A New Speech Enhancement Method Based on Nonnegative Low-rank and Sparse Decomposition

Jiayi Sun¹, Chengli Sun²* and Yi Hong²

¹Nanchang No.2 Middle School, Nanchang, Jianxi Province, 330031, China
²School of Information, Nanchang Hangkong University, Nanchang, Jianxi Province, 330063, China
*Corresponding author’s e-mail: sunchengli@nchu.edu.cn

Abstract. Enhancement of speech degraded by strong noises is a highly difficult task. In this paper, a nonnegative low-rank and sparse matrix decomposition (NLSMD) based speech enhancement method is given to address this problem. The proposed method is motivated with assumptions that in time-frequency (T-F) domain, since power spectrum of many types of noise with different frame are often correlative, noise can be assumed with a low-rank structure, while speeches are often sparse in T-F units. Based on these assumptions, we formulate the speech enhancement as a NLSMD problem, and design an objective function to recover speech component. Compared with traditional methods, the NLSMD-based method does not require a speech activity detector for noise density estimation. Experimental results show the proposed method can achieve better performance over many traditional methods in strong noise conditions, in terms of yielding less residual noise and lower speech distortion.

1. Introduction
Speech enhancement (SE) is useful in many applications, such as robot dialogue, speech-to-text, hearing-aid equipment. The goal of SE is to ameliorate the auditory quality by reducing noise [1]. Over past four decades, many research works has conducted on this area. SE methods can be operated in time domain, frequency domain, and T-F domain. A representative time domain SE method is signal subspace method [2]. Typical Frequency SE domain methods include spectral subtraction [3], minimum mean square error (MMSE [4]) and Wiener noise filter[5], and time-frequency domain SE methods involve the application of T-F matrix decomposition [6, 7] or Wavelet toolkits [8].

In traditional SE approaches, speech activity detection (SAD) is often employed to estimate the noise signal density using silence clips. However, current SAD technology is imperfect in highly noisy occasion, which may severely degrade SE performance [9]. In this work we propose a new SE approach based on NLSMD. The main idea behind NLSMD-based method is motivated by the robust principal component analysis (RPCA) theory [10] and its derivants. RPCA states that if a corrupted observation matrix is the summation of a low-rank component and a sparse component, these two components can be recovered using low-rank and sparse matrix decomposition (LSMD). Since noise power spectrum within different frame are often correlative, noise can be assumed of a low-rank structure, while speeches are often sparse in T-F units. Thus the LSMD can be used for SE in T-F domain[6, 7, 11-13].

In this work, we propose a NLSMD algorithm for speech/noise separation by imposing rank and sparsity constraints. The NLSMD-based method has two good properties: (1) NLSMD-based method
is a non-parametric method with no specific assumptions of speech and noise distribution. It needs not a training stage like deep learning approaches. (2) It can recover speech and noise component simultaneously and bypass the SAD process. This is useful in the scene where noise is strong and hardly be estimated.

This paper is organized as follows. Section 2 briefly reviews the theoretical background and related works of SE. In Section 3, we introduce NLSMD theory and its solving algorithm, and then we present NLSMD-based SE system. Experimental results are demonstrated in Section 4. Finally, conclusions are given in section 5.

2. Related work

2.1. Nonnegative matrix factorization (NMF)

Non-negative matrix factorization (NMF) [14] is a matrix decomposition algorithm which focuses on analysis of matric representations which are nonnegative. Given a observation matrix $X \in \mathbb{R}^{N \times K}$ and a positive integer R. The goal of NMF is find two non-negative matrices $U \in \mathbb{R}^{N \times R}$ and $V \in \mathbb{R}^{R \times K}$, whose product can well represent matrix $X$.

$$X \approx UV^T$$

To estimate the factorization matrices, Itakura-Saito (IS) divergence is often used for source separation and speech denoising.

$$D_{IS}(X | UV^T) = \sum_{ij} \left( \frac{X_{ij}}{X^*_{ij}} \log \frac{X_{ij}}{X^*_{ij}} - 1 \right)$$

NMF-IS can be solved by alternating updates of $U$ and $V$ as in [15]:

$$U \leftarrow U \odot \frac{X}{UV^T} \odot V, \quad V \leftarrow V \odot \frac{U^T \odot X}{U^T \odot V^T}$$

Where the symbol $\odot$ is an element-wise multiplication, all divisions and $(\cdot)^2$ are element-wise operations. Since the IS divergences are more sensitive to small-energy observations, leading these two divergences are more suitable for spectrogram analysis.

