applied sciences

MDPI

Article

Multiple Speech Source Separation Using Inter-Channel Correlation and Relaxed Sparsity

Maoshen Jia [1,*,†,‡], Jundai Sun [1,†,‡] and Xiguang Zheng [2]

[1] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; sunjundai@emails.bjut.edu.cn
[2] Faculty of Engineering & Information Sciences, University of Wollongong, Wollongong NSW2522, Australia; xz725@uow.edu.au
* Correspondence: jiamaoshen@bjut.edu.cn; Tel.: +86-150-1112-0926
† Current address: Beijing University of Technology, No. 100, Pingleyuan, Chaoyang District, Beijing, China.
‡ These authors contributed equally to this work.

Received: 5 December 2017; Accepted: 14 January 2018; Published: 16 January 2018

Abstract: In this work, a multiple speech source separation method using inter-channel correlation and relaxed sparsity is proposed. A B-format microphone with four spatially located channels is adopted due to the size of the microphone array to preserve the spatial parameter integrity of the original signal. Specifically, we firstly measure the proportion of overlapped components among multiple sources and find that there exist many overlapped time-frequency (TF) components with increasing source number. Then, considering the relaxed sparsity of speech sources, we propose a dynamic threshold-based separation approach of sparse components where the threshold is determined by the inter-channel correlation among the recording signals. After conducting a statistical analysis of the number of active sources at each TF instant, a form of relaxed sparsity called the half-K assumption is proposed so that the active source number in a certain TF bin does not exceed half the total number of simultaneously occurring sources. By applying the half-K assumption, the non-sparse components are recovered by regarding the extracted sparse components as a guide, combined with vector decomposition and matrix factorization. Eventually, the final TF coefficients of each source are recovered by the synthesis of sparse and non-sparse components. The proposed method has been evaluated using up to six simultaneous speech sources under both anechoic and reverberant conditions. Both objective and subjective evaluations validated that the perceptual quality of the separated speech by the proposed approach outperforms existing blind source separation (BSS) approaches. Besides, it is robust to different speeches whilst confirming all the separated speeches with similar perceptual quality.

Keywords: multiple speech source separation; sparsity; B-format microphone

1. Introduction

Source separation is a major research area in both signal processing and social internet of things. The information obtained by sound source separation can be widely used for speech enhancement, sound scene reconstruction, and spatial audio production [1–5]. In addition, source separation appears as the central problem of speech recognition and speaker identification problems as well [6–9]. There are several categories of source separation techniques.

Stochastic methods, such as independent component analysis (ICA), rely on a statistical assumption, i.e., mutual statistical independence of sources. They have been widely used in blind source separation (BSS) techniques to recover the sources from mixtures in a determined case [10]. In an overdetermined case, ICA is combined with principal component analysis (PCA) to reduce the dimension of the mixtures, or with least-squares (LS) to minimize the overall mean-square error (MSE) [10,11]. For the most

common underdetermined case, where there are fewer mixtures than sources, sparse representations of the sources are usually employed to increase the likelihood of sources being disjointed [3]. The underlying principle of all existing ICA methods is to achieve mutual independence among separator outputs. The relative success of this approach is mainly due to the convenience of the corresponding mathematical setting, provided by the independence assumption, for algorithm development. Another important factor is the applicability of the independence assumption to a relatively large subset of BSS application domains. However, the ICA-based separation schemes require large amounts of data recorded in a stationary acoustic condition to provide a reasonable estimate of model parameters. In addition, they impose a permutation problem due to misalignment of the individual source components [12–14].

The second group of separation methods is based on adaptive algorithms that optimize a multichannel filter structure according to the signal properties. In other words, an alternative geometric demixing strategy is derived based on the capability of a microphone array for directional acquisition or beamforming. This procedure achieves source separation by steering the beam pattern of the microphone array towards the desired source, thus filtering out the interferences regardless of their signal nature [15,16]. The underlying hypothesis is that the sources are uncorrelated; this assumption is vulnerable to reverberation so the beamformer can mitigate or cancel the desired signal in acoustic reverberation. An additional limiting factor is the spatial resolution for resolving closely located sources. Furthermore, unlike the ICA approach, beamforming requires precise information about the microphone array configuration and the desired source location. Recent work considers a non-linear mixture of beamformers which incorporates the sparsity of the spectrotemporal coefficients to address underdetermined demixing [17]. The application of this method is limited to anechoic mixtures and the performance is degraded due to reverberation.

The other major categories of the separation techniques are based on the sparseness of speech signals in the time-frequency (TF) domain. They assume that the sources approximately meet the W-disjoint orthogonality (W-DO) [18] in the TF domain, i.e., there is at most one sound source active at a certain TF instant. As a result, they achieve separation by partitioning the TF representations of the mixtures belonging to the same speech source. Such groupings can be based on time and phase delays [18] obtained from processing spaced microphone array recordings or intensity-based direction of arrival (DOA) estimates obtained from co-located (spatial) microphone recordings and using microphone directivity. The principle virtue of the W-DO based methods is that they are more computationally efficient compared to stochastic-based methods [19].

