Circuits, Systems, and Signal Processing (2023) 42:5338–5357
https://doi.org/10.1007/s00034-023-02350-1

Single-Channel Blind Source Separation using Adaptive Mode Separation-Based Wavelet Transform and Density-Based Clustering with Sparse Reconstruction

Mina Kemiha[1] · **Abdellah Kacha**[1]

Received: 16 August 2022 / Revised: 5 March 2023 / Accepted: 6 March 2023 /
Published online: 19 April 2023

Abstract

In this paper, the problem of single-channel blind source separation (SCBSS) is addressed using a novel approach that combines the adaptive mode separation-based wavelet transform (AMSWT) and the density-based clustering with sparse reconstruction. The proposed method is performed in the time–frequency domain and in a reverberant environment. First, using the Fourier transform, the amplitude spectrum of the observed mixture signal is obtained. Then, using variational scaling and wavelet functions, the AMSWT is used to adaptively extract spectral intrinsic components (SICs). To obtain a better time–frequency resolution, the AMSWT is applied to each mode. Thus, the SCBSS problem is transformed into a non-underdetermined. Then, for each frequency bin, the density-based clustering, reformulated to the eigenvector clustering problem, is performed to estimate the mixing matrix. Finally, sparse reconstruction is used to reconstruct the sources. The proposed method is evaluated using objective measures of separation quality. A computational complexity evaluation based on time consumption is also performed. Simulation results show that the proposed method is very effective for solving the SCBSS problem and provides better separation performances than the reference methods. However, the proposed method is computationally expensive.

Keywords Single-channel blind source separation · Adaptive mode separation-based wavelet transform · Eigenvector clustering · Sparse reconstruction

✉ Mina Kemiha
mina.kemiha@univ-jijel.dz

Abdellah Kacha
akacha@ulb.ac.be

[1] Radiation Physics and Applications Laboratory, University of Jijel, BP 98 Ouled Aissa, 18000 Jijel, Algeria

1 Introduction

Blind source separation (BSS) aims to separate the source signals from the mixed signals without any information. The BSS has been applied in many areas such as medical imaging and engineering [2, 44], astrophysics [40], image processing [11], geophysical data processing [35], speech processing [21–23, 32], detection and radar localization [28], communication systems [43], automatic transcription of speech [13], musical instrument identification [34], mechanical flaw detection [18], multichannel telecommunications [38], multispectral astronomical imaging [33] and speech recognition [4].

In the literature, the BSS methods are classified as being linear and nonlinear, instantaneous and convolutive, and overcomplete and underdetermined. The convolutive mixture model of BSS is an effective way to represent the speech signal mixing mechanism in a reverberant environment [7, 30]. The BSS problem can be formulated either in the time domain or in the frequency domain. The BSS can be also treated in the joint time–frequency (TF) domain where computationally efficient BSS algorithms are available.

In most situations and for many practical uses, only one-channel recording of mixture signals is available. This particular instance of the underdetermined source separation problem called single-channel source separation (SCSS) has been the subject of many studies. To address the single-channel audio source separation problem, numerous strategies have been proposed in the literature [12]. In [14], the authors attempted to combine the maximum-likelihood estimation and nonnegative matrix factorization (NMF) based on the Itakura–Saito divergence measurement. The short-time Fourier transform (STFT) representation of the observed single-channel signal has been subjected to the NMF-based approach in [39]. The method requires the use of extra training data. A combination of the empirical mode decomposition (EMD) and independent component analysis (ICA), as well as wavelet transformations, has been suggested in [31]. However, the wavelet transform needs some specified basis functions to represent a signal, and there is no rigorous mathematical theory underpinning the EMD or its improved algorithms [20]. The bark scale aligned wavelet packet decomposition has been introduced in [26] where the separation step has been performed using the Gaussian mixture model (GMM), which was employed before the Fourier transform. In [45], the authors proposed the variational mode decomposition (VMD) method to solve the SCBSS problem. The separation process is performed using joint approximate diagonalization based on the fourth-order cumulant matrices. In [36], the authors presented a novel method for SCBSS in a noisy environment. The method is based on selecting the TF units of signal presence and computing the mixture spectral amplitude. The separation process is performed using TF masking. In [25], an adaptive signal separation has been proposed. The method uses a time-varying parameter that adapts locally to instantaneous frequencies and a linear chirp (linear frequency modulation) to model the signal components. The single-channel technique has been explored for muscle artifact removal from multichannel EEG [6].

The classic TF representation is computed using the STFT. The STFT does not reflect the time-varying information. Moreover, it yields a time–frequency representation with only uniform time and frequency resolution. A new adaptive mode separation-based wavelet transform (AMSWT) has been proposed in [24] based on [10, 16]. The AMSWT method involves solving a recursive optimization problem to adaptively extract spectral intrinsic components (SICs). The limited support of each spectral mode is implemented to establish the spectral boundaries for wavelet bank configuration. Then, the spectral boundaries of the created wavelet bank configuration are used to highlight the spectral information. The AMSWT strategy is fully adaptive in the sense that one does not require prior knowledge.

In [41], a new method to solve the underdetermined BSS problem for convolutive mixture has been proposed. The method operates in the time–frequency domain, and it combines the density-based grouping and sparse source reconstruction. The density-based clustering is introduced to estimate the mixing matrix, which is converted to an eigenvector clustering issue. The rank-one structure of the local covariance matrices of the mixture TF vectors is first used to extract the eigenvectors. By combining weight clustering and density-based clustering, the eigenvectors are subsequently grouped and tweaked to provide an approximated mixing matrix. The source reconstruction is transformed into a l_p-norm minimization using an iterative Lagrange multiplier method. The Lagrange multiplier used to solve optimization problems under constraints aims to enforce the constraint, while the quadratic penalty improves the convergence. In the iterative formula, both the primal and dual variables are updated iteratively.

In this paper, a new method to solve the SCBSS problem is proposed. The method combines the AMSWT [24] and density-based clustering with the sparse reconstruction method introduced in [41]. The method is performed in three stages: (i) The amplitude spectrum of the observed mixture signal is obtained using STFT. The convolution in the time domain can be approximated by a multiplication in the STFT domain. (ii) A better TF resolution is obtained using the variational scaling and wavelet functions, which are applied to the spectral intrinsic components (SICs) extracted adaptively using the AMSWT. By creating virtual multichannel signals of the TF representation, the underdetermined single-channel problem is transformed to a non-underdetermined problem. (iii) For each TF representation and each frequency bin, the density-based clustering, which is converted to an eigenvector clustering problem, and the sparse reconstruction, which is converted to a minimization problem, are, respectively, performed to estimate the mixing matrix and sources.

The BSSeval toolbox [15] is used to evaluate the proposed method's performance. The evaluation is performed in terms of many criteria such as source-to-distortion ratio (SDR), source-to-artifact ratio (SAR) and source-to-interference ratio (SIR). The proposed method is compared to the variational mode decomposition (VMD) method [45], adaptive spectrum amplitude estimator and masking method [36] and the nonnegative tensor factorization of modulation spectrograms method [3].

The following sections make up the remaining content. The SCBSS problem formulation is presented in Sect. 2. The AMSWT method, the density-based clustering method and the source reconstruction are the main focus of Sect. 3. Simulation results are presented in Sect. 4. Finally, conclusions and discussions are given in Sect. 5.

