Speech denoising in the presence of Impulsive Noise

Ruibo Zhang
Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
zhangruibox@outlook.com

Hongqing Liu
Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
hongqingliu@outlook.com

Zhen Luo
Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

Yi Zhou
School of Communications and Information Engineering
Chongqing University of Posts and Telecommunications
zhouy@cqupt.edu.cn

ABSTRACT

This work addresses speech denoising problem in the presence of impulsive noise in transform domains. The impulsive noise, in this work, is modeled by an unknown sparse vector so that it can be actively suppressed. The speech signal is sparsely represented by the wavelet domain. To achieve the simultaneous speech recovery and the noise suppression, a joint estimation is devised based on the fact they have sparse representations in different domains. To efficiently solve the problem, the alternating direction method of multipliers (ADMM) is adopted to obtain the solution. Simulation results demonstrate the superior performance of the proposed approach.

KEYWORDS

Speech denosing, impulsive noise, sparsity, joint estimation, ADMM.

ACM Reference format:
Ruibo Zhang, Hongqing Liu, Zhen Luo, and Yi Zhou. 2017. Speech denoising in the presence of Impulsive Noise. In *Proceedings of 10th EAI International Conference on Mobile Multimedia Communications, Chongqing, China, July 2017 (MOBIMEDIA'17),* 6 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The speech denoising is a fundamental problem in the applications of speech enhancement and speech recognition [2, 7, 10]. Its main objective is to recover the clean speech from received measurements that are corrupted by noise, or interferences, or both. In this study, noise as a negative factor to the clean speech is considered. Speaking of noise, the widely discussed one is Gaussian because of its simplicity and power to model large of numbers of independent events. Many approaches are developed over the years to handle the speech denoising problem in the presence of Gaussian noise. The spectral subtraction [2] is a popular choice due to its computational efficiency, which performs subtraction of a noise spectral estimate from a noisy speech spectrum. Based on the assumptions that the speech and noise signals are Gaussian distributed, Wiener filtering [10] performs filtering of noisy speech signal by using a filter designed using the minimum mean-square error criterion. In recent studies, based on sparse coding sparse code shrinkage analyzes a statistical method that is shown to be very closely connected to the wavelet shrinkage method. In [7], for Gaussian noise, by soft thresholding of the sparse components, reduction of Gaussian noise is achieved. The advantage of sparse coding method is that the shrinkage nonlinearities can be adapted to the data.

In practice, we often encounter an another type of noise that in distribution has thicker tail than that of Gaussian distribution. This type of noise is impulsive noise and is usually modeled by α-stable distribution [8]. The difficulty of dealing with this noise comes from the fact that its second moment is infinite, and therefore the Gaussian distribution assumption based approaches will fail. To handle impulsive noise, the most popular option is to utilize ℓ_p-norm because $\|x\|_p$ is finite when $p < \alpha$ [11]. However, the noise reduction ability is not superior. Recent studies show that by utilizing the sparse property of the impulsive noise in time domain, the noise can be actively suppressed, and as a result, the noise suppression ability is greatly enhanced [5, 9].

The main objective of this work is to perform the speech recovery in the presence of impulsive noise by utilizing sparse transform domains. To that end, the time domain is still utilized to sparsely represent the noise. For speech signal, in

MOBIMEDIA'17, Chongqing, China

DOI: 10.1145/nnnnnnn.nnnnnnn

DOI 10.4108/eai.13-7-2017.2270053

this work, the wavelet domain is chosen because of its ability of modelling nonstationary signal and providing a sparse representation. As a result, a joint estimation approach is devised to simultaneously estimate the signal and suppress the noise in transform domains. To efficiently solve the optimization problem, the solver based on alternating direction method of multipliers (ADMM) [3] is developed.

2 PROBLEM FORMULATION

In this study, the noise corrupted speech signal is considered, given by

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{n}$ respectively represent the noisy speech, clean speech, and impulsive noise. For more information on the α-stable distribution or the impulsive noise, the intersected readers are referred to [5, 8, 9]. In Figure 1, the recorded clean speech, its wavelet coefficients, and impulsive noise with $\alpha = 1$ are presented. In the wavelet domain, the speech does have a sparse representation since most coefficients are small, whereas in the time domain, the sparse property of the impulsive noise is also observed because of few large spikes and many small amplitudes.

