Efficient Two-stage Speaker Identification based on Universal Background Models

S. Billeb*†‡, C. Rathgeb†‡, M. Buschbeck*, H. Reininger* and K. Kasper*‡

* atip - Advanced Technologies for Information Processing GmbH, Frankfurt, Germany
{stefan.billeb,michael.buschbeck,herbert.reininger}@atip.de

† da/sec - Biometrics and Internet Security Research Group, Darmstadt, Germany
‡ Department of Computer Science, Hochschule Darmstadt, Germany
{christian.rathgeb,klaus.kasper}@h-da.de

Abstract: Conventional speaker identification systems are already field-proven with respect to recognition accuracy. Since any biometric identification requires exhaustive $1:N$ comparisons for identifying a biometric probe, comparison time frequently dominates the overall computational workload, preventing the system from being executed in real-time. In this paper we propose a computational efficient two-stage speaker identification system based on Gaussian Mixture Model and Universal Background Model. Binarized voice biometric templates are utilized to pre-screen a large database and thereby reduce the required amount of full comparisons to a fraction of the total. Experimental evaluations demonstrate that the proposed system is capable of significantly accelerating the response-time of the system and, at the same time, identification performance is maintained, confirming the soundness of the scheme.

1 Introduction

Biometrics represent a rapidly evolving field of research and large-scale biometric systems are already deployed in commercial and governmental applications [JRP04, JFR08]. Since biometric data does not have any predefined sorting order, large-scale biometric identification systems have to compare a biometric probe to an entire database of biometric instances (gallery). The computational requirement of this comparison scheme, which is referred to as $1 : N$ system (the probe is compared to N stored instances), highly depends on comparison speed as well as on the number of instances stored in the database, i.e. real-time identification represents a challenging task. Focusing on biometric identification different mechanisms have been proposed in order to reduce the response time of the system. Indexing techniques have been proposed for different biometric characteristics, e.g. in [JPDG08, MPCG05]. However, throughout literature these techniques have been evaluated on rather small dataset, leaving scalability doubtable. In addition, serial combinations have been proposed, pre-pending computationally efficient algorithms to conventional identification systems in order to extract a subset of candidates, e.g. for iris based on binary feature vectors in [GRC09].

A binary representation of biometric features offers two major advantages: firstly a more compact storage of biometric templates and secondly a rapid comparison of biometric templates. Binary biometric templates facilitate comparison of thousands of templates per second, per single CPU core, i.e. even if binary representations of biometric data may cause a loss of information (and, thus, cause a decrease in biometric performance) these are suitable to be applied for pre-screening purposes in biometric identification systems. That is, compressed binary templates can be applied to carry out $1 : N$ comparisons while more sophisticated comparators are employed to compare according subsets of original templates. While different binarization methods have been suggest for voice biometric data, e.g. in [AB10, BBMA11], to our knowledge, no serial combinations based on binary voice templates have been considered so far.

The contribution of the present paper is twofold: (1) a scalable binarization technique for voice biometric data based on Gaussian Mixture Model (GMM) and Universal Background Model (UBM) is presented. (2) Obtained binary templates are utilized to efficiently perform $1 : N$ comparisons and return a short-list of top candidates in a speaker identification system. Implementing a serial combination of feature representations and according comparators in a single-instance scenario, computational effort required for biometric identification is significantly reduced. On a database which comprises voice samples of 339 subjects a speed-up of more than 95% is achieved, maintaining identification rates of the original system.

The remainder of this paper is organized as follows: in Sect. 2 the underlying GMM-UBM-based voice recognition system is described and in Sect. 3 the proposed two-stage identification system is introduced. Subsequently, experimental evaluations are presented in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 GMM-UBM Voice Recognition System

In past years numerous approaches to speaker authentication have been proposed: for a detailed review on existing literature the reader is referred to [KL10, LVH$^+$11]. The described system, which is considered a representative state-of-the-art speaker authentication system, is adapted in order to obtain a binary representation of voice data. Given a biometric observation $\mathbf{O}$ a total number of $\mathbf{K}$ feature vectors $\mathbf{o}_k\,[t]$ of length $\boldsymbol{T}$ are obtained in the feature extraction process, with $k = 1, \ldots, K$ and $t = 1, \ldots, T$. Feature vectors are modelled as a realization of a GMM by adapting the means of the UBM to the estimated means for the speaker using MAP adaptation. The GMM of a subject u is represented as the supervector $\boldsymbol{\xi}^{(u)}$ containing the mean vectors for each Gaussian distribution in the model.

