30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Feature extraction for murmur detection based on support vector regression of time-frequency representations

J. Jaramillo-Garzón, A. Quiceno-Manrique, I. Godino-Llorente and C.G. Castellanos-Dominguez

***Abstract*— This paper presents a nonlinear approach for time-frequency representations (TFR) data analysis, based on a statistical learning methodology - support vector regression (SVR), that being a nonlinear framework, matches recent findings on the underlying dynamics of cardiac mechanic activity and phonocardiographic (PCG) recordings. The proposed methodology aims to model the estimated TFRs, and extract relevant features to perform classification between normal and pathologic PCG recordings (with murmur). Modeling of TFR is done by means of SVR, and the distance between regressions is calculated through dissimilarity measures based on dot product. Finally, a k-nn classifier is used for the classification stage, obtaining a validation performance of 97.85%.**

I. INTRODUCTION

Cardiac mechanical activity is appraised by auscultation and processing of heart sound records (known as phonocardiographic signals - PCG), being an inexpensive and non-invasive procedure. Although the importance of classic auscultation methods has decreased due to its inherent restrictions (performance of human ear with its physical limitations, subjectivity of the examiner, etc), the PCG has preserved its importance in pediatric cardiology, cardiology, and internal diseases, evaluating congenital cardiac defects [1], and in primary home health care, where an intelligent stethoscope with decision support abilities would be of a great value [1].

In recent years computer-based PCG analysis methods have been the subject of considerable effort. Indeed, the automatic detection of cardiac murmurs strongly depends on the appropriate features (data representation), which mostly are related to timing, morphology and spectral properties of heart sounds. Cardiac murmurs are non-stationary signals and exhibit sudden frequency changes and transients, but it is common to assume linearity of the feature sets extracted from heart sounds (time and spectral characteristics, frequency representation with time resolution, and parametric modeling). In the analysis of biomedical data, such as PCG, time frequency representations (TFR) have been proposed to investigate the dynamic properties of spectral parameters during transient physiological or pathological episodes. Moreover, TFR not only distinguishes murmurs of different kinds intuitively, but also offers quantitative data [2].

In [3] the Choi-Williams distribution (CWD) was found to be an adequate representation to the description, feature extraction and classification of a significant data set of PCG signals in order to evaluate valvular pathologies.

Furthermore, given the detailed time and frequency resolution of time-frequency distribution, trainable automatic classifiers can easily be overwhelmed by the complexity of this input representation. An ultimate goal of the time-frequency or time-scale signal analysis is to have the best possible representation for the subsequent automatic classification of signals. Unfortunately, many of the highest resolution time-frequency techniques produce a high degree of detail in all regions of the time-frequency plane without regard to the need for such minutia in each region. This is because the signal to be classified has been globally over-parameterized by the time-frequency representation and thus the model would be severely under-trained [4]. In this way, we use SVR to model the TFR, as was suggested in [5], where SVR is used to model an optimization surface oriented to the detection of murmurs in PCG signals.

Our goal in this paper is to apply the time-frequency representation in PCG signals and use the SVR for modeling the surfaces, in order to reduce the amount of data in the original TFR. With the obtained support vectors, it is possible to employ dissimilarity measures based on dot products in order to obtain an estimation of distance between models of TFRs and therefore calculate the features that will be useful in the classification stage with a k-nn classifier.

This paper is organized as follows: Section II explains the background methods supporting the proposed methodology, namely, support vector regression (SVR) and dissimilarity based classification; section III explains our proposed dissimilarity measure between regression models, tests and experimental conditions, and finally, results and conclusions are presented in sections IV and V, respectively.

II. BACKGROUND METHODS

A. Statistical Learning by Support Vector Regression

The Support Vector Regression (SVR), which is grounded on Structural Risk Minimization theory [6], is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. Given N input sample sets, $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, 2, ..., N$, and N corresponding scalar output values, $y_i \in \mathbb{R}$, $i = 1, 2, ..., N$, the aim is to find a regression function of the form:

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

J. Jaramillo-Garzón, A. Quiceno-Manrique and C.G. Castellanos-Dominguez are with Control and Digital Signal Processing Group, Universidad Nacional de Colombia, sede Manizales, Colombia.
{jajaramilog,afquicenom,cgcastellanosd}@unal.edu.co
I. Godino-Llorente is with Circuits and Systems Engineering Department, Universidad Politécnica de Madrid, España.
igodino@ics.upm.es

978-1-4244-1815-2/08/$25.00 ©2008 IEEE.

