Digital Signal Processing 77 (2018) 63–76

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Improving reconstruction of time-series based in Singular Spectrum Analysis: A segmentation approach

M.C.R. Leles [a,b,*], J.P.H. Sansão [a,b], L.A. Mozelli [a,c], H.N. Guimarães [d]

[a] CELTA – Center for Studies in Electronics Engineering and Automation, UFSJ – Federal University of São João del-Rei, Rod. MG 443 km 7, 36420-000, Ouro Branco, MG, Brazil
[b] PPGEE – Graduate Program in Electrical Engineering, UFMG – Federal University of Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
[c] Department of Electronics Engineering – UFMG, Brazil
[d] Department of Electrical Engineering – UFMG, Brazil

ARTICLE INFO

Article history:
Available online 21 November 2017

Keywords:
Singular spectrum analysis
Non-stationary signals
Segmentation

ABSTRACT

Singular Spectrum Analysis (SSA) is a powerful non-parametric framework to analysis and enhancement of time-series. SSA may be capable of decomposing a time-series into its meaningful components: trends, oscillations and noise. However, if the signal under analysis is non-stationary, with its spectrum spreading and varying in time, the reliability of the reconstruction is guaranteed only when many elementary matrices are used. As a consequence, the capability to discriminate dominant structures from time-series may be impaired. To circumvent this issue, a new method, called overlap-SSA (ov-SSA), is proposed for segmentation, analysis and reconstruction of long-term and/or non-stationary signals. The raw time series is divided into smaller, consecutive and overlapping segments, and standard SSA procedures are applied to each segment with the resulting series being concatenated. This variation of SSA seeks to: improve reconstruction and component separability for non-stationary time-series; enable the analysis for large datasets, avoiding the issues of concatenation of many segments; and present some benefits of the segmentation in terms of better time–frequency characterization. These advantages are illustrated in several synthetic and experimental datasets.

1. Introduction

Singular Spectrum Analysis (SSA) is a non-parametric approach that can be used to decompose the original time-series into additive structures, promoting an efficient way to separate meaningful components, such as trends or oscillations, whilst discarding noise. It has been used into a plethora of applications and fields: time-series reconstruction [17], frequency estimation [27,16], trend extraction [29], time-series forecasting [18], trends and harmonic extraction [30], and many others [15].

The technique consist in four major steps. The *Immersion*, which maps the raw time-series into a Hankel matrix, called trajectory matrix. Parameter L, the embedding dimension, is used to assemble the trajectory matrix. The *Singular Value Decomposition* (SVD), which factorizes the trajectory matrix into a set of L elementary matrices, $A_1, A_2, \ldots, A_L$. In *Grouping*, occurs the combination of a subset of elementary matrices, $A_{\{I\}} = \sum_{i \subset I} A_i$, that capture the desired structures. Finally, *Diagonal Averaging* transforms the resultant matrix from grouping into the reconstructed time-series.

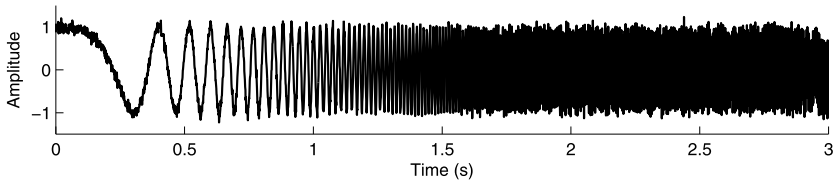
Classical implementations of SSA suffer from some drawbacks, one of which will be deepened in the sequence. Consider a synthetic signal depicted in Fig. 1(a), which is an exponential chirp signal[1] with addition of noise (SNR = 20 dB).

If a classical SSA approach is used, whose mathematical details will be provided latter in Section 3.1, one may attempt to capture the main features of this signal on decomposition stage. Lets consider two groupings, one resorting to the first 10 components and the other considering the first 20 components, and their respective reconstructed time-series, portrayed in Fig. 1. It is noticeable, as time increases, that reconstructed time-series are less trustworthy. Likewise, as more components are introduced into grouping stage, the reconstructed time-series becomes more reliable. Therefore, the only way to achieve a reasonable reconstruction, in this

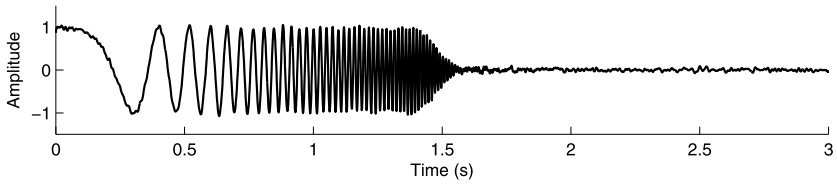
* Corresponding author at: CELTA – Center for Studies in Electronics Engineering and Automation, UFSJ – Federal University of São João del-Rei, Rod. MG 443 km 7, 36420-000, Ouro Branco, MG, Brazil.
E-mail addresses: mleles@ufsj.edu.br (M.C.R. Leles), joao@ufsj.edu.br (J.P.H. Sansão), mozelli@cpdee.ufmg.br (L.A. Mozelli), hnguimar@gmail.com (H.N. Guimarães).

https://doi.org/10.1016/j.dsp.2017.10.025
1051-2004/

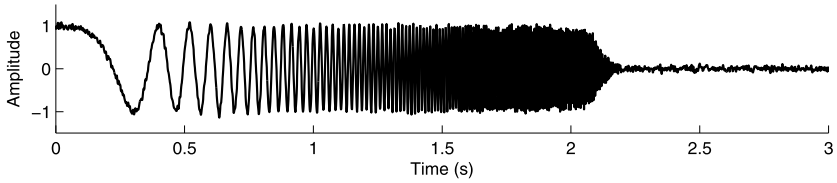
[1] To generate this signal *Matlab*® command `chirp(t,0.5,2,150, 'q');` is used.



(a) Cosine signal with variable frequency.



(b) Reconstructed time-series using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,10\}}$.



(c) Reconstructed time-series using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,20\}}$.

Fig. 1. Original time-series and reconstructed ones by SSA using $L = 60$ and two groupings.

example, is by resorting to a large number of components. However, as the number of components increases, it becomes more difficult to distinguish dominant structures from each other.

This example shows a shortcoming of classical SSA: non-stationary signals. The concept of "non-stationarity" is often defined in different ways. In this paper, non-stationarity is invoked to refer to deterministic signals whose spectral properties are time-variant [33, Chapter 11].

Other drawback appears when applying SSA to large sequential datasets, typically associated with very long time-series, arising from telecommunications, finance, bioinformatics or web mining. Usage of SSA for the entire time-series may lead to cumbersome issues, due to computation of SVD of a high dimensional matrix, when basic (off-the-shelf) numeric packages are adopted. Alternative implementations were proposed by Korobeynikov [26] which are computationally more efficient and may reduce this impact. For the Multivariate Singular Spectral Analysis (M-SSA) a solution for this drawback was proposed by Pukenas [35].

An efficient approach to handle at once both these issues, non-stationarity and large data, can be segmenting. Yiou et al. [44] proposed a local approach of SSA, called MS-SSA (Multi-Scale SSA). It seeks to perform a joint time–frequency analysis, analogously to wavelet transform. However this capability is obtained at the cost of great computational effort. Rekapalli and Tiwari [36] proposed another modification of SSA algorithm called Windowed SSA (WSSA). The idea is to perform segmentation of the dataset and compute SSA for each segment. However, there is no treatment on how to join the segments properly.

In this paper, a new algorithm based on the SSA is proposed. Like many others variations of the standard SSA, as discussed latter in this paper, the objective is to improve the application of SSA to some common situations that may occur in various fields of signal processing. The algorithm is not intended to improve the SSA for every applications but is intended to be useful for:

- improve reconstruction and component separability for non-stationary time-series, like the aforementioned example;
- improve the analysis for large datasets, avoiding the issues of concatenation of many segments;
- discuss and present some advantages of the segmented analysis in terms of better time–frequency characterization.

The remaining of this paper is organized as follows: Section 2 provides the mathematical framework concerning standard SSA analysis; Section 3 epitomizes the main features about the proposed method[2]; Section 4 discuss several examples to illustrate the benefits provided by the ov-SSA, considering synthetic and experimental datasets; Section 5 presents some remarks about the proposed method; and finally, Section 6 lays down the conclusion and future perspectives.

2. SSA

In this section, a brief description of SSA methodology is given, according to Golyandina et al. [11].

2.1. Embedding

Time-series $\mathbf{x} = (x_0, x_1, \cdots, x_n, \cdots, x_{N-1})^T$, with length N, represents the signal under analysis. The mapping of this signal into a matrix $\mathbf{A}$, of dimension $L \times K$, assuming $L \leq K$, is called *immersion*, and can be defined as:

$$\mathbf{A} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{K-1} \\ x_1 & x_2 & \cdots & x_K \\ \vdots & \vdots & & \vdots \\ x_{L-1} & x_L & \cdots & x_{N-1} \end{bmatrix}, \tag{1}$$

where L is the window length, or embedding dimension, and $K = N - L + 1$. $\mathbf{A}$, is a *Hankel* matrix, called the trajectory matrix [11].

[2] Complexity analysis and computational details of the proposed algorithm are provided in the unpublished paper by Leles et al. [28], which is companion for this Joint Special Issue.

2.2. Singular value decomposition

Factorization of the trajectory matrix $\mathbf{A}$, using Singular Value Decomposition (SVD), yields to:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^T, \tag{2}$$

where $R = \text{rank}(\mathbf{A}) \leq L$. Matrices $\mathbf{U}$ and $\mathbf{V}$ form an orthonormal system, so that $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ for $i \neq j$ and $\| \mathbf{u}_r \| = 1$. $\mathbf{\Sigma}$ is a diagonal matrix $\in \mathbb{R}^{L\times K}$, whose diagonal elements $\{\sigma_r\}$ are the singular values of $\mathbf{A}$. The eigenvalues of $\mathbf{A}\mathbf{A}^T$ are given by $\lambda_r = \sigma_r^2$.

The principal components can be obtained by [41]:

$$\mathbf{w}_r = \sigma_r \mathbf{v}_r = \mathbf{A}^T \mathbf{u}_r, \tag{3}$$

where $\mathbf{w}_r$ is a $K \times 1$ vector. The $w_k^{(r)}$ elements of $\mathbf{w}_r$ are given by

$$w_k^{(r)} = \sum_{l=0}^{L-1} u_l^{(r)} x_{l+k}, \quad k = 0, 1, \cdots, K-1, \tag{4}$$

where $u_l^{(r)}$ is $(l+1)$-th element of eigenvector $\mathbf{u}_r$ and $w_k^{(r)}$ is $(k+1)$-th element of the vector $\mathbf{w}_r$.

