Support Vector Machine Regression for Robust Speaker Verification in Mismatching and Forensic Conditions

Ismael Mateos-Garcia, Daniel Ramos, Ignacio Lopez-Moreno, and Joaquin Gonzalez-Rodriguez

ATVS – Biometric Recognition Group,
Escuela Politecnica Superior, Universidad Autonoma de Madrid,
C. Francisco Tomás y Valiente 11, 28049 Madrid, Spain
{ismael.mateos,daniel.ramos,ignacio.lopez,
joaquin.gonzalez}@uam.es

Abstract. In this paper we propose the use of Support Vector Machine Regression (SVR) for robust speaker verification in two scenarios: *i)* strong mismatch in speech conditions and *ii)* forensic environment. The proposed approach seeks robustness to situations where a proper background database is reduced or not present, a situation typical in forensic cases which has been called *database mismatch*. For the mismatching condition scenario, we use the NIST SRE 2008 core task as a highly variable environment, but with a mostly representative background set coming from past NIST evaluations. For the forensic scenario, we use the Ahumada III database, a public corpus in Spanish coming from real authored forensic cases collected by Spanish Guardia Civil. We show experiments illustrating the robustness of a SVR scheme using a GLDS kernel under strong session variability, even when no session variability is applied, and especially in the forensic scenario, under database mismatch.

Keywords: Speaker verification, forensic, GLDS, SVM classification, SVM regression, session variability compensation, robustness.

1 Introduction

Speaker verification is currently a mature technology which aims at determine whether a given speech segment of unknown source belongs to the identity of a claimed individual or not. Among the most important challenges of a speaker verification system is the robustness to the mismatch in conditions between training and testing utterances, being its compensation a main factor for the improvement of system performance. Recently, this task has been carried out by the use of data-driven session variability compensation techniques based on factor analysis, which have become the state of the art in these technologies as can be seen in the periodic NIST Speaker Recognition Evaluations (SRE) [1]. Such techniques can be applied to the best-performing systems working at the spectral level, mainly based on Gaussian Mixture Models (GMM) [2] and Support Vector Machines (SVM) [3], increasing their robustness and accuracy. Among all the different compensation variants, the Nuisance

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 484–493, 2009.

Attribute Projection (NAP) [4] has been used for SVM modelling techniques, presenting the advantages of simplicity and efficiency with respect to other more sophisticated approaches [5]. In particular, NAP has demonstrated its usefulness in systems based on SVM Classification (SVC) using Generalized Linear Discriminant Sequence (GLDS) kernel [3]. Although SVC-GLDS performance is slightly worse than other modelling approaches such as GMM or GMM-SVM [6], it constitutes an additional source of information about speaker identity, and can be combined with other systems by means of fusion [7].

Despite of their unquestionable success, factor analysis techniques still present important challenges to face. The use of such compensation techniques is strongly conditioned to the availability of databases for training the algorithms involved. In real applications the availability of development data in desirable conditions is unfortunately unfrequent. In many situations the technology developers tune their systems with databases coming from a different environment from the conditions of the operational data. This is very typical in forensics, where in each case the conditions of the recordings to analyze are extremely variable in terms of acoustic environment, channel, speaking style, emotional state, language, etc. It is almost impossible to think in the availability of a background database for all the combination of conditions in a possible case. This mismatch in the conditions between background data for system tuning and operational data has been coined *database mismatch* in a recent work [8], and constitutes an important challenge to face in the current state of the art.

In this paper we propose the use of Support Vector Machine Regression (SVR) using a GLDS kernel for robust speaker verification under strong mismatch and forensic conditions. In order to show the adequacy of our approach, we use two different speech databases: *i)* NIST SRE 2008, presenting strong mismatching conditions; and *ii)* Ahumada III, a public database in Spanish coming from authored real forensic cases and collected by Spanish Guardia Civil, which also presents different conditions than NIST databases typically used for background modelling and session variability compensation. This paper is organized as follows. First, the new approach SVM regression is introduced in Section 2. Section 3 presents the proposed SVR-GLDS system for speaker verification. In Section 4, experiments are presented in the two proposed scenarios. Results show the adequacy of SVR-GLDS for robust speaker verification, even when no session variability compensation is performed. Finally, conclusions are drawn in Section 5.

