State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software

Benoît G. B. Fauve, Driss Matrouf, Nicolas Scheffer, Jean-François Bonastre, *Senior Member, IEEE*, and John S. D. Mason

***Abstract*—This paper illustrates an evolution in state-of-the-art speaker verification by highlighting the contribution from newly developed techniques. Starting from a baseline system based on Gaussian mixture models that reached state-of-the-art performances during the NIST'04 SRE, final systems with new intersession compensation techniques show a relative gain of around 50%. This work highlights that a key element in recent improvements is still the classical maximum *a posteriori* (MAP) adaptation, while the latest compensation methods have a crucial impact on overall performances. Nuisance attribute projection (NAP) and factor analysis (FA) are examined and shown to provide significant improvements. For FA, a new symmetrical scoring (SFA) approach is proposed. We also show further improvement with an original combination between a support vector machine and SFA. This work is undertaken through the open-source ALIZE toolkit.**

***Index Terms*—Channel compensation, factor analysis, nuisance attribute projection, speaker verification.**

I. Introduction

As indicated by the growing number of participants in the international NIST speaker recognition evaluations (SREs) [1], text-independent automatic speaker verification (ASV) has experienced an increasing interest in recent years. At the same time, meaningful improvements in performance have been achieved with error rates roughly halving over the last three NIST SRE campaigns. These two phenomena are not directly linked, since the best performances invariably come from a limited number of sites, mostly with previous experience in such campaigns. And any exceptions to this observation tend to relate to site combinations, where perhaps one or more of the partners have previous SRE experience. The key point here is the high level of research and technology investment required to reach and remain close to the ever moving state of the art. In this context, combined effort across sites can clearly help; the simplest and most common example of this is the sharing of system scores with score-level fusion. At the other extreme of such cooperation is open-source software, the potential of which is far more profound.

This paper describes the most recent developments in ASV that lead to state-of-the-art performance as judged by the latest NIST SREs. The realization has taken place in the context of the open-source ALIZE toolkit [2], [3]. Key developments include support vector machines (SVMs) [4], [5], the associated nuisance attribute projection compensation (NAP) [6], and factor analysis (FA) [7], [8]. We trace individually their influences on ASV performance. As a starting point we choose a baseline that gave state-of-the-art performance in the NIST'04 SRE, namely the ALIZE-based LIA (University of Avignon) submission [9].

By incorporating one-by-one SVMs, NAP, and FA into the ALIZE environment, we are able to highlight interesting comparison and performance benefits. Of particular note here is the interpretation of FA; we highlight this since we believe the symmetrical realization presented here gives superior results to those published to date.

The contributions of this paper include the following:

- an account of the ASV evolution from 2004 to today in terms of improvement in accuracy that can be attributed to SVMs, NAP and FA;
- a comparative interpretation of the three approaches above;
- a new symmetrical variant of FA (SFA) that, for the given protocols and to the authors' knowledge, lead to results that match the best that have been published to date.

Whereas NIST publications such as [10]–[12] describe evolutions more in terms of raw results, where both systems and databases change each year, here we highlight the contributions of specific approaches, keeping the database usage constant: NIST'04 for background data, NIST'05 for development, and NIST'06 for evaluation. This helps in direct evaluation of the underlying technology, neutralizing to some extent database nuances.

Moreover, the dissemination aspect of this work is deemed of particular importance in relation to factor analysis. Its contribution to ASV accuracy proves to be indeed significant, yet there is relatively little in the open literature on FA (in relation to ASV) following on from the original work of Kenny [8].

The outline of this paper is as follows. In Section II, we describe the experimental protocols and conventions. Section III presents the baseline Gaussian mixture model (GMM) system submitted by LIA to NIST'04 SRE. In Section IV, we explain our implementation of SVM for speaker verification and present results with various sequence kernels. Session variability compensation is discussed in Section V, and implementations are

Manuscript received February 15, 2007; revised June 5, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mari Ostendorf.

B. G. B. Fauve and J. S. D. Mason are with Speech and Image Research, School of Engineering, Swansea University, Swansea SA2 8PP, U.K. (e-mail: b.g.b.fauve.191992@swansea.ac.uk; j.s.d.mason@swansea.ac.uk).

D. Matrouf and J.-F. Bonastre are with the University of Avignon LIA/CNRS, 84911 Avignon Cedex 9, France. (e-mail: driss.matrouf@lia.univ-avignon.fr; jean-francois.bonastre@lia.univ-avignon.fr).

N. Scheffer was with the University of Avignon LIA/CNRS, 84911 Avignon Cedex 9, France. He is now with the Star Lab, SRI International, Menlo Park, CA 94025 USA (e-mail: nicolas.scheffer@lia.univ-avignon.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.902877

presented in Section VI for NAP and Section VII for FA and its new SFA variant. In Section VIII, we discuss overall improvements and results are validated on the NIST'06 primary task before drawing conclusions in Section IX.

II. Experimental Setup and Conventions

A. ALIZE Toolkit

This work is centered on the ALIZE toolkit [3], developed by the LIA in the framework of the French Research Ministry Technolangue[1] programme. ALIZE comes from the collaboration done around the ELISA consortium [13], which grouped the efforts of several European laboratories in order to participate in NIST evaluation campaigns (mainly NIST-SRE and NIST Rich Transcription evaluation) from 1998 to 2004. ALIZE/SpkDet is a package within ALIZE tailored specifically to ASV; ALIZE/SpkDet was one of the chosen reference systems in the framework of the Biosecure Network of Excellence;[2] it was also used by several institution for NIST'06 SRE. Feature extraction comes from SPro [14]. All these softwares are distributed through open-source licences.

B. Experimental Protocol

The experimental protocol is fixed throughout the paper. The male part of the NIST'05 primary task (1conv4w-1conv4w) is used for development (DevSet). For this condition, one side of a 5-min conversation is given for testing and the same amount for training. All background training data, for the universal background model (UBM), T-norm [15], NAP, and FA come from the NIST'04 database. This procedure leaves the NIST'06 free for validation. The final comparisons are made on the NIST'06 core (required) condition which includes multiple languages (rather than the English only common condition).