The r-th elementary matrix, $\mathbf{A}_r$, an unitary rank $L \times K$ matrix, can be written as

$$\mathbf{A}_r = \sigma_r \mathbf{u}_r \mathbf{v}_r^T = \mathbf{u}_r \mathbf{w}_r^T. \tag{5}$$

2.3. Grouping

The grouping step is the procedure of arranging the r indices $\{1, 2, \cdots, R\}$ into M disjoint subsets I_m, with $m = \{1, 2, \cdots, M\}$, and $I_m \subset \{1, 2, \cdots, R\}$. For a set of elementary matrices $\{\mathbf{A}_r \mid r \in I_m\}$, the resulting matrix from this grouping is:

$$\mathbf{A}_{\{I_m\}} = \sum_{r\in I_m} \mathbf{A}_r = \mathbf{A}_{\{.\}}, \tag{6}$$

where $\{.\}$ designated the indices of I_m set.

So, the trajectory matrix can be rewritten as a sum of M resultant grouping matrices:

$$\mathbf{A} = \sum_{m=1}^{M} \mathbf{A}_{\{I_m\}}. \tag{7}$$

Each group is intended to represent an additive component of the original signal, like a trend, a oscillatory component or noise.

2.4. Diagonal averaging

The purpose of this step is to recover a time series, $\widetilde{\mathbf{x}}_r$, of length N, from a elementary matrix $\mathbf{A}_r$, of dimension $L \times K$. One can calculate the diagonal averaging in N antidiagonals of $\mathbf{A}_r$, according to Vautard et al. [41]:

$$\widetilde{\mathbf{x}}_n^{(r)} = \begin{cases} \frac{1}{n+1} \sum_{i=0}^{n} u_i^{(r)} w_{n-i}^{(r)} & \text{for } 0 \leq n < L-1, \\ \frac{1}{L} \sum_{i=0}^{L-1} u_i^{(r)} w_{n-i}^{(r)} & \text{for } L-1 \leq n < K, \\ \frac{1}{N-n} \sum_{i=n-K+1}^{N-K} u_i^{(r)} w_{n-i}^{(r)} & \text{for } K \leq n < N. \end{cases} \tag{8}$$

Finally, this step can be easily extended to any matrix resulting from the grouping process [11].

2.5. Some guidelines in SSA parameters selection

Separability and signal extraction are main concepts in SSA. The former is related to the capacity to distinguish components whereas the latter concerns the possibility of extracting a desired characteristic from the original time-series. Therefore, SSA decomposition can only be successful if the resulting additive components of the series are approximately separable from each other. The step of parameter selection is driven by these objectives besides the nature of the application under analysis.

2.5.1. SSA parameters selection

Since the SSA is a non-parametric, or a model-free, method is does not make any assumptions about the process that generated the observed time-series. Nevertheless, as in many signal processing tools, a preliminary analysis of the time-series can be beneficial to enhance the performance and aid in the interpretation of the results, once reconstruction depends on the choice of the components from the covariance decomposition of the trajectory matrix. In many applications, this selection is based on the previous knowledge and experience of the researcher and on a variety of practical rules [34].

Four distinct methods can be pointed to guide the choices[3] of L and I_m [38]: i) the analysis of the singular values spectra; ii) the investigation of the pairwise eigenvector scatterplots; iii) the analysis of the periodogram of the original series (and eigenvectors); and iv) the analysis of the weighted correlation matrix, which is a usual metric of separability.[4]

In the sequel, methods that are less dependent on the research intervention are briefly presented.

2.5.2. Embedding dimension

The embedding dimension L is an essential SSA parameter. In general, large values of L provide a more refined decomposition into elementary components, resulting in better separability.

General guidelines were proposed in SSA Literature regarding the choice of L if no previous information were taken into account. Hassani et al. [21] propose the median of $1, 2, \ldots, N$ as the selection of the window length. Wang et al. [42] introduced the SSA technique into the field of blind source separation and a method for selecting an optimal window length was proposed. Khan and Poskitt [25] suggest a criteria that tends to favor the selection of $L \ll N/2$. On the other hand, Golyandina [10] recommends choosing $L \sim N/2$. This difference is justified by distinct objectives: Golyandina [10] prioritizes separability whereas Khan and Poskitt [25] favor signal extraction.

It is relevant to point out that for the same time-series different values can be attributed, driven by the objective. For instance, consider the Death Series,[5] usually adopted as example. Focusing in separability and interpretation of the obtained components [17] suggests $L = 24$. However, when the goal is signal extraction and forecasting Hassani et al. [18] proposed $L = 14$. Wang et al. [43] indicate the close value $L = 16$, based in blind source separation (BSS) technique.

Finally, as pointed out by Golyandina and Zhigljavsky [15] "*There is no universal rule for the selection of the window length. However, there are several general principles for the selection of the window length L that have certain theoretical and practical grounds*".

[3] For a detailed discussion about this topic, see for instance, Hossein [17].

[4] If w-correlation is small then the corresponding series are almost w-orthogonal. For a detailed discussion, see [11, Section 1.5].

[5] Monthly accidental deaths in the USA between 1973 and 1978.

Table 1
Summarizing the parameters selection in SSA methodology.

Embedding dimension	Grouping	Automatic
Priori information[a] ($L = kT$)	eigenvalue spectra[b]	SSD[c]
Separability[d] ($L \sim N/2$)	eigenvector periodogram[e]	PSO[f]
Signal extraction[g] ($L << N/2$)	w-correlation[h]	EXSSA[i]
Automatic approaches[j]	grouping size[k]	SSA-CT[l]

Notes: [a] for instance, extraction of a periodic component with period T where k is an integer; [b] or other heuristic method Hassani [17]; [c] Bonizzi et al. [7]; [d] general recommendation Golyandina et al. [11]; [e] oscillations [3] and trends [2]; [f] Abdollahzade et al. [1]; [g] general recommendation Golyandina et al. [11]; [h] Alonso and Salgado [5]; [i] Papailias and Thomakos [34]; [j] Wang et al. [42], Khan and Poskitt [25], Hassani et al. [21], Golyandina [10]; [k] Alharbi and Hassani [4]; [l] Hassani et al. [18].

2.5.3. Grouping

Another issue is related to the choice of elementary matrices used for the purpose of reconstruction, on the grouping step.

Guidelines for choosing the grouping size, in general applications, can be found in Alharbi and Hassani [4]. In this context, there is a binary approach based in an assumption that the series is formed by a signal plus noise. Since the signal components are often dominating, the only parameter of grouping is the number of the leading components, i.e., the grouping size. On the other hand, Hassani et al. [18] indicate that a binary separation may not lead to optimal results. They consider an approach that analyzes all eigenvalues and select the ones that carry more information.

Alexandrov [2] and Alexandrov and Golyandina [3] propose methods to automatic grouping selection based on periodogram of eigenvectors for trend and harmonic components extraction, respectively. Alonso and Salgado [5] applied cluster techniques for grouping the elementary components based on k-means. A method in which the eigentriples are adaptively selected using a delayed version of the data is proposed in Sanei et al. [39].

2.5.4. Automatic selection

Initiatives for the automatic choice of SSA parameters have been developed in the recent years. Bonizzi et al. [7] proposed the Singular Spectrum Decomposition (SSD), a fully automatic method for parameter selection focusing on the frequency spectra. Abdollahzade et al. [1] applied a Particle Swarm Optimization (PSO) algorithm for fine tuning of SSA parameters. Papailias and Thomakos [34] used all relevant components of the covariance decomposition via exponential smoothing of the covariance eigenvalues introducing an "automated" SSA reconstruction procedure. Hassani et al. [19] also rely on an optimization method, using Colonial Theory. For each L, the algorithm searches the optimal R eigenvalues for the cost function given by RMSE (root mean squared error) of reconstruction. Every L is investigated.

2.5.5. Summarizing the SSA parameter selection

Table 1 summarizes some concepts discussed. For a detailed discussion, see references therein.

2.5.6. Different versions of SSA

SSA can solve a plethora of problems in a great variety of areas of knowledge [11, Section 1.3]: smoothing, filtration, noise reduction, extraction of trends, extraction of seasonality components, extraction of cycles with small and large periods, extraction of periodicities in the form of modulated harmonics, finding structure in short time series, analysis of data sets that are unevenly sampled or contain missing data, besides several other applications.

However, in specific areas or applications, variations of the original SSA can produce better results. In the following some examples are presented.

Golyandina and Shlemov [13] suggested two modifications of SSA (Oblique SSA and SSA with derivatives), which can considerably improve the separability and thereby the reconstruction accuracy. Kalantari et al. [24] proposed a variant of SSA which provides better reconstruction and forecasts results when outliers are presented in the analyzed time-series. Hassani et al. [20] applied the multivariate SSA for forecasting some of the major UK industrial production indexes. Golyandina and Shlemov [14] proposed a combination of SSA with a subspace-based parametric approach, which provides better results for extraction of polynomial (especially, linear) trends. The hurdle of mixed time series components may sometimes be overcome by the use of the Sequential SSA [15, Section 2.5.5]. Moskvina and Zhigljavsky [32] proposed a SSA based algorithm for change-point detection and subspace tracking. Rodrıguez-Aragón and Zhigljavsky [37] proposed a variation of SSA for image processing. Two approaches to applying SSA to time series with missing data were proposed by Schoellhamer [40] and Golyandina and Osipov [12].

Among the aforementioned variations, a common goal is to improve a shortcoming of the original SSA. In the following Sections a new variations is proposed and several examples are provided to illustrated its advantages.

3. A new variation of SSA: the overlap SSA (ov-SSA)

SSA stands out among other methods for some reasons. For instance, is more prone to success for short time-series [41] and it is adaptive to the underlaying data [23]. The idea in this paper is to improve SSA by exploring both these features, which may be achieved by resorting to segmentation.

Obviously, the procedure of segmentation produces smaller time-series, which are favorable to SSA. Secondly, if a time-series is non-stationary, transient oscillations can be evidenced locally in a given segment, whereas they could go unnoticed in the general computation of the whole time-series. Because the SSA is adaptive, it has the flexibility to detect such pattern changes whilst retains information that is common to more segments.

This is the key motivation in the proposed methodology. The original time-series, with length N, is divided into smaller and consecutive segments, of fixed length Z. For each segment, standard SSA algorithm is computed and a local time-series is reconstructed.

The innovative aspect of the proposed paper consists in the procedure to joint the consecutive segments. Since the SSA suffers from boundary effects, the extreme points in the left and right edges are not meaningful for the purpose of reconstruction, so only an inner subset of samples q is considered meaningful to represent the local time-series.

To solve this issue, a modification of the overlap-save method is introduced. Consecutive segments overlap each other and the initial and final samples of a given segment are discarded. This allows to concatenate samples that have been preserved, generating a reconstructed time series of the same size as the original series.

This can be better understood with the aid of Fig. 2. Consider a local segment, with length Z. All samples within this segment are used to computed the SSA but only q samples are considered meaningful (lighter grey tones), resulting in $\bar{L} = (Z - q)/2$ discarded samples (dark grey tone). Notice that the local segments Z used for analysis overlap each other, by the percentage $\beta = 100(Z - q)/Z$. However, the local segments q that are reconstructed do not overlap and can be easily concatenated in sequence to produce the whole reconstructed time-series, according to the light grey tones at the bottom of Fig. 2.

