Two Stage Approaches for the Detection and Suppression of Typed Keystrokes in Speech Signals

Rizwan Ullah
Department of Electronic Engineering and Information Science, USTC, Hefei, Anhui, 230027 China

Renjie Tong
Department of Electronic Engineering and Information Science, USTC, Hefei, Anhui, 230027 China

Yawar Ali Sheikh
Department of Electronic Engineering and Information Science, USTC, Hefei, Anhui, 230027 China

Zhongfu Ye
Department of Electronic Engineering and Information Science, USTC, Hefei, Anhui, 230027 China

ABSTRACT

In recent decades, keystroke suppression has got a particular attention due to the increasing use of laptops and computers to capture audio in various communication scenarios such as meetings, audio/video instant messaging etc. In many of these situations, a unique problem of additive keystroke transient noise is faced. Because of the non-stationary, short time and abrupt nature of the keystroke transient, it has been a challenging task for many years. In this paper, two new two-stage approaches for the suppression of keystrokes are proposed. In the first stage the speech is estimated using supervised sparse non-negative factorization, which is common in both of the methods. Then, in the second stage, keystrokes are detected and are suppressed by replacing the corrupted speech frames with the corresponding estimated speech frames obtained in the first stage using two new techniques, which is the core contribution of this work. Experimental results show that the proposed approaches exhibit good performance without significantly degrading the quality of speech.

Keywords

Single channel speech enhancement, short time Fourier transform, supervised sparse non-negative matrix factorization, correlation, keystrokes suppression, thresholding technique.

1. INTRODUCTION

Reducing highly nonstationary keystroke transients has been a goal of long standing in the field of single channel speech enhancement. It has got potential applications in many areas of speech processing including speech/talker recognition, signal processing for communication and hearing aids. The goal of keystrokes suppression is twofold: to improve the perceived quality and the intelligibility of speech by attenuating the interferer without substantially degrading the speech to reduce the listener fatigue [2].

Mainly the speech enhancement algorithms depend on the specific application, the type of noise involved and the statistical relation between the clean speech and the type of noise involved. Therefore it is very difficult, although not impossible to find a versatile algorithm that can work in all applications and in every real time environment [3].

Keystrokes suppression is a challenging task, whose energy is widely spread across the frequency domain, appearing in the form of short time clicks or vertical strips in speech spectrogram, and has an immense effect on the overall speech quality. In a big hall with large number of sub offices, different people typing on keyboards, the keystrokes become troublesome for many people including the people who use hearing aids. Nowadays, Laptops and personal computers are increasingly being used as recording devices to capture meeting, interviews, video conferencing, voice over internet protocol communications and lectures for archival purposes using the laptop's local mic. In these scenarios, the user frequently also uses the same laptop to make notes. Because of the close proximity of the laptop's mic to its keyboard, the captured speech signal is severely distorted by the impulsive sounds, the user's keystrokes generate, which degrades the quality of speech [4]. Also reducing such transients are important for speech recognition or speech perception and audio/video instant messaging or someone typing during a voice call. These degradations lower the intelligibility of speech signal, so that listener's ability to understand the message suffers. Thus, there is a need to suppress the keystroke noise in recorded speech signals that results in significant perceptual improvement.

Over the years, a variety of different noise reduction algorithms have been proposed such as spectral subtraction, statistical based methods, Wiener filtering and subspace techniques, the performance of which are mostly dependent on the estimated noise statistics in the absence of speech activity. One of the earlier approaches to noise reduction is spectral subtraction, but some annoying artifact called musical noise is often observed in the post processed signal [5]. Previously this problem was solved by modeling the transient noise as isolated impulse noise [1]. Nevertheless, the method has one drawback that it is unable to suppress long-time transient noise.