Performances are assessed using DET plots and measured in terms of equal error rate (EER) and minimum of detection cost (minDCF). The cost function is calculated following NIST criteria [10].

III. GMM Baseline

A state-of-the-art GMM–UBM system was submitted to NIST'04 SRE by LIA and is described in [9]. Most of the configuration details described in [9] are maintained.

The chosen speech/nonspeech detection parameters result in an average of 30% of frames being retained on 5-min-long recordings where the speaker of interest speaks for 2.5 min in average. The features are based on 19 linear filter-bank derived cepstra. The original LIA baseline system [9] uses 16 static and 16 first-order delta, giving a feature order F of 32. The model size C is 512. Various combinations of static, delta, double delta, and delta energy lead to the results shown in Table I. We retain a configuration corresponding to $F = 50$, $19 + 19\Delta + 11\Delta\Delta + \Delta E$. The complete set of double deltas brings little or no improvement. Such a configuration has been validated by Swansea university (UWS) during the NIST'06 SRE. This arrangement is used throughout the paper as a base for comparison and is referred to as the "GMM baseline."

[1]http://www.technolangue.net/.

[2]http://www.biosecure.info.

TABLE I
Summary of Simple Cepstral Feature Enhancements, on Static, Delta, Double Delta, and Delta Energy on the LIA04 System. T-Norm (DevSet). The Last System is Chosen as the GMM Baseline for This Work

Feature input	**F**	**EER(%)**	**Relative improvement** (%)
$16 + 16\Delta$	32	9.64	*LIA04 baseline*
$16 + 16\Delta + \Delta E$	33	9.32	3.36
$19 + 19\Delta + \Delta E$	39	9.08	5.88
$19 + 19\Delta + 11\Delta\Delta + \Delta E$	50	8.67	10.08

IV. Sequence Kernels and Linear SVM

A. General Framework

The SVM approach offers an alternative classification strategy to the widely used GMM and has been investigated by many in the context of ASV; see for example [4], [16]–[18].

Here, the LibSVM library[3] has been used to integrate SVM functionalities into ALIZE. This accommodates sequence kernels defined as

$$K(X,Y) = \Phi(X)^t \mathbf{R}^{-1} \Phi(Y) \tag{1}$$

where $\Phi(X)$ is a high-dimensional vector representation of sentence X, and $\mathbf{R}^{-1}$ a diagonal normalization matrix. For the remainder of this paper, E refers to the size of the expansion Φ.

We roughly follow the training, model compaction, and testing methods described by Campbell *et al.* in [4]. As with sequence kernels, the "kernel trick" is moved into the definition of the sequence expansion; consequently, a linear SVM suffices with input vectors $\mathbf{R}^{-(1/2)}\Phi(X)$.

Recalling that ASV is a two-class problem, then all expansion vectors corresponding to a given speaker in the training mode are labeled, for example, $+1$ and are confronted individually by expansions from a cohort of other speakers (loosely termed the impostor cohort) with the label -1. The result of the training is the definition of a separating hyperplane

$$f(\mathbf{x}) = \sum_{i=1}^{N_{SV}} \alpha_i t_i \mathbf{R}^{-1/2} \Phi(X_i)\mathbf{x} + d \tag{2}$$

based on N_{SV} support vectors and where t_i represent the ideal output, $\sum_{i=1}^{N_{SV}} \alpha_i t_i = 0$, and d is an offset. Then, the classifier model can be compacted as

$$\mathbf{w}_X = \begin{bmatrix} \sum_{i=1}^{N_{SV}} \alpha_i t_i \mathbf{R}^{-1} \Phi(X_i) \\ d \end{bmatrix} \tag{3}$$

enabling the evaluation of $f(\mathbf{x})$ with a simple dot product. Indeed, in the testing phase, the expansion of the test segment is augmented by the value 1, and then a dot product between the two vectors of dimension $E + 1$ is performed to produce a verification score

[3]http://www.csie.ntu.edu.tw/~cjlin/libsvm.

$$\text{Score}_{\text{SVM}}(X,Y) = f\left(\mathbf{R}^{-1/2}\Phi(Y)\right) = \left[\Phi(Y)^t 1\right]\mathbf{w}_X. \quad (4)$$

Because $\mathbf{R}^{-1/2}$ is already integrated in $\mathbf{w}_X$, it is not required in the calculation of $f(\mathbf{R}^{-1/2}\Phi(Y))$.

This flexible framework can now be used for various popular SVM approaches to ASV; it requires only the definition of the kernel in (1) through

$$\begin{cases} X \rightarrow \Phi(X) \\ \mathbf{r}^{-1/2} \end{cases} \quad (5)$$

where $\mathbf{r}$ is the diagonal of $\mathbf{R}$ with dimension E.

B. Generalized Linear Discriminant Sequence Kernels

For the generalized linear discriminant sequence (GLDS) as described in [4], the polynomial expansion $\Phi_P(\mathbf{f}_t)$ is calculated. This leads to a vector of dimension $\binom{F+p}{p}$ made up of all possible monomials of degree p from the original feature vector $\mathbf{f}_t$ of size F.

The average polynomial expansion

$$\Phi_{\text{GLDS}}(X) = \overline{\Phi}_P(X) = \frac{1}{T_X}\sum_{t=1}^{T_X}\Phi_P(\mathbf{f}_t) \quad (6)$$

over the T_X frames in sentence X is taken, and

$$\mathbf{r}_{\text{GLDS}} = \frac{1}{N_c}\sum_{n=1}^{N_c}\frac{1}{T_n}\sum_{i=1}^{T_n}\Phi_P(\mathbf{f}_i).*\Phi_P(\mathbf{f}_i) \quad (7)$$

is calculated on the impostor cohort of size N_c and sentence size T_n. Each component of $\mathbf{r}$ is the mean of the square values over all frame expansions from the impostor cohort. An interesting observation is that $\mathbf{r}$ captures the dynamics of polynomial coefficients.

