Denoising sound signals in a bioinspired non-negative spectro-temporal domain

C. E. Martínez\textsuperscript{a,c}, J. Goddard\textsuperscript{b}, L. E. Di Persia\textsuperscript{a,d}, D. H. Milone\textsuperscript{a,d}, H. L. Rufiner\textsuperscript{a,c,d}

\textsuperscript{a}Research Institute for Signals, Systems and Computational Intelligence, sinc(i) Facultad de Ingeniería, Universidad Nacional del Litoral - CONICET
CC217, Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina
\textsuperscript{b}Dpto. de Ingeniería Eléctrica, UAM-Iztapalapa, México
\textsuperscript{c}Laboratorio de Cibernética, Facultad de Ingeniería-Universidad Nacional de Entre Ríos
\textsuperscript{d}CONICET, Argentina

Abstract

The representation of sound signals at the cochlea and auditory cortical level has been studied as an alternative to classical analysis methods. In this work, we put forward a recently proposed feature extraction method called approximate auditory cortical representation, based on an approximation to the statistics of discharge patterns at the primary auditory cortex. The approach here proposed estimates a non-negative sparse coding with a combined dictionary of atoms. These atoms represent the spectro-temporal receptive fields of the auditory cortical neurons, and are calculated from the auditory spectrograms of clean signal and noise. The denoising is carried out on noisy signals by the reconstruction of the signal discarding the atoms corresponding to the noise. Experiments are presented using synthetic (chirps) and real data (speech), in the presence of additive noise. For the evaluation of the new method and its variants, we used two objective measures: the perceptual evaluation of speech quality and the segmental signal-to-noise ratio. Results

*Corresponding author
Email address: cmartinez@fich.unl.edu.ar (C. E. Martínez)
show that the proposed method improves the quality of the signals, mainly under severe degradation.

Key words: approximate auditory cortical representation, sound denoising, non-negative sparse coding

1. Introduction

In previous years, several techniques of signal analysis 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 signal analysis techniques in adverse environments is far from that of a normal human listener [2]. 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 [3, 4]. Their ability to efficiently solve challenging signal representation problems could be exploited in order to develop new audio and speech processing techniques.

For many years, researchers in the field of signal processing have greatly benefited from the use of methods inspired by human sensory mechanisms. Some examples of this for audio and speech encoding were mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) coefficients [5]. Auditory representations of sound at the cochlea have been widely studied. Different mathematical and computational models have been developed that allow the approximate estimation of the so-called early auditory spectrogram [6, 7]. These investigations have enabled an accurate modeling of the discharge patterns of the auditory nerve [8, 9].
Although less known, the underlying mechanisms at the level of the auditory cortex have also been studied and modeled [10]. In experimental conditions—given a sound signal—a pattern of activations can be found at the primary auditory cortex that encodes a series of meaningful cues contained in the signal. This cortical representation seems to use two principles: the need for very few active elements in the representation and the statistical independence between these elements [11]. This behavior of the cortical neurons could be emulated using the fundamentals of sparse coding (SC) [12], the independent component analysis (ICA) [13] and the notion of spectro-temporal receptive fields (STRF). The STRF are defined as the optimal linear filter that convert a time-varying stimulus into the firing rate of an auditory cortical neuron, so that it responds with the largest possible activation [14]. These concepts have led to the development of a number of contemporary auditory models that incorporate different auditory phenomena, for example neural timing information [15], modeling of spectral and temporal content in the cochlear response [9]. A very complete and recent review on biologically-inspired models for speech processing is given in [16].

A number of works have explored the use of auditory models for building robust speech/speaker recognition systems. In [17], a model of auditory perception (PEMO [18]) is used to obtain the features in a digit recognition system, after processed with well-established algorithms for speech enhancement (for example, the Ephraim and Malah estimator [19]). In [20], authors proposed the use of the model of Li [21] as a front-end in a hidden Markov model-based speech recognizer. Here, the speech is first pre-processed with state-of-the-art enhancement algorithms ([19, 22] and others). More recently,
different modifications of the MFCC representation were introduced (noise suppression, temporal masking and others) and compared to standard MFCC and PLP coefficients for speech recognition [23]. As can be seen, these efforts were mainly devoted –differently from our speech enhancement point of view– to build new feature extraction schemes for the recognizers while maintaining standard techniques for the enhancement itself.