An approach using nonlocal filtering has been proposed for keystrokes suppression [6–8]. First, the keystrokes are enhanced using a modified speech estimator. Then, the geometric structure of the keystrokes are learnt using diffusion maps, which is utilized to estimate the keystroke PSD using nonlocal diffusion filtering. Finally, the keystrokes are suppressed and the speech is enhanced by Optimally

Modified Log-Spectral Amplitude (OM-LSA) filter equipped with an estimate of the keystroke PSD. But the main drawback of this method is the assumption that the same keystroke pattern appears several times in the measurement. Thus, a single keystroke is generally not identified as interference, and hence not suppressed [9]. Recently, there has been an increasing interest in using sparse representation and dictionary learning methods for speech enhancement. But using only these methods may lead to low speech quality. Therefore the nonnegative sparse coding technique is combined with two new techniques namely (a) Thresholding based technique and (b) Correlation based technique, which lead to a good quality estimated speech. Experimental results show that the proposed methods outperform several state of the art methods as well as transient suppression algorithms.

2. PROBLEM DESCRIPTION

The keystrokes corrupted noisy signal in the short time Fourier transform (STFT) domain is usually linearly modeled as

$$X(f,n) = S(f,n) + N(f,n) \qquad (1)$$

where $X(f,n)$, $S(f,n)$ and $N(f,n)$ are the noisy signal, speech and noise spectrograms at frequency bin f and frame number n, respectively. The main motivation for using the STFT in noise suppression applications is that there exist synthesis formulae by which a time series can be exactly reconstructed from STFT representation [10].

3. PROPOSED METHODOLOGY

The single channel source separation problem usually involves solving the following problem:

$$\min_{V,H} \|S - VH\|_F^2 + \lambda\|H\|_1 \text{ s. t. } V_{i,j} \geq 0, H_{i,j} \geq 0 \qquad (2)$$

where S is the magnitude spectrogram of the signal, V is the dictionary matrix, H is the weight matrix, λ is the sparsity weight, $V_{i,j}$ denotes the $i-th$, $j-th$ element of the dictionary matrix and $H_{i,j}$ denotes the $i-th, j-th$ element of the weight matrix.

In this paper, two two-stage approaches for the keystrokes suppression in speech signals have been proposed. Stage-I is the same for both of the approaches. In stage-I sparse non-negative matrix factorization (SNMF) is used to estimate the speech signal. Then, in the second stage two methods are proposed for further enhancement of speech signal. The first one is thresholding technique (SNMF-TT), which is applied to the noisy speech signal to detect the keystrokes. The detected keystrokes are suppressed by replacing the corresponding frames from the estimated speech signal acquired from SNMF. In the second method (SNMF-CR), correlation is taken between the noisy signal and the resultant estimated signal acquired from stage-I, and based on low correlation coefficient, noise corrupted frames in the original noisy speech signal are replaced with the corresponding estimated speech signal obtained from stage-I. Thus, the noise free speech frames remain unchanged in both of the methods, due to which the speech quality and intelligibility is significantly improved, which is the main contribution towards the keystrokes suppression in this paper.

4. TRAINING AND ENHANCEMENT USING SUPERVISED SNMF

This stage consists of training stage and de-noising stage.For the training stage, the availability of clean speech spectrogram S_{speech}, of size $n_f \times n_{st}$ and a speech-free noise spectrogram S_{noise}, of size $n_f \times n_{nt}$ are assumed, while n_f representing the number of frequency bins and n_{st} and n_{nt} represents the number of speech frames and noise frames, respectively. Since different objective functions lead to different variants of NMF, but Kulback-Leibler (KL) divergence between X and VH, was found to work well for speech separation and enhancement problems [9] [11], so, for the proposed method, the focus is only on KL-divergence as given below [12]:

$$D(X\|VH) = \sum_{i,j}\left[X_{i,j}log\frac{X_{i,j}}{(VH)_{i,j}} - X_{i,j} + (VH)_{i,j}\right]$$

$$\forall ij, V_{i,j}, H_{i,j} \geq 0, \qquad (3)$$

where $X_{i,j}$ denotes the $i-th$, $j-th$ element of the magnitude spectrogram of the noisy speech, $V_{i,j}$ denotes the $i-th, j-th$ element of the dictionary matrix and $H_{i,j}$ denotes the $i-th, j-th$ element of the weight matrix. To train the speech and noise dictionaries, sparse NMF is separately performed on magnitude of the STFT of the available training data to find the basis for speech and noise.

$$S_{train} = V_{speech}H_{speech} \qquad (4)$$

$$N_{train} = V_{noise}H_{noise} \qquad (5)$$

where V_{speech} and V_{noise} denote the dictionaries of speech and noise, respectively, each of size $n_f \times n_b$, where n_b denotes the number of basis vectors to represent each source. H_{speech} denotes the weight matrix of speech and H_{noise} denotes the weight matrix of noise. The dictionaries are kept fixed and the weight matrices are discarded as they are not being used in the proposed approach.

