

1 Denoising sound signals in a bioinspired
2 non-negative spectro-temporal domain[☆]

3 C. E. Martínez^{*,a,c}, J. Goddard^b, L. E. Di Persia^{a,d}, D. H. Milone^{a,d}, H. L.
4 Rufiner^{a,c,d}

5 ^a*Research Institute for Signals, Systems and Computational Intelligence, sinc(i)*

6 *Facultad de Ingeniería, Universidad Nacional del Litoral - CONICET*

7 *CC217, Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina*

8 ^b*Dpto. de Ingeniería Eléctrica, UAM-Iztapalapa, México*

9 ^c*Laboratorio de Cibernética, Facultad de Ingeniería-Universidad Nacional de Entre Ríos*

10 ^d*CONICET, Argentina*

11 **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

Preprint submitted to *Journal of Digital Signal Processing* (C. E. Martínez) December 26, 2014

show that the proposed method improves the quality of the signals, mainly under severe degradation.

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

14 **1. Introduction**

15 In previous years, several techniques of signal analysis have been applied
16 to audio and speech denoising with relatively good results in controlled con-
17 ditions [1]. However, it is widely known that the performance of these signal
18 analysis techniques in adverse environments is far from that of a normal hu-
19 man listener [2]. On the other hand, there is an increasing number of new
20 signal processing paradigms that promise to deal with more complex situ-
21 ations. This is the case with sparse coding and compressed sensing [3, 4].
22 Their ability to efficiently solve challenging signal representation problems
23 could be exploited in order to develop new audio and speech processing tech-
24 niques.

25 For many years, researchers in the field of signal processing have greatly
26 benefited from the use of methods inspired by human sensory mechanisms.
27 Some examples of this for audio and speech encoding were *mel frequency*
28 *cepstral coefficients* (MFCC) and *perceptual linear prediction* (PLP) coeffi-
29 cients [5]. Auditory representations of sound at the cochlea have been widely
30 studied. Different mathematical and computational models have been devel-
31 oped that allow the approximate estimation of the so-called *early auditory*
32 *spectrogram* [6, 7]. These investigations have enabled an accurate modeling
33 of the discharge patterns of the auditory nerve [8, 9].

34 Although less known, the underlying mechanisms at the level of the au-
35 ditory cortex have also been studied and modeled [10]. In experimental con-
36 ditions –given a sound signal– a pattern of activations can be found at the
37 primary auditory cortex that encodes a series of meaningful cues contained in
38 the signal. This cortical representation seems to use two principles: the need
39 for very few active elements in the representation and the statistical inde-
40 pendence between these elements [11]. This behavior of the cortical neurons
41 could be emulated using the fundamentals of *sparse coding* (SC) [12], the *in-*
42 *dependent component analysis* (ICA) [13] and the notion of *spectro-temporal*
43 *receptive fields* (STRF). The STRF are defined as the optimal linear filter
44 that convert a time-varying stimulus into the firing rate of an auditory cor-
45 tical neuron, so that it responds with the largest possible activation [14].
46 These concepts have led to the development of a number of contemporary
47 auditory models that incorporate different auditory phenomena, for example
48 neural timing information [15], modeling of spectral and temporal content in
49 the cochlear response [9]. A very complete and recent review on biologically-
50 inspired models for speech processing is given in [16].

51 A number of works have explored the use of auditory models for build-
52 ing robust speech/speaker recognition system. In [17], a model of auditory
53 perception (PEMO [18]) is used to obtain the features in a digit recognition
54 system, after processed with well-stablished algorithms for speech enhance-
55 ment (for example, the Ephraim and Malah estimator [19]). In [20], authors
56 proposed the use of the model of Li [21] as a front-end in a hidden Markov
57 model-based speech recognizer. Here, the speech is first pre-processed with
58 state-of-the-art enhancement algorithms ([19, 22] and others). More recently,

59 different modifications of the MFCC representation were introduced (noise
60 suppresion, temporal masking and others) and compared to standard MFCC
61 and PLP coefficients for speech recognition [23]. As can be seen, these ef-
62 forts were mainly devoted –differently from our speech enhancement point
63 of view– to build new feature extraction schemes for the recognizers while
64 mantaining standard techniques for the enhancement itself.

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

80 In this work, this approach is adapted to a non-negative matrix factor-
81 ization (NMF) framework. A non-negative auditory cortical representation
82 is used in order to propose a novel sound denoising algorithm. NMF is a
83 recently developed family of techniques for finding parts-based, linear rep-

84 representations of non-negative data [26, 27, 28, 29]. These models deal with
85 the temporal continuity of the signals (which is also found in our auditory
86 spectrograms), such as slow variation of pitch in speech and music through
87 consecutive frames, and were applied to monaural source separation. Re-
88 garding the speech processing applications, semi-supervised/supervised ap-
89 proaches were reported [30, 31, 32, 33]. In these systems, first statistical
90 models for clean speech/noise are estimated. Then, the input signal is ana-
91 lyzed to obtain the denoised version, which is then applied to the recognition
92 block.

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

97 In the biologically-inspired context, the NMF use data described by using
98 just additive components, e.g. a weighted sum of only positive STRF atoms.
99 This new model still retains its biological analogy, in spite of the fact that
100 positive STRF implies only non-inhibitory behaviour. Thus, positive coeffi-
101 cients could be interpreted as firing rates of excitatory cortical neurons. The
102 new proposal of a non-negative auditory cortical denoising algorithm also
103 differs from previous work in the sense that now two STRF dictionary are
104 estimated from clean and noisy signals separately. Then, the dictionaries are
105 combined in a mixed dictionary containing the most representative atoms
106 for each case, obtaining a better representation of the important features of
107 sound and noise for the denoising stage.

108 The organization of the paper is as follows. Section 2 presents the meth-

109 ods that give the signal representation in the approximate auditory cortical
110 domain. Section 3 outlines the proposed technique to perform the signal de-
111 noising in this domain. Section 4 presents the experimental framework and
112 data used in the following experimentation. Section 5 shows the obtained
113 results and the discussions. Finally, Section 6 summarizes the contributions
114 of the paper and outlines future research.

115 2. Sound signal representation

116 2.1. Early auditory model

117 Mesgarani and Shamma [10] proposed a model of sound processing carried
118 out in the auditory system based on psychoacoustic facts found in physio-
119 logical experiments in mammals. The main idea behind the model is first
120 to obtain a representation of the sound in the auditory system. Then, they
121 further decompose this representation to its spectral and temporal content
122 in the cochlear response.