2 Support Vector Machine (SVM) Regression

SVR approach for GLDS speaker verification has been recently proposed by the authors in [9]. In the SVR case the goal is more general than in the widely extended SVC approach. Regression aims at learning a n-dimensional function from the data and classification aims at obtaining a classification boundary. In regression, the vector labels, y_i, are seen as a function of x_i, $g_n(x_i) = y_i$. In a binary classification problem, such as speaker verification, $g_n(\cdot)$ is a discrete function with just two values: $g_n(x_{\text{target}}) = +1$ and $g_n(x_{\text{nontarget}}) = -1$. SVR will try to find the discrete function $f(\cdot) \simeq g_n(\cdot)$.

The main difference between SVC and SVR is the loss function. SVC penalizes the situation where $f(\cdot) < g_n(\cdot)$, but as SVR aims at estimating a function, it also penalizes $f(\cdot) > g_n(\cdot)$. The loss function should consider such effect, and there are different options in the literature. A popular choice is the ε-insensitive loss function [10], where vectors are penalized when $\left|f(\cdot) - g_n(\cdot)\right| > \varepsilon$. The objective hyperplane in the SVR case will then be:

$$w = \min\left(\frac{1}{2}w^T \cdot w + C\frac{1}{m}\sum \xi_{c,i} + \xi'_{c,i}\right) .$$
$$\text{subject to: } \begin{cases} 0 \le f(x_i) - y_i \le \xi_{c,i} + \varepsilon \\ 0 \le y_i - f(x_i) \le \xi'_{c,i} + \varepsilon \end{cases} \tag{1}$$

If we compare these criteria with SVC in Equation (2), we observe some differences. We have the SVC penalty variable, $\xi_{c,i}$, for those vectors for which $f(x_i) > g_n(x_i) + \varepsilon$, and a new variable $\xi'_{c,i}$ for those ones for which $f(x_i) < g_n(x_i) - \varepsilon$.

$$w = \min\left(\frac{1}{2}w^T \cdot w + C\frac{1}{m}\sum \xi_{c,i}\right) .$$
$$\text{subject to: } 0 \le \xi_{c,i} \le 1 - y_i f(x_i) \tag{2}$$

The loss functions, $f'_{loss}(x_i)$ (SVR) centered at $f(x_i) = g_n(x_i)$ and $f_{loss}(x_i)$ (SVC) at $f(x_i) = y_i$, are defined in (3) and shown in Fig. 1.

$$f'_{loss}(x_i) = \max\left\{0, \left|y_i \cdot f(x_i)\right| - \varepsilon\right\} .$$
$$f_{loss}(x_i) = \max\left\{0, 1 - y_i \cdot f(x_i)\right\} . \tag{3}$$

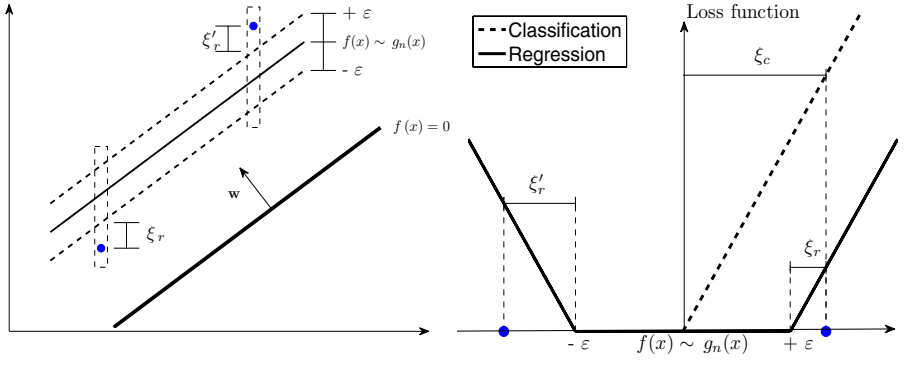


Fig. 1. SVR vs. SVC: boundaries and loss functions

3 SVR-GLDS for Speaker Verification

We propose to use SVR with a ε-insensitive loss function for the speaker verification task. Recently, the authors showed the performance of this novel approach over the core task of NIST SRE 2006 [9], a telephone scenario, obtaining good results in comparison with SVC.