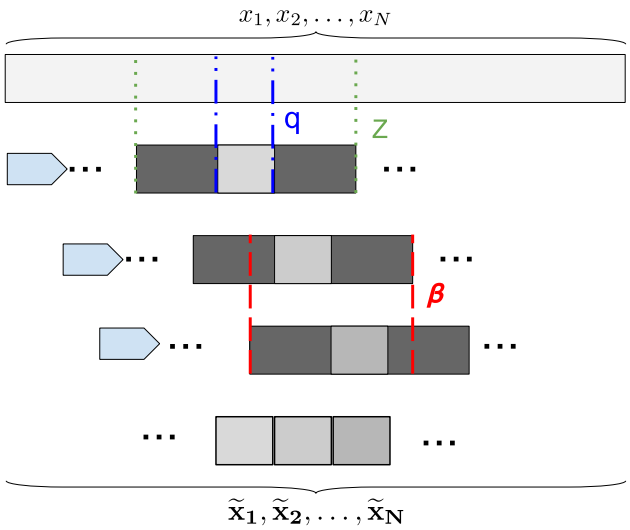


Fig. 2. Segmented SSA.

3.1. Algorithm

The pseudocode of this method is presented in Algorithm 1. Refer to Leles et al. [28] to obtain an implementation of this algorithm and detailed instructions on how to use it, including some examples.

Algorithm 1: ov-SSA: the segmented version of the standard SSA, a modification proposed in this paper.

Data: The original time-series, $(x_1, x_2, \ldots, x_N)$.
Input: Standard SSA parameters L (the immersion dimension) and $\{I_m\}$ (the grouping); The segmented version parameters Z (local segment length) and q (the amount of samples shifted).
Output: Time-series reconstructed by the segmented SSA algorithm, $\widetilde{\mathbf{x}}$.
initialization;
$\bar{L} \leftarrow (Z - q)/2$;
$\mathcal{P} = \lfloor (N - Z)/q \rfloor + 1$;
$s \leftarrow (x_1, x_2, \ldots, x_Z)^T$;
$\hat{s} \leftarrow SSA(s, L, \{I_m\})$;
$\widetilde{\mathbf{x}} \leftarrow (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{\bar{L}+q})^T$;
for $p \leftarrow 2$ **to** $\mathcal{P} - 1$ **do**
  $\rho \leftarrow (p-1)q + 1$;
  $s \leftarrow (x_\rho, x_{\rho+1}, \ldots, x_{\rho+Z})^T$;
  $\hat{s} \leftarrow SSA(s, L, \{I_m\})$;
  $\widetilde{\mathbf{x}} \leftarrow (\widetilde{\mathbf{x}}, \hat{s}_{\rho+\bar{L}+1}, \hat{s}_{\rho+\bar{L}+2}, \ldots, \hat{s}_{\rho+\bar{L}+q})^T$;
end
$p \leftarrow p + 1$;
$\rho \leftarrow (p-1)q + 1$;
$s \leftarrow (x_\rho, x_{\rho+1}, \ldots, x_{\rho+N})^T$;
$\hat{s} \leftarrow SSA(s, L, \{I_m\})$;
$\widetilde{\mathbf{x}} \leftarrow (\widetilde{\mathbf{x}}, \hat{s}_{\rho+\bar{L}+1}, \hat{s}_{\rho+\bar{L}+2}, \ldots, \hat{s}_N)^T$;

3.2. Resuming the discussion about non-stationary signals

Now, in possession of the proposed algorithm, the first example discussed in this paper can be analyzed again. Fig. 3 compares the original time-series and two reconstructed time-series by ov-SSA. The moving window has length $Z = 128$ and the overlap is 93.75% ($q = 8$). The immersion parameter is the same used by the classical SSA, $L = 60$. Two groupings are considered. Figs. 3(a) and 3(b) show that using the first 2 and first 6 components, respectively, the reconstruction is better than the standard SSA, recall Fig. 1.

Details for the time interval 1.5 to 3 seconds are provided in Figs. 3(c) to 3(f). Within this interval, the standard SSA, using dozens of components, is unable to reconstruct any significant feature of the original signal. The use of only 2 elementary matrices $\mathbf{A}_{\{1,2\}}$ improves the overall result, but in the last segment, the error is more pronounced due to border effect. Nevertheless, a small increase in the quality of components, from 2 to 6, can capture all the relevant features of the chirp signal, as portrayed in Fig. 3(b).

These visual cues are confirmed by some metrics. The reconstruction errors are revealed in Fig. 3(g). The observed error, which has amplitude ten times smaller than the original signal, can be regarded as the additive noise, since this simulations considers a SNR of 20 dB. The proposed segmented analysis indeed discriminates relevant from the meaningless information.

This example serves to motivate the conjecture that segmentation can improve the discrimination capability of the SSA. This is going to be illustrated in the sequel by more complex examples.

4. Results and discussions

In this section the proposed algorithm is applied to synthetic and experimental datasets in order to illustrate more advantages. The reconstruction performance can be quantified by standard error metrics, such as Mean Absolute Error (MAE), according to:

$$\text{MAE} = \frac{1}{N}\sum_{n=0}^{N-1} |x_n - y_n|, \tag{9}$$

where x_n is a sample of the original time-series at instant n, y_n is the reconstructed sample at this same instant and N indicates the size of the series.

4.1. Time-series reconstruction: synthetic time-series

The first example in this paper is further inspected to allow a more insightful analysis. The original time-series is considered a noise free version of the chirp signal, labeled as c. Then, standard SSA grouping, considering $L = 60$, is greatly raised $\mathbf{A}_{\{1,2,\cdots,42\}}$. In this case, considering only c, standard SSA can achieve a reconstruction error in the same magnitude of the one obtained by ov-SSA with $L = 60$. Notice that for this example, if 60 components are used in the grouping stage the reconstruction is precise, i.e., it produces the exact original time-series and no component is discriminated, including noise.

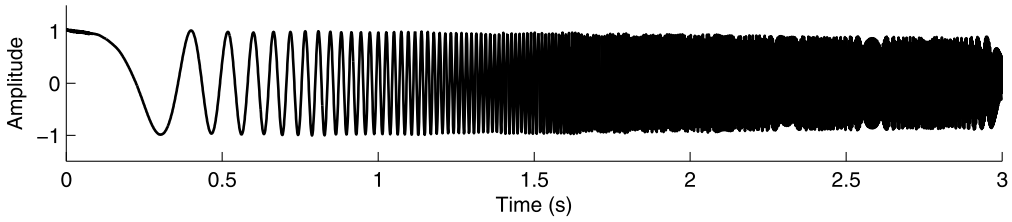
Then, the noisy version of the chirp signal is considered, $x = c + \epsilon$. The noise intensity is progressively raised such that $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, with $\sigma_\epsilon^2 = (1, 2, \ldots, 5)/100$, providing SNR = 7.34 dB, 4.30 dB, 2.21 dB, 1.13 dB and 0.10 dB.

A total of 100 simulations were performed for each of the previous noise intensities. These time-series x are used as inputs for the standard SSA and ov-SSA, resulting in several reconstructed time-series y. For each simulation the MAE is computed, according to Equation (9). The results are compiled as boxplot graphics show in Fig. 4, for both SSA and ov-SSA.

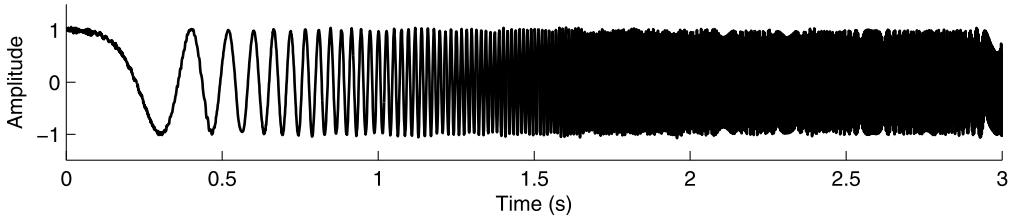
4.2. Time-series reconstruction: experimental time-series

A dataset comprising three-million years of mutually consistent records of surface air temperature is considered. This is an example of a large time-series, sampled at 100 year interval, totalizing 30,000 data points. For a discussion about this time-series see Bintanja and van de Wal [6]. To promote a fair comparison with the WSSA method, introduced by Rekapalli and Tiwari [36], the linear trend was removed.

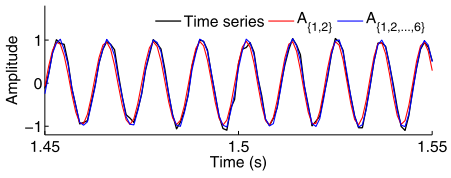
Fig. 5 shows the original time-series and the reconstructions provided by standard SSA, WSSA, and ov-SSA methods. Embedding Dimension and grouping were: $L = 200$ and $A_{\{1,2,\cdots,6\}}$, respectively. For the WSSA and the proposed method the segment length was $Z = 4096$. These are the same parameters used in Rekapalli



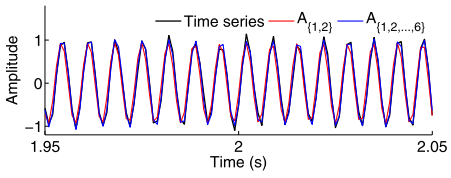
(a) Reconstructed ov-SSA time-series using $L = 60$ and $\mathbf{A}_{\{1,2\}}$.



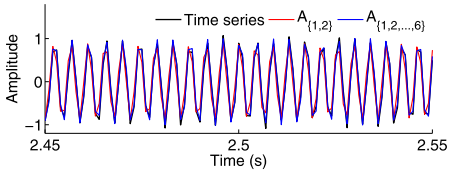
(b) Reconstructed ov-SSA time-series using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,6\}}$.



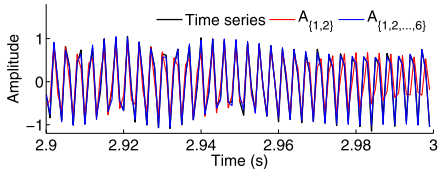
(c) First Detail.



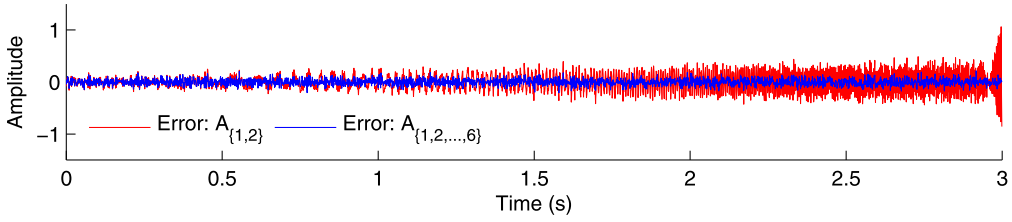
(d) Second Detail



(e) Third Detail.