2 Convolutive Mixture Model

Let $\mathbf{x}(t) = [x_1(t), .., x_M(t)]^T$ be a vector of M observed sources obtained via the mixing of N independent sources $\mathbf{s}(t) = [s_1(t), .., s_N(t)]^T$. The BSS problem aims to estimate the N sources from the M mixtures. The convolutive mixture occurs through the propagation of the sound through space and multiple paths caused by reflections from different objects, especially in rooms and closed environments. The convolutive mixture is modeled as follows:

$$x_j(t) = \sum_{i=1}^{N} \sum_{k=0}^{K-1} h_{ji}(k) s_i(t-k),\ j = 1, \ldots, M \tag{1}$$

The matrix form is given as:

$$\boldsymbol{x}(t) = \boldsymbol{H} * \boldsymbol{s}(t) = \sum_{k=0}^{K-1} \boldsymbol{H}_k \boldsymbol{s}(t-k) \tag{2}$$

where h_{ji} denotes the impulse response from source i to sensor j, and $\boldsymbol{H}$ is an MxN matrix that contains the kth filter coefficients.

In most cases and for many practical purposes, only one-channel recording is accessible. Numerous studies have examined this instance known as single-channel source separation. In this case, $M = 1$. The convolutive SCBSS in the time–frequency domain is described as follows:

$$\boldsymbol{X}(t, f) = \sum_{i=1}^{N} X_i(t, f) \tag{3}$$

where $X_i(t,f)$ is the STFT of $x_i(t)$.

The conventional source separation techniques are ineffective in this scenario. The issue in SCBSS might be viewed as a single observation combined with numerous unidentified sources.

3 Single-Channel Blind Source Separation Method

The different steps of the proposed method for single-channel blind source separation are summarized by the flowchart shown in Fig. 1.

The spectrum of the observed signal is obtained by the STFT. The convolution in the time domain is transformed into a multiplication in the STFT domain. The AMSWT approach is used to obtain an optimal spectral mode separation. Thus, the SCBSS problem is transformed into a non-underdetermined problem by establishing virtual multichannel signals of the TF representation of the observed signals. Then, the M time–frequency representations of the mixture are divided into Q nonoverlapping blocks.

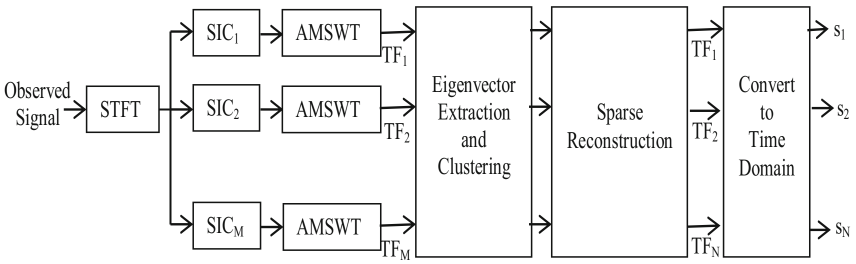


Fig. 1 Proposed method for single-channel blind source separation.

As a preprocessing step at the mixing matrix estimation stage, the TF representation of the observed signal is whitened for each frequency bin $\mathbf{x}_d$. The whitening process is performed using the eigenvector matrix $\mathbf{U}_\mathbf{x}$ and the eigenvalue matrix $\mathbf{\Sigma}_\mathbf{x}$ of $E(\mathbf{x}_d{\mathbf{x}_d}^H)$, and it is expressed by ${\mathbf{x}_d}^w = \mathbf{\Sigma}_\mathbf{x}^{-1/2}\mathbf{U}_\mathbf{x}^H\mathbf{x}_d$.

The estimation of the mixing matrix is reformulated into an eigenvector clustering issue. The ambiguity of scaling is solved by rescaling the estimated mixing matrix by restricting the first row. The order of the reconstructed sources is aligned, by grouping the nearby source TF vectors based on their correlation, in terms of power ratio, to resolve the permutation ambiguity [10].

The post-processing stage involves de-whitening the predicted mixing matrix by $\widehat{\mathbf{H}} = \mathbf{U}_\mathrm{x}\mathbf{\Sigma}_\mathbf{x}^{1/2}\widetilde{\mathbf{H}}$**.** Then, the source reconstruction is reformulated into a sparse minimization problem, whose solution was achieved using an initialization-corrected iterative Lagrange multiplier approach.

Finally, the estimated sources are obtained in the TF domain, which are transformed into the time domain using the modified method proposed in [27].

3.1 Adaptive Mode Separation-Based Wavelet Transform

The STFT is a TF representation, which has an even bandwidth distribution across all frequency channels and suffers from the TF resolution limitation due to the fixed window size. The speech signal is substantially nonperiodic and nonstationary. Therefore, the use of the STFT will result in mistakes, particularly when complex transitory phenomena like voice mixing occur in the signal.

The AMSWT performs a time–frequency analysis using variational scaling and wavelet functions. The method is built on the alternating direction method of multiplier (ADMM) solver [19], which then defines a bank of variational scaling functions and wavelets depending on the spectral boundaries that have been defined. As a result, the approximation coefficients are obtained as the inner product of the analyzed signal x with variational scaling function. The inner product of the analyzed signal x with variational wavelets yields the detail coefficients, which are expressed as:

$$W_x(t,0) = \langle x, \varnothing_1 \rangle = \int x(\tau)\overline{\varnothing}_1(\tau - t)d\tau \tag{4}$$

and

$$W_x(t,k) = \langle x, \psi_k \rangle = \int x(\tau)\overline{\psi}_k(\tau - t)d\tau \tag{5}$$

where x is the input signal.

In [24], under the amplitude-modulated and frequency-modulated (AM-FM) assumption, the intrinsic modes $u(t)$ have distinguishable features in the frequency domain. Using the ADMM solver, the spectral modes can be adaptively obtained similarly to the intrinsic mode functions (IMFs) extraction, to estimate compact modes:

$$\underset{u_k,\omega_k}{min} \begin{Bmatrix} \sum_k \|\partial_t\Big[\Big(\delta(t) + \frac{j}{\pi t}\Big) * u_k(t)\Big]e^{j\omega_k t}\|_2^2 \\ s.t. \sum_K u_k = x(t) \end{Bmatrix} \tag{6}$$

where $x(t)$ is the signal to be decomposed under the constraint that over all modes' summation should be equal to the input signal, $\delta(.)$ is the Dirac impulse and $\Big(\delta(t) + \frac{j}{\pi t}\Big) * u_k(t)$ denotes the original data and its Hilbert transform. The variables u_k, ω_k and k denote the modes, their central frequencies and the mode number, respectively.

The spectral segmentation boundary number can be determined empirically as follows:

$$\widetilde{K} = min\big\{n \in \mathbb{Z}^+ | n \geq 2\rho \ln N\big\} \tag{7}$$

where N is the signal length and ρ is the scaling exponent determined by the detrended fluctuation analysis (DFA).

