SPARSE AND INDEPENDENT REPRESENTATIONS OF SPEECH SIGNALS BASED ON PARAMETRIC MODELS

Hugo L. Rufiner, Luis F. Rocha[*]

Facultad de Ingeniería
Univ. Nac. de Entre Ríos
Argentina

John Goddard Close[†]

Depto Ing. Eléctrica
Univ. Autónoma Metropolitana
México

ABSTRACT

Recently methods for obtaining sparse representations of a signal using overcomplete dictionaries of waveforms have been studied, often motivated by the way the brain seems to process certain sensory signals. Algorithms have been developed using either a specific criterion to choose the waveforms occurring in the representation from a fixed dictionary, or to construct them as part of the method. In the case of speech signals, most approaches do not take into consideration the important temporal correlations that exist; these are known to be well approximated using linear models. Incorporating this type of a priori knowledge of the signal can facilitate the search for a suitable solution and also help with the interpretation of the representation found. In the present paper a method is proposed for obtaining a sparse representation using a generative parametric model. An example, using speech signals, is given reporting the method's efficacy for different coding costs and sparsity measures.

1. INTRODUCTION

Speech signals are amongst the most studied natural signals, however there are several problems relating to their analysis and recognition which still have not been satisfactorily solved. In the last few years several researchers have taken a different approach to traditional signal processing, where it is assumed that the signals derive from linear time invariant systems with second order statistics. The new approaches give rise to techniques based on higher order statistics and include *Independent Component Analysis* (ICA) and methods to obtain sparse representations of a signal. These approaches provide a new way of phrasing the solution to these problems.

Generally speaking a sparse coding of information is one in which only a small number of descriptors are used from a large set [1], that is, only a fraction of the elements are actively used to describe a typical pattern. In the case of a signal $\mathbf{s}$, a representation in terms of a dictionary $\mathbf{\Phi}$, or collection of parameterized waveforms $(\phi_\gamma)_{\gamma\in\Gamma}$, usually signifies a decomposition of the form:

$$\mathbf{s} = \sum_{\gamma\in\Gamma} \phi_\gamma a_\gamma = \mathbf{\Phi a} \qquad (1)$$

where $\mathbf{a}$ is the vector of coefficients. Sparseness, in this context, 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. [2]). Commonly used dictionaries are the traditional Fourier sinusoids (frequency dictionaries), Dirac functions, Wavelets (time-scale dictionaries), Gabor functions (time-frequency dictionaries), Polynomials or combinations of these.

An overcomplete dictionary implies many possible representations, and the following methods have been proposed for attaining a sparse representation from a fixed dictionary: *Basis Pursuit* [3], *Matching Pursuit* [4] and *Best Orthogonal Basis* [5]. A first attempt to apply this type of analysis to speech signals using a fixed dictionary of wavelet packets appeared in [6], giving promising results.

Further there are methods (c.f. [1]), often of a statistical nature, which also construct the waveforms appearing in (1). In this case the coefficients are usually assumed to be statistically independent, and while this does not necessarily imply sparsity of the representation, it can be achieved using a proper choice for the prior probability of the coefficients e.g. Laplacian. With this approach there are important connections to ICA [7]. In fact, it is interesting to note that whilst sparsity and independence are different criteria, they can often produce similar solutions (c.f. [8]). There are many applications of these techniques to such fields as: "natural" image analysis [9], audio and music signals [10], general biomedical signals [11] and Automatic Speech Recognition [12].

The majority of the available techniques for finding the bases do not include restrictions related to the temporal struc-

[*] The first two authors wish to thank the Facultad de Ingeniería, UNER, and Facultad de Ingeniería, UBA, Argentina for its support

[†] The last author is grateful to CONACYT, Mexico for financing this research under Project 31929-A

ture of speech signals [13, 14, 15]. In the present paper an algorithm is proposed which includes these restrictions by assuming that the elements of the dictionary are represented by an *autoregressive*(AR) filter. This type of representation has been applied extensively, and successfully, in the past to speech signals.

The paper is organized as follows: in section 2 the waveforms and algorithm used are briefly explained. In section 3 the tests employed for judging the representational efficiency of this framework are discussed and the results obtained using an example involving speech signals are given in section 4. Finally section 5 provides some conclusions concerning the paper.

