Denoising audio signals in the non-negative auditory cortical domain

C. Martínez[1,2⋆], J. Goddard[4], L. Di Persia[1,3], D. Milone[1,3] and H. Rufiner[1,2,3]

[1] Research Center for Signals, Systems and Computational Intelligence (SINC(i))
Dpto. Informática, Facultad de Ingeniería - Universidad Nacional del Litoral
CC217, Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina
[2] Laboratorio de Cibernética, Fac. de Ingeniería-Universidad Nacional de Entre Ríos
[3] CONICET, Argentina
[4] Dpto. de Ingeniería Eléctrica, UAM-Iztapalapa, México

Abstract. In this work, a biologically-inspired denoising method for audio signals is presented, which takes advantage of an approximation to the acoustical signal representation at the auditory cortical level. It is based on an optimal dictionary of atoms, estimated from early auditory spectrograms, and the Basis Pursuit algorithm to approximate the cortical activations. The proposed approach employs non-negative sparse coding to pursue a simplified denoising algorithm which exploits *a priori* information from both clean signals and noise. The method was applied to artificial signals constructed from simultaneous chirps, corrupted with additive noise. Results showed that using an objective quality measure, the method proposed here can improve the audio quality when it is applied to noisy signals.

1 Introduction

In previous years, the classic techniques of signal analysis, for example spectral subtraction, have been applied to audio and speech denoising with relatively good results in controlled conditions [1]. However, it is widely known that the performance of these techniques in adverse environments is far from that of a normal human listener. On the other hand, there is an increasing number of new signal processing paradigms that promise to deal with more complex situations. This is the case with sparse coding and compressed sensing [2]. Their ability to efficiently solve challenging signal representation problems could be exploited in order to develop new audio and speech processing techniques.

⋆ Corresponding author: cmartinez@fich.unl.edu.ar

For many years, researchers in the field of signal processing have benefited from the use of methods inspired by human sensory mechanisms. An examples of this for data encoding are the well-known *perceptual linear prediction* coefficients. Also, auditory representations of audio signals at the cochlea have been widely studied. Different mathematical models have been developed that allow the estimation of the so-called *early auditory spectrogram*. These investigations enabled an accurate modeling of the discharge patterns of the auditory nerve [3].

Although less well known, the underlying mechanisms at the level of the auditory cortex have also been studied and modeled [4]. Given a sound signal, a pattern of activations can be found at the primary auditory cortex, which encodes a series of meaningful cues contained in the signal. This representation seems to use two principles: the need for very few active elements and the statistical independence between them [5]. This behavior of the cortical neurons could be emulated using the fundamentals of *sparse coding* (SC), the *independent component analysis* (ICA) and the notion of *spectro-temporal receptive fields* (STRF), defined as the required optimal stimulus so that an auditory cortical neuron responds with the largest possible activation [6].

In a previous work [7], time-frequency representations of the auditory spectrograms of speech signals were used to estimate an optimal dictionary with the Noise Overcomplete ICA (NOCICA) algorithm [8]. Each two-dimensional atom can be thought as a STRF. Then, the *approximated auditory cortical representation* (AACR) was computed using Matching Pursuit (MP), as the set of activations that form a particular pattern. The AACR approach was applied to a phoneme classification task in clean and noisy conditions, showing the advantages and robustness of the method[5].

In this work, a non-negative matrix factorization (NMF) framework for auditory cortical representation is used in order to propose a novel audio denoising algorithm. NMF is a recently developed family of techniques for finding parts-based, linear representations, of non-negative data [10] (like our auditory spectrograms). This means

[5] This concept of *cortical representation* is slightly different from the one applied in neuroscience, where studies about brain activity involves to analyze the cortical areas that are mainly stimulated by viewing images or listening words [9]

that the data is described by using just additive components, e.g. a weighted sum of only positive STRF atoms. This new model still retains its biological analogy, in spite of the fact that positive STRF implies only non-inhibitory behaviour. The proposed algorithm takes advantage of a mixed overcomplete dictionary that combines atoms estimated from clean and noisy signals. The idea of using a cortical model for audio denoising was also proposed by Shamma in a recent work [4]. The main differences with our approach are that his cortical representation uses the concept of spectrotemporal modulation instead of sparse coding and the way he incorporates information about signal and noise.

