Measurement 60 (2015) 121–128

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

A principal component-based algorithm for denoising in single channel data (PCA for denoising in single channel data)

Antonio Mauricio F.L. Miranda de Sá [a,*], José Manoel de Seixas [b], José Dilermando Costa Junior [a], Danton Diego Ferreira [c], Augusto S. Cerqueira [d]

[a] Biomedical Engineering Program – COPPE, Federal University of Rio de Janeiro, RJ, Brazil
[b] Signal Processing Lab – COPPE/POLI, Federal University of Rio de Janeiro, RJ, Brazil
[c] Engineering Department, Federal University of Lavras, MG, Brazil
[d] Electrical Engineering Program, Federal University of Juiz de Fora, MG, Brazil

ARTICLE INFO

Article history:
Received 11 July 2013
Received in revised form 25 September 2014
Accepted 29 September 2014
Available online 18 October 2014

Keywords:
Principal Component Analysis
Single channel denoising
Power quality
Electromyography
Electrocardiography

ABSTRACT

A denoising technique for single channel data is proposed. By assuming the observed signal to be the mixture of two unknown uncorrelated sources, an expression for the principal components (PC) of the set constituted by the signal and its *k*-sample delayed version is derived. The expression does not require matrix manipulations and may be hence useful when both speed and memory usage are crucial. The second PC was found to be a suitable estimate of one of the sources. Illustrations are provided for a simulated voltage signal corrupted by harmonics and transient disturbances as well as for a real electromyographic signal with electrocardiographic interference. A comparison with a standard, wavelet-based method for denoising is also provided.

1. Introduction

In signal processing applications, one might be concerned in obtaining sources that had been previously mixed. Examples of interest arise in many distinct areas, such as in biomedical, radar and communication engineering, as well as in oceanography, atmospheric science, speech and acoustics. In the so-called blind source separation techniques, it is assumed that the sources have been mixed and that no prior information about the mixing process is available [1]. An example of such is the Independent Component Analysis (ICA), which assumes the sources to be statistically independent and generally non-Gaussian [2]. In order to assess independence, ICA takes into account higher-order statistics. For Gaussian signals, the Principal Component Analysis (PCA), which takes into account only the first two statistical moments, is widely applied, particularly in data analysis and signal/image compaction applications [3]. For Gaussian signals, only the first two moments are different from zero, and hence PCA is a suitable technique for analyzing them.

Furthermore, since PCA is often used as a preprocessing stage in many ICA algorithms, its implementation generally results in faster algorithms in comparison with full ICA. In addition, although linear techniques are being increasingly less used in the last years, PCA may be used together with some other nonlinear technique to improve the performance of this latter. An interesting example may be found in [4], where PCA was applied to Empirical Mode Decomposition to isolate the cardiac information in the processing of cardiovascular signals.

* Corresponding author at: Biomedical Engineering Program – Federal University of Rio de Janeiro, P.O. Box 68510 – CEP: 21941-972, Brazil. Tel.: +55 21 2562 8630; fax: +55 21 2562 8591.

E-mail addresses: amfls@peb.ufrj.br (A.M.F.L. Miranda de Sá), seixas@lps.ufrj.br (J.M. de Seixas), dilermando@peb.ufrj.br (J.D. Costa Junior), danton@deg.ufla.br (D.D. Ferreira), augusto.santiago@ufjf.edu.br (A.S. Cerqueira).

http://dx.doi.org/10.1016/j.measurement.2014.09.079
0263-2241/

Therefore, the proposal of new techniques and algorithms based on PCA for dealing with source separation is justifiable, particularly in applications where both speed and memory usage are crucial, e.g. in online processing and embedded systems.

In most blind source separation applications, it is imposed that the number of signal observations should be equal to that of independent sources, which simplifies the modeling but represents a limitation in applications when only a single channel of a given random process is available. This situation arises, for instance, in the analysis of biomedical signals, such as the Electrocardiogram (ECG) or the surface Electromyogram (EMG) of a given muscle, as well as in Power Quality (PQ) applications, when one is interested in monitoring the power line that may be suffering from disturbances. Therefore, some techniques have been proposed so far for the so-called underdetermined mixtures (i.e. when the number of observations is smaller than the number of sources) and to the particular case when only one observation is available, for which the single channel ICA (SCICA) has been successfully applied [5]. Due to its efficiency in dealing with signals that exhibit time- or frequency-varying features, the wavelets have been considered a standard tool in many denoising applications [6]. However, the performance of such wavelet-based denoising techniques is somehow dependent on the choice of a suitable mother wavelet function and decomposition levels, which often requires some prior information on the signal shape [7]. This may hamper the use of wavelets in some applications where there is few or even no information about the signals and/or the mixing process.