C. GMM Supervector Linear Kernel (GSL)

The development of metrics in GMM space [19] led to the idea of using Kullback–Liebler divergence to define new sequence kernels based on GMM supervectors [18].

A new expansion is defined as

$$\Phi_{\text{GSL}}(X) = \mathbf{m}_X = \begin{bmatrix} \mathbf{m}_X^1 \\ \cdots \\ \mathbf{m}_X^i \\ \cdots \\ \mathbf{m}_X^C \end{bmatrix} \quad (8)$$

which is the supervector comprising the values of means $\mathbf{m}_X^i$ taken from the GMMs, trained on utterance X. Each GMM has C components and, with an acoustic feature vector of size F, this gives a $\Phi_{\text{GSL}}(X)$ of size CF.

The weight and variance parameters from the universal background model (UBM) are used to define $\mathbf{r}$ with

$$\mathbf{r}_{\text{GSL}}^{-\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{\Sigma}_1^{-\frac{1}{2}} \\ \cdots \\ \sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-\frac{1}{2}} \\ \cdots \\ \sqrt{\lambda_C}\boldsymbol{\Sigma}_C^{-\frac{1}{2}} \end{bmatrix}. \quad (9)$$

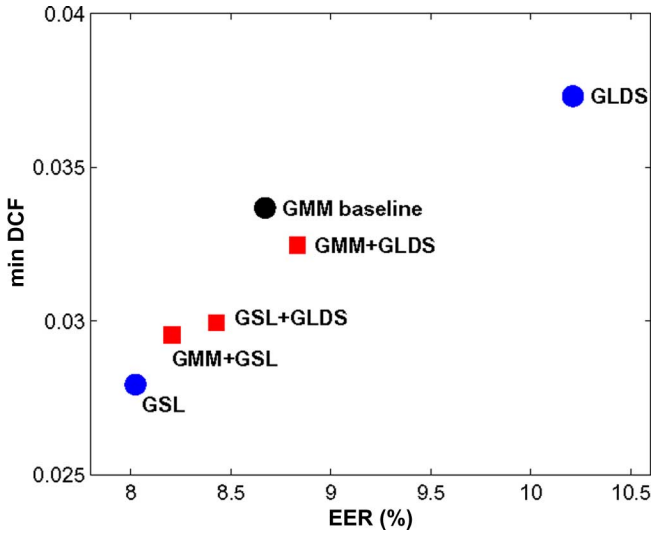


Fig. 1. EER and minDCF performances for GMM, GSL, and GLDS systems and the three associated direct fusions (DevSet).

It is possible to consider other sequence kernels using this general framework. For example, in [20], a weight-based Fisher kernel has been successfully tested with promising results.

D. Performance

We run a series of experiments designed to compare the performances of the GMM, GLDS, and GSL systems. All systems use the same test, train, GMM, and feature configurations. This point is important not only for direct comparisons of the three systems but also in assessing the potential of any fusion strategies. It is clear, for example, that additional benefits are likely to come from simultaneously tailoring different front-end cepstra to the three systems fusion.

The GMM baseline is used. So GMM models have $C = 512$ components and features of size $F = 50$. This leads to supervectors from the GMM means of size 25600. For the GLDS system, an expansion of degree $p = 3$ is used leading to a supervector size of $\binom{50+3}{3} = 23426$. The impostor cohort is based on 219 files from NIST'04 database.

Results are presented Fig. 1. Scores are T-normalized, and for the fusion cases, equal weights are applied.

Fig. 1 shows that the GSL system, harnessing the SVM and the supervectors of GMM maximum *a posteriori* (MAP) adapted means, gives the best performance. This arrangement gives the best scores overall and does not show any improvement from simple (unweighted) fusion with the other system scores. This would seem, perhaps, to cast some doubt on the idea that the generative GMM and the discriminative GLDS systems fuse well by bringing complementary information; though here we have only minimal evidence from just one configuration. Moreover, changing the front end of one of the systems is likely to show different fusion characteristics in line with that reported by Campbell *et al.* [21].

V. Session Variability Modeling

Some of the most important developments in ASV over recent years relate to strategies that address the "mismatch factor."

This term groups the effects linked to the differences between recording sessions due to transmission channel mismatch, additive noise, linguistic content, and speaker variability. Established techniques to handle these problems operate at either the feature level [22], [23], into the model [24], or at the score level with for example H-norm [25] and T-norm [15].

More recently, new approaches have been proposed by Kenny [7] with FA in a generative framework and by Solomonoff *et al.* [6] with NAP for the SVM framework. In these approaches, the goal is to directly model the mismatch rather than to compensate for their effects as it was done with H-norm and T-norm. This involves estimating the variabilities from a large database in which each speaker is recorded in multiple sessions. A somewhat similar idea is the feature mapping of [22] for channel effects, but contrary to this former approach which requires the background information to be labeled into a discrete number of channels, the new approaches of NAP and FA work on a continuous modeling of intersession information.

The underlying hypothesis is that a low-dimensional "session variability" subspace exists with only limited overlap on speaker-specific information. In the next section, we present NAP, and FA will be presented in Section VII.