In the binarization step, which builds upon a technique which has been similarly applied to on-line signatures [RMACC12], the supervector is binarized by comparing $\boldsymbol{\xi}^{(u)}$ component-wise against a modelled large population with a supervector $\overline{\boldsymbol{\mu}}$. Then, the latter vector is used as a feature-based biometric template and applied to perform identification. In the following subsection the employed feature extraction and the corresponding binarization technique, which are depicted as part of Fig. 1, are described in detail.

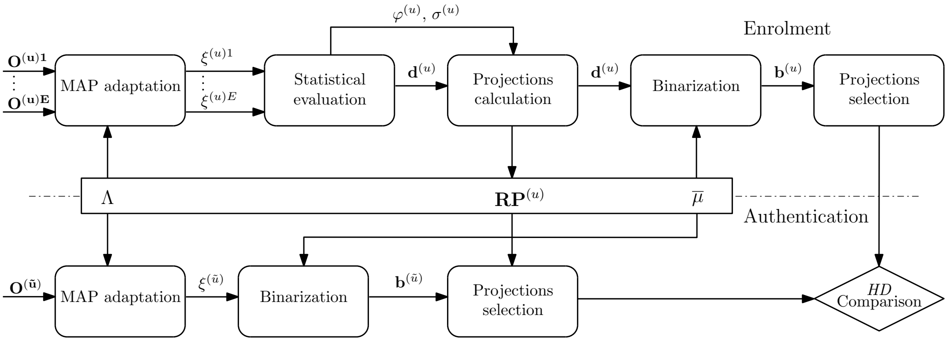


Figure 1: Flowchart enrolment and a single authentication of the proposed binarization scheme based on a GMM-UBM speaker identification system.

2.1 Feature Extraction

In order to extract features from an acquired speech signal the signal has to be decomposed into its frequency components using the FFT [RB76] at a window function of 20-30 ms size, thus, receiving one feature vector per window. As previously mentioned, several methods to extract features from the resulting magnitude spectrum have been suggested, e.g. LPC [MZR96] or PLP [SVC93], however, MFCC have shown to be powerful and difficult to improve upon in practice [Rey94, KL10].

We obtain 12 coefficients and the Log-energy value for each frame and the first and second-order derivative, i.e. in total we obtain a feature vector with 39 components per frame. The obtained vectors are further processed by cepstral mean subtraction as well as feature warping [PS01].

In order to derive an individual speaker model GMM from the general UBM, we only adapt the means of the UBM using the MAP adaptation. The UBM $\mathbf{\Lambda}$ represents a stochastic model containing a finite mixture of multivariate Gaussian components. Assume that E biometric observations $\mathbf{O}^{(u)e}$, $e = 1, \ldots, E$ of a user u are available during enrolment. We let the a posteriori probability for the component i, with $i = 1, \ldots, I$, of $\mathbf{\Lambda}$ be,

$$P\left(i|\mathbf{x}_T, \mathbf{\Lambda}\right) = \frac{\mathbf{w}_i g(\mathbf{x}_T|\boldsymbol{\mu}_i, \sum_i)}{\sum_{j=1}^{J} \mathbf{w}_j g(\mathbf{x}_T|\boldsymbol{\mu}_j, \sum_j)}, \tag{1}$$

where $\mathbf{w}_i$ represents the mixture weights, $\boldsymbol{\mu}_i$ the means, $\mathbf{x}_T$ the training vectors for the desired model, and $g(\mathbf{x}_T|\boldsymbol{\mu}_i, \sum_i)$ the component's Gaussian density. Further, for the training vectors $\mathbf{x}_T$, we compute the Maximum-Likelihood for the mean parameters as,

$$E_i(\mathbf{x}_T) = \frac{1}{n_i}\sum_{j=1}^{J} Pr\left(i|\mathbf{x}_j, \mathbf{\Lambda}\right)\mathbf{x}_j, \text{ with weight } n_i = \sum_{j=1}^{J} Pr\left(i|\mathbf{x}_j, \mathbf{\Lambda}\right). \tag{2}$$