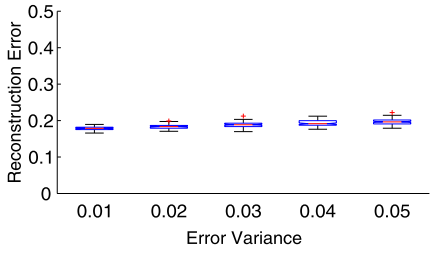


(f) Another Detail.

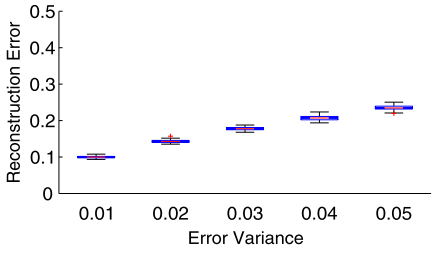


(g) Reconstruction error.

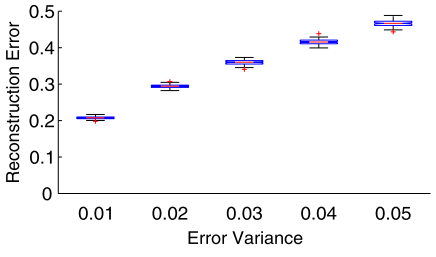
Fig. 3. Reconstructed time-series by SSA using $L = 60$ and two groupings for the same time-series considered in Fig. 1(a) along with the reconstruction error.



(a) Boxplot error: ov-SSA using $L = 60$ and $\mathbf{A}_{\{1,2\}}$.



(b) Boxplot error: ov-SSA using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,6\}}$.



(c) Boxplot error: SSA using $L = 60$ and $\mathbf{A}_{\{1,2,\cdots,42\}}$.

Fig. 4. Comparison of reconstruction error for chirp time-serie of both ov-SSA and SSA for different noise variance.

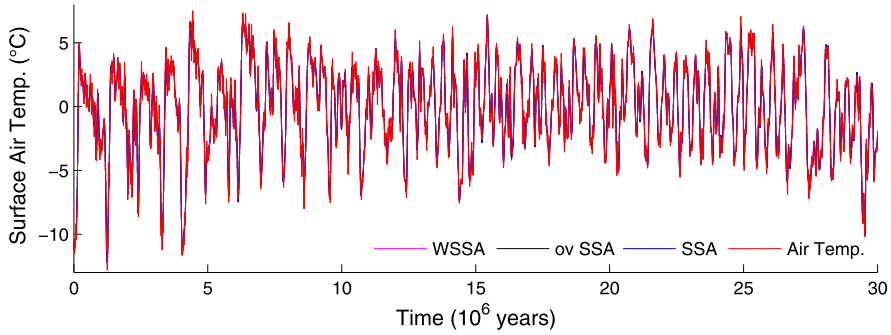
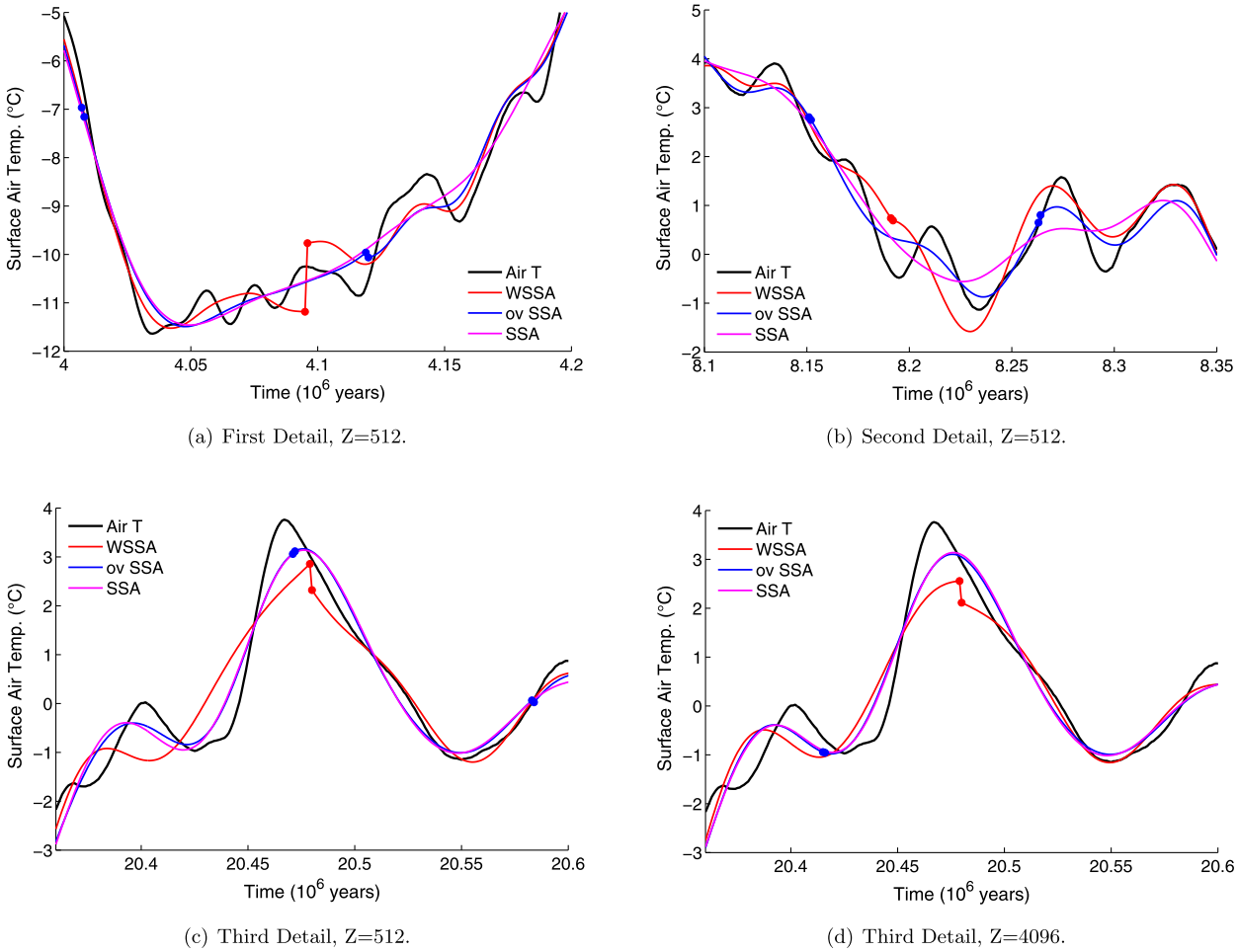


Fig. 5. Surface air temperature and reconstructed time-series according to: standard SSA, WSSA and ov-SSA.

Table 2
Mean Absolute Error for the reconstructed time series whose time intervals are depicted in Figs. 5 and 6.

	Fig. 5		Fig. 6(a)		Fig. 6(b)		Fig. 6(c)	
	$Z = 512$	$Z = 4096$	$Z = 512$	$Z = 4096$	$Z = 512$	$Z = 4096$	$Z = 512$	$Z = 4096$
ov-SSA	0.2486	0.2978	0.3533	0.3675	0.2767	0.3832	0.3004	0.3313
WSSA	0.2724	0.2984	0.3617	0.4048	0.3483	0.3997	0.4300	0.4024
SSA	0.3002		0.3646		0.3915		0.3332	



(a) First Detail, Z=512.

(b) Second Detail, Z=512.

(c) Third Detail, Z=512.

(d) Third Detail, Z=4096.

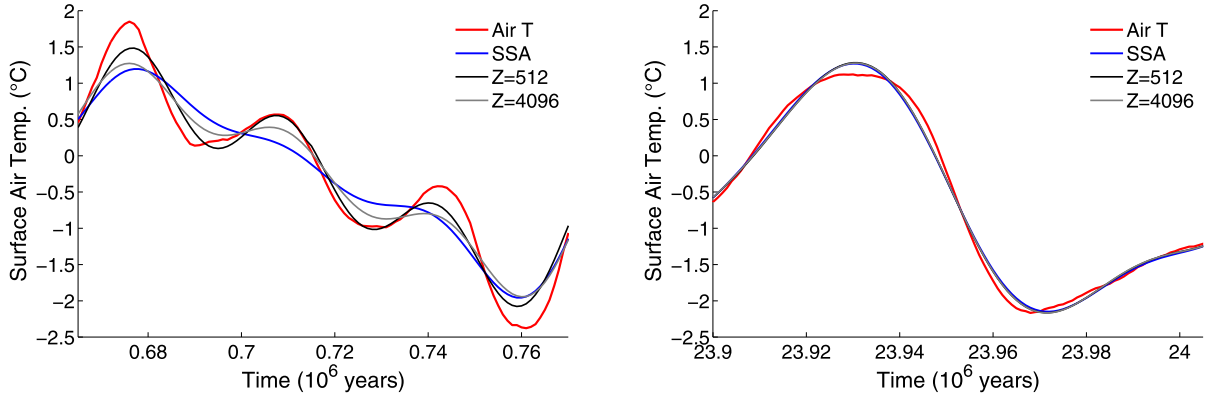
Fig. 6. Some details of the concatenation issues of WSSA that are improved by ov-SSA.

and Tiwari [36]. The amount of overlap between consecutive segments was set to 80%, which is exclusive for the proposed method.

A quantitative analysis was carried out in Tables 2 and 3. They show the MAE for the complete time series of Fig. 5 and the excerpts illustrated in Figs. 6 and 7.

4.3. Discussion about time-series reconstruction

Fig. 4 gives more insight concerning the ability to discriminate components. The comparison indicates that ov-SSA consistently produced smaller reconstructed errors than standard SSA as the noise variance is allowed to change.



(a) High-frequency captured. (b) Amplitude consistency.

Fig. 7. Original time-series (red) and reconstructed ones by SSA (blue) and ov-SSA (black) using $L = 60$ and two groupings. Standard SSA detects mainly trend whereas ov-SSA captures high frequency oscillations (a). No discrepancy between both methods (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Mean Absolute Error for the reconstructed time series whose time intervals are depicted in Fig. 7.

	Fig. 7(a)		Fig. 7(b)	
	$Z = 512$	$Z = 4096$	$Z = 512$	$Z = 4096$
ov-SSA	0.1567	0.2256	0.0731	0.0730
SSA	0.3047		0.0746	

The analysis of Figs. 4(a) and 4(b) indicates another trend. The ov-SSA with 2 components is less affected by the noise variance than the ov-SSA with 6 components, being somewhat more robust. Although ov-SSA with 6 components starts with the lowest error of all methods, insofar as the noise variance increases, is ov-SSA with 2 components that presents the smallest error among all methods.

An explanation that can be drawn is that considering larger groupings may not be an effective alternative to improve reconstruction by standard SSA, insofar the ability to separate information from noise can be impaired. To the extent that more components are considered, more meaningless information becomes part of the reconstruction process, leading to deviations.

