



Support Vector Regression for Speaker Verification

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

ATVS (Biometric Recognition Group), C/ Francisco Tomas y Valiente 11,
Universidad Autonoma de Madrid, E28049 Madrid, Spain
{ignacio.lopez, ismael.mateos, daniel.ramos, joaquin.gonzalez}@uam.es

Abstract

This paper explores Support Vector Regression (SVR) as an alternative to the widely-used Support Vector Classification (SVC) in GLDS (Generalized Linear Discriminative Sequence)-based speaker verification. SVR allows the use of a ϵ -insensitive loss function which presents many advantages. First, the optimization of the ϵ parameter adapts the system to the variability of the features extracted from the speech. Second, the approach is robust to outliers when training the speaker models. Finally, SVR training is related to the optimization of the probability of the speaker model given the data. Results are presented using the NIST SRE 2006 protocol, showing that SVR-GLDS yields a relative improvement of 31% in EER compared to SVC-GLDS. **Index Terms:** speaker verification, GLDS, SVM classification, SVM regression

1. Introduction

Speaker verification has been dominated in the last decade by systems working at the spectral level of the speaker identity [1]. Techniques like Gaussian Mixture Models (GMM) [2] or Support Vector Machines (SVM) using Generalized Linear Discriminant Sequence (GLDS) kernels [3] have demonstrated its superiority to higher level approaches [1, 4]. In recent years, hybrid approaches such as GMM-SVM systems [5] and channel compensation techniques like factor analysis [6] or nuisance attribute projection [7] have led to a significant improvement of the state-of-the-art performance.

One of the techniques which have yielded a good performance at the spectral level is SVM-GLDS speaker verification [3]. Using this technique, parameters are mapped to a high-dimensional space via a GLDS kernel function. Then, a SVM classifier is used in order to discriminate genuine users from impostors at that high dimensional space. The performance of SVM-GLDS speaker verification systems has demonstrated to be similar to the GMM modelling. Also, the fusion of SVM-GLDS classification with other approaches at the spectral level significantly improves performance [3].

SVMs have demonstrated their efficiency and accuracy in solving two main problems: *i*) discriminating among classes (classification) and *ii*) function estimation (regression). In the former the objective is to compute a class for every feature extracted from the data. In the latter the aim is finding a good approximation to a function of the features. In this sense, regression is a more general approach than classification, as a class label is indeed a function of the features. As speaker verification is essentially a binary class problem, most popular schemes are

This work was partially funded by the Spanish Ministry of Education under project TEC2006-13170-C02-01.

based on SVM classifiers (SVC). However, as we will show, a more general and robust approach can be adopted by using SVM regression (SVR). In this paper we propose the use of SVR for speaker verification using a GLDS kernel. Reported results using NIST SRE 2006 experimental protocol show a significant improvement of SVR-GLDS versus SVC-GLDS.

This work is organized as follows. SVM classification and regression is introduced in Section 2, highlighting their main differences. SVM regression for GLDS speaker verification (SVR-GLDS) is presented in Section 3. In Section 4, Experiments showing the adequacy of the proposed technique are presented. Finally, conclusions are drawn in Section 5.

2. Support Vector Machine Classification and Regression

SVM derive from the Vapnik's statistical learning theory [8], and since 1994 they have been largely used for pattern recognition due to its excellent generalization properties. For instance, a well known effect of SVM is that the number of observations and its dimensionality do not affect to SVM generalization [9]. These properties, added to the efficiency and elegance of kernel methods [10], make SVM giving an excellent performance in many different tasks. The good discrimination of SVM-based speaker verification systems [3, 5] supports this fact.

In this section we describe the use of Support Vector Machines for both classification and regression. We compare both methods and we highlight the main differences between them.

2.1. Support Vector Machine Classification (SVC)

Suppose we have l vectors $x_i \in \mathbb{R}^n$ from two different classes. Each class is labelled as $y_i \in [+1, -1]$. The classification problem consist in assigning each x_i to its corresponding class y_i . The SVC approach finds an optimal hyperplane \mathbf{w} which separates \mathbb{R}^n in two regions: vectors in one of the regions will be assigned to the class +1 and the rest to the class -1. We define the scoring function $f(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$, which measures the distance of each vector to the separating hyperplane \mathbf{w} :

$$f(x) = \langle \mathbf{w}, x \rangle + b \tag{1}$$

where b is a learned offset parameter. If the data set $D = \{(x_1, y_1), (x_2, y_2) \dots (x_l, y_l)\}$ is linearly separable, $f(\cdot)$ will be positive for all values of x_i where $y_i = +1$ and negative otherwise.