In a previous work [24], the approximate auditory cortical representation (AACR) which is a set of activations computed using matching pursuit (MP) on a discrete dictionary of bidimensional atoms, was presented. These atoms represent the STRF of the auditory cortical neurons. The AACR intends to model the global statistical characteristics of the discharge patterns in the auditory cortex, in a phenomenological rather than a physiological way. This technique provides an approximated representation of the speech signal at the auditory cortical level. It has proved to be beneficial with respect to standard spectro-temporal techniques given the fact that at this higher level in the auditory path, some aspects of the acoustic signal that arrives at the eardrum have been reduced or eliminated [16]. Among these superfluous aspects are the temporal variability of the signal and the relative phase of acoustic waveforms [25]. This approach was then applied to a phoneme classification task in both clean and noisy conditions, showing the advantages of the intrinsic robustness of the sparse coding achieved.

In this work, this approach is adapted to a non-negative matrix factorization (NMF) framework. A non-negative auditory cortical representation is used in order to propose a novel sound denoising algorithm. NMF is a recently developed family of techniques for finding parts-based, linear rep-
resentations of non-negative data [26, 27, 28, 29]. These models deal with
the temporal continuity of the signals (which is also found in our auditory
spectrograms), such as slow variation of pitch in speech and music through
consecutive frames, and were applied to monaural source separation. Re-
garding the speech processing applications, semi-supervised/supervised ap-
proaches were reported [30, 31, 32, 33]. In these systems, first statistical
models for clean speech/noise are estimated. Then, the input signal is ana-
lyzed to obtain the denoised version, which is then applied to the recognition
block.

In [34] two sparse dictionaries are obtained directly from spectrograms of
clean speech and noise. Then, a representation of the noisy speech is obtained
by a linear combination of a small number of both type of exemplars, in order
to feed a robust speech recognizer.

In the biologically-inspired context, the NMF use data 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. Thus, positive coe-
ficients could be interpreted as firing rates of excitatory cortical neurons. The
new proposal of a non-negative auditory cortical denoising algorithm also
differs from previous work in the sense that now two STRF dictionary are
estimated from clean and noisy signals separately. Then, the dictionaries are
combined in a mixed dictionary containing the most representative atoms
for each case, obtaining a better representation of the important features of
sound and noise for the denoising stage.

The organization of the paper is as follows. Section 2 presents the meth-
ods that give the signal representation in the approximate auditory cortical domain. Section 3 outlines the proposed technique to perform the signal denoising in this domain. Section 4 presents the experimental framework and data used in the following experimentation. Section 5 shows the obtained results and the discussions. Finally, Section 6 summarizes the contributions of the paper and outlines future research.

2. Sound signal representation

2.1. Early auditory model

Mesgarani and Shamma [10] proposed a model of sound 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, they further decompose this representation to its spectral and temporal content in the cochlear response.

While the complete model of Shamma consists of two stages, in this work only the first stage was used. This stage produces the auditory spectrogram (AS), an internal cochlear representation of the pattern of vibrations along the basilar membrane.

In the following, subscript 'ch' stands for cochlear, 'an' for auditory nerve and 'hc' for hair cell. The first part of the model is implemented by a bank of 128 cochlear filters \( x_{ch} \) that process the temporal signal \( s(t) \) and yield the outputs

\[
x_{ch}^k(t, f) = s(t) \otimes h_{ch}^k(t, f),
\]

(1)
where \( h^k_{ch} \) is the impulse response of the \( k \)-th cochlear filter [10]. This is a bank of overlapping constant-Q (QERB = 5.88) bandpass filters with center frequencies (CF) that are uniformly distributed along a logarithmic frequency axis, over 5.3 octaves (24 filters/octave, 0-4 kHz). The CF of the filter at location \( l \) on the logarithmic frequency axis (in octaves) is defined as

\[
f_l = f_0 2^l \text{ (Hz)},
\]

where \( f_0 \) is a reference frequency of 1 kHz [10]. The quantity and frequency distribution of the filters proved to be satisfactory for the discrimination of important acoustic clues and for an appropriate reconstruction of speech signals [9].

