World Multiconference on Systemics, Cybernetics and Informatics

July 22-25, 2001
Orlando, Florida, USA

PROCEEDINGS

Volume XVIII

Papers Content and Author Index

Organized by IIIS

International
Institute of
Informatics
and Systemics

Member of the International Federation of Systems Research

IFSR

Co-organized by IEEE Computer Society
(Chapter: Venezuela)

EDITOR
Nagib Callaos

Speech Processing

Automatic Target Recognition – Invited Session
Organizer: Dalton Rosario (USA)

Basis Pursuit applied to Speech Signals

J. Goddard, A. E. Martínez, F. M. Martínez, H.L. Rufiner[(1)]
Depto. de Ingeniería Eléctrica, U.A.M-I, Mexico
[(1)] Lab. Cibernética, F.I.-U.N.E.R., Argentina

ABSTRACT

In the present paper the method of Basis Pursuit with an overcomplete dictionary is applied to a set of phonemes chosen from the TIMIT database. Emphasis is placed on an analysis of the sparseness of the representation, and its relation to the preservation of the important acoustic cues for the phoneme group. A comparison with other methods, such as Method of Frames, Matching Pursuit and Best Orthogonal Basis is also given.

Keywords: Basis Pursuit, Sparse representations, Speech Analysis.

1. INTRODUCTION

In recent years a large number of papers have been devoted to the study of different ways of representing signals using dictionaries of appropriate functions [1-8]. A dictionary D is just a collection of parameterized waveforms $(\phi_\gamma)_{\gamma\in\Gamma}$, and a representation of the signal s in terms of D is usually a decomposition of the form:

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

Some commonly used dictionaries are the traditional Fourier sinusoids (frequency dictionaries), Dirac functions, Wavelets (time-scale dictionaries), Gabor functions (time-frequency dictionaries), or combinations of these.

Different methods, such as Method of Frames (MOF) [1], Matching Pursuit (MP) [2], Best Orthogonal Basis (BOB) [3] and Basis Pursuit (BP) [4], have been proposed for obtaining a decomposition. An important criterion for choosing a method consists in obtaining a sparse representation of the signal. This means that one would like only a 'few' of the coefficients, a_γ in (1), to be different from zero.

In [4], Chen et al propose a method, called Basis Pursuit, which is designed to produce such a sparse representation. They phrase the problem of finding a suitable representation as one of optimization with respect to the l_1 norm. More precisely, if the signal s has length n and there are p waveforms in the dictionary, then the problem to solve is:

$$\min \|a\|_1 \text{ subject to } \Phi a = s \qquad (2)$$

where a is a vector in $\Re^n$ representing the coefficients and Φ is a p x n matrix giving the values of the p waveforms in the dictionary.

This problem can be converted to a standard linear program (with only positive coefficients) by making the substitution $a \leftarrow [u, v]$ and solving (c.f. [4]):

$$\min 1^T[u, v] \text{ subject to } [\Phi,-\Phi][u, \mathrm{v}] = s,\ 0 \le u, v \qquad (3)$$

This formulation can be solved efficiently and exactly with interior point linear programming methods.

Chen et al give a number of artificial examples showing the benefits of their method, in terms of sparsity and super-resolution, compared to the corresponding representations found by MOF, MP and BOB. However a systematic study of the technique of Basis Pursuit applied to 'real world' data does not seem to have been conducted.

In the present paper this investigation is extended to the field of speech signals. Different speech signals present both highly transient and stationary behavior, so the use of the above approach is attractive in allowing the waveforms in a dictionary to adapt to the particular signal under consideration and thereby extracting the relevant features. The possibility of super-resolution and sparsity also presents advantages over traditional representations, such as Fourier.

A preliminary study of BP is conducted using a set of phonemes chosen from the TIMIT database. It includes vowels and consonants which exhibit these transient and stationary characteristics. The representations obtained from the of BP, MP, BOB and MOF are compared. Further the effect of choosing the most important coefficients, in terms of their numerical value, is studied with regard to the quality of the new signal obtained.

The paper is organized as follows: in the following section the data and dictionary are described, results are then presented and finally a short discussion and conclusions are given.

2. THE DATA AND DICTIONARY

The experiments conducted in this paper used the following set of phonemes:

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner, J. Goddard, A. E. Martínez & F. M. Martínez; "Basis Pursuit applied to Speech Signals"
Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001). Jul, ul, 2001.