VI. Nuisance Attribute Projection

A. NAP Framework

The goal of NAP is to project out a subspace from the original expanded space, where information has been affected by nuisance effects. This is performed by learning on a background set of recordings, without explicit labeling, from many different speakers' recordings. The most straightforward approach is to use the difference between a given session and the mean across sessions for each speaker. This information is now pooled across speakers to form a combined matrix. An eigenvalue problem is solved on the corresponding covariance matrix to find the dimensions of high variability for the pooled set. Given a set of recordings in their expansion form

$$\left\{\Phi_{(1,s_1)}\cdots\Phi_{(h_1,s_1)}\cdots\Phi_{(1,s_{N_s})}\cdots\Phi_{(h_{N_s},s_{N_s})}\right\} \tag{10}$$

from N_s different speakers with h_i different sessions for the ith speaker s_i the average expansion for each speaker is calculated and then removed from all the corresponding examples

$$\widetilde{\Phi}_{(l,s_i)} = \Phi_{(l,s_i)} - \overline{\Phi}_{s_i} \tag{11}$$

for $l \in [1\ h_i]$. The following matrix:

$$\mathbf{M} = \left[\widetilde{\Phi}_{(1,s_1)}\cdots\widetilde{\Phi}_{(h_1,s_1)}\cdots\widetilde{\Phi}_{(1,s_{N_s})}\cdots\widetilde{\Phi}_{(h_{N_s},s_{N_s})}\right] \tag{12}$$

of dimension $E \times N$ with $N = h_1 + \ldots + h_{N_s}$ represents all the intersession variations from the average speaker positions in the high-dimensional expended space.

To identify the subspace of dimension K where the variations are the largest, we then calculate the K eigenvectors with the highest eigenvalues of the corresponding covariance matrix $\mathbf{M}\mathbf{M}^t$. As this matrix tends to be large (of size $E \times E$), a robust inversion strategy is necessary. Here, singular value decomposition from SVDLIBC[4] toolkit has been used. The resulting eigenvectors are orthogonal and normalized. They form a base to the subspace where channel variations are the most pronounced and are concentrated into a matrix $\mathbf{S}$ of size $E \times K$.

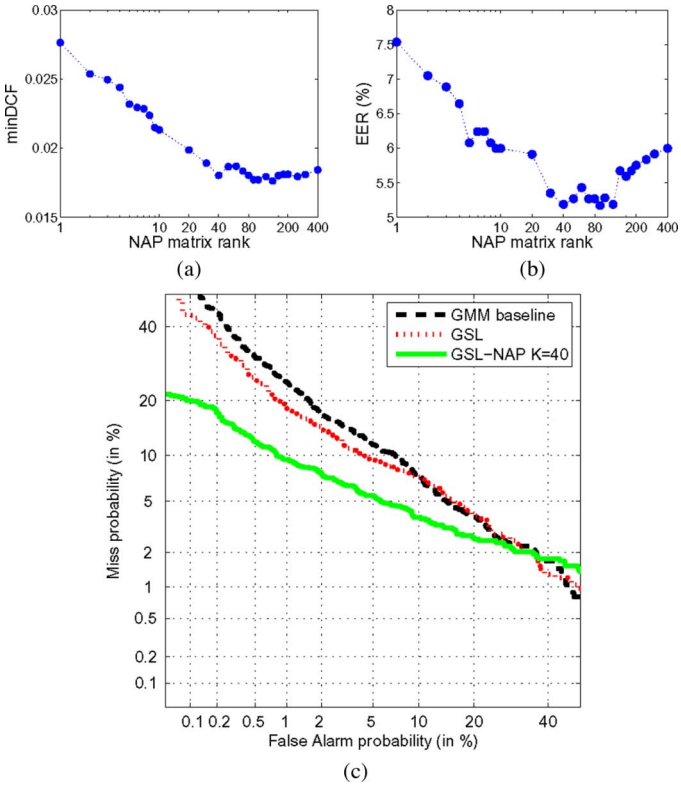


Fig. 2. NAP performances. (a) and (b) show evolution in terms of minDCF and EER, respectively, according to the rank of the NAP projection matrix. (c) shows DET plots with $\mathrm{K} = 40$ and original GSL.

In an ASV trial, this matrix is used on utterance X to project out the channel subspace

$$\hat{\Phi}_X = P(\Phi_X) = (\mathbf{I} - \mathbf{S}\mathbf{S}^t)\Phi_X \tag{13}$$

where $\mathbf{I}$ is the identity matrix. Once more the $\mathbf{S}\mathbf{S}^t$ matrix is of size $E \times E$, so the projection is calculated in the following manner for computational efficiency:

$$\hat{\Phi}_X = \Phi_X - \mathbf{S}(\mathbf{S}^t\Phi_X). \tag{14}$$

The resulting vectors are then used in a similar SVM framework to that described in Section IV.

B. Performance

Experiments are conducted with exactly the same setup as for the GSL in Section III. The NAP matrix is learned on the complete male set of NIST'04 database. After removing a few speechless files, we have 2938 training sessions for 124 different male speakers which lead to an average of 23.7 recordings per speaker.

Fig. 2 shows results in terms of minDCF and EER against the projection matrix rank. A DET plot in Fig. 2(c) is also given for direct comparison with the GSL.

[4]http://www.tedlab.mit.edu/~dr/SVDLIBC/.

The level of improvement attributable to NAP is large especially in the low false alarm probability zone with the minDCF values falling from 0.0279 to 0.0169. Fig. 2(a) and (b) show evolution of, respectively, EER and minDCF according to the rank of the NAP matrix. With a logarithmic scaling for the rank we observe an exponential improvement up to rank values around 40 and then relatively stable performance for higher values, even though the EER does increase again from 140 onward.

VII. Factor Analysis

We now introduce factor analysis which shares a similar concept to that of NAP. Introduced into speaker verification by Kenny [7], [8] and Vogt [26], it operates on generative models with traditional statistical approaches (such as EM) to model intersession variabilities.

