ADAPTIVE SPEECH ENHANCEMENT USING SPARSE PRIOR INFORMATION

*Zhimin Xiang and Yuantao Gu**

State Key Laboratory on Microwave and Digital Communications
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, CHINA

ABSTRACT

In recent years, sparse representation is adopted to improve the quality of noise corrupted speech. However, the representation of noise is also found to be sparse in some special cases, which degrades the performance of sparsity based speech enhancement. An adaptive speech enhancement algorithm using sparse prior information is proposed in this paper. In the proposed method, speech enhancement is casted to an optimization problem, where linear prediction (LP) residual and DCT coefficients are combined and adopted as the representation of speech to ensure that noise is dense in the such domain. Other features, including speech energy, noise energy, and interframe correlation are also considered as constraints to improve the quality and intelligibility of recovered speech. Experiment results show that the proposed algorithm exceeds the reference algorithms in various noise scenarios, especially, in the cases of narrowband noise and low SNR.

Index Terms— Adaptive speech enhancement, sparse representation, linear prediction, energy constraint, interframe correlation

1. INTRODUCTION

In practical systems, such as voice communication and speech recognition, noise is almost inevitable. Noise degrades the performance of these systems dramatically. For example, speech recognition accuracy likely suffers greatly in the presence of noise. Therefore, it is essential to reduce noise effectively with signal processing techniques. Over the past four decades, a number of speech enhancement algorithms have been developed. Though some of them have been utilized in commercial schemes, there is still a gap between the human desire and ready-made technologies.

This paper focuses on the enhancement for speech corrupted by additive noise in single channel systems. The conventional speech enhancement algorithms can be roughly divided into three categories [18]: spectral-subtractive algorithms [1–3], statistical-model-based algorithms [4–6], and subspace algorithms [7–9]. Most of them focus on the spectrum estimation of noise or speech.

Recently, sparse representation is extensively investigated. Sparsity is an important feature of speech, which has been found approximately sparse in some transform domains, for example, DCT domain and wavelet domain [10], and over redundant dictionaries of clean speech exemplars [11, 12]. Linear prediction (LP) residual of speech has also been found sparse [15–17]. Sparse prior information has been introduced into speech coding and speech enhancement [10–14], which achieve good performance. In [10], the sparsity of DCT coefficients is adopted for speech enhancement. The enhancement algorithm proposed in [11] consideres the sparsity over the redundant dictionary of clean speech exemplars. In [14], the prior information that the DCT-II coefficients of speech are sparse over the redundant dictionary is used to improve the speech quality.

One implicit assumption in these enhancement algorithms [10, 11, 14] is that the representation of speech in a transform domain or over a redundant dictionary is sparse, while that of noise is dense. Based on this assumption, clean speech can be recovered by finding the sparse representations. However, some kinds of noise are also found sparse in the above representation scenarios, which results in degradation of enhancement performance. For example, since coefficients of car interior noise are sparse in DCT domain, the speech enhancement performance for car interior noise is not as good as that in other noisy background [10]. In addition, some features, for example, speech energy and the interframe correlation, are not considered sufficiently in the available speech enhancement algorithms [10–14], which probably hinders the further improvement of speech quality.

In this paper, an adaptive speech enhancement algorithm using sparse prior information (ASESI) is proposed. In this algorithm, LP residual is adopted as the representation of speech in order to keep the parameters of speech sparse while that of noise dense. The energy constraint and interframe correlation are also adopted into the proposed algorithm to improve the quality and the intelligibility of recovered speech. The clean speech is recovered by finding the component whose LP residual and DCT coefficients are both sparse under the energy and the interframe correlation constraints, i.e., by solving an optimization problem. This formalized problem can be solved with numerous existing methods. LP coefficients, speech energy, and interframe correlation, which are used to recover speech, are the distinctive features for each frame, which reveals the adaptive behavior of the proposed algorithm. Experimental results confirm that a wide range of noise can be reduced effectively through the proposed approach.

2. SPARSITY OF LP RESIDUAL

This work considers the following LP model,

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n) \tag{1}$$

where $x(n), r(n), a_k$, and K denote the speech signal, the LP residual, the LP filter coefficients and order, respectively. The LP coefficients are estimated by minimizing the least squares of LP residual.

