A Novel Weighted Sparse Component Analysis for Underdetermined Blind Speech Separation

Yudong He, Baeck Hyun Woo, and Richard H.Y. So

***Abstract*—Sparse component analysis (SCA) is a popular underdetermined blind speech separation (UBSS) method. It models all sources to have an identical distribution. As speeches do not have identical distribution, SCA performs suboptimal. Some studies have improved the performance of SCA by weighting the sources through a reweighting scheme. However, they are not UBSS methods because they assume that the mixing process is known. This paper proposes a novel weighting scheme, called sparse spatial component analysis (SSCA) without the need to know the mixing process. In SSCA, weights, sources, and the parameters for modeling the mixing process are jointly optimized, making it a UBSS method. Simulation experiments show that for instantaneous mixtures, SSCA outperforms SCA and reweighted SCA, improving the source-to-distortion ratio (SDR) by 4 dB and reducing the computational time by 40%. Further, experiments using real-world recordings reveal that SSCA outperforms multichannel non-negative matrix factorization and full-rank covariance analysis (FCA) in terms of SDR. The speed of SSCA is 200% faster than FCA.**

***Index Terms*—blind source separation, underdetermined linear system, sparse component analysis, weighted l_1 minimization.**

I. Introduction

UNDERTERMINED blind speech separation (UBSS) aims to separate speech sources from their mixtures where the number of sources exceeds the number of microphones and the mixing process is unknown. Sparse component analysis (SCA) [1]–[6], multichannel non-negative matrix factorization (MNMF) [7]–[11], and full-rank spatial covariance analysis (FCA) [12]–[14] are mainstream UBSS methods.

MNMF and FCA have problems with inconsistency between the assumed and actual distribution. Speech sources and spatial images (i.e., the contribution of sources to mixture channels) are assumed to be Gaussian distributed in MNMF and FCA, respectively. However, the actual distribution of speech in the time-frequency (TF) domain is non-Gaussian, more closely resembling a Laplace distribution assumed in SCA [15]. Hence, the performances of MNMF and FCA are suboptimal. On the other hand, SCA is limited by its assumption, made for mathematical convenience, that the sources are identically distributed which ignores the unique patterns that speech exhibits in the TF domain [15]. Several studies [16]–[18] addressed this issue and formulated reweighted SCA (RWSCA) by adopting Cande's reweighting scheme [19] to weight sources and thereby improved the performance of SCA. However, these methods are not UBSS methods because they assume that the mixing process is known. In summary, weighted SCA has a more appropriate speech model than the current mainstream analytic UBSS methods, but due to the lack of innovation in the weighting scheme, few weighted SCA methods have been proposed as qualified UBSS methods.

In this paper, we propose a weighted SCA, called sparse spatial component analysis (SSCA), with a novel weighting scheme to solve UBSS without knowing the mixing process. We note that the strong sparsity (i.e., disjointness) of speech in the TF domain greatly facilitates the construction of effective weights by simply projecting the signals received by the microphone array onto the microphone gains related to source directions. We formulate two weighted l_1 minimization algorithms to estimate sources from the instantaneous and convolutive mixtures. For the convolutive algorithm, the weights, sources, and the parameters used to model the mixing process are jointly optimized in each iteration. Therefore, SSCA does not require extra methods to solve the complex mixing process estimation problem; instead, UBSS is solved in a single optimization framework. Simulation experiments reveal that for instantaneous mixtures, SSCA outperforms SCA and reweighted SCA, enhancing the source-to-distortion ratio (SDR) by 4 dB and reducing computational time by up to 40%. Experiments using real-world recordings from a public dataset demonstrate that the proposed SSCA outperforms SCA [4], MNMF [7], [9], [10] and FCA [12] in terms of SDR. In addition, the speed of SSCA is 200% faster than FCA.

The rest of the paper is organized as follows. In Section II, we introduce SCA. Section III presents our proposed SSCA. Simulated experiments validating the effectiveness of our proposed method are described in Section IV. Finally, we conclude in Section V.

