SINGLE-CHANNEL SPEAKER-DEPENDENT SPEECH ENHANCEMENT EXPLOITING GENERIC NOISE MODEL LEARNED BY NON-NEGATIVE MATRIX FACTORIZATION

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

SINGLE-CHANNEL SPEAKER-DEPENDENT SPEECH ENHANCEMENT EXPLOITING GENERIC NOISE MODEL LEARNED BY NON-NEGATIVE MATRIX FACTORIZATION

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

[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]Imaging Science Lab, Technicolor, France

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

ABSTRACT

This paper considers the single-channel speech separation problem given a noisy observation recorded by a microphone. More precisely, we focus on the speaker-dependent approach where spectral characteristic of target speech is learned in advance from a clean example. In training process, we propose to learn a generic spectral model for noise source by collecting various types of environmental noise via the established non-negative matrix factorization framework. In speech enhancement process, we propose to combine two existing group sparsity-inducing penalties in the optimization function and derive the corresponding algorithm for parameter estimation based on multiplicative update (MU) rule. Experiment result over mixtures containing different real-world noises confirms the effectiveness of our approach.

Index Terms— Speaker-dependent speech enhancement, non-negative matrix factorization, group sparsity, generic spectral model.

1. INTRODUCTION

Speech enhancement has been an active research topic for decades as it plays an important role in many domains such as telecommunication and robotics [1]. The problem becomes much harder in single-channel case, as compared to multichannel case, since spatial information about audio sources is missing, and thus any prior information about either speech or noise source can be very helpful. Among various denoising settings, we aim to speaker-dependent scenario in the paper where clean speech example is assumed to be available a priori. This use case is very popular *e.g.,* in robotics where controller's voice is often known.

The considering approach is adapted from audio source separation technique [2], where speech and noise are considered as two distinct sources appearing in a noisy observation. More precisely, a recent paper by Sun and Mysore [3] proposed to train an universal speech model by non-negative matrix factorization (NMF) [4] from different speakers and applied it for the single-channel speech separation. Similarly, Badawy *et. al.,* exploited generic NMF spectral models for all audio sources in the context of on-the-fly source separation [5]. Motivated from those above-mentioned works, we propose in this paper to train a specific spectral model for speech and a generic model for noise source in training phase. Then in the speech enhancement phase, we exploit these pre-learned NMF-based 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, on the contrary to [3] where speech model was *universal* and noise model was updated during the separation process, we consider noise model as universal and speech model is fixed during the separation process. Secondly, compared to [3] and [5] where either block sparsity-inducing penalty or component-sparsity-inducing penalty was used, we propose a combination of these two penalties which would offer better estimating the parameters in the model fitting in this paper. For the rest of the paper, Section 2 summarizes the NMF-based supervised signal separation approach as a background. We then present the proposed speaker-dependent speech enhancement approach in Section 3 followed by experimental evaluation in Section 4. Finally a conclusion is presented in Section 5.

2. NMF MODEL AND BASELINE ALGORITHM FOR SIGNAL SEPARATION

Let us start by considering a single-channel signal separation problem with two sources (speech and noise). Denoting by $\mathbf{X} \in \mathbb{C}^{F\times M}$, $\mathbf{S} \in \mathbb{C}^{F\times M}$, and $\mathbf{N} \in \mathbb{C}^{F\times M}$ 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:

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

Let $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ be 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 $\mathbf{W} \in \mathbb{R}^{F\times Q}$ and $\mathbf{H} \in \mathbb{R}^{Q\times M}$, respectively, 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, which offers scale invariant property [6]:

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

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 denotes frequency bin index and time frame index,