During the enhancement stage, the basis are fixed, V_{speech} and V_{noise}, and concatenate them to form V_{all} of size $n_f \times 2n_b$, and only H is updated. The noisy speech magnitude spectrogram X is decomposed as follows

$$X = [V_{speech}, V_{noise}]H \qquad (6)$$

Finally, to reconstruct the de-noised speech magnitude spectrogram, the basis matrix is multiplied with its corresponding activation matrix i.e. the top n_b rows of H.

$$S_e = V_{speech}H_s \qquad (7)$$

$$N_e = V_{noise}H_n \qquad (8)$$

where S_e represents the estimated speech magnitude spectrogram and H_s denotes the sub matrix inside H which corresponds to the speech basis.

$$H = \begin{bmatrix} H_s \\ H_n \end{bmatrix} \qquad (9)$$

The noisy signal's phase and inverse STFT is used to obtain the estimated speech signal in time domain.

5. TYPED KEYSTROKES DETECTION AND SUPPRESSION

Keystrokes are in the form of short duration impulse noise whose energy is distributed through a whole range of frequencies appearing in the form of strips in the spectrogram. To suppress the keystrokes efficiently, it is

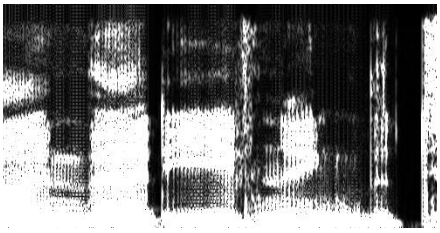

Fig 1: Clean speech spectrogram

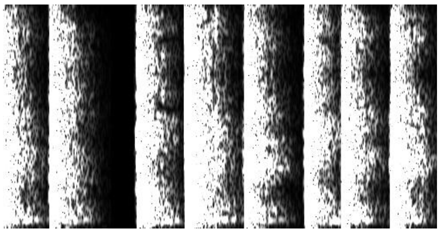

Fig 2: Speech-free noise spectrogram

better to detect keystrokes as much as possible even when both speech and keystroke noise are present in the same frame. In this stage, two approaches have been proposed, namely: (a) STFT thresholding based approach (b) Correlation based approach. Spectrograms of one second of clean speech, typical speech-free noise and noisy speech are shown in Fig. 1, Fig. 2 and Fig. 3, respectively.

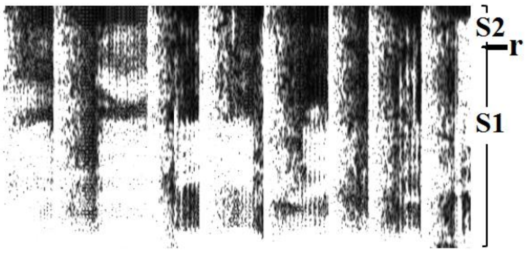

Fig 3: Noisy speech spectrogram

5.1 Detection and replacement by T-F thresholding technique

It can be seen from the Fig. 1, Fig. 2 and Fig. 3, that the energy of the noise frames are more widely distributed than the speech energy across frequency axis, which is the key motivation for developing the T-F thresholding technique. In this technique, the whole spectrogram is horizontally divided into two parts. The lower portion 'S1' is greater than the upper portion 'S2', as shown in Fig. 3. The spectrogram dividing value is represented by r. Then the norm of each spectral vector is computed. The norm of a vector x can be represented as

$$\|x\|_2 = \sqrt{\sum_{i=1}^{N} |x_i|^2} \qquad (10)$$

where $|.|$ denotes the norm of a vector and N is the total number of samples in each frame. For each spectral vector, the norm of S1 is divided by the norm of S2 and is represented by G. Then a threshold is set and G is compared to the threshold value as follows:

$$G = \frac{norm(S1)}{norm(S2)} \leqslant T \qquad (11)$$

where T denotes the threshold, and can vary if the spectrogram division value is changed. If for a specific frame, the value of G is less than T, then the frame is detected as keystroke corrupted frame, otherwise, it is detected as clean speech frame.