In this work, the matrix $\mathcal{W}$ is used to indicate the wavelet decomposition and $\mathcal{W}^{\mathrm{T}}$ denotes the wavelet reconstruction. With these notations, the received noisy speech in (1) becomes

$$\begin{aligned} \mathbf{y} &= \mathcal{W}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{n} \\ &= \mathcal{W}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{s} + \mathbf{e}, \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}$ denotes the wavelet vector that is sparse, $\mathbf{s}$ is noise sparse vector that represents the large spikes in $\mathbf{n}$, and $\mathbf{e}$ is the remodelling residue, and see [5, 9] for more information on the noise reformulations. To simultaneously recover the speech and suppress the noise, the following optimization problem is devised, given by

$$\|\mathbf{y} - \mathcal{W}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{s}\|_2 + \lambda\|\boldsymbol{\alpha}\|_0 + \tau\|\mathbf{s}\|_0, \tag{3}$$

where $\|\cdot\|_0$ is ℓ_0-norm that is known to enforce sparse solutions [4]. To efficiently obtain the solution in (3), the convex relaxation is usually applied because ℓ_0-norm is NP hard. To perform convex relaxation, ℓ_1-norm is utilized to replace the ℓ_0-norm. That is,

$$\|\mathbf{y} - \mathcal{W}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{s}\|_2 + \lambda\|\boldsymbol{\alpha}\|_1 + \tau\|\mathbf{s}\|_1. \tag{4}$$

By solving (4), one achieves the objective of simultaneous speech recovery and noise suppression. In what follows, an approach is developed based on ADMM method.

3 APPROACH BASED ON ADMM

It is seen from (4) that the optimization in terms of variables of $\boldsymbol{\alpha}$ and $\mathbf{s}$ is separable. To efficiently solve the optimization problem in (4), a two-step iterative process is utilized. First, suppose that $\boldsymbol{\alpha}$ is known, the sub-problem becomes a convex optimization problem with variables $\mathbf{s}$. Second, the sub-problem also is another convex optimization one with variable $\boldsymbol{\alpha}$. In each step, the ADMM is utilized to efficiently obtain the solution.

In the first step, the estimation in terms of $\mathbf{s}$ is rewritten in its equivalent form as

$$\text{minimize } \|\mathbf{y} - \mathbf{s}\|_2^2 + \tau\|\mathbf{s}\|_1 \tag{5}$$

with variable $\mathbf{s}$. For this problem, the ADMM steps for estimating the $\mathbf{s}$ at lth iteration are given by

$$\begin{aligned} \mathbf{s}^{l+1} &= \text{minimize}_{\mathbf{s}}\big(\|\mathbf{y} - \mathbf{s}\|_2^2 + (\rho/2)\|\mathbf{s} - \mathbf{z}^l + \mathbf{u}^l\|_2^2\big) \\ \mathbf{z}^{l+1} &= \text{minimize}_{\mathbf{z}}\big(\tau\|\mathbf{z}\|_1 + (\rho/2)\|\mathbf{s}^{l+1} - \mathbf{z} + \mathbf{u}^l\|_2^2\big) \\ \mathbf{u}^{l+1} &= \mathbf{u}^l + \mathbf{s}^{l+1} - \mathbf{z}^{l+1}. \end{aligned} \tag{6}$$

In the $\mathbf{s}$-step of (6), setting the derivative of the cost function with respect to $\mathbf{s}$ to zero produces

$$-2\mathbf{y} + 2\mathbf{s} + \rho(\mathbf{s} - \mathbf{z}^l + \mathbf{u}^l) = 0. \tag{7}$$

Rearranging the terms in (7) obtains

$$(\rho + 2)\mathbf{s} = \rho(\mathbf{z}^l - \mathbf{u}^l) + 2\mathbf{y}. \tag{8}$$

By the use of (8), the estimate of $\mathbf{s}$ in a closed-form expression is

$$\mathbf{s} = (\rho(\mathbf{z}^l - \mathbf{u}^l) + 2\mathbf{y})/(\rho + 2). \tag{9}$$

In the $\mathbf{z}$-step of (6), based on subdifferential calculus, the estimate of $\mathbf{z}$ is obtained by componentwise soft thresholding as

$$\mathbf{z}^{l+1} = \mathbf{T}_{\lambda/\rho}\big(\mathbf{s}^{l+1} + \mathbf{u}^l\big), \tag{10}$$

where the soft thresholding operator $\mathbf{T}$ is defined by

$$\mathbf{T}_{\lambda/\rho}(a) = \begin{cases} a - \lambda/\rho, & a > \lambda/\rho \\ 0, & |a| < \lambda/\rho \\ a + \lambda/\rho, & a < -\lambda/\rho. \end{cases} \tag{11}$$

It is known that soft thresholding is the proximity operator of the ℓ_1 norm, see [3].

