4 Experimental Methodology

In this chapter, we describe the methodology used to perform text-independent speaker verification experiments in this thesis. Three databases, Banca, Polyvar and NIST, are used in the following to compare systems. Two baseline models are considered: a GMM based system described in Section 2.2 and summarized in Figure 2.1 and an SVM with GLDS kernel based system described in Section 2.5 and summarized in Figure 2.3.

The outline of this chapter goes as follows. In Section 4.1, we describe the general methodology to use the databases. In Section 4.2, the databases are described and the baseline results are given for each of them.

4.1 Methodology

For both GMM and SVM based systems, the feature extraction, described in Section 2.3, is computed using the same procedure, as follows. The original waveforms are sampled every 10ms with a window size of 20ms. For all databases, each sentence is parameterized using 24 triangular band-pass filters with a DCT transformation, computed using (2.20), of order 16, complemented by their first derivative (delta), the log-energy and the delta-log-energy, for a total of 34 coefficients. A simple silence detector based on an unsupervised bi-Gaussian model is used to remove all silence frames. A bi-Gaussian model is learned using the ML criterion except for the NIST database. Since this database is noisy, the bi-Gaussian model is first learned on a random recording with land line microphone and adapted for each new sentence using the MAP algorithm with a MAP adaptation factor of $\lambda = 0.5$ in (2.6). All frames were normalized in order to have zero mean and unit variance. The NIST database being a telephone based database, the signal is thus band-pass filtered between 300 and 3400 Hz.

While the log energy is important in order to remove the silence frames, it is known not to be appropriate for the task of discrimination between clients and impostors. This feature was thus removed after removing the silences, but its first derivative was kept. Hence, the models are trained with 33 (34-1) features.

In order to select the various hyper-parameters (such as the number of Gaussians, the MAP adaptation factor, etc.), two different client populations are used: one for the development and one for the test set. We use the development set as follows; for each value of the hyper-parameter to tune, we train the client models using the training data available for each client. We then select the value of the hyper-parameter that optimizes the EER on the clients and impostors trials of the development set. Finally, we train the models on the test set using these hyper-parameters and measure the performance of the system.

All databases contain some accesses to enroll the world model. These accesses are also used as negative examples for discriminant models. The T-normalization models are the client models of the development set. When T-normalization is performed on the development set a leave-one-out cross-validation procedure (Devroye and Gyorfi, 1997) is applied in order not to bias the results: the model corresponding to the claimed identity is removed from the T-normalization model list.

4.2 Databases

In order to compare the systems presented here, three databases are used: Polyvar, Banca, NIST. All of the three databases have their own specificity that justifies their use.

Banca

The English part of the *Banca* database (Bailly-Baillière et al., 2003) contains a development and a test set of 26 clients each (13 men and 13 women) as well as another population of 60 speakers (30 females and 30 males) used to train the world model. The world model is the concatenation of two gender dependent world models. This database contains three recording conditions defined as

L. Devroye and L. Gyorfi. *A Probabilistic Theory of Pattern Recognition*. Springer, 20 February 1997.

E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 625–638. Springer-Verlag, 2003.

controlled (acquired in an office with only one person), degraded (acquired in several offices of several people) and adverse (recorded in a public area) and is provided with 7 different protocols. We have chosen to use protocol P, which we consider the most realistic: only one controlled session is available to train the client model and 546 balanced test accesses in controlled, degraded and adverse conditions were used per population. Even if this database is small, it is still interesting because of the several recording conditions, and because the impostors pronounce the same sentence as the client.

All hyper-parameters of the GMM based baseline system are tuned: the number of ML iterations to train the world model, the number of iterations for the MAP adaptation, the number of Gaussians, the variance flooring factor and the MAP adaptation factor. All were selected on the development set to minimize the EER and are given in Table 4.1. The hyper-parameters for the SVM GLDS kernel are given in Table 4.2. When we vary C, from a certain value up to ∞ we keep a maximum of support vectors (this corresponds to the optimal solution found on the development set for all databases). We will use in the following the notation $\rightarrow \infty$ to express this.

