TYPICALITY EXTRACTION IN A SPEAKER BINARY KEYS MODEL

Pierre-Michel Bousquet, Jean-François Bonastre

University of Avignon (LIA), France

ABSTRACT

In the field of speaker recognition, the recently proposed notion of "Speaker Binary Key" provides a representation of each acoustic frame in a discriminant binary space. This approach relies on an unique acoustic model composed by a large set of speaker specific local likelihood peaks (called specificities). The model proposes a spatial coverage where each frame is characterized in terms of neighborhood. The most frequent specificities, picked up to represent the whole utterance, generate a binary key vector. The flexibility of this modeling allows to capture non-parametric behaviors. In this paper, we introduce a concept of "typicality" between binary keys, with a discriminant goal. We describe an algorithm able to extract such typicalities, which involves a singular value decomposition in a binary space. The theoretical aspects of this decomposition as well as its potential in terms of future developments are presented. All the propositions are also experimentally validated using NIST SRE 2008 framework.

Index Terms: speaker modeling, binary keys, speaker recognition

1. INTRODUCTION

In [1, 2, 3] a new approach for speaker recognition, denoted "Speaker Binary Key", was presented. Contrary to classical speaker recognition based on statistical modeling of the speaker information [4][5], this approach proposes to handle directly each piece of speaker specific information in a binary space. Each coefficient of this binary space corresponds to a targeted piece of speaker-specific information which could be present (the coefficient is equal to 1) or non present (the coefficient is equal to 0) in a given acoustic frame or acoustic segment. This new approach allows to exploit temporal or sequential information as a binary vector is extracted for each acoustic frame. It also focuses on speaker specific information in a non-parametric way as each coefficient of the binary space models speaker-specific information. This approach contrasts with classical speaker recognition systems where the models represent a global acoustic space (for a given speaker for example). One of the main problem of statistical modeling, the data missing problem (and the under/over training of the models) is decreased here as, for a given direction of the space, only the presence or non presence of a given specificity is evaluated. Finally, this approach is also computationally efficient as binary arithmetic requires fewer amount of memory or computational time compared to classical statistical approaches.

This paper introduces a new concept of "typicality" between binary keys in order to increase the discriminant abilities of the approach as well as the Speaker Binary Keys robustness to channel mismatches. This concept is inspired from the search for inherited similarities in the DNA (genomics). Here, similarities pertain to the speaker or session characteristics. They are gathered from the links between the specificity models, i.e. the dimensions of the binary space. Section 2 will briefly present the Speaker Binary Key approach. The heart of this work, the typicality concept, is presented in Section 3. Section 4 is dedicated to the experimental part of this work. Finally, section 5 will presents some conclusive remarks as well as some ideas for future works.

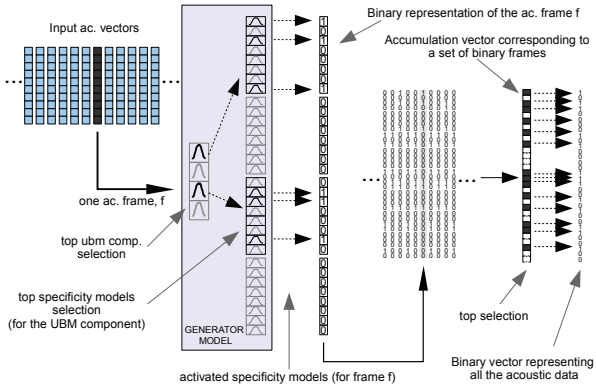

Fig. 1. Overview of the binarization process

2. OVERVIEW OF THE SPEAKER BINARY KEY APPROACH

Figure 2 presents a schematic view of the Speaker Binary Key approach. It relies on an unique acoustic model trained a priori during the development phase, denoted "Generator model", and on a binarization process. The Generator model is composed by a large set of speaker specific local likelihood peaks, called "specificity" models, associated with a classical UBM model. Each of these specificities models emphasizes a given user specific information for this particular acoustic space region. Each specificity model is represented by a Gaussian and is tied with one of the UBM component. The role of the UBM is only to tie each input frames with one or several Gaussian components, i.e. one or few acoustic regions. The specificity models are gathered from a training set comparable to the one used for the UBM training but, here, a specificity model is trained from data related to an unique speaker, in order to emphasize the specific information corresponding to this speaker. The Generator model training is detailed in [2, 3].

