N Gaussian mixtures, where $N = N_s \cdot N_g$, the number of Gaussians mixtures in the UBM times the number of anchor speakers. Note that the different speech recordings used in this work were first parameterized into an n-dimensional stream of feature vectors, all individually normalized to zero mean and unity variance.

The computational power required to get the KBM model is split between the UBM training and the *anchor* speakers adaptation phase. As we will see in the experimental section, good performances are achieved with very small UBM models, therefore being fast to compute. Still, if bigger models are needed or lots of training data are required, the UBM can be trained offline in a big machine and then reused for all applications regardless of small differences in acoustic conditions by tuning the *anchor* models adaptation ratio.

2.2. Obtaining speaker binary keys

We define speaker binary key as an N-dimensional binary vector, $\mathbf{v}_f = v_1, \ldots, v_N$, $v_f \epsilon \{0, 1\}$ where N is the number of Gaussian mixtures in the KBM model. Setting any position in $v_f[i]$ to value 1(TRUE) indicates that the ith Gaussian in the KBM coexists in the same region of the acoustic space with the acoustic sequence being modeled. We also define an accumulator vector $\mathbf{v}_c = v_1, \ldots, v_N$, $v_c \epsilon \mathbb{N}^1$ initialized to 0(FALSE), where each position $v_c[i]$ also represents the same Gaussian Mixture from the KBM as $\mathbf{v}_f$.

Given one or more test utterances (previously parameterized in the same way as in the training set) we obtain the speaker binary key in two steps. First, for each acoustic frame in the utterances we compute its likelihood given each of the Gaussians in the KBM model and select the top Θ_1 percent Gaussians with highest likelihood values. It is important to highlight that Θ_1 is not a (decision) threshold. It is more a parameterization meta parameter which defines the amount of information extracted from each frame (similar role than the dimensionality of acoustic vectors). For the selected Gaussians we increase the corresponding $\mathbf{v}_c$ positions by 1. Intuitively, this procedure projects the acoustic location of each acoustic frame from the feature space into the space of KBM Gaussians and keeps only those components with highest impact.

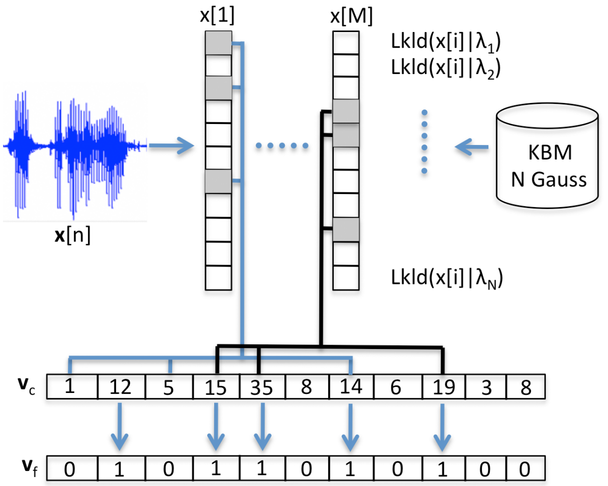

Figure 1: *Steps involved in obtaining the binary key for a speaker.*

When all frames have been processed, each position $v_c[j]$ in the accumulator vector contains the relative importance of Gaussian j in modeling the test utterances we have processed. The conversion from $\mathbf{v}_c$ to $\mathbf{v}_f$ is straightforward by setting the top Θ_2 percent positions in $\mathbf{v}_c$ with highest values to TRUE, and to FALSE when otherwise. This process is quite fast as it only requires the evaluation of the KBM model with the test data and the partial sorting of the results, which can be efficiently implemented using standard programming libraries. Figure 1 illustrates the process to obtain a speaker binary key from an input signal $\mathbf{x}[n]$ with a KBM model of $N = 11$ with $\Theta_1 \simeq 30\%$ and $\Theta_2 \simeq 45\%$.

This process looks like a component selection based on highest occupancies (regarding the KBM) except that the likelihood values are only used to select the components on a per frame basis. For each frame and component, a FALSE or TRUE decision is taken, which allows to switch from a likelihood-based numerical domain to a binary domain.