The organization of the paper is as follows. Section 2 presents the methods that produce the signal representation in the approximated auditory cortical domain. Section 3 outlines the proposed denoising technique. Section 4 presents the experimental framework and data used in experimentation. Section 5 shows the obtained results and the discussions. Finally, Section 6 summarizes the contributions and outlines future research.

2 Sparse representation of the signal

2.1 Early auditory model

Shamma *et al* proposed a model of audio processing carried out in the auditory system based on psychoacoustic facts found in physiological experiments in mammals. The main idea behind the model is first to obtain a representation of the sound in the auditory system. Then, it further decompose this representation to its spectral and temporal content in the cochlear response [4].

While the complete model of Shamma consists of two stages, in this work only the first stage was used. It first produces the *auditory spectrogram*, an internal cochlear representation of the pattern of vibrations along the basilar membrane. This part of the model is implemented by a bank of $K = 128$ cochlear filters that process the temporal signal s and yield the outputs by convolution

$$x^k_{\text{ch}} = s * h^k, \tag{1}$$

where h^k is the impulse response of the k-th bandpass filter. These outputs are transduced into auditory-nerve patterns using:

$$x_{\text{an}}^k = g_{\text{hc}}\left(\partial_t x_{\text{ch}}^k\right) * \mu_{\text{hc}}, \tag{2}$$

where the derivative ∂_t represents the velocity fluid-cilia coupling (highpass filter effect), g_{hc} the nonlinear compression in the ionic channels (sigmoid function of the channel activations) and μ_{hc} the hair-cell membrane leakage modeling the phase-locking decreasing on the auditory nerve (lowpass filter effect) [4]. Finally, the lateral inhibitory network is approximated by a first-order derivative with respect to the tonotopic (frequency) axis, which is then half-wave rectified as

$$x_{\text{lin}}^k = \max\left(\partial_f x_{\text{an}}^k, 0\right). \tag{3}$$

The output at each frequency band is then obtained by integrating this signal over a short window w, modeling a further loss of phase locking, as

$$x^k = x_{\text{lin}}^k * w. \tag{4}$$

Finally, the time-frequency representation at the early stage is composed in a matrix $\mathbf{x}$ by the set of K frequency-ordered outputs obtained.

2.2 K-SVD algorithm for non-negative sparse coding

The representation of a signal $\mathbf{x} \in \mathbb{R}^N$ (in column vector form) is given by a linear combination of the atoms found by the auditory model, in the form

$$\mathbf{x} = \mathbf{\Phi}\mathbf{a}, \tag{5}$$

where $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$ is the dictionary of M atoms and $\mathbf{a} \in \mathbb{R}^M$ are the coefficients that represents $\mathbf{x}$ in terms of $\mathbf{\Phi}$. The sparsity is included when the solution is restricted to

$$\min_a \|\mathbf{a}\|_0, \tag{6}$$

where $\|\cdot\|_0$ is the l^0 norm that counts the number of non zeros entries of the vector.

In order to find the required representation, two problems have to be jointly solved: the estimation of a sparse representation and the inference of a specialized dictionary. The coefficients found with

methods such as Basis Pursuit (BP) or MP give both atoms and activations with positive and negative values [11]. However, in some applications it could be useful to work only with positive values, thus providing the method with the ability to explain the data from the controlled addition of (positive) atoms. This is the objective of *non-negative matrix factorization* methods.

