ANALYSIS OF SPANISH SYNTHESIZED SPEECH SIGNALS USING SPECTRAL AND BASIS PURSUIT REPRESENTATIONS

F. M. Martinez[1], J. C. Goddard, A[2]. M. Martinez[2]
[1]Biomedical Engineering, [2] Computer and Systems, Department of Electric Engineering,
Universidad Autonoma Metropolitana, Mexico City, Mexico

In speech, the sounds involved in an utterance are not produced independently of one another, but rather reflect the result of a complex process of sound concatenation. It is important to study these coarticulation effects in representations of speech signals since, for example, their cues can be helpful in the development of robust speech recognition systems. Representational tools, such as the spectrogram, are useful for visualizing spectral characteristics along the time axis; most of these tools are based on second-order statistics, and it is interesting to consider other methods which might be useful in studying the problem of coarticulation. In recent years, sparse signal representations using suitable dictionaries of functions seem to provide an attractive alternative. With this alternative in mind, the present paper applies spectral and basis pursuit techniques to spanish synthesized signals. The results on a reduced vocabulary show that some prosodic and coarticulatory cues can be obtained from the basis pursuit method compared to the spectral representation.

I. INTRODUCTION

In speech, it is well known that sounds are not produced in isolation, but influence and affect one other. This complex process of sound concatenation is called coarticulation. Coarticulation is related to the speed and the coordination of the movements of the vocal fold. For example, a vowel (V) produced between two nasal consonants (C), such an /m/, presents modifications in its spectral representation due to the effects of these adjacent consonants. In a CV segment with a stop consonant, the spectral representation shows the obstruction of the air flow and then the release of the accumulated air at the moment of the closing and opening of the vocal fold. Graphical representations of these acoustic events sometimes lack clarity when analyzing changes within the same speaker, so different types of analysis and representation methods could prove useful.

In the field of speech processing there are several methods of analysis and representation. The spectrogram is an efficient tool to visualize the spectral characteristics of the signal along the time axis; it is based on second-order statistics. Alternative representations might improve this one, for example, by showing 'hidden' elements hard to identify in the spectrograms. The information obtained by higher order statistics, for example, might be useful since it may provide clues about new model configurations of speech signals that could be better than existing ones [1]. The most important problem with this kind of alternative signal analysis is a lack of understanding of their properties when applied to speech signals. Furthermore, the computational load is usually greater compared to the traditional second-order statistics based analysis [1], [2].

In a preliminary exploration of alternative techniques, this paper presents spectral and basis pursuit representations of speech signals using two different synthesized voices, spanish and mexican; spectrogram and basis pursuit algorithms were applied to the signals in order to study coarticulation effects.

The paper is organized as follows: in the next section the words selected, the effects to be analyzed and the speech representation methods are described. Results are presented and a discussion and conclusions are given.

II. METHODOLOGY

The data selected and the representation methods are described in this section.

A. Data

A set of two spanish words was selected. These words presented different combinations of spanish sounds, e.g. CV segments (stop-vowel or fricative vowel), CVV segments (liquid semivowel-diphthong) or CVC segments (liquid-vowel-fricative, stop-vowel-fricative, nasal-vowel, nasal).

For each word three different synthesized utterances were produced using mbrola [3]; tables 1 and 2 shows the acoustic characteristics of each utterance for the figures presented in this paper. The idea of studying synthesized utterances was because of the control available for specifying certain acoustic events precisely.

Table 1: Acoustic characteristics of the synthesized word 'areas'

phoneme name	duration (ms)	position of the pitch target (%)	pitch value (Hz)
a	108/108/108	30/30/30	130/130/130
r	50/50/50	- /50/ -	- / - / -
e	90/90/90	-/100/-	-/150/130
a	90/90/130	-/99/30	- / - / -

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online)

s	110/110/110	99/99/99	80/80/80

Table 2: Acoustic characteristics of the synthesized word 'gatos'

phoneme name	duration (ms)	position of the pitch target (%)	pitch value (Hz)
g	50/50/50	- / - / -	- / - / -
a	120/90/90	- /90/90	- /100/150
t	85/85/85	- / - / -	- / - / -
o	90/35/35	90/60/60	100/123/350
s	110/110/110	- / - / -	- / - / -

For each utterance the spectrogram and the basis pursuit representations were obtained.

