Sparsity and low-rank amplitude based blind Source Separation

Fangchen Feng, Matthieu Kowalski

SPARSITY AND LOW-RANK AMPLITUDE BASED BLIND SOURCE SEPARATION

Fangchen Feng and Matthieu Kowalski

L2S, Univ Paris-Sud-CNRS-CENTRALESUPELEC, Université Paris-Saclay, Gif-sur-Yvette, France

ABSTRACT

This paper presents a new method for blind source separation problem in reverberant environments with more sources than microphones. Based on the sparsity property in the time-frequency domain and the low-rank assumption of the spectrogram of the source, the STRAUSS (SparsiTy and low-Rank AmplitUde based Source Separation) method is developed. Numerical evaluations show that the proposed method outperforms the existing multichannel NMF approaches, while it is exclusively based on amplitude information.

Index Terms— Blind source separation, sparsity, low-rank, multichannel non-negative matrix factorization

1. INTRODUCTION AND STATE OF THE ART

Blind source separation (BSS) recovers source signals from a number of observed mixtures without knowing the mixing system. Separation of the mixed sounds has several applications in the analysis, editing, and manipulation of audio data [1]. In the real-world scenario, convolutive mixture model is considered to take the room echo and the reverberation effect into account:

$$x_i(t) = \sum_{j=1}^{N} a_{ij}(t) * s_j(t),\ \forall i = 1, 2, \ldots, M, \tag{1}$$

where s_j is the j-th source and x_i is the i-th mixture. N and M are the number of source and microphones respectively (M can be smaller than N). a_{ij} is the impulse response from the j-th source to the i-th microphone.

With the help of the Short Time Fourier Transform (STFT), the wideband model (1) can be well approximated by the narrowband model as:

$$\tilde{x}_i(f, \tau) \simeq \sum_{j=1}^{N} \hat{a}_{ij}(f)\tilde{s}_j(f, \tau) \tag{2}$$

where f and τ are respectively the frequency and the time index. $\tilde{x}_i(f, \tau)$ and $\tilde{s}_j(f, \tau)$ are the STFT coefficients of $x_i(t)$ and $s_j(t)$ respectively. $\hat{a}_{ij}$ is the Fourier transform of the impulse response a_{ij}.

The source separation problem for convolutive mixtures is a big challenge, especially in the under-determined setting ($M < N$). In this case, the sparsity property of signals is a popular choice. For example, the DUET method [2] proposes a sparsity based time-frequency masking technique which is efficient for anechoic mixtures. In [3], the authors combine the sparsity with the ICA (Independent Component Analysis) to solve the separation problem for each f. However, given the permutation ambiguity of the estimation, an extra step is necessary to determine the order of the estimation in each frequency band.

At the same time, there is a growing interest for applying Non-negative Matrix Factorization (NMF) based methods to music signals. After being used in [4] for polyphonic music transcription, NMF has been applied in [5] for single-channel source separation. The method developed in [5] is based on the assumption that the magnitude coefficient matrix of the observation can be modeled as a linear combination of basis functions. Let $\mathbf{X}^i \in \mathbb{R}_+^{L_F \times L_T}$ such that $\mathbf{X}^i_{f,\tau} = |x_i(f, \tau)|$, where L_F and L_T are the total number of frequency bands and time frames respectively, the NMF reads[1]:

$$\mathbf{X}^i = \mathbf{V}^i\mathbf{H}^i, \tag{3}$$

where $\mathbf{V}^i \in \mathbb{R}_+^{L_F \times K}$ and each column of $\mathbf{V}^i$ contains one pattern (frequency motif) ; $\mathbf{H}^i \in \mathbb{R}_+^{K \times L_T}$ and each row of $\mathbf{H}^i$ contains the activation coefficient for the corresponding pattern. K is the rank of $\mathbf{X}^i$. The NMF can be achieved through minimizing the measure of fit:

$$D(\mathbf{X}, \mathbf{VH}) = \sum_{f=1}^{L_F}\sum_{\tau=1}^{L_T} d(\mathbf{X}_{f,\tau}, [\mathbf{VH}]_{f,\tau}), \tag{4}$$

where $(x, y) \mapsto d(x, y)$ is a scalar cost function. Popular choices for audio source separation are Kullback-Leibler (KL) divergence [6] and Itakura-Saito (IS) divergence [7].

