ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

PHYSICA A

www.elsevier.com/locate/physa

Statistical method for sparse coding of speech including a linear predictive model

Hugo L. Rufiner[a,1,*], John Goddard[b,2], Luis F. Rocha[c,3], María E. Torres[a,1]

[a]*Facultad de Ingeniería, Univ. Nac. de Entre Ríos, Ruta Prov. 11, Km 10, E3100XAD, Oro Verde, Entre Ríos, Argentina*
[b]*Depto Ing. Eléctrica, Iztapalapa, Univ. Autónoma Metropolitana, México*
[c]*Fac. Ingeniería, Univ. de Buenos Aires, Argentina*

Received 21 November 2005

Abstract

Recently, different methods for obtaining sparse representations of a signal using dictionaries of waveforms have been studied. They are often motivated by the way the brain seems to process certain sensory signals. Algorithms have been developed using a specific criterion to choose the waveforms occurring in the representation. The waveforms are choosen from a fixed dictionary and some algorithms also construct them as a part of the method. In the case of speech signals, most approaches do not take into consideration the important temporal correlations that are exhibited. It is known that these correlations are well approximated by linear models. Incorporating this a priori knowledge of the signal can facilitate the search for a suitable representation solution and also can help with its interpretation. Lewicki proposed a method to solve the noisy and overcomplete independent component analysis problem. In the present paper we propose a modification of this statistical technique for obtaining a sparse representation using a generative parametric model. The representations obtained with the method proposed here and other techniques are applied to artificial data and real speech signals, and compared using different coding costs and sparsity measures. The results show that the proposed method achieves more efficient representations of these signals compared to the others. A qualitative analysis of these results is also presented, which suggests that the restriction imposed by the parametric model is helpful in discovering meaningful characteristics of the signals.

Keywords: Speech analysis and representation; Sparse coding; Linear predictive coding; Basis pursuit; Matching pursuit; Independent component analysis

*Corresponding author. Tel.: +54 343 4975078, +54 343 4975077x126; fax: +54 343 4975100, +54 343 4975101x105.
E-mail address: lrufiner@bioingenieria.edu.ar (H.L. Rufiner).

[1]Supported by Universidad Nacional de Entre Ríos (UNER) and Agencia Nacional de Promoción Científica y Tecnológica, Argentina, under Project PICT-2002 11-12700.
[2]Supported by CONACYT, Mexico under Project 31929-A.
[3]Supported by Facultad de Ingeniería, UBA, Argentina.

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.11.029

1. Introduction

Speech signals are amongst the most studied natural signals. In the field of speech signal analysis and modeling, important advances have been made concerning their representation, such as linear predictive coding [1], cepstral and mel frequency cepstral coefficients (MFCC) with delta and acceleration coefficients [2], and RASTA-PLP [3]. However, there are several problems which have not been satisfactorily solved. Machines are far from human performance in certain speech analysis and recognition tasks [4]. Machine performance degrades rapidly when faced with background noise, variations between different speakers, and even changes in the speaking rate of a single speaker. Humans, by comparison, are able to overcome these difficulties with apparent ease.

In the last few years several researchers have taken a different approach to traditional signal processing. These new formulations give rise to techniques based on non-linear systems and higher-order statistics, including independent component analysis (ICA) and methods to obtain sparse representations (SR) of a signal. They provide new ways of phrasing the solution of the problem of signal modeling or representation. One underlying idea is that of representing the signals involved using only a few significant characteristics; that is, as a sparse representation of only a few basic waveforms. Sometimes the waveforms are specified from the outset, and sometimes they are also found as part of the method. Super-resolution is an important property discussed in Ref. [5], where examples are given showing its advantages compared to traditional methods. Because of the intrinsic robustness of sparse representations, denoising techniques can be directly integrated into the process [6]. There are applications of these techniques to different fields, such as: natural image analysis [7,8], audio and music signals [9], general biomedical signals [10–12] and automatic speech recognition (ASR) [13,14].

Many of these new formulations are motivated by biological considerations related to the way natural images and sounds are coded within the brain. It is known that the neural code itself, which is based on spike trains, is sparse [15]. Several works use this principle as a model for the representations generated in the receptive fields of the visual and auditory cortex. It has been suggested that, using this principle, the sensory systems have been adapted in order to process the signals of their environment in an optimal and efficient way [16].

Some of the most commonly used speech representations nowadays, such as MFCC or RASTA-PLP, incorporated biologically inspired characteristics and have provided significant improvements to the performance of artificial systems. Other aspects, such as those mentioned in the previous paragraph, have not yet been taken into account even if they might provide better solutions to the speech representation problem.

In the present paper, a new method for the SR of speech signals is proposed using a generative parametric model. The main idea is to consider only waveforms derived from linear predictive models, obtaining in this way the explicit inclusion of the temporal correlations between the samples.

The new method proposed here is discussed in Section 3, and the other methods used in the paper are sketched in Section 2. Sparseness can be defined in different ways, and the tests employed for judging representational efficiency are also presented in Section 4. The proposed method is then compared to other techniques on artificial data and real speech signals and the results are given in Section 5. Finally, Section 6 provides some conclusions concerning the paper.

2. Sparse representation of signals

Given a signal $\mathbf{s} \in \mathbb{R}^N$, we consider a representation in terms of a dictionary $\mathbf{\Phi}$ as a decomposition of the form:

$$\mathbf{s} = \sum_{j=1}^{M} \boldsymbol{\phi}_j a_j = \mathbf{\Phi}\mathbf{a}, \tag{1}$$

where $\mathbf{a} \in \mathbb{R}^M$ is the coefficients vector, $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$, with $M \geqslant N$, is a collection of waveforms or *atoms* $\boldsymbol{\phi}_j$ and both ($\mathbf{a}$ and $\mathbf{\Phi}$) usually unknown.

Some authors use the term "basis" instead of "dictionary"; however, as the set of atoms may not be linearly independent, the latter is to be preferred.

When there are more waveforms in the dictionary than samples $\mathbf{s}$, i.e., $M > N$ (referred to as an overcomplete dictionary), or when the waveforms do not form a basis, then there will be non-unique representations of the signal. In this situation a suitable criterion is required to select only one of them. In this context, sparseness often refers to the criterion of choosing a representation with "as few non-zero coefficients as possible" (typically using the ℓ_0 norm), although several other criteria have been introduced (c.f. Ref. [17]).

The problem of a sparse representation of $\mathbf{s}$ with respect to ℓ_0 could be stated as follows:

$$\min \|\mathbf{a}\|_0 \text{ subject to } \mathbf{\Phi a} = \mathbf{s}. \tag{2}$$

It is important to note that in the overcomplete case mentioned above, although Eq. (1) is linear, the coefficients a_γ chosen as the solution correspond to a non-linear function of the data $\mathbf{s}$: $\mathbf{s} \to \{a_\gamma\}$.

In order to solve the SR problem, it can be split into two sub-problems: inference and learning. The first one consists in finding the representation coefficients $\mathbf{a}$ which satisfy a given sparsity criterion. The second one involves finding the optimal dictionary $\mathbf{\Phi}$ to represent the data. The last one is usually the most complex of both.

Given a set of scalar or multi-index parameters Γ, sometimes it is useful to employ atoms of parameterized waveforms $(\boldsymbol{\phi}_\gamma)_{\gamma\in\Gamma}$. Well-known fixed dictionaries of parametrized waveforms are the traditional Fourier sinusoids (frequency dictionaries), Dirac functions, wavelets (time-scale dictionaries), Gabor functions (time–frequency dictionaries), polynomials or combinations of them. Different methods have been proposed for obtaining a SR from a fixed dictionary (inference problem solution): basis pursuit (BP) [5], matching pursuit (MP) [18] and best orthogonal basis (BOB) [19]. The method of frames (MOF) [20] also gives a representation, but usually it is not sparse. In Section 2.1 some of these methods are briefly reviewed for comparative purposes.

A first attempt to apply this type of representation to speech signals using a fixed dictionary of wavelet packets appeared in Ref. [21], giving promising results in terms of an accurate localization of spectro-temporal acoustic phonetic cues with very few coefficients, even in the presence of some noise. Taking advantage of this property, a simple heuristic denoising method was introduced in Ref. [22].

Other methods to solve the problem of sparse representation—often of a statistical nature (c.f. Ref. [23])—additionally construct the waveforms appearing in (1), providing the solution to the learning problem. In this case the coefficients are assumed to be statistically independent. The sparsity of the representation can be achieved appropriately choosing the prior probability distribution of the coefficients (e.g. Laplacian). This approach has important connections to higher order statistics and ICA [24,25]. It is interesting to note that whilst sparsity and independence are different criteria, they can often produce similar solutions (c.f. Ref. [26]). Additionally, sparse codes generally have low entropy, which leads them to be optimal from an information theory point of view [27].

The statistical methods used to build the optimum dictionaries will be reviewed in Section 2.2. They will be the basis for developing the alternative method proposed in this work. An algorithm is introduced which includes restrictions related to the temporal structure of speech signals, assuming that the elements of the dictionary are represented by the impulse responses of autoregressive (AR) filters, which have been extensively and successfully applied to speech. A preliminary evaluation of this parametric approach was presented in Ref. [28]. In Section 3, the ideas behind this new algorithm are developed.

Once the representations are found, different tests can be applied to estimate the sparsity and coding cost of them. This will allow us to compare quantitatively the different methods used to obtain the representations. However, a qualitative analysis of the found dictionaries is also useful in order to test if they are able to find meaningful characteristics of the signals. These concepts are presented and discussed in Section 4.

2.1. Representation based on fixed dictionaries

In this section we present some methods to solve the inference problem based on dictionaries fixed in advance. So, the goal is here to find the representation coefficients $\mathbf{a}$ which satisfy a given sparsity criterion. In this case $\mathbf{\Phi} = \{\boldsymbol{\phi}_\gamma\}_{\gamma\in\Gamma}$ is always known.

2.1.1. Basis pursuit

The sparse representation problem in (2) is NP-hard.[4] As an alternative for obtaining a SR of $\mathbf{s}$ in (1), Chen proposed the BP method [5]. They phrase the problem of finding a suitable representation as one of optimization with respect to the ℓ_1 norm (which could be assumed as an approximation of ℓ_0). More precisely, the problem to solve is

$$\min \|\mathbf{a}\|_1 \text{ subject to } \mathbf{\Phi a} = \mathbf{s}. \tag{3}$$

This problem can be converted to a standard linear program problem and can be solved efficiently and exactly with interior point methods [5].