To estimate $\boldsymbol{\alpha}$, the optimization problem now is

$$\text{minimize } \|\mathbf{y} - \mathcal{W}\boldsymbol{\alpha} - \hat{\mathbf{n}}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{12}$$

By symmetry, the ADMM steps of $\boldsymbol{\alpha}$ at lth iteration are

$$\begin{aligned} \boldsymbol{\alpha}^{l+1} &= \text{minimize}_{\boldsymbol{\alpha}}\big(\|\mathbf{y} - \mathcal{W}\boldsymbol{\alpha} - \hat{\mathbf{n}}\|_2^2 + (\rho/2)\|\boldsymbol{\alpha} - \mathbf{g}^l + \mathbf{v}^l\|_2^2\big) \\ \mathbf{g}^{l+1} &= \text{minimize}_{\mathbf{g}}\big(\lambda\|\mathbf{g}\|_1 + (\rho/2)\|\boldsymbol{\alpha}^{l+1} - \mathbf{g} + \mathbf{v}^l\|_2^2\big) \\ \mathbf{v}^{l+1} &= \mathbf{v}^l + \boldsymbol{\alpha}^{l+1} - \mathbf{g}^{l+1}. \end{aligned} \tag{13}$$

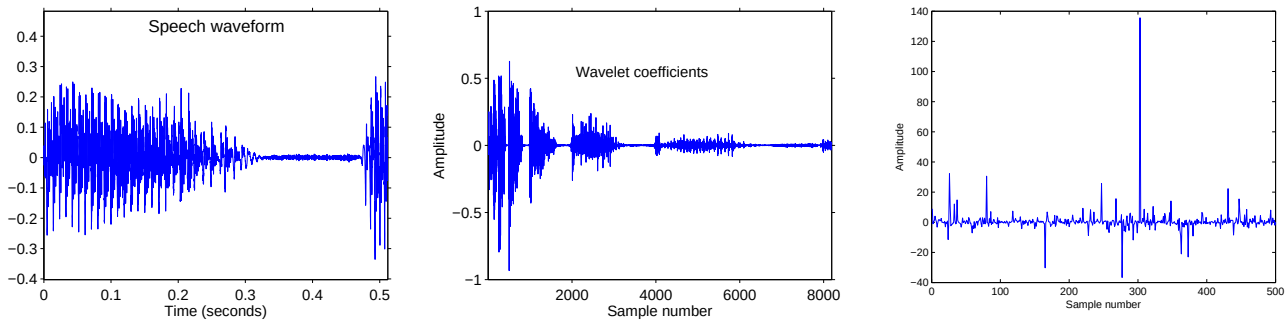
By the procedures developed in (7)-(11), the estimates for $\boldsymbol{\alpha}$ and $\mathbf{g}$ can be respectively calculated by

$$\boldsymbol{\alpha} = (\rho\mathbf{I} + 2\mathcal{W}^{\mathrm{H}}\mathcal{W})^{-1}(\rho(\mathbf{g}^l - \mathbf{v}^l) + 2\mathcal{W}^{\mathrm{H}}(\mathbf{y} - \hat{\mathbf{n}})). \tag{14}$$

$$\mathbf{g}^{l+1} = \mathbf{T}_{\lambda/\rho}\big(\boldsymbol{\alpha}^{l+1} + \mathbf{v}^l\big). \tag{15}$$

To summarize, the steps of solving optimization problem in (4) are presented in Table 1.

For comparison purposes, the joint greedy algorithms in [6] to solve (3) are also conducted. During the study, it is found that StagewiseWeak orthogonal matching pursuit (SWOMP) [1] requires no prior information of the sparse degree of the signal, but using a threshold to select atoms.



(a) Clean speech signal **(b) Wavelet coefficients** **(c) Noise samples of $\alpha = 1$**

Figure 1: Illustration of transforms.

Table 1: Algorithm 1: ADMM based approach.

Objective function: $\|\mathbf{y} - \mathcal{W}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{s}\|_2 + \lambda\|\boldsymbol{\alpha}\|_1 + \tau\|\mathbf{s}\|_1$.
Outputs: Estimates of $\mathbf{s}$ and $\boldsymbol{\alpha}$
Initialization: $l = 1$
Repeat
$l = l + 1$
Step 1: Produce the estimate of $\mathbf{s}$ by (6)-(11)
Step 2: Produce the estimate of $\boldsymbol{\alpha}$ by (13)-(15)
Step 3: Update the residual $\tilde{\mathbf{y}} = \tilde{\mathbf{y}} - (\mathcal{W}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{s})$
Until $l > T$ {maximum iteration} or norm$(\tilde{\mathbf{y}}) < \delta$ {predefined threshold}.