One of the main advantages of using the SVR approach in the GLDS space relates to the use of support vectors for SVM training. On the one hand, SVC uses support vectors which are near the boundary between classes, where the vectors use to be scarce. Moreover, variability in the conditions of speech may significantly change the final hyperplane, introducing undesired variability and therefore performance degradation. On the other hand, SVR selects support vectors from areas where there is a higher concentration of vectors. Thus, the SVC hyperplane may be more sensitive than SVR to outliers, noisy vectors, etc. In this sense, SVR can present a more robust performance than SVC against outlier support vectors due to extreme conditions in some speech utterances.

Another advantage of the SVR approach relies on the use of the ε parameter. There are some works in the literature [10] that relate the ε parameter to the noise or variability of the function estimate. Following such assumptions, we proved in a previous work [9] that tuning ε allows us to adapt the SVR training process to the variability in the expanded feature space.

4 Experiments

4.1 SVM-GLDS Systems

Both ATVS SVC-GLDS and SVR-GLDS systems are based on a GLDS kernel as described in [3]. Feature extraction is performed based on audio files processed with Wiener filtering[1]. The front-end consists on the extraction of 19 MFCC plus deltas. As a first stage to avoid session variability compensation, CMN *(Cepstral Mean Normalization)*, RASTA filtering and feature warping are performed. A third degree polynomial expansion GLDS kernel is performed on the whole observation sequence, and a separating hyperplane is computed between the training speaker features and the background model. NAP is applied for session variability compensation according to [4]. Finally, the T-Norm score normalization technique is applied. We have used the LibSVM library[2] for training both SVM algorithms.

The background set for system tuning is a subset of databases from previous NIST SRE evaluations, including telephone and microphone channels. The T-Norm cohorts were extracted from the NIST SRE 2005 target models, 100 telephone models and 240 microphone models. NAP channel compensation was trained using recordings belonging to NIST SRE 2005 speakers which are present in both telephone and microphone data.

[1] A Wiener filtering implementation is available at Berkeley Webpage: http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/qio

[2] Software available at LibSVM webpage: http://www.csie.ntu.edu.tw/~cjlin/libsvm

4.2 Databases and Experimental Protocol

Experiments have been performed using two different databases. First, the NIST SRE 2008 [1] constitutes a highly mismatching environment. Second, Ahumada III represents real forensic casework speech in conditions different to those of the background data [8].

NIST SRE 2008 database and protocol represents a real challenge in terms on session variability. The training and test conditions for the core task include not only conversational telephone speech data but also conversational speech data recorded over microphone channels involving an interview scenario, and additionally, for the test condition, conversational telephone speech recorded over a microphone channel. The evaluation protocol defines the following training conditions: 10 seconds, 1 *(short2)*, 3 and 8 conversation sides and long conversation; and the following test condition: 10 seconds, 1 (*short3)* conversation side and long conversation. Each "short" conversation, either recorded over a telephone or a microphone, has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Interview segments contain about 3 minutes of conversational speech recorded by a microphone, most of the speech generally spoken by the target speaker. In our case the experiments followed the core task, namely *short2* training conditions, and *short3* test condition *(short2-short3)*.

Taking into account the test and train channel types, the evaluation protocol can be divided in 4 conditions: *tlf-tlf* (37050 trials)*, tlf-mic* (15771 trials)*, mic-mic* (34046 trials) and *mic-tlf* (11741 trials).