The user's model is then characterized by the MAP adapted means $\widehat{\boldsymbol{\mu}}_i(\mathbf{x}_T)$ as,

$$\widehat{\boldsymbol{\mu}}_i(\mathbf{x}_T) = \boldsymbol{\alpha}_i E_i(\mathbf{x}_T) + (1 - \boldsymbol{\alpha}_i)\boldsymbol{\mu}_i, \tag{3}$$

where $\boldsymbol{\alpha}_i$ are fixed relevance factors [YLML12]. The adapted means are then combined to form the supervector $\boldsymbol{\xi}^u$ (biometric template) for user u

$$\boldsymbol{\xi}^{(u)} = [\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \ldots, \widehat{\boldsymbol{\mu}}_I], \tag{4}$$

where I defines the number of components in the UBM. The MAP adaptation can also be performed for a single biometric observation.

2.2 Binary Template Generation

Our goal is to generate a binary representation of the extracted voice template. As mentioned earlier, a binary representation allows the reduction of storage as well as a linear acceleration of identification speed in a large scale biometric system. In order to create a binary vector from the user's template we compare the derived supervector from the UBM against the inter-class mean of a large population $\overline{\boldsymbol{\mu}}$. The population $\overline{\boldsymbol{\mu}}$ has to be estimated during a training phase of the system, i.e. $\overline{\boldsymbol{\mu}}$ represents a fixed system parameter.

At the time of enrolment of user u we record a number of E observations and process them by applying the MAP adaptation to derive the vectors $\boldsymbol{\xi}^{(u)e}$ with Z coefficients ($Z = 39I$). The mean vector $\mathbf{d}^{(u)}$ is defined as,

$$\mathbf{d}^{(u)} = \frac{1}{E}\sum_{e=1}^{E}\boldsymbol{\xi}^{(u)e}, \tag{5}$$

such that the binary representation $\mathbf{b}^{(u)}$ can be estimated as,

$$\mathbf{b}^{(u)}[z] = \begin{cases} 0, \text{ if } \mathbf{d}^{(u)}[z] < \overline{\boldsymbol{\mu}}[z] \\ 1, \text{ if } \mathbf{d}^{(u)}[z] \geq \overline{\boldsymbol{\mu}}[z] \end{cases} \quad z = 1, \ldots, Z. \tag{6}$$

Since $\boldsymbol{\xi}^{(u)}$ can be obtained by a single biometric observation it is possible to directly generate a binary representation $\mathbf{b}^{(u)}$ from a probe. Moreover, it may be required to generate binarized templates of a pre-defined size v, that comprise only bits exhibiting the highest possible discriminativity. According fixed-length binary vectors can be obtained by estimating the reliability measures $\boldsymbol{\varphi}^{(u)}[z]$,

$$\boldsymbol{\varphi}^{(u)}[z] = \frac{\left|\mathbf{d}^{(u)}[z] - \overline{\boldsymbol{\mu}}[z]\right|}{\boldsymbol{\sigma}^{(u)}[z]}, \tag{7}$$

with the variance of the z-th feature estimated during the enrolment of user u defined as,

$$(\boldsymbol{\sigma}^{(u)}[z])^2 = \frac{1}{E-1}\sum_{e=1}^{E}(\boldsymbol{\xi}^{(u)e}[z] - \mathbf{d}^{(u)}[z])^2. \tag{8}$$

This measure assigns greater relevance to those features which lie further away from the population's mean than others. The v most discriminative features (largest values) are then indexed by storing a bit mask pointing at these values which is referred to as relevant projection $\mathbf{RP}^{(u)}$, i.e. this bit mask contains 1s at positions of the v most discriminative features.