In this paper, PCA is applied for noise reduction in single channel data, in a procedure similar to SCICA, i.e. by adding time-delayed versions of the signal to obtain a set of input signals that allows source extraction. An alternative expression for the principal component estimation is also derived for the particular case when the sources are uncorrelated and wide-sense stationary. Examples are provided with a simulated power signal corrupted by two simultaneous disturbances and with a surface EMG signal corrupted by ECG interference. A comparison with a standard wavelet based-method is also provided.

2. Technical considerations

Suppose that samples of a discrete random variable, $x_1[n]$, result from the summation of two zero-mean, uncorrelated sources, $s_1[n]$ and $s_2[n]$, arisen from wide-sense stationary processes. If $x_1[n]$ is the unique signal available, similarly to the approach commonly used in the SCICA, a k-delayed version of $x_1[n]$ (i.e. $x_2[n] = x_1[n-k]$) may be used as a second signal so that the set $\{x_1[n], x_2[n]\}$ may be used as input to a blind source separation technique. For such a case, the covariance matrix will be given as:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{x1}^2 & \sigma_{x1x2} \\ \sigma_{x2x1} & \sigma_{x2}^2 \end{bmatrix} = \begin{bmatrix} E\left[(s_1[n]+s_2[n])^2\right] & E[(s_1[n]+s_2[n])\,(s_1[n-k]+s_2[n-k])] \\ E[(s_1[n]+s_2[n])\,(s_1[n-k]+s_2[n-k])] & E\left[(s_1[n-k]+s_2[n-k])^2\right] \end{bmatrix} \tag{1}$$

where $E[\bullet]$ denotes expectation. Since $s_1[n]$ and $s_2[n]$ are assumed to be uncorrelated, all products in the non-principal diagonal in (1) will be zero, except for $E[s_j[n]s_j[n-k]]$ ($j = 1,2$). This fact together with the wide-sense stationarity of the sources (which turns the elements in the diagonal to be the same) lead to:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{s1}^2 + \sigma_{s2}^2 & R_{s1s1}[k] + R_{s2s2}[k] \\ R_{s1s1}[k] + R_{s2s2}[k] & \sigma_{s1}^2 + \sigma_{s2}^2 \end{bmatrix}, \tag{2}$$

where σ_{s1}^2 and σ_{s2}^2 are the variances of $s_1[n]$ and $s_2[n]$, respectively, and $R_{sjsj}[k]$ ($j = 1,2$) are the auto-correlation functions at lag k.

In order to obtain the eigenvectors and eigenvalues of $\mathbf{\Sigma}$, one must solve the equation:

$$(\mathbf{\Sigma} - \lambda \mathbf{I})\vec{v} = 0, \tag{3}$$

which is equivalent to solve $\det((\mathbf{\Sigma} - \lambda \mathbf{I})) = 0$ and leads to:

$$\lambda = \sigma_{s1}^2 + \sigma_{s2}^2 \pm (R_{s1s1}[k] + R_{s2s2}[k]). \tag{4}$$

It is interesting to notice that both values of λ are greater than zero, since $R_{sjsj}[k] \le \sigma_{sj}^2$. Such values reflect the energy of the principal components and by replacing them in (3), leads the eigenvectors to be found as:

$$\begin{aligned} \vec{v}_1 &= \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \\ \vec{v}_2 &= \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right). \end{aligned} \tag{5}$$

As it can be clearly seen, the eigenvectors in (5) do not depend on the elements of the covariance matrix in (2), which is unusual in PCA applications. However, this result reflects a property of Toeplitz symmetric matrices [8]. Therefore, the first principal component is proportional to the sum of $x_1[n]$ and $x_2[n]$, whilst the second principal component is proportional to the difference between such signals, i.e.:

$$\begin{aligned} y_1[n] &= \frac{1}{\sqrt{2}}(x_1[n] + x_2[n]) = \frac{1}{\sqrt{2}}(x_1[n] + x_1[n-k]) \\ y_2[n] &= \frac{1}{\sqrt{2}}(x_1[n] - x_2[n]) = \frac{1}{\sqrt{2}}(x_1[n] - x_1[n-k]) \end{aligned} \tag{6}$$

For real data applications, the covariance matrix is often not available and must be replaced by an estimate. However, assuming the mixing sources to be uncorrelated and samples of zero-mean, wide-sense stationary random processes, the principal components may still be estimated with expression (6).