II. Sparse Component Analysis

A. Mixing Models

We consider two mixing models. The first is the so-called instantaneous mixing model which assumes a linear superposition of source signals, i.e.,

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

where $\mathbf{X}$ is an $M \times T$ matrix containing mixed signals $x_m(t)$, with $m = 1, \cdots, M$ and $t = 1 \cdots, T$, M being the number of microphones and T being the number of signal samples. $\mathbf{S}$ is an $N \times T$ matrix composed of source signals $s_n(t)$, with $n = 1, \cdots, N$, N being the number of sources. The number of microphones is less than the number of sources, i.e., $M < N$, under the underdetermined condition. The mixing matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{M \times N}$ contains real-valued microphone gains and is

Y. He, B. H. Woo, and R.H.Y. So are with the Department of Industry Engineering & Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong (e-mail: yhebh@connect.ust.hk).

assumed to be unknown in UBSS tasks. The instantaneous mixing model is applicable in anechoic environments where arrival delays among microphones can be neglected [20].

When arrival delays and reverberation are significant, a more realistic convolutive mixing model is introduced. This model can be simplified by narrowband approximation, which assumes instantaneous mixing in each frequency band [21]:

$$\mathbf{X}_f = \mathbf{A}_f \mathbf{S}_f, \tag{2}$$

where $\mathbf{X}_f$ is an $M \times Q$ complex matrix containing the short-time Fourier transform (STFT) coefficients $\{x_m(q,f)\}$ of mixed signals at the frequency band f, with $q = 1, \cdots, Q$ and $f = 1, \cdots, F$, Q being the total number of time frames and F being the total number of frequency bands. Similarly, $\mathbf{S}_f$ is an $N \times Q$ complex matrix composed of STFT coefficients $\{s_n(q,f)\}$ of source signals. The complex-valued mixing matrix $\mathbf{A}_f \in \mathbb{C}^{M \times N}$ contains frequency-wise mixing parameters related to the direction of arrival of sources. After estimating $\mathbf{S}_f$, the time domain source signals $\mathbf{S}$ can be obtained via inverse STFT.

Following previous works, we make the following assumptions of $\mathbf{A}$. For the instantaneous mixing model, we assume $\mathbf{A}$ can be estimated by another method. Thus, when $\mathbf{S}$ is estimated (by a method introduced in Section II-B), $\mathbf{A}$ is fixed. For the convolutive mixing model, $\mathbf{A}$ and $\mathbf{S}$ are jointly estimated (by a method introduced in Section II-C), that is, $\mathbf{A}$ is not fixed but alternatively optimized along with $\mathbf{S}$. Our contribution works with both approaches.

B. Two-step SCA

When the instantaneous mixing model is considered, SCA is usually achieved in two steps [1], [2]. The first step is to estimate the mixing matrix $\mathbf{A}$. The second step is to estimate source signals using the estimated mixing matrix. Under the underdetermined condition, the second step cannot be trivially accomplished by simple inversion. SCA reconstructs sources by maximizing posterior probability. It is often assumed that the prior distribution of the phase of the STFT coefficients of sources is uniform and the magnitude is independently Laplacian distributed with zero mean and unit variance, i.e.,

$$p(|s_n(q,f)|) = \frac{1}{2}\exp(-|s_n(q,f)|),\ \forall (q,f). \tag{3}$$

Accordingly, maximizing the posterior of sources results in an equality-constrained l_1 minimization problem

$$\begin{aligned} \min_{\mathbf{S}} \quad & \|\mathbf{S}\mathbf{\Psi}\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{A}\mathbf{S}, \end{aligned} \tag{4}$$

where $\mathbf{\Psi} \in \mathbb{C}^{T \times QF}$ is the STFT matrix, and $\|\cdot\|_1$ is the l_1 norm, defined as the entry-wise absolute sum. This problem has been solved in [16].

C. Iterative two-step SCA

When the convolutive mixing model is applied, following the state-of-the-art works of SCA [4], we optimize $\mathbf{A}_f$ and $\mathbf{X}_f$ jointly in each iteration, with the following objective function

$$\min_{\mathbf{A}_f, \mathbf{S}_f} \frac{1}{2}\|\mathbf{X}_f - \mathbf{A}_f\mathbf{S}_f\|_F^2 + \lambda\|\mathbf{S}_f\|_1 + \iota_C(\mathbf{A}_f), \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, λ is a trade-off parameter to balance the sparse regularization and the data fidelity. The term $\iota_C(\mathbf{A}_f)$ is used to ensure that each column of $\mathbf{A}_f$ has unit magnitude, with the following expression

$$\iota_C(\mathbf{A}_f) = \begin{cases} 0, & \text{if } \|\mathbf{a}_{f,n}\|_2 = 1,\ n = 1, 2, \cdots, N \\ +\infty, & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathbf{a}_{f,n}$ is the n-th column of $\mathbf{A}_f$. The problem (5) has been solved in [4].

