# Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint

Hien-Thanh Duong, Quoc-Cuong Nguyen, Cong-Phuong Nguyen, Thanh-Huan Tran, Ngoc Q. K. Duong

# Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint

Hien-Thanh T. Duong[1,2], Quoc-Cuong Nguyen[1,3], Cong-Phuong Nguyen[1,3], Thanh-Huan Tran[4], Ngoc Q. K. Duong[5]

[1]International Research Institute MICA, Hanoi University of Science and Technology, Vietnam
[2]Faculty of Information Technology, Hanoi University of Mining and Geology, Vietnam
[3]Department of Instrumentation and Industrial Informatic, Hanoi University of Science and Technology, Vietnam
[4]Faculty of Information Technology, Hanoi University of Industry, Vietnam
[5]Imaging Science Lab, Technicolor, France

Email: duongthihienthanh@humg.edu.vn, cuong.nguyenquoc@hust.edu.vn, phuong.nguyencong@hust.edu.vn, huantt-fit@haui.edu.vn, quang-khanh-ngoc.duong@technicolor.com

## ABSTRACT

This paper addresses a challenging single-channel speech enhancement problem in real-world environment where speech signal is corrupted by high level background noise. While most state-of-the-art algorithms tries to estimate noise spectral power and filter it from the observed one to obtain enhanced speech, the paper discloses another approach inspired from audio source separation technique. In the considered method, generic spectral characteristics of speech and noise are first learned from various training signals by non-negative matrix factorization (NMF). They are then used to guide the similar factorization of the observed power spectrogram into speech part and noise part. Additionally, we propose to combine two existing group sparsity-inducing penalties in the optimization process and adapt the corresponding algorithm for parameter estimation based on multiplicative update (MU) rule. Experiment results over different settings confirm the effectiveness of the proposed approach.

## Keywords

Speech enhancement, audio source separation, nonnegative matrix factorization, multiplicative update, spectral model, group sparsity.

## 1. INTRODUCTION

In real-world recordings, desired speech signal is usually mixed with environmental noise as well as other unwanted interferences such as background music. Thus speech enhancement is needed to cancel the background noise and make speech signal more intelligible [1, 2]. It offers a wide range of applications in phone communication, sound post-production, and robotics. As a more insight example, in order to improve the robustness of the automatic speech recognition (ASR) systems in noisy environments, a speech enhancement algorithm is often incorporated before ASR to preprocess noisy signals [3, 4]. In order to enhance such corrupted speech, the background noise spectral power usually needs to be estimated first in order to derive a filtering function (*e.g.,* either in the form of spectral subtraction or Wiener filtering) for the desired noise reduction techniques [5, 6].

This paper considers another denoising approach motivated from audio source separation technique [7], where speech and noise are considered as two distinct sources that need to be separated from the noisy observation. More precisely, a recent paper by Sun and Mysore [8] proposed to learn an universal speech model by non-negative matrix factorization (NMF) [9] from different speakers and applied it for the single-channel speech separation. The term *universal model* is also in analogy to the universal background models for speaker verification addressed in [10]. This idea of using a generic spectral model was then exploited in the context of *on-the-fly* source separation [11, 12] where any kind of audio sources can be separated with the guidance from its examples collected from a search engine. Motivated from those above-mentioned works, we propose in this paper to learn two generic spectral models for speech and background noise independently in advance. Then in the enhancement phase, we exploit these pre-learned models together with a novel sparsity constraint to guide the factorization of the mixture spectrogram into speech part and noise part. Note that our proposed approach differs from the prior works in several aspects as follows. Firstly, compared to [8] where only the universal speech model was pre-learned and noise model was adapted during the separation process, we consider to learn the universal noise model also since noisy examples can be easily collected in advance and it would potentially improve the separation quality. Secondly, compared to [8] and [11, 12] where either block sparsity-inducing penalty or component-sparsity-inducing penalty was used, we propose in this paper a combination of these two penalties which would offer better estimating the parameters in the model fitting.

