Proceedings of the Second Joint EMBS/BMES Conference
Houston, TX, USA • October 23-26, 2002

Preserving acoustic cues in Speech Denoising

Hugo L. Rufiner[1,2], Luis F. Rocha[2], John Goddard Close[3]
[1]Facultad de Ingeniería, Univ. Nac. de Entre Ríos, Argentina
[2]Facultad de Ingeniería, Univ. Buenos Aires, Argentina
[3]Depto Ing. Eléctrica, Univ. Autónoma Metropolitana, México

***Abstract*—Recently methods for obtaining sparse representations of signals from overcomplete dictionaries have been studied and some interesting connections between them and biological sensory processing have been observed. Given that a representation of this type is designed to have only a small number of coefficients different from zero, one might intuitively expect that they would also show a certain robustness to additive noise. In the present paper an example, using speech signals, is examined by comparing several denoising techniques, including a simple method proposed by the authors. The method maintains a sparse representation of the signal and is found to preserve important acoustic cues necessary for phoneme identification.**

***Index Terms*—Sparse representation, denoising, speech.**

I. Introduction

Generally speaking, a *sparse representation* (SR) of a signal using an overcomplete dictionary of waveforms signifies that only a 'small' number of the waveforms will be actively used to represent the signal. As an overcomplete dictionary is employed, their will in fact be many possible representations and some criterion has to be selected to decide which representation is actually chosen. This is precisely where representational sparseness can be attained. Typically, conditions are placed on the coefficients in the representation; this can be achieved either through phrasing the problem as one of simultaneously minimizing the approximation of the representation to the signal and the size of the coefficients, in appropriate norms (c.f. [1]), or by assuming them to be statistically independent and choosing a proper prior probability e.g. Laplacian (c.f. [2]). Among the available methods for achieving SR, and of relevance in the present paper, are *Basis Pursuit* (BP), *Matching Pursuit* (MP), or *Best Orthogonal Basis* (BOB). These and similar techniques have been extensively applied to such fields as: 'natural' image analysis, audio and music signals, general biomedical signals and automatic speech recognition (c.f. for example [2], [3]). In particular in [4], a preliminary study of SR was applied to 'clean' speech signals using different dictionaries. There it was found that important acoustic cues in the signals could be preserved with as few as 15 waveforms. This suggests that SR may also retain similar properties for 'noisy' speech signals.

In the present work the behavior of SR in the presence of additive noise for speech signals is examined. A simple method for signal denoising is proposed and compared with other denoising techniques such as Wavelet Denoising (WD) with hard thresholding.

II. Methods and Speech Data

A. Sparse Representations and Noise

A representation of a signal s, in terms of a dictionary Φ, or collection of parameterized waveforms $(\phi_\gamma)_{\gamma\in\Gamma}$, usually signifies a decomposition of the form:

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

where ε is an additive Gaussian noise term. Wavelets Packets (WP) and Cosine Packets (CP) constitute parameterized collections of atoms with a rich variety of behaviors. Different methods have been proposed for obtaining a decomposition in the form of equation (1), and in the present paper SR are of particular interest. In Chen *et al.* a number of artificial examples are given showing the benefits of their methods (BP and BP denoising), in terms of sparsity, super-resolution and noise robustness, compared to the corresponding representations found by MP, MOF and BOB. In [4], BP and MP performed similarly when applied to clean speech signals, whilst the others performed relatively poorly.

B. Denoising Techniques

Denoising data by wavelet thresholding consists in applying a discrete wavelet transform to the original data, thresholding the detail wavelet coefficients, then inverse transforming the thresholded coefficients to obtain the denoised data [5]. The hard threshold involves setting to zero those coefficients whose absolute values are below a certain number, whereas the soft threshold also shrinks the remaining coefficients towards zero. In Chen *et al.* a method for BP denoising (BPDN) is proposed when Gaussian noise is assumed. This corresponds to the solution of $\hat{a} = \arg\min_a \left\{ \|\Phi a - s\|^2 + \lambda |a| \right\}$, where λ is a weight factor. Chen *et al.* also proposed denoising methods using MP (MPDN) and BOB (BOBDN).