2. REPRESENTATION BASED ON PARAMETRIC MODELS

In this section a more general framework is assumed where (1) is rewritten to include an additive Gaussian noise term ε as follows:

$$\mathbf{s} = \mathbf{\Phi}\mathbf{a} + \varepsilon \tag{2}$$

Following terminology used in ICA, (2) 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 base $\mathbf{\Phi}$ of size $N \times M$, with $M \geq L$.

One problem with modeling speech signals using this framework is that in the time domain (2) represents an instantaneous mixture, whereas the phenomenon is better described by a convolutive mixture. As a simple analogy, the ϕ_j can be imagined to be the characteristic states of a linear model of the vocal tract for different phonemes; a particular speech signal can then be obtained by “adding” the most important characteristics together. To try to implement this idea, the present paper restricts the waveforms used in the dictionary to those of the form:

$$\phi_{i,j} = \sum_{p=1}^{P} \phi_{i-p,j} c_{p,j} + \delta_i a_j \tag{3}$$

This restriction allows the explicit inclusion of the temporal correlation of the samples of each atom using the coefficients $c_{p,j}$. This means that the problem can be phrased as overcomplete ICA with noise and with certain restrictions on the mixing matrix. These restrictions represent a particular case of convolutive mixtures in the time domain, or can be formulated in the z domain of the original variables. In this case the convolution becomes a product and the ϕ_j's can be expressed as:

$$\Phi_j(z) = \frac{1}{C_j(z)} \tag{4}$$

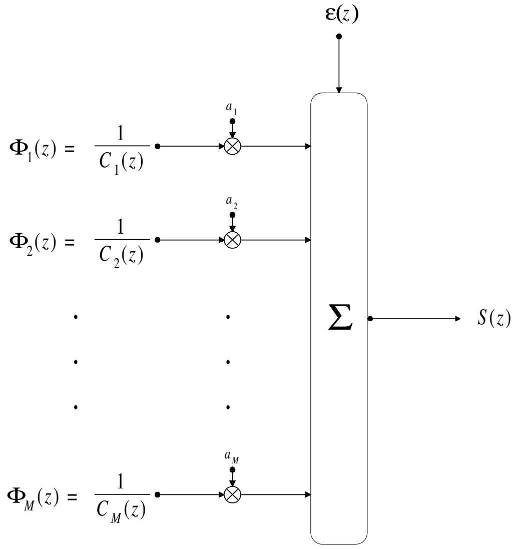
ε(z)
a_1
$\Phi_1(z) = \frac{1}{C_1(z)}$
a_2
$\Phi_2(z) = \frac{1}{C_2(z)}$
Σ
$S(z)$
a_M
$\Phi_M(z) = \frac{1}{C_M(z)}$

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

where $C_j(z)$ correspond to a polynomial in z of order P with coefficients $c_{p,j}$. The corresponding generative model is shown in Figure 1.

The approach taken in the present paper for finding the coefficients a_j and parametric waveforms $\phi_{i,j}$ is to use the techniques described in Lewicki and Olshausen [16] and Lewicki and Sejnowski [17]. This involves assuming that the coefficients are statistically independent, each with a Laplacian prior probability. Gradient ascent is applied to both the coefficients and waveforms using suitable updating rules given in [16, 17]. Finally, as the matrix $\mathbf{\Phi}$ should satisfy the restrictions placed on the columns by (3), the problem arises as to how to estimate the coefficients $c_{p,j}$. The approach adopted here used:

$$\frac{\partial MSE(\phi_j, \hat{\phi}_j)}{\partial c_{p,j}} = 0 \tag{5}$$

where MSE is the mean square error between the ϕ_j and the $\hat{\phi}_j$ approximated using (3). For the solution, the method of autocorrelation is used, and then $\mathbf{\Phi}$ is replaced by its parametric version $\hat{\mathbf{\Phi}}$. 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 P whilst $\log P(\mathbf{s}|\mathbf{\Phi})$ is increased. Further, the data $\mathbf{s}$ can also be approximated by a parametric model to help the convergence (which is the same as replacing them by the impulse responses of equivalent AR systems).