When the W-DO of simultaneously occurring speech signals is met, DOA estimates performed in the TF domain will correspond to the location of a true speech source. In practice, simultaneously occurring speech signals are not strictly W-DO for all TF instants, and the separated speech signals using these sparse-based approaches applied to the mixture suffer from musical and crosstalk distortion. This is a result of the non-sparse components (i.e., the TF components derived from more than one source) combining in the mixture, leading to unpredictable DOA estimates that do not correspond to true DOA estimates. The non-sparse TF component is discarded, causing musical distortion of the separated source. Further, if three frontal sources of equal energy are considered (one directly in line with the array and two at equal angles but opposite sides of the array), the non-sparse components contributed by the left and right sources may lead to the same DOA estimate as the middle source. This causes crosstalk distortion, where the separated sources contain spectral content from more than one source at the corresponding TF.

Considering this situation, a collaborative blind source separation method is proposed by using pair location-informed coincident microphone arrays in [20]. This method can jointly separate simultaneous speech sources based on TF source localization estimates from each microphone recording. The musical and crosstalk distortion is effectively reduced by the combination of the microphone pair and vector decomposition. However, if there are three or more speech sources, the vector decomposition will get more difficult and the computation will increase exponentially.

In previous work [21], we achieved an effective multiple sound source localization method by applying "single source zone detection". The method proposed in [21] provides the possibility and necessary parameters for further separation of the corresponding sound sources, including the source number and the corresponding DOA estimations. In this paper, in contrast to existing methods, only a B-format microphone with four channels is used to separate the TF components of the signals. We firstly measure the proportion of overlapped components among multiple sources and find that there exist many overlapped TF components with increasing source number. Thus, a multiple speech source separation method by using inter-channel correlation and half-K assumption is proposed. Specifically, considering the relaxed sparsity of speech sources, we propose a dynamic threshold-based separation approach of sparse components where the threshold is determined by the inter-channel correlation among the recording signals. Thereafter, after conducting a statistical analysis of the number of the active sources at each TF instant, it is concluded that no more than half of the sources are active in a certain TF bin among simultaneously occurring speech sources. By applying the assumption, the non-sparse components are recovered by using the extracted sparse components combined with vector decomposition and matrix factorization. Eventually, the final TF coefficients of each source will be recovered by the synthesis of sparse and non-sparse components.

The remainder of the paper is organized as follows: Section 1 presents the signal model and the limitations of the W-DO assumption. Section 2 introduces the proposed separation method. Experimental results are presented in Section 3, while conclusions are drawn in Section 4.

2. Formulation of The Problem

2.1. Signal Model

The multiple source separation problem is to separate the simultaneously occurring sources from their mixtures with no (BSS) or very limited (semi-BSS) prior knowledge about the mixing process or the sources. In this paper, we focus on the latter. Considering both the size of the microphone array and the preserved spatial parameter-integrity of the original signal, a B-format microphone with four channels, i.e., Front Left Up (FLU), Front Right Down (FRD), Back Left Down (BLD), and Back Right Up (BRU), is considered.

We suppose a scene where B-format microphone is located in the center of the room, as shown in Figure 1. There exist a certain number of speakers in the horizontal plane of the microphone with different angles relative to the center of the B-format microphone [22], i.e., point O. In the discrete TF domain, the pressure signal recorded in free-field by ch channel can be written as:

$$S_{ch}(n,l) = \sum_{i=1}^{K} H_{ch,i}(n,l) \cdot S_i(n,l) \tag{1}$$

where S_{ch} represents one of the four recording signals by the B-format microphone (i.e., s_{FLU}, s_{FRD}, s_{BLD}, s_{BRU}). S_i is one of the K sources with an orientation pair (r_i, μ_i), where radius r_i is the distance of source i with respect to O, and μ_i is the azimuth of source i with respect to x-axis. n and l represent the frame number and the frequency index, respectively. $H_{ch,i}$ is the transfer function from the i_{th} source to channel ch of B-format microphone. Assuming the free-field model, the recording signals $\bigcup\{S_{ch}\}$ can be transformed to the B-format, which consists of one omnidirectional (W) and three figure-of-eight directional (X, Y, Z) channels, i.e., S_W, S_X, S_Y, S_Z.