This can be achieved by introducing the so called ε-*insensitive loss function*:

$$|y - f(\mathbf{x})|_\varepsilon := \max\{0, |y - f(\mathbf{x})| - \varepsilon\} \tag{2}$$

which does not penalize errors below some $\varepsilon \geq 0$.
The training process of the SVR is to find an optimal vector $\mathbf{w}$ by solving the convex optimization problem [6]:

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ \text{Subject to} & \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x_i}\rangle - b \leq \varepsilon + \xi_i \\ \langle \mathbf{w}, \mathbf{x_i}\rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \\ \xi_i,\ \xi_i^* \geq 0 \end{cases} \end{array} \tag{3}$$

Preselected constants, $\varepsilon > 0$ and $C > 0$, are the insensitivity value and the tradeoff between the smoothness of the SVR function and the total training error, respectively.
Solving (3) results in the so called *support vector expansion*:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \langle \mathbf{x}_i, \mathbf{x}\rangle + b \tag{4}$$

showing that $\mathbf{w}$ can be completely described as a linear combination of the training patterns $\mathbf{x}_i$, with α_i, $i = 1, ..., N$ being the Lagrange multipliers introduced for solving the problem.
When the approximation function can not be linearly regressed, it is necessary to introduce a mapping function from the input space to a high dimensional feature space $\Phi : \mathbf{x} \mapsto \Phi(\mathbf{x})$, in such a way that the function $f(\cdot)$ between the output and the mapped input data points can now be linearly regressed in the feature space. This can be done implicitly through the use of a *kernel function* $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y})\rangle$ for replacing the dot products in (4):

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{5}$$

Kernel functions are characterized by Mercer conditions, and a commonly used one is the Gaussian radial basis function (RBF) [7]:

$$k(x, y) = exp(-\gamma\|\mathbf{x} - \mathbf{y}\|^2) \tag{6}$$

with the parameter γ being a pre-selected constant.

B. Dissimilarity-based Classification

Suppose a set of *prototype objects*

$$\mathcal{R} := \{p_1, p_2, ..., p_r\} \tag{7}$$

called the *representation set*, and suppose a dissimilarity measure $d(\cdot, \cdot)$, computed or derived from the objects. Such a dissimilarity measure must be nonnegative and to obey the reflexivity condition, $d(x, x) = 0$, but it might be non-metric. An object x is represented as a vector of the dissimilarities computed between x and the prototypes from $\mathcal{R}$:

$$D(x, \mathcal{R}) = [d(x, p_1), d(x, p_2), ..., d(x, p_r)] \tag{8}$$

Then, for a training set $\mathcal{T}$ of m objects, a classifier can be built on the $m \times r$ dissimilarity matrix $D(\mathcal{T}, \mathcal{R})$ relating all training objects to all prototypes [8].

There exists a number of ways to select the representation set $\mathcal{R}$. One method that has achieved good results is *Linear Programming* (LP). In this method, the selection of prototypes is done automatically by training a properly formulated separating hyperplane

$$f(D(x, \mathcal{R})) = \sum_{j=1}^{r} w_j d(x, p_j) + w_0 = w^T D(x, \mathcal{R}) + w0 \tag{9}$$

in a dissimilarity space $D(T, \mathcal{R})$. In this approach, a sparse solution w is obtained, which means that many weights w_j become zero. The objects from the initial set $\mathcal{R}$ ($\mathcal{R} = \mathcal{T}$, for instance), corresponding to nonzero weights are the selected prototypes, so the representation set $\mathcal{R}_{LP}$.