Aharon *et al* introduced the K-SVD as a generalization of the *k-means* clustering algorithm to solve the representation problem [12]. Moreover, they included a non-negative version of the BP algorithm, named NN-BP, for producing non-negative dictionaries. The method solves the problem

$$\min_{a} \|\mathbf{x} - \mathbf{\Phi}^L \mathbf{a}\| \quad s.t.\ \mathbf{a} \geq 0, \tag{7}$$

where a sub-matrix $\mathbf{\Phi}^L$ –that includes only a selection of the L largest coefficients– is used. In the dictionary updating, this matrix is forced to be positive by calculating

$$\min_{\phi_k, a^k} \|\mathbf{E}^k - \phi_k a^k\| \quad s.t.\ \phi_k, a^k \geq 0, \tag{8}$$

for each one of the k selected coefficients, with $\mathbf{E}^k$ being the error matrix (residual between the signal and its approximation with the k-th atom and respective activation being updated). The final algorithm was called NN-K-SVD [12].

3 Auditory cortical denoising

The main idea of this work is that the audio containing the desired sound and the noise signals can be projected to an approximated auditory cortical space, where the meaningful features of each one could easily be separated. The signals being analyzed could be decomposed into more than one (possibly overcomplete) dictionary containing a rough approach to all the features of interest. More precisely, the method is based on the decomposition of the signal into two parallel STRF dictionaries, one of them estimated from clean signals and the other one from noise signals. The estimation of both dictionaries is carried out after obtaining the respective two-dimensional early auditory spectrograms.

Given that this type of representation is essentially non-negative, a natural way to obtain both the dictionary and the cortical activations is to use an algorithm like the above outlined NN-K-SVD. This is especially true in the case of denoising applications, where forcing non-negativity on both the dictionary and the coefficients may help to find the building blocks of the signals [12]. Although in a previous work we obtained the AACR using the NOCICA algorithm in the context of a classification task [7], preliminary results in a denoising task encourage us to further explore these ideas.

Fig. 1 shows a diagram of the proposed method, which consists of two stages. In the *forward* stage, the auditory spectrogram is first obtained. Then, using a combined dictionary with the most representative atoms of signal and noise, the auditory cortical activations that best represent the noisy signal (including both clean and noisy activations) are calculated by means of the non-negative version of the BP algorithm. In the *backward* stage, the auditory spectrogram is reconstructed by taking the inverse transform from only the coefficients corresponding to the signal dictionary, discarding those of the noise dictionary. Finally, the denoised signal in the temporal domain is obtained by the inverse ear model [4]. The proposed method is named NNCD, which stands for *non-negative cortical denoising*.

The reconstruction of the auditory spectrogram from the cortical response is direct because it only consist of a linear transformation. However, a perfect reconstruction of the temporal signal from the auditory spectrogram is impossible because of the nonlinear operations of the early stage. Nevertheless, objective and subjective quality tests shown that the resulting quality is not degraded [3].

4 Experimental framework

A series of tests were carried out to demonstrate the capabilities of the proposed technique. A first series of experiments were first carried out on artificial "clean" signals constructed by a mixture of chirps and pure tones. Noises with different frequency distributions were additively added to the signals at several signal to noise ratios (SNRs) and a second series of experiments were developed with these data. The proposed technique was then applied to obtain the