The rest of the paper is organized as follows. In Section 2



*SoICT 2015, December 03-04, 2015, Hue City, Viet Nam*



DOI: http://dx.doi.org/10.1145/2833258.2833276

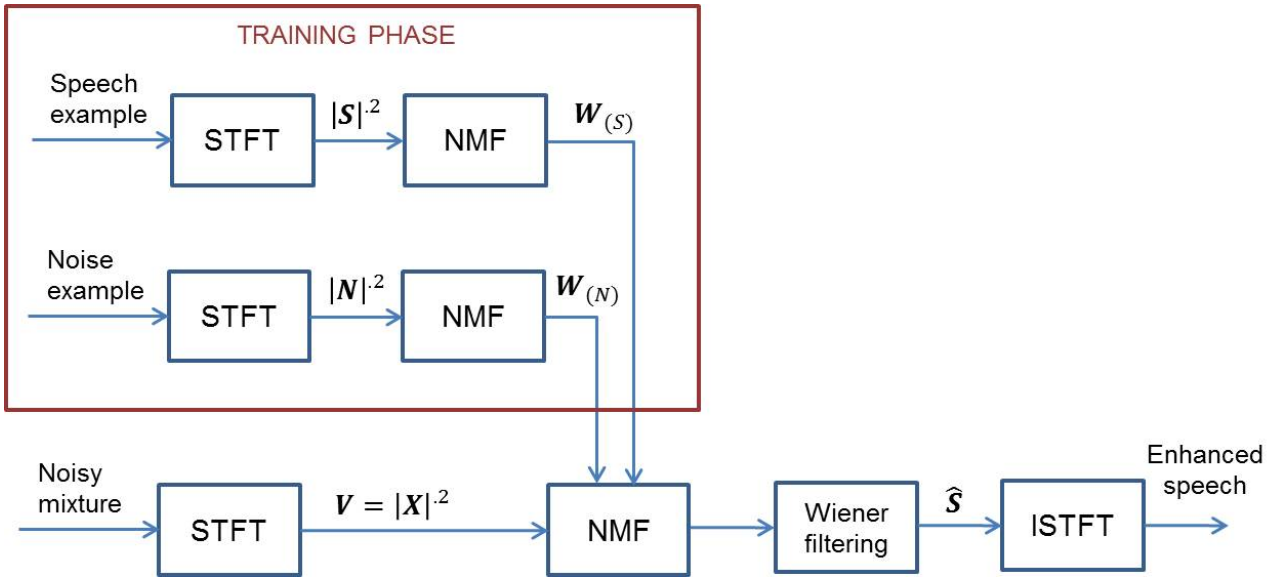


**Figure 1: General workflow of the baseline supervised speech and noise separation approach.**

we summarize the background of the supervised audio signal separation approach based on NMF model. We then present the proposed non-supervised speech separation approach in Section 3 where the formulation of the generic spectral model and the sparsity-inducing penalties are both described. Section 4 is devoted to experiment in order to validate the effectiveness of the proposed approach. Finally we conclude in Section 5.

## 2. BACKGROUND OF NMF-BASED SIGNAL SEPARATION

We summarize in this section a standard supervised signal separation method based on NMF as one of the most popular model for audio signal [13, 8]. The general pipeline works in the frequency domain after the short-time Fourier transform (STFT) transform and consists in two phases as shown in Fig. 1: (1) learning NMF source spectral models from some training examples, and (2) decomposing the observed mixture with the guide of the pre-learned models.

Let us start by considering a single-channel signal separation problem with 2 sources (speech and noise). Let $\mathbf{X} \in \mathbb{C}^{F\times M}$, $\mathbf{S} \in \mathbb{C}^{F\times M}$, and $\mathbf{N} \in \mathbb{C}^{F\times M}$ be the complex-valued matrices of the short-time Fourier transform (STFT) coefficients of the observed mixture signal, the speech signal, and the noise signal, respectively, where $F$ is the number of frequency bins and $M$ the number of time frames. The mixing model is written as

$$\mathbf{X} = \mathbf{S} + \mathbf{N}. \tag{1}$$

Denoting $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ the power spectrogram of the mixture where $|\mathbf{X}|^{\cdot p}$ is the matrix with entries $[\mathbf{X}]_{il}^{p}$. NMF aims at decomposing the $F\times M$ non-negative matrix $\mathbf{V}$ into two non-negative matrices $\mathbf{W} \in \mathbb{R}^{F\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times M}$, respectively, where $K$ is usually chosen to be smaller than $F$ or $M$, such that the divergence between $\mathbf{V}$ and $\mathbf{WH}$ is minimized in some senses. Popularly for audio, this decomposition is done by minimizing the Itakura-Saito divergence, subject to the constraints $\mathbf{W}, \mathbf{H} \geq 0$, which offers scale invariant property, as [14]