2.1.2. Matching pursuit

Mallat and Zhang [18] proposed a general method to approximate the solution to problem (2). Sparsity is directly included by choosing an appropriate number of terms. Given an initial approximation $\mathbf{s}^{(0)} = \mathbf{0}$ and an initial residual $\mathbf{R}^{(0)} = \mathbf{s}$, a sequence of approximations is iteratively constructed. At step k the parameter $\gamma = \hat{\gamma}$ is selected, such that the atom $\boldsymbol{\phi}_{\hat{\gamma}}^{(k)}$ best correlates with the residual $\mathbf{R}^{(k)}$, and a scalar multiple of this atom is added to the approximation at step $k-1$, obtaining:

$$\mathbf{s}^{(k)} = \mathbf{s}^{(k-1)} + a_{\hat{\gamma}}^{(k)} \boldsymbol{\phi}_{\hat{\gamma}}^{(k)}, \tag{4}$$

where $a_{\hat{\gamma}}^{(k)} = \langle \mathbf{R}^{(k-1)}, \boldsymbol{\phi}_{\hat{\gamma}}^{(k)} \rangle$ and $\mathbf{R}^{(k)} = \mathbf{s} - \mathbf{s}^{(k)}$. After m steps an approximation to (1) is obtained, with residue $\mathbf{R} = \mathbf{R}^{(m)}$. It is said that MP constitutes a greedy solution to the SR problem, thereof it has the same advantages and disadvantages of these type of optimization methods.[5]

As was mentioned before, other methods exist which also provide a suitable dictionary as part of the solution. This framework allows the inclusion of a particular model for the atoms of the dictionary and will be described in the following section. They are named as optimal dictionaries because in order to find the atoms a data-driven optimization problem has to be solved.

2.2. Representation based on optimal dictionaries

In this section a more general framework is assumed, where (1) is rewritten to include an additive Gaussian noise term $\boldsymbol{\varepsilon}$ as follows:

$$\mathbf{s} = \mathbf{\Phi a} + \boldsymbol{\varepsilon}. \tag{5}$$

Following terminology used in ICA, (5) is referred to as the generative model, to signify that one generates the signal $\mathbf{s} \in \mathbb{R}^N$ from a set of hidden sources a_j, arranged as a state vector $\mathbf{a} \in \mathbb{R}^M$, using a mixing matrix or dictionary $\mathbf{\Phi}$ of size $N \times M$, with $M \geqslant N$.

The a_j are initially assumed to be statistically independent, with a joint a priori distribution:

$$P(\mathbf{a}) = \prod_{j=1}^{M} P(a_j). \tag{6}$$

If $\mathbf{\Phi}$ is known and $\mathbf{s}$ is given, the state vector $\mathbf{a}$ can be estimated via the Bayes's rule by considering the posterior distribution:

$$P(\mathbf{a}|\mathbf{\Phi}, \mathbf{s}) = \frac{P(\mathbf{s}|\mathbf{\Phi}, \mathbf{a}) P(\mathbf{a})}{P(\mathbf{s}|\mathbf{\Phi})}. \tag{7}$$

A maximum a posterior (MAP) estimation of $\mathbf{a}$ reads as follows:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} [\log P(\mathbf{s}|\mathbf{\Phi}, \mathbf{a}) + \log P(\mathbf{a})]. \tag{8}$$

[4] In Computational complexity theory a non-deterministic polynomial-time hard problem.
[5] Greedily minimizes $\|\mathbf{s} - \mathbf{\Phi a}\|_2$.

If the posterior is sufficiently smooth, the maximum can be found applying gradient ascent. The solution depends on the form of the distribution chosen for the noise term $\boldsymbol{\varepsilon}$ and sources a_j, giving rise to different methods for finding the coefficients. Lewicki and Olshausen [6] proposed an a priori distribution of Laplacian type:

$$P(a_j) = N\mathrm{e}^{-\rho_j|a_j|}, \tag{9}$$

where ρ_j is given and, if the noise is Gaussian, this leads to the following rule for updating $\mathbf{a}$:

$$\Delta\mathbf{a} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}_{\varepsilon}\,\boldsymbol{\varepsilon} - \boldsymbol{\rho}^{\mathrm{T}}|\mathbf{a}|, \tag{10}$$

where $\boldsymbol{\Lambda}_{\varepsilon}$ is the inverse of the noise covariance matrix $\mathscr{E}[\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}]$, with $\mathscr{E}[\cdot]$ denoting the expected value, and $\boldsymbol{\rho} = \{\rho_j\}$.

Until now, a statistical method has been used to solve the inference problem. In what follows the learning problem is similarly solved. To estimate the value of $\boldsymbol{\Phi}$, the following objective function can be maximized [6]:

$$\widehat{\boldsymbol{\Phi}} = \arg\max_{\boldsymbol{\Phi}}[\mathscr{L}(\mathbf{s},\boldsymbol{\Phi})], \tag{11}$$

where $\mathscr{L} = \mathscr{E}[\log P(\mathbf{s}|\boldsymbol{\Phi})]_{P(\mathbf{s})}$ is the likelihood of the data. This likelihood can be found by marginalizing the following product of the conditional distribution of the data—given the dictionary—and the coefficients, together with the coefficients a priori distribution:

$$P(\mathbf{s}|\boldsymbol{\Phi}) = \int_{\mathbb{R}^M} P(\mathbf{s}|\boldsymbol{\Phi},\mathbf{a})P(\mathbf{a})\,\mathrm{d}\mathbf{a}, \tag{12}$$

where the integral is over the M-dimensional state space of $\mathbf{a}$.

The objective function (11) can be maximized using gradient ascent with the following update rule for the matrix $\boldsymbol{\Phi}$ [29]:

$$\Delta\boldsymbol{\Phi} = \eta\boldsymbol{\Lambda}_{\varepsilon}\mathscr{E}[\boldsymbol{\varepsilon}\mathbf{a}^{\mathrm{T}}]_{P(\mathbf{a}|\boldsymbol{\Phi},\mathbf{s})}, \tag{13}$$

where η is a learning coefficient (between 0 and 1). The problem at this point is how to calculate this update rule, given that it involves solving the following integral:

$$\mathscr{E}[\boldsymbol{\varepsilon}\mathbf{a}^{\mathrm{T}}]_{P(\mathbf{a}|\boldsymbol{\Phi},\mathbf{s})} = \int_{\mathbb{R}^M} (\mathbf{s} - \boldsymbol{\Phi}\mathbf{a})\,\mathbf{a}^{\mathrm{T}}P(\mathbf{a}|\boldsymbol{\Phi},\mathbf{s})\,\mathrm{d}\mathbf{a}. \tag{14}$$

As the dimension of $\mathbf{a}$ increases, the previous integral becomes analytically intractable and different authors have proposed approximation methods in order to compute it. Lewicki and Sejnowski [30] used a multivariate Gaussian approximation to the posterior distribution around its maximum $\hat{\mathbf{a}}$:

$$P(\mathbf{a}|\boldsymbol{\Phi},\mathbf{s}) \approx \sqrt{\frac{|\mathbf{H}|}{(2\pi)^M}}\,\mathrm{e}^{-1/2(\mathbf{a}-\hat{\mathbf{a}})^{\mathrm{T}}\mathbf{H}(\mathbf{a}-\hat{\mathbf{a}})}, \tag{15}$$

where $\hat{\mathbf{a}}$ is the mean value, and $\mathbf{H}^{-1}$ is the covariance, being $\mathbf{H}$ the Hessian of the log-posterior evaluated in $\hat{\mathbf{a}}$:

$$\mathbf{H} = -\nabla\nabla^{\mathrm{T}}\log P(\mathbf{a}|\Phi,\mathbf{s}). \tag{16}$$

It provides a good approximation for $\mathbf{a}$ close to $\hat{\mathbf{a}}$. This result provides a solution of (13) given by

$$\Delta\boldsymbol{\Phi} = \eta\boldsymbol{\Lambda}_{\varepsilon}(\hat{\boldsymbol{\varepsilon}}\hat{\mathbf{a}}^{\mathrm{T}} - \boldsymbol{\Phi}\mathbf{H}^{-1}), \tag{17}$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{s} - \boldsymbol{\Phi}\hat{\mathbf{a}}$.

In order to obtain the dictionary and the coefficients (Eqs. (10) and (17)), in this paper we use the implementation proposed by Lewicki and Olshausen [6] (using Lewicki's noise overcomplete ICA code or NOCICA), where computational details can be found. In the following section this method will provide the basis for introducing additional temporal statistical information in order to obtain a better estimate of $\boldsymbol{\Phi}$ in the case of speech signals.

3. Representation based on optimal dictionaries including linear predictive models

The method described in Section 2.2 finds an optimal dictionary for a general class of signals. However, one problem in modeling speech using this framework is that it ignores all the information concerning the temporal correlation between samples present in this type of signals.

In order to take advantage of this temporal correlation, we consider the atoms $\boldsymbol{\phi}_j$ associated with the characteristic states of a linear model of the vocal tract for different phonemes. Then, a particular speech signal can be obtained by "adding" only the most important characteristics together. A good approximation to this behavior will be achieved when coefficients have an a priori Laplacian distribution given by Eq. (9). To take advantage of temporal correlation, we approximate the waveforms used for the dictionary $\boldsymbol{\Phi}$ in (5) by

$$\hat{\phi}_{i,j} = -\sum_{q=1}^{Q} \phi_{i-q,j}\, c_{q,j} + \delta_i \mathrm{g}_j, \tag{18}$$

where the time index moves forward in the direction of the rows i of $\boldsymbol{\Phi}$, i.e., along each column or atom j, $c_{q,j}$ are the linear predictor coefficients and g_j are the corresponding gain coefficients. Observe that $\hat{\phi}_{i,j}$ corresponds to $\hat{\phi}_j[i]$ in the classical notation for time series. In a similar fashion $\phi_{i-q,j}$ should be seen as $\phi_j[i-q]$ and δ_i is the delta sequence $\delta[i] = [1, 0, ..., 0]^{\mathrm{T}}$ for all time indices i.