Ahumada III consists of authorized conversational speech acquired by the Acoustic and Image Processing Department of Spanish Guardia Civil from real forensic cases. The acquisition procedure uses two of the systems and procedures followed by Guardia Civil. As its present release, the recording procedure considered consists of digitalized analog magnetic recordings from GSM mobile calls, from those recordings of this type received in the last ten years, those authorized (case by case) by the corresponding judge after a trial and added to a database registered in the Spanish Ministerio del Interior, known as Base de Datos de Registros Acústicos (BDRA)[3]. In future releases of the database, speech will be included from digital wiretaps recorded directly from Spanish mobile telephone operators, the system known as SITEL (nationwide digital interception system).

Ahumada III Release 1 (Ah3R1)[4] consists of 61 speakers from a number of real cases with GSM BDRA calls across Spain, with a variety of country of origin of speakers, emotional and acoustic conditions, and dialects in the case of Spanish speech. There is no variability dimension is gender, as all of them are male speakers. All 61 speakers in Ah3R1 have two minutes of speech available from a single phone call to be used as unquestioned (control) recording, with the purpose of model enrollment or voice characterization. Additionally, ten speech segments for 31 speakers and five segments for speakers are included for testing issues, each one from a different call. Such fragments present between 7 and 25 seconds of speech, with an average

[3] With reference public scientific file number 1981420003 from Spanish Guardia Civil, Orden Ministerial INT/3764/2004 de 11 de noviembre.

[4] Ahumada III is publicly available for research purposes under license agreement and conditions (contact: http://atvs.ii.uam.es).

duration of 13 seconds. An evaluation protocol has been generated consisting in computing all possible scores from models trained with the enrollment utterances and test segments in the database (27084 trials).

4.3 Results

Strong Mismatching Conditions in NIST SRE 2008. The performance of SVC-GLDS over NIST SRE 2008 is first evaluated with two different configurations: *i)* without including any compensation technique, and *ii)* including a NAP compensation scheme. This shows the effect of compensating variability using NAP with a suitable background database. Table 1 shows the performance of the system detailed per condition. Results are presented both as EER *(Equal Error Rate)* and DCF_{min} as defined in NIST SRE [1]. It is observed that the performance of the system significantly improves when NAP is added, both for EER and DCF_{min} values. The improvement is bigger when strong channel mismatch occurs (*tlf-mic* or *mic-tlf* conditions).

Table 1. EER and DCF_{min} in NIST SRE 2008 short2-short3, for SVC-GLDS and SVR-GLDS with and without NAP session variability compensation

		tlf-tlf	tlf-mic	mic-mic	mic-tlf
SVC	EER	13.8	24.1	17.4	23.5
	DCF_{min}	0.054	0.075	0.075	0.078
SVC + NAP	EER	10.2	**13.9**	**13.0**	15.3
	DCF_{min}	0.047	**0.053**	**0.057**	**0.059**
SVR	EER	10.0	15.1	15.4	16.4
	DCF_{min}	0.045	0.055	0.065	0.064
SVR + NAP	EER	**9.6**	14.3	13.8	**15.0**
	DCF_{min}	**0.045**	0.053	0.060	0.062

In order to use the proposed SVR-GLDS system, tuning the ε parameter is firstly required, and the variation of its performance with respect to such parameter is presented in Table 2. As we saw in [9] the system performance significantly changes as a function of this parameter.