III. The Proposed Method

A. Source Estimation by Weighted l_1 Minimization

The primary limitation of SCA is that it models all sources using the same distributions. In the proposed SSCA, we model sources using a family of Laplace distributions with zero means and variances varying over time and frequency, i.e.,

$$p(|s_n(q,f)|) = \frac{w_n(q,f)}{2}\exp(-w_n(q,f)|s_n(q,f)|) \tag{7}$$

where $w_n(q,f) > 0$ is a parameter/weight controlling the variance (the smaller $w_n(q,f)$, the larger the variance). Adopting this prior distribution, the optimization problem (4) is modified as

$$\begin{aligned} \min_{\mathbf{S}} \quad & \|\mathbf{S}\mathbf{\Psi}\|_{\mathbf{W},1} \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{A}\mathbf{S}, \end{aligned} \tag{8}$$

where $\mathbf{W} = [w_n(q,f)] \in \mathbb{R}_+^{N \times QF}$ is the weight matrix containing positive weight values, and $\|\cdot\|_{\mathbf{W},1}$ is the weighted l_1 norm, defined as the weighted absolute sum (i.e., for any matrix $\mathbf{Z}$, $\|\mathbf{Z}\|_{\mathbf{W},1} = \sum_{i,j} w_{ij}|z_{ij}|$). The optimization problem (5) is modified as

$$\min_{\mathbf{A}_f, \mathbf{S}_f} \frac{1}{2}\|\mathbf{X}_f - \mathbf{A}_f\mathbf{S}_f\|_F^2 + \lambda\|\mathbf{S}_f\|_{\mathbf{W}_f,1} + \iota_C(\mathbf{A}_f), \tag{9}$$

where $\mathbf{W}_f \in \mathbb{R}_+^{N \times Q}$ is the weight matrix for the frequency band f.

B. Weight Construction

An appropriate $\mathbf{W}$ depends on the value of $|s_n(q,f)|$. Two observations inspire a simple method to estimate $|s_n(q,f)|$. The first observation is that the mixing model (2) can be reformulated as a vector sum

$$\mathbf{x}(q,f) = \sum_n \mathbf{a}_{f,n} s_n(q,f), \tag{10}$$

where $\mathbf{x}(q,f) = [x_1(q,f), \cdots, x_M(q,f)]^T$ is a vector comprising M STFT coefficients of the mixed signals at TF point (q,f). It can be seen from (10) that $\mathbf{x}(q,f)$ is a linear combination of column vectors of the mixing matrix $\mathbf{A}_f$. The second observation is that most TF points are nearly contributed by a single source, a nature of speech called W-disjoint orthogonality, which was defined early [22] and re-utilized recently [23]. Combining these two observations, a relevant estimation of $|s_n(q,f)|$ can be obtained by projecting $\mathbf{x}(q,f)$ onto the column vector $\mathbf{a}_{f,n}$, i.e.,

$$\hat{c}_n(q,f) = |\mathbf{x}^T(q,f)\mathbf{a}_n| / (\mathbf{a}_n^T\mathbf{a}_n), \tag{11}$$

where $\hat{c}_n(q,f)$ denotes an estimate of $|s_n(q,f)|$ and we omit the frequency index of $\mathbf{a}_{f,n}$ for the sake of brevity. When the W-disjoint orthogonality holds well, such an estimator is accurate as explained below. When only the n_1-th source contributes energy at a TF point, $\hat{c}_{n_1}(q,f)$ is exactly equivalent to $|s_{n_1}(q,f)|$ and greater than the estimates of other sources, i.e., $\hat{c}_{n_1}(q,f) > \hat{c}_{n_2}(q,f), \forall n_2 \neq n_1$. The estimator $\hat{c}_n(q,f)$ remains robust when W-disjoint orthogonality does not strictly hold. If two sources contribute energy (say the n_1-th source and the n_2-th source) and $\|\mathbf{a}_{n_1}\|_2 = \|\mathbf{a}_{n_2}\|_2$, $\hat{c}_{n_1}(q,f) \geq \hat{c}_{n_2}(q,f)$ if $|s_{n_1}(q,f)| \geq |s_{n_2}(q,f)|$.