In this paper, it is assumed that the initial state of filter, i.e., $x(n) = 0, \forall n \leq 0$. Hence, according to (1), LP residual vector $\mathbf{r} =$

The corresponding author of this paper is Yuantao Gu (gyt@tsinghua.edu.cn).

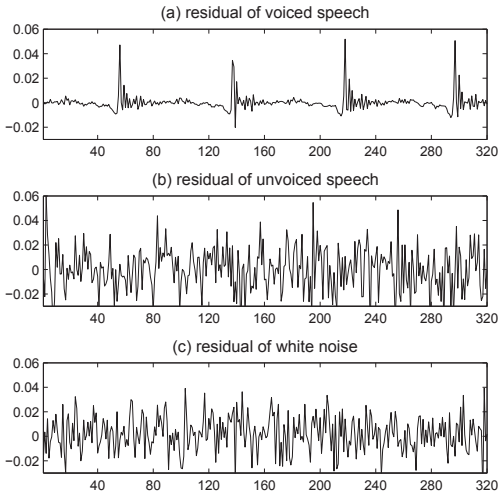

Fig. 1. Example of the residual of voiced speech, unvoiced speech and white noise. The LP filter order is 10 and the frame length is 320.

$[r(1), r(2), \cdots, r(N)]^{\mathrm{T}}$ can be expressed using the speech vector $\mathbf{x} = [x(1), x(2), \cdots, x(N)]^{\mathrm{T}}$ and the LP coefficient matrix as

$$\mathbf{r} = \mathbf{A}\mathbf{x}, \tag{2}$$

where $\mathbf{A}$ is the $N \times N$ matrix derived as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -a_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & & & \vdots \\ -a_K & \cdots & -a_1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -a_K & \cdots & -a_1 & 1 & 0 & \cdots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & -a_K & \cdots & -a_1 & 1 \end{bmatrix}.$$

For sparsity based speech enhancement algorithms, the enhancement performance closely depends on the sparsity distinction between the representation of speech and that of noise. Considering the LP residual of voiced frame is approximately sparse, as can be seen from Fig.1(a), it is highly possible that the spectrum of noise is significantly different from the spectrum of voiced speech. Hence, the representation of noise is very likely to be dense. This can be seen from Fig.1(c). Above all, the significant sparsity distinction can be guaranteed by adopting the LP residual as the representation, which probably leads to an improvement of enhancement performance. In Section 4.1, an experiment will be conducted to evaluate the sparsity of the representation of voiced speech and noise.

From Fig.1(b), it can be seen that the representation of unvoiced speech is also dense, which corresponds to its noise-like feature. Therefore, such frame is probably not recovered accurately. Since the quality and intelligibility of speech is mainly decided by the voiced speech, the focus of recovering voiced speech will not heavily reduce the performance of the proposed algorithm.

3. THE PROPOSED ALGORITHM

In this section, the main contribution of this work, an adaptive speech enhancement algorithm using sparse prior information, is introduced in detail. The proposed algorithm adopts LP residual as one of the sparse representation of speech, considering it is feasible and advantageous, as analyzed in the previous section. To make full use of the sparsity of speech, DCT coefficients are also included to contribute as a measurement. The proposed algorithm aims to recover the clean speech, whose LP residual and DCT coefficients are both sparse, via solving an optimization problem under a series of constraints. The optimization problem is formulated in the first subsection, then its solution is introduced in the second one.

3.1. Formulation of the optimization problem

Let $\mathbf{y} = [y(1), \cdots, y(N)]^{\mathrm{T}}$ be the noisy speech,

$$\mathbf{y} = \mathbf{x} + \mathbf{e}, \tag{3}$$

where $\mathbf{e}$ denotes the additive noise.

