Basis Pursuit Decomposition: An Analysis of Spanish Words

Fabiola M. Martinez[1]*, John C. Goddard*[1]*, Alma E. Martinez*[1]*, Hugo L. Rufiner*[2]

[1]Department of Electric Engineering, Universidad Autonoma Metropolitana, Mexico City, Mexico
[2] Bioengineering Faculty, Universidad Nacional Entre Rios, Argentina
fmml@xanum.uam.mx

Abstract

Time-frequency representations (TF) of speech signals are commonly used to visualize the dynamic behavior. However the choice of the basis functions employed in the TF has an important effect on the number of non-zero coefficients occurring in the representation of the signal. Signals with different morphologies, such as vowels or fricatives, could benefit from the use of different bases for each of their representations. Different methods, such as matching and basis pursuit (bp), have been introduced to find representations of signals using combined bases. One drawback of these techniques is that the computational demands and computing times are liable to be far greater than those of more traditional methods. The present paper analyzes the performance of bp applied to Spanish words. In particular, early stopping times and their relationship to the number of coefficients found are studied. The quality of the reconstructed signals is also considered both quantitatively and qualitatively.

1. Introduction

Time-frequency representations (TF) of speech signals are commonly used to facilitate the visualization of their dynamic behavior. However, the choice of the basis functions employed in the TF has an important effect on the number of non-zero coefficients occurring in the representation of the signal. Furthermore, signals with different morphologies, such as quasi-periodic vowels or noise-like fricatives, could benefit from the exploitation of different bases for each of their representations. Different methods, such as matching pursuit [1] best orthogonal basis [2] and basis pursuit [3], have been introduced to find representations of signals using combined bases (also termed dictionaries). These so called overcomplete representations tend to produce sparse representations, frequently requiring only a few non-zero coefficients [4,5]. One drawback of these techniques is that the computational demands, and computing times, are liable to be far greater than those of more traditional methods. The present paper analyzes the performance of basis pursuit applied to different Spanish words. In particular, early stopping times, and their relationship to the number of coefficients found, are studied. The quality of the reconstructed signals is also considered both quantitatively and qualitatively. It is found that a good approximation is obtained from the sixth iteration of the basis pursuit decomposition algorithm.

The paper is organized as follows: in the methodology section a description of the BP decomposition and evaluation tools is presented. Graphical and numerical results of the speech signals analyzed are shown, and finally a discussion of the relevant aspects found and conclusions are given.

2. Methodology

In this section basis pursuit principles and algorithms are presented. Also the compression procedure, the quality measurements as well as the data set are described.

2.1 Basis Pursuit

In the last few years a number of papers have been devoted to the study of different ways of representing signals using dictionaries of suitable functions [1, 6]. A dictionary *D* is a collection of parameterized waveforms $(\phi_\gamma)_{\gamma\in\Gamma}$, and a representation of the signal *s* in terms of *D* is 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.

It is common to take the criterion of sparseness in the representation of the signal; here this means that a 'few' of the coefficients a_γ in (1) are to be different from zero.

Chen et al [7] propose a method, called Basis Pursuit (BP), which is designed to produce such a sparse representation. The principle of BP is to find a representation of the signal whose coefficients have minimal l_1 norm. A suitable representation is then found by an optimization method. More precisely, if the signal *s* has length *n* and there are *p* waveforms in the dictionary, then the problem to solve is:

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

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

BP requires the solution of a convex optimization problem with inequality constrains of the type

$$\min\ c^T \mathbf{x} \quad \text{subject to } \mathrm{A}\mathbf{x}=\mathbf{b} \text{ and to } \mathbf{x} \geq 0 \qquad (3)$$

where $c^T \mathbf{x}$ is the objective function, A**x**=**b** is the collection of equality constraints, and **x** ≥ 0 is the non-negativity constraint [8]. As Chen et al show, this can be reformulated in terms of linear programming techniques. Furthermore, the problem can be converted to a standard linear program, with only positive coefficients, by making the substitution *a* ← [*u*,*v*] and solving

$$\min l^T [u,v] \quad \text{subject to } [\Phi,-\Phi][u,v] = s,\ 0 \leq u,v \qquad (4).$$

BP is seen as a principle rather than an algorithm. In order to compute the solutions to the optimization problem, the simplex and interior point algorithms can be employed.