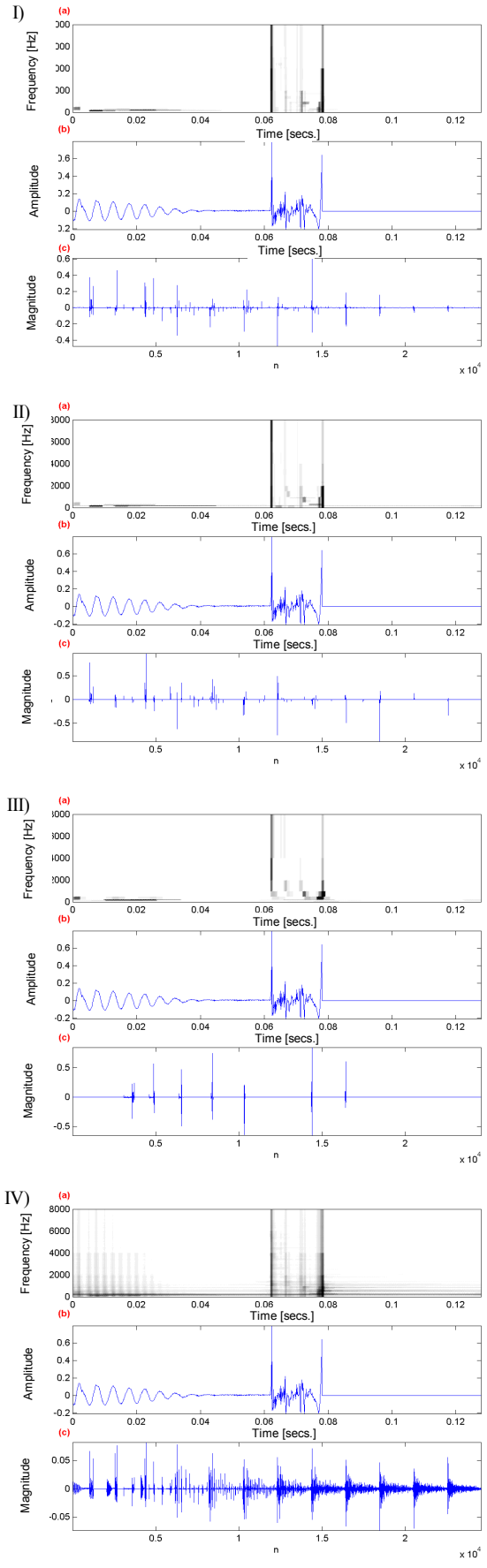

Figure 1: Comparison of different representations for phoneme /b/: I) BP, II) MP, III) BOB, IV) MOF, (a) Phase Plane, (b) Sonogram, (c) Coefficients.

/eh/, /ih/, /b/, /d/, /p/, /t/, /f/, /s/

corresponding to the speaker in the region \timit\train\dr1\fcjf0\. The choice was based on selecting a set of phonemes most representative of the classes of vowels, voiced and unvoiced stops, and fricatives. Each phoneme was extracted according to the phonetic labeling given in TIMIT, and its length was adjusted to be equal to a power of two, as required by the algorithms, by truncating or including a part of the following phoneme. This resulted in signals which varied between 1024 and 2048 samples. The signals retained their original sampling frequency of 16 KHz.

The same overcomplete dictionary was used for all the experiments, and consisted of the wavelet packet dictionary (of depth 11 or 12 depending on the signals length) based on Symmlets with 8 vanishing moments.

3. RESULTS

BP, MP, BOB and MOF were applied to all the phonemes using the Atomizer software developed by Chen.

Figure 1 shows the representations obtained using the different methods applied to the phoneme /b/. The phase plane, sonogram and coefficients are given for each. Figure 2 shows the same using BP applied to the phoneme /s/.

Figure 3 gives the original and reconstructed signal for the phoneme /p/ using the 15 most significant coefficients found with BP. The corresponding waveforms from the dictionary are also shown.

Figure 4 illustrates the mean square error (MSE) obtained for each phoneme and applying each method using the 15 most significant waveforms in each case.

Table 1 gives the percentage of coefficients left after thresholding with 5% of the maximum absolute value.

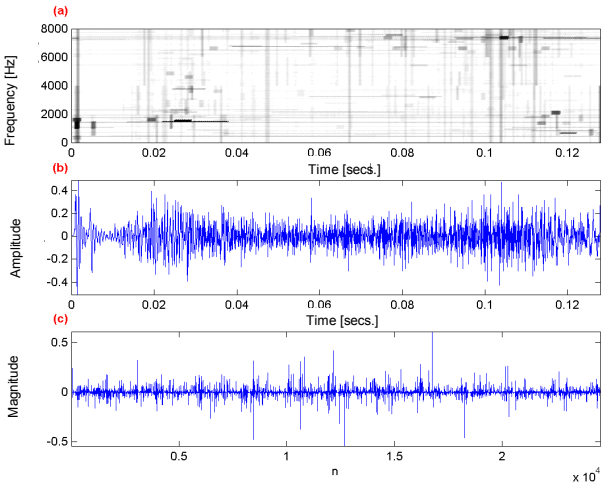

Figure 2: BP Representation of /s/ phoneme.