Based on the NMF model, the separation is then performed by clustering to regroup the patterns corresponding to the sources. Automatic clustering of the components is a difficult task [5]. Some unsupervised clustering methods have been proposed [8, 9], but their performance is limited [5]. Supervised clustering based on pattern recognition produces better results [10, 11], but requires a training stage.

The authors of [12–14] have generalized the single channel NMF based methods to the multichannel case by extending the generative model leading to the Itakura-Saito divergence. In the multichannel scenario, the regroupment is tackled by exploiting the spatial characteristics based on the covariance model [15]. These algorithms are shown to have good performance for music sources. However, they suffer from the high initialization sensitivity and high computational costs.

In this paper, we study the narrowband blind separation problem (2) using the two assumptions:

- Sparsity: for each observed time-frequency (t-f) bin, only one source contributes to the mixtures.
- Low rank: each source can be decomposed using NMF.

We then propose a new multichannel NMF method by using only the amplitude of the STFT coefficients of the observations, called STRAUSS (SparsiTy and low-Rank Amplitude based Source Separation). Although it is mentioned in [12] that phase information is crucial for separation, we show that by concentrating on the amplitude information, the proposed algorithms outperform the state of the art approaches.

[1] NMF can also be applied to the squared modulus, but we stick here to the modulus for the sake of simplicity

The rest of the paper is organized as follows. The proposed STRAUSS approach is presented in Section 2. Section 3 reports experimental results on the source separation. Section 4 concludes the paper.

2. PROPOSED METHOD

In this section, we develop the proposed STRAUSS method that takes advantage of both the sparsity and the low-rank assumptions. First, the spectrograms of the observation are decomposed into components using newly developed algorithms. Then, the regroupment is performed using the spectral clustering technique. At last, the estimated source signals are reconstructed via Wiener filter.

2.1. Sparsity and low-rank assumption

The sparsity assumption for audio t-f coefficients is a popular choice for source separation and is shown to have promising results for instantaneous mixtures [16]. One extreme case of the sparsity assumption is that for a given f and τ, there is only one activated (dominant) source [2, 17]. This assumption can be written as follow:

Assumption 1 (Sparsity)**.** *For each time-frequency index f, τ, only one source is active, such that*

$$\tilde{x}_i(f,\tau) = \hat{a}_{ij^*}(f)\tilde{s}_{j^*}(f,\tau),\ \forall i$$

where j^* is the index of the activated source for the given f, τ.

We denote by Θ_{j^*} the set that contains all the index f, τ where the source j^* is activated.

The low-rank assumption used by NMF methods lead to efficient BSS algorithms in the convolutive setting for audio signals [14]. This second assumption reads:

Assumption 2 (Low rank)**.** *For all sources, and for all time-frequency index,*

$$|\tilde{s}_j(f,\tau)| = \sum_{k=1}^{K_j} v_j^k(f)h_j^k(t)\ ,\ v_j^k(f), h_j^k(t) \geq 0$$

where $K_j << \min\{L_F, L_T\}$ is the rank of the j-th source.

Combining Assumptions 1 and 2, one has for all mixtures i and ℓ, and for all $(f,\tau) \in \Theta_{j^*}$

$$\begin{aligned}\sqrt{|\tilde{x}_i(f,\tau)||\tilde{x}_\ell(f,\tau)|} &= \sqrt{|\hat{a}_{ij^*}(f)||\hat{a}_{\ell j^*}(f)|}\cdot|\tilde{s}_{j^*}(f,\tau)|,\\ &= \sum_{k=1}^{K_{j^*}} \sqrt{|\hat{a}_{ij^*}(f)||\hat{a}_{\ell j^*}(f)|}\cdot v_{j^*}^k(f)h_{j^*}^k(\tau)\\ &= \sum_{k=1}^{K_{j^*}} \tilde{v}_{j^*}^k(f)h_{j^*}^k(\tau) \qquad (5)\end{aligned}$$

where $\tilde{v}_{j^*}^k(f) = \sqrt{|\hat{a}_{ij^*}(f)||\hat{a}_{\ell j^*}(f)|}\cdot v_{j^*}^k(f)$. The last equation in (5) shows the next proposition:

Proposition 1. *Let the positive matrix $\mathbf{X}^{i\ell} \in \mathbb{R}_+^{L_F\times L_T}$ such that $\mathbf{X}_{f,\tau}^{i\ell} = \sqrt{|\tilde{x}_i(f,\tau)||\tilde{x}_\ell(f,\tau)|}$. Let $K = \sum_{j=1}^N K_j$. Then, $\mathbf{X}^{i\ell}$ has a low-rank structure and can be factorized such that:*

$$\mathbf{X}^{i\ell} = \tilde{\mathbf{V}}^{i\ell}\mathbf{H}$$

with

$$\tilde{\mathbf{V}} \in \mathbb{R}_+^{L_F\times K},\ \tilde{\mathbf{V}}_{f,k}^{i\ell} = \sqrt{|\hat{a}_{ij^*}(f)||\hat{a}_{\ell j^*}(f)|}\cdot v_{j^*}^k(f), f \in \Theta_{j^*}$$

and

$$\mathbf{H} \in \mathbb{R}_+^{K\times L_T},\ \mathbf{H}_{k,\tau} = h_{j^*}^k(\tau), \tau \in \Theta_{j^*}$$

One can remark that for all (i,ℓ), the factorizations of $\mathbf{X}^{i\ell}$ share the same activation matrix $\mathbf{H}$, up to a permutation. Moreover, this proposition shows that each column of $\tilde{\mathbf{V}}^{i\ell}$ contains a frequency motif that comes from only one source, say j^*, and is a weighted version of the corresponding pattern $v_{j^*}^k(f)$. The challenge now is to perform a suitable clustering on $\tilde{\mathbf{V}}^{i\ell}$ for regroupment.

2.2. Joint-NMF and Clustering

From Proposition 1, one has that:

$$\mathbf{X}^{ii} = \tilde{\mathbf{V}}^{ii}\mathbf{H}, \quad \mathbf{X}^{i\ell} = \tilde{\mathbf{V}}^{i\ell}\mathbf{H}$$

with

$$\begin{aligned}\tilde{\mathbf{V}}_{f,k}^{ii} &= |\hat{a}_{ij^*}(f)|v_{j^*}^k(f)\\ \tilde{\mathbf{V}}_{f,k}^{i\ell} &= \sqrt{|\hat{a}_{ij^*}(f)||\hat{a}_{\ell j^*}(f)|}\cdot v_{j^*}^k(f)\end{aligned}$$

then, for all k,

$$\frac{\tilde{\mathbf{V}}_{f,k}^{ii}}{\tilde{\mathbf{V}}_{f,k}^{i\ell}} = \sqrt{\frac{|\hat{a}_{ij^*}(f)|}{|\hat{a}_{\ell j^*}(f)|}}. \qquad (6)$$

The proposed idea is then to perform a joint-NMF of the observed matrices $\mathbf{X}^{i\ell}$, sharing the same activation matrix $\mathbf{H}$, and then performing the clustering on the ratios of the obtained pattern matrices $\tilde{\mathbf{V}}^{i\ell}$.

2.2.1. Joint-NMF

Sticking to the stereo setting for the sake of simplicity, i.e. $M = 2$, we first perform the following joint-NMF:

$$\begin{aligned}\tilde{\mathbf{V}}^{11}, \tilde{\mathbf{V}}^{22}, \tilde{\mathbf{V}}^{12}, \mathbf{H} = \underset{\mathbf{V}^{11},\mathbf{V}^{22},\mathbf{V}^{12},\mathbf{H}}{\operatorname{argmin}} & D(\mathbf{X}^{11}, \mathbf{V}^{11}\mathbf{H})\\ + D(\mathbf{X}^{22}, \mathbf{V}^{22}\mathbf{H}) &+ D(\mathbf{X}^{12}, \mathbf{V}^{12}\mathbf{H})\end{aligned} \qquad (7)$$

where $D(\mathbf{X}, \mathbf{Y})$ can be IS or KL divergence. Such a minimization can be tackled by using multiplicative update rules adapted from classical NMF decomposition (See Appendix A for details).

