A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification

Driss Matrouf[1], *Nicolas Scheffer*[1], *Benoit Fauve*[2], *Jean-François Bonastre*[1]

[1]LIA, University of Avignon, France
[2]Speech and Image Research Group, University of Wales Swansea, UK

Abstract

For a few years, the problem of session variability in text-independent automatic speaker verification is being tackled actively. A new paradigm based on a factor analysis model have successfully been applied for this task. While very efficient, its implementation is demanding. In this paper, the algorithms involved in the eigenchannel MAP model are written down for a straightforward implementation, without referring to previous work or complex mathematics. In addition, a different compensation scheme is proposed where the standard GMM likelihood can be used without any modification to obtain good performance (even without the need of score normalization). The use of the compensated supervectors within a SVM classifier through a distance based kernel is also investigated. Experiments results shows an overall 50% relative gain over the standard GMM-UBM system on NIST SRE 2005 and 2006 protocols (both at the DCFmin and EER).

Index Terms: Speaker Verification, Session variability, Factor Analysis, GMM Supervectors.

1. Introduction

The use of Gaussian Mixture Models (GMM) in a GMM-UBM framework has been a standard in the speaker verification [1]. For a couple of years, new techniques that take session variability (or speaker intra-variability) into account have emerged, allowing to increase system performances drastically.

Among the approaches aiming at reducing the effect of session variability, feature mapping was often used alongside channel-labelled data, with the assumption of a discrete channel space. The novelty brought by the factor analysis model is that it assume the channel (or session) variability space to be continuous. In this model, the session variability effect is incorporated in the speaker model through session-dependent GMM mean supervectors offsets, constrained in a low dimensional subspace.

The work in this paper is based on a theoretical framework proposed by Kenny [2], and the reduced model in [3], called eigenchannel MAP. This paper has two main objectives. The first one is to propose a straightforward implementation of the factor analysis model. This model involves complex mathematics, and a lot of referral to other works if someone is willing to implement it. This paper can be seen as auto-sufficient in order to implement the whole approach. The second objective is to propose a new compensation scheme. In contrast to previous works, the presented strategy aims to subtract the session effect in both training and testing data, whereas others focus in compensating the target model with the test session effect. Therefore, several advantages emerge: a highly performing system without the need of score normalization, a standard GMM likelihood function that can be applied without any modification and the possibility of the resulting supervectors to be use directly in a SVM classifier. This association between the factor analysis and SVM allows to benefit from the FA decomposition power and SVM classification power.

The paper is organized as follows: in section 2, the factor analysis decomposition (eigenchannel MAP estimation) used in this work is described. In section 3, estimation formulas and algorithms are given to allow a direct implementation of this model. In section 4, the adopted strategy for session compensation is detailed as well as the score computation process. Section 5 details the experimental protocol while section 6 presents the results of the approach.

2. Session variability modeling

A speaker model can be decomposed into three different components: a speaker-session-independent component, a speaker dependent component and a session dependent component. A GMM mean supervector is defined as the concatenation of the GMM component means. Let D be the dimension of the feature space, the dimension of a supervector mean is MD where M is the number of Gaussian in the GMM. A speaker and session independent model is usually estimated in speaker verification to represent the inverse hypothesis: the UBM model. Let this model being parameterized by $\theta = \{\mathbf{m}, \mathbf{\Sigma}, \alpha\}$. In the following, (h, s) will indicate the session h of the speaker s. The factor analysis model, in our case the eigenchannel MAP estimator, can be written as:

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \tag{1}$$

where $\mathbf{m}_{(h,s)}$ is the session-speaker dependent supervector mean, $\mathbf{D}$ is $MD \times MD$ diagonal matrix, $\mathbf{y}_s$ the speaker vector (a MD vector), $\mathbf{U}$ is the session variability matrix of low rank R (a $MD \times R$ matrix) and $\mathbf{x}_{(h,s)}$ are the channel factors, a R vector (theoretically $\mathbf{x}_{(h,s)}$ does not dependent on s). Both $\mathbf{y}_s$ and $\mathbf{x}_{(h,s)}$ are normally distributed among $\mathcal{N}(0, I)$. $\mathbf{D}$ satisfies the following equation $\mathbf{I} = \tau\mathbf{D}^t\mathbf{\Sigma}^{-1}\mathbf{D}$ where τ is the *relevance factor* required in the standard MAP adaptation ($\mathbf{D}\mathbf{D}^t$ represents the *a priori* covariance matrix of $\mathbf{y}_s$).