These 128 filter outputs are transduced into auditory-nerve patterns \( x_{an} \) using

\[
x_{an}^k(t, f) = g_{hc} \left( \partial_t x_{ch}^k(t, f) \right) \otimes \mu_{hc}(t),
\]

where \( \partial_t \) 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 \( \mu_{hc} \) the hair-cell membrane leakage modeling the phase-locking decreasing on the auditory nerve (lowpass filter effect) [10]. Finally, the lateral inhibitory network is approximated by a first-order derivative with respect to the tonotopic (frequency) axis [10], which is then half-wave rectified as

\[
x_{lin}^k(t, f) = \max \left( \partial_t x_{an}^k(t, f), 0 \right).
\]

The AS is then obtained by integrating this signal over a short window, modeling a further loss of phase locking. Figure 1 shows an scheme of the auditory model as used in this work.
2.2. Sparse coding of auditory spectrogram

We now suppose that the representation of any bidimensional slide signal \( x \in \mathbb{R}^{m \times n} \) obtained from the early auditory model in (4) is given by a linear combination of atoms representing the STRFs, in the form

\[
x = \Phi a,
\]

(5)

where \( \Phi \in \mathbb{R}^{m \times n \times M} \) is the dictionary of \( M \) bidimensional atoms and \( a \in \mathbb{R}^M \) is the target representation. The 2-D basis functions of the dictionary are vectorized as \( \Phi = [\Phi_1 \ldots \Phi_M] \) with \( \Phi_i \in \mathbb{R}^{mn \times 1} \). Then, (5) can be alternatively written as \( \tilde{x} = \sum_{1 \leq i \leq M} \Phi_i a_i \). The desired sparsity is included when the solution is restricted to

\[
\min_a \|a\|_0,
\]

(6)

where \( \| \cdot \|_0 \) is the \( l^0 \) norm, which counts the number of non-zero entries of the vector. This is a NP-complete problem so several approximations were proposed [35].
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 based on basis pursuit (BP) or MP give both atoms and activations with positive and negative values [36, 37]. 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 (only positive) atoms. This is the objective of non-negative matrix factorization methods.

2.3. NN-K-SVD algorithm

As it was mentioned in Section 1, there are several approaches to obtain a nonnegative atomic sparse decomposition of data. Among them, in this work the method proposed in [38] is selected given its simplicity, excellent performance in other applications (for example, image classification [39]) and the possibility to explicitly set the number of sparse components to use in the approximation.

Aharon et al introduced the K-SVD as a generalization of the k-means clustering algorithm to solve the sparse representation problem given a set of signals $\mathbf{x}$ to be represented [38]. Moreover, they included a non-negative version of the BP algorithm, named NN-BP, for producing non-negative dictionaries. The method solves the problem

$$\min_{\mathbf{a}} \| \mathbf{x} - \Phi^{L} \mathbf{a} \|_2^2 \quad s.t. \; \mathbf{a} \geq 0,$$

(7)

where a sub-matrix $\Phi^{L}$ 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

$$\min_{\vec{\phi}_k, a_k} \| \mathbf{E}^k - \vec{\phi}_k a_k \|_2^2 \quad \text{s.t. } \vec{\phi}_k, a_k \geq 0,$$

(8)

for each one of the $k$ selected coefficients. The error matrix $\mathbf{E}^k$ is the residual between the signal and its approximation with the $k$-th atom $\vec{\phi}_k$ and its respective activation $a_k$ being updated.

The dictionary itself and the activation coefficients are calculated from the SVD of $\mathbf{E}^k = \mathbf{U} \Sigma \mathbf{V}^T$. This decomposition is then truncated to null the negative entries. Finally, the atoms and activations are obtained as the rank-one approximation with the first left and right singular vector as $\vec{\phi}_k = \mathbf{u}_1$ and $a_k = \mathbf{v}_1$. The complete algorithm, called NN-K-SVD for short [38], is illustrated in the Appendix.

3. Denoising methods

3.1. Non negative cortical denoising

The main idea of the proposed method is that sound and noise signals can be projected to an approximate 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 here proposed 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. The estimation of both dictionaries is carried out after obtaining the respective two-dimensional early auditory spectrograms for each type of signals, as was explained before. Given that this type of representation is non-negative, a natural way to obtain both the dictionary and

[10] Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. E. Martínez, J. Goddard, L. Di Persia, D. H. Milone & H. L. Rufiner; "Denoising sound signals in a bioinspired non-negative spectro-temporal domain".

the cortical activations is to use an algorithm that obtains a representation
with non-negative constraints. 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 different type of
signals [38]. Among the several NMF models reported in literature (some of
then summarized in Section 1), we chose for our purposes the above outlined
NN-K-SVD.