$$\min_{\mathbf{H}\geq 0, \mathbf{W}\geq 0} D(\mathbf{V}\|\mathbf{WH}), \tag{2}$$

where $D(\mathbf{V}\|\hat{\mathbf{V}}) = \sum_{f=1}^{F}\sum_{m=1}^{M} d_{IS}(\mathbf{V}_{fm}\|\hat{\mathbf{V}}_{fm})$, with $\hat{\mathbf{V}} = \mathbf{WH}$, $f$ and $m$ denoting frequency bin index and time frame index, respectively, and

$$d_{IS}(x\|y) = \frac{x}{y} - \log(\frac{x}{y}) - 1. \tag{3}$$

The parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are usually initialized with random non-negative values and are iteratively updated via the well-known multiplicative update (MU) rules [14].

In the supervised setting, spectral model for speech and noise, denoted by $\mathbf{W}_{(S)}$ and $\mathbf{W}_{(N)}$, respectively, is first learned from the corresponding training examples by optimizing similar criterion as (2). Then spectral model for two sources $\mathbf{W}$ is obtained by

$$\mathbf{W} = [\mathbf{W}_{(S)}, \mathbf{W}_{(N)}]. \tag{4}$$

In the testing phase (speech enhancement process), this spectral model $\mathbf{W}$ is fixed, and the time activation matrix $\mathbf{H}$ is estimated via the MU rule. Note that $\mathbf{H}$ is also partitioned into blocks as

$$\mathbf{H} = [\mathbf{H}_{(S)}^{T}, \mathbf{H}_{(N)}^{T}]^{T}, \tag{5}$$

where $\mathbf{H}_{(S)}$ and $\mathbf{H}_{(N)}$ denotes a block characterizing the time activations for speech and noise, respectively, and $\mathbf{A}^{T}$ the transposition of matrix $\mathbf{A}$.

Once the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are obtained, the speech and noise STFT coefficients are computed by Wiener filtering as

$$\hat{\mathbf{S}} = \frac{\mathbf{W}_{(S)}\mathbf{H}_{(S)}}{\mathbf{WH}} \odot \mathbf{X}, \tag{6}$$

$$\hat{\mathbf{N}} = \frac{\mathbf{W}_{(N)}\mathbf{H}_{(N)}}{\mathbf{WH}} \odot \mathbf{X}, \tag{7}$$

where $\odot$ denotes the element-wise Hadamard product and the division is also element-wise. And finally, the time domain source estimates are obtained via the inverse STFT.

All this speech enhancement workflow (STFT + NMF + Wierner filtering + ISTFT) is depicted in Fig. 1.

# 3. PROPOSED SPEECH SEPARATION APPROACH

Given the signal processing workflow with the NMF model presented in Section 2, we will first introduce the formulation of the generic spectral model for both speech and noise in Section 3.1. We then describe the model fitting with two state-of-the-art sparsity-inducing penalties considered in [10, 11] in Section 3.2. Finally, we present the proposed mixed sparsity-inducing penalty and the derived parameter estimation in Section 3.3. *Note that the considered approach can be actually understood as non-supervised in the paper's context even though it involves some training. This is because the training phase only learns the generic models from different types of example signals.*

## 3.1 Generic spectral model formulation

Assuming that some examples of speech signal and noise are available (this is actually feasible since audio recordings of speech and noise exist in many databases). Let the spectrogram of $p$-th speech examples and $q$-th noise examples be denoted by $\mathbf{V}_S^p$ and $\mathbf{V}_N^q$, respectively. First, $\mathbf{V}_S^p$ and $\mathbf{V}_N^q$ are used to learn the NMF spectral model, denoted by $\mathbf{W}_{(S)}^p$ and $\mathbf{W}_{(N)}^q$, respectively, by optimizing the criterion (similar to (2)):

