A SOLUTION TO RESIDUAL NOISE IN SPEECH DENOISING WITH SPARSE REPRESENTATION

Yongjun He[1,2]*, Jiqing Han*[1]*, Shiwen Deng*[1]*, Tieran Zheng*[1]*, Guibin Zheng*[1].

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
{heyongjun,jqhan}@hit.edu.cn

ABSTRACT

As a promising technique, sparse representation has been extensively investigated in signal processing community. Recently, sparse representation is widely used for speech processing in noisy environments; however, many problems need to be solved because of the particularity of speech. One assumption for speech denoising with sparse representation is that the representation of speech over the dictionary is sparse, while that of the noise is dense. Unfortunately, this assumption is not sustained in speech denoising scenario. We find that many noises, e.g., the babble and white noises, are also sparse over the dictionary trained with clean speech, resulting in severe residual noise in sparse enhancement. To solve this problem, we propose a novel residual noise reduction (RNR) method which first finds out the atoms which represents the noise sparely, and then ignores them in the reconstruction of speech. Experimental results show that the proposed method can reduce residual noise substantially.

Index Terms— Sparse representation, speech denoising, residual noise, basis pursuit denoising.

1. INTRODUCTION

Environmental noises not only reduce the intelligibility of speech, but also degrade the performances of speech processing systems. Although many methods are proposed, speech denoising remains to be a challenge. The difficulty arises from the nature of real-world noises that are often non-stationary and potentially speech-like, thereby inducing a significant and variable spectral overlap between speech and noises.

In the past few decades, sparse representation is extensively investigated and provides possible solutions for speech denoising. Current researches show that the neurosensory systems encode stimuli by activating only a small number of neurons out of a large population at the same time [1] [2]. In the light of these results, researchers bring forward the conception of "sparsity" and find that many signals, including speech, can be approximated to be sparse[1] [3]. Recently, speech processing algorithms which achieve improved performance by using sparse signal models [4]-[7] encourage researches to make more and further explorations.

In sparse representation, signals are represented with a set of atoms (elementary signals), and the collection of atoms is called a dictionary. By a sparse representation, we mean that the representation accounts for most or all information of a signal, with a linear combination of only a small number of atoms. In exploiting the sparsity of signal, a lot of methods are proposed; however, in this paper, we focus on the basis pursuit denoising (BPDN) [8] because of its advantage in speech denoising. With this method, the spare representation of a signal is obtained by solving the following optimization problem:

$$\min_{\mathbf{y}} \|\mathbf{y}\|_1 \quad \text{subject to } \|Y - \mathbf{\Phi}\mathbf{y}\|_2 \leq \varepsilon \tag{1}$$

where Y is the noisy observation, $\mathbf{\Phi}$ is the dictionary trained with clean speech, $\mathbf{y}$ is the sparse representation of Y over dictionary $\mathbf{\Phi}$, $\varepsilon > 0$ is the error tolerance, $\|.\|_1$ and $\|.\|_2$ denote the l_1-norm and l_2-norm respectively. For an appropriate Lagrange multiplier λ, the solution to (1) is precisely the solution to the unconstrained optimization problem:

$$\mathbf{y} = \arg\min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{1}{2}\|Y - \mathbf{\Phi}\mathbf{y}\|_2^2, \tag{2}$$

where λ is the Lagrange multiplier. Just like other sparseness optimization problems, BPDN also contains a sparseness term $\|\mathbf{y}\|_1$ and a reconstruction error term $\|\mathbf{Y} - \mathbf{\Phi}\mathbf{y}\|_2^2/2$. The regularization parameter $\lambda > 0$ trades off the costs of sparsity and the reconstruction error. Specifically, larger λ results in more sparse solutions.