Therefore, it provides certain advantage over the OMP. The SWOMP algorithm is conduced in the following step:
(1) *Initialize* the residue and let the index be a empty set of $\varnothing$.
(2) *Calculate* the inner product. If it is greater than the threshold value, corresponding columns in Φ are selected.
(3) *Estimate* the signal using the new support by least squares (LS) and update the residue.
(4) *Iterate* until certain stopping criterion is met.

By adopting the same concept in [6], the joint SWOMP (JSWOMP) is provided in Table 2 to solve (3) to simultaneously perform speech recovery and noise suppression. For JOMP and JCoSaMP algorithms, the interested readers are referred to [6].

4 NUMERICAL STUDIES

In this section, numerical studies are presented to demonstrate the performance of the proposed joint estimation method. In the all simulations, the maximum iteration and the predefined threshold are 500 and 10^{-8}, respectively. The clean speech signal shown in Figure 1a is utilized, and the α-stable distribution with $\alpha = 1.5$ and different values of γ is used to generate the impulsive noise at different SNR levels, which is defined by SNR $= 10\log_{10}(\frac{P_{\mathrm{sig}}}{2\gamma^{2/\alpha}})$, where P_{sig} indicates the signal power. The recovered speech and noise are provided in Figures 2, 3, 4 5, respectively, and it is seen that ADMM is able to reconstruct the speech and noise accurately, demonstrated in Figure 2b. For joint greedy algorithms such as JCoSaMP, JSWOMP, and JOMP, the recovered speech obtained by JCoSaMP and JSWOMP approaches are close to the clean one, whereas JOMP produces a over sparse solution. To quantitatively access the performance of the proposed method, two performance measures of Segmental signal-to-noise ratio (SegSNR) and perceptual evaluation of speech quality (PESQ) are employed, and their results are presented in Figures 6 and 7, respectively, by average of 50 independent runs. Inspecting Figures 6 and 7 reveals that the ADMM based approach outperforms the joint greedy algorithms, which agrees with the conclusion obtained from the recovered speech. In the greedy algorithms, the JCoSaMP performs the best and JOMP produces the worst indexes, which is also consistent with results in [6].

5 CONCLUSION

In this work, a joint estimation approach of impulsive noise suppression and speech recovery is developed. By utilizing the time domain, the sparse property of the impulsive noise is revealed and clean speech is sparsely represented by wavelet domain. The joint optimization is able to simultaneously recover the speech and suppress the noise in the transfer domains. To efficiently solve the optimization problem, the original optimization problem is decomposed into two sub-problems in which each sub-problem is solved by

Table 2: JSWOMP algorithm.

Inputs: $\mathbf{y}$, Φ, $\mathbf{I}$, T {the number of iterations},
a {the threshold parameter, the default is 0.5}.
Outputs: Estimates of $\mathbf{x}$ and $\mathbf{s}$
Initialization: $t = 1$, $\mathbf{r} = \mathbf{y}$, $\mathbf{r}_{\boldsymbol{\alpha}} = \mathbf{0}$ {the residue for $\boldsymbol{\alpha}$},
$\mathbf{r}_s = \mathbf{0}$ {the residue for $\mathbf{s}$}, $\boldsymbol{\alpha}_0$, and $\mathbf{s}_0$.
Repeat
t=t+1
Step 1: Estimate the signal $\boldsymbol{\alpha}$ as $(\boldsymbol{\alpha}^t, \mathbf{r}_{\boldsymbol{\alpha}}^t) \leftarrow \text{SWOMP}(\mathbf{y}, \mathbf{r}, \mathcal{W})$
Step 2: Estimate the signal $\mathbf{s}$ as $(\mathbf{s}^t, \mathbf{r}_s^t) \leftarrow \text{SWOMP}(\mathbf{y}, \mathbf{r}, \mathbf{I})$
Step 3: Update the global residue as $\mathbf{r} = \mathbf{y} - \mathbf{r}_{\boldsymbol{\alpha}}^t - \mathbf{r}_s^t$
Until $t > T$ {maximum iteration} or $\text{norm}(\mathbf{r}) < \delta$ {predefined threshold}.