A. Model Decomposition

With h indicating the session-dependent terms and s indicating the speaker-specific terms, then the speaker model can be viewed in the mean supervector space as a combination of three different components: 1) speaker-session-independent component $(\mathbf{m})$, 2) speaker-dependent component $(\mathbf{D}\mathbf{y}_s)$, and 3) session-dependent component $\mathbf{U}\mathbf{x}_{(h,s)}$. The speaker-session model can be written as

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)} \tag{15}$$

where $\mathbf{m}$ is the mean supervector coming from the UBM, D is a $CF \times CF$ diagonal matrix, and $\mathbf{U}$ is a $CF \times K$ matrix. Both the speaker offset $\mathbf{y}_s$ and session offset $\mathbf{x}_s$ are assumed to follow a standard normal distribution $\mathcal{N}(0, I)$. The key point of factor analysis is to isolate these three basic components. The training set is important, as it should include many examples to identify accurately the subspace where intersession variability is high. Given the fixed parameters $(\mathbf{m}, \mathbf{D}, \mathbf{U})$, a training algorithm is designed to perform the decomposition of the three basic components, while maximizing the likelihood of the data given the decomposition. Notice that in the case where $\mathbf{U} = 0$, the speaker-session model corresponds to the classical MAP adaptation.

The matrix $\mathbf{U}$ as commented in [5] is theoretically similar to the channel matrix $\mathbf{S}$ of NAP [see (13) in Section VI] in the way it captures intersession variability and so highlights the same subspace.

B. Verification Score Computation

Traditionally, in the GMM–UBM approach, the target speaker GMM is derived from the UBM model by updating only the mean parameters using a MAP algorithm. Given a segment of speech Y to test and a targeted speaker s with a model learned from speech T, the speaker verification score is estimated as

$$\text{score} = \frac{llk\left(Y|\mathbf{m}_{(h_T,s_T)}\right)}{llk(Y|\mathbf{m})} \tag{16}$$

where $llk(Y|\mathbf{m})$ is the average of the likelihood function from the GMM defined by $\mathbf{m}$ over all frames from Y.

By using the session factor analysis decomposition, we obtain

$$\mathbf{m}_{(h_Y,s_Y)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{s_Y} + \mathbf{U}\mathbf{x}_{h_Y} \tag{17}$$

and

$$\mathbf{m}_{(h_T,s_T)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{s_T} + \mathbf{U}\mathbf{x}_{h_T}. \tag{18}$$

To compensate the session component in the score computation, two strategies are possible.

1) Traditional Approach: In the first strategy, the test speaker is assumed to have the same identity as the target. In this case, $\mathbf{y}_{s_Y}$ (speaker component in test) is not estimated, but it is assumed to be equal to the speaker component in the target speaker $\mathbf{y}_{s_T}$. Only the channel component $(\mathbf{U}\mathbf{x}_{h_Y})$ is estimated in the test Y. To compensate the channel mismatch while calculating the score, the channel component in the target mean supervector $(\mathbf{U}\mathbf{x}_{h_T})$ is replaced by the one estimated in the test $(\mathbf{U}\mathbf{x}_{h_Y})$. The vector $\mathbf{m}$ from the UBM in the score equation remains unchanged. This strategy is adopted by Kenny [7] and Vogt [26] and has provided good results during the last NIST campaigns. However, it shows some skew. In practice, the world model needs to be compensated for session mismatch in the same way as the target model. If a world model compensation is not applied, a negative difference between the likelihoods of the test data given the target model and given the word model can be observed: the likelihood estimated on the target model can be larger than the one estimated on the world model. This observation helps explaining why both [7] and [26] obtain better results when some score normalization techniques (like ZT-norm) are applied.

2) Alternative Approach: To take into account this observation, we propose [27] an alternative strategy in which all the sessions are considered and treated separately. For each session, the session mismatch and the speaker component are estimated independently of all other sessions. So, the channel mismatch can simply be eliminated from each session. However, the session mismatch is estimated in the model space, and the session compensation (for the test) must be performed in the feature space. To do the latter, we adopt a strategy used in Vair [28], namely

$$\hat{\mathbf{t}} = \mathbf{t} - \sum_{g=1}^{M} p(g|\mathbf{t})\mathbf{U}_g\mathbf{x}_{h_Y} \tag{19}$$

where $\mathbf{t}$ is a frame of size F, $p(g|\mathbf{t})$ is the Gaussian occupation probability of the component g, and $\mathbf{U}_g$ is a subset of $\mathbf{U}$ corresponding to g. Hence, two options are available in order to compensate the session mismatch.

- 2.1) *Feature Space Compensation*. All the compensations are performed in the feature space. This option is interesting because it operates in the feature space and is independent of the classifier.
- 2.2) *Symmetrical Compensation*. The target models are compensated by eliminating the session mismatch directly in the model and the compensation in the test is performed in the feature space. We call this new approach symmetrical factor analysis (SFA).

TABLE II
EER(%) AND $\mathrm{minDCF}(\times 100)$ FOR SYMMETRICAL FA FOR RANK K BETWEEN 0 AND 100 OF MATRIX $\mathbf{U}$. NO T-NORM (DEVSET)

K	0	20	30	40	50	60	100
minDCF	3.95	2.14	2.08	1.97	2.02	2.04	2.00
EER	7.07	4.84	4.42	4.08	4.62	4.22	4.23

TABLE III
PERFORMANCE EVOLUTION IN EER(%) AND $\mathrm{minDCF}(\times 100)$. GMM BASELINE AND FA IMPLEMENTATION WITH FEATURE (FA FEAT) AND SYMMETRICAL (SFA) APPROACHES (DEVSET)

	devSet	
	minDCF	EER
GMM baseline	3.37	8.67
FA feat no norm	2.13	**4.05**
FA feat T-norm	2.09	4.70
SFA no norm	**1.96**	**4.13**
SFA T-norm	**1.94**	4.38

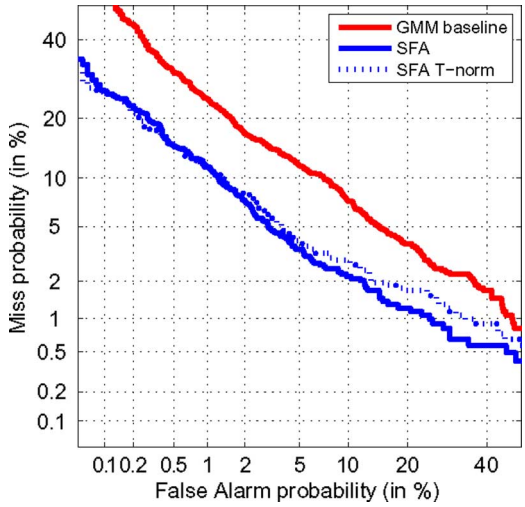


Fig. 3. DET plots for 1) GMM baseline and symmetrical FA with rank $\mathrm{K} = 40$ and 2) without and with T-normalization (DevSet).