Table 4.1. Summary of the hyper-parameters for GMM based systems on the Banca database

# of ML Iterations	# of MAP Iterations	# of Gaussians	MAP Factor: λ in (2.6)	Variance Flooring in [%]
25	5	400	0.5	60

Table 4.2. Summary of the hyper-parameters for the SVM based system on the Banca database ($\rightarrow$ means “tends to”).

Degree of the GLDS kernel	C in (2.9)
3	$\rightarrow \infty$

The SVM system is based on a GLDS kernel of degree 3, as originally proposed by Campbell et al. (2005). T-normalization was not performed because all recordings were done using only one microphone.

W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 2005.

Table 4.3. Results on the Banca database: GMMs and SVMs

	GMMs	SVMs
HTER [%]	1.39	6.94
95% Confidence	±1.03	±2.15

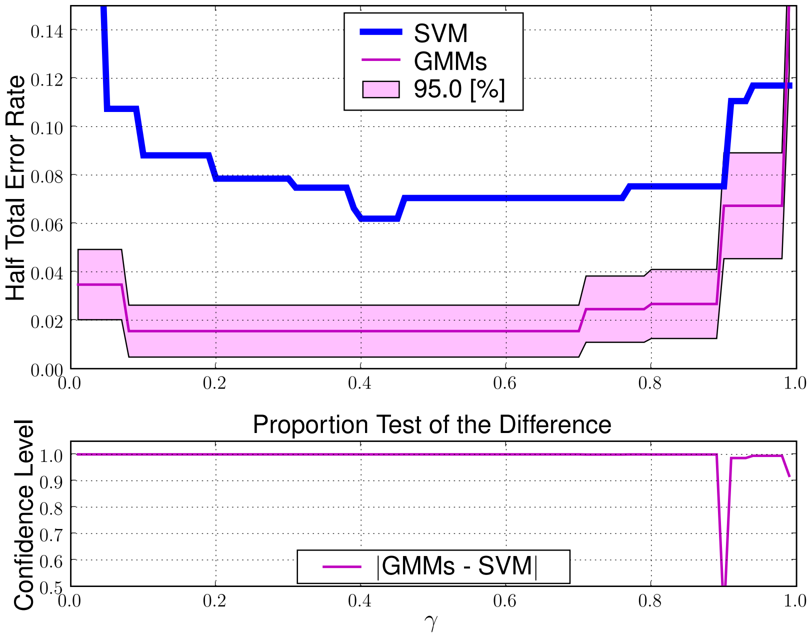

Figure 4.1. EPC curves on the test set of the Banca database: GMMs and SVMs.

Figure 4.1 and Table 4.3 show the results on the Banca database. We can see that the GMM based system outperforms significantly the SVM based system.

Polyvar

The Polyvar telephone database (Chollet et al., 1996), contains a development and a test set of 19 clients (12 men and 7 women) each, as well as another population of 56 speakers (28 men and 28 women) used to train the world model. The world model is the concatenation of two gender dependent world models. For each client, a training set contains 5 repetitions of 17 words (composed of 3 to 12 phonemes each), while a separate test set contains on average 18 repetitions of the same 17 words, for a total of 6000 utterances, as well as on average 12000 impostor utterances. Each client has 17 models, one for each word, and only 5 sequences are available to train each model. As in the original protocol, only same word accesses are kept.

The hyper-parameters of GMM based systems where tuned using the same method as for the Banca database, minimizing the EER over the development set and Table 4.4 gives a summary of the obtained hyper-parameters.

Table 4.4. Summary of the hyper-parameters for GMM based systems on the Polyvar database

# of ML Iterations	# of MAP Iterations	# of Gaussians	MAP Factor: λ in (2.6)	Variance Flooring in [%]
25	5	200	0.2	10

The SVM system is, once again, based on a GLDS kernel of degree 3 originally proposed by Campbell et al. (2005). T-normalization was not performed because all recordings were done with the same kind of telephone (land line ISDN). The hyper-parameters of the SVM based system are the same as for the Banca database and are given in Table 4.2.

Figure 4.2 and Table 4.5 show that SVMs and GMMs should be considered as equivalent for most values of $\gamma < 0.7$ while the SVM based system outperforms the GMM based system for most values of $\gamma > 0.7$.

G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais. Swiss french polyphone and polyvar: telephone speech databases to model inter- and intra-speaker variability. IDIAP-RR 01, IDIAP, 1996.