III. Experimental Outline

A. Dissimilarity measure between regressions

Given two modeled TFRs, the next step is the calculus of some dissimilarity measure between regression functions. In this case, a distance measure is proposed as follows. First, the dot product between functions is computed:

$$\langle f_1(\cdot), f_2(\cdot)\rangle = \langle \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \cdot), \sum_{j=1}^{N} \beta_j k(\mathbf{x}_j, \cdot)\rangle \tag{10}$$

(10) can be expressed as [9]:

$$\langle f_1, f_2\rangle = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{11}$$

Leading to the matrix form:

$$\langle f_1, f_2\rangle = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\beta} \tag{12}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are column vectors containing the α_i, $i = 1, ..., N$ and $\beta_i, i = 1, ..., N$ coefficients, and $\mathbf{K}$ is the *kernel matrix*, a positive semidefinite and symmetric matrix such that its entries are $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.
Now, it is possible to derive a distance function between regressions, as:

$$\begin{aligned} d(f_1, f_2) &= (\langle f_1 - f_2, f_1 - f_2\rangle)^{1/2} \\ &= \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} \end{aligned} \tag{13}$$

Although the calculus of the entire kernel matrix is a high demanding computational task, it is clear that the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors are sparse, so just a few entries of the kernel matrix have to be computed.

B. Database

The database used in this work is made up with 22 de-identified adult subjects (16 normals and 6 with murmur). Eight recordings were taken from each patient, corresponding to the four traditional focuses of auscultation (mitral, tricuspid, aortic and pulmonary areas) in the phase of post-expiratory and post-inspiratory apnea. An electronic stethoscope (`WelchAllyn® Meditron` model) was used to acquire the heart sounds simultaneously with a standard 3-lead ECG. Both signals were digitized at 44.1 kHz with 16-bits per sample.

Furthermore, in order to select beats without artifacts and another type of noise that can degrade the performance of the algorithms, a visual and audible inspection was carried out by cardiologists, and 402 individual beats were extracted, 201 for each class, using a R-peak detector.

C. Time-frequency representation

Time-frequency representations of PCGs were obtained using two methods: Short Time Fourier Transform (STFT) and Choi-Williams distribution (CWD). Both representations map the signal energy into the time-frequency plane. STFT is easy to compute, but it has resolution problems, caused by the uncertainty principle, when the signal is windowed. On the other hand, CWD provides better resolution than STFT and it would be able to improve the characterization of PCG signals.

Both STFT and CWD belong to the general class of time-frequency distributions (Cohen class), and they can be computed using the expression (14). To calculate the different distributions, it is only necessary to change the kernel $\phi(\theta,\tau)$.

$$\begin{aligned} P(t,\omega) = & \frac{1}{4\pi^2} \int\int\int e^{-j\theta t - j\tau\omega + j\theta u}\phi(\theta,\tau) \\ & \cdot s^*\left(u - \frac{1}{2}\tau\right) s\left(u + \frac{1}{2}\tau\right) du d\tau d\theta \qquad (14) \end{aligned}$$

where $s(u)$ is the signal to be analyzed, τ is the time delay and θ is the frequency lag.

The kernels to compute the TFRs are shown in (15), (15a) is used to compute STFT, and (15b) for CWD. In (15a), $h(u)$ is the windowing function. On the other hand, in (15b) σ is a positive parameter controlling the kernel concentration around the origin of the time and frequency lag plane and, hence, the overall amount of smoothing.

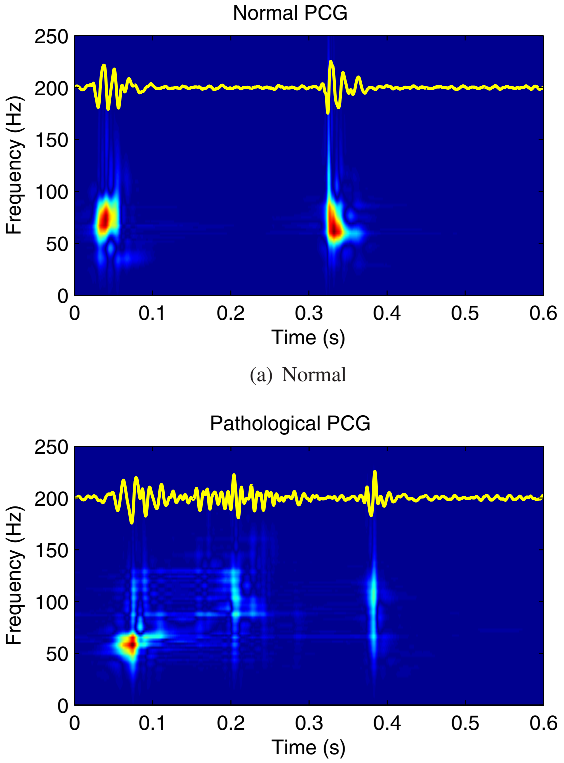
$$\phi_{STFT}(\theta,\tau) = \int h^*(u - \frac{1}{2}\tau)h(u + \frac{1}{2}\tau)e^{-j\theta u}du \qquad (15a)$$