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

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

$$x_{\text{ch}}^k(t, f) = s(t) \otimes h_{\text{ch}}^k(t, f), \quad (1)$$

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

$$f_l = f_0 2^l \text{ (Hz)}, \quad (2)$$

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

140 These 128 filter outputs are transduced into auditory-nerve patterns x_{an}
 141 using

$$x_{\text{an}}^k(t, f) = g_{\text{hc}} \left(\partial_t x_{\text{ch}}^k(t, f) \right) \otimes \mu_{\text{hc}}(t), \quad (3)$$

142 where ∂_t represents the velocity fluid-cilia coupling (highpass filter effect),
 143 g_{hc} the nonlinear compression in the ionic channels (sigmoid function of
 144 the channel activations) and μ_{hc} the hair-cell membrane leakage modeling
 145 the phase-locking decreasing on the auditory nerve (lowpass filter effect)
 146 [10]. Finally, the lateral inhibitory network is approximated by a first-order
 147 derivative with respect to the tonotopic (frequency) axis [10], which is then
 148 half-wave rectified as

$$x_{\text{lin}}^k(t, f) = \max \left(\partial_f x_{\text{an}}^k(t, f), 0 \right). \quad (4)$$

149 The AS is then obtained by integrating this signal over a short window,
 150 modeling a further loss of phase locking. Figure 1 show an scheme of the
 151 auditory model as used in this work.

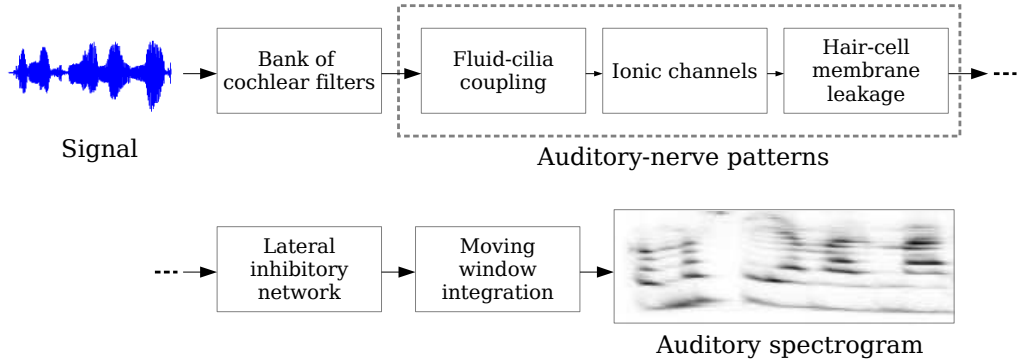


Figure 1: Early auditory model.

152 *2.2. Sparse coding of auditory spectrogram*

153 We now suppose that the representation of any bidimensional slide signal
 154 $\mathbf{x} \in \mathbb{R}^{m \times n}$ obtained from the early auditory model in (4) is given by a linear
 155 combination of atoms representing the STRFs, in the form

$$\mathbf{x} = \mathbf{\Phi} \mathbf{a}, \quad (5)$$

156 where $\mathbf{\Phi} \in \mathbb{R}^{m \times n \times M}$ is the dictionary of M bidimensional atoms and $\mathbf{a} \in$
 157 \mathbb{R}^M is the target representation. The 2-D basis functions of the dictionary
 158 are vectorized as $\mathbf{\Phi} = [\vec{\Phi}_1 \dots \vec{\Phi}_M]$ with $\vec{\Phi}_i \in \mathbb{R}^{[mn] \times 1}$. Then, (5) can be
 159 alternatively written as $\vec{x} = \sum_{1 \leq i \leq M} \vec{\Phi}_i a_i$. The desired sparsity is included
 160 when the solution is restricted to

$$\min_a \|\mathbf{a}\|_0, \quad (6)$$

161 where $\|\cdot\|_0$ is the l^0 norm, which counts the number of non zeroes entries of
 162 the vector. This is a NP-complete problem so several approximations were
 163 proposed [35].

164 In order to find the required representation, two problems have to be
 165 jointly solved: the estimation of a sparse representation and the inference
 166 of a specialized dictionary. The coefficients found with methods based on
 167 *basis pursuit* (BP) or MP give both atoms and activations with positive and
 168 negative values [36, 37]. However, in some applications it could be useful to
 169 work only with positive values, thus providing the method with the ability to
 170 explain the data from the controlled addition of (only positive) atoms. This
 171 is the objective of *non-negative matrix factorization* methods.

172 2.3. NN-K-SVD algorithm

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

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

$$\min_a \|\mathbf{x} - \Phi^L \mathbf{a}\|_2^2 \quad s.t. \quad \mathbf{a} \geq 0, \quad (7)$$

184 where a sub-matrix Φ^L that includes only a selection of the L largest coeffi-
 185 cients is used. In the dictionary updating, this matrix is forced to be positive

186 by calculating

$$\min_{\vec{\phi}_k, a^k} \|\mathbf{E}^k - \vec{\phi}_k a^k\|_2^2 \quad s.t. \quad \vec{\phi}_k, \vec{x}^k \geq 0, \quad (8)$$

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

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

196 3. Denoising methods

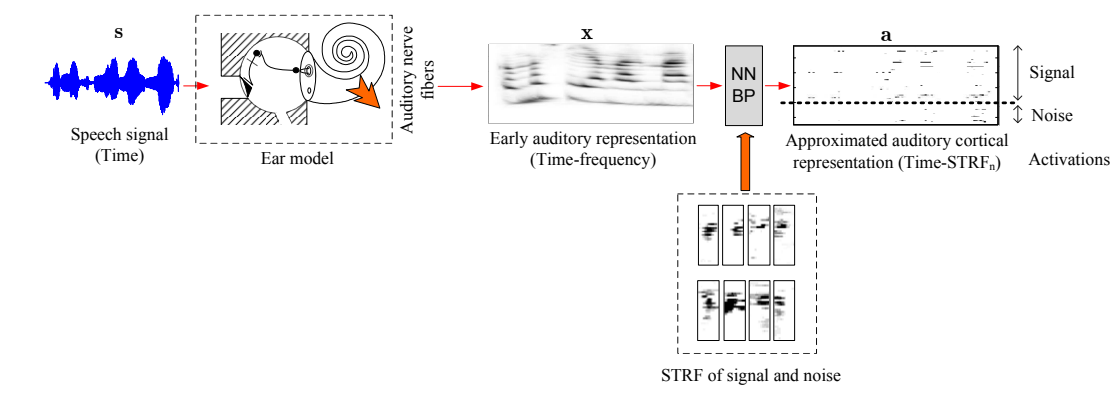
197 3.1. Non negative cortical denoising

198 The main idea of the proposed method is that sound and noise signals can
199 be projected to an approximate auditory cortical space, where the meaning-
200 ful features of each one could easily be separated. The signals being analyzed
201 could be decomposed into more than one (possibly overcomplete) dictionary
202 containing a rough approach to all the features of interest. More precisely,
203 the method here proposed is based on the decomposition of the signal into
204 two parallel STRF dictionaries, one of them estimated from clean signals and
205 the other one from noise. The estimation of both dictionaries is carried out
206 after obtaining the respective two-dimensional early auditory spectrograms
207 for each type of signals, as was explained before. Given that this type of rep-
208 resentation is non-negative, a natural way to obtain both the dictionary and