With the sparse representation $\mathbf{y}$, the clean speech can be enhanced with

$$\hat{X} \approx \mathbf{\Phi}\mathbf{y}. \tag{3}$$

One assumption of speech denoising with sparse representation is that the representation of the clean speech is sparse while that of the noise is dense over the dictionary. If this assumption is true, the sparse decomposition in (2) cannot find atoms to represent the noise; as a result, the noise is discarded as residual error. Then the clean speech can be reconstructed with (3). Unfortunately, the above-mentioned assumption is far from the truth in speech processing. With a well-designed method, we can obtain a dictionary, over which the representation of speech is enough sparse. The problem is that the representation of the noise over this dictionary may be also sparse (see experiments). One typical example is the white noise, which is much like unvoiced speech. If a dictionary can represent unvoiced speech sparely, the same thing also happens to white noise. Another example is the babble noise, which is speech itself and can be represented sparsely over any dictionary trained with clean speech.

According to the analysis in section 2, more residual noise remains in the reconstructed speech if the noise is also spare over the speech dictionary. To solve this problem, we propose a novel

This research is partly supported by the National Natural Science Foundation of China under grant No. 91120303 and 61071181, the Scientific Research Fund of Heilongjiang Provincial Education Department under grant No. 12511096

[1]The "sparsity" in this paper means "approximate sparsity".

method RNR, which first finds out those atoms representing the noise sparsely and then ignores them with a mask in the reconstruction of speech. Further experiments show that the proposed method can reduce the residual noise substantially.

2. ANALYSIS OF SPEECH DENOISING WITH BPDN

In this section, we will show that residual noise remains in the enhanced speech by BPDN if the noise is also sparse. In the preprocessing, the speech signal is split into overlapping frames. A speech frame Y corrupted by additive noise can be modeled with

$$Y = X + V \tag{4}$$

where X and V are the clean speech and noise, respectively. Equation (4) can hold to be true both in the time- and frequency- domains. If the noise is absent, the sparse representation of X is obtained by

$$\mathbf{x}_0 = \arg\min_{\mathbf{x}} \lambda \|\mathbf{x}\|_1 + \frac{1}{2}\|X - \mathbf{\Phi}\mathbf{x}\|_2^2 . \tag{5}$$

When the speech X is corrupted by the noise, the task in speech denoising is to recover X from the noisy speech Y. In the speech denoising with sparse representation, it is assumed that the noise is dense over the dictionary $\mathbf{\Phi}$, and therefore cannot be represented sparsely. In other words, $\mathbf{x}_0$ is also the expected sparse representation of Y, i.e.,

$$\mathbf{x}_0 = \arg\min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{1}{2}\|Y - \mathbf{\Phi}\mathbf{y}\|_2^2 , \tag{6}$$

then the clean speech can be reconstructed with

$$X \approx \mathbf{\Phi}\mathbf{x}_0 \tag{7}$$

where the noise is discarded as residual error and the noise reduction is achieved.

In fact, the above assumption is far from the truth, since the noise is often sparse over the dictionary $\mathbf{\Phi}$. As a result, it is impossible to obtain $\mathbf{x}_0$ in decomposing Y. Suppose the result of the decomposition is $\mathbf{y}_0$, i.e.,

$$\mathbf{y}_0 = \arg\min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{1}{2}\|Y - \mathbf{\Phi}\mathbf{y}\|_2^2 \tag{8}$$

where $\mathbf{y}_0$ not only represents the speech, but also represents part of the noise. In the reconstruction, the enhanced speech is obtained by

$$\hat{X} \approx \mathbf{\Phi}\mathbf{y}_0 \tag{9}$$

If $\mathbf{x}_0$ is viewed as the expected representation, $\mathbf{y}_0$ always can be written as

$$\mathbf{y}_0 = \mathbf{x}_0 + \mathbf{e} \tag{10}$$

where $\mathbf{e} = \mathbf{y}_0 - \mathbf{x}_0$, then (9) can be rewritten as

$$\hat{X} \approx \mathbf{\Phi}\mathbf{x}_0 + \mathbf{\Phi}\mathbf{e} \approx X + \mathbf{\Phi}\mathbf{e} \tag{11}$$

Compared with the reconstructed result in (7), the residual noise $\mathbf{\Phi}\mathbf{e}$ remains in (11). If $\mathbf{\Phi}\mathbf{e}$ contains high energy, the enhanced speech is still noisy. That is why speech denoising with sparse representation often does not remove the noise completely, especially when the noise can be represented sparsely by the clean dictionary.