2.2. Robust principal component analysis

Principal component analysis (PCA) is a classical signal analysis approach used for exploring underlying data structure. It assumes that the given high-dimensional observations lie in a low-dimensional linear subspace. Suppose the given data are arranged as the columns of a large matrix $M \in \mathbb{R}^{N \times K}$, PCA aims to find a low-rank matrix $L$ with the following constrained optimization:

$$\min_{L} \frac{1}{2} \| M - L \|_F^2, \text{ s.t. rank}(L) \leq r$$

Where $r \ll \min(N,K)$ is the low-rank constraint for the $L$, and $\| \cdot \|_F$ is the Frobenius norm. Equation (4) can be solved using singular value decomposition. PCA performs well in the present of small Gaussian noise. However, it will drop down under strong noise or outliers, even if that corruption affects only very few of the observations. Wright [10, 16] propose a new theory called RPCA to overcome the limitations of PCA. The theoretical idea of RPCA is to decompose observation matrix $M$ into $M=L+S$, where $S \in \mathbb{R}^{N \times K}$ is a sparse matrix with less non-zero coefficients. The unknown matrices $L$ and $S$ can be obtained by minimizing the following formula

$$\min \| L \|_F + \lambda \| S \|_1, \text{ s.t. } M = L + S,$$
where \( \| \| \) represents the matrix nuclear norm which is defined as the sum of all singular values of \( L \) [17]. \( \| \| \) represents the 11-norm which is defined as summation of absolute values of matrix elements of \( S \). When the observations are full of small noise, the equality in (5) no longer holds, Zhou and Tao proposes GoDec algorithm [18] to address this problem, which can be formulated as

\[
\min_{L,S} \| M - L - S \|_F^2,
\]

s.t. \( \text{rank}(L) \leq r \), \( \text{card}(S) \leq h \),

(6)

Where \( \text{card}(S) \) is the cardinality of \( S \), i.e. the number of nonzero coefficients of \( S \). Compared with RPCA in which the rank and cardinality parameters are previously unspecified, GoDec can be regard as a conditional LSMD method in which the parameters \( r \) and \( h \) are user defined beforehand to restrict matrix decomposition.

3. NLSMD based speech enhancement method

3.1 Problem formulation

Let \( m(t) = s(t) + l(t) \) be the corrupted speech consisting of clean signal \( s(t) \) and noise \( l(t) \). After short-time Fourier transform (STFT), we get its complex spectrum form as

\[
M(n,k) = S(n,k) + L(n,k)
\]

(7)

Where \( k \in \{0,\ldots,K-1\} \) and \( n \in \{0,\ldots,N-1\} \) is frequency index and time-frame index respectively. By stacking spectral magnitude vectors as columns frame-by-frame, we obtain the T-F representation of (7) being

\[
M = L + S
\]

(8)

Problem 1. (NLSMD): Given noisy matrix \( M = L + S + E \), where \( L \) is a low-rank matrix with rank constraint \( r \), \( S \) is a sparse matrix with sparse constraint \( h \), \( E \) is a residual noise matrix. \( L \) and \( S \) correspond noise matrix and enhanced speech matrix respectively. The matrices \( L \) and \( S \) can be recovered by the minimizing the following reconstruction error:

\[
\min_{L,S} \| M - L - S \|_F^2,
\]

s.t. \( \text{rank}(L) \leq r \), \( \| S \|_0 \leq h \), and \( L \geq 0 \).

(9)

Since noise spectrogram is often of low-rank structure, we can assign a small integer to \( r \). On the contrary, speech spectrogram is usually sparse in T-F bins. Thus, by means of NLSMD, the noise and speech components can be separated as far as possible. Notes the NLSMD is different from RPCA, where the constraints for \( L \) and \( S \) are not specified beforehand, and it is also different from GoDec, where \( L \) and \( S \) are not imposed nonnegative constraints.