According to [24], (6) is solved using a quadratic penalty term; the parameter λ denotes the Lagrangian multiplier used to render the problem unconstrained

$$\begin{aligned} L\left(u_k, \omega_k, \lambda\right) = \eta \sum_k \left\| \delta_t \left[\left(\delta\left(t\right) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{j\omega_k t} \right\|_2^2 \\ + \langle \lambda, x - \sum_K u_k \rangle + \left\| x - \sum_K u_k \right\|_2^2 \end{aligned} \tag{8}$$

Therefore, u_k is determined recursively as

$$\widehat{u}_k^{n+1}(\omega) = \frac{X(\omega) - \sum_{i \neq j} \widehat{u}_i^{n+1}(\omega) + \frac{\widehat{\lambda}^n}{2}}{1 + 2\eta(\omega - \omega_k^n)^2} \tag{9}$$

where $X(\omega)$, $\widehat{u_i}(\omega)$ and $\widehat{\lambda}(\omega)$ denote, respectively, the Fourier transforms of the input signal $x(t)$, the mode function $u_i(t)$ and $\lambda(t)$. η denotes the balancing parameter of the data-fidelity constraint. The center frequencies ω_k^{n+1} are updated as the center of gravity of the corresponding mode's power spectrum using the following equation

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left|\widehat{u}_k^{n+1}(\omega)\right|^2 d\omega}{\int_0^\infty \left|\widehat{u}_k^{n+1}(\omega)\right|^2 d\omega} \tag{10}$$

As a result, rather than using a predefined wavelet bank, we create adaptive wavelet banks based on spectral modes and their corresponding center frequencies, which represent the intrinsic components.

Authors in [24] used the mode bandwidth and central frequencies to define the boundaries between each mode; however, in the literature, some authors just used the average of the two central frequencies as spectral boundary, which ignores the spectrum distribution.

Consider the kth mode with an average frequency ω_k and a spectral bandwidth β_k. Then, the boundary $\mathbf{\Omega}_k$ between the kth mode and the $(k+1)th$ mode is given by

$$\mathbf{\Omega}_k = \frac{\omega_k + \frac{\beta_k}{2} + \omega_{k+1} - \frac{\beta_{k+1}}{2}}{2} \tag{11}$$

where $\mathbf{\Omega}_0 = 0$ and $\mathbf{\Omega}_k = \pi$.

The authors apply the same principle used in the production of both Littlewood—Paley and Meyer's wavelets [9] for variational scaling functions and wavelets based on spectral boundaries. $\widehat{\varnothing}_k$ and $\widehat{\psi}_k$ are, respectively, defined by the following equation, with γ is the parameter that ensures no overlap between the two consecutive transitions.

$$\widehat{\varnothing}_k = \begin{cases} 1, \omega \le (1-\gamma)\mathbf{\Omega}_k \\ \cos\left(\frac{\pi}{2}\alpha(\gamma, \mathbf{\Omega}_k)\right), (1-\gamma)\mathbf{\Omega}_k \le \omega \le (1+\gamma)\mathbf{\Omega}_k \\ 0 \; otherwise \end{cases} \tag{12}$$

and

$$\widehat{\psi}_k = \begin{cases} 1, (1+\gamma)\mathbf{\Omega}_k \le \omega \le (1-\gamma)\mathbf{\Omega}_{k+1} \\ \cos\left(\frac{\pi}{2}\alpha(\gamma, \mathbf{\Omega}_{k+1})\right), (1-\lambda)\mathbf{\Omega}_{k+1} \le \omega \le (1+\lambda)\mathbf{\Omega}_{k+1} \\ \sin\left(\frac{\pi}{2}\alpha(\gamma, \mathbf{\Omega}_k)\right), (1-\lambda)\mathbf{\Omega}_k \le \omega \le (1+\lambda)\mathbf{\Omega}_k \\ 0 \; otherwise \end{cases} \tag{13}$$

where $\alpha(\gamma, \mathbf{\Omega}_k) = \beta\{\left(\frac{1}{2\gamma\mathbf{\Omega}_k}\right)\left[|\omega| - (1-\gamma)\mathbf{\Omega}_k\right]\}]$ and $\beta(\nu)$ is an arbitrary function defined as follows:

$$\beta(\nu) = \begin{cases} 0, \nu \le 0 \\ 1, \nu > 1 \\ \beta(\nu) + \beta(1-\nu) = 1, 0 < \nu < 1 \end{cases} \tag{14}$$

The adaptive mode separation-based wavelet transform algorithm is summarized as follows:

Step 1: Time–frequency representation
Input: Observed mixture.

- Using the Fourier transform, obtain the amplitude spectrum signal.
- Obtain the appropriate spectrum spectral modes (segments). Execute the first inner loop and the second inner loop to update u_k according to (9), and update ω_k according to (10), respectively
- Compute the proper spectral boundaries using (11). Then, using (12) and (13), the bank of variational scaling functions and wavelets based on the spectral boundaries is defined.
- Finally, using (4) and (5), respectively, apply variational scaling and wavelet functions to each mode to obtain the time–frequency distribution.

Output: time–frequency distribution of the observed mixture.

3.2 Density-Based Clustering

In [41], the authors introduced the eigenvector clustering as an alternative to estimate the mixing matrix. The eigenvector clustering is based on two factors, which are the local density ρ_q and the minimum distance δ_q that may be taken between the point q and any additional points with a higher density. They are given, respectively, by

$$\rho_q \triangleq \sum_{k \neq q} e^{-\frac{\upsilon_{qk}^2}{\tau_c^2}} \tag{15}$$

and

$$\delta_q = \min_{k:\rho_k > \rho_q} (\upsilon_{qk}) \tag{16}$$

where the region for each data point is defined by a cut-off distance τ_c, and υ_{qk} are the elements of the similarity matrix:

$$\boldsymbol{V} \triangleq \begin{bmatrix} \upsilon_{11} & \cdots & \upsilon_{1Q} \\ \vdots & \ddots & \vdots \\ \upsilon_{Q1} & \cdots & \upsilon_{QQ} \end{bmatrix} \tag{17}$$

From the eigenvectors $\boldsymbol{A}$ whose elements are $\boldsymbol{a}_q$, the similarity matrix $\boldsymbol{V}$ is generated as follows:$\upsilon_{qk} = \|\boldsymbol{a}_q - (\boldsymbol{a}_q^H \boldsymbol{a}_k)\|_F^2$, where $\|.\|_F$ denotes Frobenius norm [29] expressed as follows:

$$\begin{aligned}\left\| a_q - (a_q^H a_k) \right\|_F^2 &= \sum_{q=1}^{Q} \sum_{k=1}^{Q} \left| a_q - (a_q^H a_k) \right|^2 \\ &= \text{trace}\left(\left(a_q - \left(a_q^H a_k \right) \right) \left(a_q - (a_q^H a_k) \right)^H \right) \end{aligned} \tag{18}$$

where $(.)^H$ denotes the conjugate transpose.

The eigenvector extraction is based on the local covariance matrix $\boldsymbol{R}_q^{\mathrm{X}} = \sum_{i=1}^{N} \sigma_{i,q}^2 h_i h_i^H$ where h_i is called the steering vector representing each direction of the mixing matrix. According to [41], there is at least one subblock indexed as q_i for which the associated local covariance $\boldsymbol{R}_{q_i}^{\mathrm{X}}$ has roughly a rank-one structure. This condition is exploited in [16] where the authors applied the eigenvalue decomposition (EVD) to the local covariance matrix $\boldsymbol{R}_q^{\mathrm{X}}$, which results in the following equation:

$$\boldsymbol{R}_q^{\mathrm{X}} = \boldsymbol{U}_q \boldsymbol{\Sigma}_q \boldsymbol{U}_q^H \tag{19}$$

where $\boldsymbol{U}_q$ and $\boldsymbol{\Sigma}_q$ denote the eigenvector matrix and eigenvalue matrix, respectively.