Before carrying out the denoising, the dictionaries corresponding to clean
signals and noise should be estimated. They are produced applying twice the
NN-K-SVD algorithm described in Section 2.3, one for each type of signal.
The dictionaries are then rearranged according to the activation for the training samples, in descending order. From these two sets, a combined dictionary containing atoms of signal and noise is used in our approach. This new dictionary is composed by the “most representative” atoms of each previous
dictionary, by selecting those with greater activation.

Fig. 2 shows a diagram of the method here proposed, which consists of
two stages. In the forward stage (Fig. 1.a), the auditory spectrogram is
firstly obtained. Then, using the combined dictionary, 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 (Fig. 1.b), the auditory spectrogram
is reconstructed by taking the inverse transform from only the coefficients
corresponding to the signal dictionary (synthesis). In this way, the denoising
of the signal is carried out in the approximate non-negative auditory cortical
domain. Finally, the denoised signal in the temporal domain is obtained by
Figure 2: Diagram of the NNCD method for denoising in the cortical domain. (a) Forward stage: cortical representation. (b) Backward stage: denoised reconstruction.
the approximate inverse ear model. 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 consists of a linear transformation. However, a perfect reconstruction of the original signal from the auditory spectrogram is impossible because of the nonlinear operations in the earlier described in Section 2.1. Shamma proposed a method to approximately invert the model and showed through objective and subjective quality tests that the resulting quality of this approximate reconstruction is not degraded [9].

The idea of using a cortical model for sound denoising was also proposed by Shamma in a recent work [10]. The main differences with our approach are that his cortical representation uses the concept of spectrotemporal modulation instead of STRF and non-negative sparse coding, and also the way he incorporates information about signal and noise.

3.2. Speech denoising configurations

We propose applying the NNCD in three different scenarios for denoising speech signals degraded by uncorrelated additive noise:

(a) “NNCD speech”: corresponds to the NNCD reconstruction from selected atoms of the speech dictionary, discarding the noise selected atoms.

(b) “Wiener/NNCD noise”: applies a Wiener filter to the noisy signal $y(t)$, where the noise estimation $n'(t)$ is given by the NNCD reconstruction from only selected atoms of the noise dictionary.
(c) “NNCD+Wiener”: applies a Wiener filter to both previously NNCD
estimations of noise $n'(t)$ and speech $s'(t)$.

In cases (b) and (c), the Wiener filter is estimated by means of the Short-
Time Fourier Transform (STFT), as $\frac{|S(\omega, \tau)|^2}{|S(\omega, \tau)|^2 + |N(\omega, \tau)|^2}$. Here, $S(\omega, \tau)$ and
$N(\omega, \tau)$ are the STFT representations of $s(t)$ and $n(t)$ respectively. Note
that in case (c), the Wiener filter is estimated from the speech signal $s'(t)$
instead of $s(t)$ [40, 41]. Fig. 3.2 shows the block diagrams of these configu-

For comparison purposes, different filtering algorithms were also imple-
mented and tested:

- iWiener: the iterative Wiener method [42]. After preliminary exper-
imentation, the number of iterations was fixed at 4.

- apWiener: the speech enhancement based on the use of the A Priori
Signal to Noise ratio in a minimum mean square error estimation, as
given in [43].

- Wavelet: sound denoising using the thresholding of wavelet coefficients.
The parameters of this process were: 5 levels of a Daubechies 8 function,
soft thresholding using the unbiased SURE estimator and rescaling
using a single estimation of level noise based on first-level coefficients
[44].

- mBand: Multi-band spectral subtraction, a method that takes into ac-
count the fact that colored noise affects the speech spectrum differently
at various frequencies [45]. The parameters of the algorithm were fixed
at 6 frequency bands with a linear spacing between bands.
Figure 3: Schematics of the three configurations proposed to apply the NNCD to speech enhancement: (a): NNCD speech only, (b) Wiener filter with noise estimation given by the NNCD, and (c) Wiener filter calculated with the estimation of signal and noise given by the NNCD.
BNMF: a recently proposed Bayesian formulation of nonnegative matrix factorization [33]. First, a mean square error estimator for the speech signal is derived, then it learns the NMF noise model online from the noisy signal (unsupervised speech denoising).

Given the nature and characteristics of the artificial/real signals, the Wavelet denoising was used in the experiments with artificial signals, where mBand and BNMF were used in the experiments with speech data.