$$\phi_{CWD}(\theta,\tau) = e^{-\theta^2\tau^2/\sigma} \qquad (15b)$$

In the implementation of TFRs, PCG signals were down-sampled to $4000Hz$ in order to reduce the amount of data to analyze and were length-normalized and zero padded to 4800 points.

After preprocessing, the TFR for each beat was computed. The STFT was estimated using a Hamming window whose length was $50ms$ or 200 points with overlap of 50% and 256 point in frequency domain. The CWD was calculated using the fast algorithm proposed in [10], with the kernel parameter $\sigma = 2$ and 256 points in frequency. The CWD of two PCG signals (normal and pathologic) is shown in Figure 1.

As a result, in STFT a matrix with 256 rows and 47 columns was obtained; whereas in CWD the size of the obtained matrix was 256×4800. Finally, the amount of data was reduced again by using 2-D down-sampling, obtaining matrices of 64 rows and 47 columns for STFT, and 64×192 for CWD.



(a) Normal

(b) With murmur

Fig. 1. Examples of Choi-Williams distributions of different PCG signals.

D. Regression model selection

The parameters ε (insensitive width) and C (error penalization) were adjusted by using the approximations given in [11]:

$$\varepsilon = 3\sigma_{noise}\sqrt{\ln N/N} \qquad (16)$$

$$C = \max|\bar{y} + 3\sigma_y|, |\bar{y} + 3\sigma_y| \qquad (17)$$

where σ_{noise} is some estimation of the noise standard deviation, N is the number of training instances, $\bar{y}$ and σ_y are the mean and standard deviation of training targets, respectively.

For adjusting the γ parameter of RBF kernel (6), the methodology proposed in [7] was used. It can be demonstrated that the generalization capacity of the SVR is a convex function depending on γ, so the optimal γ can be found by minimizing the generalization error with some optimization method for quasi convex functions.

E. Dissimilarity based classification

In order to design the dissimilarity based classifier, an initial representation set $\mathcal{R}$ of 200 signals (100 of each class) was extracted from the database. Then, the distances among all objects in the representation set were calculated, constructing the 200×200 dissimilarity matrix $D(\mathcal{R},\mathcal{R})$.

TABLE I

CONFUSION MATRIX WITH STFT REPRESENTATION

		Machine	
		With murmur	Normal
Label	With murmur	151	5
	Normal	5	150

TABLE II

CONFUSION MATRIX WITH CWD REPRESENTATION

		Machine	
		With murmur	Normal
Label	With murmur	159	4
	Normal	3	160

The linear programming method described in section II-B was then applied over the dissimilarity space, obtaining a final representation set $\mathcal{R}_{LP}$ of 91 prototypes for STFT y 76 prototypes for CWD. The remaining objects in each case were returned to the training sets for the classification stage.

IV. CLASSIFICATION RESULTS

Using the dissimilarity matrices $D(\mathcal{T}, \mathcal{R}_{LP})$, a $1 - nn$ classifier was trained and validated using the *leave one out* schema. The confusion matrices for each TFR case are shown in tables I and II:
It is remarkable that the number of samples in the STFT case was of 311, while for the CWD case was of 326 because the number of selected prototypes was different.

Table III shows the classification accuracy with specificity and sensitivity for each case.