The extracted vector denoted $\mathbf{a}_q$ corresponds to the largest eigenvalue of $\boldsymbol{\Sigma}_q$ and also represents the first eigenvector in $\boldsymbol{U}_q$. To obtain the eigenvector matrix $\mathbf{A} \triangleq [\mathbf{a}_1, \ldots, \mathbf{a}_Q]$, the eigenvector extraction is done subblock wisely.

According to [41], the global maximum in the density indexed as q^* has a minimum distance δ_{q^*} defined as follows:$\delta_{q^*} = \underset{q,k=1,\ldots,Q}{max} (\upsilon_{qk})$ if $\rho_{q^*} = \underset{q=1,\ldots,Q}{max} (\rho_q)$(20)

The two components are multiplied together to provide the following score:

$$\gamma_q = \rho_q \times \delta_q \tag{21}$$

To get $\{\gamma_q\}_{q=1}^{Q}$, the scores from (20) are applied to all of the subblocks. The obtained scores are then arranged in a decreasing order. As a result, the eigenvectors with the greatest N scores define the clusters, which are denoted by $\mathbf{C} \triangleq [\boldsymbol{c}_1, ..., \boldsymbol{c}_N]$.

As mentioned in [41], it would be difficult to cluster the eigenvectors using solely the density-based strategy described above. To address this issue, a weight clustering approach to further tune the projected clusters is used [42]. The procedure of the weighted eigenvector clustering can be summarized in three steps.

First, the eigenvector is weighted by a kernel function defined as follows:

$$\boldsymbol{b}_{qk} \triangleq e^{\omega_{qk}^2/\tau_0^2} \boldsymbol{a}_q \tag{22}$$

where $k = 1, .., N$ and $\omega_{qk} = \|\boldsymbol{a}_q - \left(\boldsymbol{a}_q{}^H \boldsymbol{c}_k\right) \boldsymbol{c}_k\|_F^2$.

Then, the weighted covariance matrix is created as:

$$\boldsymbol{R}_k^{\mathrm{b}} = \sum_{q=1}^{Q} \boldsymbol{b}_{qk}\boldsymbol{b}_{qk}{}^{H} \tag{23}$$

Finally, the EVD is applied to the weighted covariance matrix $\boldsymbol{R}_k^{b}$:

$$\boldsymbol{R}_k^{\mathrm{b}} = \boldsymbol{U}_{qk}\boldsymbol{\Sigma}_{qk}\boldsymbol{U}_{qk}{}^{H} \tag{24}$$

As an updated of cluster $\boldsymbol{c}_k$ where $k = 1, ..., N$, the eigenvector that corresponds to the largest eigenvalue from the equation (24) is extracted.

The mixing matrix estimation algorithm is summarized as follows:

```
Step 2 : Mixing matrix estimation
Input : X which represents the TF representation of the observed signal whose elements $\mathbf{x}_d$.
    For each block q ∈ Q do
        • Calculate the local covariance matrix of $\mathbf{R}_{\mathrm{q}}^{\mathrm{X}}$ using $\widehat{\mathbf{R}}_{f,q}^{\mathrm{X}} = \frac{1}{p}\sum_{d=q(P-1)+1}^{qP} \mathbf{x}_{f,d}\mathbf{x}_{f,d}^{H}$
        • Construct the eigenvector matrix A using (19).
    End
        • Using the eigenvector matrix A, compute the similarity matrix defined by (17)
    For each block q ∈ Q do
        • Calculate the local density $\rho_{\mathrm{q}}$ and the minimum distance $\delta_{\mathrm{q}}$ and the score $\gamma_{\mathrm{q}}$ using (15), (16) and (21), respectively
    End
        • Calculate $\delta_{\mathrm{q}*}$ using (20), then, obtain the score sequence $\Upsilon = [\gamma_1, \ldots, \gamma_Q]$.
        • To obtain the score sequence $\Upsilon$, record the eigenvector matrix with the same permutation of decreasing alignment. So, to get the estimated clusters $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$, truncate the first $N$ reordered eigenvectors.
    For $k = 1$ to $N$ do
    For each subblock q ∈ Q do
        • Calculate the weighted eigenvector $\mathbf{b}_{\mathrm{qk}}$ using $\mathbf{a}_{\mathrm{q}}$ and $\mathbf{c}_{\mathrm{k}}$, then calculate $\mathbf{R}_{\mathrm{qk}}^{\mathrm{b}}$ using (22) and (23), respectively.
        • Calculate $\widetilde{\mathbf{h}}_{\mathrm{k}}$ using (24)
    End
    End
Output: Estimated mixing matrix $\widehat{\mathbf{H}}$ .
```

3.3 Source Reconstruction

In [41], the sparsity-based method has been introduced as an alternative to reconstruct the estimated source signal using a l_p-norm-based minimization measurement. (The convergence is guaranteed for $0 < p < 1$.) The method consists of converting the source reconstruction problem to a sparse reconstruction minimization problem. A designed iterative Lagrange multiplier approach with an appropriate initialization procedure is used to solve this minimization problem.

The source reconstruction is performed to find the sparsest term of s_d. For this, the maximum posterior likelihood of s_d is expressed as:

$$\begin{aligned} &\underset{s_d}{max} \prod_{i=1}^{N} P\left(\left|s_{i,d}\right|\right) \\ &s.t.\mathbf{x}_d = \widehat{\mathbf{H}} s_d \end{aligned} \tag{25}$$

where the complex-valued super-Gaussian distribution $P\left(\left|s_{i,d}\right|\right)$ is given by:

$$P\left(\left|s_{i,d}\right|\right) = p\frac{\gamma^{1/p}}{\Gamma(\frac{1}{p})}e^{-|s_{i,d}|^p} \tag{26}$$

where p and γ control the shape and variance of the probability function, Γ denotes the gamma function and $\widehat{\mathbf{H}}$ represents the estimated mixing matrix.

The problem returns to solve the equivalent optimization problem given as follows:

$$\begin{aligned} &\underset{s_d}{min} \sum_{i=1}^{N} |s_{i,d}|^p \\ &s.t.\mathbf{x}_d = \widehat{\mathbf{H}} s_d \end{aligned} \tag{27}$$

The Lagrange multiplier method is used to solve the optimization problem. Hence, the problem is reformulated to an unconstrained optimization problem as follows:

$$\underset{s_d,\alpha}{min} \mathcal{F}(s_d, \alpha) \triangleq \sum_{i=1}^{N} |s_{i,d}|^p + \alpha^H(\mathbf{x}_d - \widehat{\mathbf{H}} s_d) \tag{28}$$

where α is the Lagrange multiplier.