4. Experimental framework

A series of experiments were carried out to demonstrate the capabilities of the proposed technique. The first of these were carried out on artificial “clean” sound signals constructed by a mixture of chirps and pure tones. Then a second series of experiments were developed to work with real data consisting of speech signals of complete sentences from a single speaker. Noises with different frequency distributions and non stationary behaviours were additively aggregated to the signals at several signal to noise ratios (SNRs). The proposed technique was then applied to obtain the denoised signals and the performance was evaluated by two objective methods: the perceptual evaluation of speech quality (PESQ) score [46] and the classical segmental signal-to-noise ratio (SNRseg) [47].

4.1. Artificial and real signals and noises

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 that 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. Experiments with this type of signals were designed just to illustrate the operation of the method, also for sanity check and to show the feasibility of the method.

The clean speech data was extracted from a widely-used database in the speech recognition field, the TIMIT corpus [48]. The data used in this work corresponds to the set of 10 speech sentences of the speaker FCJF0 in dialectic region number 1. Sentences have a mean length of 5 seconds.

Two kinds of noise with different frequency content were used. On the one hand, the 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), and on the other hand voice babble and street noises 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 [49]. The street noise corresponds to an outdoor recording and was taken from the Aurora database [50]. In all the experiments, the noise was first conveniently resampled to the same rate and resolution of the clean signals. The noisy signals were obtained by additively mixing the signals at different SNRs.

4.2. Combined clean-noisy dictionary estimation

First, the auditory spectrograms of clean signals were obtained. Then, the training data for the estimation of the dictionaries was extracted by means
of a sliding time-frequency windowing using frames of 64 ms in length with an overlapping of 8 ms.

The dictionaries were generated using complete dictionaries. For the artificial data, 512 atoms of size $64 \times 8$ were calculated. Here, the 64 coefficients correspond to a downsampled version of the original 128 coefficients representing the range 0-4 kHz, while the 8 columns correspond each to a window of 8 ms. For speech data, based on preliminary experiments, the number of columns was reduced to 4, given that with 8 windows the dictionary learning process becomes computationally very intensive. Thus, in this case, the dictionaries have 256 atoms of size $64 \times 4$.

For the artificial data, $1/10$ of the total number of signals was used as training data (100 random selected chirp signals). For the estimation of noise dictionaries, the same ratio of $1/10$ was used as the balance of training/test data. For the speech sentences, a 10-fold leave-one-out method was applied, where each partition consisted on 9 sentences for train and 1 sentence for test.

From each dictionary, the most active atoms were collected. Then, they were combined to form new dictionaries with atoms containing both clean and noisy features. The reported results consist of the mean value obtained for the 10 partitions.

4.3. Denoised signals quality estimation

For the speech denoising experiments, two well-known objective speech quality measures were evaluated: the PESQ score and the segmental signal-to-noise ratio (SNRseg).
The PESQ score is an objective quality measure introduced by the International Telecommunication Union (ITU) as a standard for evaluation of speech quality after transmission over communication channels [46]. 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 mean opinion score (MOS) [51] and robust automatic speech recognition results [52]. 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.

The segmental signal-to-noise ratio is another quality measure here evaluated. It was obtained as the frame-based average SNR value calculated from the original and the processed signals. Here, short segments of 15-20 ms are used (instead of the whole signals). This time domain measure was computed as in [47], using the MATLAB code provided in [53].

5. Results and discussions

5.1. Non-negative STRF dictionaries

Fig. 4 shows a selection of STRFs from a combined dictionary. Here, the most active (best trained) atoms are presented, 64 atoms for chirp signals and 8 atoms for white noise signals.

It can be clearly seen the features captured by the STRFs in each dictionary are the more prominent ones contained in the training signals. For the first group, some atoms (see, for example, number 2, 3 and 4 in the first row) capture portions of pure tones or chirp signals, while others show the combination of them. For the second group, the atoms show mainly the high
Figure 4: Example of spectro-temporal receptive fields (STRF) estimated from the early auditory representation of artificial signals and white noise signals, showing the most active atoms of each dictionary (left). A single atom with axis labels and colorbar is also showed (right). The top 8 rows show the 64 most important STRF for clean signals, whereas the last row shows the respective STRF for the noise signals. The dimensions of each atom follow the setup outlined in Section 4.2.

energy characteristics of the noise signals. Thus, in the context of sparse coding given in Section 2.2, each segment of the input signal can be represented by a linear combination of selected atoms from these dictionary.