BP-simplex algorithm starts from any linearly independent collection of *n* atoms in the dictionary, which is deemed the current decomposition. Then, the current decomposition is iteratively improved by swapping atoms for new atoms in order to improve the objective functional. By application of anti-cycling rules, there is a way to select swaps that guarantees convergence to an optimal solution.

BP-interior point algorithm starts from a solution to the overcomplete representation problem containing many more than *n* nonzero elements. One iteratively modifies the coefficients, maintaining feasibility, but applying a transformation that has the effect of successively sparsifying the vector ***a***. At some iteration the vector ***a*** has less than or equal to *n* significantly nonzero entries that correspond to the atoms that appear in the final solution. In the paper BP-interior algorithm was employed.

In the BP decomposition a wavelet symmlet-8 was employed, as parameterized waveforms, to form the dictionary; one reason for choosing this dictionary is the approximate resemblance it has to averaged voiced speech waveforms.

2.2 Compression procedure

The BP decomposition delivers a large coefficient vector with a few nonzero elements. Since it is desirable to obtain a compact representation of the whole signal as an alternative to the traditional methods, a threshold value was defined to keep those coefficients of highest importance. The speech signal recovery is performed with the set of coefficients retained after applying the threshold. Donoho and Johnstone [9] proposed a global threshold,

$$\lambda \approx \sigma\sqrt{2\log n} \qquad (5).$$

This is applied to threshold the coefficients and then reconstruct an estimated signal. Although this threshold is mostly used in denoising signals that were estimated by discrete wavelet transform, it was found to be a good reference.

2.3 Quality Measurements

In speech enhancement it is important to achieve algorithms which enhances the quality of the perceived signal [10]. From the research in this area several methodologies have been proposed, yet most of them have not been able to succeed because of their lack of correlation with speech perception. The listeners have an intuitive understanding of speech quality, they are capable to detect the speech signal among several noise sources surrounding it. Since this paper is aimed to analyze the recovery of a signal from a reduced set of coefficients, some of the speech quality measurements used in speech enhancement were used in order to assess the performance of the decomposition method.

2.3.1 Objective quality assessment

An objective speech quality measure shows the level of distortion for each frame across time. These methods are based on mathematical measures between the original and the coded speech signal.

a. *Weighted Spectral Slope Measure (WSS)*: developed by Klatt in 1982, it is based on an auditory model; since the human ear selects frequency contents non linearly, this model takes 36 overlapping filters with bandwidths that increase progressively to estimate the smoothed short-time spectrum of the speech signal. This measure finds a weighted difference between the spectral slopes in each band. The weight shows the closeness to a spectral peak or valley and if it is the largest in the spectrum. For a given frame *j* the WSS in decibels is found as

$$d_{WSS}(j) = K_{spl}\left(K - \hat{K}\right) + \sum_{k=1}^{36} w_a(k)\left(S(k) - \hat{S}(k)\right)^2 \quad (6)$$

where K, $\hat{K}$ are related to the overall sound pressure level of the original and coded utterances and K_{SLP} is a parameter which can be varied to increase performance.

b. *Log-Likelihood Ratio Measure (LLR)*: also referred as the Itakura distance and is defined by

$$d_{LLR}\left(\bar{a}_d, \bar{a}_\phi\right) = \log\left(\frac{\bar{a}_d R_\phi \bar{a}_d^T}{\bar{a}_\phi R_\phi \bar{a}_\phi^T}\right) \quad (7)$$

where a_ϕ is the coefficient vector for an original frame of speech obtained using linear prediction (LP), and a_d is the corresponding processed speech coefficient vector.

2.2.2 *Subjective quality assessment*

Subjective measurements are based on the opinion of a listener or a group of listeners of the quality of the utterance. Early methods evaluated speech intelligibility using *modified rhyme test (MRT)*, and *diagnostic rhyme test (DRT)*. Here listeners are presented with rhyming words which differ in their leading consonantal phonemes. Other methods, called quality tests, distinguish among speech systems of high intelligibility. A direct method, the *mean opinion score* (MOS), assesses the degree of quality by rating the speech signal under test on a five-point scale where the subjective impressions of a listener are assigned a numerical value [11].