Now lets take a look to the experimental data. Some time intervals of Fig. 5 are detailed in Figs. 6 and 7. They intend to highlight some distinguishable aspects of each reconstruction method.

Fig. 6 reveals the boundary effects of the method proposed by Rekapalli and Tiwari [36]. When two consecutive segments are joined there is a noticeable discontinuity. This effect becomes more clear as the length of the segments are made shorter, as the comparison of Figs. 6(c) and 6(d) evinces, where Z assumes 512 and 4096, respectively.

In contrast, the proposed method minimizes this kind of artifact, providing a more smooth transition among segments, in all cases investigated. The proposed method is somewhat more robust, since the size of the sliding window does not impact much in the discontinuity effect (notice Figs. 6(a) to 6(c), which represents the same length Z for different intervals).

Table 2 indicates that such discontinuities result in a worse reconstruction. In overall, the reconstruction error provided by WSSA is greater than the one provided by the proposed approach. The ov-SSA was able to decrease the error by 9% and 17% compared to the WSSA and standard SSA, respectively, considering the entire data. For some intervals, the decrease was more relevant, by 21% and 29%, respectively.

Another difference can be seen in Fig. 7, where the proposed method fits better the original time-series than the classical SSA. This is confirmed by Table 3, related to the detail shown in Fig. 7(a). The reconstruction error was decreased almost by half.

It seems that the classical SSA is able to capture long term components, present over most of the signal. The proposed method still can capture these components, but is also able to identify more information present locally, corresponding to short time movements. These findings are not always verifiable, as shown by Fig. 7(b), in which both methodologies achieve the same reliability in the reconstruction, confirmed by the close reconstruction errors found in Table 3.

By setting the parameters accordingly, the results from both the standard SSA and the WSSA can be recovered. The first case is achieved by not segmenting the time-series, i.e., setting the segment size equal to the time-series length, $Z = N$. The second situation corresponds to no overlap, i.e., setting $q = Z$. Therefore, the proposed method is more general.

In short, these examples have show that the ov-SSA can achieve a compromise between improved reconstruction and the ability to separate relevant information from meaningless ones. It also revealed that ov-SSA promotes a more smooth transitions among segments in comparison with other segmentation methods. The improvements of ov-SSA, in terms of computational complexity for large datasets, are discussed in detail in [28].

In the sequel, other advantages of the proposed approach are presented, besides improving time-series reconstruction.

4.4. Improving time–frequency characterization

In this section, the main objective is to explore additional information revealed by the segmented approach of the SSA. There may be situations in which the improvement in reconstruction is negligible but the analysis using the ov-SSA still can be beneficial. Towards this end, additional synthetic and experimental data are analyzed.

4.4.1. Time–frequency characterization: synthetic time-series

A synthetic signal is proposed according to:

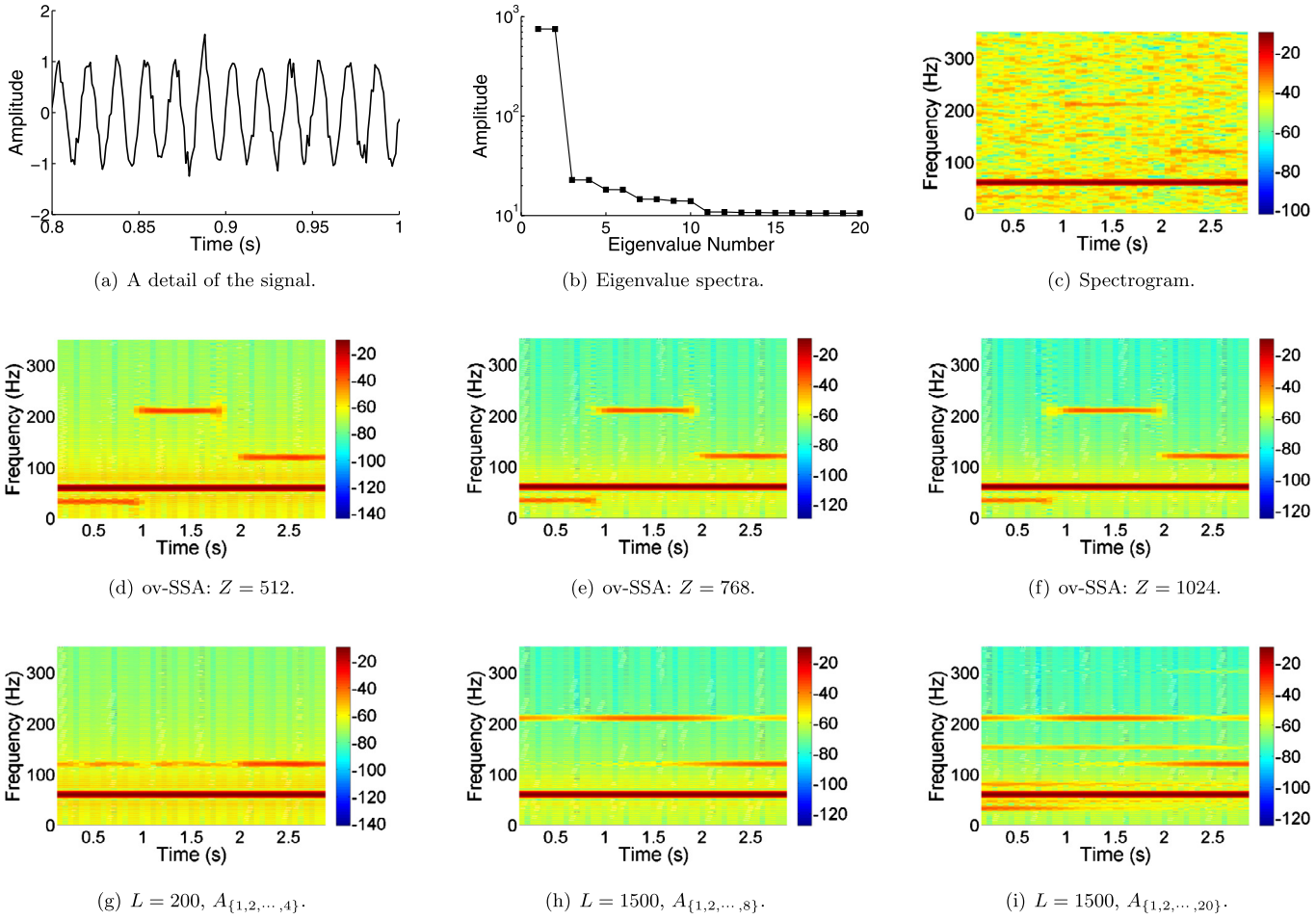
$$x(t) = \sin(2\pi f t) + h(t) + \xi, \qquad f = 60\,\text{Hz}, \quad 0 \le t \le 3, \tag{10}$$

where $h(t)$ is transitory signal given by:

$$h(t) = \begin{cases} 0.05\sin(2\pi f_1 t), & 0 \le t \le 1; \\ 0.05\sin(2\pi f_2 t), & 1 \le t \le 2; \\ 0.05\sin(2\pi f_3 t), & 2 \le t \le 3. \end{cases} \tag{11}$$

For the transitory frequencies $f_1 = 33$ Hz, $f_2 = 210$ Hz, and $f_3 = 120$ Hz, such that the signal ξ is noise (SNR = 15 dB), the synthetic time-series is shown in Fig. 8(a). Its eigenvalue spectra is portrayed in Fig. 8(b).

The spectrogram allows to analyze the energy within the frequency–time distribution. Fig. 8(c) depicts the computation of



(a) A detail of the signal. (b) Eigenvalue spectra. (c) Spectrogram.

(d) ov-SSA: $Z = 512$. (e) ov-SSA: $Z = 768$. (f) ov-SSA: $Z = 1024$.

(g) $L = 200$, $A_{\{1,2,\cdots,4\}}$. (h) $L = 1500$, $A_{\{1,2,\cdots,8\}}$. (i) $L = 1500$, $A_{\{1,2,\cdots,20\}}$.

Fig. 8. Analysis of the synthetic signal proposed by Equation (10). Figs. (a) to (c) correspond to the original time-series simulated: a detail of the signal (a); its eigenvalue spectra (b); the spectrogram of the whole signal (c). Figs. (d) to (f) correspond to the spectrogram of the reconstructed time-series of ov-SSA varying the segment length Z, but keeping constant $L = 200$ and $A_{\{1,2,3,4\}}$, whereas (g) to (i) correspond to the classical SSA, varying the immersion dimension and grouping.

Table 4
Mean Absolute Error, Equation (9). These results correspond to the reconstructed time-series whose spectrograms are displayed in Fig. 8.

ov-SSA			SSA		
Fig. 8(d)	Fig. 8(e)	Fig. 8f	Fig. 8(g)	Fig. 8(h)	Fig. 8(i)
0.0985	0.0986	0.0988	0.1011	0.1002	0.0979

the Short-Time Fourier Transform (STFT) for the original time-series shown in 8(a).

Figs. 8(d), 8(e) and 8(f) present the spectrogram of reconstructed time-series by ov-SSA using $L = 200$ and $A_{\{1,2,3,4\}}$, for three segments length: $Z = 512$, $Z = 768$, and $Z = 1024$, respectively.

Figs. 8(g), 8(h) and 8(i) shows the spectrogram of reconstructed time-series for SSA approach. In Fig. 8(g) it was used the same parameters as for ov-SSA, i.e., $L = 200$ and $A_{\{1,2,3,4\}}$. In Fig. 8(h) the choice was $L = N/2 = 1500$, and $A_{\{1,2,\cdots,8\}}$, following the rule of thumb from [10]. In Fig. 8(i) more elementary matrices were used in grouping stage, $A_{\{1,2,\cdots,8,\cdots,20\}}$.

Table 4 depicts the MAE of reconstructed time-series for both approaches.

4.4.2. Time–frequency characterization: experimental time-series

As an example of experimental time-series, recorded voice sample was chosen from the American English Vowels Database[6] [22].

[6] The recordings are freely available at http://homepages.wmich.edu/~hillenbr/.

Table 5
Mean Absolute Error for different vowels.