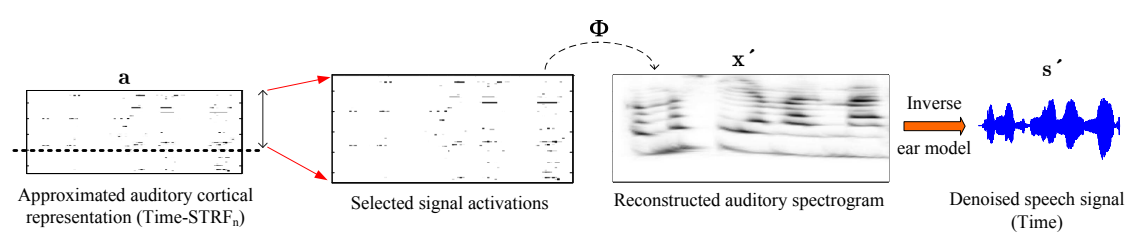
209 the cortical activations is to use an algorithm that obtains a representation
210 with non-negative constraints. This is especially true in the case of denois-
211 ing applications, where forcing non-negativity on both the dictionary and
212 the coefficients may help to find the building blocks of the different type of
213 signals [38]. Among the several NMF models reported in literature (some of
214 then summarized in Section 1), we chose for our purposes the above outlined
215 NN-K-SVD.

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

224 Fig. 2 shows a diagram of the method here proposed, which consists of
225 two stages. In the *forward* stage (Fig. 1.a), the auditory spectrogram is
226 firstly obtained. Then, using the combined dictionary, the auditory cortical
227 activations that best represent the noisy signal (including both clean and
228 noisy activations) are calculated by means of the non-negative version of the
229 BP algorithm. In the *backward* stage (Fig. 1.b), the auditory spectrogram
230 is reconstructed by taking the inverse transform from only the coefficients
231 corresponding to the signal dictionary (synthesis). In this way, the denoising
232 of the signal is carried out in the approximate non-negative auditory cortical
233 domain. Finally, the denoised signal in the temporal domain is obtained by



(a)



(b)

Figure 2: Diagram of the NNCD method for denoising in the cortical domain. (a) Forward stage: cortical representation. (b) Backward stage: denoised reconstruction.

234 the approximate inverse ear model. The proposed method is named NNCD,
235 which stands for *non-negative cortical denoising*.

236 The reconstruction of the auditory spectrogram from the cortical response
237 is direct because it only consists of a linear transformation. However, a
238 perfect reconstruction of the original signal from the auditory spectrogram
239 is impossible because of the nonlinear operations in the earlier described in
240 Section 2.1. Shamma proposed a method to approximately invert the model
241 and showed through objective and subjective quality tests that the resulting
242 quality of this approximate reconstruction is not degraded [9].

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

248 3.2. *Speech denoising configurations*

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

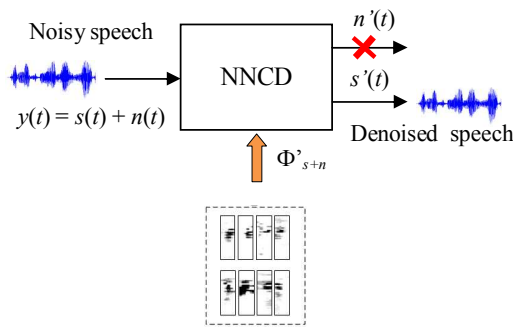
- 251 (a) “NNCD speech”: corresponds to the NNCD reconstruction from se-
252 lected atoms of the speech dictionary, discarding the noise selected
253 atoms.
- 254 (b) “Wiener/ NNCD noise”: applies a Wiener filter to the noisy signal $y(t)$,
255 where the noise estimation $n'(t)$ is given by the NNCD reconstruction
256 from only selected atoms of the noise dictionary.

257 (c) “NNCD+Wiener”: applies a Wiener filter to both previously NNCD
258 estimations of noise $n'(t)$ and speech $s'(t)$.

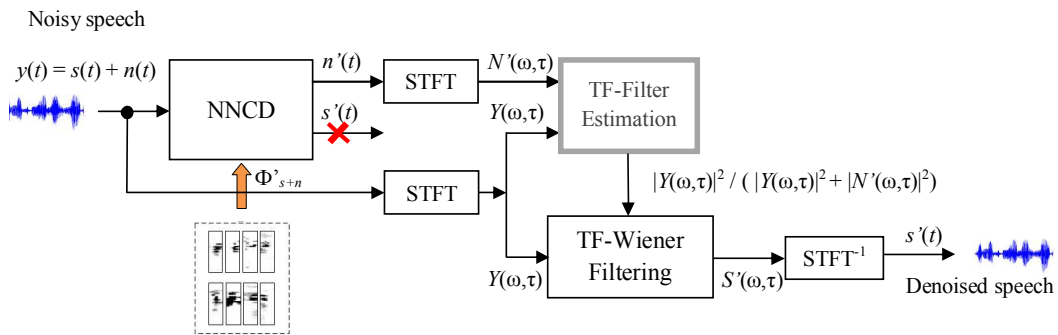
259 In cases (b) and (c), the Wiener filter is estimated by means of the Short-
260 Time Fourier Transform (STFT), as $\frac{|S(\omega, \tau)|^2}{|S(\omega, \tau)|^2 + |N(\omega, \tau)|^2}$. Here, $S(\omega, \tau)$ and
261 $N(\omega, \tau)$ are the STFT representations of $s(t)$ and $n(t)$ respectively. Note
262 that in case (c), the Wiener filter is estimated from the speech signal $s'(t)$
263 instead of $s(t)$ [40, 41]. Fig. 3.2 shows the block diagrams of these configu-
264 rations.

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

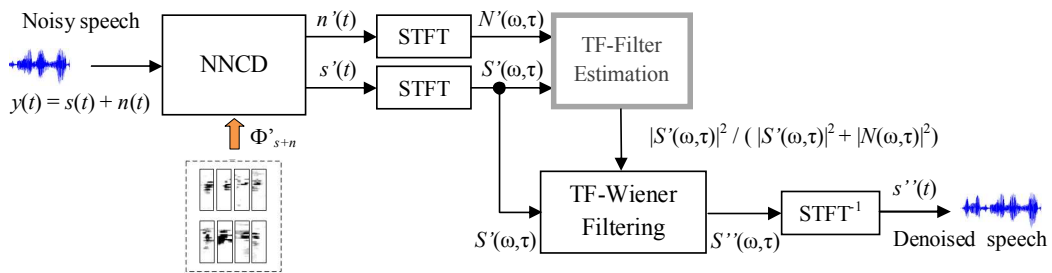
- 267 • iWiener: the iterative Wiener method [42]. After preliminary experi-
268 mentation, the number of iterations was fixed at 4.
- 269 • apWiener: the speech enhancement based on the use of the *A Priori*
270 *Signal to Noise ratio* in a minimum mean square error estimation, as
271 given in [43].
- 272 • Wavelet: sound denoising using the thresholding of wavelet coefficients.
273 The parameters of this process were: 5 levels of a Daubechies 8 function,
274 soft thresholding using the unbiased SURE estimator and rescaling
275 using a single estimation of level noise based on first-level coefficients
276 [44].
- 277 • mBand: Multi-band spectral subtraction, a method that takes into ac-
278 count the fact that colored noise affects the speech spectrum differently
279 at various frequencies [45]. The parameters of the algorithm were fixed
280 at 6 frequency bands with a linear spacing between bands.