In order to infer about the applicability of expression (6) even in cases where the assumptions made for deriving it may not fully apply, two examples are provided in the next section. In the first one, voltage signals corrupted by both transitory and harmonics were simulated, while in the second one, signals reflecting muscle contraction were

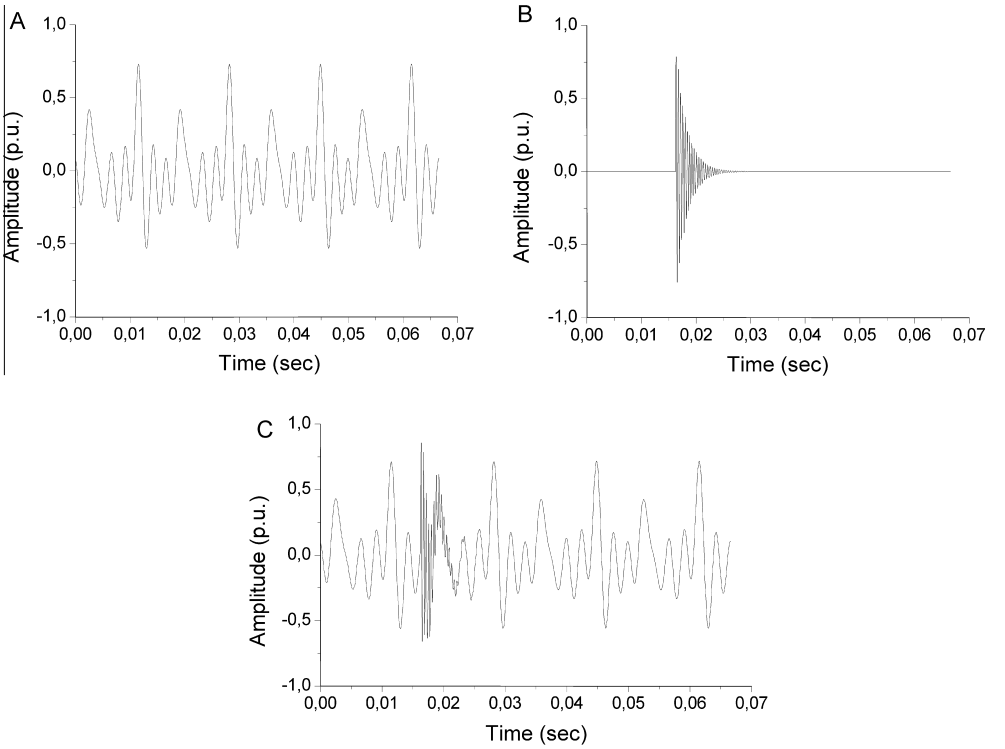

Fig. 1. Simulated multiple disturbance in power signals (1C): transient (1B) with harmonics (1A).