We can further write $\mathbf{e}$ as

$$\mathbf{e} = \mathbf{e}_+ + \mathbf{e}_- \tag{12}$$

where $\mathbf{e}_+$ contains the coefficients which increase the energy of the reconstructed speech and $\mathbf{e}_-$ for the opposite situation, then (11) is written as

$$\hat{X} \approx X + \mathbf{\Phi}\mathbf{e}_+ + \mathbf{\Phi}\mathbf{e}_- . \tag{13}$$

There are two types of distortions in (13). The first one is $\mathbf{\Phi}\mathbf{e}_+$, which increases the energy of the reconstructed speech; the second one is $\mathbf{\Phi}\mathbf{e}_-$ resulting in a energy decreasing of the reconstructed speech.

Although the above analysis is for BPDN, the result also applies to other sparse optimization problems. For example, the sparseness term in (2) is replaced by the l_p-norm ($0 \le p < 1$), or the reconstruction error is replaced by other distance measures, such as the Kullback–Leibler divergence.

3. THE PROPOSED METHOD

As analyzed in the previous section, the reason for the residual noise in sparse reconstruction is that the noise is also sparse over the speech dictionary. Properly speaking, on the premise of sparsity, some atoms in the speech dictionary can represent the noise observations with a low residual error. If we can find out such atoms and ignore them in the reconstruction, the residual noise can be suppressed.

Given an over-complete dictionary $\mathbf{\Phi} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N]$ with $\mathbf{a}_k$ ($k = 1, ..., N$) as its atoms, equation (11) can be rewritten as

$$\begin{aligned}\hat{X} &\approx \mathbf{\Phi}\mathbf{x}_0 + \mathbf{\Phi}\mathbf{e} \\ &= \sum_{i=1}^{I} x_i \mathbf{a}_i^s + \sum_{j=1}^{J} e_j \mathbf{a}_j^e\end{aligned} \tag{14}$$

where $\mathbf{\Psi}^s = \{\mathbf{a}_1^s, \mathbf{a}_2^s, ..., \mathbf{a}_I^s\} \subset \mathbf{\Phi}$ and $\mathbf{\Psi}^e = \{\mathbf{a}_1^e, \mathbf{a}_2^e, ..., \mathbf{a}_J^e\} \subset \mathbf{\Phi}$ are used to represent the clean speech and residual noise, respectively, $x_i \in \mathbf{x}_0$, $v_j \in \mathbf{e}$, I and J are the numbers of non-zero coefficients in $\mathbf{x}_0$ and $\mathbf{e}$, respectively. The ideal denoising result is

$$X_{ideal} = \sum_{i=1}^{I} x_i \mathbf{a}_i^s . \tag{15}$$

In this case, the noise is removed completely; however, under the influence of the noise, we now only obtain $\hat{X}$, which contains $\sum_{j=1}^{J} e_j \mathbf{a}_j^e$. Therefore, we have the reason to believe that the atoms in $\mathbf{\Psi}^e$ can represent the noise sparely. If $\mathbf{\Psi}^s \cap \mathbf{\Psi}^e = \phi$ and we can find out all atoms in $\mathbf{\Psi}^e$, an idea reconstruction in (15) can be obtained. In our method, we attempt to find out the atoms in $\mathbf{\Psi}^e$ but not in $\mathbf{\Psi}^s$, and then set their coefficients as zeros. First of all, we need to make clear what will happen if the coefficient of an atom is set as zero.

For an atom $\mathbf{a}$ with its coefficient set as zero, there are three cases:

(a) If $\mathbf{a} \in \mathbf{\Psi}^e - \mathbf{\Psi}^s$, the residual noise can be reduced;

(b) If $\mathbf{a} \in \mathbf{\Psi}^s - \mathbf{\Psi}^e$, the speech energy should be missing, resulting in a distortion;

(c) If $\mathbf{a} \in \mathbf{\Psi}^s \cap \mathbf{\Psi}^e$, the residual noise can be reduced at the expense of speech energy missing.