W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 2005.

Table 4.5. Results on the Polyvar database: GMMs vs SVMs.

	GMMs	SVMs
HTER [%]	4.77	4.49
95% Confidence	±0.33	±0.32

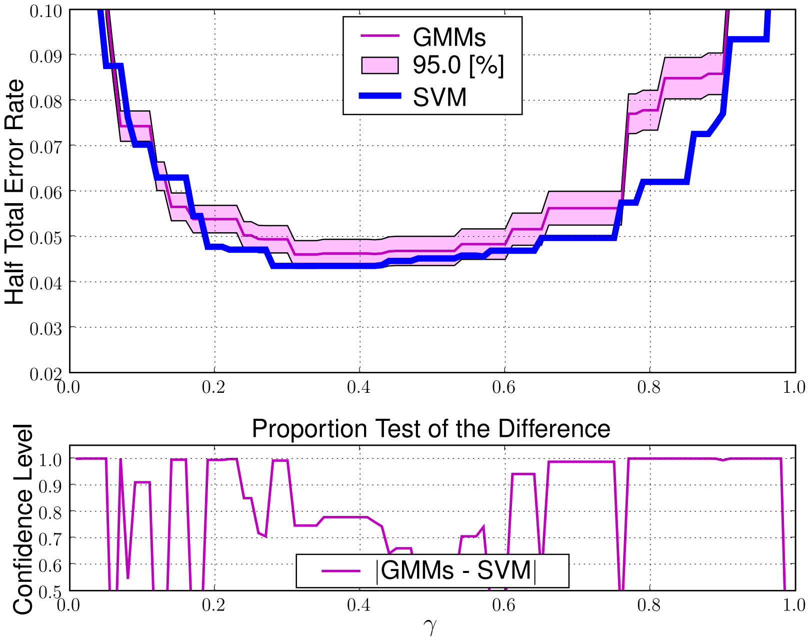

Figure 4.2. EPC curves on the test set of the Polyvar database: GMMs vs SVMs.

NIST

The NIST database is a subset of the database that was used for the *NIST 2002 and 2003 Speaker Recognition Evaluation*, which comes from the second release of the cellular switchboard corpus, Switchboard Cellular - Part 2, of the Linguistic Data Consortium. This data was used as test set while the world model data and the development data comes from previous NIST campaigns. For both development and test clients, there were about 2 minutes of telephone speech used to train the models and each test access was less than 1 minute long. Only female data are used and thus only a female world model is used. The development population consisted of 100 females, while the test set is composed of 191 females. 655 different records are used to compute the world model or as negative examples for the discriminant models. The total number of accesses in the development population is 3931 and 17578 for the test set population with a proportion of 10% of true target accesses. Only test accesses between 15 and 45 seconds are considered as the primary condition in the NIST campaign (see `http://www.nist.gov/speech/tests/spk/2003` for the evaluation plan).

Table 4.6 gives a summary of the hyper-parameters used for GMM based experiments after selection based on minimizing EER on the development set. T-normalization is performed using (5.43) for the GMM based system. Figure 4.3 shows the improvement obtained by the T-normalization and justifies the use of score normalization for the GMM based system on NIST database. No score normalization procedure is applied for SVMs GLDS based kernel due to the computational cost and the small expected improvement as explained later in Chapter 5.

Table 4.6. Summary of the hyper-parameters for GMMs based systems on the NIST database

# of ML Iterations	# of MAP Iterations	# of Gaussians	MAP Factor: λ in (2.6)	Variance Flooring in [%]
25	5	100	0.5	60

The hyper-parameters of SVMs based system are given in Table 4.2 and are once again the same as the two precedent databases.

Table 4.7. EPC curves on the test set of the NIST database: SVM v.s. GMM + T-norm

	GMMs + T-norm	SVM
HTER [%]	8.68	11.06
95% Confidence	±0.84	±1.05

Figure 4.4 and Table 4.7 show that the SVM based system outperforms the GMM based system for small values of γ and that the GMM based system outperforms SVM based system for the other values of γ.