Assuming $|s_n(q,f)|$ follows the distribution (7), the maximum likelihood estimator (MLE) of $w_n(q,f)$ is $1/|s_n(q,f)|$. Compute the weight according to the MLE formula and get $w_n(q,f) = 1/(\hat{c}_n(q,f)+\sigma)$, where σ is introduced to avoid infinite weight in the case of $\hat{c}_n(q,f) = 0$. However, we found that such weight did not exhibit stable performance in practice, potentially due to a suboptimal choice of σ. To circumvent tedious tuning of σ, we propose to set $w_n(q,f)$ proportional to the sine of the angle between $\mathbf{x}(q,f)$ and $\mathbf{a}_n$:

$$\begin{aligned} w_n(q,f) &= \|\mathbf{a}_n\|_2 \sqrt{1 - \left(\frac{\hat{c}_n(q,f)\|\mathbf{a}_n\|_2}{\|\mathbf{x}(q,f)\|_2}\right)^2} \\ &= \|\mathbf{a}_n\|_2 \sqrt{1 - \left(\frac{|\mathbf{x}^T(q,f)\mathbf{a}_n|}{\|\mathbf{a}_n\|_2\|\mathbf{x}(q,f)\|_2}\right)^2}, \end{aligned} \tag{12}$$

where we multiply the sine with $\|\mathbf{a}_n\|_2$ to remove the contribution of the magnitude of $\mathbf{a}_n$ to the weight.

C. Optimization

1) Two-step SSCA: Define $f_1(\mathbf{S}) = \|\mathbf{S}\mathbf{\Psi}\|_{\mathbf{W},1}$, $f_2(\mathbf{S}) = 0$ if $\mathbf{X} = \mathbf{A}\mathbf{S}$, and $f_2(\mathbf{S}) = +\infty$, otherwise. The constrained problem (8) can be reformulated as an unconstrained problem

$$\min_{\mathbf{S}} f_1(\mathbf{S}) + f_2(\mathbf{S}), \tag{13}$$

which can be directly solved by the Douglas-Rachford (DR) algorithm [24]. The DR algorithm is presented in Algorithm 1.

Algorithm 1: The Douglas-Rachford Algorithm

Initialize: $\mathbf{Z}^{(0)} \in \mathbb{R}^{N\times T}$ and a thresholding parameter $\gamma > 0$
$k = 0$;
repeat
  $\mathbf{S}^{(k)} = \text{prox}_{\gamma f_2}(\mathbf{Z}^{(k)})$;
  $\mathbf{Y}^{(k)} = \text{prox}_{\gamma f_1}(2\mathbf{S}^{(k)} - \mathbf{Z}^{(k)})$;
  $\mathbf{Z}^{(k+1)} = \mathbf{Z}^{(k)} + \mathbf{Y}^{(k)} - \mathbf{S}^{(k)}$;
  $k = k + 1$;
until *convergence*;
return $\mathbf{S}^{(k)}$

In Algorithm 1, $\text{prox}_{\gamma f_i}$ is called the proximal operator associated with the function f_i, defined as

$$\text{prox}_{\gamma f_i}(\mathbf{Z}) := \underset{\mathbf{U}}{\text{argmin}}\, f_i(\mathbf{U}) + \frac{1}{2\gamma}\|\mathbf{U} - \mathbf{Z}\|_F. \tag{14}$$

In our case, the proximal operator of f_1 is given by

$$\text{prox}_{\gamma f_1}(\mathbf{Z}) = \text{prox}_{\gamma\|\cdot\|_{\mathbf{W},1}}(\mathbf{Z}\mathbf{\Psi})\mathbf{\Psi}^H, \tag{15}$$

with an entry-wise soft-thresholding function

$$\text{prox}_{\gamma\|\cdot\|_{\mathbf{W},1}}(\mathbf{Z}) = [\frac{z_{ij}}{|z_{ij}|}(|z_{ij}| - \gamma w_{ij})^+], \tag{16}$$

where $(\cdot)^+ = \max(\cdot, 0)$, and $(\cdot)^H$ is the conjugate transpose. The proximal operator of f_2 is derived as in [25]:

$$\text{prox}_{\gamma f_2}(\mathbf{Z}) = \mathbf{Z} + \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{X} - \mathbf{A}\mathbf{Z}). \tag{17}$$