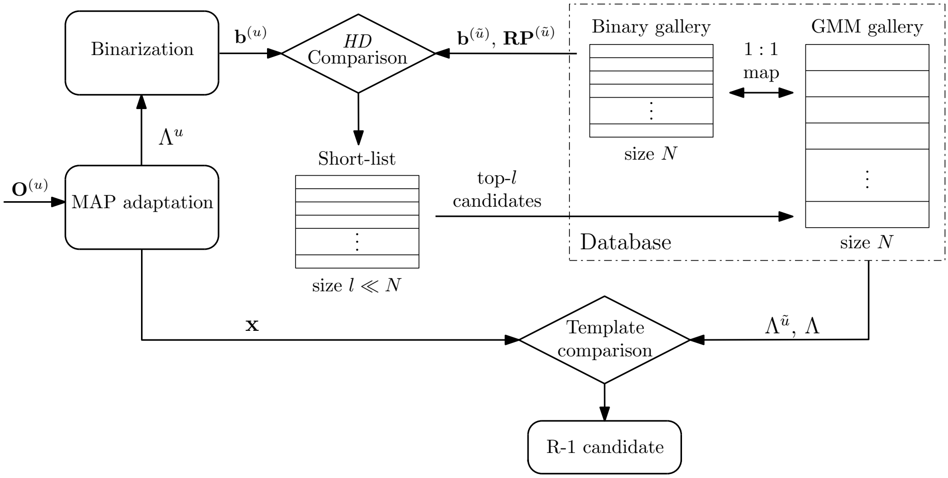


Figure 2: Basic operation mode of the proposed two-stage speaker identification system based on universal background models.

3 Two-stage Identification System

The proposed two-stage speaker identification system is depicted in Fig. 2 and comprises two components: pre-screening and template comparison. During pre-screening a computationally efficient Hamming distance-based comparator is applied to extracted binary templates in order to return a short-list of top candidates, which most likely match the presented probe. Subsequently, the original templates of the short-list candidates are considered for the more complex comparison of the original system. Both steps are described in detail in the following subsections.

3.1 Pre-screening

Comparisons between binary biometric probes and gallery templates are implemented by the simple Boolean exclusive-OR operator (XOR) applied to a pair of binary vectors, masked (AND'ed) by the relevant projection of the gallery template. The XOR operator $\oplus$ detects disagreements between any corresponding pair of bits while the AND operator $\cap$ ensures that only most discriminative bits (with respect to the gallery template) are considered. For a binary template $\mathbf{b}^{(u)}$ of user u, a gallery template $\mathbf{b}^{(\tilde{u})}$ of user $\tilde{u}$ and the corresponding relevant projection $\mathbf{RP}^{(\tilde{u})}$ we compute the fractional Hamming distance (*HD*) as a measure of the dissimilarity,

$$HD(\mathbf{b}^{(u)}, \mathbf{b}^{(\tilde{u})}) = \frac{||(\mathbf{b}^{(u)} \oplus \mathbf{b}^{(\tilde{u})}) \cap \mathbf{RP}^{(\tilde{u})}||}{v}, \tag{9}$$

where v is equal to the norm of $\mathbf{RP}^{(\tilde{u})}$, $v = ||\mathbf{RP}^{(\tilde{u})}||$.

The computational efficient HD-based comparator is utilized to pre-screen the entire database. For this purpose N pair-wise comparisons are performed, where N is the number of users registered with the identification system, resulting in a vector $\mathbf{S}$ of dissimilarity scores $s = HD$. Subsequently, scores in $\mathbf{S}$ are sorted in descending order to obtain the set $\mathbf{S}' = \{s_1, s_2, \dots, s_N | \forall i, j, i < j : s_i \leq s_j, \}$. Finally, the top-$l$ candidates, i.e. the candidates which the first l scores in $\mathbf{S}'$ point at, are returned.