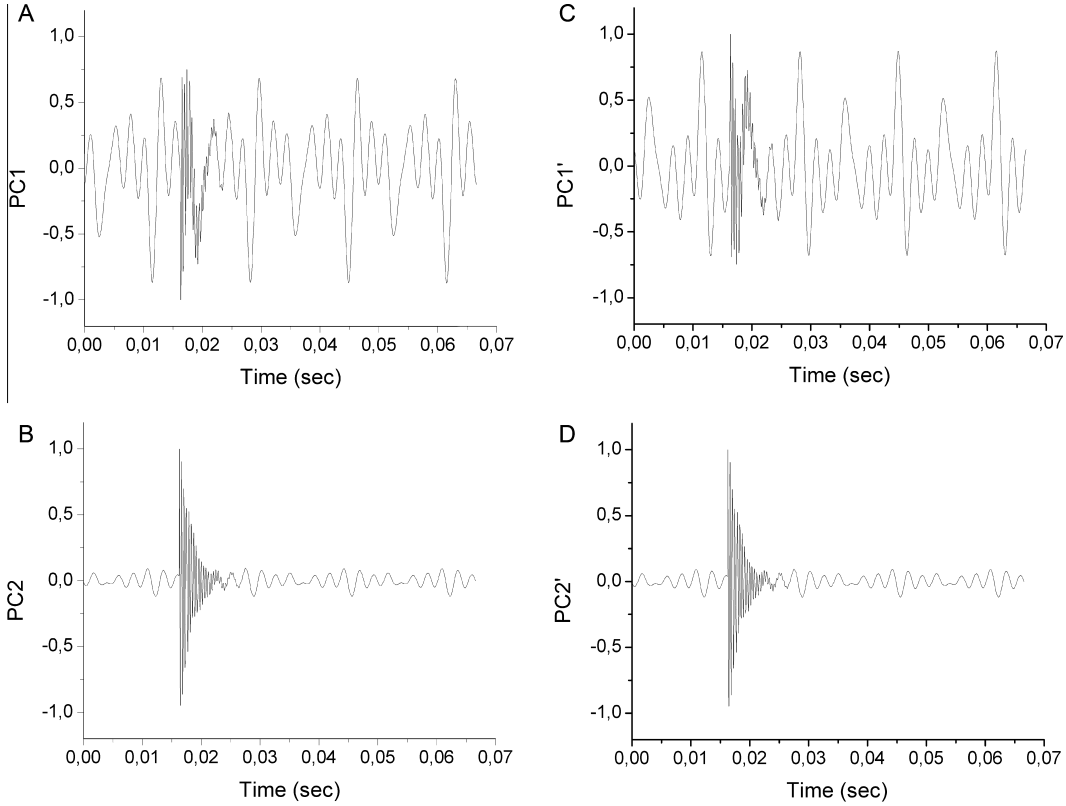

Fig. 2. Principal components obtained by both SVD PCA (A and B) and expression (6) (C and D) for the multiple disturbance signal shown in Fig. 1C.

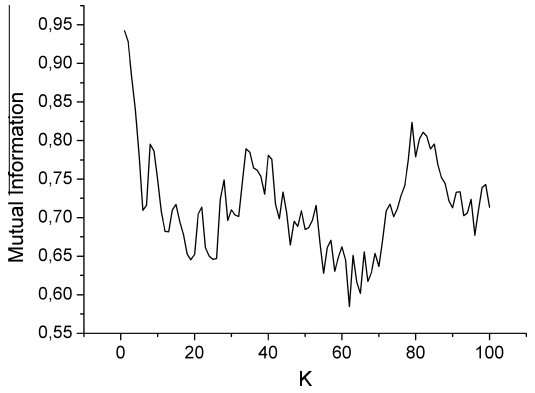

Fig. 3. Mutual information values between the second principal component extracted for the set constituted by the signal shown in Fig. 1C and a *k*-delayed version of it and the oscillatory transient component of the signal (Fig. 1B) for different values of *k*.

collected at the trunk, where interference of electrical activity from the heart is known to occur.

3. Results

3.1. Power system simulated signals

An interesting field of research in which Power Quality (PQ) concerns is in the analysis of voltage and current signals corrupted by multiple disturbances [9]. The correct characterization of PQ disturbances is an important requirement in order to find their causes and to take corrective actions. Recent works have applied sophisticated filtering methods, such as filter banks or wavelets, for the correct estimation and quantification of PQ [10,11], as well as to support PQ disturbance detection and classification methods as an important preprocessing tool [12-15]. The core idea of these works is to decompose multiple disturbances into single ones in order to facilitate their analysis. Despite their good performance, these filtering methods are considered to be computationally complex for implementation when a real-time monitoring is required.

The first step in PQ analysis may be the removal of the fundamental component (e.g. 60 Hz), and then, the remaining signal is decomposed into different frequency bands. Fig. 1C shows an example of a simulated multiple PQ disturbance, which comprises an oscillatory transient (Fig. 1B) and harmonics (Fig. 1A) (the fundamental component has already been removed).

The principal components (PCs) of the signal shown in Fig. 1C and its *k*-delayed version were computed according to expression (6) for distinct *k*-values and are shown in Fig. 2 for $k = 1$ (Fig. 2C and D). They were also obtained through single value decomposition (SVD), such as the MATLAB routine PRINCOMP applies (Fig. 2A and B). It can be noted that both methods led to very similar results. Actually, except for phase inversion in the first PC, both methods lead to the same results. This is an indication that theses disturbances satisfy the assumptions that led expression (1) to be simplified to (2). It can also be noted that the second PC exhibits much reduced interference from the harmonics, providing a suitable estimation of the oscillatory transient source shown in Fig. 1B.

Since in this simulated case the mixing sources are known, it is possible to evaluate the degree of similarity between the second PC and the transient source, as well as to determine the *k*-value that leads to the best estimation of this latter. This was carried out by estimating the mutual information (MI) between the signals based on the algorithm proposed in [16]. The normalization proposed in [17] was applied, resulting in MI-values ranging from 0 (total absence of mutual information) to 1 (maximum mutual information). It was decided to use MI rather than the linear correlation, because it provides a more