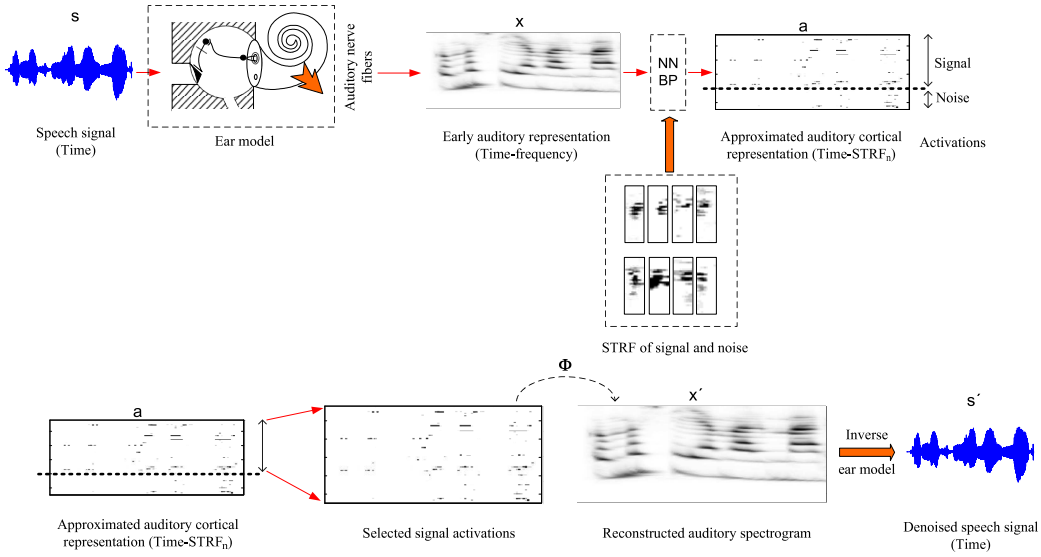

Fig. 1. Diagram of the proposed NNCD method for denoising in the cortical domain. Top: forward stage (cortical representation). Bottom: backward stage (denoised reconstruction).

denoised signals and the performance was evaluated by an objective method: the raw PESQ scores.

4.1 Test signals and noise

A total of 1000 artificial signals were obtained by concatenating 7 different subsignal segments of 64 ms each at a sampling frequency of 8 kHz. Each segment consisted of the random combination of up or down chirps and pure tones. In order to restrict all the possible combinations of these features so a relatively simple dictionary was able to represent them, the spectrogram was divided in two frequency zones, below and above 1200 Hz. Inside each zone only one of the features could occur. Also, the frequency slopes of the chirps are fixed in each zone.

Two kinds of noise with different frequency content were additively mixed. On one hand, white noise, which exhibits a relatively high frequency content with a non-uniform distribution in the early auditory spectrogram (due to its logaritmic frequency scale). On the other hand, speech babble with mainly low frequency content in that

representation. The white noise was generated by a HF radio channel and the babble noise was recorded in a crowded indoor ambient, both taken from the NOISEX-92 database [13].

4.2 Cortical representation

The auditory spectrograms of clean signals were obtained and the training data for the estimation of the dictionaries was extracted as a series of sliding time-frequency windows without overlapping. The same considerations apply to the estimation of noise dictionaries. The dictionaries were generated with 512 atoms of size 64×8 (complete dictionaries). Here, the 64 coefficients correspond to a downsampled version of the original 128 coefficients representing the range 0-4 kHz. The 8 columns correspond each to a window of 8 ms.

From each dictionary, the most active atoms were collected and combined to form dictionaries with 256 atoms containing both clean and noisy features.

4.3 Quality measurement

The PESQ score is an objective quality measure introduced by the International Telecommunication Union (ITU) as a standard for evaluation of speech quality after transmition over communication channels [14]. It uses an auditory representation based on bark scale to compare the original and distorted speech signals. It has been shown to be very well correlated with perceptual tests using MOS [15] and robust speech recognition results [16]. Although only artificial signals were experimented in this work, we decided to use the PESQ score given the similarities found in the energy variation of chirps and speech formants, when analyzing their spectrograms.

The measure is calculated frame by frame, after frame delay alignment and gain compensation, thus the method is insensitive to time-varying delay and scaling. The signals are then compared in the auditory domain using cognitive models to nonlinearly weight the differences and produce two perceptually weighted time-frequency differences: the masking thresholds of the human hearing (D), and the amounts of frequency contents that are introduced by the transmission method (A, manifested as musical noise). They are then