By this way, the corrupted frames in STFT domain are detected. Every detected frame is then replaced by the corresponding estimated frames from the estimated speech signal S_e obtained from stage-I and kept the phase same. The speech samples, not corrupted by the Keystrokes remain unchanged. Thus, there is a significant improvement in the quality and intelligibility of the speech signal.

Table 1. Keystrokes suppression performance comparison among various algorithms

Input SNR	Spectral subtraction		Enhanced-OMLSA		SNMF-CR		SNMF-TT	
	PESQ	Output SNR	PESQ	Output SNR	PESQ	Output SNR	PESQ	Output SNR
-10 dB	0.597	1.889	0.882	1.821	1.174	11.744	1.194	15.167
-5 dB	0.738	2.865	1.072	2.776	1.330	13.148	1.365	16.726
0 dB	0.952	3.849	1.342	3.646	1.499	14.364	1.559	18.651
5 dB	1.242	4.816	1.681	4.616	1.767	16.077	1.7945	20.908
10 dB	1.563	5.570	1.974	5.429	1.973	17.354	2.014	22.805

5.2 Keystroke suppression by correlation

To have minimum distortion and to replace only the noisy frames, the vector by vector correlation between the noisy signal X and the estimated signal S_e obtained from stage-I, is computed in T-F domain using magnitude spectrogram.

$$corr(\boldsymbol{X}, \boldsymbol{S_e}) = \frac{\sum_{i=1}^{N}(X_i - \overline{\boldsymbol{X}})(S_{e_i} - \overline{\boldsymbol{S_e}})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{\boldsymbol{X}})^2}\sqrt{\sum_{i=1}^{N}(S_{e_i} - \overline{\boldsymbol{S_e}})^2}} \qquad (12)$$

and $\overline{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N} X_i$

where N is the total number of samples in each frame, $\boldsymbol{X}$ is the noisy speech vector and $\boldsymbol{S_e}$ is the estimated speech vector in the STFT domain, obtained from stage-1. Based on a low correlation coefficient value between the absolute value of the estimated speech frame and the absolute value of noisy frame, corrupted frames in the noisy signal X are replaced by the corresponding estimated frames from S_e, while keeping the phase unchanged i.e. if the value of the correlation coefficient is less than ε, where ε denotes the threshold, then the frame is replaced from S_e to X, otherwise there is no change in X.

The final estimated speech in time domain is obtained by taking inverse STFT of the resultant estimated speech.

6. EXPERIMENTAL RESULTS

In this section, the performance of the proposed algorithms is evaluated and compared. Two objective speech quality measures are used to measure the performance of the tested algorithms, i.e., segmental SNR and perceptual evaluation of speech quality (PESQ).

In simulations, the speech signal data is downloaded from http://www.speech.cs.cmu.edu/cmu_arctic/ having 16 KHz sampling rate. The noise data is obtained from the websites https://www.pond5.com/, http://soundbible.com/, http://www.soundjay.com/, https://www.freesound.org/, and are re-sampled to 16 KHz. The STFT transform, with frame length of 512 and Hanning analysis window is implemented. We choose a sparsity weight "λ" of 1.5 on H and 100 factors for the SNMF. The spectrogram dividing value "r" is set to 252, and the value of "T" has been chosen to be 200, experimentally. The value of " ε" is kept at a fixed value of 0.76 and the value of "N" is 257. The experiment group consists of 10 male and 10 female speeches and are evaluated each under input SNR of -10dB, -5dB, 0dB, 5dB and 10dB. In every experiment, 6-seconds long mixture segment is used for testing. For the training, 50-seconds long speech and noise segments are used. For every experiment, absolutely different testing and training data is used, i.e., none of the training data is used in testing and vice versa. Moreover, the results of 20 speakers are averaged under one input SNR.