The implicit solution of the problem is given as follows:

$$s_d = \Psi^{-1}(s_d)\widehat{\mathbf{H}}^H\left(\widehat{\mathbf{H}}\Psi^{-1}(s_d)\widehat{\mathbf{H}}^H\right)^{-1}\mathbf{x}_d \tag{29}$$

where $\Psi^{-1}(s_d) \triangleq \begin{bmatrix} |s_{1,d}|^{2-p} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & |s_{N,d}|^{2-p} \end{bmatrix}$

The iterative scheme to obtain the solution s_d is given as follows:

$$\widehat{s}_d^{(iter+1)} = \begin{cases} \Psi^{-1}\left(\widehat{s}_d^{(\text{iter})}\right)\widehat{\mathbf{H}}^H\left(\widehat{\mathbf{H}}\Psi^{-1}\left(\widehat{s}_d^{(\text{iter})}\right)\widehat{\mathbf{H}}^H\right)^{-1}\mathbf{x}_d if \|\widehat{s}_d^{(\text{iter})}\|_0 \geq M \\ \Psi^{-1}\left(\widehat{s}_d^{(\text{iter})}\right)\widehat{\mathbf{H}}^H\left(\widehat{\mathbf{H}}(\Psi^{-1}\left(\widehat{s}_d^{(\text{iter})}\right) + \epsilon\mathbf{I})^{-1}\widehat{\mathbf{H}}^H\right)^{-1}\mathbf{x}_d elseif \|\widehat{s}_d^{(\text{iter})}\| < M \end{cases} \tag{30}$$

The source reconstruction algorithm is summarized as follows:

```
Step 3: Estimation of the TF representation of the sources
Input: Time–frequency representation of the observed signal denoted $\mathbf{X}$ whose elements $\mathbf{x}_d$ and estimated mixing matrix $\widehat{\mathbf{H}}$
    For each frequency bin do
      • Initialize the sources as $\widehat{s}_d^{(0)} = \sum_{j=1}^{C_N^M} \omega_j \mathrm{y}_{j,d}$
      Repeat
      • Update $\widehat{s}_d^{(\text{iter})}$ using (30)
      • $iter = iter + 1$
      Until $\|\widehat{s}_d^{(\text{iter})}\|_p^{\ p} - \|\widehat{s}_d^{(\text{iter+1})}\|_p^{\ p}$ is less than a given threshold.
      End
      Aware that $\|\widehat{s}_d^{(\text{iter})}\|_p^{\ p} \triangleq \sum_{i=1}^{N} |s^{(\text{iter})}{}_{i,d}|^p$ .
Output: time–frequency representation of the estimated sources.
```

For each frequency bin d, since the iterative method computes successive approximations to the solution of the problem, the stopping criterion minimizes the iterative absolute error. The tolerance or threshold of the stopping criteria is determined to guarantee the best algorithm performance without resulting in a high computing time.

4 Simulation Results

To investigate the effectiveness of the proposed method, numerical simulations have been performed in a reverberant environment. The TIMIT database [37] and NOIZEUS database [1] were used to build the speech dataset, which was chosen at random (available online). The sampling rate of the speech signals is $f_s = 16$kHz, and the speakers might be either female or male. Using the technique outlined in [17], the propagation environment is simulated as a reverberant room shown in Fig. 2.

The room impulse response from the source i to the sensor is illustrated in Fig. 3. By adjusting the reverberant time, a variety of convolutive mixed signals can be produced. It is crucial to evaluate the transmission duration of the signal decay to 60 dB to reflect the room reverberation.

As an illustration, let the three sources shown in Fig. 4a. The three sources are convolutedly mixed in the virtual room shown in Fig. 1 using the room impulse responses shown in Fig. 2. The observed single-channel signal is shown in Fig. 4b. Figure 4c shows a frame of 1024 sample length of the observed signal. The obtained modes are shown in Fig. 4d. For this example, the decomposition of the observed signal results

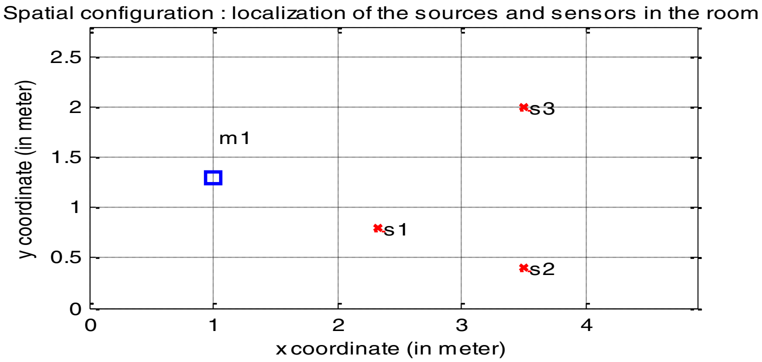


Fig. 2 Sources—microphone configuration.

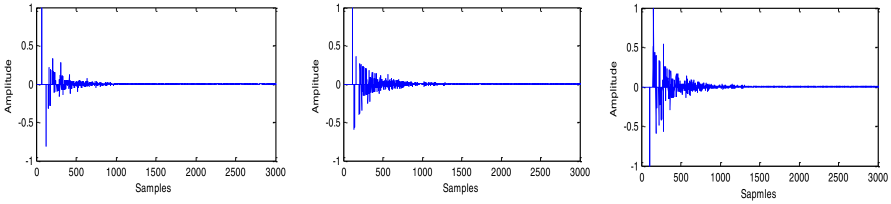


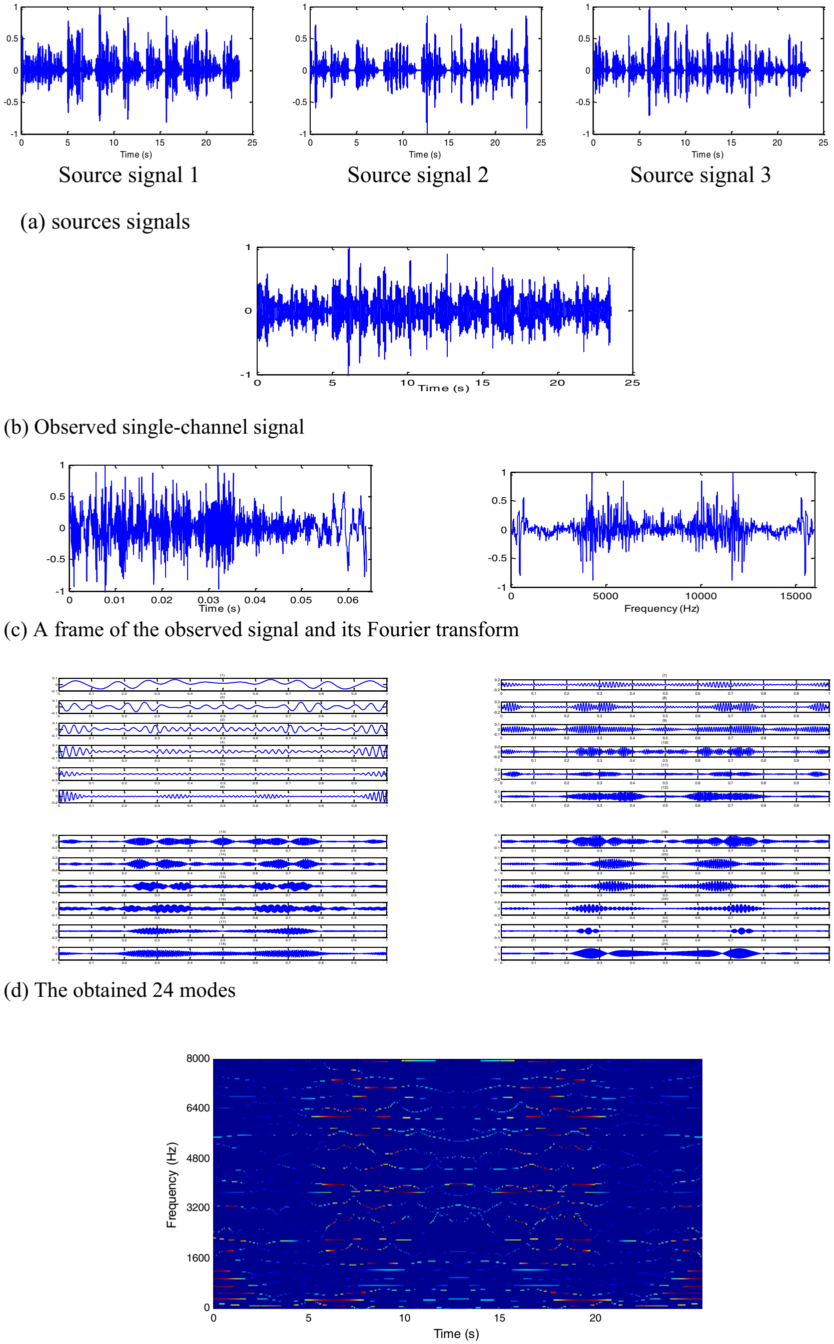
Fig. 3 Room impulse responses from the source i to the microphone.