In this way we impose a new restriction on the optimization problem (11) that allows the explicit inclusion of the temporal correlation of the samples of each atom by using the coefficients $c_{q,j}$. This means that the problem to be solved can be phrased as one of overcomplete ICA with noise and certain restrictions on the mixing matrix. These restrictions include the approximation of this matrix's columns by a linear prediction model. We denominate this new method linear prediction ICA (LP-ICA). This framework represents a particular case of convolutive mixtures in the time domain and can be formulated in the z domain of the original variables. In the latter case the convolution becomes a product and the $\boldsymbol{\phi}_j$'s can be expressed as

$$\boldsymbol{\Phi}_j(z) = \frac{\mathrm{g}_j}{C_j(z)}, \tag{19}$$

where $C_j(z) = 1 + \sum_q c_{q,j}\, z^{-q}$. This is obtained applying the z transform to Eq. (18). The corresponding generative model is shown in Fig. 1.

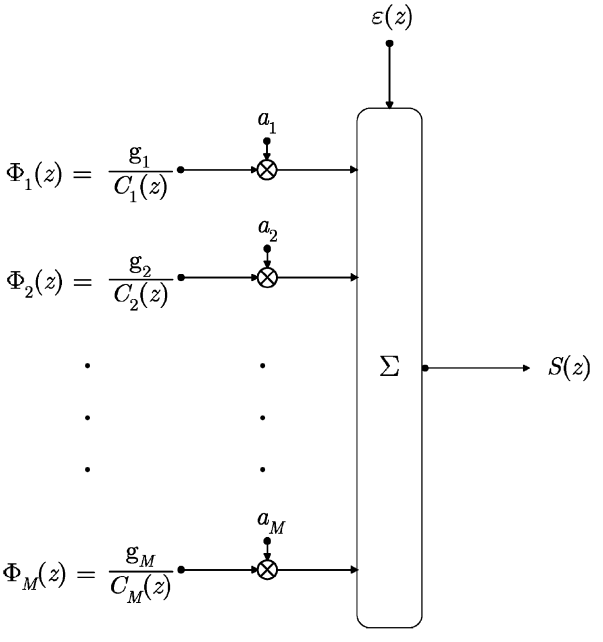

Fig. 1. Generative model for the speech signals in the z domain. This is a particular case of convolutive mixtures.

In order to solve this parametric ICA problem, it is necessary to find the representation coefficients, the waveforms and the parametric approximation. These issues can be handled separately. The approach taken in the present paper for finding the coefficients a_j and parametrically modeled waveforms $\phi_{i,j}$ is to use the techniques described in the previous section [6,30], including a parametric approximation step [28]. This is done iteratively.

At each iteration k the matrix $\mathbf{\Phi}^{(k)}$ has to simultaneously satisfy the restrictions imposed on the columns by (18) and the maximization of the likelihood in (11). Once $\mathbf{\Phi}^{(k)}$ is estimated, the coefficients $c_{q,j}^{(k)}$ can be computed using the usual stationary hypothesis by

$$\frac{\partial \mathscr{E}[\|\boldsymbol{\phi}_j^{(k)} - \hat{\boldsymbol{\phi}}_j^{(k)}\|_2]}{\partial c_{q,j}^{(k)}} = 0, \tag{20}$$

i.e., minimizing the mean square error (MSE) between $\boldsymbol{\phi}_j^{(k)}$ and $\hat{\boldsymbol{\phi}}_j^{(k)}$ approximated using (18), which are the j-column vectors $\{\phi_{i,j}^{(k)}\}$ and $\{\hat{\phi}_{i,j}^{(k)}\}$, respectively.

In order to solve (20), instead of the autocorrelation method we used in Ref. [28], here we adopt Prony's method [31]. This method has the ability to recover the impulse response that better matches a given sequence, and it improves the performance of the former method. Once the linear predictor coefficients are obtained, $\hat{\mathbf{\Phi}}^{(k)}$ is evaluated with (18). At this point $\mathbf{\Phi}^{(k)}$ should be replaced by its parametric version $\hat{\mathbf{\Phi}}^{(k)}$. To make sure that this change is not too disruptive in the first steps of the algorithm, the complexity of the model is gradually diminished using the order Q if $\log P(\mathbf{s}|\mathbf{\Phi}^{(k)})$ increases. Moreover, if the approximation exceeds an error threshold for some atom $\boldsymbol{\phi}_j^{(k)}$, it remains unchanged. In this way, when iterations are finished, we obtain an estimated dictionary $\widehat{\mathbf{\Phi}}$.

Summarizing, the solution of the problem can be described in terms of the following LP-ICA algorithm:

Initialize $k = 0$ and $\mathbf{\Phi}^{(0)}$ randomly
Initialize order of parametric approximation $Q = Q_{ini}$
REPEAT
  $k = k + 1, l = 0$
  Initialize $\mathbf{a}^{(k,0)}$ with pseudoinverse solution of (1): $\mathbf{\Phi}^{(k-1)\dagger}\mathbf{s}$
  REPEAT
    $l = l + 1$
    Compute $\Delta\mathbf{a}^{(k,l)}$ using (10)
    $\mathbf{a}^{(k,l)} = \mathbf{a}^{(k,l-1)} + \Delta\mathbf{a}^{(k,l)}$
  UNTIL termination conditions
  Compute $\Delta\mathbf{\Phi}^{(k)}$ using (17)
  $\mathbf{\Phi}^{(k)} = \mathbf{\Phi}^{(k-1)} + \Delta\mathbf{\Phi}^{(k)}$
  Compute $c_{q,j}^{(k)}$ via (20)
  Compute $g_j^{(k)}$ equalizing energy of $\boldsymbol{\phi}_j^{(k)}$ and $\hat{\boldsymbol{\phi}}_j^{(k)}$
  Compute $\hat{\mathbf{\Phi}}^{(k)}$ using (18)
  If $\mathrm{MSE}(\boldsymbol{\phi}_j^{(k)}, \hat{\boldsymbol{\phi}}_j^{(k)}) >= \vartheta$ THEN $\hat{\boldsymbol{\phi}}_j^{(k)} = \boldsymbol{\phi}_j^{(k)}$
  If $|\log P(\mathbf{s}|\hat{\mathbf{\Phi}}^{(k)}) - \log P(\mathbf{s}|\mathbf{\Phi}^{(k)})| < \zeta$ THEN $\mathbf{\Phi}^{(k)} = \hat{\mathbf{\Phi}}^{(k)}$, ELSE $Q = Q - 1$
  If $Q < Q_{min}$ THEN $Q = Q_{min}$
UNTIL termination conditions

where ϑ and ζ are predefined constants that control the speed and degree of the parametric approximation. These thresholds are empirically set, adjusted for each data base in order to assure the algorithms convergence. Furthermore, there are predefined constants Q_{ini} and Q_{min} that fix the initial and the minimum allowed value for Q. Q_{ini} was set to a high value ($Q_{ini} \simeq N/2$, with N the signal length), and for Q_{min} low values have been selected ($Q_{min} = 4$ or 6). Termination conditions are based on a predetermined number of iterations.[6]

[6]In the algorithm $\mathbf{s}$ stands for a randomly selected subset from all the signal data.

4. Tests to estimate sparsity and coding costs

It is important to establish some criteria to evaluate the obtained representation. As a "good model" generally is equivalent to a "good representation" of the signals involved, the evaluation of the generative model proposed in this work is an "indirect" one.

There are several ways to measure the representational effectiveness; this means to measure how good the coefficients **a** code the data **s** using the specified generative model. Among the available methods there are essentially two groups: those related to the dispersion of the coefficients with respect to a norm, and those derived from the coefficients statistics. To estimate the sparsity and coding costs of a representation, in the present paper five of the most common ones are employed. From the first group the norms ℓ_0, *minvol* and ℓ_1 have been selected, and the kurtosis $\mathscr{K}$, the entropy $\mathscr{H}$, and the number of bits *#bits* have been chosen from the second group of methods. As a measure of the reconstruction capability of the method proposed here, the value of the averaged MSE over the patterns has been evaluated. The tests used in the paper are briefly discussed in the following two subsections.

4.1. Sparsity measures

An obvious measure is simply to count the number of non-zero terms. This is exactly what the zero norm ℓ_0 does [32]:

$$\|\mathbf{a}\|_0 = \#\{j : a_j \neq 0\}.$$

However, this measure is highly sensitive to small perturbations of zero elements. To ameliorate the strict effect of the zero norm, the minimum volume norm [17] was introduced. This norm is defined as

$$minvol(\mathbf{a}) = \sum_{j=1}^{M} \frac{a_j^2}{a_j^2 + v},$$

where $0 < v < \min_j(a_j)$ is a small number of machine precision order. This gives an approximation to ℓ_0 when a small threshold is used to decide if a coefficient is considered non-zero, and it possesses better properties.

The ℓ_1-norm is often used as a practical approximation to the ℓ_0-norm in optimization problems (c.f. Refs. [5,32]). Minimizing with respect to it is the same as requiring that the coefficients have a Laplacian prior probability, which is the case here.

Guspí and Introcaso [17] analyze other possibilities and their use in finding sparse solutions for general indeterminant linear systems is similar to (5).

4.2. Coding costs and other useful measures

One way to quantify the dispersion from the statistical viewpoint is to use the 4th order moment or kurtosis. If the a_j are considered as r.v. then

$$\mathscr{K}(a_j) = \frac{\mathscr{E}[a_j^4]}{\mathscr{E}[a_j^2]^2} - 3.$$

The kurtosis $\mathscr{K}$ is not a norm but it provides a good measure of dispersion for symmetric unimodal distributions. The kurtosis generally increases when the entropy decreases, that is why sometimes it can be related to statistical independence, although this interpretation requires care [27]. It is often also used as a measure of "non-Gaussianity". If its value is positive one speaks of super-Gaussians, and sub-Gaussians if it is negative. The principal problem is that it is sensitive to outliers. In the present paper, the average of the kurtoses is calculated for each dimension of the vector **a**.

The entropy $\mathscr{H}$ is often used to measure the coding efficiency of **s** in terms of **a**:

$$\mathscr{H}(a_j) = -\sum_i p_i(a_j) \log p_i(a_j).$$

It has more to do with statistical independence (c.f. Ref. [27]) than sparsity. On minimizing the sum of the $\mathscr{H}(a_j)$ the mutual information between the coefficients a_j is destroyed, which, with suitable restrictions, gives statistically independent coefficients [27]. Even though it is a different criterion, it is also desirable to obtain low entropies for the sparse codes, and both criteria can be applied with good results [33]. One has to be careful when using $\mathscr{H}$ for estimating the coding cost given that low entropy can be obtained even though the data are not well represented by the dictionary. However, if the dictionary does generate the data, then $\mathscr{H}$ is a reasonable coding cost measure [30]. This can be known by controlling the approximation error. In the present paper the average of the entropies is calculated for each dimension of $\mathbf{a}$ (with the estimator version proposed in Ref. [34]).