B. Spectrogram

The spectrogram algorithm splits the signal into overlapping segments and applies a window [4]. For each segment it computes the discrete-time Fourier transform for a given length (nfft) to produce an estimate of the short-term frequency contents. The matlab algorithm to compute the spectrograms was applied to the signals [5]. The spectrogram is computed from

$$\Gamma_y(\omega) = 2\pi \sum_{k=-\infty}^{\infty} |C_k|^2 \delta\left(\omega - k\frac{2\pi}{N}\right)$$

where $\Gamma_y(\omega)$ is the power density spectrum for a periodic signal *y(n)*, C_k the associated coefficients [6]

The set of parameters used is the following:

- sampling frequency = 16 KHz
- nfft = 256
- hamming window of nfft length
- no overlap between windows.

C. 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 [7], [8]. 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. In this paper a wavelet symmlet was employed.

An important criterion for choosing a method consists in obtaining a sparse 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 [9] propose a method, called Basis Pursuit (BP), which is designed to produce such a sparse representation. A suitable representation is found by 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.

It can be shown that the problem can be converted to a standard linear program, with only positive coefficients, by making the substitution $a \leftarrow [u,v]$ and solving

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

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

III. Results

Figs 1 and 2 show the spectrogram and BP representations of the synthesized utterances of the word 'areas' produced by a spanish voice. Figs 3 and 4 show the spectrogram and BP representations of the synthesized utterances of the word 'gatos' produced by a mexican voice.

In each of the figures the segment to be analyzed is selected between the lines: the fricative-diphthong (CVV) segment for the word 'areas' and the vowel-stop-vowel (VCV) segment for the word 'gatos'.

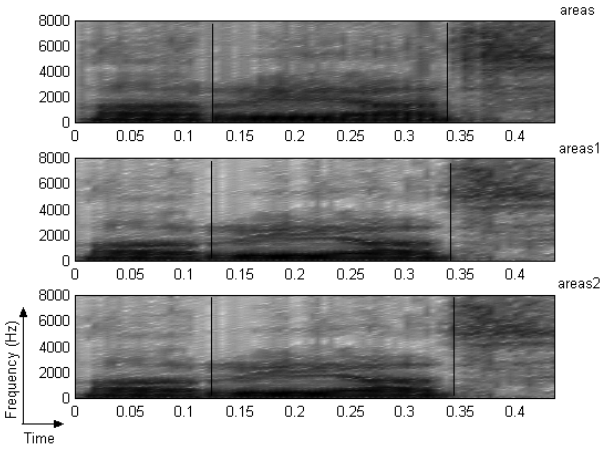

Fig.1 Spectrogram of the word "areas" (spanish synthesized voice)

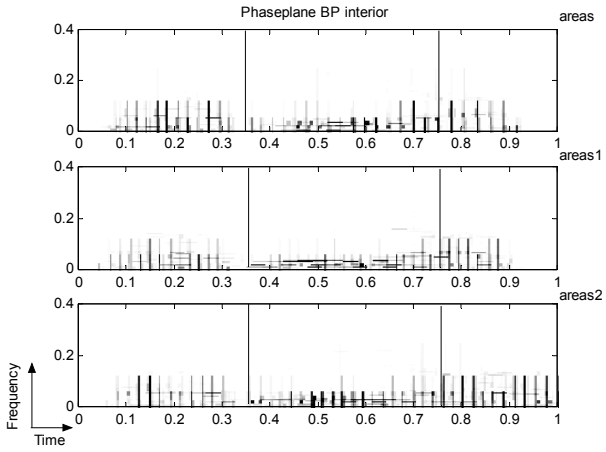

Fig.2 Phaseplane Basis Pursuit of the word "areas" (spanish synthesized voice)