However, there are many effects which may cause overlapping between classes, e. g. noise, channel effects, intra- and inter- class variability, etc. Therefore, some vectors will be incorrectly classified. In this case, we will have two different criteria for finding \mathbf{w} : *i*) maximizing the margin between classes

and *ii*) minimising a loss function proportional to missclassified vectors. A weighting factor C controls the relevance of one criteria against the other, as it can be seen in the following formula:

$$\mathbf{w} = \arg \min_{\mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} + C \frac{1}{m} \sum \xi_{c,i} \right) \quad (2)$$

subject to $0 \leq \xi_{c,i} \leq 1 - y_i f(x_i)$

Here, $\xi_{c,i}$ is a slack variable associated to the non-optimally classified vector i ($i \in \{1, \dots, m\}$) in a classification problem, and it will only be non-zero for those x_i which make $y_i \cdot f(x_i) < 1$. Notice that if $0 < y_i \cdot f(x_i) < 1$, x_i will be correctly classified but its associated $\xi_{c,i}$ value will be different to 0. Thus, for classification problems the loss function is defined as:

$$f_{loss}(x_i) = \max\{0, 1 - y_i \cdot f(x_i)\} \quad (3)$$

Non-linear classification can be solved by using $\phi(x_i)$ instead of x_i . The function $\phi(\cdot)$ maps each vector to a higher dimensional feature space where vectors are linearly separable. As SVM only require the inner product of the vectors in the features space $\langle \phi(x_i), \phi(x_j) \rangle$, we define the kernel function as:

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (4)$$

The kernel function $k(x_i, x_j) \in \mathbb{R}$ allows us to compute $\langle \phi(x_i), \phi(x_j) \rangle$ without explicitly mapping each vector into the high dimensionality space. This is known as the kernel trick.

2.2. Support Vector Machine Regression

In the regression problem, y_i is not a class but any other function of x_i . Therefore SVR can be used to learn n -dimensional functions $g_n(\cdot)$ such as

$$g_n(x_i) = y_i \quad (5)$$

The goal in the regression problem is to approximate $f(\cdot) \simeq g_n(\cdot)$. Notice that, although $g_n(\cdot)$ can take either continuous or discrete values, the SVR approximation will always be a continuous function. In the SVR case, the C parameter is used to control how much we need to approximate $f(\cdot)$ to $g_n(\cdot)$.

Regarding the loss function, the main difference of regression with respect to classification is that errors are penalized not only when $f(\cdot) < g_n(\cdot)$ but also when $f(\cdot) > g_n(\cdot)$. Therefore, the loss function has to be modified in order to take a different behavior than in the classification case because, for classification errors, this penalty was only applied when $y_i \cdot f(x_i) - 1 < 0$.

A popular loss function for regression is the ε -insensitive loss function [11]. This function tolerates some degree of mismatch by the use of an margin controlled by the ε parameter. As errors only occur when $|f(\cdot) - g_n(\cdot)| > \varepsilon$, the SVR training goal is to find \mathbf{w} such as:

$$\min \left(\frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} + C \frac{1}{m} \sum (\xi_{r,i} + \xi'_{r,i}) \right) \quad (6)$$

subject to $\begin{cases} 0 \leq f(x_i) - y_i - \varepsilon \leq \xi_{r,i} \\ 0 \leq y_i - f(x_i) - \varepsilon \leq \xi'_{r,i} \end{cases}$

As it can be seen, two different slack variables are introduced for regression: $\xi_{r,i}$ for those vectors for which $f(x_i) > g_n(x_i) + \varepsilon$, and $\xi'_{r,i}$ for those ones that $f(x_i) < g_n(x_i) - \varepsilon$. The loss function is now defined as:

$$f'_{loss}(x_i) = \max\{0, |y_i - f(x_i)| - \varepsilon\} \quad (7)$$

Figure 1 illustrates $f'_{loss}(\cdot)$ and its differences with $f_{loss}(\cdot)$.