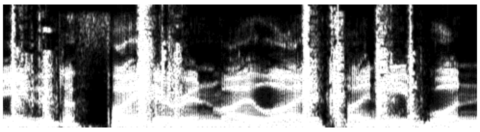

Fig. 4(a) Noisy signal spectrogram

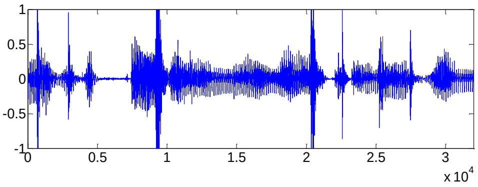

Fig. 4(b) Noisy signal in time domain

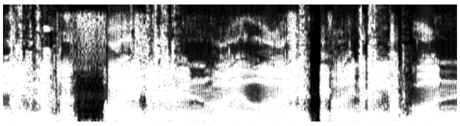

Fig. 5(a) Spectrogram of denoised signal using SNMF-CR

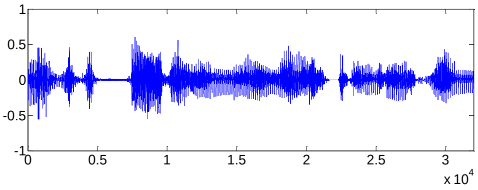

Fig. 5 (b) Denoised signal using SNMF-CR in time domain

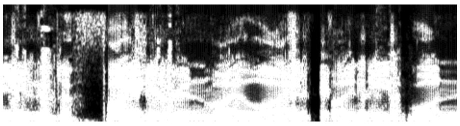

Fig. 6(a) Spectrogram of denoised signal using SNMF-TT

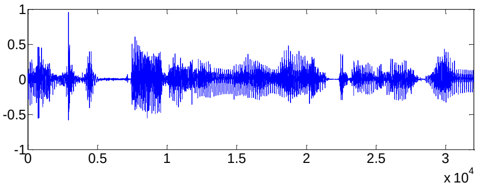

Fig. 6(b) Denoised signal using SNMF-TT in time domain

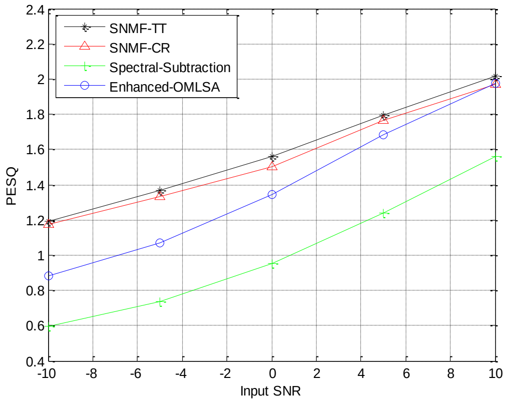

Fig. 7 PESQ value vs. input SNR for Keystroke suppression

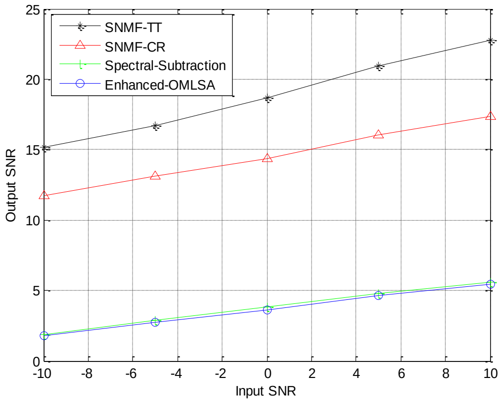

Fig. 8 Output SNR vs. input SNR for Keystroke suppression

The performance of proposed algorithms is compared with the enhanced-OMLSA [9] and spectral subtraction algorithm [14]. The experimental results in T-F domain as well as in time domain are shown in Fig. 4(a), Fig. 4(b), Fig. 5(a), Fig. 5(b), Fig. 6(a) and Fig. 6(b) with the signal length of two seconds. The comparison results are shown in table 1, Fig.7 and Fig.8 in terms of PESQ and output SNR values with the frame length of 512. From the results, it can be seen that the proposed algorithms outperform other algorithms and works best for the keystroke transient suppression without substantially degrading the audible quality of the speech signal.

In some scenarios, and also it can be seen from the output SNR values, that SNMF-TT has shown best performance of the compared algorithms. But the key motivation for presenting SNMF-CR is the sufficiency of single threshold selection which makes it more practical than SNMF-TT.