in 24 modes. The TF representations of the obtained modes are shown in Fig. 4e. A comparison between the STFT of the estimated frame and the original frame of the observed signal is shown in Fig. 4f. The estimated sources are shown in Fig. 4g. A comparison between the TF representations of original sources and estimated sources is shown in Fig. 4h.

As can be seen, the estimated sources are highly similar to the original sources. The proposed method based on the AMSWT method and density-based clustering with sparse reconstruction provides an accurate estimate of the source signals and results in a spectral content located with high accuracy.

The BSSeval toolbox [15] is used to analyze the performance of the proposed approach. The estimated sources are expressed as $\widehat{s} = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}$ for the objective performance criteria measurement, where s_{target} refers to the source signals, e_{interf} stands for interference from other sources, e_{noise} stands for distortion brought on by noise and e_{artif} includes all other artifacts introduced by the separation algorithm.

The parameter p of the l_p-norm-based minimization method can have a significant impact on source reconstruction performance [41]. Many tests have been performed using different values of p to assess its effect on the source-to-distortion ratio (SDR) using the given dataset. Table 1 displays the obtained SDRs for the parameter p varying from 0.1 to 0.9 by a step of 0.2.



Source signal 1

Source signal 2

Source signal 3

(a) sources signals

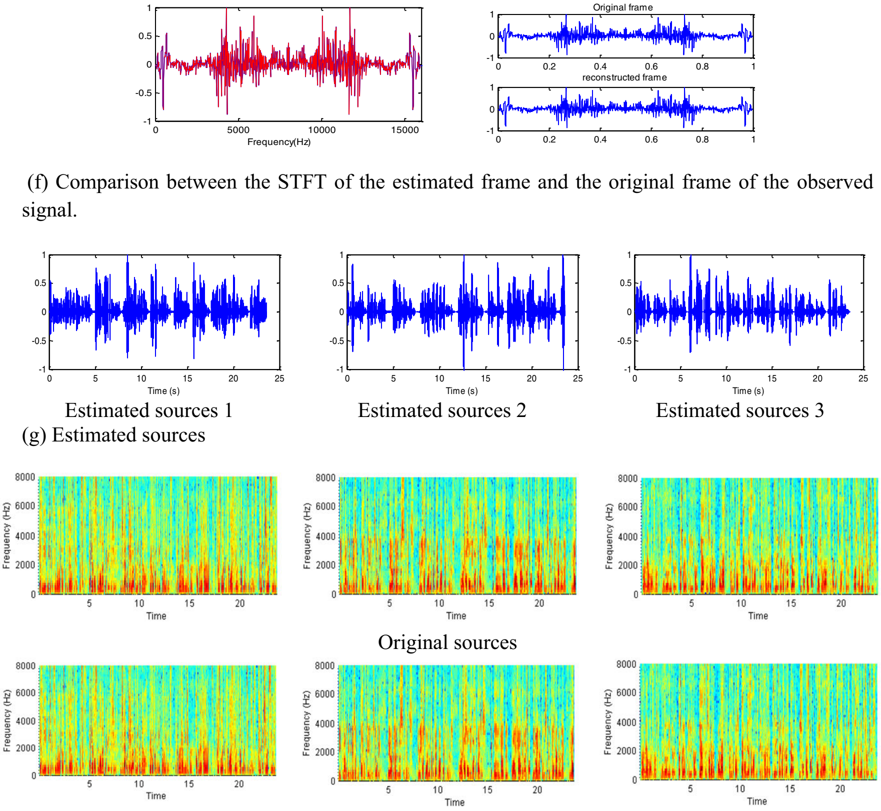
(b) Observed single-channel signal

(c) A frame of the observed signal and its Fourier transform

(d) The obtained 24 modes

(e) The duration of the signal presented as an example is 24 sec

Fig. 4 Illustration example of the single-channel separation of a convolutive mixture of three speech signals based on the proposed method.



(f) Comparison between the STFT of the estimated frame and the original frame of the observed signal.

Estimated sources 1 Estimated sources 2 Estimated sources 3

(g) Estimated sources

Original sources

Estimated sources

(h) Comparison between the TF representation of original sources and estimated sources original

Fig. 4 continued

Table 1 SDRs evaluation for different values of the parameter p

Parameter p value	0.1	0.3	0.5	0.7	0.9
SDR	9.73	10.02	10.42	14.12	13.39

As can be seen, the SDR marginally increases as p increases and reaches its maximum when $p = 0.7$. The parameter p is set to 0.7 in the subsequent experiments. Changing the value of the parameter p to take advantage of the source sparsity proves that the sparse reconstruction based on l_p-norm-based minimization method is very effective.

The estimated sources' performances are evaluated using the SDR, the source-to-artifact ratio (SAR) and the source-to-interference ratio (SIR) criteria, and compared with the performances of the estimated sources obtained via the VMD method [45], adaptive spectrum amplitude estimator and masking method [36] and the nonnegative tensor factorization of modulation spectrograms method [3]. The SDR, SAR and SIR are defined as follows:

$$\mathrm{SDR} = 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}}\|^2} \tag{31}$$

$$\mathrm{SAR} = 10\log_{10}\frac{\|s_{\mathrm{target}} + e_{\mathrm{interf}} + e_{\mathrm{noise}}\|^2}{\|\mathrm{e}_{\mathrm{artif}}\|^2} \tag{32}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}}\|^2} \tag{33}$$

Figure 5 shows the mean square error (MSE) between the original signal and the estimated sources obtained via the proposed method and reference methods. The comparison is performed for different reverberation conditions where the reverberation time is varied from 100 ms to 500 ms by a step of 50 ms. As observed, the proposed method provides the smallest MSE even in a highly reverberant environment.