5.2. Artificial signals denoising

Our scheme for denoising was applied using the representation discussed above. 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. 5 shows the short-time Fourier transform (STFT) for a clean (top), noisy with white noise at SNR=0 dB...
Figure 5: Example of the denoising of an artificial 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.

Table 1 shows the PESQ scores obtained of denoising the artificial signals. For all cases, there was an increase in the PESQ score when the NNCD was applied to the noisy signals and our method also outperformed the results obtained with the baseline. 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).

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. For reference, the PESQ obtained using the NNCD method in the same conditions as in Table 1 but on clean signal (SNR=∞) was 2.105. The result is almost identical to the one of the auditory model, showing that no additional degradation was introduced. This is because the number of selected coefficients in the NN-K-SVD method is enough to preserve the quality of the reconstructed signal. In this way, the method not only provides a good enhancement in the noisy case but also preserve the signal when there is no noise. The PESQ values greater than the model distortion (for example, 2.16 for white noise at SNR=12 dB) are pointing out that small amount of noise are beneficial for the quality of the signal obtained. This effect might be due to the stochastic resonace, which concern to non-linear systems (like our proposal) \[54\].

In order to demonstrate the benefits of using the auditory representation of the signal, an experiment replacing this model with the short-time Fourier transform was carried out. Here, two dictionaries trained with clean chirp signals and white noise were obtained. Then, the NNCD method was applied in the same conditions as in Table 1 for noisy signals at SNR=0 dB. The PESQ obtained was 1.27, which is better than the wavelet denoising (0.87)
Table 1: Raw PESQ scores obtained for artificial signals. The NNCD scheme applied was the scenario (a) given in Section 3.2. In bold face, the best result obtained for each experimental condition.

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR (dB)</th>
<th>Signal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy</td>
<td>Wavelet</td>
<td>NNCD</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1.93</td>
<td>1.79</td>
</tr>
<tr>
<td>White</td>
<td>6</td>
<td>1.40</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.69</td>
<td>0.87</td>
</tr>
<tr>
<td>Voice babble</td>
<td>12</td>
<td>1.82</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1.23</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.56</td>
<td>0.53</td>
</tr>
<tr>
<td>Model distorsion:</td>
<td>2.11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

but lower than the result obtained using the NNCD method (1.99). This result would be supporting the intrinsic robustness of the sparse representation when using the auditory model.

5.3. Speech denoising

In Fig. 6, a subset of 64 atoms from the dictionary trained with speech data is shown. It can be seen that different particularities of the signals are learned, for example, onset events (see atoms number 1 and 3 in the first row), offset (atom number 5 in the first row), combination of formants (atoms number 2 and 7 in the first row), energy spreading in a wide frequency range possibly given by fricative phonemes (atom number 1 in the last line), etc.

Fig. 7 shows an example of the denoising of real data signals corresponding to speech data. The clean signal corresponds to the sentence /She had
your dark suit in greasy wash water all year/ (shown in the top spectrogram).

The signal is then contaminated with white noise at SNR=0 dB. The effects of the noise can be seen in the middle spectrogram, where almost every important speech feature has been masked by the noise. The denoising scheme, however, is able to recover the most prominent formants and to reduce the energy noise as shown in the bottom spectrogram.

For the measures of PESQ and SNRseg, a 10-fold cross validation procedure was applied by training a dictionary with 9 signals and testing with the remaining one. In each case, white and street noise were added with SNR of 12, 6 and 0 dB. The results are summarized in Table 2 and 3. They show the mean and standard deviation of PESQ and SNRseg scores obtained for the cross validation scheme, being tested on the three different scenarios in...
Figure 7: Example of the auditory cortical denoising result of a speech signal contaminated with white noise at SNR=0 dB. The spectrograms (STFT) of the clean signal (top), the noisy signal (middle) and the denoised reconstructed signal (bottom) are shown. The acoustic signal at the top of the figure is given as reference.
the application of NNCD and compared with different baseline methods (see
Section 3.2). For each experimental condition, the method that obtained the
best denoising quality is emphasized in boldface.

It can be seen that state-of-the-art method performs better only at very
high SNR (12 dB), while the NNCD method achieves good results in re-
alistic conditions when the energy noise increases at lower SNR. Here, our
method obtains the larger differences in the PESQ and SNRseg scores be-
tween the noisy and denoised signals. For example, in the case of white noise
at SNR=0 dB the method improves the PESQ from 1.63 up to 2.12 and
SNRseg from -2.77 to 4.56. With respect to the other denoising methods,
the NNCD approach performs better for both measures, PESQ and SNRseg,
under real and very high non-stationary noise, like the street noise used in
these experiments. As an example, it can be seen an improvement in PESQ
at SNR=0dB from 1.79 up to 2.24 and in SNRSeg from -3.54 up to 3.94. This
type of noise presents a more complex structure, which could be captured by
our approach.