3. Straightforward implementation of the factor analysis model

The success of the factor analysis model relies on a good estimation of the $\mathbf{U}$ matrix, thanks to a sufficiently high amount of data, where a high number of different recordings per speaker is available.

Notations: Let $\mathbf{A}$ be a $MD \times K$ matrix formed by concatenating vertically M matrices of dimensions $D \times K$. Let us denote $\{\mathbf{A}\}_{[g]}$ the g^{th} matrix in $\mathbf{A}$ (usually corresponding to the g^{th} component in the model).

3.1. General statistics

General statistics on the data have to be computed to estimate the latent variables and the $\mathbf{U}$ matrix of equation 1. These are the zero order and first order statistics with respect to the UBM model[1]. Let $\mathbf{N}_s$ and $\mathbf{N}_{(h,s)}$ be vectors containing the zero order speaker-dependent and session-dependent statistics respectively (both of dimensions M for a particular utterance), precisely:

$$\mathbf{N}_s[g] = \sum_{t\in s} \gamma_g(t);\ \mathbf{N}_{(h,s)}[g] = \sum_{t\in(h,s)} \gamma_g(t), \qquad (2)$$

where $\gamma_g(t)$ is the *a posteriori* probability of Gaussian g for the observation t. In the equation $\sum_{t\in s}$ means the sum over all frames belonging to the speaker s and $\sum_{t\in(h,s)}$ means the sum over all frames belonging the session h of speaker s. Let $\mathbf{X}_s$ and $\mathbf{X}_{(h,s)}$ be respectively vectors containing the first order speaker-dependent and session-dependent statistics. For an utterance, the dimensions of $\mathbf{X}_s$ and $\mathbf{X}_{(h,s)}$ are equal to $M \times D$:

$$\{\mathbf{X}_s\}_{[g]} = \sum_{t\in s} \gamma_g(t)\cdot t;\ \{\mathbf{X}_{(h,s)}\}_{[g]} = \sum_{t\in(h,s)} \gamma_g(t)\cdot t \qquad (3)$$

3.2. Latent variables estimation

In the following, the estimation of the channel factors and the speaker vector are given. Let $\overline{\mathbf{X}}_s$ and $\overline{\mathbf{X}}_{(h,s)}$ be respectively speaker and channel dependent statistics defined as follows:

$$\begin{aligned} \{\overline{\mathbf{X}}_s\}_{[g]} &= \{\mathbf{X}_s\}_{[g]} - \sum_{h\in s} \mathbf{N}_{(h,s)}[g]\cdot\{\mathbf{U}\mathbf{x}_{(h,s)}\}_{[g]} \\ \{\overline{\mathbf{X}}_{(h,s)}\}_{[g]} &= \{\mathbf{X}_{(h,s)}\}_{[g]} - \{\mathbf{m}+\mathbf{D}\mathbf{y}_s\}_{[g]}\cdot\sum_{h\in s}\mathbf{N}_{(h,s)}[g] \end{aligned} \qquad (4)$$

$\overline{\mathbf{X}}_s$ is used to estimate the speaker vector (session effects are removed) while $\overline{\mathbf{X}}_{(h,s)}$ is used to estimate channel factors (speaker effects are removed)

Let $\mathbf{L}_{(h,s)}$ be $R\times R$ matrix, and $\mathbf{B}_{(h,s)}$ a vector of dimension R, both defined as:

$$\begin{aligned} \mathbf{L}_{(h,s)} &= \mathbf{I} + \sum_{g\in\text{UBM}} \mathbf{N}_{(h,s)}[g]\cdot\{\mathbf{U}\}^t_{[g]}\cdot\mathbf{\Sigma}^{-1}_{[g]}\cdot\{\mathbf{U}\}_{[g]} \\ \mathbf{B}_{(h,s)} &= \sum_{g\in\text{UBM}} \{\mathbf{U}\}^t_{[g]}\cdot\mathbf{\Sigma}^{-1}_g\cdot\{\overline{\mathbf{X}}_{(h,s)}\}_{[g]}, \end{aligned} \qquad (5)$$

where $\mathbf{\Sigma}_g$ is the covariance matrix of the g^{th} UBM component. By using $\mathbf{L}_{(h,s)}$ and $\mathbf{B}_{(h,s)}$, $\mathbf{x}_{(h,s)}$ and $\mathbf{y}_s$ can be obtained by using the following equations:

$$\begin{aligned} \mathbf{x}_{(h,s)} &= \mathbf{L}^{-1}_{(h,s)}\cdot\mathbf{B}_{(h,s)} \\ \{\mathbf{y}_s\}_{[g]} &= \frac{\tau}{(\tau+\mathbf{N}_s[g])}\cdot\mathbf{D}_g\cdot\mathbf{\Sigma}^{-1}_g\cdot\ \{\overline{\mathbf{X}}_s\}_{[g]}, \end{aligned} \qquad (6)$$

where $\mathbf{D}_g = \frac{\mathbf{\Sigma}^{1/2}_g}{\sqrt{\tau}}$, τ is the MAP relevance factor (14.0 in our experiments).

3.3. Intersession matrix estimation

The matrix $\mathbf{U}$ can be estimated line by line, with $\{\mathbf{U}\}^i_{[g]}$ being the i^{th} line of $\{\mathbf{U}\}_{[g]}$ then:

$$\mathbf{U}^i_{[g]} = \mathbf{LU}^{-1}_g\cdot\mathbf{RU}^i_g, \qquad (7)$$

Algorithm 1: Estimation algorithm of **U**

```
For each speaker $s$ and session $h$ : $\mathbf{y}_s \leftarrow 0$ , $\mathbf{x}_{(h,s)} \leftarrow 0$
$\mathbf{U} \leftarrow random$ (U is initialized randomly);
Estimate statistics: $\mathbf{N}_h$, $\mathbf{N}_{(h,s)}$, $\mathbf{X}_s$, $\mathbf{X}_{(h,s)}$ (eq.2 and 3);
for $i = 1$ to $nb\_iterations$ do
    for all $s$ and $h$ do
        Center statistics: $\overline{\mathbf{X}}_s$, $\overline{\mathbf{X}}_{(h,s)}$ (eq.4);
        Estimate $\mathbf{L}^{-1}_{(h,s)}$ and $\mathbf{B}_{(h,s)}$ (eq.5);
        Estimate $\mathbf{x}_{(h,s)}$ and $\mathbf{y}_s$ (eq.6);
    end
    Estimate matrix $\mathbf{U}$ (eq. 7 and 8) ;
end
```

Algorithm 2: Estimation of latent variables on a single utterance

```
Result: Estimate $\mathbf{m}_{(h_0,s_0)}$, $\mathbf{y}_{s_0}$, $\mathbf{x}_{(h_0,s_0)}$
$\mathbf{y}_{s_0} \leftarrow 0$ , $\mathbf{x}_{(h_0,s_0)} \leftarrow 0$ ;
Estimate statistics: $\mathbf{N}_{s_0}$, $\mathbf{N}_{(h_0,s_0)}$, $\mathbf{X}_{s_0}$, $\mathbf{X}_{(h_0,s_0)}$ (eq.2 and 3);
for $i = 1$ to $nb\_iterations$ do
    Center statistics: $\overline{\mathbf{X}}_{s_0}$, $\overline{\mathbf{X}}_{(h_0,s_0)}$ (eq.4);
    Estimate $\mathbf{L}^{-1}_{(h_0,s_0)}$ and $\mathbf{B}_{(h_0,s_0)}$ (eq.5);
    Estimate $\mathbf{x}_{(h_0,s_0)}$ and $\mathbf{y}_{s_0}$ (eq.6);
end
```

where $\mathbf{RU}^i_g$ and $\mathbf{LU}_g$ are given by:

$$\begin{aligned} \mathbf{LU}_g &= \sum_s\sum_{h\in s}(\mathbf{L}^{-1}_{(h,s)} + \mathbf{x}_{(h,s)}\mathbf{x}^T_{(h,s)})\cdot\mathbf{N}_{(h,s)}[g] \\ \mathbf{RU}^i_g &= \sum_s\sum_{h\in s}\{\overline{\mathbf{X}}_{(h,s)}\}_{[g]}[i]\cdot\mathbf{x}_{(h,s)} \end{aligned} \qquad (8)$$

The algorithm 1 presents the adopted strategy to estimate the session variability matrix with the above developments (the standard likelihood function can be used to asses the convergence), while the algorithm 2 refers to the case of the estimation of $\mathbf{x}$ and $\mathbf{y}$ for a single utterance (a single iteration is performed in practice). The majority of the computational time is spent in Cholesky decompositions of $R\times R$ matrices in eq. 6 and 7. This process can however be parallelized.