Another important measure is given by the coding cost. $\mathscr{H}$ can be used to calculate it by means of the smallest number of bits required for coding the patterns (Shannon's theorem):

$$\#bits \geqslant \mathscr{H}_{bits}(a_j) = -\sum_i p_i(a_j)\log_2 p_i(a_j). \tag{21}$$

The MSE of the reconstructed signal averaged by all the patterns gives an idea of the required fitting capability:

$$\mathrm{MSE}(\mathbf{s}-\mathbf{\Phi a}) = \langle \|\mathbf{s}-\mathbf{\Phi a}\|_2 \rangle.$$

Its value is bounded below by the noise variance σ_ε in our generative model (5).

4.3. Methods employed for qualitative analysis of the dictionaries

In order to perform a qualitative analysis of the obtained dictionaries, in Section 5 we will use two methods that we briefly describe here.

To analyze and compare the dictionary atoms we use their spectrograms. In this case a compromise between the width of the time window and the overlap is accomplished in order to adequately identify events in both time and frequency. The spectrograms are ordered with a one-dimensional self-organizing Kohonen map in such a way that those which appear to be more similar are also closer together. Finally, some of the intermediate atoms are eliminated in order to show only the most important ones.

To obtain a global time–frequency view of the obtained dictionaries, we use a time–frequency ($T-F$) "tile covering". To create the corresponding figures a method with ellipses, similar to that described in Ref. [16], is used. The temporal extent of each atom is measured using the width required to cover 95% of their power. Frequency width is measured using the spectral bandwith at 10 dB down from the peak. Atoms that are not localized (where the main spectral peak account for less than 50% of the total power) are omitted from the plot.

5. Results and discussion

In this work two kinds of experiments have been performed: one using artificial data, and the other with real speech data. In the first case, the synthesis process or "direct problem" has been controlled using the generative model so that the solution of the "inverse problem" would be known in advance. For the real speech data, the representations and the different dictionaries obtained for the different types of phonemes have been compared with the main characteristics of each phonetic class.

5.1. Artificial data

The proposed parametric method LP-ICA and NOCICA method described in Sections 2 and 3 were applied to artificial data and the tests described in Section 4 were evaluated.

The artificial data set was generated from the parametric version of the generative model (5). The parameters have been set so that the generated data closely resembles samples taken from vowel phonemes, in accordance with the interpretation of the generative model as a speech synthesizer. Therefore, the dictionary elements were chosen as functions with 2 poles in the z domain. For the election of the $c_{q,j}$ coefficients,

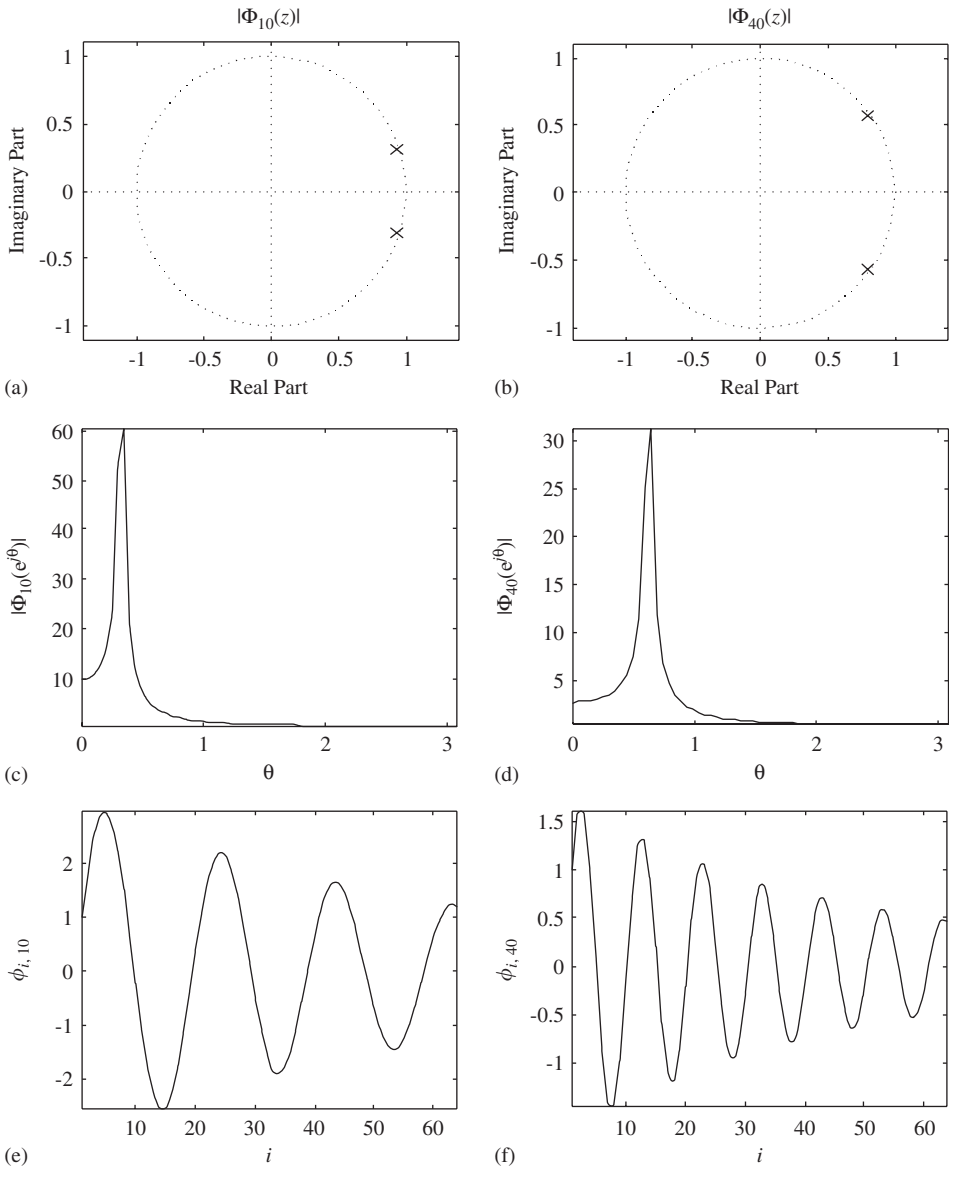

Fig. 2. Pole-zero diagrams (a, b), spectra (c, d) and temporal signals (e, f) for two atoms of the dictionaries used to generate the artificial data with the generative model.

frequency values taken from the first two formants of the five Spanish vowels, pronounced in an isolated and sustained fashion by different speakers [35], were used. In this way, the atoms constitute damped sinusoids with frequencies which are equivalent to the characteristic resonances of the vocal tract for the production of these vowels. The sampling frequency used was 8000 Hz. The case considered at this stage consisted of 64 atoms of 64 samples each (64×64). With this assembled dictionary, coefficients were generated with independent Laplacian distributions, and atoms were mixed using (5), producing the set of data or signals for the experiments (a total of 1000 frames with 64 samples each). A small amount of noise was added, with a Gaussian distribution and zero mean (SNR = 80 dB).[7] Examples of both, the atoms and the generated signals, can be seen in Figs. 2 and 3, respectively.

[7]This SNR was used in this work since we were not considering the robustness of the representation.

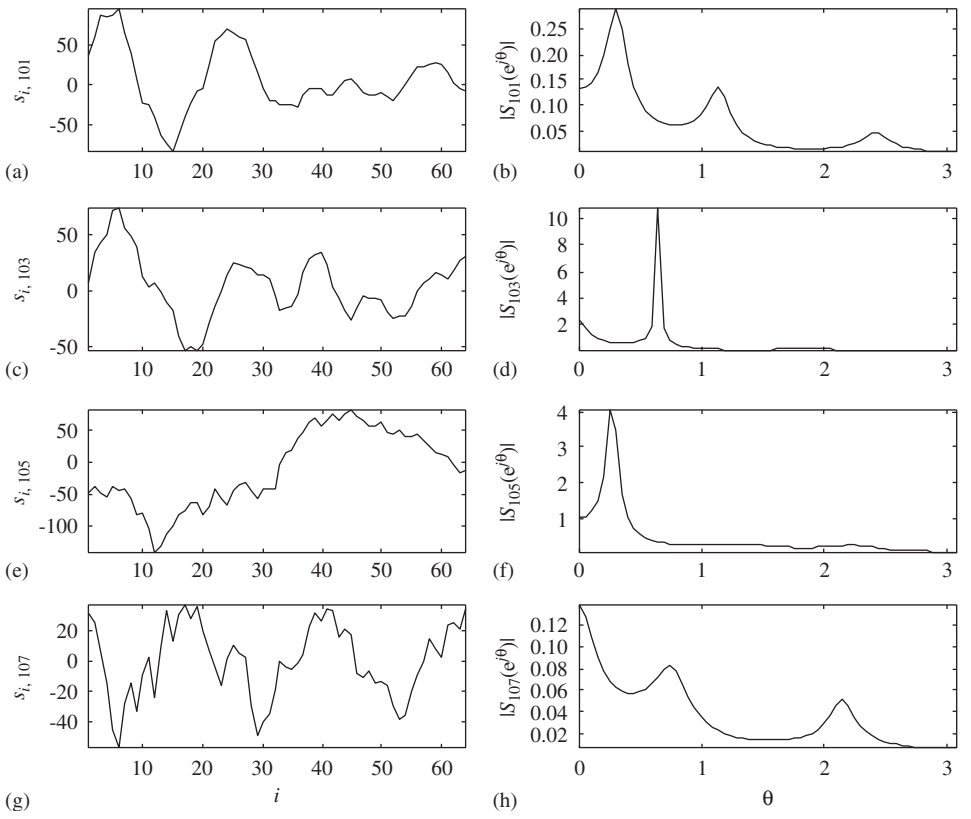

Fig. 3. Signal examples (a, c, e, g) and their corresponding spectra (b, d, f, h) generated using the dictionary of the previous figure.