Table 2. EER and DCF_{min} in NIST SRE 2008 short2-short3, for different values of ε in SVR-GLDS without NAP session variability compensation

		$\varepsilon = 0.05$	0.1	0.2	0.4	0.8
tlf-tlf	EER	**9.9**	10.0	10.9	13.5	13.9
	DCF_{min}	0.046	**0.045**	0.047	0.052	0.054
tlf-mic	EER	16.9	**15.1**	16.6	23.8	24.0
	DCF_{min}	0.059	**0.055**	0.063	0.074	0.075
mic-mic	EER	15.7	**15.4**	15.9	16.8	17.4
	DCF_{min}	**0.064**	0.065	0.067	0.074	0.075
mic-tlf	EER	17.0	**16.4**	18.8	22.8	23.6
	DCF_{min}	**0.063**	0.064	0.066	0.078	0.078

In most cases $\varepsilon = 0.1$ significantly improves the system performance, which is very similar to the optimum value in cases where it is seen at $\varepsilon = 0.05$. The optimal value of the parameter is coherent with the experiments presented in [9] using telephone speech in NIST SRE 2006 database and protocol. Thus, without NAP compensation, system tuning of the ε parameter seems robust over different databases, and should be performed one time and not for each one of the four conditions.

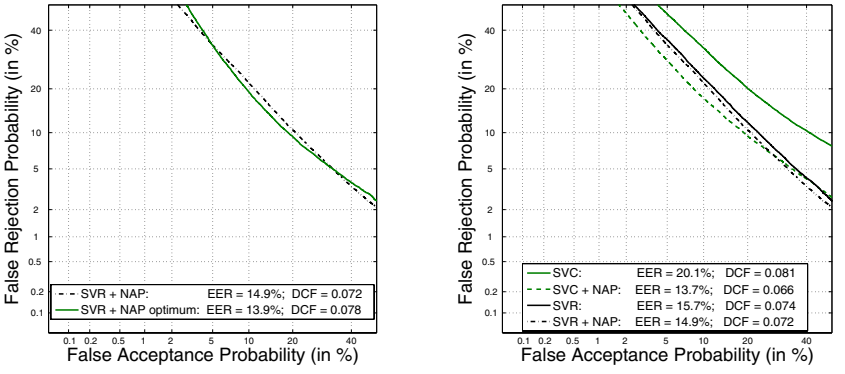
Next approach shows the performance of applying the same NAP compensation scheme to SVC-GLDS and SVR-GLDS systems. As the NAP transformation changes the properties of the expanded space a ε tuning is required before using the proposed system, the compensated parameters vectors will be significantly different to the previous ones. Table 3 shows the performance for different values of ε.

In this case the optimal value for the ε parameter varies depending on the condition. The optimal value observed for the non-compensated feature space was $\varepsilon = 0.1$, we will use this value in the rest of experiments. Fig. 2 *a)* presents a comparison between the performance of SVR-GLDS + NAP with $\varepsilon = 0.1$ and the optimal selection of ε for each one of four the conditions. The performance is similar.

Table 3. EER and DCF_{min} in NIST SRE 2008 short2-short3, for different values of ε in SVR-GLDS with NAP session variability compensation

		$\varepsilon = 0.05$	0.1	0.2	0.4	0.8
tlf-tlf	EER	9.7	**9.6**	10.1	10.2	10.2
	DCF_{min}	0.046	**0.045**	0.046	0.047	0.047
tlf-mic	EER	17.0	14.3	**13.3**	13.9	13.9
	DCF_{min}	0.059	0.053	**0.052**	0.053	0.053
mic-mic	EER	15.5	13.8	13.4	**13.0**	**13.0**
	DCF_{min}	0.062	0.060	0.057	**0.057**	0.057
mic-tlf	EER	17.1	**15.0**	15.7	15.3	15.3
	DCF_{min}	0.062	0.062	0.061	**0.059**	0.059

Finally, we compare the performance of the two approaches, SVC-GLDS and SVR-GLDS, with and without NAP compensation scheme. Table 1 shows the comparison in EER and DCF_{min} values for each condition and Fig. 2 *b)* shows the global DET curves of the systems. The system with the best performance in most part of the cases is SVC-GLDS + NAP, obtaining a relative improvement in EER of 31% and 19% in DCF_{min} value. However, the proposed system, SVR-GLDS, presents a similar performance before and after channel compensation. This has the advantage that there is no need of using NAP to obtain similar performance as SVC-GLDS + NAP. It is worth noting that if no channel compensation could be applied because the non-availability of a background database, the SVC-GLDS performance worsens significantly, especially when strong session mismatch occurs (*tlf-mic* and *mic-tlf*). If a suitable database is available, NAP may significantly improve the performance, but if such database is not available or the representative data is scarce, SVR-GLDS seems a convenient option for obtaining robustness. The latter may be the case in many real applications, such as the forensic environment. Moreover, if a suitable database is available SVR-GLDS + NAP provides just a reduced improvement, in both EER and DCF_{min} values (5% and 3% respectively), with respect to SVR-GLDS.