4. Hybrid domain session compensation for the verification task

The following paragraph details the strategy employed to perform the session variability compensation. The verification task is defined as follows. A speaker $\mathbf{s}_{tar}$ is enrolled by the system with his training data $Y_{\mathbf{s}_{tar}}$. Given a sequence of speech frames $\mathcal{Y} = \{y_1 \dots y_T\}$ and the speaker $\mathbf{s}_{tar}$, the speaker verification task consists in determining if $\mathcal{Y}$ was spoken by $\mathbf{s}_{tar}$ or not. Using the factor analysis decomposition in both training and testing data, one can write:

$$\begin{aligned} \mathbf{m}_{(\mathbf{h}_{tar},\mathbf{s}_{tar})} &= \mathbf{m}+\mathbf{D}\mathbf{y}_{\mathbf{s}_{tar}}+\mathbf{U}\mathbf{x}_{\mathbf{h}_{tar}}, \\ \mathbf{m}_{(\mathbf{h}_{test},\mathbf{s}_{test})} &= \mathbf{m}+\mathbf{D}\mathbf{y}_{\mathbf{s}_{test}}+\mathbf{U}\mathbf{x}_{\mathbf{h}_{test}}. \end{aligned} \qquad (9)$$

where the speaker $\mathbf{s}_{tar}$ in the training data and $\mathbf{s}_{test}$ in the testing data have been distinguished. To deal with the session variability, the strategy adopted by [4, 3] assumes that the test speaker have the same identity as the target speaker, *i.e.* $\mathbf{y}_{\mathbf{s}_{test}} = \mathbf{y}_{\mathbf{s}_{tar}}$. The channel component $\mathbf{U}\mathbf{x}_{\mathbf{h}_{test}}$ of the test

[1] All posterior probabilities are computed on the UBM model

segment is estimated under this assumption. Indeed, the session component in the target model $\mathbf{Ux}_{\mathbf{h}_{tar}}$ is replaced by the one estimated in the test data $\mathbf{Ux}_{\mathbf{h}_{test}}$. The world model in the score equation remains unchanged. This strategy has several drawbacks: the target speaker model is changing for each test and performance are good when using score normalization techniques.

In this work, a hybrid domain normalization strategy is proposed, aiming to withdrawn the session component in the test and training data, that is:

$$\mathbf{m}_{\mathbf{s}_{tar}} = \mathbf{m} + \mathbf{D}\mathbf{y}_{\mathbf{s}_{tar}};\ \mathbf{m}_{\mathbf{s}_{test}} = \mathbf{m} + \mathbf{D}\mathbf{y}_{\mathbf{s}_{test}}. \qquad (10)$$

In this strategy, speakers are assumed to be different and are treated separately. The following paragraphs explain how to achieve this strategy in a system, when using a LLR-based verification approach or a kernel distance-based verification approach.

4.1. LLR based scoring

The speaker verification score is an expected log-likelihood ratio:

$$LLK(\mathcal{Y}|\mathbf{m}_{(\mathbf{h}_{tar},\mathbf{s}_{tar})})) - LLK(\mathcal{Y}|\mathbf{m}), \qquad (11)$$

where $LLK(\cdot|\cdot)$ indicate the average of the log-likelihood function over all frames. Here, GMMs have the same covariance matrices as well as the same mixture weights (both dropped from the equation for clarity). Two session compensation approaches can be adopted: the first one is to perform compensation at the frame level, where session compensation can be seen as a front-end, the second one that we propose is an hybrid compensation, that is the session variability is subtracted from the target speaker model (model domain) and the compensation in the testing data is performed at the frame level (feature domain). The following formula is used to remove the session effect for each frame t (also successfully used by [5]):

$$\hat{t} = t - \sum_{g=1}^{M} \gamma_g(t) \cdot \{\mathbf{U} \cdot \mathbf{x}_{\mathbf{h}_{test}}\}_{[g]}. \qquad (12)$$

4.2. Kernel based scoring and SVM modeling

By using equation 10, the factor analysis model estimates supervectors containing only speaker information, normalized with respect to the session variability. In [6], the authors proposed a probabilistic distance kernel that computes a distance between GMMs, well suited for a SVM classifier. Let $\mathcal{X}_\mathbf{s}$ and $\mathcal{X}_{\mathbf{s}'}$ be two sequences of speech data corresponding to speakers **s** and **s**$'$, the kernel forumlation is given below.