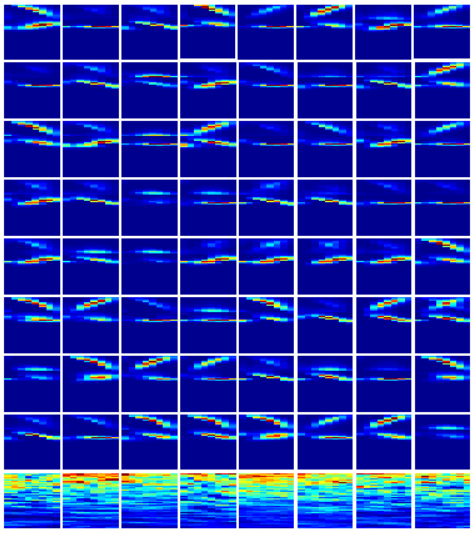

Fig. 2. Example of spectro-temporal receptive fields (STRF) calculated from the early auditory representation of artificial signals and white noise signals, showing the most active atoms of each dictionary. The top 8 rows show the 64 most important STRF for clean signals, whereas the last row show the respective STRF for the noise signals. The dimensions of each atom follow the setup outlined in Section 4.2.

integrated over frequency using different p-norms and combined to produce a single value, the raw PESQ score, defined as $4.5-\alpha D-\beta A$, with $\alpha = 0.1$ and $\beta = 0.0309$. The measure has an ideal value of 4.5 for clean signals with no distortion, and a minimum of -0.5 for the worst case of distortion.

5 Results and discussions

5.1 Non-negative STRF dictionaries

Fig. 2 shows a selection from a dictionary where the 64 most active atoms for chirp signals and 8 atoms for white noise signals are presented. It can be clearly seen the features captured by the STRF corresponding to each dictionary are, respectively, the combination of chirp and pure tones and the noise characteristics that are more prominent in the training signals.

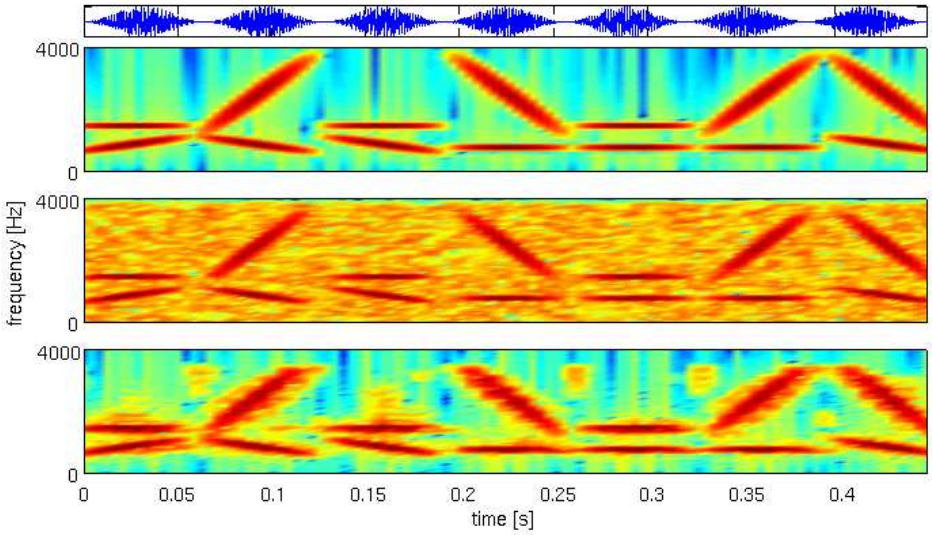

Fig. 3. Example of the denoising of an artifical signal with a combination of 7 windowed segments of random chirps and pure tones. The spectrograms (STFT) of the clean signal (top), a noisy version obtained by the addition of white noise at SNR=0 dB (middle) and the denoised signal (bottom) are shown. The temporal signal at the top of the figure is given as reference.

5.2 Denoising of artificial signals

Our scheme for denoising was applied using the NNCD approach. The reconstruction of the denoised auditory spectrogram was obtained by selecting only the clean atoms from the 32 greatest activations selected by the NN-BP algorithm. Fig. 3 shows a well-known analysis, the short-time Fourier transform (STFT) for a clean (top), noisy with white noise at SNR=0 dB (middle) and denoised signal (bottom), with the temporal signal above the clean spectrogram. In the spectrogram shown at the bottom, the effects of the denoising carried out in the cortical representation can be seen, where the most important features are reconstructed.