(a)



(b)



(c)

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.

281 • BNMF: a recently proposed Bayesian formulation of nonnegative ma-
282 trix factorization [33]. First, a mean square error estimator for the
283 speech signal is derived, then it learns the NMF noise model online
284 from the noisy signal (unsupervised speech denoising).

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

288 4. Experimental framework

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

300 4.1. Artificial and real signals and noises

301 A total of 1000 artificial signals were obtained by concatenating 7 differ-
302 ent subsignal segments of 64 ms each at a sampling frequency of 8 kHz. Each
303 segment consisted of the random combination of up or down chirps and pure

304 tones. In order to restrict all the possible combinations of these features so
305 that a relatively simple dictionary was able to represent them, the spectro-
306 gram was divided in two frequency zones, below and above 1200 Hz. Inside
307 each zone only one of the features could occur. Also, the frequency slopes of
308 the chirps are fixed in each zone. Experiments with this type of signals were
309 designed just to illustrate the operation of the method, also for sanity check
310 and to show the feasibility of the method.

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

315 Two kinds of noise with different frequency content were used. On the
316 one hand, the white noise, which exhibits a relatively high frequency content
317 with a non-uniform distribution in the early auditory spectrogram (due to its
318 logarithmic frequency scale), and on the other hand voice babble and street
319 noises with mainly low frequency content in that representation. The white
320 noise was generated by a HF radio channel and the babble noise was recorded
321 in a crowded indoor ambient, both taken from the NOISEX-92 database [49].
322 The street noise corresponds to an outdoor recording and was taken from the
323 Aurora database [50]. In all the experiments, the noise was first conveniently
324 resampled to the same rate and resolution of the clean signals. The noisy
325 signals were obtained by additively mixing the signals at different SNRs.

326 *4.2. Combined clean-noisy dictionary estimation*

327 First, the auditory spectrograms of clean signals were obtained. Then, the
328 training data for the estimation of the dictionaries was extracted by means

329 of a sliding time-frequency windowing using frames of 64 ms in length with
330 an overlapping of 8 ms.

331 The dictionaries were generated using complete dictionaries. For the arti-
332 ficial data, 512 atoms of size 64×8 were calculated. Here, the 64 coefficients
333 correspond to a downsampled version of the original 128 coefficients repre-
334 senting the range 0-4 kHz, while the 8 columns correspond each to a window
335 of 8 ms. For speech data, based on preliminary experiments, the number of
336 columns was reduced to 4, given that with 8 windows the dictionary learn-
337 ing process becomes computationally very intensive. Thus, in this case, the
338 dictionaries have 256 atoms of size 64×4 .

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

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

349 *4.3. Denoised signals quality estimation*

350 For the speech denoising experiments, two well-known objective speech
351 quality measures were evaluated: the PESQ score and the segmental signal-
352 to-noise ratio (SNRseg).

353 The PESQ score is an objective quality measure introduced by the In-
354 ternational Telecommunication Union (ITU) as a standard for evaluation of
355 speech quality after transmission over communication channels [46]. It uses
356 an auditory representation based on bark scale to compare the original and
357 distorted speech signals. It has been shown to be very well correlated with
358 perceptual tests using *mean opinion score* (MOS) [51] and robust automatic
359 speech recognition results [52]. The measure has an ideal value of 4.5 for
360 clean signals with no distortion, and a minimum of -0.5 for the worst case of
361 distortion.

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

367 5. Results and discussions

368 5.1. Non-negative STRF dictionaries

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

372 It can be clearly seen the features captured by the STRFs in each dic-
373 tionary are the more prominent ones contained in the training signals. For
374 the first group, some atoms (see, for example, number 2, 3 and 4 in the first
375 row) capture portions of pure tones or chirp signals, while others show the
376 combination of them. For the second group, the atoms show mainly the high

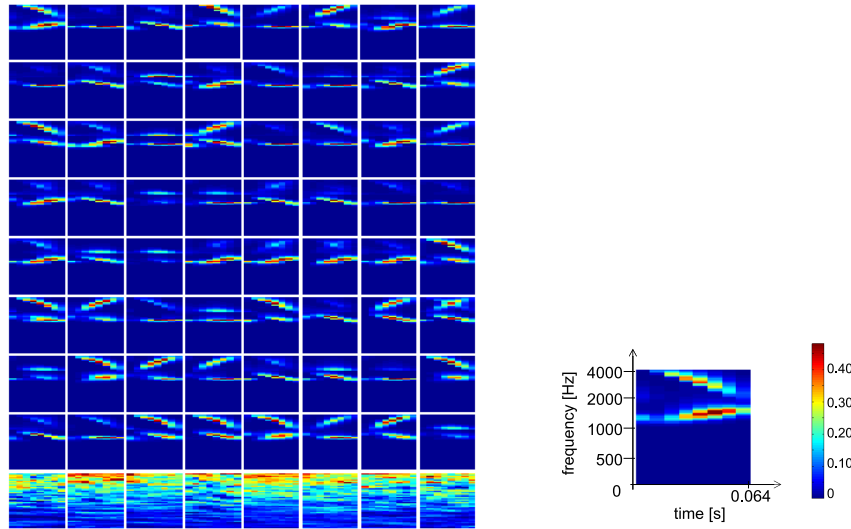


Figure 4: Example of spectro-temporal receptive fields (STRF) estimated from the early auditory representation of artificial signals and white noise signals, showing the most active atoms of each dictionary (left). A single atom with axis labels and colorbar is also showed (right). The top 8 rows show the 64 most important STRF for clean signals, whereas the last row show the respective STRF for the noise signals. The dimensions of each atom follow the setup outlined in Section 4.2.