6. Conclusions

A new denoising method of audio signals was presented, inspired by the
biological processing carried out at the primary auditory cortical level. The
method obtains a sparse coding of the spectrogram at cochlea level using
a non-negative approach. The atoms of the dictionary are calculated from
clean signals and noise. Then, the denoising signal is obtained by inverting
the model using only the atoms corresponding to the signal, discarding the
noise activations.
Table 2: Mean raw PESQ scores obtained for speech sentences from the TIMIT corpus. The 'W' and 'S' on the left column stand for White and Street noise. The three scenarios for the NNCD based speech enhancement given in Section 3.2 are denoted as (a), (b) and (c). In bold face, the best quality for each case. For reference, the score for the clean signal after transformation to the cortical domain and reconstruction back to the time domain is 2.15.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Signal</th>
<th>Noisy</th>
<th>iWiener</th>
<th>apWiener</th>
<th>mBand</th>
<th>BNMF</th>
<th>NNCD (a)</th>
<th>NNCD (b)</th>
<th>NNCD (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td></td>
<td>2.25 (0.14)</td>
<td>2.59 (0.15)</td>
<td>2.53 (0.15)</td>
<td><strong>2.66 (0.21)</strong></td>
<td>2.41 (0.10)</td>
<td>2.46 (0.08)</td>
<td>2.31 (0.14)</td>
<td>2.52 (0.08)</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>1.92 (0.13)</td>
<td>2.19 (0.08)</td>
<td>2.17 (0.09)</td>
<td>2.18 (0.12)</td>
<td>2.18 (0.10)</td>
<td>2.26 (0.08)</td>
<td>1.97 (0.12)</td>
<td><strong>2.36 (0.05)</strong></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>1.63 (0.18)</td>
<td>1.86 (0.15)</td>
<td>1.84 (0.16)</td>
<td>1.84 (0.18)</td>
<td>1.80 (0.09)</td>
<td>1.99 (0.13)</td>
<td>1.67 (0.17)</td>
<td><strong>2.12 (0.10)</strong></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>2.57 (0.13)</td>
<td>2.61 (0.13)</td>
<td>2.73 (0.13)</td>
<td><strong>2.86 (0.11)</strong></td>
<td>2.30 (0.14)</td>
<td>2.67 (0.11)</td>
<td>2.65 (0.12)</td>
<td>2.71 (0.11)</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>2.21 (0.10)</td>
<td>2.18 (0.12)</td>
<td>2.39 (0.09)</td>
<td>2.49 (0.11)</td>
<td>2.06 (0.16)</td>
<td>2.45 (0.07)</td>
<td>2.30 (0.09)</td>
<td><strong>2.51 (0.05)</strong></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>1.79 (0.13)</td>
<td>1.76 (0.15)</td>
<td>2.00 (0.10)</td>
<td>2.11 (0.09)</td>
<td>1.82 (0.13)</td>
<td>2.14 (0.08)</td>
<td>1.89 (0.11)</td>
<td><strong>2.24 (0.06)</strong></td>
</tr>
</tbody>
</table>