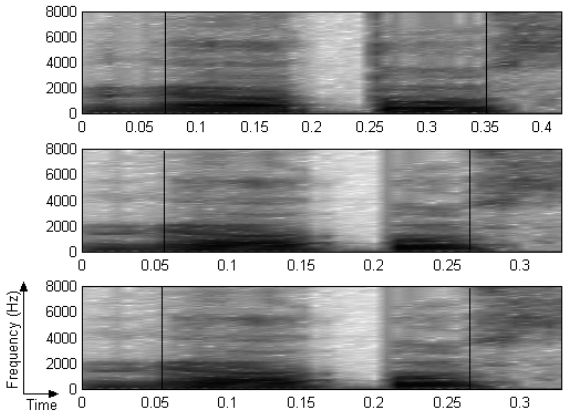

Fig.3 Spectrogram of the word "gatos" (mexican synthesized voice)

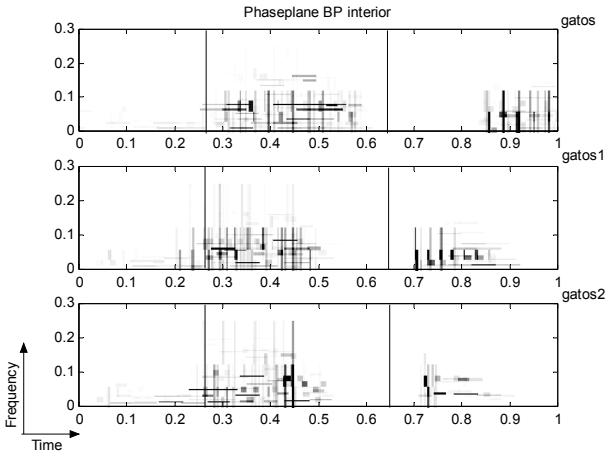

Fig.4 Phaseplane Basis Pursuit of the word "gatos" (mexican synthesized voice)

IV. Discussion

The spectral representations show a very similar behavior for the three utterances in both examples. In the case of 'areas' (Fig 1) the CVV segment presents changes in energy and the effect of the diphthong (/ea/) can be seen in the vowel formants. In Fig 3, the spectral VCV segment remains without changes in the three utterances and the vowel configurations do not seem to be different.

The basis pursuit diagrams (Figs 2 & 4) show noticeable changes among the utterances for the same word. In Fig 2 the diphthong segment BP coefficients are presented in different arrays while in Fig 4, the VCV segment clusters its BP coefficients in different locations of time. It is important to notice that the number and energy level of the coefficients are also different for each utterance even for the same word.

One of the disadvantages that the basis pursuit interior point algorithm presents is the amount of processing time, since there are an important number of calculations involved. Table 3 presents the BP-Interior algorithm time durations (in seconds) for the example words.

Table 3: Time duration (secs) of BP-Interior Algorithm applied to the speech signals

Word	Spanish voice	mexican voice
Areas	539.8	302.47
Gatos	526.3	250.99

V. Conclusion

Spectral and basis pursuit representations were applied to synthesized speech signals in order to identify differences in coarticualtory effects among three utterances of the same word.

The study of basis pursuit applied to speech analysis shows possible advantages of the method over traditional approaches. These advantages present themselves 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.

Further work in the development of atomic decompositions and higher order analysis tools will be addressed in the future.

REFERENCES

[1] C. L. Nikias, A. P. Petropulu, *Higher-order Spectral Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1993, pp. 1–5.

[2] F. Martinez, H. Rufiner, J. Goddard and A. Martinez, "Analysis of Spanish Speech Signals using Higher Order Statistics", IFMBE Proceedings of Third Latinamerican Congress on Biomedical Engineering, Joao Pessoa Brazil, 2004.

[3] MBROLA Project, speech synthesis available in http://tcts.fpms.ac.be/synthesis/mbrola.html

[4] Oppenheim, A.V., and R.W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989, pp. 713-718

[5] *Signal Processing Toolbox (4.0) User´s Guide,* The Math Works, Natick, MA, 1998, pp. 0-363-0-367.

[6] J. R. Deller, J.H. Hansen and J.G. Proakis, *Discrete-time Processing of Speech Signals*, IEEE Press, 2000.

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

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

[9] S.S. Chen, D.L. Donoho and M. A. Saunders, "Atomic Decomposition by Basis Pursuit", SIAM Journal of Scientific Computing, Vol.20, pp. 33-61, 1998.