For the binarization process, a transformation $T : R^n \rightarrow N^m$ between an n-dimensional (acoustic) feature vector and an m-dimensional binary vector is applied on each acoustic frame. It allows to project each individual acoustic vector into the binary space. 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 process is performed independently for each UBM component, i.e. for the specificity models tied with this component. 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. An unique binary vector able to represent a particular speaker is then obtained by majority voting: a value 1 is set for the vector locations corresponding to the highest number of 1 in the sequence of binary vectors, while a value of 0 is given otherwise.

As highlighted in [2, 3], the proposed approach inherits a part of its concepts from both anchor-based modeling [6] and UBM component posterior probabilities [7].

3. TYPICALITY EXTRACTION

The vectors produced by the binary key system indicate, for each of the N specificities of the model, if it has been picked to represent a given session. Dimensions of the binary space, the specificities, are categorical ('factor') variables with only two levels *yes* or *no*.

3.1. Similarities and typicalities

A simple proximity between two sessions according to a given specificity is obtained using the boolean AND operator. Since the two levels are coded 0 and 1, a similarity between two binary vectors v_1, v_2 is computed [1] by the formula:

$$S(v_1, v_2) = v_1.v_2 = \sum_{i=1}^{N} (v_1)_i (v_2)_i \quad (1)$$

where the scalar product . is equivalent in the binary space to the AND operator for categorical variables. This similarity is normalized by division by N (or by the maximal number of picked specificities to provide a $[0, 1]$ value).

Here, our aim is to reveal specificity subsets linked by a discriminant relation. We expect that computing a new similarity between two binary vectors according to such typicalities will improve the decision process. We introduce below the notion of typicality. A typicality is a subset $\mathcal{L}$ of specificities $\{1, ..., N\}$ from which it is possible to improve the discrimination by adding complementary information on vector proximities. A typicality can contain information on session, speaker, total variability, or impostor distribution. Mathematically, we assume that this relation between linked specificities allows to compute on them a similarity based on a "full-product" instead of the scalar product. For two binary vectors, their values 0 or 1 for the specificities subset corresponding to a typicality $\mathcal{L}$ are first selected, then a value of similarity between the two sub-vectors according to $\mathcal{L}$ is computed by crossed multiplications between all their binary values (sum of all the products between pairs of values). The similarity of two vectors according to $\mathcal{L}$ is the full-product:

$$S_{\mathcal{L}}(v_1, v_2) = \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} (v_1)_i (v_2)_j \quad (2)$$

This similarity considers that every combination of selected specificities for this typicality indicates a "trace" of an useful effect, and is more powerful that the scalar product which proceeds only "dimension by dimension".

3.1.1. Full-product fast computation

The previous similarity can be efficiently computed. If the sub vectors of the $\mathcal{L}$-indices of v_1 (resp. v_2) contain n_1 (resp. n_2) non-zero values, the full-product between v_1 and v_2 is equal to $n_1 \times n_2$.

[1] using a criterion derived from the Sokal & Michener distance.

Given a set $\{\mathcal{L}\}$ of l typicalities, we introduce the $l \times N$ matrix L such as its i^{th} row is the Kronecker's vector $\delta_{\mathcal{L}_i}$ of the i^{th} typicality and the diagonal matrix D filled with the l typicalities cardinals. Then a normalized similarity between two vectors according to this set of typicalities can be computed by:

$$S_{\{\mathcal{L}\}}(v_1, v_2) = v_1.L^t D L v_2 \quad (3)$$

3.2. Singular value decomposition and categorical variables

3.2.1. Nuisance Attributes Projection and categorical variables

The Nuisance Attributes Projection (NAP) procedure [8] estimates session variability as a subspace of intermediate rank r obtained using principal axes (eigenvectors having the largest eigenvalues) of the within-class covariance matrix, and projects the vectors onto the orthogonal complementary subspace, assumed to be the speaker space. Let U denotes the $r \times N$ matrix of the r leading eigenvectors. The projection on the previous subspace uses the matrix $I - U^t U$ and the similarity between two session-compensated vectors becomes:

$$S(v_1, v_2) = \left(I - U^t U\right) v_1 . \left(I - U^t U\right) v_2 \quad (4)$$

which, since $\left(I - U^t U\right)$ is a projection matrix, is simply rewritten as:

$$S(v_1, v_2) = v_1.v_2 - U v_1 . U v_2 \quad (5)$$

This technique appears inadequate in a binary model, due to the categorical nature of the specificities variables. The multiple correspondence analysis (MCA) is an equivalent-to-PCA technique for categorical variables. It builds an hypertable of contingency and a Burt's table to load frequencies of each level pairs. Then a χ^2-metric distance allows to compare levels of the variables and compute principal variability axes, onto which the train vectors are projected. But in the special case of binary variables (all variables have only two levels), one demonstrates that the MCA is equivalent to a PCA on the binary coded vectors. We present below the results of a NAP technique applied into our binary space.

3.3. A typicality extractor using eigenvectors binarization

3.3.1. Binarization algorithm

It is possible to extract subsets of typicalities from the previous eigenvectors matrix U. Often, in the field of data analysis, the variability principal axes have to be "explained" by the initial variables. It means that each principal axis summarizes most highly correlated groups of initial dimensions (for us the specificities). To determine the proportion of each specificity in the generation of a principal axis, the set of n training vectors is projected onto this axis. The $\mathbb{R}^n$ random-vector obtained is a well-known *principal component*. Then, covariances or correlations between this principal component and the $\mathbb{R}^n$ random vectors of each dimensional variable (here the specificities) are computed.

Let V be a $N \times r$ matrix of r eigenvectors v_j and $cov(s_k, p_j)$ denote the covariance between the k^{th} specificity s_k and the j^{th} principal component p_j. The specificities that maximize $|cov(s_k, p_j)|$ are considered as the most important to "explain" the j^{th} eigenvector. It is straightforward to show that the $N \times r$ matrix $\{cov(s_k, p_j)\}_{j,k}$ can be quickly computed thanks to the formula:

$$\{cov(s_k, p_j)\}_{j,k} = V \Lambda_r \quad (6)$$

where Λ_r is the r maximal eigenvalues diagonal matrix. As explained above, the most positive and negative values of each column

of this table indicate the most involved specificities in the generation of this variability principal axis.

To binarize V, we replace its initial values with 1 for the most positive of $V\Lambda_r$, -1 for the most negative, 0 otherwise. But let us consider the problem of independence between effects: if a given specificity is non-zero for too many eigenvectors, these eigenvectors move away from orthogonality [2] and their effects tend to mix themselves. It is getting hard to add their values in an unique homogeneous score. Therefore, we choose to distribute a small and constant amount q of -1 and 1 for each specificity. This decision is the key of our algorithm: in this way, scalar-products between eigenvectors will be close to 0, thus eigenvectors close to orthogonality. Having to binarize a $N \times r$ matrix V of r eigenvectors v_j, we propose the following algorithm:

Compute $M = V\Lambda_r = \left\{\lambda_k v_{k_j}\right\}_{j,k}$
For each specificity s_k
  set to 1 the q most positive values of the k^{th} row of M
  set to -1 the q most negative values of the k^{th} row of M
  set to 0 all other values of the k^{th} row

where q is a constant amount of values by specificity. M is a discretized matrix containing only $-1, 0$ or 1 values. Let B and B' be the $N \times r$ matrices defined by:

$B_{ij} = 1$ if $M_{ij} = 1$ and 0 otherwise
$B'ij = 1$ if $M_{ij} = -1$ and 0 otherwise

M is equal to $B - B'$, and B and B' are binary matrices. Each column of B indicates a group of strongly correlated specificities, hence a typicality. Each column of B' indicates another typicality. Subtraction $B - B'$ mentions an "opposition" between these two typicalities.

3.3.2. Binary eigenvectors spectrum

The amount of non-zero values in the successive binary eigenvectors can be tapped in the same way as eigenvalues. These values, which turn out to be decreasing, can be interpreted as an energy spectrum. It also turns out that this "spectrum" accentuates the energy of principal variabilities, concentrating meaningful eigenvectors in a smaller set than its common continuous version. For example, the binary spectrum of matrix U used in our experiments is negligible beyond the 50^{th} value.

3.3.3. Validation of the extractor

We have now to prove that specificity subsets extracted by our algorithm verify typicality properties. As explained above, a typicality contains a discrimination power, and this power can be measured using "full-product" between pairs of vectors. To test this assumption, we experiment a NAP technique with a similarity of the form (5), but where the variability matrix is replaced by the binary eigenvectors matrices of typicalities. The validity -and quality- of these typicalities will be evaluated by the performance computed on our test sets.