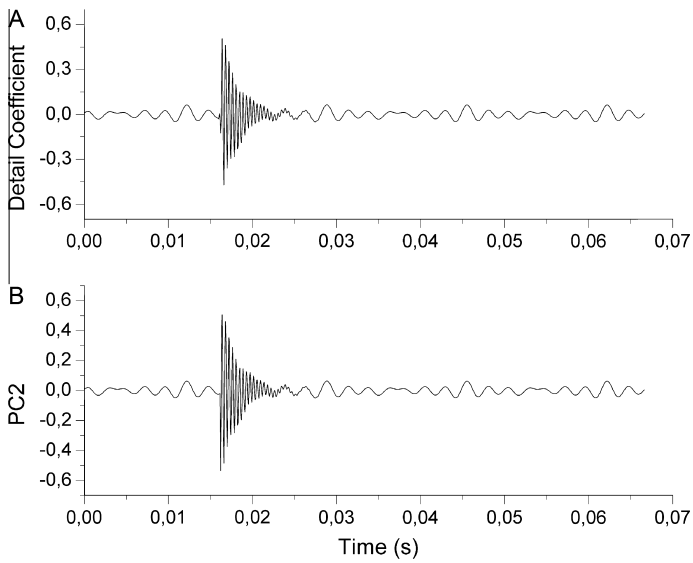

Fig. 4. Detail coefficient vector of the wavelet decomposition of signal in Fig. 1(A) and the second principal component obtained according to (6) for the observation set formed by this same signal and its one-shifted version (B).

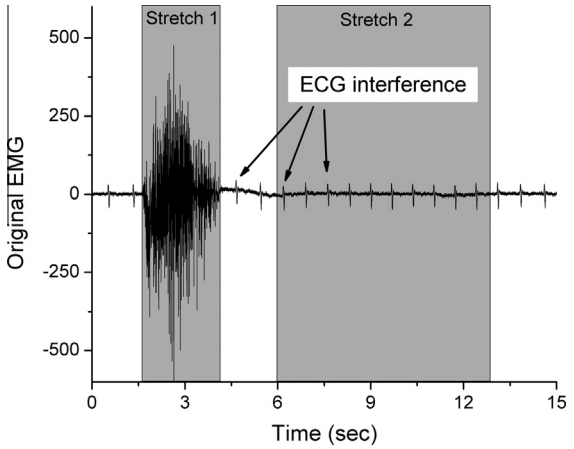

Fig. 5. EMG signal collected on lumbar erectors with ECG interference.

complete measure of dependence between both signals by taking into account higher-order statistics. The result is shown in Fig. 3, which depicts the normalized MI-values between the second PC and the transient signal in Fig. 1B for distinct k-values. It can be clearly noted that the higher MI-value (MI = 0.9423) occurs for $k = 1$ and hence the result shown in Fig. 2 with just a single sample shift [i.e. $k = 1$ in expression (6)] reflects the best estimation of the transitory source.

In order to provide a comparison with a standard method, a single-level one-dimension wavelet decomposition was performed in the simulated signal shown in Fig. 1C. It was decided to use the mother wavelet function Daubechies 5, due to the suitable results observed with it in comparison with other mother functions. Fig. 4A shows the detail coefficient vector of the wavelet decomposition (i.e. the parcel related to the high-frequency behavior of the original signal). The second PC obtained according to (6) with $k = 1$ is shown again in Fig. 4B to allow a comparison. As it can be clearly noted, both methods led to similar results, i.e. a suitable estimate of the oscillatory transient parcel of the simulated signal. The Mutual Information between the detail coefficient vector and the original oscillatory disturbance signal in Fig. 1B was found to be smaller than that for the second PC (MI = 0.9019). This indicates, in this particular case, that the PC-based method proposed exhibits a slightly superior performance. It is worth pointing out however that the wavelet-decomposition is affected by the mother wavelet function used and that a better performance could have been possibly found if extensive further investigations had been carried out to obtain a more suitable mother wavelet function.

3.2. Biomedical signals

The electromyogram (EMG) is the signal recorded through surface electrodes that results from the action potential summation occurring in motor units in the muscle tissues [18]. The electrocardiogram (ECG), which reflects the electrical activity from the heart, may interfere with the EMG signals when these latter are recorded on the trunk muscles [18,19]. The presence of ECG in EMG signals