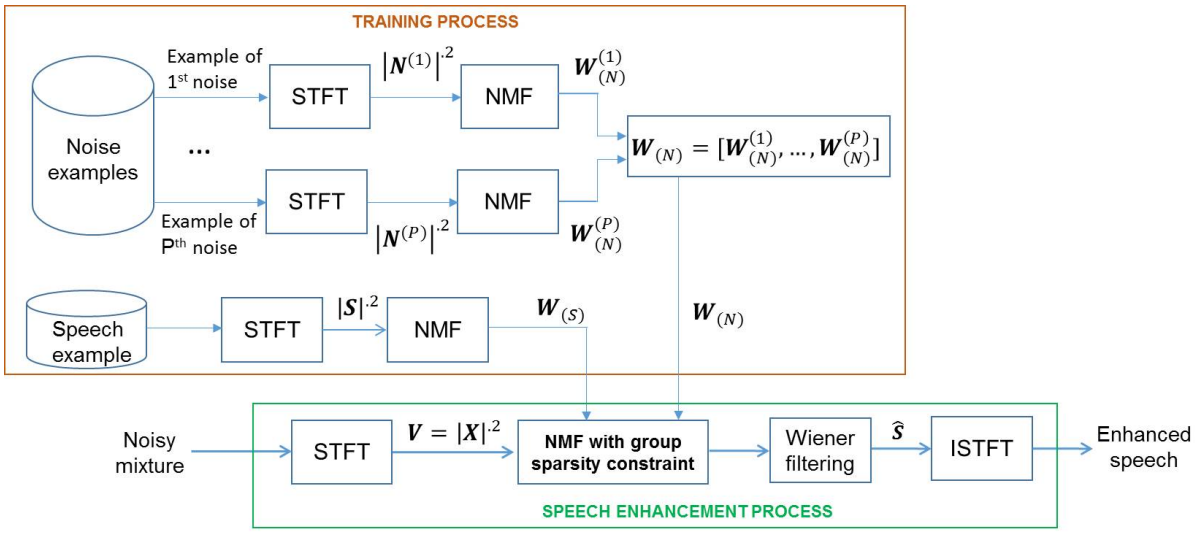

Fig. 1. General workflow of the proposed speaker-dependent speech enhancement approach.

respectively, and $d_{IS}(x\|y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$. 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 [6] as

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right)}{\mathbf{W}^T (\mathbf{WH})^{\cdot -1}} \tag{3}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right) \mathbf{H}^T}{(\mathbf{WH})^{\cdot -1} \mathbf{H}^T} \tag{4}$$

where $\mathbf{A}^T$ the transposition of matrix $\mathbf{A}$, $\odot$ denotes the element-wise Hadamard product, the power and the division is also element-wise.

In training phase of the supervised setting, spectral model for speech and noise, denoted by $\mathbf{W}_{(S)}$ and $\mathbf{W}_{(N)}$, respectively, is firstly learned from the corresponding training examples by optimizing similar criterion as (2). Then spectral model for two sources $\mathbf{W}$ is obtained by $\mathbf{W} = [\mathbf{W}_{(S)}, \mathbf{W}_{(N)}]$. In 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 as (3). Note that $\mathbf{H}$ is also partitioned into two block as $\mathbf{H} = [\mathbf{H}_{(S)}^T, \mathbf{H}_{(N)}^T]^T$, where $\mathbf{H}_{(S)}$ and $\mathbf{H}_{(N)}$ denotes block characterizing the time activations for speech and noise, respectively.

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

$$\hat{\mathbf{S}} = \frac{\mathbf{W}_{(S)}\mathbf{H}_{(S)}}{\mathbf{WH}} \odot \mathbf{X}. \tag{5}$$

And finally, the estimated time domain speech signal are obtained via the inverse STFT.

3. PROPOSED SPEAKER-DEPENDENT SPEECH SEPARATION ALGORITHM

General workflow of the proposed supervised approach for speech separation is shown in Fig. 1. In the following paragraphs, we first present the NMF-based models for speech and noise, which are learned during training process, in Section 3.1. We then describe the model fitting with the proposed group sparsity constraint for the speech enhancement process in Section 3.2. Finally, we derive the corresponding parameter estimation algorithm in Section 3.3.

3.1. Pre-trained models for speech and noise

3.1.1. Speaker-dependent speech spectral model

Assuming that $\mathbf{V}_{(S)} = |\mathbf{S}|^{\cdot 2}$ is the spectrogram of a clean speech example. Speech model $\mathbf{W}_{(S)}$ is learned given $\mathbf{V}_{(S)}$ by optimizing the criterion (similar to (2)):

$$\min_{\mathbf{H}_{(S)} \geq 0, \mathbf{W}_{(S)} \geq 0} D(\mathbf{V}_{(S)} \| \mathbf{W}_{(S)}\mathbf{H}_{(S)}), \tag{6}$$

where $\mathbf{H}_{(S)}$ is the corresponding time activation matrix.