2) Iterative two-step SSCA: To solve the problem (9), we generalize the N-Regu algorithm [4] and present it in Algorithm 2. When the weight matrix $\mathbf{W}$ is an all-ones matrix where all elements are equal to one, our algorithm reduces to the original N-Regu algorithm. The function $\mathcal{P}_C(\mathbf{A})$ is given by the column-wise normalization:

$$\mathbf{a}_n = \frac{\mathbf{a}_n}{\|\mathbf{a}_n\|_2},\ n = 1, 2, \cdots, N. \tag{18}$$

The function $\mathcal{W}$ is to calculate the weight matrix with entries defined by (12). Since $\mathbf{A}_f^{(k)}$ can not accurately estimate the true $\mathbf{A}_f$ in the early stage of iteration, we make $\mathbf{W}^{(k)}$ gradually approach the result of function $\mathcal{W}$ by using the factor k/k_{max} where k_{max} is the maximum number of iterations.

Algorithm 2: The Weighted N-Regu Algorithm

Initialize: $\mathbf{A}_f^{(0)} \in \mathbb{C}^{M\times N}$, $\mathbf{S}_f^{(0)} \in \mathbb{C}^{N\times Q}$, $\mathbf{W}_f^{(0)} \in \mathbb{R}_+^{N\times Q}$, and a trade-off parameter $\lambda \geq 0$
$L^{(0)} = \|\mathbf{A}_f^{(0)^H}\mathbf{A}_f^{(0)}\|_2$, $k = 0$;
repeat
  $\mathbf{V}^{(k)} = -\mathbf{A}_f^{(k)^H}(\mathbf{X}_f - \mathbf{A}_f^{(k)}\mathbf{S}_f^{(k)})$;
  $\mathbf{S}_f^{(k+1)} = \text{prox}_{\lambda/L^{(k)}\|\cdot\|_{\mathbf{W}_f^{(k)},1}}(\mathbf{S}_f^{(k)} - \frac{1}{L^{(k)}}\mathbf{V}^{(k)})$;
  $\mathbf{A}_f^{(k+1)} = \mathcal{P}_C(\mathbf{X}_f\mathbf{S}_f^{(k+1)^H})$;
  $L^{(k+1)} = \|\mathbf{A}_f^{(k+1)^H}\mathbf{A}_f^{(k+1)}\|_2$;
  $\mathbf{W}_f^{(k+1)} = \frac{k}{k_{\text{max}}}(\mathcal{W}(\mathbf{X}_f, \mathbf{A}_f^{(k+1)}) - \mathbf{W}_f^{(0)}) + \mathbf{W}_f^{(0)}$;
  $k = k + 1$;
until *convergence*;
return $\mathbf{S}^{(k)}$

IV. Simulation Experiments

We conducted two experiments to evaluate the proposed SSCA, implemented on Ubuntu 18.04.4 LTS operating system with an Intel Core i9-9900K CPU @ 3.60GHz X 16 and 62.7GiB memory. The first experiment used instantaneous mixtures of clean speech signals from the dataset of SiSEC2011 [26], which had a 16 KHz sampling rate and included male and female voices, mainly in English. Instantaneous mixtures were generated by mixing three to six speech signals ($3 \leq N \leq 6$) into two-channel mixtures ($M = 2$) using random mixing matrices. These matrices had independent, identically distributed Gaussian entries with zero mean and unit variance, which were then taken as absolute values and normalized to ensure each entry was positive and each column vector had a unit magnitude. The STFT window length was set to 1024 samples (i.e., 64 ms) with

TABLE I
UBSS RESULTS ON INSTANTANEOUS MIXING MIXTURES (BOLD NUMBERS INDICATE THE BEST PERFORMANCE AND THOSE NOT SIGNIFICANTLY DIFFERENT FROM IT, PAIRED T-TEST AT THE 5% SIGNIFICANCE LEVEL)