The proposed optimization problem is to solve a balanced minimization,

$$\hat{\mathbf{x}} = \min_{\mathbf{z} \in \mathbb{R}^N} \lambda \|\tilde{\mathbf{A}}\mathbf{z}\|_1 + (1-\lambda)\|\mathbf{D}\mathbf{z}\|_1, \tag{4}$$

subject to three constraints as follows

$$\alpha_1 \tilde{E}_{\mathbf{x}} \le \|\mathbf{z}\|_2^2 \le \alpha_2 \tilde{E}_{\mathbf{x}}, \tag{5}$$

$$\|\mathbf{y} - \mathbf{z}\|_2^2 \le \tilde{E}_{\mathbf{e}}, \tag{6}$$

$$\|\hat{\mathbf{x}}_2' - \mathbf{z}_1\|_2^2 \le \epsilon \tilde{E}_{\mathbf{x}}, \tag{7}$$

where $\hat{\mathbf{x}}$ and $\mathbf{z}$ denote the recovered speech and its candidate, $\mathbf{D}$ denotes DCT matrix, $\lambda, \alpha_1, \alpha_2$, and ϵ are parameters, respectively. Other parameters and the meanings of (4), (5), (6), and (7) will be explained later.

The objective function (4) ensures that both LP residual and DCT coefficients of the recovered speech are sparse, where λ is a factor used to balance these two types of sparsity. In the ideal situation, the LP coefficients of clean speech should be applied for the best performance. However, $\mathbf{A}$ is obviously not available in practice. Hence, its estimate, $\tilde{\mathbf{A}}$, is adopted in (4), while the estimation method will be introduced in detial in Section 3.2.

In order to improve the accuracy, the energy of the recovered speech is constrained to be close to that of the clean speech, i.e., the first constraint of (5), where α_1 and α_2 describe the degree of approximation. To the same reason that the clean speech is not available, the clean speech energy is replaced by its estimate, $\tilde{E}_{\mathbf{x}}$. The second constraint of (6) is rather similar to the first one. It puts a condition on the estimated noise, which is demanded not to be too large, using an estimated noise energy, $\tilde{E}_{\mathbf{e}}$.

Please notice that the consecutive speech frames are overlapped in the proposed model. Consequently, one may ready to demand that the overlapping fragments of the recovered speech frames are as close as possible, as described in (7), where $\hat{\mathbf{x}}_2'$ and $\mathbf{z}_1$ denote the last P samples of the previous recovered speech $\hat{\mathbf{x}}'$ and the first P samples of $\mathbf{z}$, respectively. Parameter ϵ is used to control the correlation level. Actually, this constraint probably results in an improvement of the intelligibility of recovered speech.

Based on the discussion above, clean speech will be recovered, if its LP residual and DCT coefficients are both sparse under the constraints of moderate energy and correlation with previous frame, which conveys to the main idea of the proposed algorithm.

3.2. Solution of the optimization problem

Based on the above analysis, the enhancement is conducted in two steps: first estimating and setting the parameters, and then solving the optimization problem.

Step 1: Estimating and setting parameters

The energy of clean speech and noise, as well as the LP coefficients of clean speech, need to be estimated to activate the solution. Iterative Wiener filtering algorithm [4] is selected to do preprocess and the idea of energy prediction is applied in the estimation.

First, the noisy speech frame is processed using iterative Wiener filtering algorithm. The energy and the LP coefficients of the processed speech are calculated and denoted by E_{w} and $\mathbf{a}_{\mathrm{w}}$, respectively. Then the estimated upper bound and lower bound of the clean speech energy of the current frame, which are denoted by E_{u} and E_{l}, are predicted based on the previous recovered frame. The energy of $\hat{\mathbf{x}}_2'$ is adopted as the lower bound. The upper bound, which is controlled by the noisy speech energy of the current frame, is obtained by

$$E_{\mathrm{u}} = E_{\mathrm{l}} + E_{\mathbf{y}_3}, \tag{8}$$

where $E_{\mathbf{y}_3}$ denotes the energy of the last $N-P$ samples of the noisy speech frame.