7. CONCLUSION AND FUTURE WORK

In this paper, two new two-stage methods for the suppression of keystrokes in speech signals are proposed and demonstrated good performance. The first stage is the same in both methods i.e. estimating the speech signal using supervised SNMF. In the thresholding technique, the second stage selects a threshold for replacing the corrupted frames while in the correlation based technique, the second stage replaces the corrupted frames on the basis of correlation between the noisy speech and the estimated speech. In both of the methods, the first stage is supervised, while the second is unsupervised. In future, SNMF can be enhanced for this particular task and a new and improved algorithm will be developed for the detection of keystrokes.

8. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61671418).

9. REFERENCES

[1] Tong, R. Zhou, Y. Zhang, L. Bao, G. and Ye, Z. A Robust Time-frequency Decomposition Model for Suppression of Mixed Gaussian-impulse Noise in Audio Signals. IEEE Transactions on Audio, Speech and Language Processing, Vol.23, No.1, Pages.69-79, Jan. 2015.

[2] Sigg, C.D. Dikk, T. Buhmann, J. M. Speech enhancement using generative dictionary learning. Audio, Speech, and Language Processing, IEEE Transactions on (Volume: 20, Issue: 6).

[3] Benesty, J. Chen, J. Huang, Y. Cohen, I. Noise Reduction in Speech Processing. Series: Springer Topics in Signal Processing, Vol. 2, 2009.

[4] Subramanya, A. Seltzer, M. L. and Acero, A. Automatic Removal of Typed Keystrokes from Speech Signals. IEEE signal processing letters, vol. 14, no. 5, may 2007.

[5] Mavaddaty, S. Ahadi, S. M. Seyedin, S. A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation. Speech Communication 76 (2016) 42–60.

[6] Talmon, R. Cohen, I. and Gannot, S. Transient noise reduction using nonlocal diffusion filters, IEEE Trans. Audio, Speech and Lang. Process., vol. 19, Issue 6, pp. 1584–1599, Aug. 2011.

[7] Talmon, R. Cohen, I. and Gannot, S. Clustering and suppression of transient noise in speech signals using diffusion maps, Proc. 36th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP11), Prague, Czech Republic, May 22-28, 2011.

[8] Talmon, R. Cohen, I. and Gannot, S. Single-channel transient interference suppression with diffusion maps. IEEE trans. on audio, speech, and lang. Process., vol. 21, no. 1, January 2013.

[9] Hirszhorn, A. Dov, D. Talmon, R. and Cohen, I. Transient interference suppression in speech signals based on the OM-LSA algorithm. Int. Workshop on Acoustic Signal Enhancement 2012, 4-6 September 2012, Aachen.

[10] (Arden) Huang, Y. Benesty, J. Audio Signal Processing for Next-Generation Multimedia Communication Systems. Bell Laboratories, Lucent Technologies, kluwer academic publishers, 2004

[11] Wilson, K. W. Raj, B. Smaragdis, P. Divakaran, A. Speech denoising using nonnegative matrix factorization with priors. 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing.

[12] Sohrab, F. Erdogan, H. Recognize and separate approach for speech denoising using nonnegative matrix factorization. 23rd European Signal Processing Conf. (EUSIPCO), Aug. 31 2015-Sept. 4 2015.

[13] Schafer, R. W. Rabiner, L. R. Digital Representations of Speech Signals. Proceedings of the ieee, vol. 63, no. 4, april 1975.

[14] Boll, S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech, and Signal Process., vol. 27, no. 2, pp. 113 – 120, Apr. 1979.

[15] Smaragdis, P. From learning music to learning to separate. In Forum Acusticum, Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge MA 02139, USA 2005.

[16] Nandhini, S. Shenbagavalli, A. Voiced/Unvoiced Detection using Short Term Processing. Int. conf. on Innovations in Information, Embedded and Communication Systems (ICIIECS-2014)

[17] Mohammadiha, N. Smaragdis, P. and Leijon, A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 10, pp.2140–2151, Oct. 2013.

[18] Luo, Y. Bao, G. Xu, Y. Ye, Z. Supervised Monaural Speech Enhancement Using Complementary Joint Sparse Representations. IEEE signal processing