$$\min_{\mathbf{H}_{(S)}^p \geq 0, \mathbf{W}_{(S)}^p \geq 0} D(\mathbf{V}_S^p \| \mathbf{W}_{(S)}^p \mathbf{H}_{(S)}^p), \tag{8}$$

$$\min_{\mathbf{H}_{(N)}^q \geq 0, \mathbf{W}_{(N)}^q \geq 0} D(\mathbf{V}_N^q \| \mathbf{W}_{(N)}^q \mathbf{H}_{(N)}^q), \tag{9}$$

where $\mathbf{H}_{(S)}^p$ and $\mathbf{H}_{(N)}^q$ are time activation matrixes. Given $\mathbf{W}_{(S)}^p$ and $\mathbf{W}_{(N)}^q$ for all speech examples $p = 1, ..., P$ and noise examples $q = 1, ..., Q$, respectively, the generic spectral models for speech and noise are constructed as

$$\mathbf{W}_S = [\mathbf{W}_{(S)}^1, \ldots, \mathbf{W}_{(S)}^P] \tag{10}$$

$$\mathbf{W}_N = [\mathbf{W}_{(N)}^1, \ldots, \mathbf{W}_{(N)}^Q]. \tag{11}$$

In the practical implementation, we may need several speech examples for different male voices and female voices (*e.g.,* $P = 4$), and examples of different types of noise such as wind sound, cafeteria, waterfall, street noise, etc.,. (*e.g.,* $Q = 5$).

## 3.2 Model fitting with state-of-the-art sparsity-inducing penalties

The generic spectral models $\mathbf{W}_S$ and $\mathbf{W}_N$ constructed in (10) and (11), respectively, become large matrices when the number of examples increases, and they are actually redundant matrices since different examples may share similar spectral patterns. Thus in the model fitting for the mixture spectrogram, sparsity constraint is naturally needed so as to fit only a subset of the large matrix $\mathbf{W} = [\mathbf{W}_S, \mathbf{W}_N]$ to the targeted source in the mixture [15, 16, 17]. In other words, the mixture spectrogram $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ is decomposed by solving the following optimization problem

$$\min_{\mathbf{H} \geq 0} D(\mathbf{V} \| \mathbf{W}\mathbf{H}) + \lambda \Omega(\mathbf{H}) \tag{12}$$

where $\Omega(\mathbf{H})$ denotes a penalty function imposing sparsity on the activation matrix $\mathbf{H}$, and $\lambda$ is a trade-off parameter determining the contribution of the penalty. When $\lambda = 0$, $\mathbf{H}$ is not sparse and the entire generic model is used as illustrated in Fig. 2a. Recent work in audio source separation has considered two penalty functions as the following.

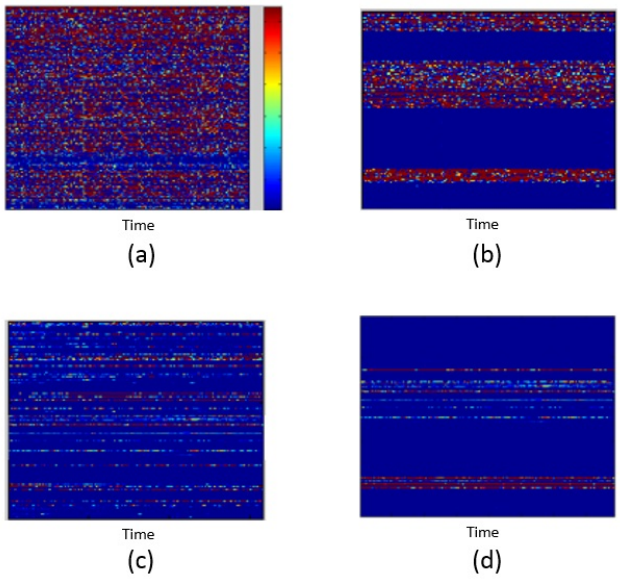


**Figure 2: Estimated activation matrix H: (a) without a sparsity constraint, (b) with a block sparsity-inducing penalty** (13)**, (c) with a component sparsity-inducing penalty** (14)**, and (d) with the proposed mixed sparsity-inducing penalty** (15)**.**