Method	# of sources	SDR (dB)	ESTOI	PESQ	Time (s)
SCA	3	9.78	0.72	2.29	21
	4	4.38	0.57	1.93	35
	5	0.25	0.45	1.63	52
	6	-0.92	0.40	**1.63**	69
RWSCA [16]	3	8.96	0.69	2.12	43
	4	3.32	0.53	1.69	75
	5	-1.33	0.37	1.28	110
	6	-2.60	0.33	1.18	140
SSCA (prop)	3	**13.74**	**0.84**	**2.80**	15
	4	**7.05**	**0.67**	**2.22**	22
	5	**2.73**	**0.54**	**1.83**	33
	6	**0.29**	**0.45**	**1.67**	41

a 50% overlap rate. The mixing matrix $\mathbf{A}$ was estimated by an existing method [27]. We compared the proposed SSCA against plain SCA and RWSCA proposed in [16]. Since tested methods are special cases of ours, we implemented them using Algorithm 1. SSCA used a weight matrix calculated by (12), the plain SCA used an all-ones matrix, and RWSCA used a weight matrix determined by a reweighting scheme. The parameter γ was set to 0.1 and the convergence condition was set to $\|\mathbf{S}^{(k)} - \mathbf{S}^{(k-1)}\|_F / \|\mathbf{S}^{(k)}\|_F \leq 0.01$. Table I displays the average values of SDR [28], extended short-time objective intelligibility (ESTOI) [29], perceptual evaluation of speech quality (PESQ) [30], and computation time over ten mixtures. SSCA achieved the best performance (up to 4 dB SDR higher than SCA) and converged faster than SCA (saving up to 40% of time). This is because weights in SSCA helped the optimization algorithm identify active sources in the early stage of iteration, reducing unnecessary optimization of inactive sources.

In the second experiment, we used real-world recordings from the SiSEC2011 *dev1* package. By combining genders and source directions, we obtained 32 two-channel mixtures of three sources with ambient reverb times of 130 ms and 250 ms. We compared the performance of the proposed SSCA optimized by Algorithm 2 with plain SCA [4] (a special case of Algorithm 2 where the weight matrix is an all-ones matrix), MNMF optimized by expectation maximization algorithm (MNMF_EM) [7], MNMF optimized by multiplicative update algorithm (MNMF_MU) [7], two state-of-the-art fast MNMF (FastMNMF1 [9] and FastMNMF2 [10]), and FCA [12]. For Algorithm 2, we set the parameter λ to 0.05, the maximum number of iterations $k_{\max}$ to 5e3, and initialized the weight matrix as an all-ones matrix, updating it every 100 iterations for stabilization. We solved the frequency permutation alignment problem using an existing method based on inter-frequency correlation [31]. For MNMF algorithms, the number of components per source was set to ten as suggested in [7]. Values of the rest of the parameters of benchmark methods are default. For all algorithms, we set the STFT window length to 2048 samples (i.e., 128 ms) with a 50% overlap rate. The results are shown in Table II. The proposed SSCA achieved the best SDR performance and was comparable to the best benchmarks (FCA and MNMF_EM) in ESTOI and PESQ while being the fastest among them.

TABLE II
UBSS RESULTS ON REAL-RECORDING MIXTURES (BOLD NUMBERS INDICATE THE BEST PERFORMANCE AND THOSE NOT SIGNIFICANTLY DIFFERENT FROM IT, PAIRED T-TEST AT THE 5% SIGNIFICANCE LEVEL)

Method	Reverb	SDR (dB)	ESTOI	PESQ	Time (s)
SCA [4]	130 ms	3.84	0.42	1.52	86
	250 ms	2.87	0.37	1.44	-
MNMF_MU [7]	130 ms	1.23	0.25	1.25	23
	250 ms	0.86	0.25	1.34	-
MNMF_EM [7]	130 ms	2.76	0.46	**2.02**	117
	250 ms	0.88	0.41	**1.76**	-
FastMNMF1 [9]	130 ms	1.92	0.36	1.60	20
	250 ms	1.30	0.33	1.59	-
FastMNMF2 [10]	130 ms	2.64	0.40	1.73	14
	250 ms	1.80	0.37	1.68	-
FCA [12]	130 ms	3.50	**0.49**	**2.03**	229
	250 ms	2.74	**0.45**	**1.86**	-
SSCA (prop)	130 ms	**4.72**	**0.48**	1.94	95
	250 ms	**3.47**	0.42	**1.84**	-
MF-FCA [14]	130 ms	5.69	0.62	2.31	4971
	250 ms	4.27	0.55	2.15	-