Figures 6, 7 and 8 show, respectively, the SDR, SAR and SIR obtained by the proposed method and reference methods for different reverberant times. As can be seen, the proposed method results in a better performance in terms of the three criteria compared to the VMD, adaptive spectrum amplitude estimator and masking and nonnegative tensor factorization of modulation spectrogram methods in a reverberant environment. The proposed method results in higher performance criteria even in a highly reverberant environment.

The proposed method has been compared to the reference methods in terms of time computing. In general, the computational complexity [5, 8] is a measure of the execution time. The Fourier transform of a signal of length N has a computational complexity of O($NlogN$) [5, 8]. Then, using variational scaling and wavelet functions, the AMSWT is introduced to adaptively extract spectral intrinsic components (SICs). The AMSWT method is built on the ADMM solver, with a computational complexity of $O\left(n^2\right)$. The density-based clustering has a computational complexity of $O\left(n^3\right)$. The

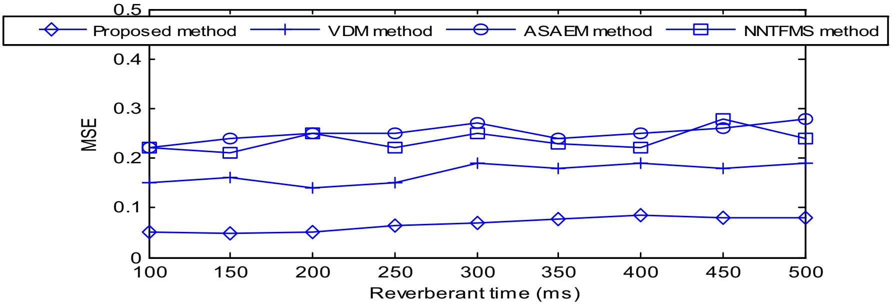


Fig. 5 Comparison in terms of mean square errors (MSEs) between the original signal and the estimated sources obtained via the proposed method and reference methods.

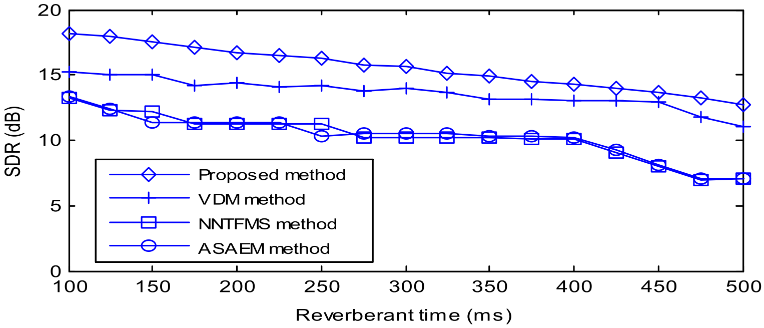


Fig. 6 Comparison in terms of SDR between the original signal and the estimated sources obtained via the proposed method and reference methods.

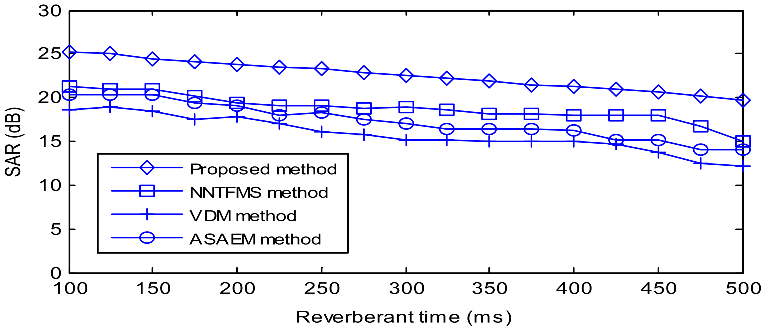


Fig. 7 Comparison in terms of SAR between the original signal and the estimated sources obtained via the proposed method and reference methods.

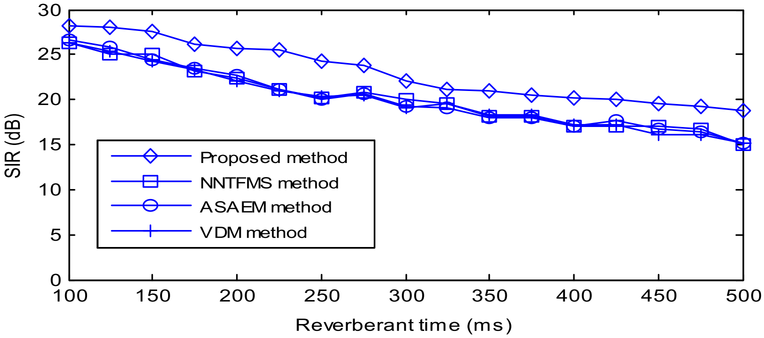


Fig. 8 Comparison in terms of SIR between the original signal and the estimated sources obtained via the proposed method and reference methods.

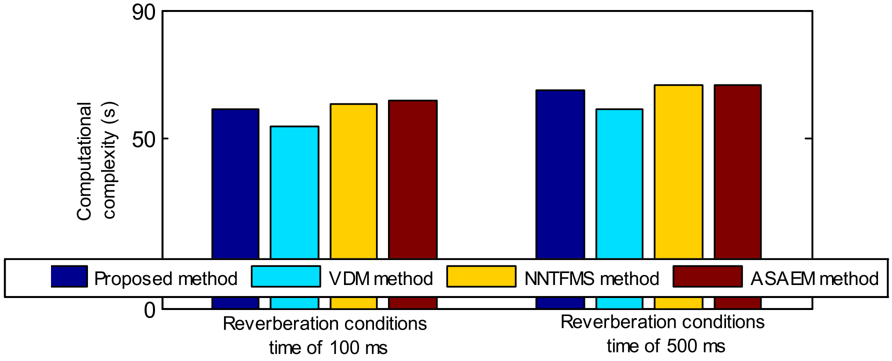


Fig. 9 Comparison between the computational complexity of the proposed method and reference methods in terms of time running (sec).

computational complexity required to compute the similarity matrix is $O\left(n^2\right)$, and the sparse reconstruction used to reconstruct the estimated source has a computational complexity of $O\left(n^3\right)$. The sparse methods are computationally expansive.

The experiments have been carried out using a PC with a 2.4 GHz processor and 4 GB of RAM. The comparison has been performed for reverberation time conditions of 100 ms and 400 ms. The obtained results are shown in Fig. 9. As can be seen, the proposed method has a computational cost than the reference methods both in a weakly reverberant environment and in a highly reverberant environment. The SCBSS methods in the time–frequency domain are computationally expensive.

5 Conclusion

A new method to solve the SCBSS problem has been presented. The method combines the adaptive mode separation-based wavelet transform (AMSWT) with adaptive mode separation and the density-based clustering with sparse reconstruction. The SCBSS problem is transformed into a non-underdetermined. The method operates in the time–frequency domain and a reverberant environment. The proposed method has been tested on speech datasets constructed from TIMIT and NOIZEUS databases for various reverberation time conditions. Simulation experiments indicate that the proposed method results in the smallest MSE and the highest values of SIR, SAR and SDR compared to the reference methods. The simulation results demonstrate the effectiveness of the proposed method to solve the SCBSS problem even in a highly reverberant environment. In terms of computational complexity, the proposed method is expensive.