The first two authors wish to thank to Facultad de Ingeniería, UNER, and Facultad de Ingeniería, UBA, Argentina for its support
Last author is grateful to CONACYT, Mexico for financing this research under Project 31929-A

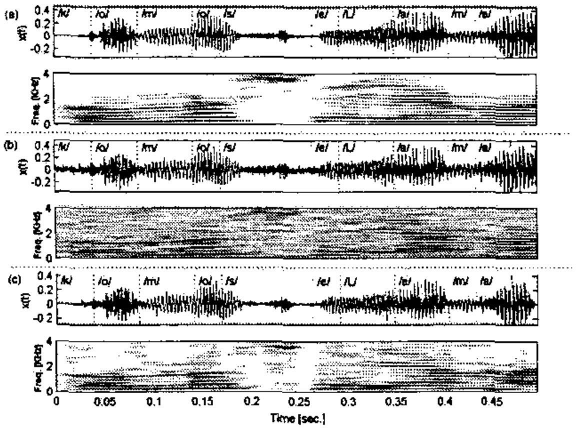

Fig. 1. Sonogram and Spectrogram of: (a) Clean Speech Signal, (b) Noisy Speech Signal (SNR 10 dB WHITE Noise), (c) HDNBP Denoised Speech Signal

The following simple heuristic denoising method (HDN) was also utilized: select an appropriate dictionary and find a SR of the signal using BP or MP. Order the waveforms by coefficient size, and retain the largest ones whose reconstructed signal had the value of it's normalized energy higher than the normalized mean squared error. This represents a compromise between the quality of the approximation and it's sparseness. The number of waveforms chosen was also limited to between 15 and 35.

C. Speech Data

A small subset of the Albayzin speech corpus [6] was used for experiments. The subset consisted of a vocabulary size of 200 words concerning Spanish geography. The clean speech data was contaminated with different kinds of noise: White and babble noise of the NOISEX-92 data base was used [7]. The noise was mixed with the speech data at different signal to noise ratios (SNR).

III. Results and Discussion

The tests with phonemes in [4], suggested that important acoustic phonetic cues could be preserved with as few as 15 atoms. When noise was added practically the same atoms were selected (although in a slightly different order). The tests in [4], also indicated that good choices for the dictionaries were CP and WP Symmlets with 8 vanishing moments. The sonograms and spectrograms of a speech utterance, together with the noisy and HDNBP versions are shown in figure 1. After HDNBP was applied to the noisy signal important acoustic cues are preserved, such as the release of the phoneme /k/, formants and duration of vowels, and the coloration of the fricative /s/. Finally a comparison of SNR_{out} in dB for the denoising techniques were tested under different noisy conditions, and presented in I. BPDN and BOBDN [1] were tested but not included because they never converged on the data. WDN and HDNBP have the bests results. In some cases WDN is better than HDNBP. It should be noted that the HDN parameters were adjusted to maximize denoised speech intelligibility, and it is well known that this is not directly related to SNRs. In most of the cases where WDN outperformed HDN acoustic distortion was present in the form of 'musical noise'. Clearly a measure including these effects should be taken into account. Further, the threshold values were fixed at a value independent from the original SNR, therefore individual tuning could improve on the results obtained.

TABLE I

Comparison of SNR_{out} (dB) of denoising techniques.

Noise	SNR_{in}	HDNBP	HDNMP	MPDN	WDN
Clean	∞	26.56	14.33	4.00	17.43
BABBLE	100	15.13	14.46	4.00	17.37
BABBLE	50	15.19	14.23	4.00	17.12
BABBLE	30	14.89	13.55	3.89	15.85
BABBLE	20	13.61	12.24	3.78	13.18
BABBLE	10	9.75	8.43	3.42	7.78
BABBLE	0	3.75	1.63	2.01	-0.27
WHITE	100	15.16	14.33	4.00	17.42
WHITE	50	15.05	14.19	4.01	17.40
WHITE	30	14.81	14.37	4.00	16.22
WHITE	20	12.18	14.17	3.96	14.09
WHITE	10	10.31	13.18	3.96	11.09
WHITE	0	5.24	10.91	3.16	6.75

Although the method proposed is simple, the important point to note is that sparse representations could offer a way of handling noise in speech processing (as well as in other fields). The authors are presently investigating the application of sparse representations to the preprocessing stage of a robust automatic speech recognition system, as well as the coding of the acoustic cues of speech signals.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] S.G. Mallat, *A Wavelet Tour of signal Processing*, Academic Press, second edition, 1999.

[6] F. Casacuberta, R. García, J. Llisterri, C. Nadeu, J. Prado, and A. Rubio, "Development of a spanish corpora for the speech research," in *Workshop on Int. Coop. & Standardisation of Speech DBs & Speech I/O Assessment Methods*, Chiavari, September 1991, pp. 26–28.

[7] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II NOISEX-92: A database and experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.