The fusion rule used to obtain the key from the individual binary vectors computed on each frame is the maximum rule. Some other classical fusion rules are possible as well as new rules based on binary arithmetics.

3. Experimental Evaluation

3.1. Evaluation Setup

In order to test the feasibility of the proposed binary key to discriminate among speakers we have performed tests using the TelVoice [9] database, which consists of over 87 speakers (42 male and 45 female) recorded through phone calls in 10 different sessions, where the time between recordings ranges from three days to more than one year. Each session consists of 10 spoken items, consisting of the speaker's cellphone number, his/her ID number, name and date of birth. Using this database for this test is challenging both because of the time

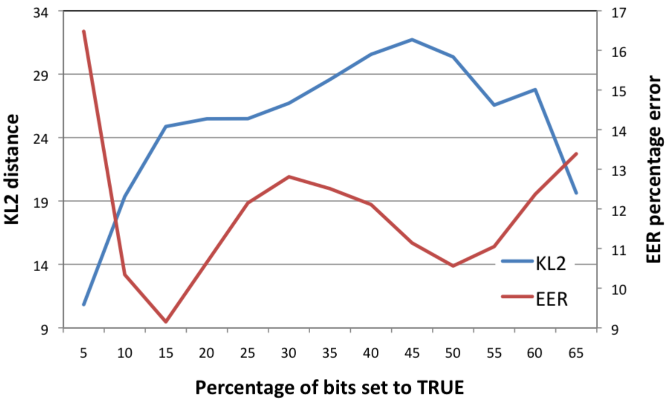

Figure 2: *KL2 and EER* w.r.t. *the percentage of final bits chosen.*

difference between sessions and of the partial text dependency of the recordings among all users. For the evaluation we split the data into 3 sets — each one with 15 female speakers and 14 male speakers — for training, development and testing. For training we used all data available for each of the speakers (104 short utterances in total) whereas for development and testing we further split each speaker's data into sets of 8 utterances for 7 sessions, and 2 sets of 8 utterances for 3 sessions. The average accumulated length per development/testing set is 30.1 seconds.

In this paper we use a simple similarity metric between any pair of binary keys defined as

$$S(\mathbf{v}_{f1}, \mathbf{v}_{f2}) = \frac{1}{N} \sum_{i=1}^{N} (v_{f1}[i] \wedge v_{f2}[i]) \tag{1}$$

where $\wedge$ indicates the logic *AND* operator between any two bits. Then, in order to evaluate how discriminative a binary key is in modeling a given speaker we use two approaches. On the one hand, we model the statistical distributions of same and different speaker similarities with normal distributions and compute the symmetrized Kullbak-Leibler distance (KL2) as in [10]. On the other hand we compute the equal error rate (EER) which indicates the minimum percentage error where miss and false alarms are equal.

3.2. Binary Key Analysis

Given a KBM trained from a UBM with $N_g = 128$ Gaussians and $N_s = 29$ *anchor* speakers (both male and female) from the training set, we set the optimum parameters Θ_1 and Θ_2 given the development data. Figure 2 shows the KL2 distance between same-speaker and different-speaker distributions and the EER error with respect to the Θ_2 percentage. Note that we strive to maximize KL2 and minimize EER. For $\Theta_2 \leq 5\%$ most selected positions correspond to Gaussians that are commonly prominent in all binary keys and therefore cannot effectively discriminate between speakers. Gaussians in this group are usually those modeling silence and noisy acoustic frames. Also, note that for $\Theta_2 \geq 50\%$ the discriminability power of the binary key gradually diminishes. We observe that the optimum values for Θ_2 are different depending on the chosen metric. Using KL2 we would select $\Theta_2 = 45\%$ while using EER it would be $\Theta_2 = 15\%$. Running a similar experiment for Θ_1 we found that $\Theta_1 = 1\%$ optimizes both KL2 and EER.