To obtain a ready-for-scoring matrix, two points have to be taken into account. First, the columns of $B - B'$ are not independent. However, if q is close to 1, the very low number of non-zero values in each column allows to obtain a matrix $(B - B')^t (B - B')$ close to be diagonal (ie orthogonality of the eigenvectors). Secondly, eigenvectors can be length-normalized to standardize their effects. Let D denote the $r \times r$ diagonal matrix such as $D_j = \sum_{j=1}^{N} \left(B_{jk} + B'_{jk}\right)$ is the number of non-zero values in the k^{th} binary eigenvector. The eigenvectors matrix becomes $(B - B')\, D^{-\frac{1}{2}}$. Using the quasi-orthogonality of those eigenvectors, the proposed similarity between two vectors is

$$S(v_1, v_2) = v_1.v_2 - v_1.Av_2 \tag{7}$$

where $A = (B - B')\, D^{-1} (B - B')^t$.

The vectors v_1, v_2 and matrices B, B' are binary, D is diagonal with only integer values. Moreover, the second term of the similarity is a weighted sum of full-products. Note that distributing more than one 1 and one -1 for each specificity ($q > 1$) moves away the eigenvectors from orthogonality but adds new informations to the matrix.

3.3.4. Fundamental result

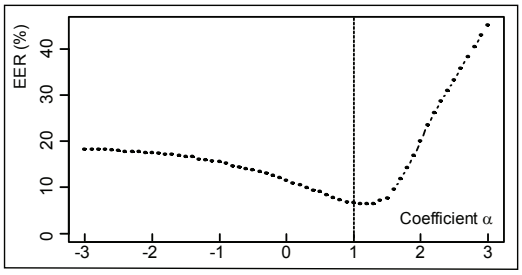

Fig. 2. Fusion $v_1.v_2 - \alpha\,(v_1.Av_2)$ of initial and binary eigenvectors matrix scores: coefficient α and resulting EER%

The relevance of the similarity (7) as a measure of proximity with respect to typicalities can be questioned: on the one hand initial real values of eigenvectors have been discretized, on the other hand the equivalence between similarities (4) and (5) is no longer verified, due to non-orthogonality of binary eigenvectors. Moreover, (7) can be seen as a fortunate fusion between two heterogeneous scores: the initial similarity $v_1.v_2$ and a new similarity $-v_1.Av_2$. To show the validity of (7), we tested the fusion of its separate terms. A fusion score is produced by varying a real coefficient α in the similarity:

$$S_{alpha}(v_1, v_2) = v_1.v_2 - \alpha\,(v_1.Av_2) \tag{8}$$

Figure 2 shows the results of this fusion for the experiment detailed below. For α varying from -3 to 3, we display the equal-error-rates (EER %) calculated from (8). The best discrimination performance are obtained for α close to 1. This clearly shows the relevance of this similarity.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

In order to validate experimentally our proposals, we used the 32768 specificities coverage previously described in [2, 3] (128 GMM-components, 256 specificities by Gaussian).

All the sessions (train and test) are binarized by majority voting, with a criterion of maximum of picked "ones" by Gaussian inversely proportional to the information quantity on this Gaussian. The number of ones per binary keys vector is fixed for all the experiments.

[2] Their scalar products move away from 0.

The variability matrices are trained using a set of 5233 sessions from 296 speakers of NIST-SRE 2004 and 2005. Speaker verification experiments are performed based upon the NIST SRE 2008 database, male speakers, for condition det 7 (telephone-telephone). This condition uses 1271 speakers, 6615 tests (439 target tests, the rest are impostor trials). Results are given in terms of equal-error-rate (EER) and the minimum DCF (an a posteriori decision).

4.2. Results

The Table 1 presents the comparison of results using three similarities: the initial similarity (eq.1) termed as "Initial", the similarity using NAP with continuous matrix (eq.5) termed as continuous NAP", and the "full binary" similarity using typicalities obtained by the binarized version of the eigenvector matrix (eq.7) termed as "binary Typicality", this last for several ranks of the matrix. The number q of 1 or -1 by specificity is fixed and equal to to 3. First, NAP technique greatly improves performance compared to initial similarity. This validates the efficiency of usual singular value decomposition techniques into our binary space. Second, results of the full binary typicality-based similarity are close to those of the traditional NAP. It is worth noting that they are achieved for very low ranks (a rank of 10, thus 20 typicalities, yields an interesting performance of EER 6.64% and minDCF 0.0347).