Table 1
Sparsity and coding costs obtained from different representations of the artificial data with the dictionary fixed in advance

Representation	ℓ_0	*minvol*	$\mathscr{K}$	*#bits*	MSE (**s**, **Φa**)
Original	0.45	0.67	2.89	3.02	2.28E − 03
NOCICA (Eq. (10))	0.23	0.74	0.98	3.32	1.26E − 03
BP	0.05	0.51	62.55	1.64	4.68E − 02
MP	0.28	0.44	8.30	2.38	2.90E − 03
DCT	0.21	0.53	0.88	3.37	0.00E + 00
DWT	0.47	0.92	0.39	3.53	0.00E + 00

With the generated data certain experiments were conducted using the methods described in the previous sections. The obtained results are displayed in Table 1, and also compared to those corresponding to cosine and wavelets bases. In this table "original" makes reference to the fact that coefficients and the dictionary used correspond to the artificial data set already generated. NOCICA, BP and MP mean that the original dictionary was used, but coefficients were calculated from the artificial data and with these methods. DCT and DWT indicate that artificial data were used to calculate the representation in terms of the most traditional transforms such as discrete cosine and dyadic wavelets transform (with Symmlet 8 mother wavelet), respectively. From this table we observe that among the traditional transformations, the least sparse representation is given by the DWT (higher values of ℓ_0 and *minvol*). This is due to the fact that the elements of the basis are quite different from the atoms used to generate the data (thereby requiring the use of many elements in the representation of the signal). It can also be noticed that it requires the largest number of bits to

code the coefficients. Observe that DCT displays a higher sparsity, this is due to the fact that the atoms are quite similar to cosine functions, even though the frequencies for the example were specially chosen. Original coefficients are located in an intermediate position. Among the specific methods, it can be seen in the table that BP achieves the sparsest data representation and requires a smaller number of bits to be coded, although with a greater margin of error than the other methods. NOCICA and MP behaved similarly. These specific methods are able to find even sparser representations than the original one. Kurtosis results are in agreement with the previous ones, recalling that, for this parameter, higher values are related to higher sparsity.

An alternative analysis which confirms these observations can be appreciated in the graphical representation shown in Fig. 4. It shows the mean of the coefficients, ordered and normalized to the maximum value, for the different representations presented in Table 1. Taking into account that, in our case, for each representation we obtain a coefficients matrix belonging to $\mathbb{R}^{64\times1000}$ (that corresponds to 1000 frames with 64 coefficients each one), for each frame the coefficients are sorted and the mean value is computed for the obtained rows. In Fig. 4 the obtained vector, normalized with the global maximum of the original coefficients matrix, for each representation are shown. A higher sparsity of the representation is characterized by a steeper fall of the corresponding curve.

The inverse problem was then solved using both parametric and non-parametric algorithms so that the dictionaries which were produced by these methods could be compared with the original one (which, as previously mentioned, was known in this artificial case). Table 2 shows the results obtained. It can be seen that all measurements, except ℓ_1, favor the proposed method. Thus, it can be said that this one achieves a sparser representation and with a smaller margin of error. Another column is included in the table with the average MSE between the original dictionary and the one obtained by both methods. The result shows that the parametric method LP-ICA estimates the original dictionary used to generate the data better than the non-parametric one. Kurtosis values are in agreement with these results. This can be corroborated by inspecting Figs. 5 and 6 where a comparison between some of the atoms obtained by both methods and the original one is given, for both the time and frequency domains. As can be observed, the NOCICA method tends to find atoms with more spectral peaks than the original ones.

As can be appreciated, the parametric method LP-ICA finds atoms which are more similar to the original ones and achieves a sparser representation than the NOCICA method. This is due to the fact that it benefits

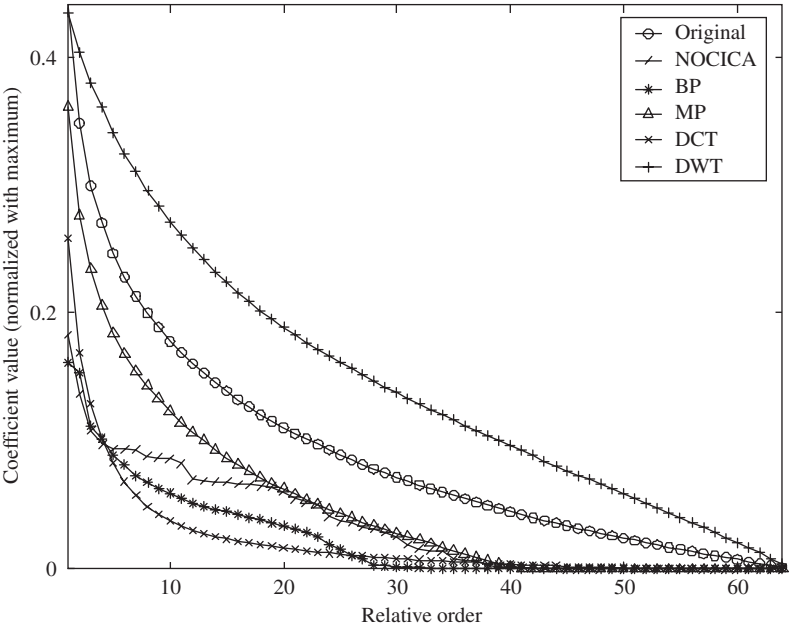

Fig. 4. The mean of the sorted coefficients normalized with the maximum value, for the different representations shown in Table 1 which are derived from a fixed dictionary (64×64).

Table 2
Sparsity and coding costs obtained from the representations of the artificial data using the different methods (including the estimation of the dictionary)

Method	ℓ_0	*minvol*	ℓ_1	$\mathscr{K}$	*#bits*	MSE $(\mathbf{s}, \mathbf{\Phi a})$	MSE$(\mathbf{\Phi}, \hat{\mathbf{\Phi}})$
NOCICA	0.45	0.63	0.60	1.14	3.37	1.28E − 04	1.3634
LP-ICA	0.20	0.34	0.85	21.86	2.39	5.38E − 06	1.0902

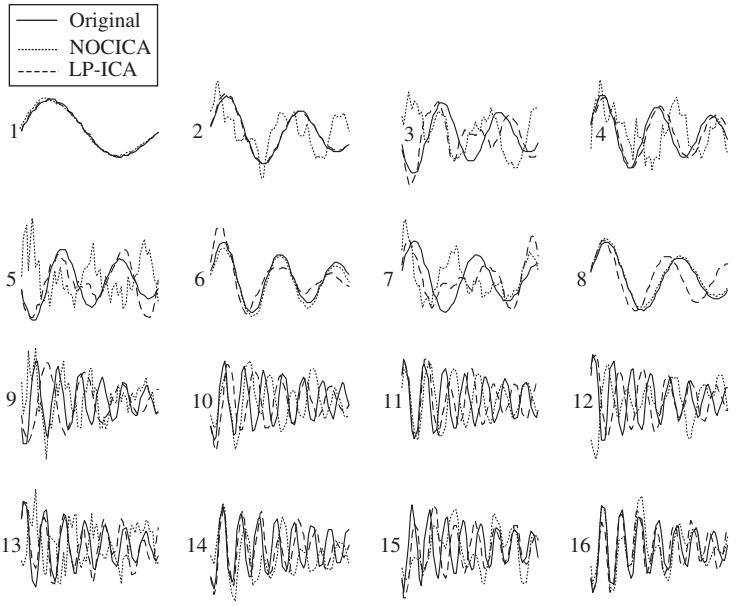

Fig. 5. A comparison among the atoms found by the different methods for the artificial data in the time domain (64 × 64).

from the a priori knowledge that the temporal structure of the atoms can be described by a simple parametric model (which is precisely the case for this example).

5.2. Real data

A subset of the Albayzin speech corpus [36] was used for the experiments. This subset consisted of 600 sentences concerning Spanish geography, with a vocabulary size of 200 words. The corpus was recorded in a studio using six male and six female speakers from the central area of Spain with an average age of 31.8 years. The average sentence lasted 3.55 s and the data were digitalized at 8 kHz using 16 bits and a μ-law sampling. From the segmentation information, frames with a size of 128 samples were extracted for five vowels (/a/, /e/, /i/, /o/ and /u/) and two consonants (/s/ and /k/), giving approximately 2000 frames each. This subset was selected for these initial experiments in order to include different phonemic classes while maintaining a small dataset. The proposed parametric method LP-ICA and the NOCICA version [6] have been applied to the data, and the tests described in Section 4 have been calculated for the 128 × 128 and the 128 × 256 cases (complete and overcomplete cases). Different experiments have been performed, training the methods with the data for each isolated phoneme, and then adding the data together (the data case ''all''). The same data were used for training the dictionaries and for testing the sparsity and coding costs.

The results obtained for the two methods are presented in Tables 3 and 4 for the complete and overcomplete case, respectively. The last two rows of each table show the significance values obtained with a paired Wilcoxon test. As can be observed, once more the proposed method LP-ICA gives in general a sparser

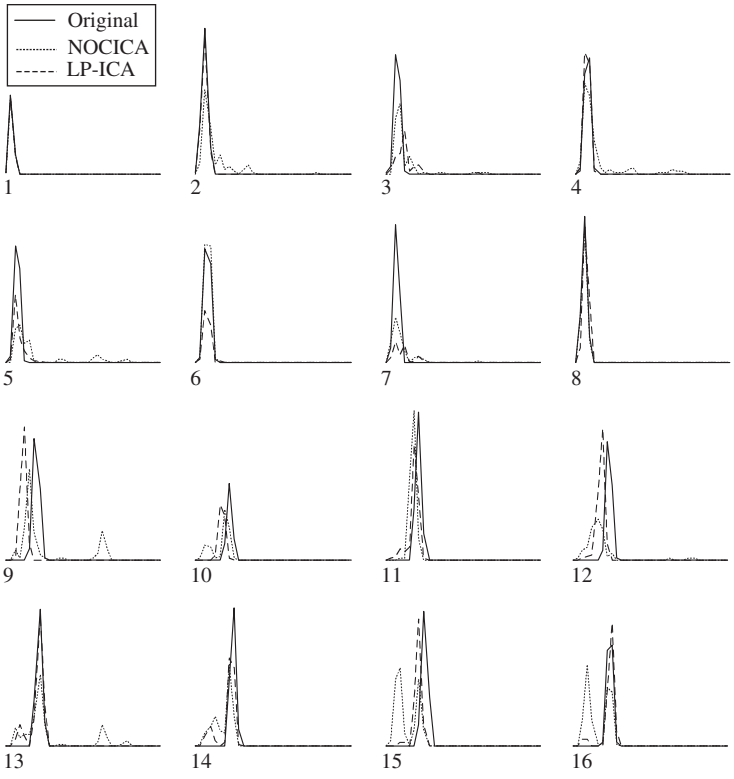

Fig. 6. A comparison among atoms' spectra found by the different methods for the artificial data (64 × 64).