References

1. NOIZEUS database. http://ecs.utdallas.edu/loizou/speech/noizeus/

2. S. Al-Baddai, and al., Combining EMD with ICA to Analyze Combined EEG-fMRI Data. in*Proceeding of the MIUA*. UK. pp 223–228. (2014).
3. T. Barker, T. Virtanen, Blind separation of audio mixtures through nonnegative tensor factorization of modulation spectrograms. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(12), 2377–2389 (2016)
4. I. Bekkerman, J. Tabrikian, Target detection and localization using mimo radars and sonars. IEEE Trans. Signal Process. **54**(10), 3873–3883 (2006)
5. M. Bouchard, F. Albu, The Gauss-Seidel fast affine projection algorithm for multichannel active noise control and sound reproduction systems. Int. J. Adapt. Control Signal Process **19**, 107–123 (2005)
6. X. Chen, A. Liu, Removing muscle artifacts from EEG data: multichannel or single-channel techniques? IEEE Sens J **16**(7), 1986–1997 (2016)
7. A. Cichocki, S. Amari, *Adaptive Blind Signal and Image Processing* (Wiley, New York, 2003)
8. I. Darazirar, M. Djendi, A two-sensor Gauss-Seidel fast affine projection algorithm for speech enhancement and acoustic noise reduction. Appl. Acoustics **96**, 39–52 (2015)
9. I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, PA, USA, 1992)
10. K. Dragomiretskiy, D. Zosso, Variational mode decomposition. IEEE Trans. Signal Process. **62**(3), 531–544 (2014)
11. B.A. Draper et al., Recognizing faces with PCA and ICA. Comput Vis Image Underst **91**, 115–137 (2003)
12. Z. Duan, Y. Zhang, C. Zhang, Z. Shi, Unsupervised single-channel music source separation by average harmonic structure modeling. IEEE Trans. Audio Speech Lang. Process **16**, 766–778 (2008)
13. A. Eronen, Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. *In proceedings of the 7thInternational Symposium on Signal processing and Its Applications*, Paris (2003).
14. C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
15. C. Févotte, R. Gribonval, E. Vincent, BSS EVAL toolbox user guide, IRISA (2005). http://www.irisa.fr/metiss/bss_eval
16. J. Gilles, Empirical wavelet transform. IEEE Trans. Sign. Process. **61**(16), 3999–4010 (2013)
17. E.A.P. Habets, Room impulse response generator. Tech. Rep. **2**(24), 1 (2006)
18. M.A. Haile, B. Dykas, Blind source separation for vibrationbased diagnostics of rotorcraft bearings. J. Vib. Control **22**(18), 3807–3820 (2016)
19. M.R. Hestenes, Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969)
20. N.E. Huang, et al., The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. in *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences* 454(1971) pp 903–995 (1998)
21. X. Huang, L. Yang, R. Song, W. Lu, Effective pattern recognition and find-density-peaks clustering based blind identification for underdetermined speech mixing systems. Multimed. Tools Appl. **77**(17), 22115–22129 (2018)
22. M. Kemiha, A. Kacha, Complex blind source separation. Circ. Syst. Sign. Pr **36**(11), 4670–4687 (2017)
23. A. Kumar, C.V. Rama Rao, A. Dutta, Performance analysis of blind source separation using canonical correlation. Circuits. Syst. Sign. Process. **37**(2), 658–673 (2018)
24. F. Li et al., Seismic time–frequency analysis via adaptive mode separation-based wavelet transform. IEEE Geosci. Remote. Sens. Lett. **17**(4), 696–700 (2020)
25. L. Li, C.K. Chui, Q. Jiang, Direct signal separation via extraction of local frequencies with adaptive time-varying parameters. IEEE Trans. Sign. Process. **70**, 2321–2333 (2022)
26. Y. Litvin, I. Cohen, Single-channel source separation of audio signals using bark scale wavelet packet decomposition. J. Sign. Process. Syst. **65**, 339–350 (2010)
27. W. Liu, S.Y. CaoChen, Applications of variational mode decomposition in seismic time-frequency analysis. Geophysics **81**(5), V365–V378 (2016)
28. X. Liu, A. Srivastava, K. Gallivan optimal linear representations of images for object recognition. *In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop* (Madison, WI, USA pp 18–20, 2003)
29. C.F. Van Loan, *Matrix computations (Johns Hopkins Studies in the Mathematical Sciences)* (The Johns Hopkins Univ. Press, MD, USA, 1996)
30. S. Makino, T.W. Lee, H. Sawada, *Blind speech separation* (Springer-Verlag, Berlin, Germany, 2007)
31. B. Mijovic et al., Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis. IEEE Trans. Biomed. Eng. **57**, 2188–2196 (2010)

32. A. Nagathil, C. Weihs, K. Neumann, R. Martin, Spectral complexity reduction of music signals based on frequency-domain reduced-rank approximations: an evaluation with cochlear implant listeners. J. Acoust. Soc. Amer. **142**(3), 1219–1228 (2017)
33. D. Nuzillard, A. Bijaoui, Blind source separation and analysis of multispectral astronomical images. Astron Astrophys Suppl Ser. **147**(1), 129–138 (2000)
34. R.B. Randall, A history of cepstrum analysis and its application to mechanical problems. Mech. Syst. Signal Process. **97**, 3–19 (2017)
35. A.K. Takahata et al., Unsupervised processing of geophysical signals: a review of some key aspects of blind deconvolution and blind source separation. IEEE Sign. Process. Mag. **29**(4), 27–35 (2012)
36. N. Tengtrairat, W.L. Woo, S.S. Dlay, B. Gao, Online noisy single-channel source separation using adaptive spectrum amplitude estimator and masking. IEEE Trans. Sign. Process. **64**(7), 1881–1895 (2016)
37. TIMIT database. https://catalog.ldc.upenn.edu/Ldc93s1.
38. J. Traa, P. Smaragdis, Multichannel source separation and tracking with RANSAC and directional statistics. IEEE/ACM Trans. Audio Speech Lang. Process **22**(12), 2233–2243 (2014)
39. B. Wang, M.D, Plumbley, investigating single-channel audio source separation methods based on non-negative matrix factorization. *In Proceedings of* the *ICA Research Network InternationalWorkshop* (Liverpool, UK, pp. 17–20 2006)
40. S. Wilson, J. Yoon, Bayesian ICA-based source separation of Cosmic Microwave Background by a discrete functional approximation. arXiv 2010, arXiv:1011.4018.
41. J. Yang, Y. Guo, Z. Yang, Under-Determined Convolutive Blind Source Separation Combining Density-Based Clustering and Sparse Reconstruction in Time-Frequency Domain. IEEE Trans. Circuits Syst.–i:Regular Papers **66**(8), 3015–3027 (2019)
42. J.J. Yang, H.-L. Liu, Blind identification of the underdetermined mixing matrix based on K-weighted hyperline clustering. Neurocomputing **149**, 483–489 (2015)
43. J. Yang, D.B. Williams, MIMO transmission subspace tracking with low rate feedback. *In proceedings of* the ICASSP, Philadelphia (2005)
44. X. Zeng et al., Fetal ECG extraction by combining single-channel SVD and cyclostationarity-based blind source separation. Int. J. Sig. Process. **6**(4), 367–376 (2013)
45. Y. Zhang, S. Qi, L. Zhou, Single-channel blind source separation for wind turbine aeroacoustics signals based on variational mode decomposition. IEEE Access **6**, 73952–73964 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.