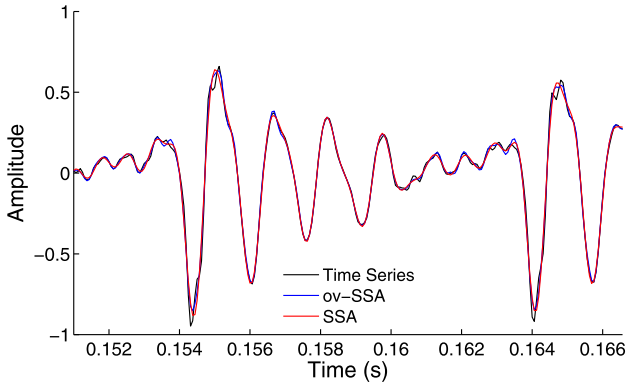
Vowel	ov-SSA				SSA
	$Z = 128$	$Z = 256$	$Z = 512$	$Z = 1024$	
/æ/	0.0183	0.0212	0.0212	0.0212	0.0218
/ɑ/	0.0100	0.0146	0.0146	0.0147	0.0155
/ɔ/	0.0044	0.0053	0.0053	0.0053	0.0056
/ɛ/	0.0134	0.0152	0.0155	0.0155	0.0158
/e/	0.0121	0.0146	0.0146	0.0148	0.0154
/ɜ/	0.0055	0.0093	0.0092	0.0095	0.0104
/ɪ/	0.0137	0.0166	0.0163	0.0163	0.0178
/i/	0.0195	0.0261	0.0262	0.0264	0.0266
/o/	0.0036	0.0046	0.0046	0.0047	0.0046
/ʊ/	0.0094	0.0116	0.0116	0.0125	0.0174
/ʌ/	0.0125	0.0144	0.0146	0.0152	0.0163
/u/	0.0028	0.0033	0.0033	0.0033	0.0035

The recording consists of male subject vowels, with 16-bit resolution and 16000 Hz sample-rate.

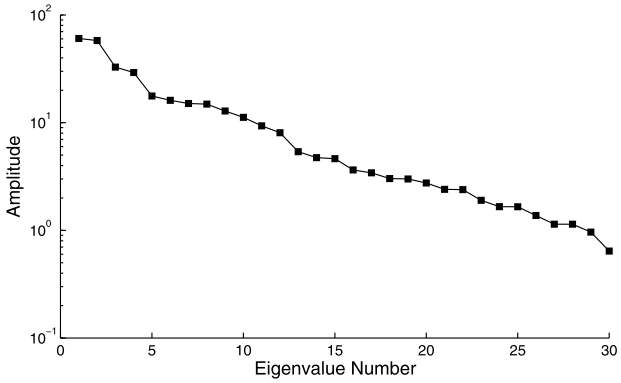
A similar analysis to the last section was conducted (refer to Fig. 9), where the vowel /ʌ/, as in the word “hud” is used. The parameters are $L = 60$ and $A_{\{1,2,\ldots,12\}}$.

Every vowel in this database was reconstructed with the standard SSA and the segmented SSA, resulting in the MAE presented in Table 5. Distinct values for parameter Z were used.

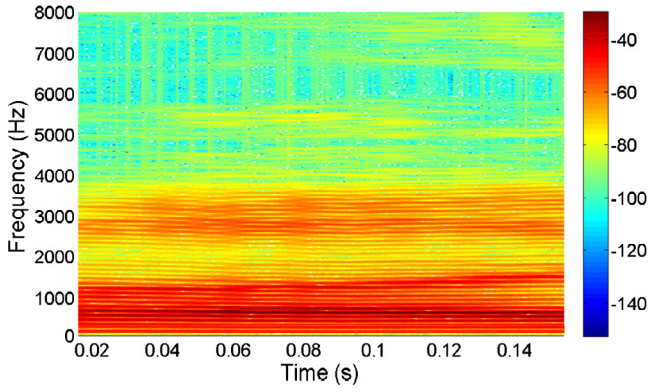
Fig. 9(f) shows the spectral density estimation using the linear prediction coefficients (LPC) [31]. The LPC was performed using



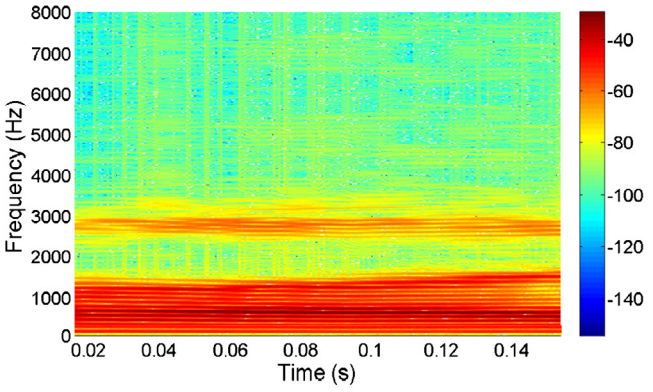
(a) Original and reconstructed time-series.



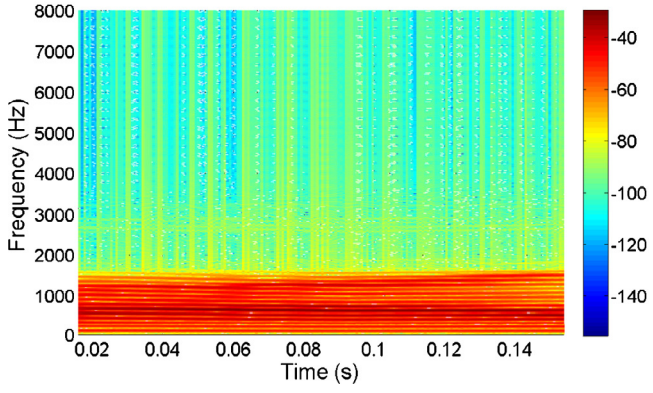
(b) Eigenvalue Spectra.



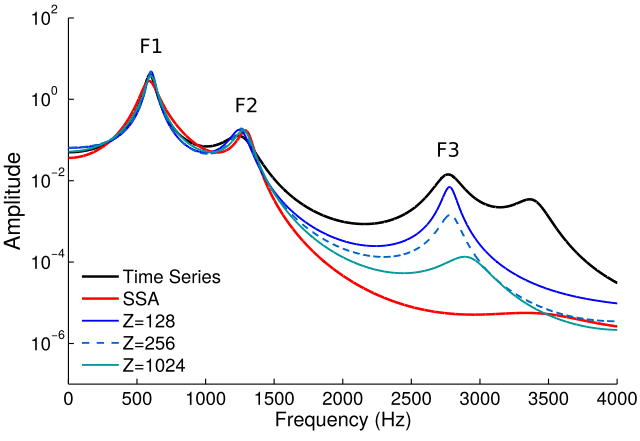
(c) Spectrogram of original time-series.



(d) ov-SSA: $Z = 128$, $L = 60$, $A_{\{1,2,\ldots,12\}}$.



(e) SSA: $L = 60$, $A_{\{1,2,\ldots,12\}}$.



(f) Power Spectral Density Estimation.

Fig. 9. Analysis of the experimental time-series. Fig. (a) shows the original time-series (blue), the reconstructed time-series by standard SSA (red) and ov-SSA (green). Fig. (b) exhibits the eigenvalue spectra. The spectrogram of the whole signal is shown in (c). Figs. (d) and (e) correspond to the spectrogram of the reconstructed time-series by ov-SSA, with parameters $Z = 128$, $L = 60$, and $A_{\{1,2,\ldots,12\}}$, and classical SSA, with $L = 60$, and $A_{\{1,2,\ldots,12\}}$, respectively. In (f) the estimation of spectral power density of all reconstructed series. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

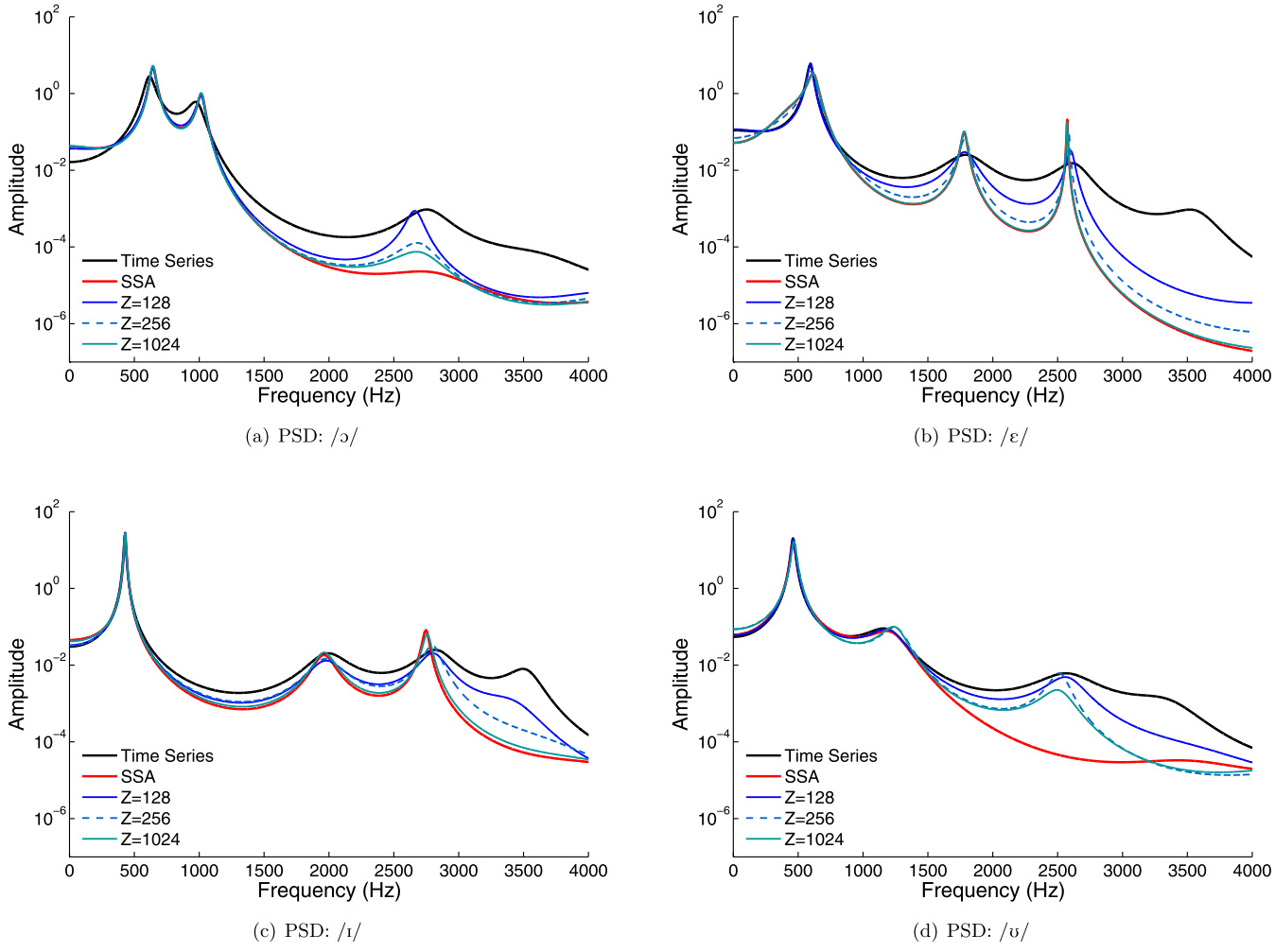
14 coefficients and models the subject vocal tract frequency response.

Spectrum peaks in Fig. 9(f) are the resonances in the vocal tract, usually referred as “formants”, which are important in vowel distinction [22,9]. The three first formants are highlighted in this figure. This information is very useful since the location of the first two or three resonant peaks are used to discriminate between vocalic sounds.