2.2.2. Clustering and source reconstruction

The ratios given by (6) can be sensitive to small numbers. In order to avoid such instabilities, we consider the following element-wise ratios:

$$\mathbf{R}^1 = \frac{\tilde{\mathbf{V}}^{11}}{\tilde{\mathbf{V}}^{12}} = \mathbf{R}^2 = \frac{\tilde{\mathbf{V}}^{12}}{\tilde{\mathbf{V}}^{22}}$$

Then, we select the elements in $\mathbf{R}^1$ and $\mathbf{R}^2$ that are close enough w.r.t a given threshold ϵ to construct a matrix $\mathbf{R}$

$$\mathbf{R}_{f,k} = \begin{cases} \frac{\mathbf{R}_{f,k}^1 + \mathbf{R}_{f,k}^2}{2} & \text{if } \left|\tilde{\mathbf{V}}_{f,k}^{11}\tilde{\mathbf{V}}_{f,k}^{22} - \left(\tilde{\mathbf{V}}_{f,k}^{12}\right)^2\right| < \varepsilon \\ 0 & \text{otherwise}\end{cases} \qquad (8)$$

We use the spectral clustering [18] on the columns of $\mathbf{R}$, where the sparse correlation coefficient[2] is used as the distance between

[2]The sparse correlation coefficient calculates the correlation coefficient only on the commun support of two vectors.

any two columns. The separated sources are then estimated from the amplitude of their STFT coefficients using a Wiener filtering [7]. The STRAUSS method is summarized in Algorithm 1:

Algorithm 1 (STRAUSS)**.**

1. *Calculate the amplitude matrices* $\mathbf{X}^{11}$, $\mathbf{X}^{22}$ *and* $\mathbf{X}^{12}$ *from the observations;*
2. *Find* $\tilde{\mathbf{V}}^{11}$, $\tilde{\mathbf{V}}^{22}$, $\tilde{\mathbf{V}}^{12}$ *and* $\mathbf{H}$ *by solving* (7) *(See Appendix A);*
3. *Calculate* $\mathbf{R}$ *according to* (8)*;*
4. *Perform the clustering for the columns of* $\mathbf{R}$ *using spectral clustering;*
5. *Reconstruct the estimated source using Wiener filtering.*

3. EXPERIMENTS

3.1. Experimental setup

We evaluated the proposed method with both IS and KL divergence for stereo music mixtures ($M = 2$) that contained three music parts ($N = 3$). The room impulse responses were simulated via the toolbox in [19]. The distance between the two microphones varied from 4 cm to 1 m. The reverberation time (RT_{60}) was set from 50 ms to 400 ms. For each case, we created 10 mixtures using sources from the datasets [20, 21]. The mixtures were down-sampled to 14.7 kHz and truncated to 8 s. We chose a tight STFT with a Hann window of length 1024 samples (69.7 ms) with 50% overlap, using the LTFAT implementation [22].

The separation performance was evaluated using the Signal-to-Distorsion Ratio (SDR), Signal-to-Interference Ratio (SIR), source Image to Spatial distortion Ratio (ISR) and Signals to Artifacts Ratio (SAR) [23]. The SDR reveals the overall quality of each estimated source. SIR indicates the crosstalk from other sources. ISR measures the amount of spatial distortion and SAR is related to the amount of musical noise. The average result over the 10 mixtures is given.

For the proposed algorithms, the rank for the NMF decomposition of the observations was set to 12, using 500 iterations of the multiplicative update rules. The parameter[3] in (8) was set to $\epsilon = 10^{-4}$.

The proposed algorithms are denoted by STRAUSS-IS and STRAUSS-KL depending on the chosen divergence for the NMF. STRAUSS approach is compared with the MNMF [12] and the "Full rank" method of [15]. All the algorithms are initialized randomly with 10 different initializations. For the proposed STRAUSS approach, we also used a deterministic initialization based on the complex SVD [24], and are denoted by STRAUSS-IS-SVD and STRAUSS-KL-SVD.