If E_{w} is in the interval of E_{l} and E_{u}, the estimates from the preprocessed frame are considered as acceptable to the recovery of clean speech, or else, the predicted values computed from the previous recovered frame are preferred. Therefore, the estimates of clean speech energy and LP coefficients used for optimization are obtained by

$$\{\tilde{E}_{\mathbf{x}}, \tilde{\mathbf{a}}\} = \begin{cases} \{E_{\mathrm{w}}, \mathbf{a}_{\mathrm{w}}\} & E_{\mathrm{l}} \le E_{\mathrm{w}} \le E_{\mathrm{u}}; \\ \{E_{\mathrm{l}}, \hat{\mathbf{a}}'\} & E_{\mathrm{w}} < E_{\mathrm{l}}; \\ \{E_{\mathrm{u}}, \hat{\mathbf{a}}'\} & E_{\mathrm{w}} > E_{\mathrm{u}}, \end{cases} \tag{9}$$

where $\hat{\mathbf{a}}'$ denotes the LP coefficients of the previous recovered frame.

Consequently, the estimate of noise energy is obtained as

$$\tilde{E}_{\mathbf{e}} = E_{\mathbf{y}} - \tilde{E}_{\mathbf{x}} \tag{10}$$

where $E_{\mathbf{y}}$ is the energy of noisy speech frame.

Parameters λ, α_1, α_2, and ϵ used in the optimization problem need to be set. According to (4), a smaller λ leads to the sparser DCT coefficients, while a larger λ leads to the sparser LP residual. To be noticed, matrix $\mathbf{D}$ used in calculated DCT coefficients is a predefined, while matrix $\tilde{\mathbf{A}}$ used in producing LP residual is an estimate. Therefore, based on our observation, λ is set small in low SNR scenario to reduce the misleading of the estimated LP coefficients. Additionally, for the noise whose DCT coefficients are also sparse, λ is set large. According to (7), a smaller ϵ leads to the stronger interframe correlation. Therefore, ϵ is set large to reduce the error propagation from the previous frame to the current frame when the SNR is low. Finally, α_1 and α_2 are set close to 1 so that the energy error between recovered speech and clean speech is kept small.

Step 2: Solving the optimization problem

The optimization problem is a constrained nonlinear programming problem, which has been extensively and intensively studied. Hence, after obtaining the parameters in (4), (5), (6), and (7), it can be solved using available methods.

4. EXPERIMENTS

In this section, experiments are conducted to verify the idea and the performance of the proposed algorithm. Clean speech is extracted from TIMIT database and downsampled at 8kHz. Various types of noise, including stationary white Gaussian noise (WGN), car interior noise (CIN), and F16 cockpit noise (FCN), are obtained from Noisex-92 database. Noisy speech is produced by adding the above mentioned noise to the clean speech at -5 dB, 0 dB, 5 dB and 10 dB, respectively. In these experiments, the frame is 32 ms (256 samples) long and overlapped for 24 ms (192 samples). Tenth-order LP analysis is adopted.

Before testing the proposed algorithm, we firstly explore the sparsity of speech and various types of noise in LP residual domain and DCT domain by signal compressibility.

4.1. Signal compressibility experiment

Signal compressibility is measured with the averaged mean squared reconstruction errors (MSRE) of signal energy compaction [10]. MSRE for a given signal $\mathbf{x}$ is defined as

$$\varepsilon(\mathbf{x}, \tau) = \|\mathbf{x} - \mathbf{\Phi}^{-1}\lceil\mathbf{X}\rfloor_\tau\|_2^2 \tag{11}$$

where

$$\mathbf{X} = \mathbf{\Phi}\mathbf{x},$$

$\mathbf{\Phi}$ is a transform matrix, and $\lceil\mathbf{X}\rfloor_\tau$ is the signal obtained by keeping its $\tau \times 100\%$ largest values and setting the rest to zeros. According to (11), for a fixed τ, the smaller the MSRE is, the sparser the representation of $\mathbf{x}$ is. For a given signal set $\mathcal{S} = \{\mathbf{x}_n, n = 1, \cdots, T\}$, the averaged MSRE is defined as

$$\bar{\varepsilon}(\mathcal{S}, \tau) = \frac{1}{T}\sum_{n=1}^{T}\varepsilon(\mathbf{x}_n, \tau) \tag{12}$$

In this experiment, the averaged MSREs of voiced speech, WGN, CIN, and FCN are compared. First, let $\mathbf{\Phi}$ be the transform matrix $\mathbf{A}$ in (2) constructed from the LP coefficients of the voiced speech. The result is shown in Fig.2. Then, let $\mathbf{\Phi}$ be the DCT matrix. The result is shown in Fig.3.