3. TESTS TO ESTIMATE SPARSITY AND CODING COSTS

There are several tests to estimate sparsity and coding costs of the representation, and in the present paper six of the most common are employed to judge the representational efficiency of the proposed method. The tests can be divided into two groups: those related to sparsity with respect to a norm, and those related to the statistics of the coefficients. In the first category are ℓ_0, minvol and ℓ_1 (using the notation given in the tables). The first, ℓ_0, counts the number of nonzero coefficients. minvol [2] is an approximation to ℓ_0 when a small threshold is used to decide if a coefficient is considered nonzero. ℓ_1 is often used as a practical approximation to the ℓ_0-norm in optimization problems (c.f. [3]). Minimizing with respect to this norm is the same as requiring that the coefficients have a Laplacian prior probability (which is the case here). In the second category are K, H and # bits. K is the kurtosis and is often used for symmetric unimodal distributions (c.f. [18]). Generally its value increases when the entropy decreases, and so it is often related to statistical independence. The entropy H is often used to measure coding efficiency. It has more to do with statistical independence (c.f. [18]) than sparsity although it is desirable that H have a low value in this case. Finally # bits represents the coding cost in terms of the number of bits, and is calculated using H.

Finally, the value of MSE, the averaged mean squared error over the patterns, is included as a measure of the reconstruction capability of the method.

4. RESULTS

A subset of the Albayzin speech corpus [19] was used for the experiments. This subset consisted in 600 sentences concerning Spanish geography, with a vocabulary size of 200 words. The corpus was recorded in a studio using 6 male and 6 female speakers from the central area of Spain with an average age of 31.8 years. The average sentence lasted 3.55 secs and the data was 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 5 Spanish vowels, giving approximately 2000 frames each. The proposed parametric method and the standard version [16] (using Lewicki's nocica code) were applied to the data, and the tests described in section 3 were calculated for the 128x128 case. The results obtained for the two methods are shown in Tables 1 and 2. The last rows of the tables show the mean values obtained for the corresponding columns averaged over all vowels. The power spectral density plots from some of the waveforms found are also shown in Figs. 2 and 3.

Table 1. Sparsity and coding costs obtained from the proposed parametric method after 150 iterations

	ℓ_0	minvol	ℓ_1	K	H	# bits	MSE
/a/	0.06	0.04	0.05	75.15	0.11	1.34	6.6E-05
/e/	0.17	0.17	0.20	28.49	0.41	1.67	1.4E-04
/i/	0.11	0.11	0.12	58.01	0.28	1.27	9.6E-05
/o/	0.11	0.11	0.13	43.78	0.28	1.45	9.7E-05
/u/	0.12	0.07	0.05	67.22	0.12	1.29	6.0E-05
	0.11	0.10	0.11	54.53	0.24	1.40	9.2E-05

Table 2. Sparsity and coding costs obtained from the standard method after 2500 iterations.

	ℓ_0	minvol	ℓ_1	K	H	# bits	MSE
/a/	0.16	0.24	0.50	34.24	1.01	1.55	6.1E-04
/e/	0.14	0.22	0.47	36.29	0.96	1.69	6.9E-04
/i/	0.13	0.19	0.36	47.47	0.78	1.38	6.7E-04
/o/	0.08	0.15	0.37	71.49	0.79	1.25	6.9E-04
/u/	0.09	0.12	0.22	74.04	0.47	0.90	6.4E-04
	0.12	0.18	0.38	52.71	0.80	1.35	6.6E-04

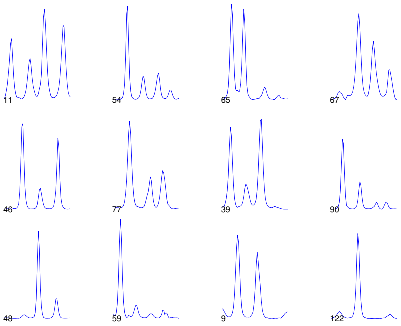

Fig. 2. Power Spectral Density plots from some of the waveforms obtained for the vowel /a/ using the proposed parametric method.