a) b)

Fig. 2. DET curves in NIST SRE 2008 short2-short3 task: *a)* SVR + NAP ($\varepsilon = 0.1$) and SVR + NAP (ε optimum); *b)* SVC, SVC + NAP, SVR and SVR + NAP

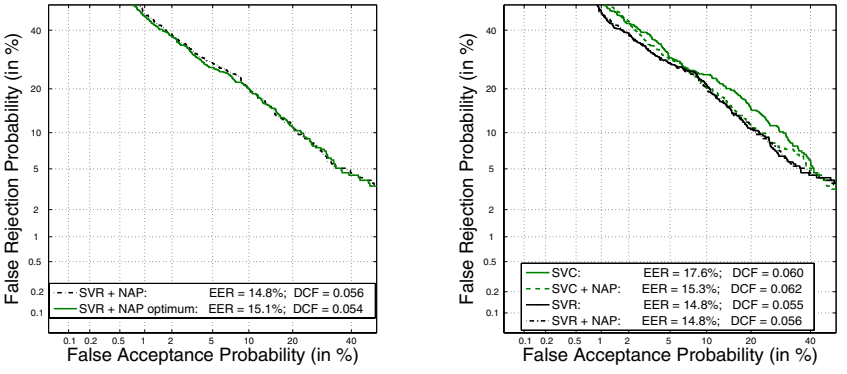
Real Forensic Conditions in Ahumada III. In order to show the performance of the proposed system in similar conditions to those found in real forensic cases, Fig. 3 *b)* shows the SVC-GLDS performance with and without including NAP compensation over Ahumada III. As we observed in NIST SRE 2008, the performance of the system improves when NAP is added, but in this case the relative improvement is significantly lower (13% versus 31% in EER). Moreover, it is observed a degradation in DCF_{min} performance after NAP compensation. The loss in NAP compensation effectiveness can be attributed to the lack of background data in operational conditions. Thus, when a high *database mismatch* is observed among the background and the operational databases, session variability compensation techniques are not only less efficient, but can also even degrade performance [8].

In order to be robust to such lack of background data, the proposed SVR-GLDS approach is used. First, we perform an experiment to show the variability of performance with respect to the ε value. Table 4 presents such results.

Table 4. EER and DCF_{min} in Ahumada III, for different values of ε in SVR-GLDS with and without NAP session variability compensation

		$\varepsilon = 0.05$	0.1	0.2	0.4	0.8
SVR	EER (%)	**14.6**	14.8	15.5	17.4	17.6
	DCF_{min}	0.055	**0.055**	0.058	0.058	0.059
SVR + NAP	EER (%)	15.1	**14.8**	15.6	15.3	15.3
	DCF_{min}	**0.054**	0.056	0.059	0.062	0.062

The system performance with and without NAP is similar, as we saw in NIST SRE 2008 (Table 2 and Table 3). The optimal ε value lays between 0.05 and 0.1, Fig. 3 *a)* shows a comparison between the performance of SVR-GLDS + NAP with these ε values, the technique is not very sensitive. The system performance is similar. Finally, the DET curves of the two approaches with and without session variability compensation are showed in Fig. 3 *b)*.



a) b)

Fig. 3. DET curves in Ahumada III: *a)* SVR + NAP ($\varepsilon = 0.1$) and SVR + NAP (ε optimum); *b)* SVC, SVC + NAP, SVR and SVR + NAP

We observe that using a forensic corpus under database mismatch conditions without any compensation scheme the SVR performance is better than SVC (relative improvement of 16% in EER and 8% in DCF_{min} value), a similar situation can be seen in Fig. 2 b) over NIST SRE 2008. Once we have included NAP the performance of SVR and SVC is similar, but slightly better for SVR. These results are different than those presented for NIST SRE 2008, where in general SVC-GLDS + NAP outperformed SVR-GLDS + NAP. In forensic case, where suitable databases are difficult to obtain SVR seems a more convenient option for obtaining robustness.