Table 3
Sparsity and coding costs obtained from different representations of the real speech data for different phonemes using NOCICA and LP-ICA, for the complete case (128 × 128)

Experiment	ℓ_0	*minvol*	ℓ_1	$\mathscr{K}$	$\mathscr{H}$	*#bits*	MSE (**s**, **Φa**)
/a/(NOCICA)	0.17	0.26	0.54	25.96	1.06	1.86	6.19E − 04
/e/(NOCICA)	0.15	0.23	0.47	35.62	0.94	1.79	6.07E − 04
/i/(NOCICA)	0.12	0.17	0.32	46.79	0.70	1.33	6.17E − 04
/o/(NOCICA)	0.11	0.16	0.35	70.70	0.76	1.24	5.80E − 04
/u/(NOCICA)	0.08	0.10	0.17	105.89	0.38	0.74	5.48E − 04
/s/(NOCICA)	0.10	0.15	0.32	37.61	0.64	1.05	7.36E − 04
/k/(NOCICA)	0.38	0.51	0.56	11.49	0.77	1.63	1.20E − 03
Median	0.12	0.17	0.35	37.61	0.76	1.33	6.17E − 04
/a/(LP-ICA)	0.16	0.24	0.50	32.46	1.00	1.58	6.03E − 04
/e/(LP-ICA)	0.02	0.11	0.76	173.23	1.92	0.85	1.09E − 03
/i/(LP-ICA)	0.11	0.16	0.33	61.82	0.75	1.16	7.85E − 04
/o/(LP-ICA)	0.03	0.09	0.50	201.22	1.22	0.86	8.63E − 04
/u/(LP-ICA)	0.06	0.08	0.16	157.96	0.36	0.68	6.52E − 04
/s/(LP-ICA)	0.11	0.18	0.40	36.50	0.82	1.16	8.25E − 04
/k/(LP-ICA)	0.03	0.06	0.19	143.69	0.50	0.59	7.01E − 04
Median	0.06	0.11	0.40	143.69	0.82	0.86	7.88E − 04
Significance	0.047	0.11	0.72	0.03	0.47	0.05	0.37
$*$ if $p<0.05$	*			*		*	

Table 4
Sparsity and coding costs obtained from different representations of the real speech data for different phonemes using NOCICA and LP-ICA, for the overcomplete case (128 × 256)

Experiment	ℓ_0	*minvol*	ℓ_1	$\mathscr{K}$	$\mathscr{H}$	*#bits*	MSE (**s**, **Φa**)
/a/(NOCICA)	0.22	0.27	0.27	12.99	0.52	1.96	5.55E − 04
/e/(NOCICA)	0.09	0.12	0.21	61.15	0.48	1.12	6.07E − 04
/i/(NOCICA)	0.06	0.08	0.15	103.55	0.35	0.76	6.06E − 04
/o/(NOCICA)	0.04	0.07	0.16	110.18	0.39	0.78	5.86E − 04
/u/(NOCICA)	0.04	0.05	0.08	134.24	0.18	0.48	5.41E − 04
/s/(NOCICA)	0.17	0.21	0.31	17.19	0.54	1.25	9.30E − 04
/k/(NOCICA)	0.21	0.24	0.23	27.17	0.42	1.04	7.80E − 04
All(NOCICA)	0.12	0.20	0.51	32.86	1.03	1.29	8.69E − 04
Median	0.11	0.12	0.21	61.15	0.45	1.08	6.07E − 04
/a/(LP-ICA)	0.07	0.12	0.30	34.75	0.62	1.31	5.71E − 04
/e/(LP-ICA)	0.06	0.10	0.28	79.81	0.61	0.99	7.80E − 04
/i/(LP-ICA)	0.04	0.07	0.21	125.19	0.49	0.71	7.02E − 04
/o/(LP-ICA)	0.06	0.09	0.17	87.28	0.38	0.76	5.87E − 04
/u/(LP-ICA)	0.00	0.02	0.15	797.53	0.33	0.20	6.16E − 04
/s/(LP-ICA)	0.10	0.13	0.20	34.75	0.41	0.88	7.20E − 04
/k/(LP-ICA)	0.04	0.06	0.13	120.10	0.34	0.56	7.02E − 04
All(LP-ICA)	0.12	0.19	0.45	36.78	0.92	1.20	8.24E − 04
Median	0.06	0.10	0.21	83.55	0.45	0.82	7.02E − 04
Significance	0.05	0.05	0.98	0.11	0.55	0.01	0.84
∗ if $p<0.05$	*	*				*	

representation, with a smaller number of bits and with a similar MSE. This difference is also more pronounced in the overcomplete case. In the parametric case the final average order Q found was 29 and 22 for the complete and overcomplete case, respectively.

An example of the waveforms from some of the atoms found in the "all" data case (128 × 256) (data mixed from all the classes) is also shown in Figs. 7 (a) and (b), both for NOCICA and for the proposed parametric method, respectively. At first sight the dictionaries found look similar (observe the marked atoms) and for this case the measures slightly favor the parametric method (see Tables 3 and 4, row "All (128 × 256)").

In order to get a better understanding of why the learned dictionaries reflect the most important features of the different phoneme types, a qualitative analysis was performed on some of them. Fig. 8 shows the spectrograms obtained from the learned dictionary atoms for the vowel /a/ (128 × 256) with both methods (see Section 4.3). Among the observed differences, one can see how NOCICA achieves a representation that encompasses not only the involved frequencies, but also some atoms which account for the "phase" or specific temporal events (look at those displaying vertical light patterns). On the other hand, given that the parametric method LP-ICA assumes that the atoms constitute impulse responses of AR filters, the relative phase aspect is ignored and only atoms tuned to specific frequencies appear. This would indicate a relative insensitivity to the phase, a desirable feature if one wants to use the dictionary as a shift invariant event detector.

Fig. 9 shows the spectrograms obtained from the learned dictionary atoms for the phoneme /s/ (128 × 256) with both methods. Here, a similar analysis to that for /a/ can be conducted, noting that there is a much more marked difference in the number of atoms tuned to a principal frequency. This is due to the fact that, in order to achieve large band widths, more complex or higher order models should be used; thus, the method finds a simpler solution, which in turn appears to be even sparser (see Tables 3 and 4 for this case). Fig. 10 illustrates these ideas, it also gives a more "global" view of the distribution of the atoms in the $T-F$ plane. One can observe the $T-F$ coverage of each one of the dictionary atoms as ellipses, as was described in Section 4.3. In the case of /a/ (upper row), it can be clearly seen that the proposed method LP-ICA offers a greater frequency resolution for low frequencies, particularly in the zone which corresponds to formants (see further on). The aspect of the aforementioned unique phase is also corroborated. On the other hand, in the case of /s/ (bottom

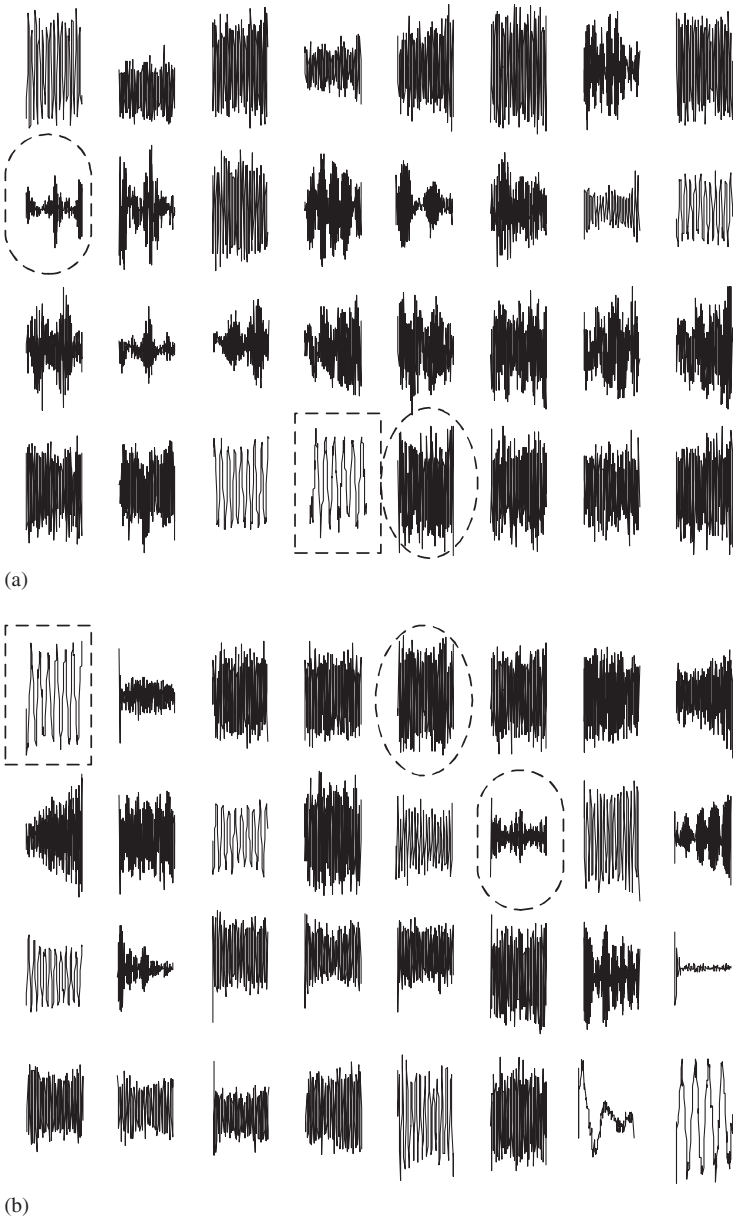

Fig. 7. Some of the waveforms obtained for all speech data using: (a) the standard NOCICA method, (b) the proposed parametric method LP-ICA.

row), it can be observed that most of the atoms found by the proposed method are tuned to a principal frequency, with greater resolution in the high-frequency zone.