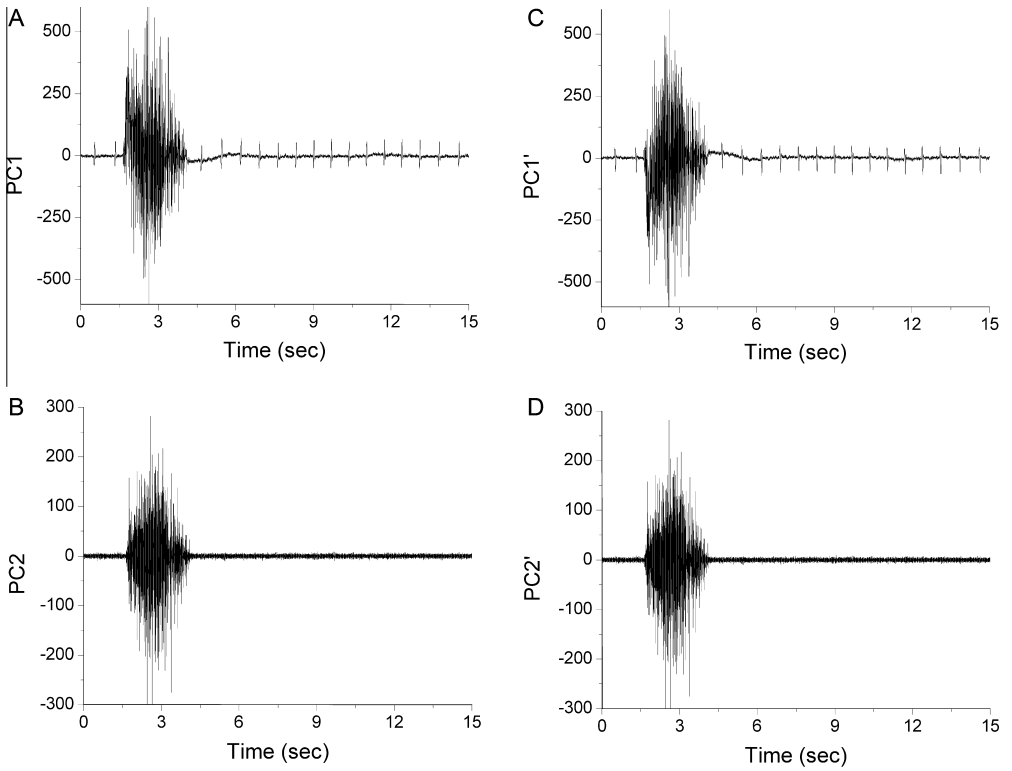

Fig. 6. Principal components obtained from both expression (6) (A and B) and SVD (C and D) and for the EMG signal with ECG interference shown in Fig. 5 and $k = 1$.

influences the amplitude and frequency measures of the true EMG and the removal of such interference is then a necessary task for the correct evaluation of the EMG.

In order to illustrate the proposed technique, an experimental EMG signal was collected from one subject at lumbar erector muscles, where ECG interference is most likely to occur. This work was submitted and approved by the Ethics Committee Research and the subject signed a consent form. A stretch of such signal is shown in Fig. 5, where one can clearly notice the ECG interference.

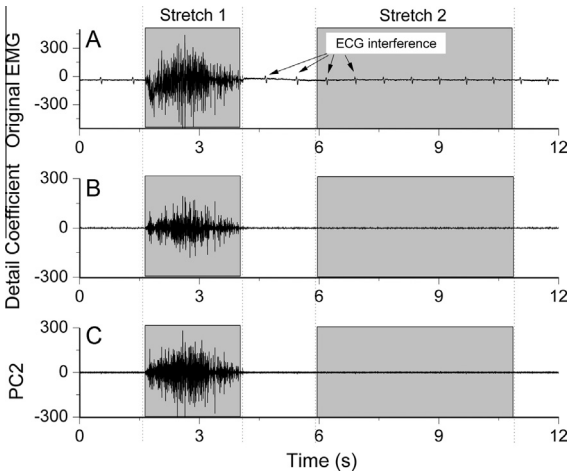

Fig. 7. Original EMG signal with the indication of the stretches used in the spectral analysis (A). Detail coefficient vector of the wavelet decomposition of this original signal (B) and the second principal component obtained according to (6) for set obtained by this same signal and a one-shifted version of it (C).

The principal components (PCs) of the set including this signal and its k-sample-shifted version were computed according to (6) and from SVD estimation, leading to virtually identical results, except for phase inversion in the first PCs (see Fig. 6, for $k = 1$). Furthermore, the first PC is very similar to the acquired EMG signal. The second PC, however, exhibits much reduced ECG interference, providing a suitable estimation for the EMG source. As in the previous example with simulated PQ signals, a single sample shift in the corrupted EMG signal led to greatest reduce of the ECG activity in the second PC.

The same wavelet decomposition performed in the simulated PQ data was carried out for the experimental EMG signal shown in Fig. 5. The result is shown in Fig. 7B together with the second PC obtained according to (6) for $k = 1$ (Fig. 7C). As it can be seen, both methods led to reduction of the electrocardiographic interference. Since the EMG signal free of ECG interference is not available in this case, in order to quantify such reduction and hence to evaluate the performance of both methods, a spectral analysis was carried out. In order to do so, two distinct stretches of the signal were considered. The first one (stretch 1) is the moment during which the maximum voluntary isometric contraction (MVIC) occurs and the second one (stretch 2) reflects a period at rest (i.e. with no muscle contraction).