$$K(\mathcal{X}_\mathbf{s}, \mathcal{X}_{\mathbf{s}'}) = \sum_{g=1}^{M} \left(\sqrt{\alpha_g}\mathbf{\Sigma}_g^{-\frac{1}{2}}\mathbf{m}_s^g\right)^t \left(\sqrt{\alpha_g}\mathbf{\Sigma}_g^{-\frac{1}{2}}\mathbf{m}_{s'}^g\right). \quad (13)$$

This kernel is valid when only means of GMM models are varying (weights and covariance are taken from the world model). $\mathbf{m}_\mathbf{s}$ is taken here from the model in eq. 10, *i.e.* $\mathbf{m}_s = \mathbf{m} + \mathbf{D}\mathbf{y}_s$.

5. Protocol

All experiments were performed using the ALIZE and LIA_SpkDet toolkit[2][7].

5.1. Database, protocols, toolkits

Speaker verification experiments, presented in section 6, are performed based upon the NIST SRE 2005 as a development set, and 2006 database for the validation set, male speakers only (referred to as 2005 and 2006 protocol). The 2005 protocol consists of 274 speakers, 9012 tests (951 target tests, the rest is impostor trials) while the 2006 protocol consists of 354 speakers, 9720 tests (741 target tests, the rest is impostor trials)[3]. Results are given in terms of equal-error-rate (EER) and the minimum of DCF (an *a posteriori* decision). Train and test utterances contain 2.5 minutes of speech in average (telephone conversation, where around 30% of speech frames per speaker have been retained). The intersession variability matrix is enrolled on the NIST-SRE-2004 database with 2938 examples with 124 speakers (around 20 iterations to reach convergence). From the same database, 200 impostors speakers are used for score normalization and negative examples in the SVM classifier.

Table 1: *Results of the baseline GMM-UBM system on the 2005 and 2006 protocol. DCFmin (x100), EER(%).*

	SRE-05		SRE-06	
	DCFmin	EER	DCFmin	EER
Nonorm	3.83	7.15	3.88	6.79
Tnorm	3.05	8.52	2.9	5.7

5.2. SVM training

The LIA_SpkDet toolkit now benefits from the LIBSVM [8] library to induce SVM and to classify instances. SVM models are trained with an infinite (very large in practice) C parameter thus avoiding classification error on the training data (hard margin behaviour). The negative labelled examples are speakers from the normalisation cohort.

5.3. Baseline GMM-UBM system

The baseline system is a standard GMM-UBM system [9]. The background model in the experiments is the same as the background model in the LIA submission in the NIST-SRE-2006 campaign (male set only)[4]. Training is performed based upon Fisher database[5], and consists of about 10 millions of speech frames. Speaker models are derived by Bayesian adaptation on the Gaussian component means, with a relevance factor of 14.

Frames are composed of 19 LFCC parameters, its derivatives, and 11 second derivatives (the frequency window is restricted to 300-3400 Hz). A normalization process is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance. The background model has 512 components whose variance parameters are floored to 50% of the global variance (0.5). It is worth noting that feature warping pre-processing was not necessary to obtain good performance. Table 1 shows the results of the baseline system and T-norm scores (Z and ZT-norm do not bring any improvement).

6. Experimental results

In the following is detailed the experimental results obtained with the implementation given in this paper.

Table 2 investigates the effect of the session variability subspace rank, without any score normalization. The best performing system have been found to have a rank of 40. It is

[2] an open-source software available at http://www.lia.univ-avignon.fr/heberges/ALIZE/

[3] 2005 protocol corresponds to the core condition, labeled as *det7* and the 2006 protocol corresponds to the core condition, labeled as *det3*

[4] NIST 2006, SRE evaluation plan, www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf

[5] Fisher English Training Speech Part 1, LDC n:LDC2004S13

Table 2: *Varying the rank subspace. 2005 Protocol. DCFmin(x100), EER(%). The best rank in terms of DCFmin is 40.*

	Subspace rank					
	0	20	40	60	80	100
DCFmin (x100)	3.83	2.05	**1.83**	1.93	1.95	1.99
EER (%)	7.15	5.1	4.42	**4.22**	4.31	4.23

Table 3: *Score normalization techniques on the factor analysis model (rank=40). Znorm always brings an improvement over the baseline. DCFmin(x100), EER(%)*

	SRE-05		SRE-06	
	DCFmin	EER	DCFmin	EER
Nonorm	1.83	4.42	1.61	2.97
Tnorm	1.84	4.72	1.29	2.83
ZTnorm	1.72	4.62	**1.18**	**2.15**
Znorm	**1.64**	**4.21**	1.46	2.33

worth noting that the proposed methodology do not need a T or Znorm to give good results compared to the classical compensation scheme. Indeed, a absolute improvement of 2 points in DCFmin is observed, as well as 2.7% at the EER.