The result in Fig.2 shows that the LP residual of speech is significantly sparser than that of noise, which confirms the advantage of adopting LP residual of speech as the representation. From Fig.3, it can be clearly seen that the DCT coefficients of CIN are sparser than that of the speech, which indicates that the speech corrupted by CIN may not be accurately recovered by the only sparsity constraint in DCT domain. This result verifies the proposed approach of combining the sparsity measurement in both LP residual domain and DCT domain. Furthermore, one may predict that the proposed algorithm works excellent in the narrow band noise scenario, i.e., CIN scenario.

4.2. Performance of the proposed algorithm

In this experiment, the performance of the proposed algorithm (ASESI) is evaluated and compared with some reference algorithms, including that using DCT coefficients as the representation (DCTNR) [10] and iterative Wiener filtering algorithm (IWF) [4]. The proposed algorithm using accurate parameters (ASESI-OPT) other than estimated values is also compared to provide an insight. The Perceptual Evaluation of Speech Quality (PESQ) [18] is adopted to evaluate the recovered speech. Please notice that though IWF algorithm

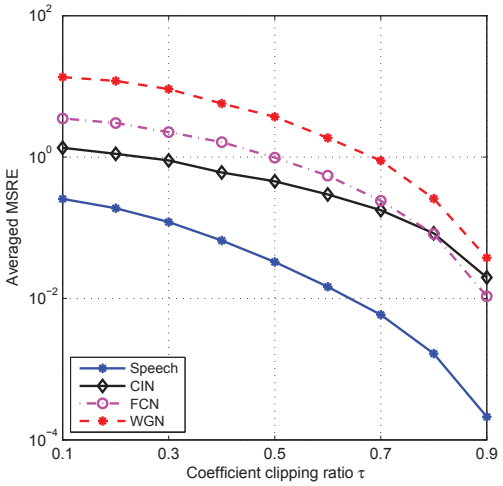

Fig. 2. Comparison of averaged MSRE between speech and various types of noise when $\mathbf{\Phi} = \mathbf{A}$.

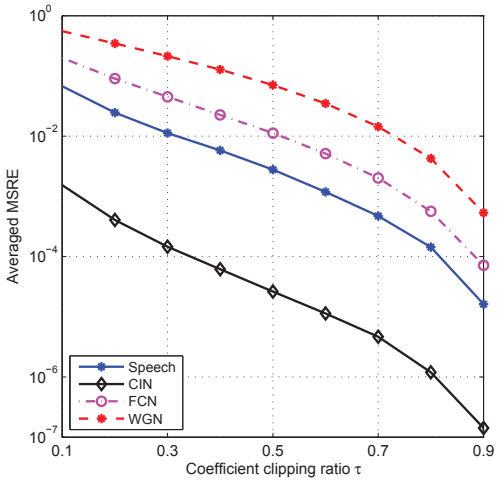

Fig. 3. Comparison of averaged MSRE between speech and various types of noise when $\mathbf{\Phi} = \mathbf{D}$.

Table 1. Values of λ in The Optimization Problem

	WGN		CIN		FCN	
	EST	OPT	EST	OPT	EST	OPT
-5dB	0.20	0.40	0.50	0.50	0.20	0.43
0dB	0.22	0.43	0.52	0.53	0.23	0.43
5dB	0.25	0.44	0.54	0.56	0.25	0.44
10dB	0.30	0.45	0.58	0.60	0.28	0.45

serves as a preprocessing step for the proposed algorithm, it still needs to be compared to demonstrate the improvement of our contribution.

According to the analysis in Section 3.2, the values of the parameters are set as follows. α_1 and α_2 are set to 0.98 and 1.2, respectively. Corresponding the SNR of -5 dB, 0 dB, 5 dB, 10 dB, ϵ is 0.3, 0.24, 0.2, 0.12 for ASESI, and set to 0.1, 0.08, 0.06, 0.06 for ASESI-OPT. The values of λ in ASESI and ASESI-OPT are shown in Table 1, where the column of EST corresponds to ASESI and OPT corresponds to ASESI-OPT. Function *fmincon* in MATLAB is used to solve the optimization problem. PESQ of each algorithm is shown in Table 2.