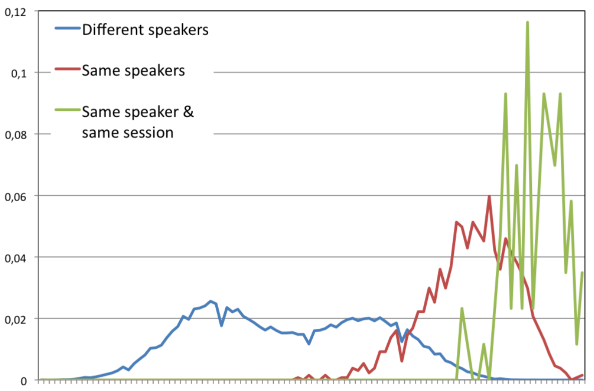

Figure 3: *Normalized histograms of same-speaker and different-speaker similarities.*

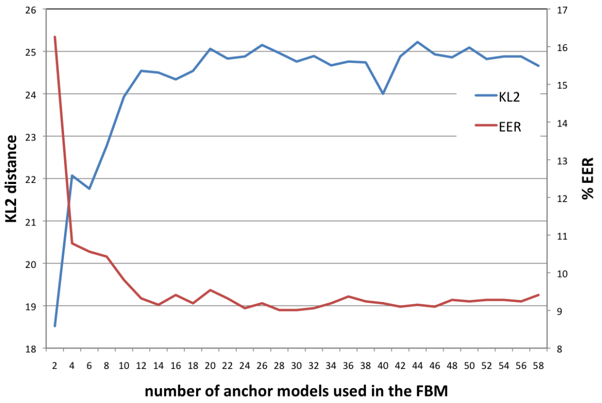

Figure 4: *Binary key discriminability* w.r.t. *the number of anchor models used in the KBM.*

In order to explain the λ_2 optimization discrepancy depending on the metric used, Figure 3 shows the histograms (normalized to have area 1) of the computed similarities between all same-speaker, same-speaker within the same session and different-speaker binary key pairs found in the development set. As expected, the same-speaker histogram has higher similarity values than the different-speaker histogram, both being easily separable by thresholding. This difference is further enhanced if we only consider the similarities between utterances of the same speaker recorded in the same session. Observe also that while the distributions for the same-speaker values are monomodal and could be modeled by a normal distribution, the different-speaker distribution follows a bimodal distribution. Given that the KL2 distance used here considers two monomodal and normal distributions, it might sometimes produce slightly biased results. This is why for further tests we use $\Theta_2 = 15\%$ as optimized using EER. Note, though, that using the KL2 metric based on monomodal distributions is still a valid metric to compare how the overall similarity distributions deviate from each other, complementary to using the EER, which focuses on the percentage of data that overlaps between both distributions.

Next, Figure 4 shows the KL2 and EER metrics computed on the development set by using a KBM obtained using a different number of *anchor* speakers. In all cases the KBM was derived from the same UBM ($N_g = 128$ Gaussians, trained on the training set). At each step, N_s was incremented in 2 new speakers (one male and one female) from data in the training and testing sets. We can see that we reach a steady state both for EER and KL2 around $N_s = 12$ speakers, with a performance comparable to the bigger models. This indicates that only with a few *anchor* speakers we sufficiently cover the acoustic space necessary to model any speaker (at least from the same acoustic conditions like in the database we are using). Furthermore, as the metrics remain fairly constant after 12 models we believe the binary key is robust to the choice of *anchor* speakers, which in this case has been done randomly (except for keeping the balance between men and women).

Table 1: *UBM size effect of the binary key and comparison with direct KBM construction from the UBM*

System	size (bits)	Dev. set		Test set	
		KL2	EER	KL2	EER
BK with 16G UBM	0.4Kb	19.4	12.4%	15.9	12.3%
BK with 32G UBM	0.9Kb	22.9	11.1%	16.0	11.8%
BK with 64G UBM	1.8Kb	24.7	9.9%	19.6	10.6%
BK with 128G UBM	3.7Kb	24.9	9.1%	18.4	10.8%
BK with 256G UBM	7.4Kb	25.4	10.5%	19.2	10.6%
BK with 512G UBM	14.8Kb	24.4	9.2%	18.8	10.1%
Minimal BK	2Kb	24.4	10.6%	18.5	11.5%
Direct 256G UBM	0.2Kb	15.6	17.7%	14.7	19.0%
Direct 512G UBM	0.5Kb	16.9	16.1%	14.1	15.2%
Direct 1024G UBM	1Kb	17.7	16.5%	14.6	15.2%
Direct 2048G UBM	2Kb	18.3	16.1%	14.8	14.4%
Direct 4096G UBM	4Kb	18.4	16.3%	14.6	14.6%