2.4 Data

Two sets of sentences taken from Latino40 Database [12] were used to carry out the analysis. Isolated Spanish words from sentences spoken by Mexican and Argentinian speakers were selected for the BP decomposition. In order to use a suitable sized dictionary, the speech signals were resampled at 8 Khz and fixed to the same size.

3 Results

The results shown are based on the proper Spanish nouns Cuba and Salinas; the words selected contain voiced elements that can be perceived easily by the human ear. Since processing time is an important aspect in speech applications, studies were conducted applying BP decompositions to the same word using different numbers of iterations. Tables 1 & 2 show the times obtained for these iterations. The increase in processing time after 6 iterations is notable.

Table 1: Processing time (s) for the BP decomposition for different iterations in "Cuba"

Itn	Time (s)
1	4.42
5	35.3
6	49.95
10	152.39
15	322.17
18	359.99
20	594.66
25	608.90

Table 2: Processing time (s) for the BP decomposition for different iterations in "Salinas"

Itn	Time (s)
1	5.11
5	36.53
6	50.57
10	150.82
15	259.85
18	320.98
20	390.85
25	493.93

The threshold values used to obtain the reduced vector of coefficients and recover the signal were set at a values of ½λ, λ and 2λ with λ equal to 0.2776. Tables 3 & 4 show the number of coefficients left after applying the threshold value at different iterations. As expected, the coefficients number increases at lower threshold values. In the case of 2λ, the number of coefficients retained is low and remains constant after the 10th iteration. Fig. 1 shows the graphical results for the noun "Salinas", time processing for the BP decomposition and number of coefficients left after thresholding.

Table 3: Number of coefficients left after applying the threshold in "Cuba"

Itn	½ λ	λ	2 λ
1	39	13	4
5	102	37	12
6	114	45	12
10	117	50	13
15	118	48	13
18	118	51	13
20	119	50	13
25	120	50	13

Table 4: Number of coefficients left after applying the threshold in "Salinas"

Itn	½ λ	λ	2 λ
1	40	9	3
5	129	31	9
6	134	36	9
10	157	42	11
15	160	45	11
18	163	45	11
20	164	45	11
25	165	46	11

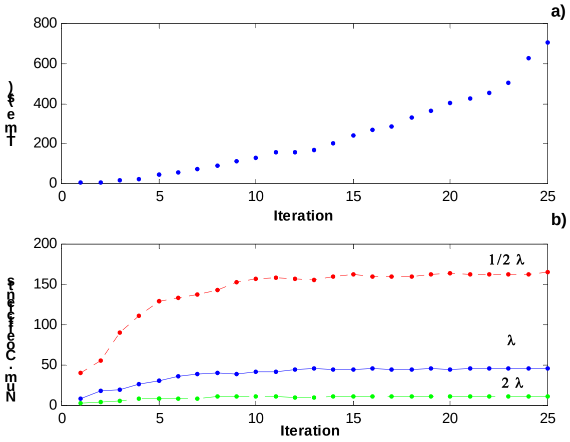

Fig. 1 a) Time elapsed (in seconds) to obtain the BP decomposition of the word “Salinas”; b) Number of coefficients left after applying the threshold value

It can be noticed that there is an important difference between the number of coefficients when a threshold of coefficients when a threshold of ½λ is applied, and the number of coefficients when λ and 2λ are used as threshold.

The quality evaluation of the speech signals recovered from the set of coefficients was performed with the Weighted Spectral Slope Measure (WSS) and the Log-Likelihood Ratio Measure (LLR). Fig. 2 shows the mean value of both measures in the word “Salinas” for the different number of coefficients left after applying the threshold. It can be observed that after approximately the sixth iteration, the change of values diminish for most of the cases; similar results are present in the word “Cuba”. In Fig. 3 the results of the subjective measure mean opinion score (MOS) are shown.