We expect to find out all atoms in case (a) to ignore their contributions, and at the same time avoid ignoring the atoms in case (b). Considering the above factors, we decompose a set of noise samples over the clean dictionary to find out the atoms which are used most frequently in the representation of the noise and used infrequently in the representation of clean speech, and then get rid of their contribution in the reconstruction. Note that if the coefficient of an atom in a

sparse representation is non-zero, we say this atom is used once. The implementation of the proposed method is summarized in Algorithm I.

In this Algorithm, $\mathbf{\Phi}$ is an over-complete dictionary with N atoms, $\mathbf{p}_d$ is a vector with $\mathbf{p}_d[n]$ be the prior probability of the nth atom, h is a threshold for $\mathbf{p}_d$, k is the number of the atoms that represent noise sparsely, $\mathbf{m}$ is a mask with elements 0 indicating the corresponding atoms are ignored and 1 for the opposite situation, P is the number of the noise frames used for calculating the mask, $\mathbf{c}$ stores the used time of each atom in the representation of noise frames, and $diag(\mathbf{m})$ stands for the diagonal matrix with its diagonal component value equal to the value of $\mathbf{m}$. The dictionary $\mathbf{\Phi}$ is obtained by training on the clean speech with the method proposed in [9], and $\mathbf{p}_d$ is obtained by reconstructing the training speech with their sparse representations and counting the used time of each atom.

The speech is processed utterance by utterance and the first P frames are supposed to be noise only. First, the noise frames are decomposed over the speech dictionary. Next, the most frequently used k atoms are found out, and then the atoms, of which prior probabilities are less than h, are masked by setting the corresponding elements in $\mathbf{m}$ as zeros. Finally, all frames of this utterance are decomposed and reconstructed, with the masked atoms ignored in the reconstruction.

Algorithm I
Input: $\mathbf{\Phi}, \mathbf{p}_d, h, k, U, P$
output: $\hat{U}$
For each utterance U
step 1: split U into overlapping frames $U_1, ..., U_M$
step 2: calculate their magnitude spectrums $Y_1, ..., Y_M$ and phases $\zeta_1, ..., \zeta_M$
step 3: decompose the first P noise frames with (2)
step 4: count the used time of each atom
  step 4.1: initialize $\mathbf{c} = [0, ..., 0]$;
  step 4.2: if the nth ($n = 1{:}N$) atom is used d times, set $\mathbf{c}[n] = d$;
step 5: calculate the mask
  step 5.1: initialize $\mathbf{m} = [1, ..., 1]$;
  step 5.2: get g =the kth largest value of $\mathbf{c}[n], n = 1{:}N$;
  step 5.3: if $\mathbf{c}[n] \geq g$ and $\mathbf{p}_d[n] < h$ $\mathbf{m}[n] = 0$;
step 6: decompose frames:
  $\mathbf{y}_i \leftarrow \min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{1}{2} \|Y_i - \mathbf{\Phi}\mathbf{y}\|_2^2, i = 1 : M$
step 7: reconstruct the enhanced speech frames:
  $\hat{X}_i = \mathbf{\Phi} diag(\mathbf{m})\mathbf{y}_i, i = 1 : M$
step 8: obtain time-domain signal $\hat{U}$ with $\hat{X}_i (i = 1 : M)$ and the noisy phases.

4. EXPERIMENTS AND RESULTS

To evaluate the propose method, experiments were conducted on the TIMIT database with all utterances down-sampled at 8 kHz. To obtain noisy speech, four noises taken from the noise-92 database, namely white, f16, babble and pink noises, were added artificially on the utterances of the TIMIT testset at -5 dB, 0 dB, 5 dB, 10 dB and 20 dB. The used dictionary which contained 1024 atoms was trained with all utterances in the TIMIT trainset. In the preprocessing, speech was Hamming windowed every 10 ms with a window width of 20 ms, and then each frame was passed through a discrete Fourier transform (DFT). Next, the magnitude spectrum