(i) **Block sparsity-inducing penalty**

In order to eliminate irrelevant training examples, *i.e.,* the ones do not contain similar spectral characteristics like the targeted source in the mixture, the following sparsity penalty function is proposed [10, 11]

$$\Omega_1(\mathbf{H}) = \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(g)}\|_1), \tag{13}$$

where $\mathbf{H}_{(g)}$ is a subset of $\mathbf{H}$ representing the activation coefficients for $g$-th block, $G$ is the total number of blocks, and $\epsilon$ is a non-zero constant. In this case, a block represents one training example and $G$ is the total number of used examples ($G = P + Q$). This penalty enforces the activation for relevant examples only while omitting the poorly fitting examples since their corresponding activation block will likely converge to zero, as visualized in Fig. 2b (similar figure can be seen also in [11]).

(ii) **Component sparsity-inducing penalty**

$$\Omega_2(\mathbf{H}) = \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \tag{14}$$

where $\mathbf{h}_k$ denotes $k$-th row of $\mathbf{H}$. As explained in [11], this penalty is motivated by the fact that only a part of the spectral model learned from an example may fit well with the targeted source in the mixture, while the remaining components in the model do not. Thus instead of activating the whole block, the penalty $\Omega_2(\mathbf{H})$ allows to select only the more likely relevant spectral components from $\mathbf{W}$. An example of $\mathbf{H}$ after convergence is shown in Fig. 2c (similar

figure can be seen also in [11]).

### 3.3 Proposed mixed sparsity-inducing penalty and derived algorithm

Inspired by the advantage of the penalty functions (13) and (14), we propose to combine them in another context of speech enhancement considered in this paper as follow:

$$\Omega_{\text{new}}(\mathbf{H}) = \alpha \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(g)}\|_1) + (1-\alpha) \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \tag{15}$$

where $\alpha$ weights the contribution of each term. (15) can be seen as the generalization of (13) and (14) in the sense that when $\alpha = 1$, (15) is equivalent to (13) and when $\alpha = 0$, (15) is equivalent to (14). Fig. 2d shows an example of the activation matrix $\mathbf{H}$ after convergence when the novel penalty (15) is used. It can be seen that some block converge to zero due to the contribution of the first term in (15), while in the remaining blocks, some components are zeros due to the second term in (15).

In order to derive the parameter estimation algorithm optimizing (12) with the proposed penalty function (15), one can rely on MU rules and the majorization-minimization algorithm. The resulting algorithm is summarized in Algorithm 1, where $\mathbf{Y}_{(g)}$ is a uniform matrix of the same size as $\mathbf{H}_{(g)}$ and $\mathbf{z}_k$ a uniform row vector of the same size as $\mathbf{h}_k$.

**Algorithm 1** NMF with mixed sparsity-inducing penalty

**Require:** $\mathbf{V}$, $\mathbf{W}$, $\lambda$, $\alpha$
**Ensure:** $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{WH}$
  **repeat**
    // Taking into account block sparsity-inducing penalty
    **for** $g = 1, ..., G$ **do**
      $\mathbf{Y}_{(g)} \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_{(g)}\|_1}$
    **end for**
    $\mathbf{Y} = [\mathbf{Y}_{(1)}^T, \ldots, \mathbf{Y}_{(G)}^T]^T$
    // Taking into account component sparsity-inducing penalty
    **for** $k = 1, ..., K$ **do**
      $\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|}$
    **end for**
    $\mathbf{Z} = [\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T]^T$
    // Updating activation matrix
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left( \frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \lambda(\alpha \mathbf{Y} + (1-\alpha)\mathbf{Z})} \right)^{\cdot \frac{1}{2}}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
  **until** convergence

## 4. EXPERIMENT

We will first describe the dataset, algorithm settings, and evaluation criteria in Section 4.1, we then discuss on the simulation result in Section 4.2.

### 4.1 Data and evaluation metrics

We created two training sets for speech and noise. Speech training set contains four different speeches[1] from two female voices and two male voices with 10 second duration each. Noise training set includes seven types of environmental noise[2]: kitchen sound, waterfall, square, metro, field sound, cafeteria, bird song with duration varies from 5 to 15 seconds. These sets were used to learn the generic spectral model for speech and noise as described in Section 3.1.