4.5. Discussion about time–frequency characterization

A brief examination of Fig. 8 indicates that both the classical SSA and the ov-SSA are able to identify the main component in the signal, the sustained oscillation at 60 Hz. However, an in-depth analysis point some shortcomings about standard SSA.

At first, the quantitative analysis carried out in Table 4 shows that the reconstructed signal by both approaches posses the MAE of same magnitude. Therefore, in the light of the results discussed



(a) PSD: /ɔ/ (b) PSD: /ɛ/

(c) PSD: /ɪ/ (d) PSD: /ʊ/

Fig. 10. Estimation of power spectral density (PSD) of some reconstructed vowels.

in the previous section, both methods achieved the same reliability. However, considering the time–frequency distribution some new insights are possible.

In Fig. 8(g) only the signal with frequency f_1 is detected. As the grouping is allowed to increase, more transient frequencies are detected, until that in Fig. 8(i) where the standard SSA identifies the vanishing signals f_1, f_2, and f_3. However, it shows another frequencies that are not meaningful. In time-series reconstructed by classical SSA with $L = 1500$, $A_{\{1,2,\cdots,8\}}$, whose spectrogram is shown in Fig. 8(h), all the relevant structures are easy to identify. However, only the time location of f_1 is precise. Locate signals f_2 and f_3 is not possible with accuracy. In short, even thought the optimal value for L was chosen, leading to the least amount of MAE for the standard SSA, some deficiencies are noticed concerning time–frequency distribution.

The proposed method provided a more accurate information. In this example, all reconstructed time-series, by distinct segment sizes, are composed mainly by the relevant signals of Equation (10). It is very noticeable that the reconstructed time-series carries a great deal of information. As seen in red tones, in Figs. 8(d) to 8(f), four distinct components were identified, representing the meaningful frequencies of the original time series. Furthermore, the time-location of these components are very consistent with the description of Equation (10).

Regarding the experimental time series, ov-SSA reconstructs the time-series with less error or at least with errors in the same magnitude of standard SSA, as Table 5 presents.

However, compare Fig. 9(c), the spectrogram of the original signal, with Figs. 9(d) and 9(e), the spectrograms of ov-SSA and standard SSA, respectively. The proposed method indicates 3 relevant frequencies (in red/orange tones) whereas the standard SSA can detect just 2. Referring to Fig. 9(f) another aspect deserves attention. As the parameter Z decreases the peak location changes.

Different vowels are characterized by the spectral peak locations. In the practical point of view, the ov-SSA provides an alternative to extract parameters for voice coding. Formant structure is kept after signal reconstruction and three spectral peaks were always recognizable by ov-SSA. Standard SSA, in contrast, extracted only two peaks in many cases. This aspect is illustrated by the PSD estimation in Fig. 10 for other vowels. Notice that since the higher frequency formants have less relative magnitude, they are harder to be estimated.

In general, ov-SSA is able to unveil more peaks than the regular SSA, offering more suitable parameters for an automatic speech recognition algorithm, for example. In Figs. 10(b) and 10(c) both the ov-SSA and standard SSA are able to identify three formants. However, in Figs. 10(a) and 10(d) the third formant was captured only by ov-SSA. Similar findings are obtained for the remaining vowels, omitted for the sake of brevity.

A quantitative analysis is carried in Table 6. The absolute errors is computed as the difference between the location of frequency peaks of the SSA methods and the location of the frequency peaks by the PSD estimation. It is important to keep in mind that PSD uses an auto-regressive method to estimate spectra. Symbol "—"

Table 6
Absolute errors. Difference between the location of frequency (Hertz) peaks of the SSA methods and the location of the frequency peaks by the PSD estimation.

	/æ/			/ɑ/			/ɔ/			/ɛ/		
	F_1 685	F_2 1807	F_3 2611	F_1 825	F_2 1024	F_3 2701	F_1 664	F_2 1024	F_3 2639	F_1 577	F_2 1811	F_3 2559
SSA	7.02	11.83	7.24	200.24	29.98	–	20.69	8.13	67.30	37.26	31.95	14.73
$Z = 128$	11.66	10.86	40.12	124.32	71.09	18.97	23.13	12.77	22.13	18.21	28.77	42.07
$Z = 256$	10.44	11.35	8.38	199.02	28.76	–	21.42	10.3	38.98	26.03	26.58	26.21
$Z = 1024$	9.46	12.57	1.63	213.18	36.09	–	20.93	9.35	34.58	37.50	30.97	13.75

	/e/			/ɜ/			/ɪ/			/i/		
	F_1 439	F_2 2126	F_3 2762	F_1 439	F_2 1220	F_3 1681	F_1 411	F_2 2034	F_3 2795	F_1 327	F_2 2257	F_3 2943
SSA	12.17	34.94	14.86	3.45	198.46	–	18.93	72.82	48.17	17.97	33.12	26.98
$Z = 128$	11.68	10.03	25.67	3.94	258.03	–	16.98	56.22	5.29	6.50	7.98	5.50
$Z = 256$	12.42	26.63	2.16	3.70	188.20	–	16.49	59.39	0.90	16.51	29.46	18.68
$Z = 1024$	12.42	28.10	4.60	3.70	206.76	–	18.44	70.87	37.68	16.75	32.15	23.57

	/o/			/ʊ/			/ʌ/			/u/		
	F_1 486	F_2 826	F_3 2631	F_1 459	F_2 1036	F_3 2625	F_1 636	F_2 1153	F_3 2756	F_1 371	F_2 793	F_3 2719
SSA	44.35	31.32	117.54	2.67	144.66	–	47.38	137.04	–	16.75	3.94	28.08
$Z = 128$	48.50	49.63	131.94	2.43	139.78	67.87	33.71	93.58	22.32	21.15	14.44	12.69
$Z = 256$	47.28	44.51	141.22	8.77	204.97	113.53	31.51	107.74	26.96	19.68	10.53	0.49
$Z = 1024$	47.04	43.77	137.55	8.53	197.15	129.39	34.93	120.93	133.16	18.71	8.09	26.13

indicated a formant that was not detected. In general, ov-SSA produces frequency estimations compatible with standard SSA and PSD. However, it can be seen that, in some cases, ov-SSA method identifies more formants than standard SSA, with a result more similar to the one produced by PSD. The ov-SSA is able to locate higher frequencies formants, which are harder to find.

5. Remarks about the proposed technique

In this section, the proposed approach is contextualize with other segmentation approaches present in the recent SSA Literature. Also, the parameters selections is discussed. For more technical details about the implementation the reader is referred to Leles et al. [28].

5.1. Comparison with segmented versions of SSA

The initiative to extend standard SSA to a local version has been proposed in other studies. Yiou et al. [44] proposed an approach called MS-SSA (Multi-Scale SSA). It seeks to perform a joint time–frequency analysis, analogously to wavelet transform, a capability missing in standard SSA. The improvement is obtained by Yiou et al. [44] with a great impact over the computational effort. To reconstruct a single sample of the original time-series at an instant, the MS-SSA algorithm applies several computations of the standard SSA, each one considering different lengths for the segment Z (see Yiou et al. [44] Fig. 11, for instance). For the next sample the procedure is repeated and so on. Another difference is the length of the reconstructed series. In the proposed approach reconstructed and original time-series are of the same size whereas in MS-SSA the reconstructed time-series is shorter.

A more efficient approach of local SSA, in terms of computational effort, appears in Rekapalli and Tiwari [36], called Windowed SSA. The idea is to compute SSA over consecutive, non-overlapping, segments of the original dataset. This approach lessen the computational effort, being useful for the long time-series studied in Rekapalli and Tiwari [36]. However, since there is no treatment about how the successive segments are connected, time-series reconstruction can be worse and some artifacts appear, as discussed in Section 4.3.

Table 7
Summarizing the range of SSA variations. this list may be merged with the SSA variations presented in column "Automatic" in Table 1.

Time-series characteristics	SSA variation
Improving the separability	[13]: Oblique SSA[a]
Outliers are presented:	[24]
Forecasting multiple time-series	[20]: Multivariate SSA
Mixed time series components	[30]: Sequential SSA
Detection of structural changes	[32]: change-point detection
Image processing	[37]: 2d-SSA
Missing data	[40] and [12]
Time-varying structures	The proposed method: ov-SSA

Finally, the proposed approach was inspired by the *overlap-save sectioning* method used in signal processing for filtering or convolution of infinite (or finite) duration waveforms [8]. Note that the standard SSA algorithm suffers from boundary effects on both sides. In the *overlap-save sectioning* method only the initial points must be discarded, because boundary effects occur only at the filtering initialization.

5.2. When the ov-SSA is recommended?

In the sequence a list of possible applications of the ov-SSA is given:

***A priori* knowledge:** When the analyst experience or the physical nature of the problem indicates that the signal is non-stationary.

Spectrogram: Application of this technique may reveal the presence of transient information hidden in the original time-series.

Structural change: In [32] a change-point detection and subspace tracking algorithm is proposed. In this sense, if the structural change in the time-series occurs, the ov-SSA may benefit from it.

Therefore, the ov-SSA is another method that enlarges the range of options in terms of variations of the standard SSA, as Table 7 summarizes. This list can be expanded with the methods presented in column "Automatic" in Table 1.

5.3. Parameter selection

The choice of the SSA window length (L) has effect on the fast/slow characteristic of decomposition. This gives rise to a trade-off between the eigenfilters ability to quickly adapt itself when an abrupt change occurs. As L increases, the estimate becomes slow, so the ability to self-adapt to the signal changes is compromised. On the other hand, small values of L affect the capability to capture harmonic components whose periods are greater than the value of this parameter. This can be a drawback of the proposed approach because signals with very large period, when compared to the segment Z, may be interpreted only as trend. The choice of parameter q can be regarded as a compromise solution between computational cost and detection of transient terms in the series. A detailed analysis concerning this effect in the reconstruction error is given in [28]. In short, the selection of parameters Z and q are related to the selection of the standard SSA parameters.

A paramount aspect revealed by the results is the robustness of the proposed method. Even if the length Z changes by 8 times, the computed MAE does not varies at the same rate, as presented by Tables 2, 3, 4 and 5. The maximum increase in the MAE of the proposed method, in relative terms, was 50%.

Although some guidelines were briefly described in Section 2.5, the discussion about systematic choice of the parameters lays outside of the scope of this article once there is an open debate in the Literature of the standard SSA. On the other hand, Golyandina and Zhigljavsky [15] states that: "Whatever the circumstances, it is always a good idea to repeat SSA analysis several times using different values of L." In the SSA-CT method proposed by Hassani et al. [18], for each L values, a grouping is obtained. Based on this, at each iteration of the proposed method, the SSA-CT, for instance,[7] could be used to SSA parameter tuning.