3.1.2. Generic spectral noise model

Assuming that $\mathbf{V}_{(N)}^{(p)} = |\mathbf{N}^{(p)}|^{\cdot 2}, 1 \leq p \leq P$ is the spectrogram of p-th noise examples. First $\mathbf{V}_{(N)}^{(p)}$ is used to learn the NMF spectral model, denoted by $\mathbf{W}_{(N)}^{(p)}$, by optimizing the criterion (similar to (2)):

$$\min_{\mathbf{H}_{(N)}^{(p)} \geq 0, \mathbf{W}_{(N)}^{(p)} \geq 0} D(\mathbf{V}_{(N)}^{(p)} \| \mathbf{W}_{(N)}^{(p)}\mathbf{H}_{(N)}^{(p)}), \tag{7}$$

where $\mathbf{H}_{(N)}^{(p)}$ is time activation matrix. Given $\mathbf{W}_{(N)}^{(p)}$ for all noise examples $p = 1, ..., P$, the generic spectral models for noise is constructed as

$$\mathbf{W}_{(N)} = [\mathbf{W}_{(N)}^{(1)}, \ldots, \mathbf{W}_{(N)}^{(P)}]. \tag{8}$$

In the practical implementation, we may need several examples of different types of noise such as wind sound, cafeteria, waterfall, street noise, etc.,. (*e.g.,* $P = 7$).

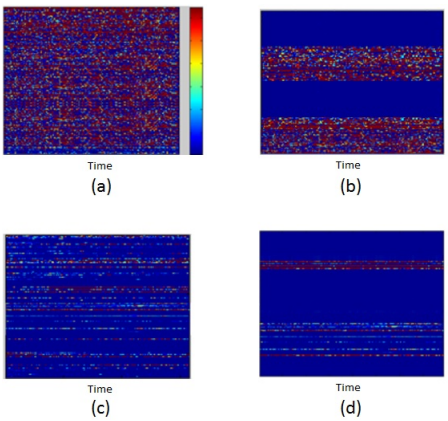

Fig. 2. Estimated activation matrix $\mathbf{H}$: (a) without a sparsity constraint, (b) with a block sparsity-inducing penalty (10), (c) with a component sparsity-inducing penalty (11), and (d) with the proposed mixed group sparsity constraint (12).

3.2. Proposed group sparsity constraint for model fitting

The generic noise model $\mathbf{W}_{(N)}$ constructed in (8) becomes a large matrix when the number of examples increases, and it is actually redundant since different examples may share some similar spectral patterns. Thus in the model fitting for the mixture spectrogram in the speech enhancement process, sparsity constraint is naturally needed so as to fit only a subset of the $\mathbf{W}_{(N)}$ to the actual noise representing in the mixture [7]. In other words, the mixture spectrogram $\mathbf{V} = |\mathbf{X}|^{.2}$ is decomposed by solving the following optimization problem

$$\underset{\mathbf{H}\geq 0}{min}\ D(\mathbf{V}\|\mathbf{WH}) + \lambda\Omega(\mathbf{H}_{(N)}) \tag{9}$$

where $\Omega(\mathbf{H}_{(N)})$ denotes a penalty function imposing sparsity on the activation matrix $\mathbf{H}_{(N)}$, and λ is a trade-off parameter determining the contribution of the penalty. When $\lambda = 0$, $\mathbf{H}_{(N)}$ 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 [5].

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

$$\Omega_1(\mathbf{H}_{(N)}) = \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(N)}^{(g)}\|_1), \tag{10}$$

where $\mathbf{H}_{(N)}^{(g)}$ is a subset of $\mathbf{H}_{(N)}$ representing the activation coefficients for g-th block, G is the total number of blocks, ϵ is a non-zero constant, and $\|.\|_1$ is ℓ_1-norm. In the considered setting, a block represents one training example and G is the total number of used examples ($G = P$). This penalty enforces the activation for relevant nosie 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 [5]).

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

$$\Omega_2(\mathbf{H}_{(N)}) = \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_{(N)}^{k}\|_1), \tag{11}$$

where $\mathbf{h}_{(N)}^{k}$ denotes k-th row of $\mathbf{H}_{(N)}$. As explained in [5], 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 $\Omega_2(\mathbf{H}_{(N)})$ allows to select only the more likely relevant spectral components from $\mathbf{W}$. An example of $\mathbf{H}_{(N)}$ after convergence is shown in Fig. 2c (similar figure can be seen also in [5]).