3.2 Solving of NLSMD

In order to optimize Eq. (9), we employ an iterative optimization strategy. At one iteration, we update the unknown parameters \( L \) and frozen \( S \), then at another iteration we update the unknown parameters \( S \) and frozen \( L \). We borrow the framework of GoDec and solve the following two subproblems of NLSMD alternatively until convergence:

\[
\begin{align*}
L_i &= \arg \min_{\text{rank}(L) \leq r, L \geq 0} \| M - L_{i-1} - S \|_F^2 \quad (a) \\
S_i &= \arg \min_{\| F \|_F \leq S, S \geq 0} \| M - L_i - S \|_F^2 \quad (b)
\end{align*}
\]

(10)

3
The subproblem (10-a) is a typical fixed-rank approximation problem, which aims to find a nonnegative low-rank matrix $L_i$ with rank constraint. NMF can be used to address this problem [10].

\[(U_i, V_i) = \text{NMF}(M - S_{i-1})\]  

\[L_i = U_i V_i^T\]  

In (10-b), $S_i$ can be estimated by following formula.

\[f(S) = \|M - L - S\|_F^2 + \lambda \|S\|_1\]  

\[= \sum_{n,k} (M_{nk} - L_{nk} - S_{nk})^2 + \lambda \sum_{n,k} |S_{nk}|^2\]

Since, $S_{nk} \geq 0$, $f(S_{nk}) = \sum_{n,k} (M_{nk} - L_{nk} - S_{nk})^2 + \lambda \sum_{n,k} S_{nk}^2$

\[\frac{\partial f(S_{nk})}{\partial S_{nk}} = -\lambda + (M_{nk} - L_{nk} - S_{nk}) = 0\]

\[S_{nk} = M_{nk} - L_{nk} - \lambda\]

In summary, we have following alternative least projection algorithm for NLSMD.

**Algorithm 1. Optimization algorithm for NLSMD**

Given $r$, $h$, $I$;
Initialize $M_0 = M$, $S_0 = [0]_{I \times r}$;

1. for $i=1,2,\ldots, I$
2. $(U_i, V_i) = \text{NMF}(M_{i-1})$;
3. $L_i = U_i V_i^T$;
4. $\hat{X}_i = M_{i-1} - L_i + S_{i-1}$;
5. $S_i = P_{\Omega}(\hat{X}_i)_{\Omega}$, $\Omega$: $(M - L_i)_{n,k \in \Omega} \geq \lambda$,
   
   $M_i = L_i + P_{\Omega}(\hat{X}_i)_{\Omega}$;
6. $i \leftarrow i + 1$;
end for

outputs $L$ and $S$

In above algorithm, lines 2 and 3 are used for updating low-rank matrix $L_i$, lines 4 and 5 are used for updating sparse matrix $S_i$. At each iteration, NLSMD find the subset of $X_i$ satisfying the given condition to build $S_i$, $M_i$ is then updated by the summation of $L_i$ and $P_{\Omega}(X_i)_{\Omega}$, where

\[P_{\Omega}(X_i)_{\Omega} = \begin{cases} X_{\Omega i}, & X_{\Omega i} \in \Omega \\ 0, & \text{otherwise}. \end{cases}\]  

\[3.3 \text{ NLSMD base speech enhancement system}\]

The noisy signal is weighted using a Hamming window and transform into frequency domain by STFT. Then the proposed NLSMD is conducted to decompose the noisy speech matrix and obtain matrix $S$. Similar to many analysis-modification-synthesis (AMS) based speech enhancement approaches [19], we combined the entries of $S$ with the phase of noisy signal to estimate the Fourier transform of $s(t)$,

\[\hat{S}(n,k) = |S(n,k)|e^{j\phi(n,k)}\]  

Finally, the enhanced speech signal $\hat{s}(t)$ is constructed by taking inverse STFT of $\hat{S}(n,k)$ and followed by an overlap-add procedure.
4. Experiments

NOIZEUS database was used to analyze the performance of NLSMD-based method. 30 Sentences (sp1~sp30) taken from NOIZEUS database were corrupted by white Gaussian white, hfchannel, street, and train noise at SNR of 0, 5, 10dB. All the signal were sampled at 8 kHz. Hamming window is used for segmenting signal into frames with window length of 300 and 40% shift percentage. Then we transformed the signal into frequency spectrum by a 1024 points STFT.