5 Conclusions

In this paper we propose a robust approach for speaker verification by means of Support Vector Machine Regression (SVR). The presented work shows that SVR using a GLDS kernel is robust to the lack of a proper background set for NAP session variability compensation, clearly outperforming Support Vector Machine Classification (SVC) in such a situation. This is in accordance with previous work of the authors, where telephone-only speech was used [9]. In this work, two much harder scenarios are proposed. First, NIST SRE 2008 core task is used as a highly mismatching database with multichannel data. Results in this scenario show similar performance among SVC and SVR when NAP is trained with a proper background dataset. However, we simulate the lack of such a database by eliminating the compensation step, and SVR clearly outperforms SVC, showing a much higher robustness. Second, Ahumada III database is used, which consists of speech from real forensic cases. In this scenario, where a background database is not available (i.e., under database mismatch), results show a much lesser effectivity of the NAP compensation technique. Moreover, SVR performs better than SVC, confirming the robustness simulated in NIST SRE 2008.

This work shows that, if a suitable background database for NAP is not available, SVR outperforms SVC, being also a better option in order to obtain robustness to

unseen conditions. Moreover, NAP may significantly improve the performance of the system, but under database mismatch its effectiveness is significantly reduced. This is especially important in forensic scenarios, where the availability of a proper database adapted to the case at hand may be almost impossible in many situations.

Future work includes the exploration of different SVR approaches for the GLDS space, such non-linear loss functions and different kernels. We will also explore the complementarity and correlation of SVR with respect to other approaches in the state of the art in speaker verification such as GMM and GMM-SVM.

Acknowledgements. This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. We also thank Lt. Cln. Jose Juan Lucena and people from the Acoustics and Image Processing Department from Guardia Civil for their important effort in collecting data for forensic purposes.

References

1. National Institute of Standards and Technology (NIST), 2008 speaker recognition evaluation plan (2008), http://www.nist.gov/speech/tests/sre/2008/index.html
2. Reynolds, D.A.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
3. Campbell, W.M., Quatieri, T.F., Dunn, R.B.: Support Vector Machines for Speaker and language Recognition. Computer Speech and Language 20, 210–229 (2006)
4. Solomonoff, A., Campbell, W.M., Boardman, I.: Advances in Channel Compensation for SVM Speaker Recognition. In: Proc. Of ICASSP, pp. 629–632 (2005)
5. Kenny, P., Oullet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing 16(5), 980–988 (2008)
6. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-Carrasquillo, P.A.: Support Vector Machines using GMM Supervectors for Speaker Verification. Signal Processing Letters 13(5), 308–311 (2006)
7. Brümmer, N., et al.: Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. IEEE Transactions on Audio, Speech and Language Processing 15(7), 2072–2084 (2007)
8. Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Lucena-Molina, J.J.: Addressing Database Mismatch in Forensic Speaker Recognition with Ahumada III: a Public Real-Casework Database in Spanish. In: Proc. Of Interspeech, pp. 1493–1496 (2008)
9. Lopez-Moreno, I., Mateos-Garcia, I., Ramos, D., Gonzalez-Rodriguez, J.: Support Vector Regression for Speaker Verification. In: Proc. Of Interspeech, pp. 306–309 (2007)
10. Smola, A.J., Schoelkopf, B.: A Tutorial on Support Vector Regression. Tech. Rep. NeuroCOLT2 Technical Report NC2-TR-1998-030, Royal Holloway College (1998)