If the dictionaries found for vowels by both methods are further analyzed, smaller peaks at other frequencies also appear, even though the larger part of the atoms' energy is located around one main frequency. This would mean that other relevant information has been coded in the atoms, which is not evident from the previous analyses (although it can be noticed after careful inspection of spectrograms in Fig. 8). It is known that formants are important to distinguish between vowels, both in the isolated case and in continuous speech. However in the latter case one also has to track changes of formantic patterns in time because the classes are not as well separated [37]. Because of its frequency range, this extra information could be associated with speech formants. This means that these methods are able to find relevant information for discrimination

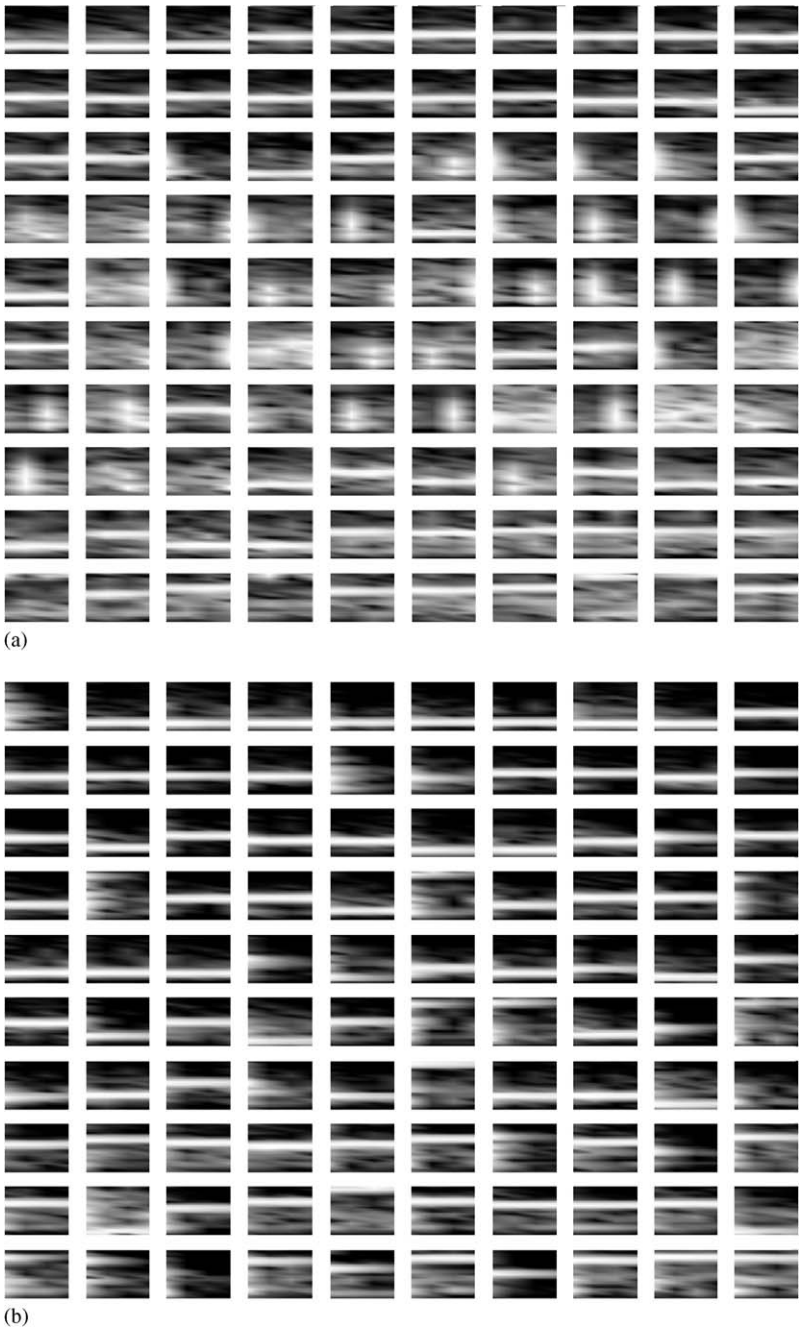

Fig. 8. Spectrograms obtained from learned dictionary atoms for vowel /a/ (128 × 256): (a) NOCICA, (b) LP-ICA. The width for time and height for frequency axis is 16 ms and 4 kHz, respectively, for each atom.

purpose using only the training data and, in the parametric case LP-ICA, this information seems to be better represented.

6. Conclusions and future work

In the present paper we have introduced a new method to obtain an independent and sparse representation based on solving an overcomplete ICA problem with noise by means of a parametric dictionary called

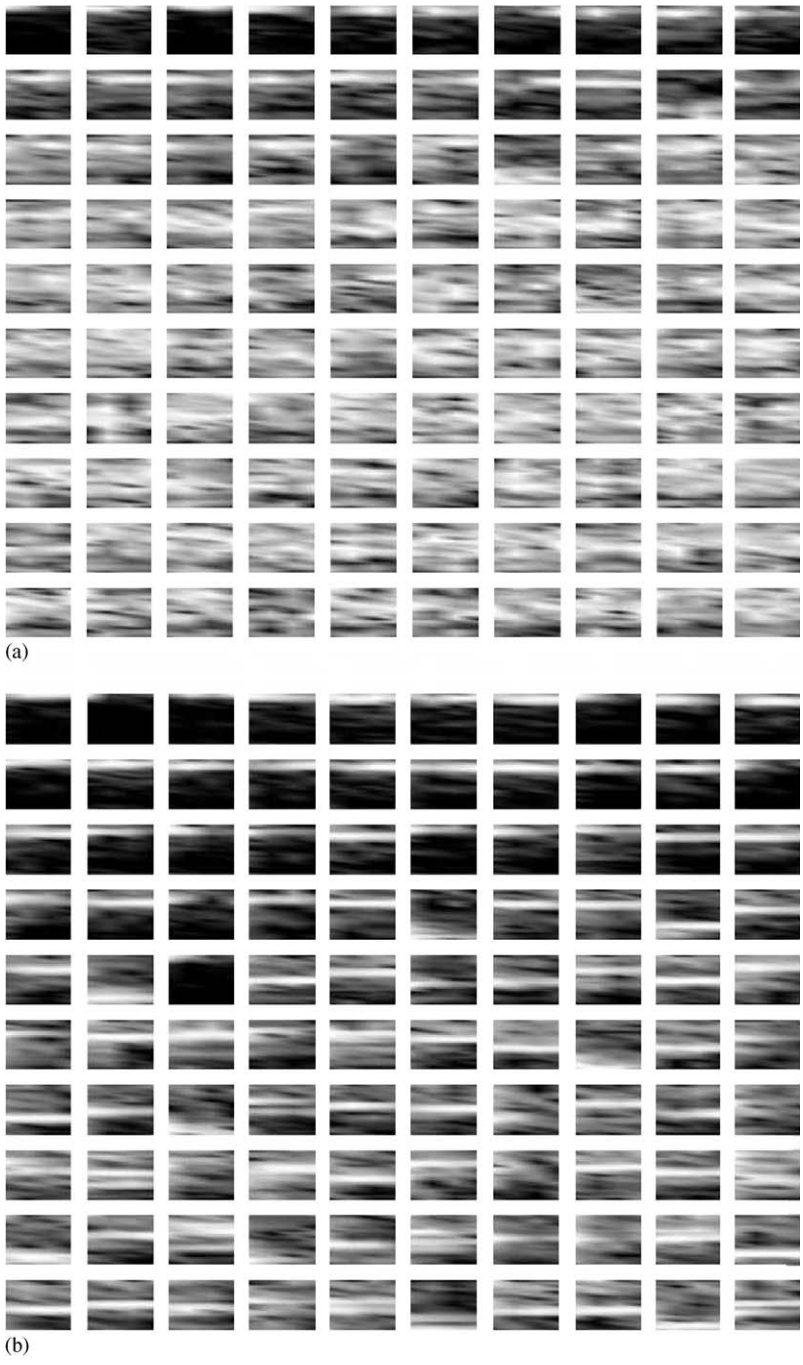

Fig. 9. Spectrograms obtained from learned dictionary atoms for the fricative /s/ (128 × 256): (a) NOCICA, (b) LP-ICA. The width for time and height for frequency axis is 16 ms and 4 kHz, respectively, for each atom.

LP-ICA. The method has been applied and compared to a standard version by calculating different sparsity measures and coding costs using artificial data and real speech examples. A qualitative analysis of the obtained dictionaries has also been performed.

It has been shown that the results obtained favor the proposed method LP-ICA, which takes advantage of the temporal correlations in the data to find a suitable and better solution.