Table 1 shows the PESQ scores obtained of denoising the articial signals. For all cases there was an increase in the PESQ score when the NNCD was applied to the noisy signals. The improvement was more marked when the noise energy was higher (SNR=0 dB) and smaller when the signals become cleaner at larger SNR (lower energy of the noise).

Table 1. Raw PESQ scores obtained for artificial signals.

Noise	SNR (dB)	Signal	
		Noisy	Denoised
White	12	1.93	2.16
	6	1.40	2.11
	0	0.69	1.99
Babble	12	1.82	2.05
	6	1.23	2.01
	0	0.56	1.91

The PESQ score for the original (clean) signal after transformation using the auditory model and reconstruction back to the time domain is 2.11. This score measures the distortion from the best quality (PESQ MOS of 4.5) that is introduced by the use of the early auditory model, which is only approximately invertible. Even if the noise is completely removed by the NNCD, there is an intrinsic error introduced by the auditory analysis method.

6 Conclusions

In this work, a biologically-inspired denoising method for sound signals was presented, based on the signal representation at the auditory cortical level. Our approach employs non-negative sparse coding to pursue a simple denoising algorithm which exploits *a priori* information from both clean and noisy signals.

The method was applied to denoising of artificial signals, in the presence of different types and levels of noise. The results demonstrate that the proposed method can improve objective quality measures, mainly in severely degraded signals.

Future direction of research will attempt to optimize the denoising at several SNRs and to explore the capabilities of this technique as a preprocessing stage in robust recognition systems.

Acknowledgements

The authors wish to thank: the ANPCyT, the UNL (with CAI+D 012-72, PAE 37122, PAE-PICT-2007-00052), the UNER (with PID

61111-2 and 6106), and the CONICET from Argentina, for their support.

References

1. Y. Hu and P.C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7-8):588–601, 2007.
2. D.L. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.
3. T. Chiu, P. Ru and S. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118(2):897–906, 2005.
4. N. Mesgarani and S. Shamma. Denoising in the domain of spectrotemporal modulations. *EURASIP Journal on Audio, Speech and Music Processing*, 2007:8 pages, 2007.
5. D. Klein, P. Konig and K. Kording. Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing*, 2003(7):659–667, 2003.
6. H. Rufiner, J. Goddard, L. Rocha and M.E. Torres. Statistical method for sparse coding of speech including a linear predictive model. *Physica A: Statistical Mechanics and its Applications*, 367:231–251, 2006.
7. H. Rufiner, C. Martínez, D. Milone and J. Goddard. Auditory cortical representations of speech signals for phoneme classification. In *MICAI 2007: Advances in Artificial Intelligence*, volume 4827 of *Lecture Notes in Computer Science*, pages 1004–1014. Springer-Verlag, 2007.
8. M. Lewicki and T. Sejnowski. Learning overcomplete representations. In *Proceedings of Advances in Neural Information Processing 10, NIPS '97*, pages 556–562. MIT Press, 1998.
9. T. Mitchell, S. Shinkareva, A. Carlon, K-M. Chang, V. Malave, R. Mason and M. Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320:1191–1195, 2008.
10. P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.
11. S. Chen, D. Donoho and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
12. M. Aharon, M. Elad and A.M. Bruckstein. K-SVD and its non-negative variant for dictionary design. In *Proceedings of the SPIE conference wavelets*, volume 5914, 2005.
13. A. Varga and H. Steeneken. Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.
14. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*, 2001.
15. A.W. Rix, J.G. Beerends, M.P. Hollier and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752, 2001.
16. L. Di Persia, D. Milone, H. Rufiner and M. Yanagida. Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing*, 88(10):2578–2583, 2008.