6. Conclusions and future works

In this paper and in Leles et al. [28], which is a companion paper for this Joint Special Issue on "Reproducible Research in Signal Processing", a new method that improves segmented analysis based on the SSA has been proposed. The method consists on dividing a time-series into smaller, consecutive, and overlapping segments. By using this strategy, the ability to separate relevant components from noise or meaningless information is improved. This was illustrated by synthetic and real datasets. The reconstructed time-series by the proposed approach are more trustworthy than the ones produced by other SSA based methods. Quantitative and qualitative analysis were carried out to corroborate this conclusion. Implementation issues, complexity analysis and sensitivity of the parameters are detailed in the companion work by Leles et al. [28].

A byproduct of the proposed method, which was initially developed for time-series reconstruction, is an enhancement of the time–frequency characterization by the SSA. In the cases studied in this paper, even when the reconstructed time-series were qualitative and quantitative similar in terms of reconstruction fidelity, more information was present in the time-series reconstructed by ov-SSA. The proposed approach, when compared with the standard SSA, produced signals that unveil more significant components, resulting in a better frequency characterization. This can be very useful for speech recognition applications, as illustrated. Also, it has been shown that proposed segmented analysis can identify with more accuracy, the time-slot where each frequency occur in the original signal, another situation in which the standard SSA can fail.

When the analyzed signal is stationary or when its harmonic components have large period, in comparison with segment Z, the efficient implementation of SSA is the right choice. However, when the analyzed signal posses a strong time-varying characteristic, the ov-SSA seems to be the recommend alternative.

As future works, the authors are interested in developing a systematic procedure for the choice of parameters and in improving the proposed algorithm by allowing the adaptive choice of the parameter Z.

Acknowledgments

We appreciate the reviewers and editors valuable comments that certainly improved this manuscript quality. We also acknowledge the Brazilian agencies CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) for partial support of this work.

References

[1] M. Abdollahzade, A. Miranian, H. Hassani, H. Iranmanesh, A new hybrid enhanced local linear neuro-fuzzy model based on the optimized singular spectrum analysis and its application for nonlinear and chaotic time series forecasting, Inf. Sci. 295 (2015) 107–125.
[2] T. Alexandrov, A method of trend extraction using singular spectrum analysis, REVSTAT Stat. J. 7 (2009) 1–22.
[3] T. Alexandrov, N. Golyandina, Automatic extraction and forecast of time series cyclic components within the framework of SSA, in: Proceedings of the Fifth Workshop on Simulation, 2005, pp. 45–50.
[4] N. Alharbi, H. Hassani, A new approach for selecting the number of the eigenvalues in singular spectrum analysis, J. Franklin Inst. 353 (2016) 1–16.
[5] F. Alonso, D. Salgado, Analysis of the structure of vibration signals for tool wear detection, Mech. Syst. Signal Process. 22 (2008) 735–748.
[6] R. Bintanja, R.S.W. van de Wal, North American ice-sheet dynamics and the onset of 100,000-year glacial cycles, Nature 454 (2008) 869–872.
[7] P. Bonizzi, J.M.H. Karel, O. Meste, R.L.M. Peeters, Singular spectrum decomposition: a new method for time series decomposition, Adv. Adapt. Data Anal. 06 (2014) 1450011.
[8] E. Brigham, Fast Fourier Transform and Its Applications, Prentice Hall, 1988.
[9] G. Fant, Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations, vol. 2, Walter de Gruyter, 1971.
[10] N. Golyandina, On the choice of parameters in singular spectrum analysis and related subspace-based methods, Stat. Interface 3 (2010) 259–279.
[11] N. Golyandina, V. Nekrutkin, A.A. Zhigljavsky, Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, 2001.
[12] N. Golyandina, E. Osipov, The "caterpillar"-SSA method for analysis of time series with missing values, J. Stat. Plan. Inference 137 (2007) 2642–2653.
[13] N. Golyandina, A. Shlemov, Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series, Stat. Interface 8 (2015) 277–294.
[14] N. Golyandina, A. Shlemov, Semi-nonparametric singular spectrum analysis with projection, Stat. Interface 10 (2017) 47–57.
[15] N. Golyandina, A.A. Zhigljavsky, Singular Spectrum Analysis for Time Series, Springer Briefs in Statistics, 2013.
[16] T. Harris, H. Yuan, Filtering and frequency interpretations of singular spectrum analysis, Physica D 239 (2010) 1958–1967.
[17] H. Hassani, Singular spectrum analysis: methodology and comparison, J. Data Sci. 5 (2007) 259–267.
[18] H. Hassani, Z. Ghodsi, E.S. Silva, S. Heravi, From nature to maths: improving forecasting performance in subspace-based methods using genetics colonial theory, Digit. Signal Process. 51 (2016) 101–109.
[19] H. Hassani, Z. Ghodsi, E.S. Silva, S. Heravi, From nature to maths: improving forecasting performance in subspace-based methods using genetics colonial theory, Digit. Signal Process. 51 (2016) 101–109.
[20] H. Hassani, S. Heravi, A. Zhigljavsky, Forecasting UK industrial production with multivariate singular spectrum analysis, J. Forecast. 32 (2013) 395–408.
[21] H. Hassani, R. Mahmoudvand, M. Zokaei, Separability and window length in singular spectrum analysis, C. R. Math. 349 (2011) 987–990.
[22] J. Hillenbrand, L.A. Getty, M.J. Clark, K. Wheeler, Acoustic characteristics of American English vowels, J. Acoust. Soc. Am. 97 (1995) 3099–3111.
[23] G.T. Jemwa, C. Aldrich, Classification of process dynamics with Monte Carlo singular spectrum analysis, Comput. Chem. Eng. 30 (2006) 816–831.
[24] M. Kalantari, M. Yarmohammadi, H. Hassani, Singular spectrum analysis based on l1-norm, Fluct. Noise Lett. 15 (2016) 1650009.

[7] Or other method for automatic parameter selection presented in Table 1.

[25] M.A.R. Khan, D.S. Poskitt, Moment tests for window length selection in singular spectrum analysis of short and long-memory processes, J. Time Ser. Anal. 34 (2013) 141–155.
[26] A. Korobeynikov, Computation- and space-efficient implementation of SSA, Stat. Interface 3 (2010) 357–368.
[27] M.C. Leles, A.S. Vale-Cardoso, M.G. Moreira, H.N. Guimarães, C.M. Silva, A. Pitsillides, Frequency-domain characterization of singular spectrum analysis eigenvectors, in: 2016 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT, IEEE, 2016, pp. 22–27.
[28] M.C.R. Leles, H.N. Guimarães, J.P.H. Sansao, L.A. Mozelli, A new algorithm in singular spectrum analysis framework: the overlap-SSA (ov-SSA), Software X (2017), https://doi.org/10.1016/j.softx.2017.11.001, in press.
[29] M.C.R. Leles, L.A. Mozelli, H.N. Guimarães, New trend-following indicator: using SSA to design trading rules, Fluct. Noise Lett. 16 (2017) 1750016.
[30] H. Mahdi, H. Hassani, H. Taibi, Sea level in the Mediterranean Sea: seasonal adjustment and trend extraction within the framework of SSA, Earth Sci. Inform. 06 (2013) 99–111.
[31] J. Markel, A.H. Gray Jr., Linear Prediction of Speech, Springer-Verlag, 1976.
[32] V. Moskvina, A. Zhigljavsky, An algorithm based on singular spectrum analysis for change-point detection, Commun. Stat., Simul. Comput. 32 (2003) 319–352.
[33] G.P. Nason, Stationary and non-stationary time series, in: H.M. Mader, S.G. Coles, C.B. Connor, L.J. Connor (Eds.), Statistics in Volcanology, The Geological Society, 2006, pp. 129–142, chapter 11.
[34] F. Papailias, D. Thomakos, EXSSA: SSA-based reconstruction of time series via exponential smoothing of covariance eigenvalues, Int. J. Forecast. 33 (2017) 214–229.
[35] K. Pukenas, Algorithm for the characterization of the cross-correlation structure in multivariate time series, Circuits Syst. Signal Process. 33 (2014) 1289–1297.
[36] R. Rekapalli, R.K. Tiwari, Windowed SSA (singular spectral analysis) for geophysical time series analysis, J. Geol. Resour. Eng. 3 (2014) 167–173.
[37] L.J. Rodrıguez-Aragón, A. Zhigljavsky, Singular spectrum analysis for image processing, Stat. Interface 3 (2010) 419–426.
[38] S. Sanei, H. Hassani, Singular Spectrum Analysis of Biomedical Signals, CRC Press, 2015.
[39] S. Sanei, T.K. Lee, V. Abolghasemi, A new adaptive line enhancer based on singular spectrum analysis, IEEE Trans. Biomed. Eng. 59 (2012) 428–434.
[40] D.H. Schoellhamer, Singular spectrum analysis for time series with missing data, Geophys. Res. Lett. 28 (2001) 3187–3190.
[41] R. Vautard, P. Yiou, M. Ghil, Singular-spectrum analysis: a toolkit for short, noisy chaotic signals, Physica D 58 (1992) 95–126.
[42] R. Wang, H.G. Ma, G.Q. Liu, D.G. Zuo, Selection of window length for singular spectrum analysis, J. Franklin Inst. 352 (2015) 1541–1560.
[43] R. Wang, H.G. Ma, J.Q. Qin, X.W. Feng, H.M. Liu, Analysis of death series by SSA based BSS technique, in: 2015 10th International Conference on Information, Communications and Signal Processing, ICICS, IEEE, 2015, pp. 1–5.
[44] P. Yiou, D. Sornette, M. Ghil, Data-adaptive wavelets and multi-scale singular-spectrum analysis, Phys. D, Nonlinear Phenom. 142 (2000) 254–290.

Michel Carlo Rodrigues Leles is an Associate Professor in the Department of Telecommunication and Mechatronics Engineering at Federal University of São João del-Rei, since 2010. As of December 2017, he is a visiting researcher at the Aeronautics Institute of Technology in the Electronics Engineering Division. His research work focuses on signal processing, time series analysis, and computational finance.

João Pedro Hallack Sansão is an Associate Professor in the Department of Telecommunication and Mechatronics Engineering at Federal University of São João del-Rei, since 2011. His research interests are disordered voice signal analysis and image processing.

Leonardo Amaral Mozelli is an Associate Professor in the Department of Electronics Engineering at Federal University of Minas Gerais, since 2017. Prior to that, he worked at Federal University of São João del-Rei, as an Associate Professor, from 2010 to 2016. He has published more than 40 scientific papers over the past decade and his current research interests include: signal processing, control theory, robotics, and technologies for sustainable development.

Homero Nogueira Guimarães received his undergraduate degree in Electrical Engineering and his Ph.D. degree in Physiology from Federal University of Minas Gerais (UFMG), Brazil, in 1991 and 1996, respectively. Since 1998, he is an Associate Professor in the Department of Electrical Engineering at UFMG. His research interests include signal processing (in particular, spectral analysis and singular spectrum analysis), trading systems, biomedical signal processing (in particular, heart rate variability and cardiotoxicity evaluation of drugs) and music composition. He is also proud to be a long-distance runner and swimmer.