For the purpose of comparison, we also developed oracle versions of the proposed algorithms: after the NMF step initialized by the complex SVD, the original sources were used as the reference for clustering. These oracle versions of the algorithms are designed to illustrate the best clustering achievable, and are denoted by STRAUSS-IS-Oracle and STRAUSS-KL-Oracle.

3.2. Source separation results

Figure 1 shows the separation results obtained with the proposed algorithms for $\text{RT}_{60} = 250$ ms with the microphone distance $d = 4$ cm.

[3] In practice, to make the algorithms more robust, we eliminated the elements in $\tilde{\mathbf{V}}^{11}$, $\tilde{\mathbf{V}}^{22}$ and $\tilde{\mathbf{V}}^{12}$ which are less than $\epsilon = 10^{-4}$

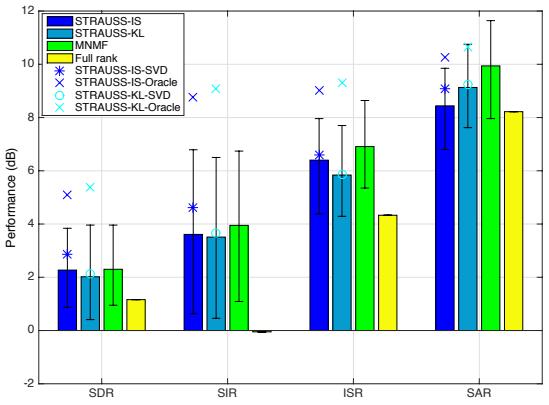

Fig. 1. Source separation performance. For STRAUSS-IS/KL and MNMF, bar represents the mean value and the error bar represents the maximum and minimum value over 10 trials.

One can notice that the results obtained by the complex-SVD initialization technique are better than the average performance with random initializations, and, for STRAUSS-IS, clearly outperform the average results of MNMF in terms of SDR and SIR. The oracle results show that the SDR, SIR, ISR and SAR can only be improved up to about 4 dB using only the amplitude of the sources.

On Figure 2 we display the performance of the STRAUSS-IS/KL-SVD approaches as a function of the reverberation time (RT_{60}) in terms of SDR and SIR with a microphone distance $d = 4$ cm, compared to the average results of the MNMF, and the Full rank method.

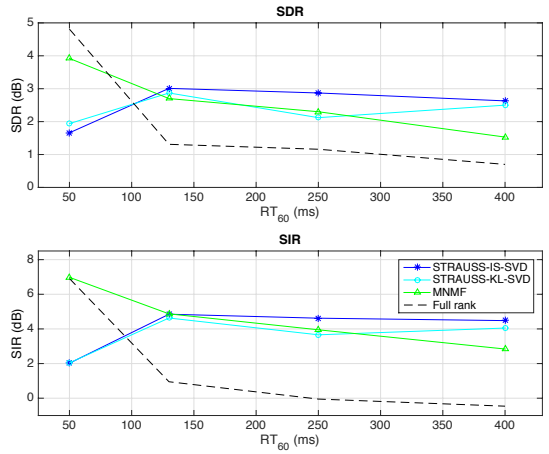

Fig. 2. Performance of the algorithms as a function of the reverberation time.

Except the low reverberation condition ($\text{RT}_{60} = 50$ ms), STRAUSS-SVD approaches appear to be robust to the reverberation time and outperform the state of the art approaches in higher reverberation situations.

Finally, Figure 3 shows the performance of the proposed algorithms as a function of the microphone distance, for a reverberation

time $RT_{60} = 250$ ms.