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

380 5.2. Artificial signals denoising

381 Our scheme for denoising was applied using the representation discussed
 382 above. The reconstruction of the denoised auditory spectrogram was ob-
 383 tained by selecting only the clean atoms from the 32 greatest activations
 384 selected by the NN-BP algorithm. Fig. 5 shows the short-time Fourier
 385 transform (STFT) for a clean (top), noisy with white noise at SNR=0 dB

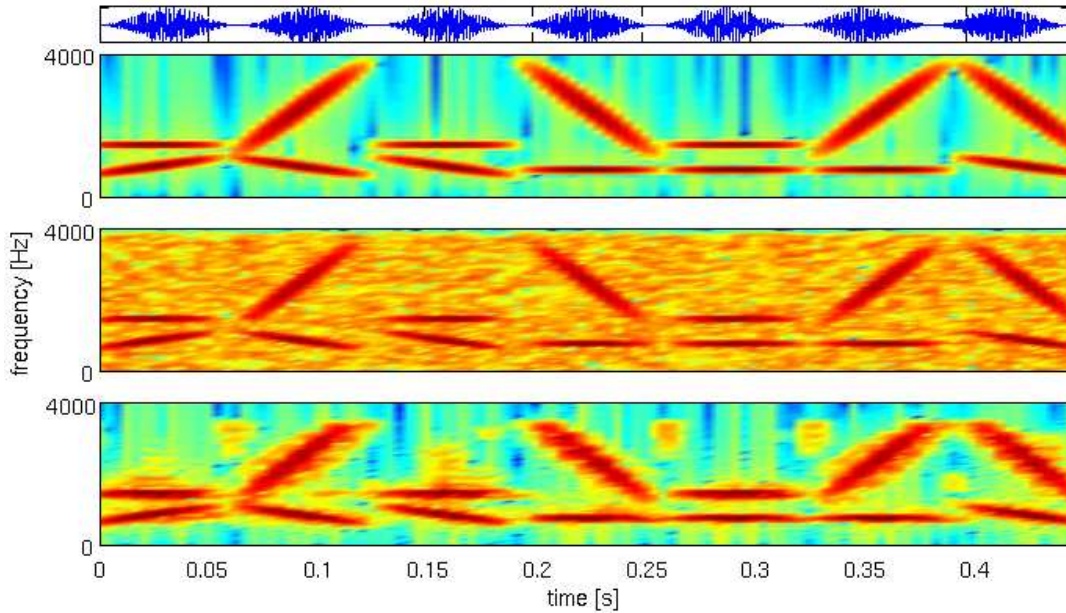


Figure 5: Example of the denoising of an artificial signal with a combination of 7 windowed segments of random chirps and pure tones. The spectrograms (STFT) of the clean signal (top), a noisy version obtained by the addition of white noise at SNR=0 dB (middle) and the denoised signal (bottom) are shown. The temporal signal at the top of the figure is given as reference.

386 (middle) and denoised signal (bottom), with the temporal signal above the
 387 clean spectrogram. In the spectrogram shown at the bottom, the effects of
 388 the denoising carried out in the cortical representation by the NNCD can be
 389 seen, where the most important features are reconstructed.

390 Table 1 shows the PESQ scores obtained of denoising the artificial signals.
 391 For all cases, there was an increase in the PESQ score when the NNCD was
 392 applied to the noisy signals and our method also outperformed the results
 393 obtained with the baseline. The improvement was more marked when the
 394 noise energy was higher (SNR=0 dB) and smaller when the signals become

395 cleaner at larger SNR (lower energy of the noise).

396 The PESQ score for the original (clean) signal after transformation using
397 the auditory model and reconstruction back to the time domain is 2.11.
398 This score measures the distortion from the best quality (PESQ MOS of
399 4.5) that is introduced by the use of the early auditory model, which is
400 only approximately invertible. Even if the noise is completely removed by
401 the NNCD, there is an intrinsic error introduced by the auditory analysis
402 method. For reference, the PESQ obtained using the NNCD method in the
403 same conditions as in Table 1 but on clean signal ($\text{SNR}=\infty$) was 2.105. The
404 result is almost identical to the one of the auditory model, showing that
405 no additional degradation was introduced. This is because the number of
406 selected coefficients in the NN-K-SVD method is enough to preserve the
407 quality of the reconstructed signal. In this way, the method not only provides
408 a good enhancement in the noisy case but also preserve the signal when there
409 is no noise. The PESQ values greater than the model distortion (for example,
410 2.16 for white noise at $\text{SNR}=12$ dB) are pointing out that small amount of
411 noise are beneficial for the quality of the signal obtained. This effect might
412 be due to the *stochastic resonance*, which concern to non-linear systems (like
413 our proposal) [54].

414 In order to demonstrate the benefits of using the auditory representation
415 of the signal, an experiment replacing this model with the short-time Fourier
416 transform was carried out. Here, two dictionaries trained with clean chirp
417 signals and white noise were obtained. Then, the NNCD method was ap-
418 plied in the same conditions as in Table 1 for noisy signals at $\text{SNR}=0$ dB. The
419 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.

Noise	SNR (dB)	Signal		
		Noisy	Wavelet	NNCD
White	12	1.93	1.79	2.16
	6	1.40	1.43	2.11
	0	0.69	0.87	1.99
Voice babble	12	1.82	1.72	2.05
	6	1.23	1.14	2.01
	0	0.56	0.53	1.91
Model distortion: 2.11				

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

423 5.3. Speech denoising

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

430 Fig. 7 shows an example of the denoising of real data signals correspond-
431 ing to speech data. The clean signal corresponds to the sentence /She had

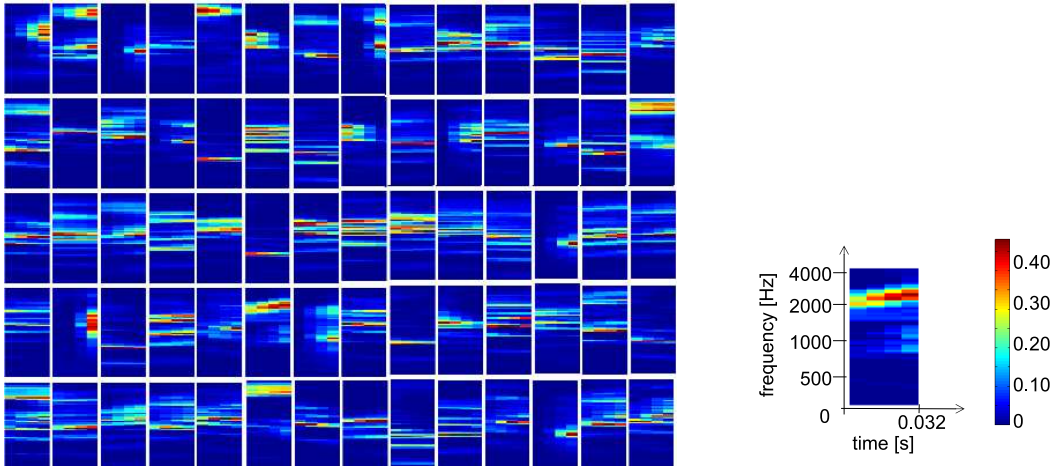


Figure 6: Examples of spectro-temporal receptive fields (STRF) calculated from the early auditory representation of speech signals (left). A single atom with axis labels and colorbar is also showed (right). The dimensions of each atom follow the setup outlined in Section 4.2.