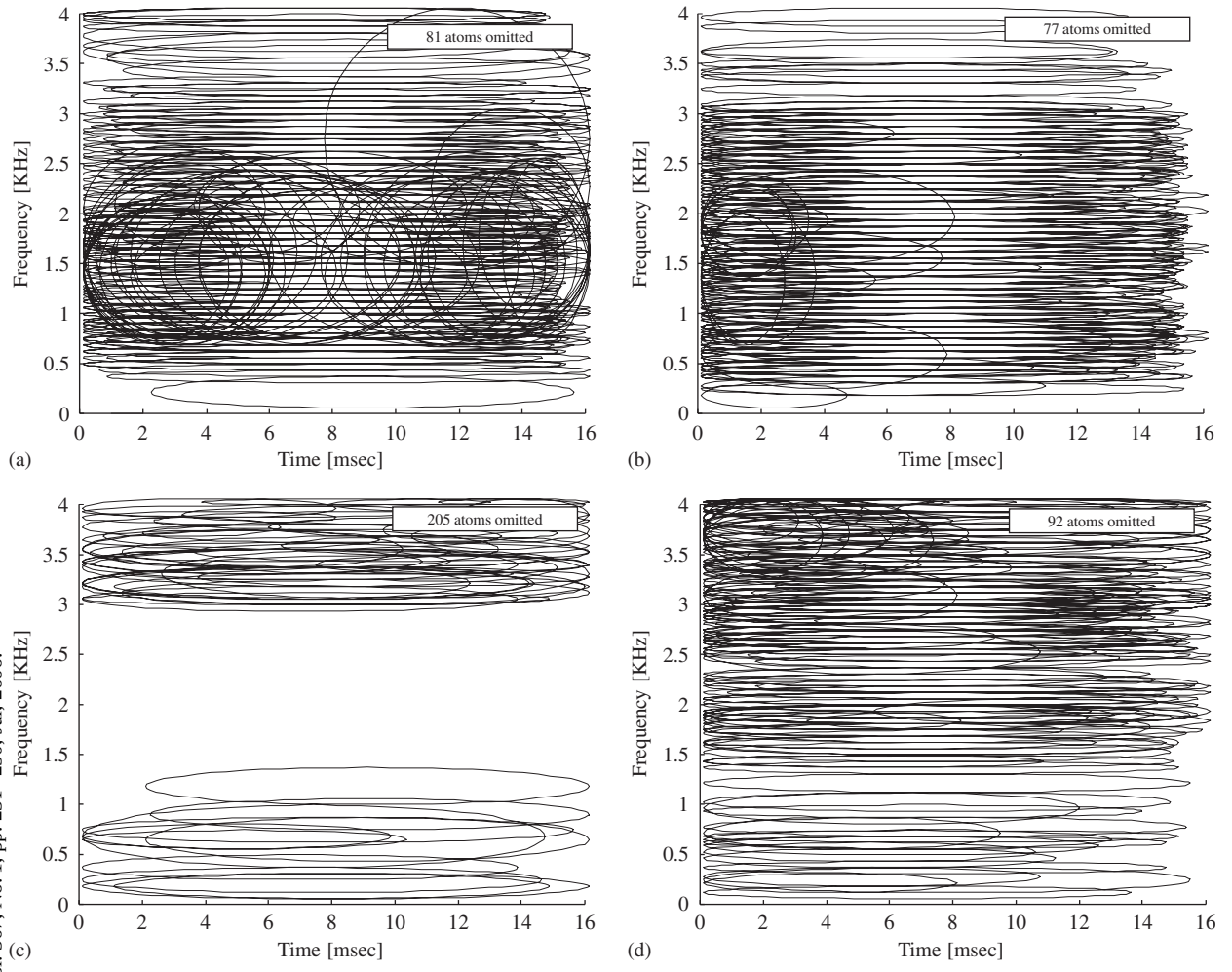

Fig. 10. Time–frequency tiles obtained from learned dictionary: (a) vowel /a/ (128 × 256) NOCICA, (b) vowel /a/ (128 × 256) LP-ICA, (c) fricative /s/ (128 × 256) NOCICA, (d) fricative /s/ (128 × 256) LP-ICA. The number of atoms that were omitted because of bad localization were indicated in each plot.

We have shown that the waveforms found by both methods also exhibit important differences derived from the restrictions imposed by the parametric approximation used by LP-ICA. These restrictions allow to discover a different representation of the data that preserves important characteristics of the different phonemes considered here, and with better sparsity. This is an important feature if the phonemes are to be statistically modeled within this framework.

The use of these dictionaries within the context of an ASR system is an area to be pursued in a future work. It is important to mention that it is possible to apply the same type of analysis to speech representations other than waveforms, avoiding in this way the possible use of ''too many'' atoms to code characteristics which may have little significance for recognition purpose such as the atoms phase. However, an important issue addressed by the LP-ICA method is the possibility of achieving phase shift invariant representations, which constitutes a known problem in other ICA-based feature extraction methods [14]. The proposed approach could take advantage of the parametric representation of the atoms in order to carry out the analysis stage using a greedy MP approximation, implemented by means of an AR matched filter bank [38]. This constitutes a fast approximate implementation to solve the inference problem with a physiological plausible basis and will be explored in a future work.

References

[1] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[2] J. Deller, J. Proakis, J. Hansen, Discrete Time Processing of Speech Signals, Macmillan Publishing, New York, 1993.
[3] H. Hermansky, N.H. Morgan, Rasta processing of speech, IEEE Trans. Speech Audio Process. 2 (4) (1994) 587–589.
[4] R.P. Lippmann, Speech recognition by machines and humans, Speech Commun. 19 (22) (1997) 1–15.
[5] S.S. Chen, D.L. Donoho, M.A. Sanders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1999) 33–61.
[6] M.S. Lewicki, B.A. Olshausen, A probabilistic framework for the adaptation and comparison of image codes, J. Opt. Soc. Am. 16 (7) (1999) 1587–1601.
[7] B.A. Olshausen, D.J. Field, Vision and the coding of natural images, Am. Sci. 88 (3) (2000) 238–245.
[8] B.A. Olshausen, D.J. Field, Natural image statistics and efficient coding, in: Proceedings of the Workshop on Information Theory and the Brain, vol. 7, Scotland, September 4–5, 1995, University of Stirling, pp. 333–339.
[9] M.D. Plumbley, S.A. Abdallah, J.P. Bello, M.E. Davies, G. Monti, M.B. Sandler, Automatic music transcription and audio source separation, Cybernetics Systems 33 (6) (2002) 603–627.
[10] T.-P. Jung, S. Makeig, T.-W. Lee, M.J. McKeown, G. Brown, A.J. Bell, T.J. Sejnowski, Independent component analysis of biomedical signals, in: Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, 2000, pp. 633–644.
[11] S. Makeig, A.J. Bell, T.-P. Jung, T.J. Sejnowski, Independent component analysis of electroencephalographic data, Adv. Neural Inform. Process. Systems 8 (1996) 145–151.
[12] S. Makeig, T.-P. Jung, A.J. Bell, T.J. Sejnowski, Blind separation of auditory event-related brain responses into independent components, Proc. Nat. Acad. Sci. USA 94 (1997) 10979–10984.
[13] J.H. Lee, H.Y. Jung, T.W. Lee, S.Y. Lee, Speech feature extraction using independent component analysis, in: Proceedings of the ICASSP, vol. 3, Istanbul, Turkey, 2000, pp. 1631–1634.
[14] O.-W. Kwon, T.-W. Lee, Phoneme recognition using ICA-based feature extraction and transformation, Signal Process. 84 (6) (2004) 1005–1019.
[15] K. Katayama, Y. Sakata, T. Horiguchi, Sparse coding for layered neural networks, Physica A 310 (3–4) (2002) 532–546.
[16] M.S. Lewicki, Efficient coding of natural sounds, Nature Neurosci. 5 (4) (2002) 356–363.
[17] F. Guspí, B. Introcaso, Soluciones ralas de sistemas lineales indeterminados, El Ingeniero en la Red 1(VII) (2000) 1–10 (Revista Electrónica FCEIyA, UNR, Argentina).
[18] S.G. Mallat, Z. Zhang, Matching pursuit with time–frequency dictionaries, IEEE Trans. Signal Process. 41 (1993) 3397–3415.
[19] R. Coifman, M.V. Wickerhauser, Entropy-based algorithms for best basis selection, IEEE Trans. Inform. Theory 38 (2) (1992) 713–719.
[20] I. Daubechies, Time–frequency localization operators: a geometric phase space approach, IEEE Trans. Information Theory 34 (4) (1988) 605–612.
[21] H.L. Rufiner, J. Goddard, A.E. Martínez, F.M. Martínez, Basis pursuit applied to speech signals, in: Fifth World Multi-Conference on Systemics, Cybernetics and Informatics (SCI), Orlando, 2001, IEEE, Silverspring, MD, pp. 517–520.
[22] H.L. Rufiner, L.F. Rocha, J.G. Close, Preserving acoustic cues in speech denoising, in: Proceedings of the Second Joint Meeting of the IEEE Engineering in Medicine and Biology Society and the Biomedical Engineering Society EMBS-BMES2002, vol. 1, Houston, 2002, pp. 288–289.
[23] B.A. Olshausen, D.J. Field, Emergence of simple cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[24] A. Hyvärinen, Survey on independent component analysis, Neural Comput. Surveys 2 (1999) 94–128.
[25] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and application, Neural Networks 13 (4–5) (2000) 411–430.
[26] N. Saito, B.M. Larson, B. Benichou, Sparsity vs. statistical independence from a best-basis viewpoint, in: A. Aldroubi, A.F. Laine, M.A. Unser (Eds.), Wavelet Applications in Signal and Image Processing VIII, Proceedings of the SPIE, vol. 4119, 2000, pp. 474–486.
[27] G.F. Harpur, Low entropy coding with unsupervised neural networks, Ph.D. Thesis, Department of Engineering, University of Cambridge, Queens' College, February 1997.
[28] H.L. Rufiner, L.F. Rocha, J.G. Close, Sparse and independent representations of speech signals based on parametric models, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, USA, 2002, pp. 989–992.
[29] S.A. Abdallah, Towards music perception by redundancy reduction and unsupervised learning in probabilistic models, Ph.D. Thesis, Department of Electronic Engineering, King's College London, 2002.
[30] M.S. Lewicki, T.J. Sejnowski, Learing overcomplete representations, in: Advances in Neural Information Processing 10 (Proceedings NIPS'97), MIT Press, Cambridge, MA, 1998, pp. 556–562.
[31] T.W. Parks, C.S. Burrus, in: Digital Filter Design, Wiley, New York, 1987, pp. 226–228.
[32] D.L. Donoho, Sparse components of images and optimal atomic decomposition, Technical report, Department of Statistics, Stanford University, December 1998.
[33] N. Saito, B. Benichou, Sparsity vs. statistical independence in adaptive signal representations: a case study of the spike process, in: G.V. Welland (Ed.), Beyond Wavelets, vol. 10, Studies in Computational Mathematics, Academic Press, New York, 2003, pp. 225–257 (Chapter 9).
[34] M.E. Torres, El procesamiento de señales ligadas a problemas no lineales, Ph.D. Thesis, Universidad Nacional de Rosario—Argentine, 1999 (Math. D. Thesis).

[35] L. Aronson, L. Rufiner, H. Furmansky, P. Estienne, Características acústicas de las vocales del español Rioplatense, Fonoaudiológica 46 (2) (July–September 2000) 12–20.
[36] F. Casacuberta, R. García, J. Llisterri, C. Nadeu, J.M. Prado, A. Rubio, Development of a Spanish corpora for the speech research, in: Workshop on International Co-operation and Standardisation of Speech Databases and Speech I/O Assessment Methods, Chiavari, 1991, pp. 26–28.
[37] J.M. Hillenbrand, M.J. Clark, T.M. Nearey, Effects of consonant environment on vowel formant patterns, J. Acoust. Soc. Am. 109 (2) (2001) 748–763.
[38] M.M. Goodwin, M. Vetterli, Matching pursuit and atomic signal models based on recursive filter banks, IEEE Trans. Signal Process. 47 (7) (1999) 1890–1902.