We added in Table II the performance of the state-of-the-art multi-frame FCA (MF-FCA) [14] as a reference, but it is important to note that this is not a fair comparison. Because MF-FCA considers time correlation between frames and adopts multi-frame input for collaborative inference while SSCA is a single-frame model. However, SSCA reduced the performance gap with MF-FCA to less than 1 dB in terms of SDR, and its calculation time is only one-50th of MF-FCA's.

V. CONCLUSION

In this paper, we have presented the novel sparse spatial component analysis (SSCA) for underdetermined blind speech separation (UBSS). The SSCA method optimizes the weighting of the sources based on their directions making it able to remove the identical source distribution requirement without prior knowledge of the mixing process. Furthermore, our SSCA model assumes that the speech follows non-identical Laplace distributions which is more appropriate than the assumption of non-identical Gaussian distributions adopted by the state-of-the-art analytic UBSS methods such as multi-channel non-negative matrix factorization (MNMF) and full-rank covariance analysis (FCA). The appropriateness is supported by superior performance in simulated experiments using instantaneous mixtures and real-world recordings. The proposed method is a single-frame model that could potentially be upgraded to a multi-frame model offering even better performance than multi-frame FCA.

ACKNOWLEDGMENT

This study is partially supported by the Hong Kong Grants Council through projects 16209823 and C5052-23G.

REFERENCES

[1] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[2] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–12, 2006.

[3] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 430–437.

[4] F. Feng and M. Kowalski, "Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrowband approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 442–456, 2018.

[5] J. Yang, Y. Guo, Z. Yang, and S. Xie, "Under-determined convolutive blind source separation combining density-based clustering and sparse reconstruction in time-frequency domain," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 8, pp. 3015–3027, 2019.

[6] Y. Xie, T. Zou, J. Yang, W. Sun, and S. Xie, "Sa-ucbss: Sparsity-based adaptive underdetermined convolutive blind source separation," *Knowledge-Based Systems*, vol. 300, p. 112224, 2024.

[7] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 3, pp. 550–563, 2009.

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

[9] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[10] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2610–2625, 2020.

[11] M. Fontaine, K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Generalized fast multichannel nonnegative matrix factorization based on gaussian scale mixtures for blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1734–1748, 2022.

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

[13] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel wiener filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1950–1965, 2021.

[14] H. Sawada, R. Ikeshita, K. Kinoshita, and T. Nakatani, "Multi-frame full-rank spatial covariance analysis for underdetermined blind source separation and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[15] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[16] S. Arberet, P. Vandergheynst, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.

[17] S. Arberet and P. Vandergheynst, "Reverberant audio source separation via sparse and low-rank modeling," *IEEE Signal Processing Letters*, vol. 21, no. 4, pp. 404–408, 2014.

[18] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 541–545.

[19] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, pp. 877–905, 2008.

[20] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *IEEE Transactions on Neural networks*, vol. 19, no. 3, pp. 475–492, 2008.

[21] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[22] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–529.

[23] Y. He, H. Wang, Q. Chen, and R. H. So, "Harvesting partially-disjoint time-frequency information for improving degenerate unmixing estimation technique," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 506–510.

[24] P. L. Combettes and J.-C. Pesquet, "A douglas–rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, 2007.

[25] M.-J. Fadili and J.-L. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *2009 16th IEEE International conference on image processing (ICIP)*. IEEE, 2009, pp. 1461–1464.

[26] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (sisec2011): - audio source separation -," in *Latent Variable Analysis and Signal Separation*, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 414–422.

[27] V. G. Reju, S. N. Koh, and Y. Soon, "An algorithm for mixing matrix estimation in instantaneous blind source separation," *Signal Processing*, vol. 89, no. 9, pp. 1762–1773, 2009.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[29] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[31] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2010.