432 your dark suit in greasy wash water all year/ (shown in the top spectrogram).
 433 The signal is then contaminated with white noise at SNR=0 dB. The effects
 434 of the noise can be seen in the middle spectrogram, where almost every im-
 435 portant speech feature has been masked by the noise. The denoising scheme,
 436 however, is able to recover the most prominent formants and to reduce the
 437 energy noise as shown in the bottom spectrogram.

438 For the measures of PESQ and SNRseg, a 10-fold cross validation proce-
 439 dure was applied by training a dictionary with 9 signals and testing with the
 440 remaining one. In each case, white and street noise were added with SNR of
 441 12, 6 and 0 dB. The results are summarized in Table 2 and 3. They show
 442 the mean and standard deviation of PESQ and SNRseg scores obtained for
 443 the cross validation scheme, being tested on the three different scenarios in

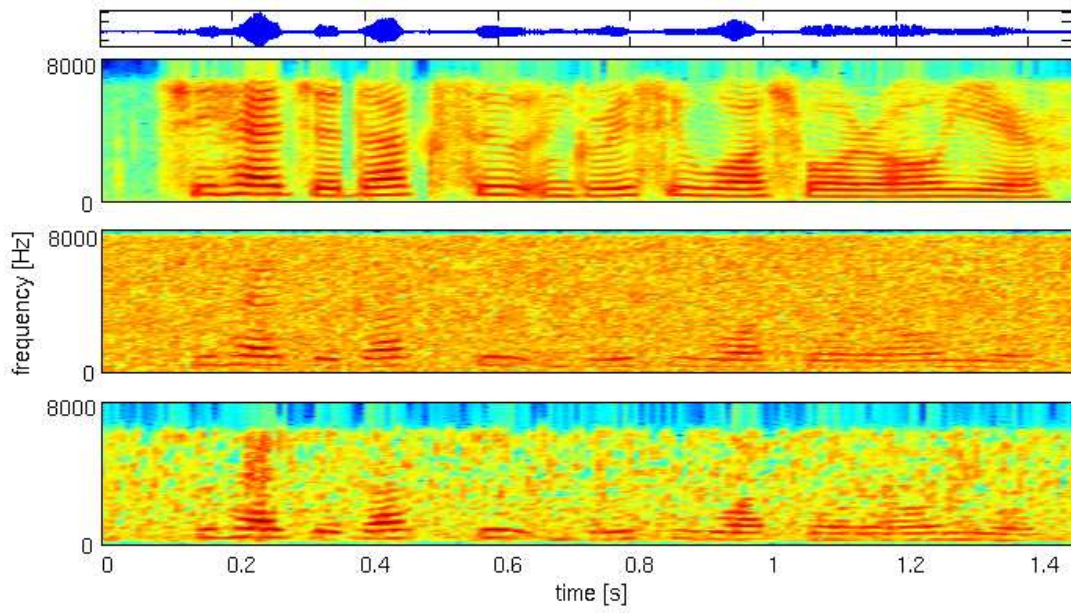


Figure 7: Example of the auditory cortical denoising result of a speech signal contaminated with white noise at SNR=0 dB. The spectrograms (STFT) of the clean signal (top), the noisy signal (middle) and the denoised reconstructed signal (bottom) are shown. The acoustic signal at the top of the figure is given as reference.

444 the application of NNCD and compared with different baseline methods (see
445 Section 3.2). For each experimental condition, the method that obtained the
446 best denoising quality is emphasized in boldface.

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

460 6. Conclusions

461 A new denoising method of audio signals was presented, inspired by the
462 biological processing carried out at the primary auditory cortical level. The
463 method obtains a sparse coding of the spectrogram at cochlea level using
464 a non-negative approach. The atoms of the dictionary are calculated from
465 clean signals and noise. Then, the denoising signal is obtained by inverting
466 the model using only the atoms corresponding to the signal, discarding the
467 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.

	SNR (dB)	Signal					NNCD		
		Noisy	iWiener	apWiener	mBand	BNMF	(a)	(b)	(c)
W	12	2.25 (0.14)	2.59 (0.15)	2.53 (0.15)	2.66 (0.21)	2.41 (0.10)	2.46 (0.08)	2.31 (0.14)	2.52 (0.08)
	6	1.92 (0.13)	2.19 (0.08)	2.17 (0.09)	2.18 (0.12)	2.18 (0.10)	2.26 (0.08)	1.97 (0.12)	2.36 (0.05)
	0	1.63 (0.18)	1.86 (0.15)	1.84 (0.16)	1.84 (0.18)	1.80 (0.09)	1.99 (0.13)	1.67 (0.17)	2.12 (0.10)
S	12	2.57 (0.13)	2.61 (0.13)	2.73 (0.13)	2.86 (0.11)	2.30 (0.14)	2.67 (0.11)	2.65 (0.12)	2.71 (0.11)
	6	2.21 (0.10)	2.18 (0.12)	2.39 (0.09)	2.49 (0.11)	2.06 (0.16)	2.45 (0.07)	2.30 (0.09)	2.51 (0.05)
	0	1.79 (0.13)	1.76 (0.15)	2.00 (0.10)	2.11 (0.09)	1.82 (0.13)	2.14 (0.08)	1.89 (0.11)	2.24 (0.06)

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.

	SNR (dB)	Signal					NNCD		
		Noisy	iWiener	apWiener	mBand	BNMF	(a)	(b)	(c)
W	12	6.98 (3.42)	8.43 (1.82)	10.04 (2.95)	6.91 (1.99)	1.59 (0.30)	5.60 (1.14)	7.63 (3.47)	5.79 (0.90)
	6	1.84 (2.54)	4.50 (1.54)	5.14 (2.12)	5.14 (2.56)	1.62 (0.31)	5.21 (0.62)	2.68 (2.52)	5.24 (0.70)
	0	-2.77 (2.00)	2.10 (0.85)	0.04 (1.92)	2.25 (0.23)	1.57 (0.16)	3.84 (0.84)	-2.01 (2.04)	4.56 (0.79)
S	12	7.10 (2.31)	6.33 (1.33)	8.67 (2.23)	7.09 (1.31)	1.54 (0.22)	5.75 (0.79)	8.24 (2.40)	5.68 (0.48)
	6	1.93 (2.24)	3.79 (1.05)	4.13 (2.40)	4.52 (1.59)	1.69 (0.36)	5.26 (0.50)	3.51 (2.15)	4.95 (0.36)
	0	-3.54 (2.27)	1.71 (0.61)	-1.19 (2.55)	2.37 (1.07)	1.57 (0.30)	3.94 (0.54)	-1.94 (2.23)	3.89 (0.33)

468 The performance of the method using synthetic and real signals with
469 additive noise was obtained through two objective quality measures. Results
470 showed that our proposed method and its variants can improve the quality
471 of sound signals, specially under severe conditions.