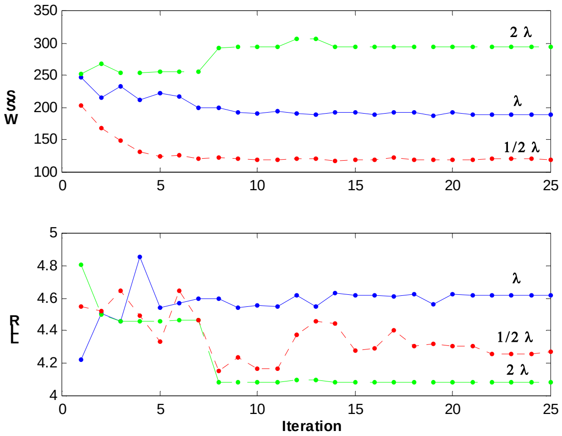

Fig. 2 Mean values of the Weighted Spectral Slope Measure (WSS) and Log-Likelihood Ratio Measure (LLR) for the word “Salinas”

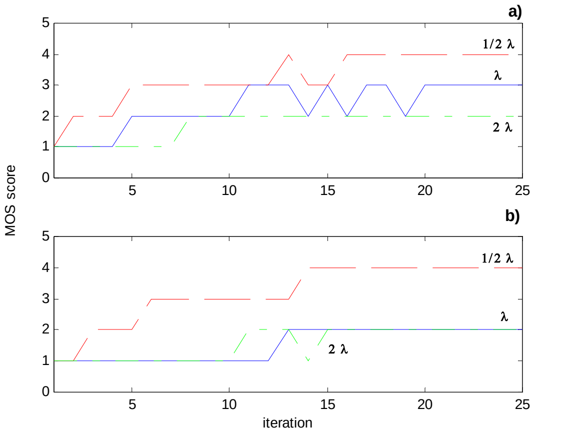

Fig. 3 MOS score obtained from the BP decomposition iterations for the proper noun a) “Cuba” and b) “Salinas”. The categories considered are: excellent (5), good (4), fair (3), poor (2) and unsatisfactory (1).

A BP decomposition procedure produces a sparse representation of the signal, where there are many zero elements; for example in Fig. 4 the BP representation of “Cuba” is shown. Those cells which appear darker are the ones which contribute the most to the information about the signal.

The characteristics in the time and frequency domains of the waveforms recovered using a different number of iterations for the words selected are shown in Fig. 5 through Fig. 8.

4 Discussion

The words taken from the Latino40 database are sampled at 16 Khz. It was necessary to resample them down to 8 Khz and fix them so that all had the same length. The dictionary used is based on the signal length; the more samples the signal has, the more complex and time demanding the BP decomposition is.

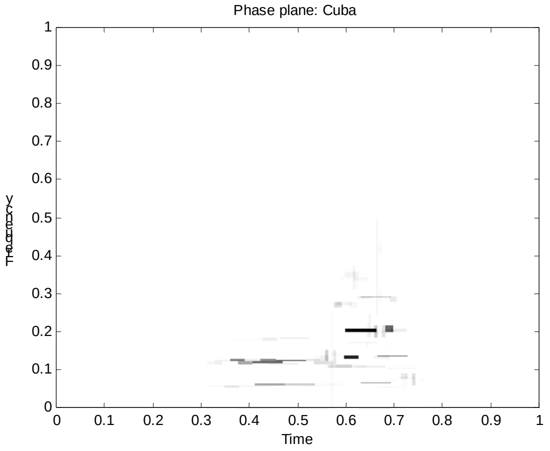

Fig. 4 Time-Frequency representation of the word “Cuba” from the Basis Pursuit decomposition after 6 iterations with a threshold value for the coefficient selection ½ λ.

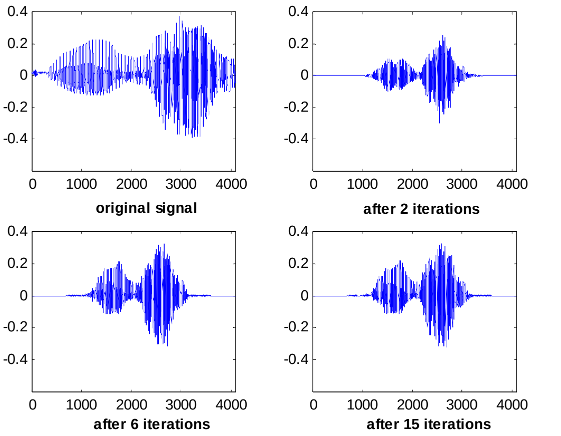

Fig. 5 signal recovery from the BP decomposition for the word "Cuba". The threshold value for the coefficient selection is ½ λ.