TABLE III

ACCURACY PERCENTAGES

	Accuracy	Sensitivity	Specificity
STFT	96.78%	96.79%	96.77%
CWD	97.85%	97.55%	98.16%

Table III shows that with CWD representation, the classification results are slightly better that with STFT representation. However, for STFT 15 prototypes more than for CWD were selected, i. e. for each sample 15 more distances have to be computed. This two facts confirms that CWD representation provides a better way of extracting the time-frequency information for PCG signals.

V. CONCLUSIONS

The results showed that the proposed methodology, based on SVR and dissimilarity measures, is able to characterize successfully the TFRs. Moreover, it is posible to classify PCG signals (normals and murmurs) using the features extracted from the TFRs, obtaining remarkable results: 97.85% of classification success over the used database.

It is difficult to compare the results obtained in this work with respect to the results of previous works, because there is no available standard databases of PCG recordings in order to evaluate the performance of the algorithms. However, the results of this work showed better performance than the results obtained with other techniques based mainly in the application of neural networks to perform the classification with an accuracy between 85% and 95% [12], [13].

The results show that CWD representation provides a better way for extracting the time-frequency information from PCG signals. This had also been shown in [3] with a classification performance of 98%, using heuristic measures on the TFR. In our work, an automatic feature finding in the time-frequency plane was performed, with similar results and an improved methodology, which had been successfully applied to EEG, ECG and voice signals, but rarely used in PCG and heart sound signals.

Future work includes development of strategies to allow for physical interpretation of the relevant features (zones of the TFD with discriminant information). Moreover, it would be interesting to evaluate the effects of noise using signals with lower signal-to-noise ratio.

VI. ACKNOWLEDGMENTS

This work has been supported by *Jóvenes Investigadores* program financed by COLCIENCIAS, and the project *Detección de los niveles de compromiso de resonancia en niños con labio y/o paladar hendido* financed by DIMA.

REFERENCES

[1] M. Tavel and H. Katz, "Usefulness of a new sound spectral averaging technique to distinguish an innocent systolic murmur from that of aortic stenosis," *The American Journal of Cardiology*, vol. 95, no. 11, pp. 902–904, 2005.

[2] E. Sejdic and J. Jiang, "Comparative study of three time-frequency representations with applications to a novel correlation method," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 2, 17–21 May 2004, pp. ii–633–6.

[3] P. Bentley, J. McDonnell, and P. Grant, "Classification of native heart valve sounds using the choi-williams time-frequency distribution," in *Proc. IEEE 17th Annual Conference Engineering in Medicine and Biology Society*, vol. 2, 20–23 Sept. 1995, pp. 1083–1084.

[4] L. Atlas, L. Owsley, J. McLaughlin, and G. Bernard, "Automatic feature-finding for time-frequency distributions," in *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 18–21 June 1996, pp. 333–336.

[5] E. Delgado, J. Jaramillo, A. Quiceno, and G. Castellanos, "Parameter tuning associated with nonlinear dynamics techniques for the detection of cardiac murmurs by using genetic algorithms," in *Proc. Computers in Cardiology*, 2007.

[6] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[7] W. Wang, Z. Xu, W. Lu, and X. Zhang, "Determination of the spread parameter in the gaussian kernel for classification and regression," *Neurocomputing*, vol. 55, no. 3-4, pp. 643–663, 2003.

[8] E. Pękalska and R. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 943–956, 2002.

[9] B. Schoelkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: The MIT Press, 2002.

[10] D. Barry and D. Barry, "Fast calculation of the choi-williams time-frequency distribution," vol. 40, no. 2, pp. 450–455, 1992.

[11] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, pp. 113–126, Jan. 2004.

[12] C. Gupta, C. Gupta, R. Palaniappan, S. Rajan, S. Swaminathan, and S. Krishnan, "Segmentation and classification of heart sounds," in *Proc. Canadian Conference on Electrical and Computer Engineering*, R. Palaniappan, Ed., 2005, pp. 1674–1677.

[13] S. Strunic, S. Strunic, F. Rios-Gutierrez, R. Alba-Flores, G. Nordehn, and S. Burns, "Detection and classification of cardiac murmurs using segmentation techniques and artificial neural networks," in *Proc. IEEE CIDM*, F. Rios-Gutierrez, Ed., 2007, pp. 397–404.