Table 3: Mean SNRseg obtained for speech sentences from the TIMIT corpus. The 'W' and 'S' on the left column stand for White and Street noise. The three scenarios for the NNCD speech enhancement given in Section 3.2 are denoted as (a), (b) and (c). In bold face, the best result for each condition. For reference, the score for the clean signal after transformation to the cortical domain and reconstruction back to the time domain is 5.41.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Signal</th>
<th>Noisy</th>
<th>iWiener</th>
<th>apWiener</th>
<th>mBand</th>
<th>BNMF</th>
<th>NNCD (a)</th>
<th>NNCD (b)</th>
<th>NNCD (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td></td>
<td>6.98 (3.42)</td>
<td>8.43 (1.82)</td>
<td><strong>10.04 (2.95)</strong></td>
<td>6.91 (1.99)</td>
<td>1.59 (0.30)</td>
<td>5.60 (1.14)</td>
<td>7.63 (3.47)</td>
<td>5.79 (0.90)</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>1.84 (2.54)</td>
<td>4.50 (1.54)</td>
<td>5.14 (2.12)</td>
<td>5.14 (2.56)</td>
<td>1.62 (0.31)</td>
<td>5.21 (0.62)</td>
<td>2.68 (2.52)</td>
<td><strong>5.24 (0.70)</strong></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>-2.77 (2.00)</td>
<td>2.10 (0.85)</td>
<td>0.04 (1.92)</td>
<td>2.25 (0.23)</td>
<td>1.57 (0.16)</td>
<td>3.84 (0.84)</td>
<td>-2.01 (2.04)</td>
<td><strong>4.56 (0.79)</strong></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>7.10 (2.31)</td>
<td>6.33 (1.33)</td>
<td>8.67 (2.23)</td>
<td>7.09 (1.31)</td>
<td>1.54 (0.22)</td>
<td>5.75 (0.79)</td>
<td><strong>8.24 (2.40)</strong></td>
<td>5.68 (0.48)</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>1.93 (2.24)</td>
<td>3.79 (1.05)</td>
<td>4.13 (2.40)</td>
<td>4.52 (1.59)</td>
<td>1.69 (0.36)</td>
<td><strong>5.26 (0.50)</strong></td>
<td>3.51 (2.15)</td>
<td>4.95 (0.36)</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>-3.54 (2.27)</td>
<td>1.71 (0.61)</td>
<td>-1.19 (2.55)</td>
<td>2.37 (1.07)</td>
<td>1.57 (0.30)</td>
<td><strong>3.94 (0.54)</strong></td>
<td>-1.94 (2.23)</td>
<td>3.89 (0.33)</td>
</tr>
</tbody>
</table>
The performance of the method using synthetic and real signals with additive noise was obtained through two objective quality measures. Results showed that our proposed method and its variants can improve the quality of sound signals, specially under severe conditions.

Future research will be devoted to further improve the performance and also investigate the application of this technique in the preprocessing stage of robust classification systems.

Acknowledgements

The authors wish to thank: the Agencia Nacional de Promoción Científica y Tecnológica (with PICT 2010-1730), the Universidad Nacional de Litoral (with CAI+D 2011 #58-511, #58-519, #58-525), the Universidad Nacional de Entre Ríos (with PID NOVEL 6121), the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) from Argentina (with PIP 2011 00284), and the SEP and CONACyT from México (with Program SEP-CONACyT CB-2012-01, No.182432), for their support.

Appendix

The pseudocode for the NN-K-SVD method is showed in Figure 8 [38].

References

Initialization: Set the NN random normalized dictionary $\Phi^{(0)} \in \mathbb{R}^{m \times n \times M}$.

Set $J = 1$ and repeat until convergence.

Sparse coding stage: use the NN version of the Basis Pursuit decomposition algorithm to calculate $a_i$ for $i = 1, \ldots, M$.

$$\min_a \| x - \Phi a \|_2^2 \quad \text{s.t.} \quad \|a\|_0 \leq L \land a \geq 0.$$ 

Dictionary update stage: for $k = 1, \ldots, L$

- Define the samples that use $\phi_k$: $\omega_k = \{i | 1 \leq i \leq M, a_i(k) \neq 0\}$.
- Compute $E_k = x - (\Phi a - \phi_k a(k))$.
- Choose only the columns corresponding to $\omega_k$, and obtain $E_{k\omega_k}$.
- Set $A = E_{k\omega_k}$, 

$$\phi_k = \begin{cases} 0, & u_1(i) < 0 \\ u_1(i), & \text{otherwise} \end{cases}$$

$$a(k) = \begin{cases} 0, & v_1(i) < 0 \\ v_1(i), & \text{otherwise} \end{cases}$$

where $u_1$ and $v_1$ are the first singular vector of $A$.

Repeat $J$ times:

$$\phi = \frac{A a}{a'^a}.$$  

Project: $\phi(i) = \begin{cases} 0, & \phi(i) < 0 \\ \phi(i), & \text{otherwise} \end{cases}$

$$a = \frac{\phi A}{\phi'^\phi}.$$  

Project: $a(i) = \begin{cases} 0, & a(i) < 0 \\ a(i), & \text{otherwise} \end{cases}$

Normalize $\phi_k$.

Set $J = J + 1$.

Figure 8: The NN-K-SVD algorithm.

29


[10] N. Mesgarani and S. Shamma. Denoising in the domain of spectrotemporal...