Two measures were used to assess performance, including segmental-SNR measure and PESQ (Perceptual Evaluation of Speech Quality) measure for evaluations. We fix r=1 as low-rank constraint in NLSMD method. We compare the proposed method with traditional SE methods, including SS (spectral subtraction) [3], Subspace method [2], WienerTPS [5], MMSENPS [20], MSSMAP [21], and CLSMD [6]. Methods of SS, WienerTPS and MMSENPS used a energy based SAD algorithm.

Table 1. Performance comparisons in terms of SegSNR

| Noise  | Method   | 0dB       | 5dB       | 10dB      |
|--------|----------|-----------|-----------|-----------|
| white  | NLSMD    | 2.253     | 3.798     | 4.971     |
|        | CLSMD    | 2.147     | 3.686     | 4.856     |
|        | MSSMAP   | 0.364     | 2.625     | 5.145     |
|        | SS       | -0.999    | 0.596     | 3.135     |
|        | Subspace | 0.466     | 3.003     |           |
|        | MMSENPS  | -0.761    | 0.466     | 1.909     |
|        | WienerTPS| -0.464    | 0.787     | 2.119     |
| hfchannel | NLSMD   | 0.751     | 2.660     | 4.286     |
|         | CLSMD    | 0.712     | 2.602     | 4.219     |
|         | MSSMAP   | 0.057     | 2.306     | 4.797     |
|         | SS       | -1.002    | 0.948     | 3.316     |
|         | Subspace | -1.486    | 1.116     | 3.784     |
|         | MMSENPS  | -0.879    | 0.366     | 1.850     |
|         | WienerTPS| -0.431    | 0.834     | 2.258     |
| street  | NLSMD    | -1.379    | 0.926     | 2.692     |
|         | CLSMD    | -1.389    | 0.918     | 2.550     |
|         | MSSMAP   | -1.707    | 0.535     | 3.512     |
|         | SS       | -0.987    | 0.711     | 3.256     |
|         | Subspace | -2.625    | -0.218    | 2.374     |
|         | MMSENPS  | -2.072    | -0.799    | 0.890     |
|         | WienerTPS| -0.835    | 0.218     | 1.798     |
| train   | NLSMD    | -0.173    | 1.649     | 3.419     |
|         | CLSMD    | -0.179    | 1.647     | 3.359     |
|         | MSSMAP   | -1.576    | 0.932     | 3.588     |
|         | SS       | -1.446    | 0.861     | 3.120     |
|         | Subspace | -2.453    | 0.162     | 2.731     |
|         | MMSENPS  | -2.488    | -0.691    | 0.868     |
|         | WienerTPS| -1.058    | 0.418     | 1.906     |

Table 2. Performance comparisons in terms of PESQ

| Noise  | Method   | 0dB       | 5dB       | 10dB      |
|--------|----------|-----------|-----------|-----------|
| white  | NLSMD    | 2.028     | 2.389     | 2.594     |
|        | CLSMD    | 2.010     | 2.385     | 2.588     |
|        | MSSMAP   | 1.986     | 2.319     | 2.655     |
|        | SS       | 1.633     | 1.850     | 2.330     |
|        | Subspace | 1.720     | 2.194     | 2.602     |
|        | MMSENPS  | 1.875     | 2.165     | 2.492     |
Table 1 shows the segSNR values, higher segSNR indicates better noise suppression performance. We can see the NLSMD have a good performance in most of conditions, especially in the low-SNR (0db and 5db). The PESQ performances were show in Table 2. PESQ was examines the overall quality of the enhanced speech. It is observed that the NLSMD achieved good performance than other compared methods in noise types of white, hfchannel, and train. These results show that NLSMD has good suppression capability along with good hearing quality.