3.2 Template Comparison

Based on the short-list returned in the pre-screening stage the probe is compared against a total number of l original gallery templates. The comparison score between the probe of user u and a gallery template of user $\tilde{u}$ is defined as the LLR of the user's GMM $\mathbf{\Lambda}^{\tilde{u}}$ and the UBM $\mathbf{\Lambda}$, which is defined as,

$$LLR(\mathbf{O}^{(u)}, \mathbf{\Lambda}^{\tilde{u}}, \mathbf{\Lambda}) = \sum_i \log\big(P(i|\mathbf{x}, \mathbf{\Lambda}^{\tilde{u}})\big) - \log\big(P(i|\mathbf{x}, \mathbf{\Lambda})\big), \tag{10}$$

where $\mathbf{\Lambda}^{\tilde{u}}$ represents the GMM of user $\tilde{u}$, $\mathbf{\Lambda}$ is the UBM, $\mathbf{w}_i$ are the mixture weights, $\boldsymbol{\mu}_i$ the means, and $\mathbf{x}$ the feature vectors extracted from $\mathbf{O}^{(u)}$. The a posteriori probability for each component $i = 1, \dots, I$ of Λ is defined analogue to Eq. 1,

$$P(i|\mathbf{x}, \mathbf{\Lambda}) = \frac{\mathbf{w}_i g(\mathbf{x}|\boldsymbol{\mu}_i, \sum_i)}{\sum_{j=1}^{J} \mathbf{w}_j g(\mathbf{x}|\boldsymbol{\mu}_j, \sum_j)}. \tag{11}$$

The LLR is a test of the hypothesis H_u, u is the target speaker, and the anti-hypothesis $H_{\overline{u}}$, u is not the target speaker. The resulting score can then be used as an ordering criterion by assuming that a higher score is more likely to be a genuine trial than an impostor trial.

3.3 Workload Reduction

Without the loss of generality, any serial combination of a conventional biometric identification system and a computationally more efficient system gains a linear speed up, with respect to the amount of comparisons performed. Assuming that a total number of N subjects are registered with the traditional identification system, the workload W for a single biometric identification can be defined by,

$$W = N(T_c + t_\varepsilon) + \delta, \tag{12}$$

where T_c represent the computational cost of a single comparison of the probe to a gallery instance, t_ε represents secondary computational costs, e.g. file access, and δ comprises all one-time secondary costs, e.g. sorting of scores or feature extraction performed on the acquired voice sample.

The proposed two-stage system reduces the overall workload to W',

$$W' = Nt_c + lT_c + \delta', \tag{13}$$

where t_c is the computational cost of a single more efficient comparison, $t_c \ll T_c$, and l is the number of top-candidates returned by the pre-screening process. Assuming that secondary computational costs are comparable, $\delta \simeq \delta'$, the overall computational cost is reduced, i.e. $W' < W$, if,

$$t_c < T_c(1 - \frac{l}{N}), \tag{14}$$

which is most likely the case for a computationally efficient pre-screening and small numbers of l, as will be demonstrated in conducted experiments. Even in case, $\delta \ll \delta'$, one-time costs become negligible for large numbers of N. That is, additional costs required for the proposed two-stage system, e.g. the MAP adaptation process, only slightly influence the response time of the identification system.

4 Experimental Evaluation

In the following subsection we define the experimental setup and compare the performance of the original identification system to the proposed two-stage scheme, with respect to biometric performance (identification rates) as well as time consumption.

4.1 Experimental Setup

Experiments are carried out on a text-independent digit-corpus database which comprises voice samples of a total number of 339 subjects. For each subject in the database at least 32 voice samples of length 3,000ms-5,000ms are available. The samples contain three to five spoken digits. At the time of enrolment 30 samples are used to generate the according models for a user. The remaining samples are applied in the identification process.

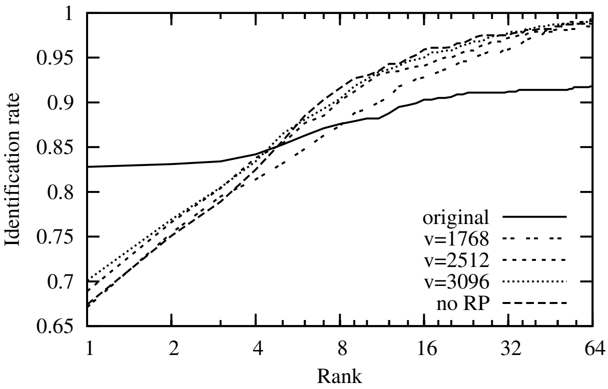
Performance is estimated in terms of (true-positive) identification rate (IR). In accordance with ISO/IEC IS 19795-1 [ISO06] the IR is the proportion of identification transactions by subjects enrolled in the system in which the subject's correct identifier is the one returned. In experiments, identification is performed in the closed-set scenario returning the rank-1 (R-1) candidate as the identified subject (without applying a specific decision threshold). By analogy, R-2 defines the proportion where the rank-1 or the rank-2 candidate represent the correct subject, and so forth. The CMC illustrates the progression of the identification rate with respect to the number of false positives.