Table 3 shows the improvement with score normalization. Znorm is the technique that always bring an improvement in both protocols, while Tnorm is only effective in 2006. Indeed, the DCFmin drops from 1.83 to 1.64 with Znorm in 2005 and from 1.61 to 1.18 with ZTnorm in 2006. While the behavior is different in both years, ZTnorm seems to be the most confident choice for score normalization.

Table 4 consists in avoiding session compensation in the feature domain by using speaker dependent supervectors only through the use of the kernel described in eq. 13 along a SVM classifier. This system obtains similar performances as the one presented above in 2005 but outperforms the classical system in 2006. It is worth noting that here Znorm does not bring any gain, and that in this case Tnorm brings most of the improvement. The improvement over the classical LLR system is of 1.64 to 1.58 in 2005, while from 1.18 to 1.03 in 2006. Comparing to results on the NIST-SRE-2006 evaluation, this system would be one of the most performing. Moreover, in terms of stability of the decision score, if the threshold is tuned on the 2005 protocol, the DCF on the 2006 protocol would be 1.23.

7. Conclusion

A factor analysis model for speaker verification have been successfully designed and proposed by [2] (section 2). This technique is complex to implement and this paper try to answer this issue by proposing a straightforward implementation without referral to other papers or complex mathematics, described in section 3. Moreover, this is part of the ALIZE/LIA_SpkDet toolkit that is freely available to the community. In section 4, an hybrid compensation scheme (both feature and model domain) is proposed. There is several advantages in favor of this strategy. First, the implementation is simpler, as the target speaker model does not change over the verification experiment and the standard likelihood computation can be employed. Second, while the classical compensation scheme brings a bias in scores (score normalization is needed to obtain good performance), this approach presents good results with native scores. Finally, the use of a SVM classifier with a proper supervector-based kernel is straightforward. For a single system, results presented with this latter on the NIST-SRE-2006 database (male speakers) would be one of the most performing in the evaluation. Indeed, in case of the kernel-based scoring, the absolute improvement over a standard GMM-UBM system is of around 2 points at the DCFmin and 6% at the EER (the same behavior can be observed on the 2005 protocol). As channel factors are independent of the Gaussian component, one perspective would be to extend the model to multiple channel factors, dependant on clusters of Gaussian.

Table 4: *GMM factor analysis supervectors with distance kernel (rank=40). 2005 and 2006 protocol. DCFmin(x100), EER(%). Tnorm brings the most significant gain.*

	SRE-05		SRE-06	
	DCFmin	EER	DCFmin	EER
Baseline (40)	1.97	4.83	1.40	2.83
Znorm	1.92	5.36	1.54	2.70
Tnorm	1.61	**4.42**	**1.03**	2.29
ZTnorm	**1.58**	4.51	1.06	**2.16**

8. References

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

[2] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice Modeling With Sparse Training Data,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, p. 345, 2005.

[3] R. Vogt, B. Baker, and S. Sridharan, “Modelling Session Variability in Text-Independent Speaker Verification,” in *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2005), Lisboa, Portugal*, 2005.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor Analysis Simplified,” in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, vol. 1, 2005.

[5] C. Vair, D. Colibro, and P. Laface, “Channel factors compensation in model and feature domain for speaker recognition,” in *Odyssey’06, the Speaker Recognition Workshop, San Juan, Puerto Rico*, Jun 2006.

[6] W. Campbell, D. Sturim, and D. Reynolds, “Support Vector Machines Using GMM Supervectors for Speaker Verification,” *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[7] J.-F. Bonastre, F. Wils, and S. Meignier, “Alize, a free toolkit for speaker recognition,” in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, Philadelphia, USA, March 2005.

[8] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, vol. 10, no. 1-3, pp. 19–41, 2000.