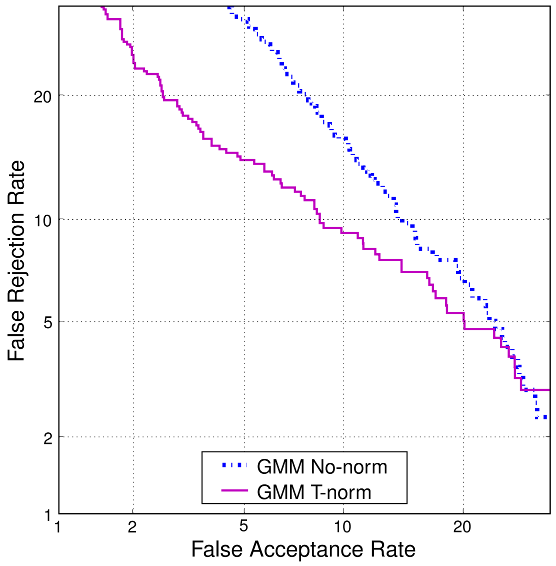

Figure 4.3. DET curves on the development set of the NIST database using, or not, the T-normalization procedure.

4.3 Conclusion

Except for the Banca database, both the SVM and GMM based systems are more or less equivalent. The SVM based system is easy to tune because the only hyper-parameters are the degree of the polynomial expansion and C in (2.9). In all cases the optimal value for degree was 3. We have also noted that the C value should be large. That means that SVMs maximize the margin without accepting examples in the margin. This can be explained by the fact that only few positive training examples are available and the cost function is not optimal for highly unbalanced class problem. In order to make use of C, the cost function should probably be modified. Even if it seems comfortable to have no hyper-parameter to tune, it also means there is no way to adjust the capacity of the SVM models, which can be important to expect improvements of the SVM performance.

The original Banca database and its protocol descriptions was published in:

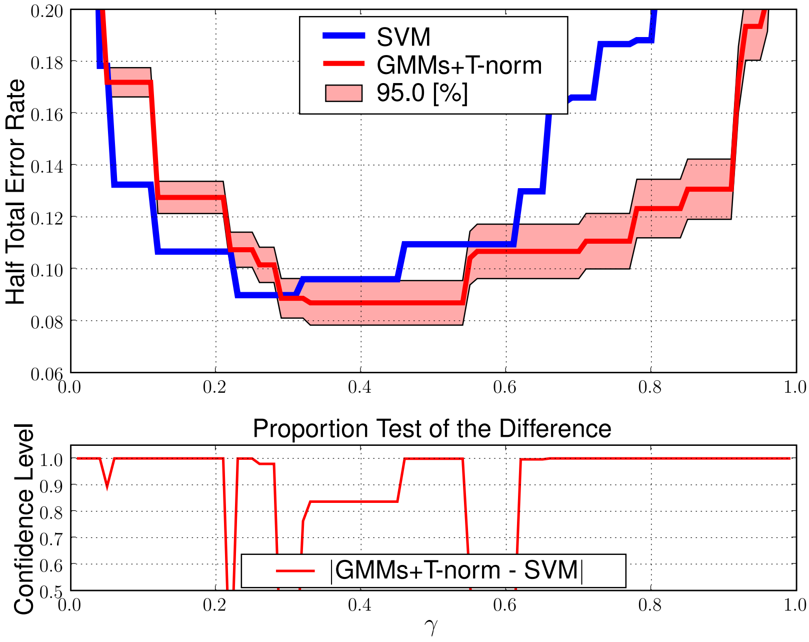

Figure 4.4. Results on the test set of the NIST database: GMMs + T-norm vs SVMs.

CONTRIB E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 625–638. Springer-Verlag, 2003

The Polyvar database and its protocol descriptions was published in:

CONTRIB F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacob, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Mariéthoz, C. Mokbel, and H. Mokbel. An overview of the picasso project research activities in speaker verification for telephone applications. In *6th European Conference on Speech Communication and Technology — Eurospeech'99*, volume 5, pages 1963–1966, Budapest, Hungary, september 1999

The SVM based system never outperformed significantly the GMM based system.

- Does that mean that non-discriminant models are the best solution for speaker verification?
- Are GMM based systems really non-discriminant?
- Is the statistical framework applicable to SVMs?
- Is T-normalization also applicable to the SVMs based approaches?

In the next chapter we address these questions in order to have a good starting point to develop new discriminant approaches.