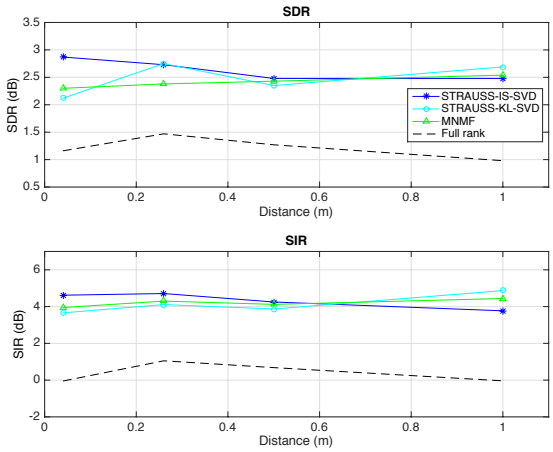

Fig. 3. Performance of the algorithms as a function of the microphone distance

Although STRAUSS-IS-SVD decreases slightly as the distance increases, it appears to be robust to the microphone distance and outperforms the state of the art approaches in most of the situations (except $d = 1$ m).

3.3. Computational time

The computational time of the STRAUSS algorithms, MNMF and the Full rank method, with $\text{RT}_{60} = 250$ ms and $d = 4$ cm, are reported in Table 1.

Table 1. Computational time for different algorithms

STRAUSS-IS	STRAUSS-KL	MNMF	Full Rank
92.6 s	36.7 s	2381.4 s	3415.4 s

It is clear that the proposed algorithms are much faster than the existing approaches.

4. CONCLUSION

We have developed a new approach, STRAUSS, based on the low-rank and sparsity assumption of the sources which concentrates on amplitude information of the STFT coefficients. Experimental results show that the derived algorithms outperform the state of the art results for the under-determined convolutive mixtures, while being twenty times faster. It is also interesting that these results are obtained without using the phase information. Compared to the oracle results, there is only 3 dB of improvement by using only the amplitude of the coefficients. As stressed in [12] we are convinced that the phase information must be used to improve the separation. This work shows that there exists a great potential to improve separation techniques of convolutive mixtures.

A. APPENDIX

Based on the algorithms proposed in [6, 25], the multiplicative update rules can be obtained for IS and KL divergence as follows:

IS joint-NMF

$$\mathbf{V}^{11} \leftarrow \mathbf{V}^{11} \circ \sqrt{\frac{(\mathbf{X}^{11}/\hat{\mathbf{X}}^{11,2})\mathbf{H}^T}{\hat{\mathbf{X}}^{11,-1}\mathbf{H}^T}} \tag{9}$$

$$\mathbf{V}^{22} \leftarrow \mathbf{V}^{22} \circ \sqrt{\frac{(\mathbf{X}^{22}/\hat{\mathbf{X}}^{22,2})\mathbf{H}^T}{\hat{\mathbf{X}}^{22,-1}\mathbf{H}^T}} \tag{10}$$

$$\mathbf{V}^{12} \leftarrow \mathbf{V}^{12} \circ \sqrt{\frac{(\mathbf{X}^{12}/\hat{\mathbf{X}}^{12,2})\mathbf{H}^T}{\hat{\mathbf{X}}^{12,-1}\mathbf{H}^T}} \tag{11}$$

$$\mathbf{H} \leftarrow \mathbf{H} \circ \sqrt{\frac{\mathbf{V}^{11}(\mathbf{X}^{11}/\hat{\mathbf{X}}^{11,2}) + \mathbf{V}^{22}(\mathbf{X}^{22}/\hat{\mathbf{X}}^{22,2}) + \mathbf{V}^{12}(\mathbf{X}^{12}/\hat{\mathbf{X}}^{12,2})}{\mathbf{V}^{11,T}\hat{\mathbf{X}}^{11,-1} + \mathbf{V}^{22,T}\hat{\mathbf{X}}^{22,-1} + \mathbf{V}^{12,T}\hat{\mathbf{X}}^{12,-1}}} \tag{12}$$