Table 1. Comparison of performance of three similarities: initial, NAP with continuous matrix and binary typicality-based (this last for several ranks of binary matrices).

score	rank	DCF$\times$100	EER %
Initial	-	5.18	11.44
continuous NAP	50	2.53	5.46
binary Typicality	1	5.10	10.52
	5	3.66	7.47
	10	**3.47**	**6.64**
	15	3.57	6.83
	25	3.42	6.60
	50	3.54	7.06
	100	3.52	7.06

Table 2 compares performance in terms of EER and needed amounts of information (in bytes) of several speaker verification systems: each of these methods uses one or more matrices to handle variabilities and we indicate the sizes in bytes of these matrices. Amount of information required by typicality system is an hundred times lower than the one required by a Factor Analysis technique. This result demonstrates that the speaker discrimination problem can be greatly reduced in terms of quantity of informations by using the binary approach.

Table 2. Comparison of performance and necessary amounts of information (matrices) from different systems: Joint Factor Analysis (JFA), Factor analysis (FA) and Typicality. No normalization applied.

	JFA	FA	Typicality
space dim.	512×50	512×50	128×256
matrices ranks	$40 + 60$	40	$10 + 10$
data type	64 bits	64 bits	1 bit
EER	2.72%	3.89%	6.64%
amounts of info. (KB)	20 480 KB	8 190 KB	81 KB

5. CONCLUSION AND PERSPECTIVES

In this paper we presented an extension to the Speaker Binary Key approach presented in [1, 2, 3]. We proposed to extend the discriminative power and the session mismatch robustness by the use of "'typicalities"'. A typicality is a subset of specificities which contains discriminant information gathered from vector proximities. The typicalities are extracted by an algorithm based on eigenvector binarization. The experimental results demonstrated the interest of such a typicality concept as our proposal improved significantly the performance (from 11.44% EER to 6.64% EER). Compared with the traditional continuous domain NAP approach presented in [2, 3], the performance of our full binary approach remains comparable (5.46% EER for NAP to be compared with 6.64% EER for the binary one). Finally, it is interesting to compare the Speaker Binary Key approach using typicality extraction with our JFA baseline. The JFA performs better (2.72% of EER vs 6.64% of EER) but it uses about 250 times more information to model the session effects than the typicality-binary approach (20480 KB vs 81 KB). This results shows that important gains in terms of computer resources are possible using our approach.

The results presented in this paper confirmed the interest of the Speaker Binary Key approach. Further works will look at the specificity extraction during the training of the Generator model, which is the core point of the approach. The typicality concept seems well adapted for this task as it allows to improve the intrinsic speaker discriminant nature of the binary space. In this paper the extracted typicalities have a "negative" effect, in the sense that they summarize a session variability which has to be compensated. A second set of future investigations will focus on "positive" typicalities, able to highlight only speakers features, considering (by analogy with genomics) one-speaker binary keys like observations of a latent inherited family.

6. REFERENCES

[1] X. Anguera and J.F. Bonastre, "Novel binary key representation for biometric speaker recognition", Interspeech 2010, MAkuhari, Japan.

[2] J.F. Bonastre, X. Anguera, G. H. Sierra, P.M. Bousquet "Speaker modeling using local binary decisions", in Proc. Interspeech 2011, Firenze, Italia.

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

[4] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability : Theory and Algorithms," CRIM, Montreal, (Report) CRIM-06/08-13, 2005.

[5] P.M. Bousquet, D. Matrouf, J.F. Bonastre,"Intersession compensation and scoring methods in the i-vectors space for speaker recognition", in Proc. Interspeech 2011, Firenze, Italia.

[6] Y. Mami and D. Charlet, "Speaker identification by location in an optimal space of anchor models", ICSLP 2002, Denver, USA.

[7] N. Scheffer, J.F. Bonastre, "UBM-GMM driven discriminative approach for speaker verification", in IEEE Odyssey - The Speaker and Language Recognition Workshop 2006, San Juan.

[8] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation", in IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, 2006, vol. 1. pp. 97-100.