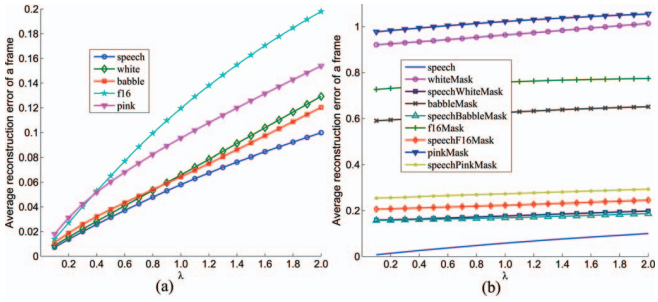

Fig. 1. The comparison of the average reconstruction errors (a) reconstruction errors for clean speech and four noises (b) reconstruction error for clean speech and four noises with RNR

of each frame was used for sparse decomposition. The SPASM tools [10] were used for dictionary training and sparse decomposition (Lasso algorithm). In the experiments, $k = 50$ and h is set as the value of the 200th smallest element in $\mathbf{p}_d$.

First, the sparsity of the four noises is tested by reconstructing noise frames with their sparse representation over the speech dictionary and comparing their average reconstruction errors with that of the clean speech. The duration of each noise used for reconstruction is three minutes, and the speech (from the TIMIT database) with the same length is used for comparison. The result is shown in Fig. 1 (a), where λ varies from 0.1 to 2.0. It is seen that the reconstruction errors of white and babble noises are very close to that of the clean speech, especially when $\lambda < 1$. This result indicates that the two noises can be sparsely represented in a residual error as low as that of the clean speech. Next, one second data of each noise is used to compute a mask (with Algorithm I, $\lambda = 1$), with which the speech and the corresponding noise are reconstructed again (Fig. 1 (b)). Taking white noise as an example, the reconstruction error of the clean speech (denoted by "speechWhiteMask") is increased in compared with that of the reconstruction without using the mask. However, a much larger error is observed in the reconstruction of white noise (denoted by "whiteMask"). Similar results can be observed in the reconstruction of other noises. These results show that by ignoring the atoms representing noises sparsely, the sparsity of noises over the speech dictionary can be reduced substantially in the expense of small speech energy missing.

One example is shown in Fig. 2, where an utterance from the TIMIT database (Fig. 2 (a)) is first distorted by the white noise at 10 dB (Fig. 2 (b)) and then used for enhancement. Fig. 2 (c) and Fig. 2 (e) are the speech enhanced with BPDN directly, while Fig. 2 (d) and Fig. 2 (f) with RNR. In the decomposition, $\lambda = 1$. It is seen that although BPDN can reduce the noise, there is still residual noise in its enhanced result. On the contrary, the result by RNR almost has no residual noise. Compared with Fig. 2 (a), the last unvoiced phoneme "s" in Fig. 2 (f) is missing, since this phoneme is much like white noise. With RNR, the atoms used for representing phoneme "s" is ignored, resulting in an energy missing.

Finally, further experiments are conducted to enhance the noisy testset on the TIMIT database. The spectral subtraction (SS) [12] is chosen for comparison. The results are summarized in Fig. 3, where the measure is the average magnitude distance of a frame (the enhanced to the clean). The parameter are set as follows: $\lambda = 8, 6, 3, 1, 0.5$ corresponding SNR= -5 dB, 0 dB, 5 dB, 10 dB, 20 dB (lower SNR needs larger λ in BPDN). The chosen λ is the best one

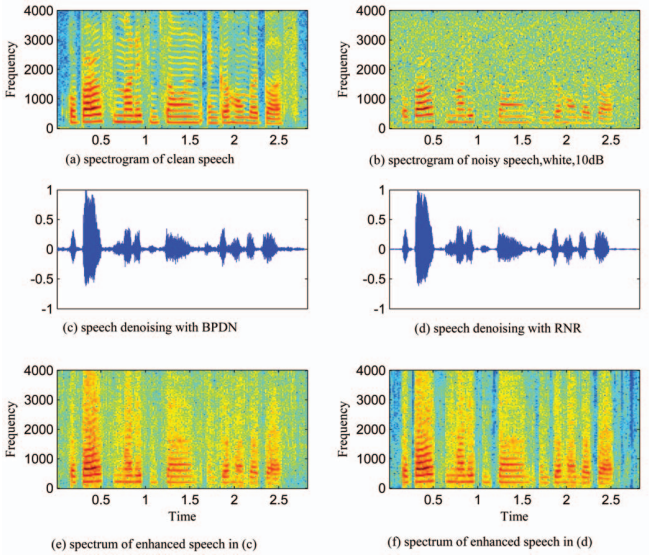

Fig. 2. A speech enhancement comparison. The sentence is from the TIMIT database and its content is "or borrow some money from someone and go home by bus".