An analysis of the dependency of the proposed binary key (BK) *w.r.t.* the size of the UBM model is shown in the top part of Table 1, for the development and testing sets. Using $N_g = 64$ Gaussians and above we obtain very similar results in both sets, indicating that we do not need very large initial models to be effective in discriminating between speakers. Given this and the results presented in Figure 4, the center line in Table 1 shows the minimal binary key ($N_g = 64$ Gauss for the initial UBM + $N_s = 12$ *anchor* speakers) that could be considered to still show competitive performances with only 96 bytes per binary key. Next, the lower part of the table shows the performance metrics for binary keys obtained in the same way as the proposed but using an KBM consisting only of a UBM trained on the training set (the same data used for the BK). We can see that, for similar sizes of the resulting binary vectors, these *direct* approaches obtain considerably worse results both in KL2 and EER.

4. Conclusions

In this article we presented a new way to develop a biometric authentication system which represents the data gathered from a given biometric captor as a binary key. The proposed approach shows several advantages. Firstly, it allows the representation of biometric samples and a user reference by a compact binary vector. Secondly, the comparison between a sample and a reference is obtained by a simple similarity metric between two binary vectors. This comparison is cost effective. Thirdly, the decision process and the corresponding threshold take place in the binary vector domain, allowing for use of straightforward and meaningful information theory approaches. Furthermore, all the processes linked to the intrinsic nature of the targeted biometric media are placed into a unique module, the binary key extractor. This module accounts for the feature extraction and all the domain-dependent modeling, including the specific variability modeling. The underlined statistical models are trained off-line and this module could be easily transposed from one biometry to another as long as the same framework (UBM/GMM and FA) is followed.

We investigated the feasibility of the proposed approach on a speaker recognition task. Even if the current implementation is very simple, we demonstrated experimentally that the Speaker Binary Keys proposed in this paper are able to discriminate between speakers. After this first step, we aim to consolidate the results obtained using a larger database. Several developments of the approach have also to be investigated. The use of FA variability modeling is easy to apply and demonstrated its potential has been demonstrated in the speaker recognition literature. This technology should be implement rapidly in the system. The selection of anchor speakers could be improved, in order to select an anchor set as small as possible but with a maximal coverage of the speaker characteristic space. To take advantage of the binary domain, a new decision process could be defined, using information theory to propose an optimal threshold corresponding to the targeted operating point. Finally, one of the main advantages of the approach proposed is certainly its ability to be applied to several other biometric fields, like face or fingerprints. It seems interesting to us to propose such an adaptation of the method (to a new domain) and to take advantage of the binary domain of the keys for multimedia fusion.

5. References

[1] 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.

[2] P. Kenny, G. Boulianne, and P.Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, p. 345, 2005.

[3] N. Brummer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230 – 275, 2006, odyssey 2004: The speaker and Language Recognition Workshop - Odyssey-04.

[4] T. Merlin, J.-F. Bonastre, and C. Fredouille, "Non directly acoustic process for costless speaker recognition and indexation," in *in Proc. COST-254 International Workshop on Intelligent Communication Technologies and Applications, with emphasis on Mobile Communications*, 1999.

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

[6] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. V. Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 2072–2084, 2007.

[7] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *in Proc. ICASSP*, vol. 1, 2006, p. H.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] L. Rodríguez-Linares, C. García-Mateo, and J. L. Alba-Castro, "On combining classifiers for speaker authentication," *Pattern Recognition Journal*, vol. 36, pp. 347–359, 2003.

[10] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *in Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.