From Table 2, it can be seen that ASESI outperforms DCTNR in all noisy conditions and performs better than IWF at low SNR, which indicates that the proposed algorithm is effective to reduce various types of noise, especially in the heavy noise scenario. Moreover, ASESI significantly outperforms DCTNR for CIN, which shows the advantage of the proposed algorithm for the noise whose representation in particular transform domains is sparse. The result also shows that the performance of ASESI-OPT is significantly better than that of DCTNR, IWF, and ASESI, which indicates that it is promising to further improve the performance of ASESI by increasing the estimation accuracy of the parameters.

Table 2. PESQ of Respective Enhancement Algorithms

		DCTNR	IWF	ASESI	ASESI-OPT
WGN	-5dB	1.6890	1.4520	1.7816	2.3932
	0dB	1.9656	1.9168	2.0002	2.6981
	5dB	2.2328	2.3041	2.3867	2.9990
	10dB	2.4842	2.7080	2.5535	3.2576
CIN	-5dB	0.7496	0.7132	1.4020	2.0751
	0dB	1.0800	1.3866	1.6403	2.4123
	5dB	1.8228	2.3940	2.2041	2.6304
	10dB	2.2442	2.8807	2.5826	3.0294
FCN	-5dB	1.3375	1.2443	1.5058	2.0809
	0dB	1.7103	1.7564	1.8061	2.3492
	5dB	2.0634	2.1891	2.1759	2.6819
	10dB	2.4824	2.7047	2.4953	2.9773

5. CONCLUSION

In this paper, an adaptive speech enhancement algorithm using sparse prior information is proposed. To keep the representation of noise dense, as well as to improve the quality and intelligibility of the recovered speech, both the LP residual and DCT coefficients are adopted and balanced as the sparsity measurement. Furthermore, the energy and interframe correlation are considered as additional constraints. Experiment results confirm that the representation of speech and noise are sparse and dense, respectively and clearly. Experiment results also confirm that noise can be reduced effectively using the proposed algorithm. Additionally, experiment results indicate that the performance of the proposed algorithm will be improved by enhancing the estimation accuracy of the parameters in the optimization model.

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[2] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", in*Proc. ICASSP2002*, vol. 4, pp. IV-4164, May 2002.

[3] H. Gustafsson, S. E. Nordholm and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799-807, Nov. 2001.

[4] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech", *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 26, no. 3, pp. 197- 210, Jun. 1978.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[6] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors", in *Proc. ICASSP2002*, vol. 1, pp. I-253-I-256, May 2002.

[7] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, July 1995.

[8] H. Yi and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, July 2003.

[9] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, Feb. 2001.

[10] W. Dalei, Z. Wei-Ping and M. N. S. Swamy, "On sparsity issues in compressive sensing based speech enhancement", in *Proc. ISCAS2012*, pp. 285-288, May 2012.

[11] J. F. Gemmeke, H. Van Hamme, B. Cranen and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition", *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272-287, Apr. 2010.

[12] C. D. Sigg, T. Dikk and J. M. Buhmann, "Speech enhancement with sparse coding in learned dictionaries", in *Proc. ICASSP2010*, pp. 4758-4761, March 2010.

[13] H. Yongjun, H. Jiqing, D. Shiwen, Z. Tieran and Z. Guibin, "A solution to residual noise in speech denoising with sparse representation", in *Proc. ICASSP2012*, pp. 4653-4656, March 2012.

[14] H. Feng, L. Tan and W. B. Kleijn, "Transform-domain wiener filter for speech periodicity enhancement",in *Proc. ICASSP2012*, pp. 4577-4580, March 2012.

[15] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction", *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 103-106, Jan. 2010.

[16] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen and M. Moonen, "Sparse linear prediction and its applications to speech processing", *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 20, no. 5, pp. 1644-1657, July 2012.

[17] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals", in*Proc. ICASSP2009*, pp. 4125-4128, Apr. 2009.

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