We evaluated the performance of the proposed algorithm via a test set containing four single-channel mixtures of speech and noise artificially mixed at 0 dB SNR. Four different noises considered are: traffic + wind sound, ocean-waves, cafeteria + guita, forest birds + car. The mixtures were sampled at 16000 Hz and have duration between 5 and 10 seconds.

Our algorithm settings are as follows. The number of iterations for MU updates in all algoritms was 100 for both training and testing. The number of NMF components was set to 32 for speech and 16 for noise. The trade-off parameter $\lambda$ determining the contribution of the sparsity-inducing penalty were tested with values ranging from 1 to 1000 for each algorithm in order to choose the best values. The factor $\alpha$ weighting the contribution of each penalty term in (15) were tested with values ranging from 0.001 to 0.9.

The speech enhancement performance was evaluated by the *source-to-distortion ratio* (SDR), and *source-to-interference ratio* (SIR) measured in dB where the higher the better. These criteria, known as BSS-EVAL metrics, have been mostly used in the source separation community and the source code is available [18, 2]. Note that we do not use two other metrics (SAR and ISR) in the BSS-EVAL since they are less important than SDR and SIR.

| Method | SDR (dB) | SIR (dB) |
|---|---|---|
| NMF-Block sparsity | 6.1 | 9.2 |
| NMF-Component sparsity | 6.3 | 9.6 |
| NMF-Proposed | **6.5** | **9.7** |

**Table 1: Average performance of source separation.**

### 4.2 Simulation result

We compare the speech enhancement performance obtained by our approach using mixed sparsity-inducing penalty in (15) (named "NMF-Proposed") with two state-of-the-art algorithms exploiting block sparsity penalty (named "NMF-Block sparsity") and component sparsity penalty only (named "NMF-Component sparsity") [10, 11]. After fine-tuning to choose the parameter which yields the highest SDR for speech[3], we set $\lambda = 5$ for NMF with block sparsity; $\lambda = 70$ for NMF with component sparsity; and $\lambda = 110$, $\alpha = 0.1$ for the proposed algorithm.

The result is averaged over all four testing mixtures for three different algorithms and shown in Table 1. One can see that the proposed algorithm offers a better speech enhancement performance in terms of both SDR and SIR compared to the two existing ones. More precisely, it gained 0.2 dB and 0.4 dB SDR higher than those of the NMF-Block and the NMF-Component, respectively.

[1]Speech files are from the International Signal Separation and Evaluation Campaign (SiSEC): *http://sisec.wiki.irisa.fr*.

[2]Some noise files are from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND): *http://parole.loria.fr/DEMAND/*.

[3]Note that among the considered criteria, SDR is the most important one since it measures the overall signal distortion.

## 5. CONCLUSION

In this paper, we presented a novel single-channel speech separation approach motivated from the recent source separation technique where generic spectral characteristics of speech signal and noise can be roughly learned in advance. We proposed to combine two existing sparsity-inducing constraints which may help resulting in better parameter estimation. In contrast with other state-of-the-art speech enhancement methods which are mostly efficient for stationary or diffuse noise, the considered approach can potentially deals with non-stationary noise such as background music thanks to the pre-learned generic spectral model. Preliminary experiment with noisy audio signals containing different types of noise confirms the effectiveness of the proposed algorithm. Future work can extend the proposed approach to multi-channel case where spatial model, such as the one considered in [19], for audio sources is incorporated. Additionally, validating the effectiveness of the proposed denoising approach for ASR would be a particular interest.

## 6. REFERENCES


[1] Jacob Benesty, Shoji Makino, and Jingdong Chen, *Speech Enhancement*, Springer, 2005.

[2] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[3] G. Kim and P. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio, Speech, Language Processing*, vol. 18, no. 8, pp. 2080–2090, 2010.

[4] J. Barkera, E. Vincent, N. Maa, H. Christensena, and P. Greena, "The pascal chime speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

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

[6] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *Signal Processing Letter*, vol. 9, no. 4, pp. 113–116, 2002.

[7] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[8] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural and Information Processing Systems 13*, 2001, pp. 556–562.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[11] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[12] D. El Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[13] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 414–421.

[14] C. Févotte, N. Bertin, and J. Durrieu, "Non-negative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[15] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparse criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[16] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[17] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Interspeech*, 2012, pp. 17–20.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1–11, 2013.