C. Performance

We now present a set of results relating to the different strategies in the FA framework.

As in Section V, for the NAP, we present performance evolution with the new SFA approach according to the intersession matrix rank (Table II). For computational reasons, results are given on fewer rank values and without T-norm. As with NAP, good results are obtained in the region of the matrix $\mathbf{U}$ rank $K = 40$.

Table III provides results with strategies 2.1 and 2.2 with and without T-norm. Both approaches have similar performance, but SFA performance is slightly better. In particular, SFA provides the best results without the need of further T-normalization, confirming the potential benefit of the new approach. This is clearly shown in Fig. 3, which plots the GMM baseline and SFA with and without T-norm. When SFA is compared to the baseline, the relative reduction of EER is about 50%.

VIII. ANALYSIS AND FINAL RESULTS

In this section, we first interpret the evolution of the state of the art over the last three NIST SREs. We then look at potential complementarity between FA and SVM-NAP and show further improvements by combining the GSL kernel with FA.

A. Evolution in Results for State-of-the-Art Performances

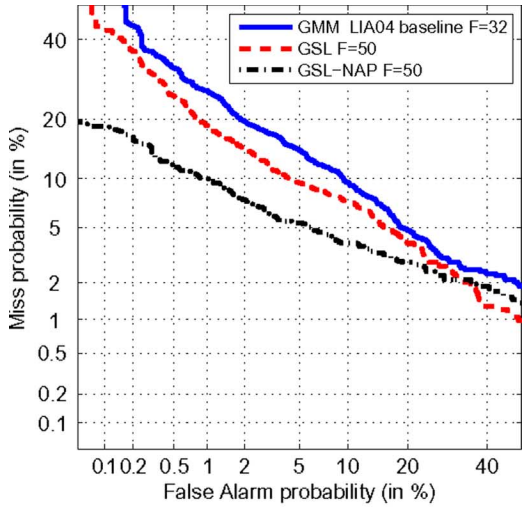
We present a series of systems described above. Fig. 4(a) shows results on the male-only development set. Fig. 4(b) and (c) gives equivalent results on NIST'06 SRE 1conv4w-1conv4w all language core condition, respectively, for male and for both genders.

The results in Fig. 4 and the performance values in Table IV show significant improvements gained from NAP and SFA techniques over the GMM system with, for example, minDCF roughly halving from GMM32 to GSL-NAP in each of the three columns. We notice a database difference between NIST'05 SRE and NIST'06 SRE. In NIST'06, improvement from session compensation occurs as well in the high false alarm probability zone of the DET plot. Beyond this difference, there is a constant improvement in the series of systems presented. Fig. 4(b) (NIST'06 male only) and Fig. 4(c) (NIST'06 both genders) show similar trends suggesting no gender dependency within the intersession compensation approach.

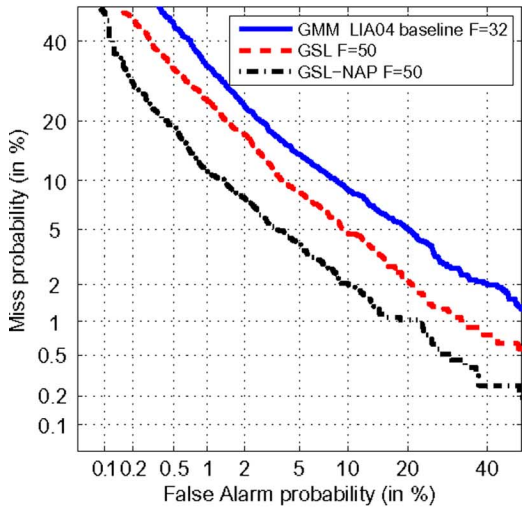
We now comment on the evolution in state-of-the-art performances over the last three NIST campaigns by highlighting key innovative development. In each case, we consider only a single system, thus avoiding the need to review the potential benefits of fusion. The EER and minDCF values relate to the NIST'06 SRE core, required condition, combined male and female, multiple languages. An overview of results for this SRE is available from [12], and this shows a small number of sites with EER in the region of 5% with two sites just below this value. The notion of state of the art is somewhat subjective, but if in this case we take as a definition the fact that it corresponds to good performances reproduced by a few sites, then we can illustrate the evolution as follows:

- **04 SRE**: Standard GMMs with T-norm widely used, e.g., *LIA04 baseline* (EER $= 9.92\%$; minDCF $= 0.0425$);
- **05 SRE**: New ideas beyond the conventional GMM framework (SVM, feature mapping,...), e.g., GSL $F = 50$ (EER $= 7.20\%$; minDCF $= 0.0335$);
- **06 SRE**: Effective exploitation of session variability through NAP and FA., e.g., GSL-NAP (EER $= 5.02\%$; minDCF $= 0.0226$).