Table 1: Identification rates of the original UBM/GMM system compared to different configurations of the pure proposed binarization.

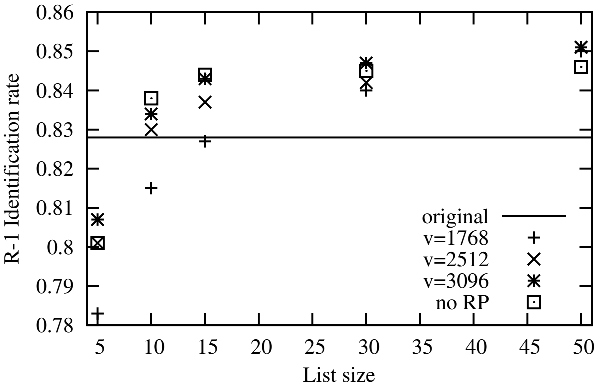
$I = 64$	Original	$v = 1024$	$v = 1512$	$v = 1768$	no $\mathbf{RP}$
R-1	72.0%	68.7%	71.3%	72.0%	67.4%
R-2	73.1%	**77.2%**	**77.8%**	**77.6%**	**76.1%**
R-5	75.1%	**85.7%**	**85.9%**	**86.2%**	**85.1%**
R-10	78.9%	**91.6%**	**91.9%**	**91.6%**	**90.6%**
$I = 128$	Original	$v = 1768$	$v = 2512$	$v = 3096$	no $\mathbf{RP}$
R-1	78.2%	72.9%	73.7%	72.9%	69.9%
R-2	79.0%	**79.3%**	**81.1%**	**80.0%**	78.9%
R-5	79.7%	**86.9%**	**88.2%**	**87.6%**	**86.4%**
R-10	84.3%	**92.6%**	**92.8%**	**92.3%**	**91.7%**
$I = 256$	Original	$v = 1768$	$v = 2512$	$v = 3096$	no $\mathbf{RP}$
R-1	82.8%	67.5%	68.0%	70.0%	67.4%
R-2	83.1%	75.0%	76.6%	76.8%	75.0%
R-5	83.4%	83.2%	**85.4%**	**86.4%**	**85.7%**
R-10	87.6%	**89.2%**	**92.2%**	**92.6%**	**92.9%**

4.2 Performance Evaluation

Table 1 summarizes identification rates obtained for different configurations of the GMM-UBM-based system as well as the binarization technique (rates where the binarization scheme outperforms the original one are marked in bold). In case no random projection is applied, the total amount of bits compared is $Z = 39I$. We did not consider models with more than $I = 256$ components, as for more components, e.g. $I = 512$, the response time of identification processes did not turn out to be practical for the original system, even for the applied database of limited size. Time consumption during operational testing caused significant delay in analysis of the resulting performance and optimizations. As can be observed from Table 1 the original system clearly outperforms the sole binarized system with respect to the R-1 identification rate. However, for the R-2 rate the binarization scheme already achieves comparable biometric performance and for R-5 and R-10 rates it even gains accuracy across different configurations. That is, the proposed binarization technique is highly suitable for pre-screening purposes. The CMCs of the best configuration of the original system ($I = 256$) and the presented binarization method utilizing different relevant projections are plotted in Fig. 4.1. As can be seen, for increasing rank values all configurations quickly outperform the original system. Further, the number of incorporated bits within the binarization scheme can be significantly reduced without a loss of accuracy. In several cases a reduction of the amount of applied bits to the v most discriminative ones even gains biometric performance. For example, for $I = 64$ the $v = 1768$ most relevant bits yield a R-1 rate of 72.0% compared to a R-1 rate of 67.4% if no $\mathbf{RP}$, i.e. $v = 2496$, is applied (see first row of Table 1).