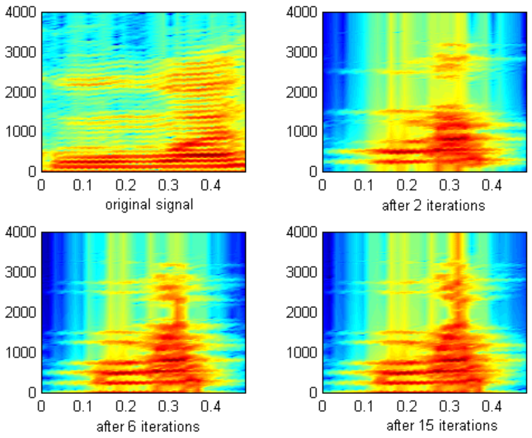

Fig. 6 Spectral representation of the word "Cuba" taken from the reconstructed signal . The threshold value for the coefficient selection is ½ λ.

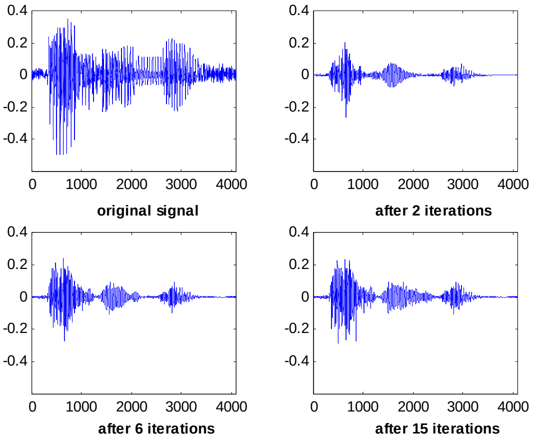

Fig. 7 signal recovery from the BP decomposition for the word "Salinas". The threshold value for the coefficient selection is λ.

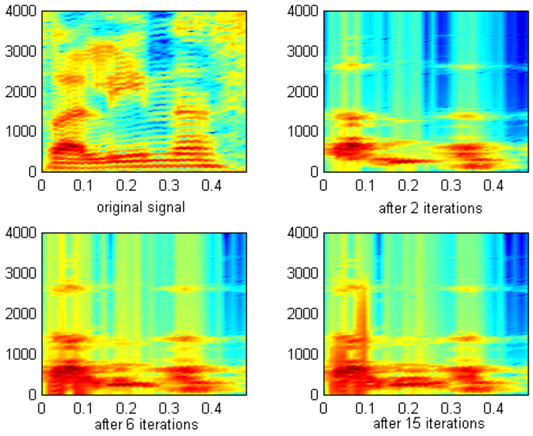

Fig. 8 Spectral representation of the word "Salinas" taken from the reconstructed signal . The threshold value for the coefficient selection is λ.

In this sense structures such as short words and syllables could be suitable to analyze; studies in large databases in English [13] and texts in Spanish [14] have shown that one-syllable words such as prepositions and articles cover an important part of the vocabulary.

A symmlet wavelet was selected to build the dictionary. This waveform has characteristics that are present in many of the voiced sound classes. Although there are several interesting waveforms that could be suitable for the BP decomposition, it might be convenient to define a special type of waveforms. These wavelets would have to present more speech characteristics, for example the sudden changes in energy as it happens to the occlusive sounds.

In the development of the BP decomposition, the amount of time involved is important. From Tables 1 & 2 it can be observed that it increases rapidly with the number of iterations; in order to implement real-time applications this is a severe limitation. This paper analyzed time consumed per BP iteration; from the evaluation measurements it can be established that after the 6th iteration it is possible to get a good tradeoff between time consumed and quality of signal.

The threshold used to obtain the reduced vector of coefficients is a constant value; since the BP decomposition presents a number of waveforms which contain different information, efforts to define suitable thresholding methods need to be addressed.