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner, J. Goddard, A. E. Martínez & F. M. Martínez; "Basis Pursuit applied to Speech Signals"
Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001). Jul, ul, 2001.

4. DISCUSSION AND CONCLUSIONS

In the present paper BP was applied to a set of phonemes chosen from TIMIT using an overcomplete dictionary. The methods of MOF, MP and BOB were also used for comparison.

In traditional representations of speech signals, such as Figure 5b for the phoneme /p/, there is a compromise between simultaneous time–frequency resolution. This can hide the important acoustic cues present in the speech signal. This is contrasted with 5a, given using BP, where it can be seen that a much better resolution for time-frequency events is obtained. This seems to hold true also for the other phonemes used in the paper. It should be noted that the localization depends on the waveforms chosen for the particular dictionary. In the case of this paper a highly overcomplete dictionary was used.

Figure 1 and Table 1 suggest that the methods of BP, MP and BOB give sufficiently sparse representations of the phoneme signals. This is not the case for MOF, as is to be expected. Both BP and MP achieve good localization of the acoustic cues in Figure 1, which are related to the duration of the phoneme /b/ and its voiced characteristics. It should be noted that MP was used with an option to select only up to the first 1000 waveforms; this imposed a sparseness restriction on the results found with it.

Figure 2 shows the case of phoneme /s/, which has a relative uniform frequency content. In this case a non-sparse representation is obtained by all the methods. This is similar to the case of /f/.

Figure 3 shows that for the phoneme /p/, the 15 most significant waveforms are able to capture the acoustic cues of the original signal adequately. It is interesting to observe the 15 waveforms found, also shown in Figure 3. Figure 4 gives a comparative study of the MSE obtained for all phonemes and all methods. BP, MP and BOB are comparable in terms of their approximation accuracy.

This preliminary study of BP applied to the field of speech signals shows possible advantages of the method over traditional approaches. These advantages present themselves in terms of the adequate localization of acoustic cues, obtained using a generally sparse representation. It is necessary to understand the effect of the choice of dictionary on the acoustic cues for speech signals. It is interesting to observe that the methods used in [5,6] also choose the waveforms.

5. ACKNOWLEDGMENTS

The first three authors wish to thank CONACYT for financing this research under Proyecto 31929-A. and the last author wish to thank Universidad Nacional de Entre Ríos for its support.

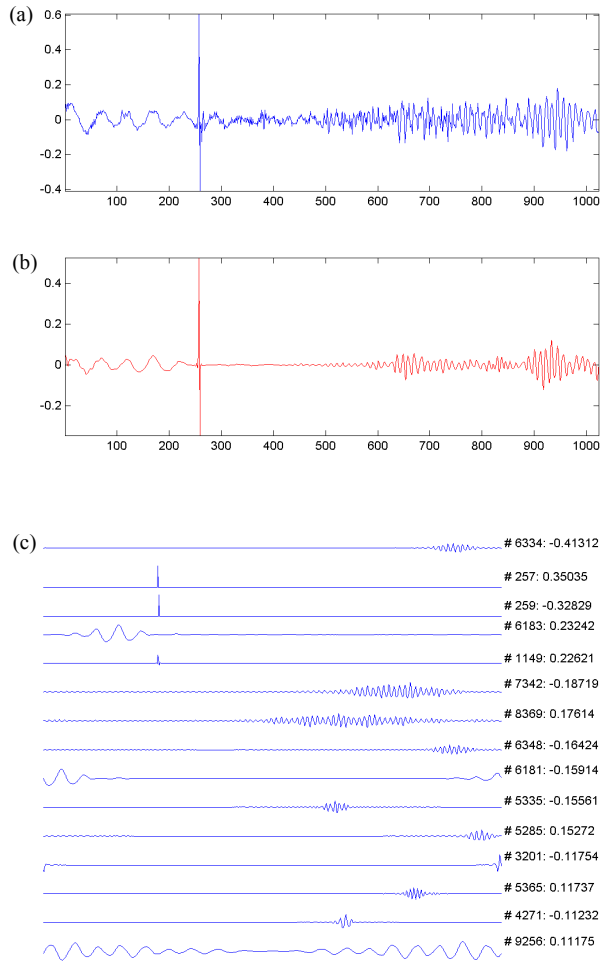

Figure 3: Reconstruction by means of the first 15 BP most important atoms for phoneme /p/: (a) Original Signal, (b) Aproximation, (c) Atoms and coefficient values used in the aproximation.