(a) CMC of binarized systems



(b) R-1 rates of two-stage system ($I = 256$)

Figure 3: CMC curve for the original system and configurations of the binarized system and R-1 rates for configurations of the two-stage identification system for $I = 256$.

For different configurations, short-list sizes and adequate relevant projections, the R-1 rates of the proposed two-stage identification system ($I = 256$) are plotted in Fig. 4.1. Table 2 summarizes obtained rates for different numbers of components, accordingly (rates where the two-stage scheme outperforms the original one are marked bold). Obviously, large values of l improve the biometric performance while convergence is reached relatively quickly, e.g. employing a number of $I = 64$ components returning more than 15 candidates does not yield any improvement. For the vast majority of configurations the presented approach clearly enhances the biometric performance of the underlying system which obtains R-1 identification rates of 72.0%, 78.2%, and 82.8% for a number of $I = 64$, 128, and 256 components, respectively (cf. Table 1).

Experiments have been performed on a system consisting of an Intel Core i7-37770 CPU with 3.4 GHz and 32 GB RAM, running CentOS 6.3 x86_64. Comparisons were not par-

Table 2: R-1 rates for different configurations of the proposed two-stage identification system.

$I = 64$	$v = 1024$	$v = 1512$	$v = 1768$	no **RP**
$l = 5$	**73.9%**	**73.4%**	**73.7%**	**73.5%**
$l = 10$	**74.1%**	**74.3%**	**74.0%**	**73.6%**
$l = 15$	**74.0%**	**74.9%**	**74.4%**	**73.8%**
$l = 30$	**73.6%**	**73.6%**	**73.8%**	**73.4%**
$l = 50$	**73.4%**	**73.5%**	**73.4%**	**73.5%**
$I = 128$	$v = 1768$	$v = 2512$	$v = 3096$	no **RP**
$l = 5$	**78.3%**	**79.3%**	**78.6%**	77.8%
$l = 10$	**79.5%**	**79.8%**	**79.5%**	**79.1%**
$l = 15$	**79.7%**	**80.1%**	**80.2%**	**79.9%**
$l = 30$	**80.2%**	**80.0%**	**80.3%**	**79.8%**
$l = 50$	**80.1%**	**80.2%**	**80.2%**	**79.9%**
$I = 256$	$v = 1768$	$v = 2512$	$v = 3096$	no **RP**
$l = 5$	78.3%	80.1%	80.7%	80.1%
$l = 10$	81.5%	**83.0%**	**83.4%**	**83.9%**
$l = 15$	82.7%	**83.7%**	**84.3%**	**84.4%**
$l = 30$	**84.0%**	**84.2%**	**84.7%**	**84.5%**
$l = 50$	**85.0%**	**85.1%**	**85.1%**	**84.6%**

allelized during time consumption tests. While optimized biometric identification systems make use of parallel distributed data processing, this difference is irrelevant since we aim at comparing the two types of techniques based on same configurations and report speed-up in percentage, since absolute values of identification speed directly relate to the size of the dataset. For the applied dataset of 339 users the obtained overall speed-ups for different sizes of l are summarized in Table 3. As can be seen, computational performance is significantly improved achieving speed-ups up to 98%. As expected, a natural trade-off between computational performance and biometric performance (accuracy) is yielded (cf. Table 2). A pair-wise comparison within the original system takes on average 42.1 ms. The HD score between two binary templates is estimated in approximately 0.007 ms. Additional one-time computational cost for the MAP adaptation requires 62.4 ms on average. As previously mentioned, secondary computational cost, e.g. file access, are inevitable for both, the original system as well as the proposed two-stage scheme, which limits the overall performance gain. Furthermore, it becomes clear that, in contrast to the presented experiments, theoretical analysis, which estimate the amount of required operations of algorithm complexity, may cloud the picture of the actual speed-up. As can be observed from Table 3, with increasing length of the top-l list the overall speed-up decreases. However, with increasing database size computational performance is further gained, i.e. for a simulated workload of $N = 3,000$ users and a pre-screening short-list of $l = 100$ candidates a speed-up of 96% is obtained. That is, the proposed approach is expected to achieve even more performance gain for large-scale databases where speaker identification still represents a critical issue.