An interesting property of SVR which does not apply for SVC is that the ε -insensitive loss function leads to a maximum-a-posteriori (MAP) estimation of \mathbf{w} [12]. It can be shown that $e^{-f'_{loss}(\cdot)}$ is proportional to $p(\mathbf{w} | D, \varepsilon)$, i. e. the posterior probability of \mathbf{w} given the data and the value of the ε margin. Therefore, by minimizing $f'_{loss}(\cdot)$ we will maximize the log-probability that $f(\cdot) = g_n(\cdot)$.

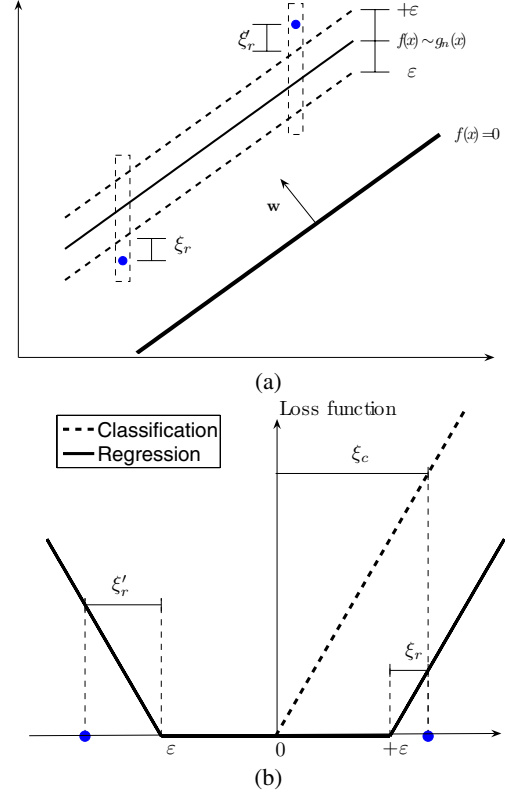


Figure 1: SVR versus SVC. Boundaries (a) and loss function (b). The loss functions are centered at $f(x_i) = y_i$ for SVC (f_{loss}) and at $f(x_i) = g_n(x_i)$ for SVR (f'_{loss}). f_{loss} penalizes x_i such as $y_i \cdot f(x_i) - 1 < 0$, while f'_{loss} penalizes x_i such as $|f(x_i) - g_n(x_i)| > \varepsilon_i$.

3. SVR-GLDS Speaker Verification

Speaker verification is a two-class classification problem. The objective is to take a decision about if a testing utterance corresponds to a claimed identity or not. In widely used SVC-GLDS speaker verification, for each SVM speaker model, the class label will take the value 1 for the target vectors belonging to the speaker and -1 for nontarget vectors from anyone else.

Our proposal is to use a SVR with an ε -insensitive loss function for classification. Thus, the SVR goal function $g_n(\cdot)$ is discrete and it only takes two different values, namely $g_n(\cdot) \in \{+1, -1\}$ for target and nontarget speakers respectively. Note that for this problem, the support vectors will not be the nearest ones to \mathbf{w} , as in classification, because in such case they would minimize $f_{loss}(\cdot)$, but they would not minimize $f'_{loss}(\cdot)$. This difference from the standard SVC makes SVR more robust against outliers or noisy vectors being used for obtaining \mathbf{w} , because in SVR the support vectors are selected from regions in the feature space where vectors of each class are more concentrated.