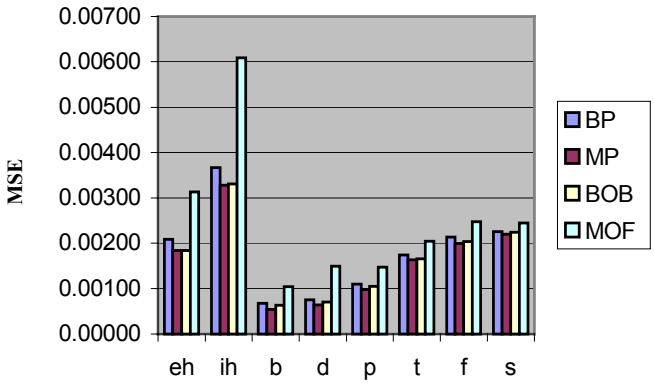

Figure 4: MSE Reconstruction with 15 atoms

Phoneme	BP	MP	BOB	MOF
/eh/	0.501	0.529	0.574	14.880
/ih/	0.417	0.586	0.559	14.190
/b/	0.423	0.370	0.533	11.690
/d/	0.515	0.497	0.630	20.810
/p/	1.713	1.651	2.228	32.420
/t/	1.546	1.306	1.758	27.640
/f/	2.120	1.851	2.201	34.230
/s/	3.715	2.897	3.984	59.770
Mean	1.369	1.211	1.558	26.954

Table 1: Sparseness of the representation

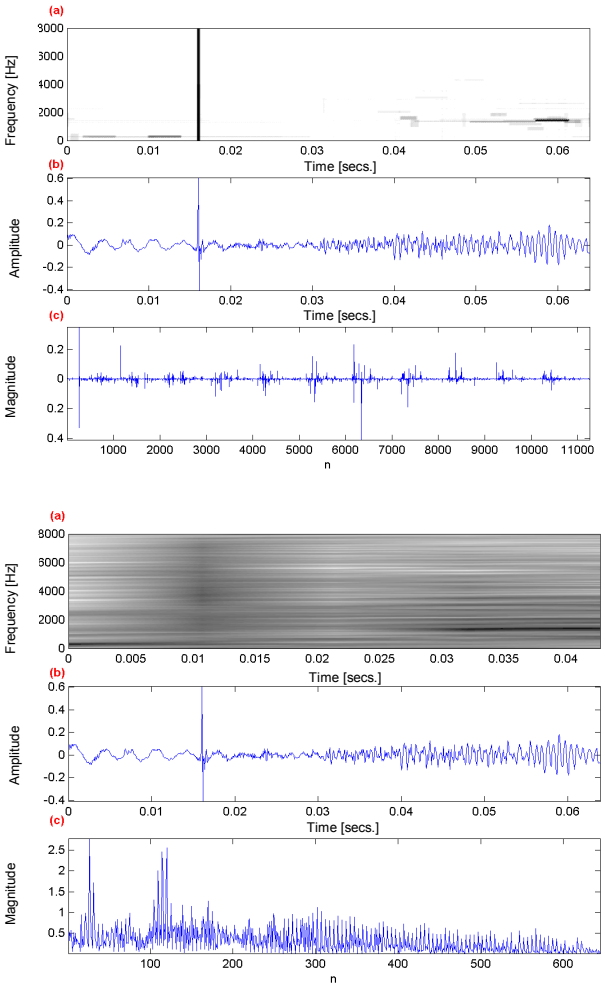

Figure 5: Sonogram and spectrogram of phoneme /p/, (a) Spectrogram, (b) Sonogram, (c) Coefficients.

6. REFERENCES

[1] Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, PA.1992.

[2] S. Mallat and Z. Zhang, "Matching Pursuit in a time-frequency dictionary", IEEE Trans. Signal Proc., vol. 41, pp. 3397-3415, 1993.

[3] R.R.Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best-basis selection", IEEE Trans. Info. Theory, vol. 38, pp. 713-718,1992.

[4] S.S. Chen, D.L. Donoho and M.A. Sanders, "Atomic decomposition by basis pursuit", SIAM Journal on Scientific Computing, vol. 20(1), pp. 33-61, 1999.

[5] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?", Vision Res., 37(23), 1997.

[6] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations", Neural Computation, 12,(2), pp. 337-365, 2000.

[7] M.M. Goodwin and M. Vetterli, “Matching Pursuit and Atomic Signal Models based on Recursive Filter Banks”, IEEE Trans. Signal Proc., Vol 47, No. 7, pp. 1890-1902, 1999.

[8] F. Girosi, "An equivalence between sparse approximation and Support Vector Machines", Neural Computation, 10(6), pp. 1455-1480, 1998.

sinc( ) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. L. Rufiner, J. Goddard, A. E. Martínez & F. M. Martínez, "Basis Pursuit applied to Speech Signals"
Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001). Jul, ul, 2001.

ISBN: 980-07-7558-7