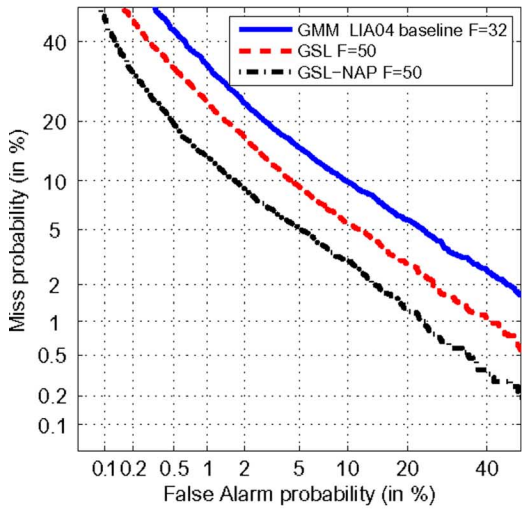
This interpretation of performance evolution highlights individual systems. Concurrent with these developments, much work has been done on fusion, an important but separate topic. Over the same three-year period, NIST SREs have seen a large growth in fusion submission incorporating score-based fusion with good effect. Here, we have focused more on the individual components and their contributions to improved performance.



(a)



(b)



(c)

Fig. 4. DET plots on three databases. (a) devSet. (b) NIST'06 core condition male only. (c) NIST'06 core condition male and female; each comparing GMM LIA04 baseline, GSL, and GSL-NAP.

The final subsection considers fusion of the two best systems, namely SFA and GSL–NAP.

TABLE IV
PERFORMANCE EVOLUTION IN EER(%) AND $\min\mathrm{DCF}(\times 100)$ FROM THE LIA04 BASELINE GMM TO LATEST COMPENSATION DEVELOPMENTS ON MALE 05, MALE 06, AND MALE AND FEMALE COMBINED 06

	male 05		**male 06**		**all 06**	
	DCF	EER	DCF	EER	DCF	EER
LIA04 baseline	3.51	9.64	4.21	9.36	4.23	9.88
GMM baseline	3.37	8.67	3.94	8.47	4.04	9.14
GSL	2.79	8.02	3.37	6.88	3.35	7.20
GSL-NAP	**1.62**	5.28	2.07	4.33	2.26	5.02
SFA	1.94	**4.38**	2.17	4.78	-	-

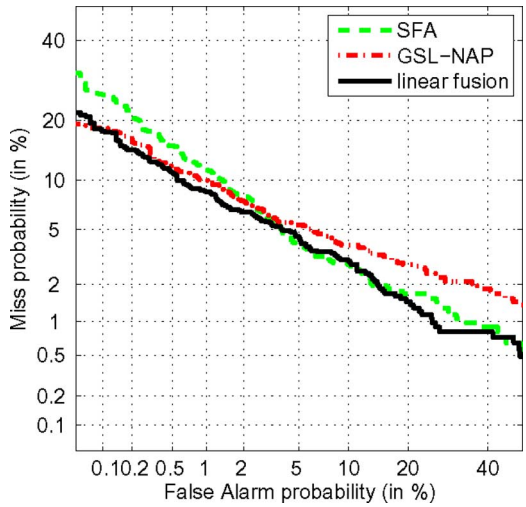


Fig. 5. DET plots for GSL–NAP, SFA, and their unweighted linear fusion, rank $= 40$, T-norm (devSet).

B. FA and NAP Combination and GSL on FA Supervectors

In Sections VI and VII, we have presented the current two main approaches for intersession compensation. Fig. 5 shows that even when operating with the same features, UBM, background intersession information, and rank, they bring improvement at different points of the DET plot. An unweighted linear fusion between these systems with T-norm shows their complementarity.

A way to exploit this complementarity is to use the discriminative SVM approach on mean supervectors $\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s$ evaluated in the factor analysis framework. Table V shows that such a system (FA–GSL) tends to follow the performance of the fusion between FA and GSL–NAP.

IX. CONCLUSION

Accuracy in ASV over recent years has improved significantly as judged by the last three NIST SREs [12]. Of course, during this interval, interest in biometrics more generally has increased as has the number of participants in NIST-style controlled evaluations. With increased interest comes increased competition and improved performances. To stay abreast of such advances inevitably calls for greater resources and investment. This can be ameliorated by cooperation across participants. Another route with enormous potential is open-source software. The work described in this paper has been achieved

TABLE V
EER(%) AND minDCF($\times 100$) FOR GSL-NAP, SFA, FUSION, AND GSL–SVM ON FA, rank $= 40$, T-NORM (DEVSET AND NIST'06 CORE MALE ONLY)

	male 05		**male 06**	
	minDCF	EER	minDCF	EER
GSL-NAP	1.62	5.28	2.07	4.33
SFA	1.94	**4.38**	2.17	4.78
fusion	**1.53**	4.70	**1.92**	**3.95**
FA-GSL	1.57	4.43	1.99	3.97

using the ALIZE/SpkDet open-source toolkit. The paper traces advances in NIST SREs by realizations of classifier (SVM) and session variability (NAP and FA) strategies which can account for the above EER improvements. These are evaluated one by one with the paper presenting original and directly comparative assessments. These illuminate the advances that have underpinned the halving of error rates mentioned above. Whereas NIST publications [12] can only trace improvement over years by showing best results of a year on the database of the year (with no account of database dependency), we highlighted evolutions of systems over the last three years and assess them on the two latest NIST databases. In addition, symmetrical FA, a refinement to factor analysis, is presented which acts independently on both the speaker model and the test frames. A traditional GMM scoring can be used with the merit to reduce the importance of score normalization. Performances are presented which are at least as good as those published to date under the given NIST SRE conditions. Finally, complementarity between SVM–NAP and FA approaches has been highlighted and a solution combining FA with GSL–SVM has been proposed. This work constitutes a solid base for further comparison between popular methods (FA and NAP) in today's ASV.

ACKNOWLEDGMENT

The authors would like to thank P. Kenny, N. Brümmer, and D. van Leeuwen for their practical interpretations of channel compensation during various workshops and seminars.

