Sparsity and Low-Rank Amplitude Based Blind Source Separation

Fangchen Feng and Matthieu Kowalski
Univ Paris-Sud-CNRS-CentraleSupelec
Gif-sur-Yvette, France
Email: forname.name@u-psud.fr

***Abstract*—Based on the sparsity property in the time-frequency domain and the low-rank assumption of the spectrogram of the sources, the STRAUSS (SparsiTy and low-Rank AmplitUde based Source Separation) method is presented. Numerical evaluations show that the proposed method outperforms the existing multichannel NMF approaches on music signals, while it is exclusively based on amplitude information. Some evaluation on speech are also presented.**

I. Model

We study the narrowband blind separation problem

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

using the two assumptions:

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.

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

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

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

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).

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

$$\sqrt{|\tilde{x}_i(f,\tau)||\tilde{x}_\ell(f,\tau)|} = \sum_{k=1}^{K_{j^*}} \tilde{W}_{j^*}^k(f)h_{j^*}^k(\tau) \tag{1}$$

where $\tilde{W}_{j^*}^k(f) = \sqrt{|\hat{a}_{ij^*}(f)||\hat{a}_{\ell j^*}(f)|}\cdot w_{j^*}^k(f)$.

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 $\tilde{\mathbf{W}}^{i\ell}$.

II. Algorithm

Sticking to the stereo setting, we first perform a joint-NMF:

$$\tilde{\mathbf{W}}^{11},\tilde{\mathbf{W}}^{22},\tilde{\mathbf{W}}^{12},\mathbf{H} = \underset{\tilde{\mathbf{W}}^{11},\tilde{\mathbf{W}}^{22},\tilde{\mathbf{W}}^{12},\mathbf{H}}{\operatorname{argmin}} D(\mathbf{X}^{11},\tilde{\mathbf{W}}^{11}\mathbf{H}) + D(\mathbf{X}^{22},\tilde{\mathbf{W}}^{22}\mathbf{H}) + D(\mathbf{X}^{12},\tilde{\mathbf{W}}^{12}\mathbf{H}) \tag{2}$$

where $D(\mathbf{X},\mathbf{Y})$ can be Itakura-Saito or Kullback-Leibler divergence.

We consider the following element-wise ratios:

$$\mathbf{R}^1 = \frac{\tilde{\mathbf{W}}^{11}}{\tilde{\mathbf{W}}^{12}} = \mathbf{R}^2 = \frac{\tilde{\mathbf{W}}^{12}}{\tilde{\mathbf{W}}^{22}} \tag{3}$$

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 $\mathbf{R}$

$$\mathbf{R}_{f,k} = \begin{cases} \frac{\mathbf{R}^1_{f,k}+\mathbf{R}^2_{f,k}}{2} & \text{if } \left|\tilde{\mathbf{W}}^{11}_{f,k}\tilde{\mathbf{W}}^{22}_{f,k} - \left(\tilde{\mathbf{W}}^{12}_{f,k}\right)^2\right| < \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

We then use the spectral clustering [5] with the sparse correlation on the columns of $\mathbf{R}$.

III. Results

The proposed method is evaluated on stereo music and speech mixtures, with 3 sources. The room impulse responses were simulated via the toolbox [4]. 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 [1], [9]. 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 for music experiments, using the LTFAT implementation [7]. The separation performance was evaluated using SDR/SIR/ISR/SAR [8].

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 [6] and the "Full rank" method of [3]. 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 [2], 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.

TABLE I
COMPUTATIONAL TIME FOR DIFFERENT ALGORITHMS

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

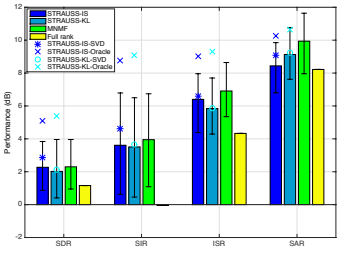

Fig. 1. Music: Source separation performance. For STRAUSS-IS/KL and MNMF, bar represents the mean value and the error bar represents the maximum and minimum value over 10 trials.

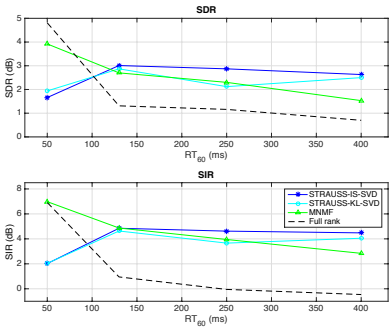

Fig. 2. Music: Performance of the algorithms as a function of the reverberation time.

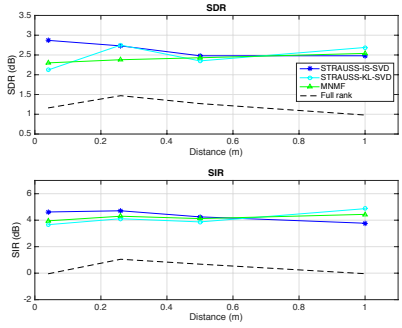

Fig. 3. Music: Performance of the algorithms as a function of the microphone distance

REFERENCES

[1] 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*, pages 414–422. Springer, 2012.

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

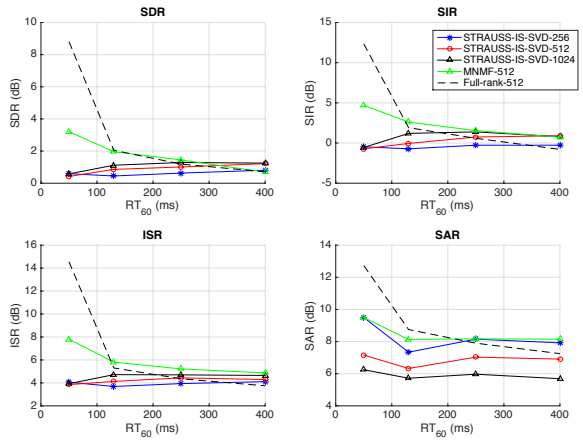

Fig. 4. Speech: Performance of the algorithms as a function of the reverberation time, with a microphone distance of 4 cm.

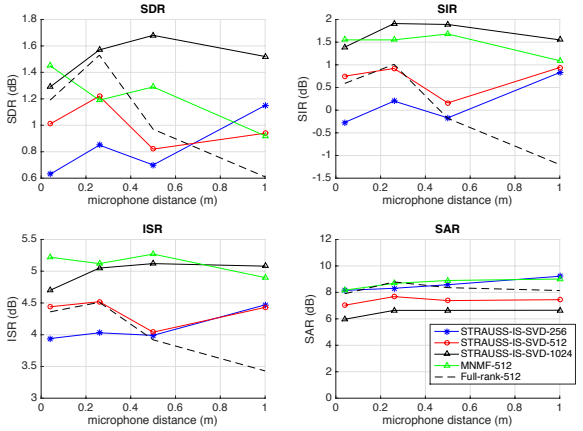

Fig. 5. Speech: Performance of the algorithms as a function of the microphone distance, with $\mathrm{RT}_{60} = 250$ ms

[3] 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*, 18(7):1830–1840, 2010.

[4] 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*, 124(1):269–277, 2008.

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

[6] 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*, 21(5):971–982, 2013.

[7] Peter L Sondergaard, Bruno Torresani, and Peter Balazs. The linear time frequency analysis toolbox. *International Journal of Wavelets, Multiresolution and Information Processing*, 10(04), 2012.

[8] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.

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