Table 3: Average speed-up across different component and short-list sizes.

List size l	5	10	15	30	50	75	100
Speed-up	98.0%	96.5%	95.1%	90.6%	84.7%	77.4%	70.0%

5 Conclusion

In this work we proposed a two-stage speaker identification system based on UBM. It has been demonstrated that the incorporation of a binarization technique enables a computationally efficient pre-screening of the database returning a short-list of top-candidates, which significantly reduces the computational workload of the entire system. The proposed approach is generic and an integration into existing systems is feasible at negligible cost (binary templates are stored efficiently). In contrast to hardware-oriented speed-up solutions, e.g. spreading the recognition workload across a set of processors, the presented scheme represents, to our knowledge, the first low-cost software solution for performing speaker identification on large-scale voice databases.

Future work will comprise applying binary voice templates within biometric template protection schemes [RU11] in order to protect the privacy of registered users.

Acknowledgement

This work has been partially funded by the Center for Advanced Security Research Darmstadt (CASED).

References

[AB10] X. Anguera and J.-F. Bonastre. A novel speaker binary key derived from anchor models. In *Proc. of INTERSPEECH'10*, pages 2118–2121. ISCA, 2010.

[BBMA11] J.-F. Bonastre, P.M. Bousquet, D. Matrouf, and X. Anguera. Discriminant binary data representation for speaker recognition. In *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'11)*, pages 5284–5287, 2011.

[GRC09] J. E. Gentile, N. Ratha, and J. Connell. An Efficient, Two-Stage Iris Recognition System. In *Proc. IEEE 3rd Int'l Conf. on Biometrics: Theory, Applications, and Systems*, pages 1–5. IEEE, 2009.

[ISO06] ISO/IEC TC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, March 2006.

[JFR08] A. K. Jain, P. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2008.

[JPDG08] U. Jayaraman, S. Prakash, Devdatt, and P. Gupta. An Indexing technique for biometric database. In *Proc. of Int'l Conf. on Wavelet Analysis and Pattern Recognition, (ICWAPR'08)*, pages 758–763, 2008.

[JRP04] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14:4–20, 2004.

[KL10] T. Kinnunen and H. Li. An Overview of Text-independent Speaker Recognition: From Features to Supervectors. *Speech Communication*, 52(1):12–40, 2010.

[LVH+11] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer. Survey and evaluation of acoustic features for speaker recognition. In *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'11)*, pages 5444–5447, 2011.

[MPCG05] A. Mhatre, S. Palla, S. Chikkerur, and V. Govindaraju. Efficient Search and Retrieval in Biometric Databases. In *Proc. of SPIE Defense and Security Symposium*, 2005.

[MZR96] R. J. Mammone, Xiaoyu Zhang, and R. P. Ramachandran. Robust speaker recognition: a feature-based approach. *IEEE Signal Processing Magazine*, 13(5):58–, 1996.

[PS01] J. Pelecanos and S. Sridharan. Feature Warping for Robust Speaker Verification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (ODYSSEY'01)*, pages 213–218, 2001.

[RB76] C. Rader and N. Brenner. A new principle for fast Fourier transformation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(3):264–266, 1976.

[Rey94] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Trans. on Speech and Audio Processing*, 2(4):639–643, 1994.

[RMACC12] E. A. Rua, E. Maiorana, J. L. Alba Castro, and P. Campisi. Biometric Template Protection Using Universal Background Models: An Application to Online Signature. *IEEE Trans. on Information Forensics and Security*, 7(1):269–282, 2012.

[RU11] C. Rathgeb and A. Uhl. A Survey on Biometric Cryptosystems and Cancelable Biometrics. *EURASIP Journal on Information Security*, 2011(3), 2011.

[SVC93] J. Smolders and D. Van Compernolle. In search for the relevant parameters for speaker independent speech recognition. In *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP'93)*, volume 2, pages 684–687 vol.2, 1993.

[YLML12] C. You, H. Li, B. Ma, and K.-A. Lee. Effect of Relevance Factor of Maximum a posteriori Adaptation for GMM-SVM in Speaker and Language Recognition. In *Proc. of INTERSPEECH'12*. ISCA, 2012.