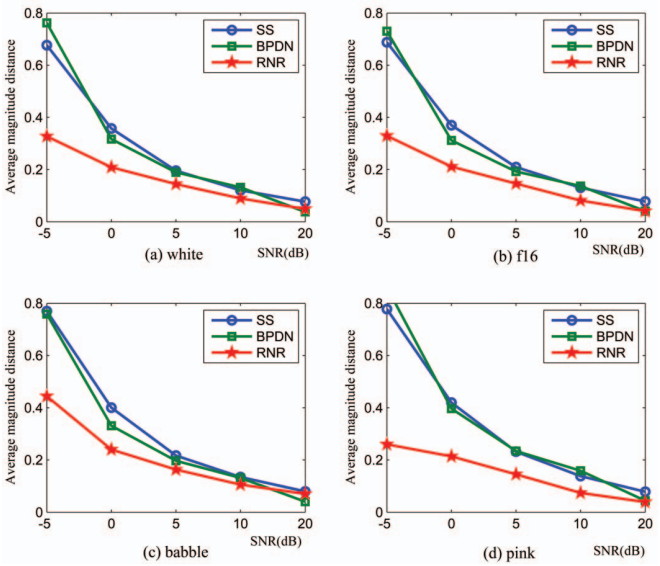

Fig. 3. The comparison of speech enhancement results. Four noises, namely white, f16, babble and pink are used to distort the clean speech at the SNRs of -5 dB, 0 dB, 5 dB, 10 dB and 20 dB.

for BPDN denoising obtained by experiment test. We can see that the performances of SS and BPDN are close to each other. BPDN outperforms SS at all SNRs in babble noise, since babble is a time-varying noise, in which the noise spectra estimation is more difficulty for SS. RNR performs better than the other two methods almost in all noise conditions, especially when the SNR is low. In white and babble noises, RNR is slightly outperformed by BPDN when the SNR is 20 dB. This may be because the energy missing under high SNR is more obvious. In summary, the advantage of RNR over the other two methods becomes larger with the decrease in SNR.

5. CONCLUSION

As shown in the experiments, the speech and the noise may be both sparse over a dictionary learning on clean speech. This problem is more severe when the noise is speech-like, such as the white (like unvoiced speech) and babble noises. Under this condition, more residual noise may remains in the enhanced speech, resulting a performance degradation or even a failure in noise reduction. To solve this problem, we proposed RNR to reduce the residual noise by ignoring the atoms which represented noise sparsely. Experiments showed that RNR performed better than BPDN and spectral subtraction.

6. REFERENCES

[1] B. Olshausen, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[2] D. Attwell and S. Laughlin, "An energy budget for signaling in the grey matter of the brain," *J. Cereb. Blood Flow Metab.*, vol. 21, no. 10, pp. 1133–1145, 2001.

[3] M. A. Davenport, M. F. Duarte, Y. C. Eldar and G. Kutyniok, "Introduction to compressed sensing," in *Compressed sensing: theory and applications*, Cambridge University Press, 2011.

[4] T. Virtane, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[5] C. D. Sigg, T. Dikk and J. M. Buhmann, "Speech enhancement with sparse coding in learned dictionaries," In *proc. ICASSP*, 2010, pp. 4758–4761.

[6] L. -L. Durrieu and L. -P. Thiran, "Sparse non-negative decomposition of speech power spectra for formant tracking," In *proc. ICASSP*, 2011, pp. 5260–5263.

[7] J. F. Gemmeke, T. Virtanen and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, Sep. 2011.

[8] S. Chen, D. Donoho and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[9] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online dictionary learning for sparse coding," In *Proc. ICML*, 2009.

[10] SPAMS tools: Available on: http://www.di.ens.fr/willow/SPAMS/downloads.html.

[11] C. L. Philipos, Speech enhancement: theory and practice, Taylor and Francis, 2007.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.