The evaluation of the recovered signals was performed by objective and subjective quality measurements; these are intended for test intelligibility in speech enhancement. The MOS score shows that the listener seems to hear the same quality of signal from some iteration on. In this kind of tests there are some factors that alter results including auditive fatigue, the number of words presented, instructions given to the listener and his/her personal interpretation of the scores.

In tests conducted, the spectral information shown in Fig. 6 & 8, for different iterations preserved sufficient information for the listener to identify the corresponding voiced sound. The spectrogram is a useful tool for visualizing how the spectral content of the recovered signal appears in each BP iteration, but this type of representation does not show all the time-frequency information that the BP graphics do.

5 Conclusions

A basis pursuit analysis to reconstruct speech signals from a reduced number of coefficients is presented. The results show that it is possible to obtain a good approximation from the sixth Basis Pursuit decomposition iteration onwards. Quality measurements indicate an adequate level of intelligibility of the reconstructed signals from a reduced set of coefficients. The study of basis pursuit applied to speech analysis shows possible advantages of this method over traditional approaches. These advantages are present in terms of the adequate localization of acoustic cues, obtained from a sparse representation. It is necessary to understand the effect of the dictionary on the acoustic cues for speech signals and to develop efficient methods to obtain the BP coefficients.

Future work on the development of atomic decompositions, quality evaluation and thresholding methods will be addressed.

6 Acknowledgements

This work is supported by Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico and Secretaría de Ciencia, Tecnología e Innovación Productiva (SECYT), Argentina through the Program of Scientific and Technological Cooperation.

7 References

[1]. Mallat S., Zhang Z, “Matching Pursuit in a time-frequency Dictionary”, *IEEE Trans. Signal Processing*, vol. 41, pp. 3397-3415, 1993.

[2]. Mallat S. G., *A Wavelet Tour of signal Processing*, Academic Press, Second Edition, 1999.

[3]. Rufiner H. L., Goddard J., Martínez A. E., Martínez F. M., “Basis pursuit applied to speech signals,” I*EEE 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI)*, Orlando, July 2001

[4]. Olshausen B. A., Field D. J., “Emergence of simple cell receptive field properties by learning a sparse code for natural images,” *Nature*, Vol. 381, pp. 607-609, 1996.

[5]. Chen S. & Donoho D., “Basis Pursuit”, *Proc. of 32nd Asilomar Conference on Signals, Systems and Computers*, Nov. 1998 California

[6]. Donoho, D.L. & Johnstone I.M., “Ideal spatial adaption by wavelet shrinkage”, *Biometrika*, 81 (1994), pp. 425-455

[7]. Ephraim Y., “Statistical-model-based speech enhancement systems”. *Proceedings of the IEEE*, 80(10):1526-1555, October 1992.

[8]. Bernstein, Jared, et al. The Latino40 Speech Database. Entropic Research Laboratory, Washington, DC. 1994.

[9]. Wu S., Kingsbury B., Morgan N., Greenberg S., “Incorporating information form syllable-length time scales into automatic speech recognition”, *Proc. of ICASSP*, Seattle EEUU, 1998, pp.721-724.

[10]. Goddard J., Martinez A., MacKynney R., Martinez F., “The syllable structure of Don Quijote” , *Proc. Of 10th International conference on Speech an d Coputer SPECOM 2005*, Patras, 17-19 October 2005, pp. 251-254.

[11]. Nobuhiko Kitawaki, Masaaki Honda & Kenzo Itoh “Speech-quality assessment methods for speech-coding systems”, *IEEE Communications Magazine*, vol. 22, no. 10, October 1984, pp. 26 - 33

[12]. Bernstein, Jared, et al. The Latino40 Speech Database. Entropic Research Laboratory, Washington, DC. 1994.

[13]. Wu S., Kingsbury B., Morgan N., Greenberg S., “Incorporating information form syllable-length time scales into automatic speech recognition”, *Proc. of ICASSP*, Seattle EEUU, 1998, pp.721-724.

[14]. Goddard J., Martinez A., MacKynney R., Martinez F., “The syllable structure of Don Quijote” , *Proc. Of 10th International conference on Speech an d Coputer SPECOM 2005*, Patras, 17-19 October 2005, pp. 251-254.