Thus, the spectrum of both stretches was estimated (Welch method of averaging periodograms with a 64-point Hanning window with 50% overlap) for the original signal as well as for the respective stretches in the signals resulting from the denoising techniques. The result is shown in Fig. 8. For the period during MVIC (Fig. 8A), where the signal-to-interference ratio is higher and the ECG interference is less evident, the proposed method based on expression (6) led to a more pronounced reduction in the spectrum within the frequency band from 0 to 25 Hz, where ECG

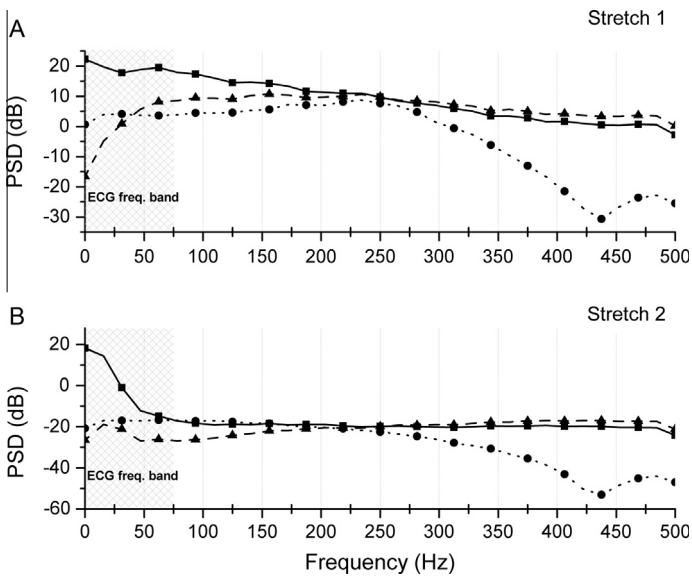

Fig. 8. Power spectrum density (PSD) estimation in decibels for stretches 1(A) and 2(B) of the signals shown in Fig. 7. The acquired EMG signal spectra are plot in continuous lines with squares, the detail coefficient of the wavelet decomposition spectra are plot in dotted lines with circles and the second principal component spectra are plot in dashed lines with triangles.

activity is known to occur. The attenuation due to wavelet decomposition is smaller within this range but it becomes much more pronounced as the frequency increases above 250 Hz, where the interference of the ECG in the EMG signal has probably vanished, since the ECG is usually assumed to be band-limited between 1 and 75 Hz (frequency range indicated in the hachured area in the figure). It is worth to point out that the signal obtained with the proposed denoising method based on the PCA tends to have spectra-values closer to the original signal as the frequency increases, which indicates that this method does not seem to alter the parcel due to the voluntary muscle contraction. This same pattern is observed in the signal at rest (Fig. 8B), but with both techniques leading to a more pronounced attenuation in the low frequency range, for which the ECG interference level is much higher in comparison with the signal during MVIC (stretch 1). These results indicate that, for the EMG with ECG interference, the proposed method outperforms the wavelet-based method for denoising and original source estimation, at least for the wavelet mother function and decomposition levels used.

4. Conclusion

In this paper, a new technique for noise reduction in single channel signals is proposed. It assumes that the investigated signal is a result from the mixture of two unknown, uncorrelated zero-mean sources. Based on this assumption, an alternative, simple expression of the principal components of the two-set built with the signal and a k-delayed version of it was derived. The second principal component was found to be an estimate of one of the mixing sources.

Illustrations of the technique were provided for a simulated power signal corrupted by harmonic and transient disturbances and the signal from surface electromyography that exhibited interference from the electrocardiogram. These simulated and experimental results indicate that in conditions where both stationarity and uncorrelation of the mixing sources are suspected to occur the PCA may be efficient for denoising single channel data.

The technique can be easily applied, since the sources are obtained based on the sum and subtraction of the single-channel signal and a k-sample delayed version of it. This technique results in a very fast implementation in comparison with standard PCA algorithms. The best performance was observed for $k = 1$. In this case, in the second principal component, one of the mixing sources is strongly attenuated. This can be intuited because such expression reflects a high-pass filter action and if $x_1[k] = s_1[n] + s_2[n]$ (with $s_2[n]$ band limited to low frequencies), then the parcel related to $s_2[n]$ will almost vanish in term $(x_1[n] - x_1[n-k])$, provided $R_{s2s2}[k]$ has not an impulse shape, which certainly applies when $s_2[n]$ is band-limited and $k = 1$. This explains the best performance for such unit lag. Therefore, unknown embedded sources in a signal might be attenuated by using expression (6). Due to the high-pass behavior of such expression, the attenuation occurs if the spectrum of one source does not overlap to the other one. Additionally, the shorter the lag, the greater the chance of stationarity assumption to be met.