Motivated by penalties mentioned above, we propose a so-called *mixed group sparsity constraint* combining (10) and (11) as

$$\Omega_{\text{new}}(\mathbf{H}_{(N)}) = \alpha \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(N)}^{(g)}\|_1) + (1-\alpha) \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_{(N)}^{k}\|_1), \tag{12}$$

where α weights the contribution of each term. (12) can be seen as the generalization of (10) and (11) in the sense that when $\alpha = 1$, (12) is equivalent to (10) and when $\alpha = 0$, (12) is equivalent to (11). Fig. 2d shows an example of the activation matrix $\mathbf{H}_{(N)}$ after convergence when the novel penalty (12) is used. It can be seen that some block converge to zero due to the contribution of the first term in (12), while in the remaining blocks, some components are zeros due to the second term in (12).

3.3. Derived algorithm for parameter estimation

In order to derive the parameter estimation algorithm optimizing (9) with the proposed penalty function (12), 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 $\mathbf{H}_{(N)}^{g}$ and $\mathbf{z}_k$ a uniform row vector of the same size as $\mathbf{h}_{(N)}^{k}$.

Algorithm 1 Parameter estimation algorithm with mixed group sparsity constraint

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

4. EXPERIMENTS

We first described the data set, algorithm settings, and evaluation criteria in Section 4.1, then the result is discussed in Section 4.2.

4.1. Data and evaluation metrics

We created two training sets including for speech and noise source. Training speech example was 10 seconds long and was made by the same person with speech in the tested mixtures [1]. Noise examples were 7 types of environmental noise[2]: kitchen sound, waterfall, square, metro, living sound, presto, bird song with duration varies from 5 to 15 seconds. These training examples were used to learn the 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 5 single-channel mixtures of speech and noise artificially mixed at 0 dB signal-to-distortion ratio. 5 different noise sources considered were forestbird+car, office sound, traffic + wind sound, oceanwaves, and park sound. 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 λ determining the contribution of the sparsity-inducing penalty were tested with values ranging from 0.001 to 1000 for each algorithm in order to choose the best values. The factor α weighting the contribution of each penalty term in (12) were tested with values ranging from 0.001 to 0.9.

The speech enhancement performance was evaluated by the *source-to-distortion ratio* (SDR), *source-to-interference ratio* (SIR), and *source-to-artifacts ratio* (SAR) measured in dB where the higher the better. These criteria, known as BSS-EVAL metrics, have been mostly used in the source separation community [8].

4.2. Simulation result

We compare the speech enhancement performance obtained by our approach using mixed group sparsity constraint in (12) (named "Proposed approach") with two state-of-the-art algorithms exploiting block sparsity penalty and component sparsity penalty only (named "State of the art") [5], which are closest to our approach. After fine-tuning to choose the parameter which yields the highest SDR for speech[3], we set $\lambda = 0.02$ for NMF with block sparsity; $\lambda = 0.04$ for NMF with component sparsity; and $\lambda = 0.02$, $\alpha = 0.4$ for the proposed algorithm.

The result for each testing mixtures is shown in Table 1. Note that due to the lack of space, Table 1 shows only the highest result obtained by either block sparsity or component sparsity constraint (the two algorithms we aim to compare with) in the "State of the art" row. One can see that the proposed algorithm offers a better speech enhancement performance in terms of both SDR, SIR and SAR compared to the existing ones. More precisely, it gained 0.2 dB SDR higher than the state-of-the-art algorithm.

Method	Mixture	SDR	SIR	SAR
State of the art	Mix 1	9.4	16.6	10.4
	Mix 2	13.9	27.8	14.2
	Mix 3	0.8	1.8	10.2
	Mix 4	2.8	4.8	8.6
	Mix 5	7.1	11.4	9.4
	Average	**6.8**	**12.5**	**10.5**
Proposed approach	Mix 1	9.6	16.7	10.4
	Mix 2	14.1	28.0	14.2
	Mix 3	0.8	1.7	10.2
	Mix 4	3.0	4.9	8.7
	Mix 5	7.3	11.5	9.6
	Average	**7.0**	**12.6**	**10.6**

Table 1. Speech separation performance (criteria are measured in dB, the higher the better).

5. CONCLUSION

In this paper, we have presented a novel speaker-dependent single-channel speech separation algorithm based on NMF formulation. The presenting approach exploits a generic noise model learned from different types of environmental noise. Additionally, we have proposed to combine two existing sparsity-inducing constraints for the parameter estimation process so as to potentially improve the separation performance. Experiment with mixtures containing different types of real-world noise confirms the effectiveness of the proposed algorithm. Future research can be devoted to extend the work to multi-channel case where a spatial model, such as the one considered in [9], for audio sources is incorporated.

[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.

6. REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.