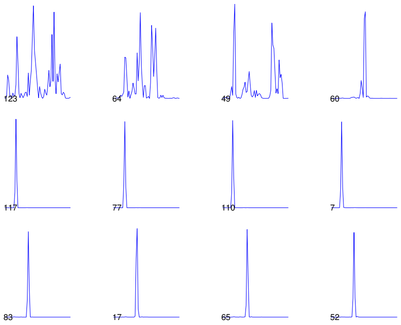

Fig. 3. Power Spectral Density plots from some of the waveforms obtained for the vowel /a/ using the standard method.

5. CONCLUSIONS

The present paper has introduced a method to obtain an independent and sparse representation based on solving an overcomplete ICA problem with noise for a specific parametric dictionary. The method was applied and compared to a standard version by calculating different sparsity measures and coding costs using a speech signal example.

As can be seen from the Tables the results obtained are, with one exception, better for the proposed method and use a significantly smaller number of iterations. This suggests that the proposed method is able to take advantage of the temporal correlations in the data to find a suitable solution quicker.

Figs. 2 and 3 of the power spectral density plots from some of the waveforms found by the two methods also exhibit important differences. Those from the proposed method, in Fig. 2, are smoother and more "complex" than those obtained by the standard method, suggesting that higher order characteristics are better captured.

An interesting property of the proposed method is that it allows for an easier way of dealing with the complexity of the waveforms given the "additive" nature of their poles.

6. REFERENCES

[1] B.A. Olshausen and D.J. Field, "Emergence of simple cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[2] F. Guspí and B. Introcaso, "Soluciones ralas de sistemas lineales indeterminados," *El Ingeniero en la Red*, vol. 1, no. VII, pp. 1–10, Mayo 2000, Revista Electrónica FCEIyA, UNR, Argentina.

[3] S.S. Chen, D.L. Donoho, and M.A. Sanders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[4] S.G. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. in Signal Proc.*, vol. 41, pp. 3397–3415, December 1993.

[5] R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, March 1992.

[6] H.L. Rufiner, J. Goddard, A.E. Martínez, and F.M. Martínez, "Basis pursuit applied to speech signals," in *5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI)*, Orlando, Julio 2001, IEEE.

[7] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[8] N. Saito, B. M. Larson, and B. Benichou, "Sparsity vs. statistical independence from a best-basis viewpoint," in *Wavelet Applications in Signal and Image Processing VIII*, A. Aldroubi, A.F. Laine, and M.A. Unser, Eds., 2000, Proc. SPIE 4119, pp. 474–486.

[9] B.A. Olshausen and D.J. Field, "Vision and the coding of natural images," *American Scientist*, vol. 88, no. 3, pp. 238–245, 2000.

[10] S. A. Abdallah and M. D. Plumbley, "Sparse coding of music signals," Preprint, 2001.

[11] T.-P. Jung, S. Makeig, T.-W. Lee, M.J. McKeown, G. Brown, A.J. Bell, and T.J. Sejnowski, "Independent component analysis of biomedical signals," http://www.cnl.salk.edu/ jung/ica.html, 2001.

[12] J.H. Lee, H.Y. Jung, T.W Lee, and S.Y. Lee, "Speech feature extraction using independent component analysis," in *Proc. ICASSP*, 2000, vol. 3, pp. 1631–1634.

[13] B.A. Pearlmutter and L.C. Parra, "A context-sensitive generalization of ica," in *International Conference on Neural Information Processing*, Hong Kong, September 1996.

[14] L.C. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, pp. 320–327, May 2000.

[15] A. Hyvärinen, "Complexity pursuit: Separating interesting components from time-series," *Neural Computation*, vol. 13, no. 4, pp. 883–898, 2001.

[16] M.S. Lewicki and B.A. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America*, vol. 16, no. 7, pp. 1587–1601, 1999.

[17] M.S. Lewicki and T.J. Sejnowski, "Learing overcomplete representations," in *Advances in Neural Information Processing 10 (Proc. NIPS'97)*. 1998, pp. 556–562, MIT Press.

[18] G. F. Harpur, *Low Entropy Coding with Unsupervised Neural Networks*, Ph.D. thesis, Cambridge University, Engineering Department, 1997.

[19] F. Casacuberta, R. García, J. Llisterri, C. Nadeu, J. M. Prado, and 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, September 1991, pp. 26–28.