KL joint-NMF

$$\mathbf{V}^{11}_{f,k} \leftarrow \mathbf{V}^{11}_{f,k} \frac{\sum_\tau \frac{\mathbf{H}_{k,\tau}\mathbf{X}^{11}_{f,\tau}}{\hat{\mathbf{X}}^{11}_{f,\tau}}}{\sum_\tau \mathbf{H}_{k,\tau}} \tag{13}$$

$$\mathbf{V}^{22}_{f,k} \leftarrow \mathbf{V}^{22}_{f,k} \frac{\sum_\tau \frac{\mathbf{H}_{k,\tau}\mathbf{X}^{22}_{f,\tau}}{\hat{\mathbf{X}}^{22}_{f,\tau}}}{\sum_\tau \mathbf{H}_{k,\tau}} \tag{14}$$

$$\mathbf{V}^{12}_{f,k} \leftarrow \mathbf{V}^{12}_{f,k} \frac{\sum_\tau \frac{\mathbf{H}_{k,\tau}\mathbf{X}^{12}_{f,\tau}}{\hat{\mathbf{X}}^{12}_{f,\tau}}}{\sum_\tau \mathbf{H}_{k,\tau}} \tag{15}$$

$$\mathbf{H}_{k,\tau} \leftarrow \mathbf{H}_{k,\tau} \frac{\sum_f \frac{\mathbf{V}^{11}_{f,k}\mathbf{X}^{11}_{f,\tau}}{\hat{\mathbf{X}}^{11}_{f,\tau}} + \sum_f \frac{\mathbf{V}^{22}_{f,k}\mathbf{X}^{22}_{f,\tau}}{\hat{\mathbf{X}}^{22}_{f,\tau}} + \sum_f \frac{\mathbf{V}^{12}_{f,k}\mathbf{X}^{12}_{f,\tau}}{\hat{\mathbf{X}}^{12}_{f,\tau}}}{\sum_f \mathbf{V}^{11}_{f,k} + \sum_f \mathbf{V}^{22}_{f,k} + \sum_f \mathbf{V}^{12}_{f,k}} \tag{16}$$

B. REFERENCES

[1] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

[2] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE transactions on*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] Shoko Araki, Shoji Makino, Audrey Blin, Ryo Mukai, and Hiroshi Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 3, pp. iii–881.

[4] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, 2003, pp. 177–180.

[5] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[6] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[7] C. Févotte, "Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition," in *Machine Audition: Principles, Algorithms and Systems*, Wenwu Wang, Ed., chapter 11. IGI Global Press, Aug. 2010.

[8] Michael A Casey and Alex Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proceedings of the International Computer Music Conference*, 2000, pp. 154–161.

[9] Shlomo Dubnov, "Extracting sound objects by independent subspace analysis," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[10] Shankar Vembu and Stephan Baumann, "Separation of vocals from polyphonic audio recordings," in *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*, 2005, pp. 337–344.

[11] Marko Helen and Tuomas Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Signal Processing Conference (EUSIPCO), 2005 13th European*. IEEE, 2005, pp. 1–4.

[12] Hideyuki Sawada, Hirokazu Kameoka, Shunsuke Araki, and Naonori Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 971–982, 2013.

[13] Simon Arberet, Alexey Ozerov, Ngoc QK Duong, Emmanuel Vincent, Rémi Gribonval, Frédéric Bimbot, and Pierre Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 1–4.

[14] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, 2010.

[15] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1830–1840, 2010.

[16] Fangchen Feng and Matthieu Kowalski, "A unified approach for blind source separation using sparsity and decorrelation," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1736–1740.

[17] Simon Arberet, Rémi Gribonval, and Frédéric Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *Signal Processing, IEEE Transactions on*, vol. 58, no. 1, pp. 121–133, 2010.

[18] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al., "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[19] Eric A Lehmann and Anders M Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[20] M Vinyes, "MTG MASS database," `http://www.mtg.upf.edu/static/mass/resources`, 2008.

[21] Shoko Araki, Francesco Nesta, Emmanuel Vincent, Zbynvek Koldovskỳ, Guido Nolte, Andreas Ziehe, and Alexis Benichoux, "The 2011 signal separation evaluation campaign (sisec2011):-audio source separation," in *Latent Variable Analysis and Signal Separation*, pp. 414–422. Springer, 2012.

[22] Peter L Sondergaard, Bruno Torresani, and Peter Balazs, "The linear time frequency analysis toolbox," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 04, 2012.

[23] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] JM Becker, Matthias Menzel, and Christian Rohlfing, "Complex SVD initialization for NMF source separation on audio spectrograms," *DAGA 2015*, 2015.

[25] Cédric Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 1980–1983.