The comparison with a standard method based on wavelet decomposition shows that the proposed technique is at least as suitable as this latter for attenuating the interference signal in the simulated voltage signal corrupted by harmonics and transient disturbances. However, the wavelet decomposition is affected by the chosen wavelet mother function and decomposition levels. In this particular case of known mixing sources, a suitable choice could have been made in order to ensure good results, but this cannot be guaranteed in real applications where the shape of the mixing sources is unknown.

Moreover, the PCA-based method proposed in the present work has a very simple implementation. Wavelet-based methods for denoising may eventually have excellent performance, but this may be some times achieved at the expense of very complex mother wavelet functions and hence of a correspondingly complex, time consuming implementation. Therefore, the much simpler computational implementation of the proposed method, together with the suitable performance of it in both simulated and real data, reinforces its use in single channel denoising applications where both speed and memory usage are crucial.

Acknowledgment

This work received financial support from the Brazilian Agencies CAPES, CNPq, FAPERJ and FAPEMIG.

References

[1] P. Comon, C. Jutten, Handbook of Blind Source Separation – Independent Component Analysis and Applications, Elsevier, Amsterdam, 2010.
[2] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[3] I.T. Jolliffe, Principal Component Analysis, second ed., Springer-Verlag, New York, 2002.
[4] E. Pinheiro, O. Postolache, P. Girão, Empirical mode decomposition and principal component analysis implementation in processing non-invasive cardiovascular signals, Measurement 45 (2012) 175–181.
[5] M.E. Davies, C.J. James, Source separation using single channel ICA, Signal Process. 87 (2007) 1819–1832.
[6] D.L. Donoho, De-noising by soft-thresholding, IEEE Trans. Inform. Theory 41 (1995) 613–627.
[7] M. Vetterli, J. Kovacevic, Wavelets and Subband Coding, Prentice-Hall, New Jersey, 1995.
[8] F. Bünger, Inverses, determinants, eigenvalues, and eigenvectors of real symmetric Toeplitz matrices with linearly increasing entries, Linear Algebra Appl. 459 (2014) 595–619.
[9] M.H.J. Bollen, P.F. Ribeiro, I.Y.H. Gu, et al., Trends, challenges and opportunities in power quality research, EUR., Trans. Electr. Power 20 (2009) 3–18.
[10] C.A. Naik, P. Kundu, Power quality index based on discrete wavelet transform, Int. J. Electr. Power Energy Syst. 53 (2013) 994–1002.
[11] Y.-J. Shin, E.J. Power, M. Grady, A. Arapostathis, Power quality indices for transient disturbances, IEEE Trans. Power Deliv. 21 (1) (2006) 253–261.
[12] D.D. Ferreira, A.S. Cerqueira, C.A. Duque, J.M. de Seixas, M.V. Ribeiro, Automatic system for classification of isolated and multiple disturbances in electric signals, SBA Control Automat. 22 (2011) 39–48.
[13] W.-M. Lin, C.-H. Wu, C.-H. Lin, F.-S. Cheng, Detection and classification of multiple power-quality disturbances with wavelet multiclass svm, IEEE Trans. Power Deliv. 23 (2008) 2575–2582.

[14] Z. Liu, Q. Zhang, Z. Han, G. Chen, A new classification method for transient power quality combining spectral kurtosis with neural network, Neurocomputing 125 (2014) 95–101.
[15] M.V. Ribeiro, J.L.R. Pereira, Classification of single and multiple disturbances in electric signals, EURASIP J. Adv. Signal Process. (2007).
[16] Z.I. Botev, J.F. Grotowski, D.P. Kroese, Kernel density estimation via diffusion, Ann. Stat. 38 (2010) 2916–2957.
[17] H. Joe, Relative entropy measures of multivariate dependence, J. Am. Stat. Assoc. 84 (2000) 157–164.
[18] G.H. Lu, J.S. Brittain, P. Holland, J. Yianni, A.L. Green, J.F. Stein, T.Z. Aziz, S.Y. Wang, Removing ECG noise from surface EMG signals using adaptive filtering, Neurosci. Lett. 462 (2009) 14–19.
[19] H.L. Butler, R. Newell, C.L. Hubley-Kozey, J.W. Kozey, The interpretation of abdominal wall muscle recruitment strategies change when the electrocardiogram (ECG) is removed from the electromyogram (EMG), J. Electromyogr. Kinesiol. 19 (2009) e102–e113.