5. Conclusions
In this work, we presented a new speech enhancement approach by transforming speech enhancement problem into a NLSMD problem. An efficient optimization algorithm was developed to estimate the nonnegative low-rank and sparse matrix by imposing rank and sparse constraints. Experiments show the proposed method is effective under low SNR conditions. The proposed method can simultaneously estimate speech and noise. It does not require a speech activity detector for noise density estimation. In future, we will further improve NLSMD based SE method by incorporating more prior knowledge of signal, such as speech continuity and auditory perception property.

Acknowledgments
This work was supported partly by National Natural Science Foundation of China (No. 61861033), Natural Science Foundation of Jiangxi (No. 20202ACBL202007) and Natural Science Foundation of Shandong (No. ZR2020MF020).

References
[1] Loizou, P.C., Speech Enhancement: Theory and Practice. (2007) New York: Taylor & Francis.
[2] Hu, Y. and P.C. Loizou, (2003) A generalized subspace approach for enhancing speech corrupted by colored Noise. IEEE Transactions on Audio, Speech and Language Processing. 11(4): 334-342.
[3] Boll, S.F., (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process, 27(2): 113-120.
[4] Ephraim, Y. and D. Malah, (1985) Speech enhancement using a minimum mean square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. 33(2): 443–445.

[5] Scalart, P. and J. Vieira-Filho, (1996) Speech enhancement based on a priori signal to noise estimation, in Proc. 21st IEEE Int. Conf. Acoust. SpeechSignal Processing. Atlanta, GA. pp. 629-632.

[6] Sun, C., Q. Zhu, and M. Wan, (2014) A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition. Speech Communication, 60: 44-55.

[7] Bando, Y., et al., (2018) Speech enhancement based on bayesian low-Rank and sparse decomposition of multichannel magnitude spectrograms. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(2): 215-230.

[8] Bahoura, M. and J. Rouat, (2006) Wavelet speech enhancement based on time-scale adaptation. Speech Communication, 2006. 48(12): 1620-1637.

[9] Manohar, K. and P. Rao, (2006) Speech enhancement in nonstationary noise environments using noise properties. Speech Communication, 48: 96-109.

[10] Candes, E.J., et al., (2011) Robust Principal Component Analysis? Journal of the ACM, 58(3): 1-37.

[11] Cohen, J.E. and N. Gillis, (2019) Nonnegative low-Rank sparse component analysis, in ICASSP. pp. 8226-8230.

[12] Ji, Y., W. Zhu, and B. Champagne, (2019) Speech enhancement based on dictionary learning and low-rank matrix decomposition. IEEE Access, 7: 4936-4947.

[13] Sun, C. and C. Yuan, (2019) Speech enhancement based on constrained low-rank sparse matrix decomposition integrated with temporal continuity regularization. Archive of Acoustics, 44(4): 681-692.

[14] Lee, D.D. and H.S. Seung, (1999) Learning the parts of objects by nonnegative matrix factorization. Nature, 401(6755): 788-791.

[15] Févotte, C., N. Bertin, and J. Durrieu, (2009) Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. Neural Computing, 21(3): 793-830.

[16] Wright, J., Y. Peng, and Y. Ma, (2009) Robust Principal Component Analysis: Exact Recovery of Corrupted Low-rank Matrices by Convex Optimization. In NIPS 2009.

[17] Candes, E.J. and T. Terence, (2010) The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory, 56(5): 2053-2080.

[18] Zhou, T. and D. Tao. (2011) GoDec: Randomized Low-rank & Sparse Matrix Decomposition in Noisy Case. in Proceedings of the 28 th International Conference on Machine Learning. Bellevue, WA, USA.

[19] Paliwal, K., K. Wojcicki, and B. (2010) Schwerin, Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Communication, 52: 450-475.

[20] Cohen, L., (2004) Speech enhancement using a noncausal a priori SNR estimator. IEEE Signal Processing Letters, 11(9): 725-728.

[21] Lu, Y. and P.C. Loizou, (2011) Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty IEEE Transactions on Audio, Speech and Language Processing. 19(5): 1123-1137.