REFERENCES

[1] A. Martin and M. Przybocki, "The NIST Speaker Recognition Evaluation Series," National Institute of Standards and Technology [Online]. Available: http://www.nist.gov/speech/tests/spk
[2] ALIZE: Open Tool for Speaker Recognition [Online]. Available: http://www.lia.univ-avignon.fr/heberges/ALIZE/
[3] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. ICASSP*, 2005, pp. 737–740.
[4] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, pp. 210–229, 2006.
[5] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, pp. I-97–I-100.
[6] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, 2005, pp. 629–632.
[7] P. Kenny and P. Demouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.
[8] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP*, 2005, pp. 637–640.
[9] J.-F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf, "NIST'04 speaker recognition evaluation campaign: New LIA speaker detection platform based on ALIZE toolkit," in *NIST SRE'04 Workshop: Speaker Detection Evaluation Campaign*, Toledo, Spain, Jun. 2004 [Online]. Available: http://www.lia.univ-avignon.fr/fich_art/611-SRE04-LIA-v2.pdf
[10] A. Martin and M. Przybocki, "NIST speaker recognition evaluation chronicles," in *Proc. Odyssey*, 2004, pp. 15–22.
[11] M. Przybocki, A. Martin, and A. Le, "NIST speaker recognition evaluation chronicles—Part 2," in *Proc. Odyssey*, 2006, pp. 1–6.
[12] M. Przybocki, A. Martin, and A. Le, "NIST speaker recognition evaluation chronicles—Part 2," in *Odyssey Workshop Presentation* [Online]. Available: http://www.speakerodyssey.com/templates/13.pdf
[13] The ELISA consortium, "The ELISA consortium. the ELISA systems for the NIST'99 evaluation in speaker detection and tracking," *Digital Signal Process.*, vol. 10, pp. 143–153, 2000.
[14] G. Gravier, SPro: Speech Signal Processing Toolkit [Online]. Available: http://www.gforge.inria.fr/projects/spro
[15] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, p. 4254, 2000.
[16] J. Mariéthoz and S. Bengio, "A Kernel trick for sequences applied to text-independent speaker verification systems," IDIAP, IDIAP-RR 77, 2005.
[17] V. Wan, "Speaker verification using support vector machines," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 2003.
[18] W. M. Campbell, D. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.
[19] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *Proc. ICSLP*, 2004, pp. 2329–2332.
[20] N. Scheffer and J.-F. Bonastre, "A UBM-GMM driven discriminative approach for speaker verification," in *Proc. Odyssey*, 2006, pp. 1–7.
[21] W. M. Campbell, D. A. Reynolds, and J. P. Campbell, "Fusing discriminative and generative methods for speaker recognition: Experiments on switchboard and NFI/TNO field data," in *Proc. Odyssey*, 2004, pp. 41–44.
[22] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, 2003, pp. II-53–II-56.
[23] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey*, 2001, pp. 213–218.
[24] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Proc. ICASSP*, 1997, pp. 1071–1074.
[25] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," in *Proc. ICASSP*, 1996, pp. 113–116.
[26] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *Proc. ICASSP*, 2006, pp. I-987–I-900.
[27] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech 2007*, accepted for publication.
[28] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. Odyssey*, 2006, pp. 1–6.

Benoît G. B. Fauve received the M.S. degree in physics from the Ecole Nationale Supérieure de Physique de Marseille (ENSPM) (now Centrale) Marseille, France, in 2001. He is currently pursuing the Ph.D. degree at Swansea University, Swansea, U.K., working on speaker verification.

In 2001, he joined Neuvoice, Ltd., as a Speech and Acoustic Monitoring Research Engineer. He collaborated on several projects related to voice biometrics in the framework of the BIOSECURE European Network of Excellence.

Driss Matrouf received the Ph.D. degree in noisy speech recognition from the LIMSI Laboratory, Paris IX University, Paris, France, in 1997.

He then joined the University of Avignon (LIA), Avignon, France as an Associate Professor. His research interests include speech recognition, language recognition, and speaker recognition. His current research interests concentrate on session and channel compensation for speech and speaker recognition. In parallel with this research activities, he teaches at LIA in the fields covering computer science, speech coding, and information theory.

Nicolas Scheffer received the M.S. degree in control systems and applied computer science from the Ecole des Mines, Nantes, France, in 2002 and the Ph.D. degree in computer science from the University of Avignon (LIA), Avignon, France, in 2006.

He is currently a Postdoctoral Fellow in the Star Lab, SRI International, Menlo Park, CA. His research interests includes text-independent speaker verification, identification, and tracking. He has also been involved in the development of the ALIZE toolkit at the LIA, an open-source software in the field of speaker detection and diarization. He was also part of the BIOSECURE European Network of Excellence that focus on multimodal biometrics.

Jean-François Bonastre (SM'05) received the Ph.D. degree in automatic speaker identification using phonetic-based knowledge from the University of Avignon (LIA), Avignon, France, in 1994 .

He then joined the LIA as an Associate Professor. In 2006, he became a member of the "Institut Universitaire de France." As a member of the Natural Language Processing Group, he developed his research in speaker characterization and recognition using phonetic, statistic, and prosodic information while teaching and lecturing on various subjects covering computer science, speech processing, audio signal classification and indexing, and biometry. In 2002, he spent one year as an Invited Professor with Panasonic Speech Technology Laboratory (PSTL).

Dr. Bonastre was the chairman of AFCP, the French-Speaking Speech Communication Association (currently a regional branch of ISCA) from 2001 to 2004. He was also the chairman of the ISCA SPLC (SPeech and Language Characterization) SIG for two years, and he joined the board of ISCA in September 2005. He organized the RLA2C ISCA/IEEE workshop in 1998 and since then has participated in the Program Committee of the "Speaker Odyssey" series of workshops.

John S. D. Mason received the M.Sc. and Ph.D. degrees in control and digital signal processing from the University of Surrey, Surrey, U.K., in 1973.

He is with the Speech and Image Group, School of Engineering, Swansea University, Swansea, U.K., where he is currently a Senior Lecturer. He has supervised over 30 Ph.D. students in signal processing, mostly in speech and speaker recognition.