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

475 **Acknowledgements**

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

483 **Appendix**

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

485 **References**

- 486 [1] Y. Hu and P.C. Loizou. Subjective comparison and evaluation of
487 speech enhancement algorithms. *Speech Communication*, 49(7-8):588–
488 601, 2007.

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 $i = 1, \dots, M$.

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

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

- Define the samples that use $\vec{\phi}_k : \omega_k = \{i | 1 \leq i \leq M, \mathbf{a}_i(k) \neq 0\}$.
- Compute $\mathbf{E}_k = \mathbf{x} - (\Phi \mathbf{a} - \vec{\phi}_k \mathbf{a}(k))$.
- Choose only the columns corresponding to ω_k , and obtain $\mathbf{E}_k^{\omega_k}$.
- Set $A = \mathbf{E}_k^{\omega_k}$,

$$\vec{\phi}_k = \begin{cases} 0, & \mathbf{u}_1(i) < 0 \\ \mathbf{u}_1(i), & \text{otherwise} \end{cases}$$

$$\mathbf{a}(k) = \begin{cases} 0, & \mathbf{v}_1(i) < 0 \\ \mathbf{v}_1(i), & \text{otherwise} \end{cases}$$

where \mathbf{u}_1 and \mathbf{v}_1 are the first singular vector of A .

Repeat J times:

$$\vec{\phi} = \frac{A \mathbf{a}}{\mathbf{a}' \mathbf{a}}. \text{ Project: } \vec{\phi}(i) = \begin{cases} 0, & \vec{\phi}(i) < 0 \\ \vec{\phi}(i), & \text{otherwise} \end{cases}$$

$$\mathbf{a} = \frac{\vec{\phi}' A}{\vec{\phi}' \vec{\phi}}. \text{ Project: } \mathbf{a}(i) = \begin{cases} 0, & \mathbf{a}(i) < 0 \\ \mathbf{a}(i), & \text{otherwise} \end{cases}$$

Normalize $\vec{\phi}_k$.

Set $J = J + 1$.

Figure 8: The NN-K-SVD algorithm.

- 489 [2] J-C Junqua and J-P Haton. Robustness in automatic speech recognition:
490 Fundamentals and applications. Kluwer Academic Publishers, 1995.
- 491 [3] M. Lewicki and T. Sejnowski. Learning overcomplete representations.
492 *Neural Computation*, 12(2):337–365, 2000.
- 493 [4] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information*
494 *Theory*, 52(4):1289–1306, 2006.
- 495 [5] J. Deller, J. Proakis, J. Hansen. *Discrete Time Processing of Speech*
496 *Signals*. Macmillan Publishing, New York, 1993.
- 497 [6] B. Delgutte. Physiological models for basic auditory percepts. In H.H.
498 Hawkins, T.A. McMullen, A.N Popper, R.R. Fay, editor, *Auditory Com-*
499 *putation*. Springer, New York, 1996.
- 500 [7] H. Rufiner, L. Rocha, and J. Goddard. Sparse and independent repre-
501 sentations of speech signals based on parametric models. In *Proceedings*
502 *of the ICSLP'02*, pages 989–992, 2002.
- 503 [8] S. Greenberg. The ears have it: The auditory basis of speech percep-
504 tion. In *Proceedings of the International Congress of Phonetic Sciences*,
505 volume 3, pages 34–41, 1995.
- 506 [9] T. Chiu, P. Ru and S. Shamma. Multiresolution spectrotemporal anal-
507 ysis of complex sounds. *Journal of the Acoustical Society of America*,
508 118(2):897–906, 2005.
- 509 [10] N. Mesgarani and S. Shamma. Denoising in the domain of spectrotem-

- 510 poral modulations. *EURASIP Journal on Audio, Speech and Music*
511 *Processing*, 2007:8 pages, 2007.
- 512 [11] D. Klein, P. König and K. Kording. Sparse spectrotemporal coding of
513 sounds. *EURASIP Journal on Applied Signal Processing*, 2003(7):659–
514 667, 2003.
- 515 [12] B. Olshausen and D. Field. Sparse Coding with an Overcomplete Basis
516 Set: A Strategy Employed by V1? *Vision Research*, 37(23):3311–3325,
517 1997.
- 518 [13] A. Hyvärinen and E. Oja. Independent component analysis: algorithms
519 and applications. *Neural Networks*, 13(4-5):411–430, 2000.
- 520 [14] F. Theunissen, K. Sen and A. Doupe. Spectro-temporal receptive fields
521 of nonlinear auditory neurons obtained using natural sounds. *Journal*
522 *of Neuroscience*, 20:2315–2331, 2000.
- 523 [15] D-S Kim, S-Y Lee and R. Kil. Auditory processing of speech signals
524 for robust speech recognition in real-world noisy environments. *IEEE*
525 *Transactions on Speech and Audio Processing*, 7(1):55–69, 1999.
- 526 [16] R. Stern and N. Morgan. Hearing is believing: Biologically-inspired
527 feature extraction for robust automatic speech recognition. *IEEE Signal*
528 *Processing Magazine*, 29(6):34–43, 2012.
- 529 [17] M. Kleinschmidt, J. Tchorz and B. Kollmeier. Combining speech en-
530 hancement and auditory feature extraction for robust speech recogni-
531 tion. *Speech Communication*, 34(1):75–91, 2001.

- 532 [18] T. Dau, D. Püschel and A. Kohlrausch. A quantitative model of the “ef-
533 fective” signal processing in the auditory system. I. Model structure. *The*
534 *Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996.
- 535 [19] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean
536 square error short-time spectral amplitude estimator. *IEEE Transac-*
537 *tions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, 1984.
- 538 [20] R. Flynn and E. Jones. Combined speech enhancement and auditory
539 modelling for robust distributed speech recognition. *Speech Communi-*
540 *cation*, 50(10):797–809, 2008.
- 541 [21] Q. Li, F. Soong and O. Siohan. A high-performance auditory feature
542 for robust speech recognition. *Interspeech*, 51–54, 2000.
- 543 [22] S. Rangachari and P. Loizou. A noise-estimation algorithm for highly
544 non-stationary environments. *Speech communication*, 48(2):220–231,
545 2006.
- 546 [23] C. Kim and R. Stern. Power-normalized cepstral coefficients (PNCC)
547 for robust speech recognition. In *Proc. of Acoustics, Speech and Signal*
548 *Processing, ICASSP*, 4101–4104, 2012.
- 549 [24] C. Martínez, J. Goddard, D. Milone and H. Rufiner. Bioinspired sparse
550 spectro-temporal representation of speech for robust classification. *Com-*
551 *puter Speech and Language*, 26:336–348, 2012.
- 552 [25] O-W Kwon and T-W Lee. Phoneme recognition using ICA-based fea-
553 ture extraction and transformation. *Signal Processing*, 84(6):1005–1019,
554 2004.