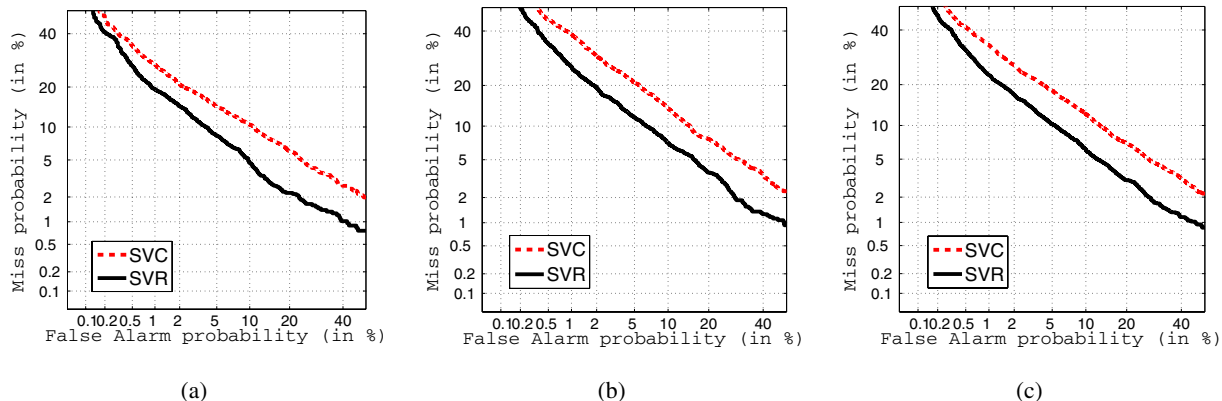


Figure 2: Comparison of SVC-GLDS and SVR-GLDS in NIST SRE 2006 ($\varepsilon = 0.1$) 1conv4w-1conv4w task for male (a), female (b) and pooled gender (c) data.

On the other hand, SVC uses a set of support vectors which are nearer the frontier between classes, where vectors of each class use to be scarce. Thus, SVC hyperplane may be more sensitive than SVR to outliers in the support vectors.

Finally, an optimal training ε -insensitive SVR requires adequate tuning of C and ε parameters. Some works in the literature [13] relate the ε parameter to the noise or variability of the function to estimate. Therefore, the optimal value of ε allows us to obtain a quantitative measure of the feature variability in speaker verification problems. Moreover, optimizing the ε parameter adapts the SVR training process to the observed variability in the data.

4. Experiments

4.1. Baseline system

Our baseline system is a SVM-GLDS speaker recognition system as described in [3]. Feature extraction obtains 19 MFCC coefficients plus deltas. In order to avoid channel mismatch effects, cepstral mean normalization is applied, followed by RASTA filtering and feature mapping (see [4] for details). The similarity computation is based on SVC [3]. A GLDS kernel expansion is performed on the whole observation sequence, and a separating hyperplane is computed between the speaker features and the background model. The system uses a polynomial expansion of degree three [14] prior to the application of the GLDS kernel. We have used the LibSVM library [15] for both SVM classification and regression. Finally, Tnorm [16] score normalization technique is performed in order to scale the scores distribution.

4.2. Database and experimental protocol

Experiments have been performed using the evaluation protocol proposed by NIST in its 2006 Speaker Recognition Evaluation (SRE) [17]. The database used in this evaluation consists of: *i*) a subcorpus of the MIXER database [18] and *ii*) a significant amount of additional multi-channel and multi-language data acquired in order to complete the corpus for the evaluation. The acquisition conditions include different communication channels (landline, GSM, CDMA, etc.), different handsets and microphones (carbon button, electret, earphones, cordless, etc.) and different languages (American English, Arabic, Spanish, Mandarin, etc.). The evaluation protocol defines the following training conditions: 10 seconds, 1, 3 and 8 conversation sides; and

the following test conditions: 10 seconds, 1 conversation side, 3 full conversations in a mixed channel and multichannel microphone data. Each conversation side has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Although there are speakers of both genders in the corpus, no cross-gender trials are defined. Details can be found in the NIST webpage (www.nist.gov/speech). In our case the experiments followed the 1 conversation side training conditions, and 1 conversation side test condition (1conv4w-1conv4w). The background set for system tuning is a subset of databases from previous NIST SREs. Trials performed using this development set follow the corresponding NIST SRE protocol. The Tnorm cohorts were extracted from the NIST 2005 SRE targets models for each training condition.

4.3. Results

First of all, we have investigated the variation of the performance of the proposed SVR-GLDS system with respect to the parameter ε as defined in Section 2.2 (Equation 6). Tables 1 and 2 show the performance for different values of ε . Results are presented both as Equal Error Rate (EER) and DCF_{min} as defined by NIST [17]. It is observed that the performance of the system significantly improves for values around $\varepsilon = 0.1$, both for EER and DCF values. Therefore, the value $\varepsilon = 0.1$ will be used for SVR for the experiments presented below.

ε	0.01	0.05	0.1	0.2	0.4	0.8
EER(%)	9.1	7.8	6.9	8.4	9.9	10.3
$DCF_{min} \cdot 10^2$	3.5	3.2	2.9	3.5	3.7	3.7

Table 1: EER and DCF_{min} in NIST SRE 2006 male 1conv4w-1conv4w, for different values of ε .