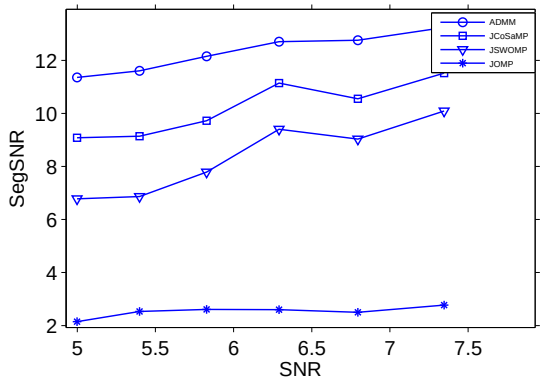


Figure 6: SegSNR of different approaches in terms of SNR.

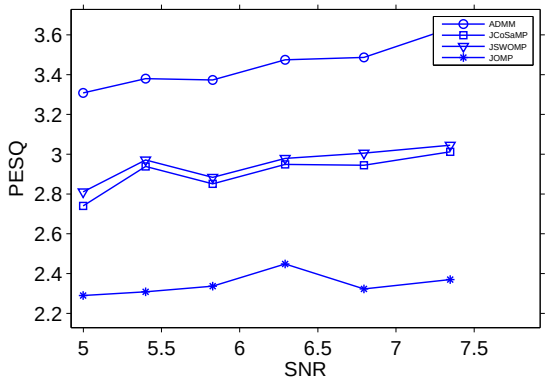


Figure 7: PESQ of different approaches in terms of SNR.

ADMM approach. The numerical simulations demonstrate that the ADMM based approach outperforms the greedy algorithms based approaches by generating much better recovered speech and higher performance measure indexes.

ACKNOWLEDGMENT

This work was jointly supported by the National Natural Science Foundation of China under Grant 61501072, by Foundation and Advanced research projects of Chongqing Municipal Science and Technology Commission under Grant cstc2015jcyjA40027.

REFERENCES

[1] T. Blumensath and M. E. Davies. 2009. Stagewise Weak Gradient Pursuits. *IEEE Transactions on Signal Processing* 57, 11 (2009), 4333–4346.

[2] S. Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*. 113–120.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* 3, 1 (2011), 1–122.

[4] E. J. Candès and T. Tao. 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* 35, 35 (2007), 2313–2351.

[5] H.Q.Liu, D. Ding, Y. Li, and Y. Zhou. 2015. Frequency estimation with missing measurements under impulsive noise. The 2015 8th International Congress on Image and Signal Processing (CISP 2015), Shenyang, China.

[6] H.Q.Liu, Y. Li, Y. Zhou, and Trieu-Kien Truong. 2016. Greedy pursuit algorithms for sparse signal reconstruction in the case of impulsive noise. IEEE International Conference on Digital Signal Processing (DSP'16), Beijing, China.

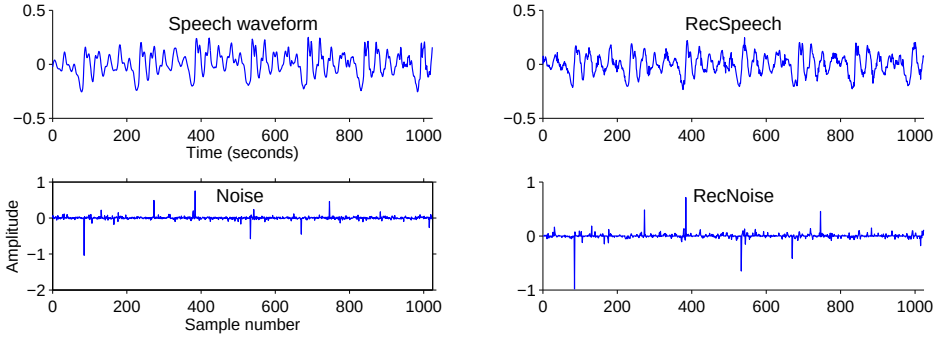
[7] A Hyvarinen. 1999. Sparse Code Shrinkage: denoising of Non-gaussian Data by Maximum Likelihood Estimation. *Neural Computation* 11, 7 (1999), 1739–1768.

[8] E. E. Kuruoglu. 1998. *Signal Processing in alpha-Stable Noise Environments: A Least l_p-Norm Approach*. Ph.D. Dissertation. Department of Engineering, University of Cambridge.