- 555 [26] P.O. Hoyer. Non-negative matrix factorization with sparseness con-
556 straints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.
- 557 [27] T. Virtanen. Monaural sound source separation by nonnegative matrix
558 factorization with temporal continuity and sparseness criteria. *IEEE*
559 *Transactions on Audio, Speech, and Language Processing*, 15(3):1066–
560 1074, 2007.
- 561 [28] F. Weninger, J. Feliu and B. Schuller. Supervised and semi-supervised
562 suppression of background music in monaural speech recordings. In
563 *Proc. of Acoustics, Speech and Signal Processing, ICASSP*, 61–64, 2012.
- 564 [29] F. Weninger, J. Feliu and B. Schuller. Source separation using regular-
565 ized NMF with MMSE estimates under GMM priors with online learning
566 for the uncertainties. *Digital Signal Processing*, 29(0):20–34, 2014.
- 567 [30] P. Smaragdis. Convolutional speech bases and their application to su-
568 pervised speech separation. *IEEE Transactions on Audio, Speech, and*
569 *Language Processing*, 15(1):1–12, 2007.
- 570 [31] K. Wilson, B. Raj and P. Smaragdis. Regularized non-negative matrix
571 factorization with temporal dependencies for speech denoising. *INTER-*
572 *SPEECH*, 411–414, 2008.
- 573 [32] R. Vipperla, S. Bozonnet, D. Wang and N. Evans. Robust speech recog-
574 nition in multi-source noise environments using convolutional non-negative
575 matrix factorization. *CHiME: Workshop on Machine Listening in Mul-*
576 *tisource Environments*, 74–79, 2011.

- 577 [33] N. Mohammadiha, P. Smaragdis and A. Leijon. Supervised and unsuper-
578 vised speech enhancement using nonnegative matrix factorization. *IEEE*
579 *Transactions on Audio, Speech, and Language Processing*, 21(40):2140–
580 2141, 2013.
- 581 [34] J. Gemmeke, T. Virtanen and A. Hurmalainen. Exemplar-based sparse
582 representations for noise robust automatic speech recognition. *IEEE*
583 *Transactions on Audio, Speech, and Language Processing*, 19(7):2067–
584 2080, 2011.
- 585 [35] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM*
586 *Journal on Computing*, 24(2):227–234, 1995.
- 587 [36] S. Chen, D. Donoho and M. Saunders. Atomic decomposition by basis
588 pursuit. *SIAM Review*, 43(1):129–159, 2001.
- 589 [37] S.G. Mallat and Z. Zhang. Matching pursuit with time-frequency dic-
590 tionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.
- 591 [38] M. Aharon, M. Elad and A.M. Bruckstein. K-SVD and its non-negative
592 variant for dictionary design. In *Proceedings of the SPIE conference*
593 *wavelets*, volume 5914, 2005.
- 594 [39] R. Zhang, C. Wang and B. Xiao. A strategy of classification via sparse
595 dictionary learned by non-negative K-SVD. In *12th IEEE International*
596 *Conference on Computer Vision Workshops (ICCV Workshops)*, 117–
597 122, 2009.

- 598 [40] Y. Huang and J. Benesty (editors). *Audio Signal Processing for next-*
599 *generation multimedia communication systems*. Kluwer Academic Press,
600 2004.
- 601 [41] D. Milone, L. Di Persia and M.E. Torres. Denoising and recognition
602 using hidden Markov models with observation distributions modeled by
603 hidden Markov trees. *Pattern Recognition*, 43(4):1577–1589, 2009.
- 604 [42] J. Lim and A. V. Oppenheim. All-pole modeling of degraded
605 speech. *IEEE Transactions on Acoustics, Speech and Signal Process-*
606 *ing*, 26(3):197–210, 1978.
- 607 [43] P. Scalart and J. Vieira Filho. Speech enhancement based on a priori
608 signal to noise estimation. In *Proc. of Acoustics, Speech and Signal*
609 *Processing, ICASSP*, volume 2, pages 629–632, 1996.
- 610 [44] D. Donoho. De-noising by soft-thresholding. *IEEE Transactions on*
611 *Information Theory*, 41(3):613–627, 1995.
- 612 [45] S. Kamath and P. Loizou. A multi-band spectral subtraction method
613 for enhancing speech corrupted by colored noise. In *Proc. of Acous-*
614 *tics, Speech and Signal Processing, ICASSP*, volume 4, pages 4164–4164,
615 2002.
- 616 [46] Perceptual evaluation of speech quality (PESQ): An objective method
617 for end-to-end speech quality assessment of narrow-band telephone net-
618 works and speech codecs. *ITU-T Recommendation P.862*, 2001.
- 619 [47] J. Hansen and B. Pellom. An effective quality evaluation protocol for

- 620 speech enhancement algorithms. In *Proc. Int. Conf. Spoken Lang. Pro-*
621 *cess.*, volume 7, pages 2819–2822, 1998.
- 622 [48] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett and N. Dahlgren.
623 DARPA TIMIT Acoustic-phonetic continuous speech corpus documen-
624 tation. Technical report, National Institute of Standards and Technol-
625 ogy, 1993.
- 626 [49] A. Varga and H. Steeneken. Assessment for automatic speech recognition
627 II: NOISEX-92: a database and an experiment to study the effect of
628 additive noise on speech recognition systems. *Speech Communication*,
629 12(3):247–251, 1993.
- 630 [50] H. Hirsch and D. Pearce. The AURORA experimental framework for
631 the performance evaluation of speech recognition systems under noisy
632 conditions. In *Proceedings of the ISCA ITRW ASR2000*, 2000.
- 633 [51] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual
634 evaluation of speech quality (PESQ)-a new method for speech quality
635 assessment of telephone networks and codecs. In *Proc. of Acoustics,*
636 *Speech and Signal Processing, ICASSP*, volume 2, pages 749–752, 2001.
- 637 [52] L. Di Persia, D. Milone, H. Rufiner and M. Yanagida. Perceptual eval-
638 uation of blind source separation for robust speech recognition. *Signal*
639 *Processing*, 88(10):2578–2583, 2008.
- 640 [53] P. Loizou. *Speech enhancement: Theory and Practice*. CRC press, 2013.

- 641 [54] M. McDonnell and D. Abbott. What is stochastic resonance? Defi-
642 nitions, misconceptions, debates, and its relevance to biology. PLoS
643 Computational Biology, 5(5), e1000348, 2009.