ε	0.01	0.05	0.1	0.2	0.4	0.8
EER(%)	11	8.6	8.5	9.7	11.9	12
$DCF_{min} \cdot 10^2$	4.1	3.5	3.6	4.2	4.7	4.8

Table 2: EER and DCF_{min} in NIST SRE 2006 female 1conv4w-1conv4w, for different values of ε .

We have also evaluated the performance of SVR-GLDS versus the SVC-GLDS baseline system. Table 3 shows the differences between them in terms of EER and DCF_{min} . It is shown that SVR obtains a relative improvement in EER of 34% for male, and 29% for female, whereas the relative improvement of the DCF_{min} value is 22% and 25% in the male and female cases respectively. Finally, Figure 2 shows the discrimination performance of SVR-GLDS versus SVC-GLDS for the male, female and pooled gender cases. We can observe a significant performance improvement at all operating points in the DET curve for all gender conditions.

	Male		Female		Pooled	
	SVC	SVR	SVC	SVR	SVC	SVR
EER(%)	10.4	6.9	12	8.5	11.3	7.8
DCF	3.7	2.9	4.8	3.6	4.3	3.3

Table 3: SVC-GLDS and SVR-GLDS systems in NIST SRE 2006 1conv4w-1con4w task. It shows the EER(%) and $DCF_{min} \cdot 10^2$ for male and female genders.

5. Conclusions

In this paper we have presented a Support Vector Machine Regression (SVR) approach for speaker verification in the GLDS kernel space. This technique presents advantages with respect to Support Vector Machine Classification (SVC). First, the loss function used is related to the variability present in the feature space. Thus, varying the parameter ε we can adapt to such variation, which may be due to inter-session variability (e. g., channel mismatch) or intra-speaker variability. Second, the technique is more robust against outliers. Finally, the regression technique optimizes the posterior probability of the model \mathbf{w} given the data and the ε parameter. Reported results have demonstrated the adequacy of support vector regression (SVR) for speaker verification with a GLDS kernel function, as significant improvements are shown both in EER and DCF_{min} . Future work includes the use of different SVR approaches for the GLDS space, such as ν -SVR [10], non-linear loss functions and different kernels. Also, the application of SVR to other SVM-based speaker recognition systems as GMM Supervectors [5], and non singular class labelling will be considered. Finally, the proposed technique will be tested in different databases in order to explore its robustness to environmental changes.

6. References

- [1] D. A. Reynolds, "An overview of speaker recognition technology," in *Proc. of ICASSP*, 2003, pp. 4072–4075.
- [2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [3] 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*, vol. 20, no. 2-3, pp. 210–229, 2006.
- [4] J. Gonzalez-Rodriguez, D. Ramos-Castro, D. Torre-Toledano, A. Montero-Asenjo, J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Fierrez-Aguilar, D. Garcia-Romero, and J. Ortega-Garcia, "On the use of high-level information for speaker recognition: the ATVS-UAM system at NIST SRE 2005," *IEEE Aerospace and Electronic Systems Magazine*, pp. 15–21, January, 2007.
- [5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters*, vol. 13(5), pp. 308–311, 2006.
- [6] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. of ICASSP*, 2004, vol. 1, pp. 37–40.
- [7] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," in *Proc. of ICASSP*, 2005, pp. 629–632.
- [8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1999.
- [9] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [10] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods and Support Vector Learning*, MIT Press, 2000.
- [11] K. Muller, A. J. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Proc. of the 7th International Conference on Artificial Neural Networks*, 1997, vol. 1327 of *Lecture Notes In Computer Science*, pp. 999–1004.
- [12] P. Sollich, "Probabilistic methods for support vector machines," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K. Müller, Eds., vol. 12, pp. 349–355. MIT Press, 1999.
- [13] A. J. Smola and B. Schoelkopf, "A tutorial on support vector regression," Tech. Rep. NeuroCOLT2 Technical Report NC2-TR-1998-030, Royal Holloway College, University of London, UK, 1998.
- [14] W. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in *Proc. of IEEE International Workshop on Neural Networks for Signal Processing*, 2000, pp. 775–784.
- [15] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [16] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.
- [17] NIST, "2006 speaker recognition evaluation plan: <http://www.nist.gov/speech/tests/spk/2006/index.htm>," 2006.
- [18] J. P. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. F. Martin, and M. A. Przybocki, "The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation," in *Proc. of Odyssey*, 2004, pp. 29–32.