[9] H. Q. Liu, Y. Li, Y. Zhou, H.-C. Chang, and T.-K. Truong. 2016. Impulsive noise suppression in the case of frequency estimation by exploring signal sparsity. *Digital Signal Processing* 57 (Jun. 2016), 34–45.

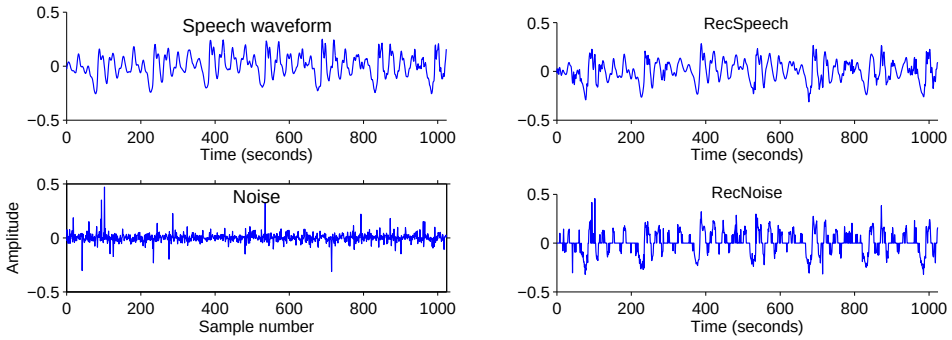
[10] I. Y. Soon and S. N. Koh. 2000. Low distortion speech enhancement. *IEE Proceedings - Vision Image and Signal Processing* 147, 3 (2000), 247–253.

[11] W.-J.Zeng, H.C.So, and X.Jiang. 2016. Outlier-robust greedy pursuit algorithms in lp-space for sparse approximation. *IEEE Transactions on Signal Processing* 64, 1 (Jan. 2016), 60–75.



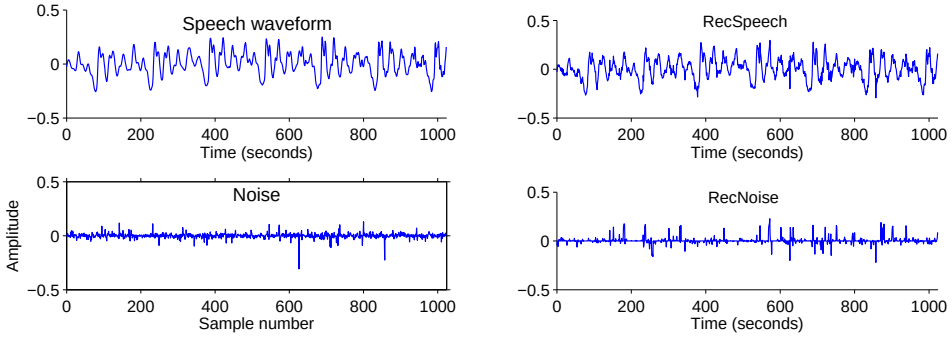
(a) The original speech and noise (b) ADMM reconstruction

Figure 2: Signal and noise reconstructions by the ADMM method.



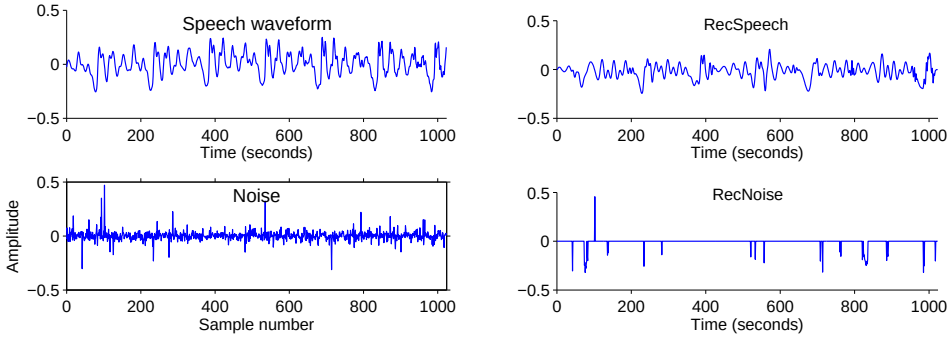
(a) The original speech and noise (b) JCoSaMP reconstruction

Figure 3: Signal and noise reconstructions by the JCoSaMP method.



(a) The original speech and noise (b) JSWOMP reconstruction

Figure 4: Signal and noise reconstructions by the JSWOMP method.



(a) The original speech and noise

(b) JOMP reconstruction

Figure 5: Signal and noise reconstructions by the JOMP method.