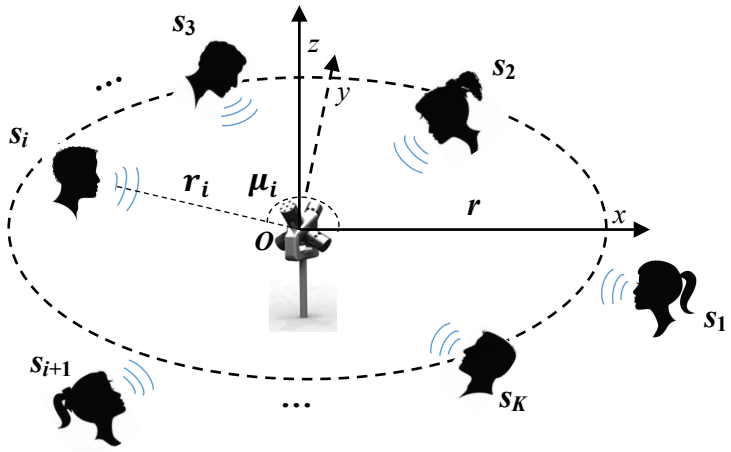

Figure 1. Illustration of the multi-source model and configuration of recording scene. The surrounding sources are numbered S_1 to S_K.

2.2. Limitation of the W-DO Assumption

W-disjoint orthogonality (W-DO) [23] reveals the sparsity speech sources, which means that only one source is active at a certain TF bin. Besides, the spatial information of source S is preserved in the B-format signal by [24,25]:

$$\begin{cases} S_W(n,l) = \frac{\sqrt{2}}{2} \cdot S(n,l) \\ S_X(n,l) = \cos\mu \cdot \cos\eta \cdot S(n,l) \\ S_Y(n,l) = \sin\mu \cdot \cos\eta \cdot S(n,l) \\ S_Z(n,l) = \sin\eta \cdot S(n,l) \end{cases} \tag{2}$$

where S is the sound source signal, and μ and η are the azimuth and elevation of the sound source, respectively, with respect to the center point O. If the W-DO property is valid, similar to [21], the estimated DOA $\hat{\mu}(n,l)$ of each TF bin can be obtained by:

$$\hat{\mu}(n,l) = arctan\frac{I_Y(n,l)}{I_X(n,l)} \tag{3}$$

where I_Y and I_X can be calculated by:

$$\begin{cases} I_X(n,l) = Re\{S_W^*(n,l) \cdot S_X(n,l)\} \\ I_Y(n,l) = Re\{S_W^*(n,l) \cdot S_Y(n,l)\} \end{cases} \tag{4}$$

where $Re\{\cdot\}$ denotes taking a real part of the argument and * denotes conjugation operation.

However, simultaneously occurring speech signals are more likely to overlap in the TF domain, which means that more than one source is active at a TF bin with certain probability [21]. Hence, if the overlapped TF bins of multiple sources occupy a excessive proportion, the localization of multiple sources will be badly influenced. Further, the source separation method based on the localization procedure is not valid again. To solve this problem, some efficient localization methods have been proposed based on single-source bins or zone detection [26]. However, these methods can not eliminate the aliasing of TF components of the multiple source signals. As a result, the aliasing of the TF components cannot be recovered completely, which leads to poor separation quality in the case of the multiple sources. These overlapped TF components are defined as non-sparse components (which cause the famous cocktail problem), while the other TF components derived from one source are defined as sparse components.

In order to verify the W-DO assumption has reduced accuracy with an increasing number of simultaneously occurring sources, we examine how many TF bins are overlapped in the TF domain.

The ratio of overlapped TF bins (ROTF) can be defined as the measure to detect the proportion of non-sparse components, i.e.,

$$ROTF = 1 - \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N}\sum_{n=1}^{N}\left(\frac{||S_i'(n,l)||_0}{||S_i(n,l)||_0}\right) \tag{5}$$

where N is the number of total frames and quasi-norm $||\cdot||_0$ counts the number of non-zero components in its argument. $S_i'(n,l)$ can be obtained by:

$$S_i'(n,l) = \begin{cases} S_i(n,l), if \frac{|S_i(n,l)|}{\left|\sum_{j=1}^{K} S_j(n,l)\right|} > \xi, l \in [1,L] \\ 0, else \end{cases} \tag{6}$$

where L is the number of STFT points in a frame. We detect the TF instant where only one source of energy is dominant among all sources, and then calculate the proportion of the these instants. Eventually, the ROTF is defined by subtracting the proportion. The ROTF implies the ratio of overlapped components among K simultaneously occurring sources. Obviously, a higher ROTF means weaker sparsity among these sources.

In order to examine the average ROTF among simultaneously occurring speech signals, statistical analysis is performed. In total, 36 sentences (the sampling frequency is 16 kHz) from the NTT [27] speech database are used for testing. In the following evaluation, all the test data is from the NTT [27] database unless otherwise stated. Each sentence was divided into a group with the other $K-1$ $(2 \leq K \leq 6)$ sentences in the time domain, resulting in K simultaneously occurring speech conditions. For $K = 2$, each sentence was divided into a group with each of the remaining 35 sentences resulting $36 \times 35 = 1260$ combinations. For $K > 2$, each sentence was randomly grouped 35 times with $K-1$ other sentences to give the same number of combinations (1260) as for $K = 2$. In addition, $\xi = 0.9$. The average length of each recording is about 8 s. Based on the aforementioned conditions, a statistical analysis of ROTF is taken. Statistical results are shown in Figure 2 with 95% confidence intervals.

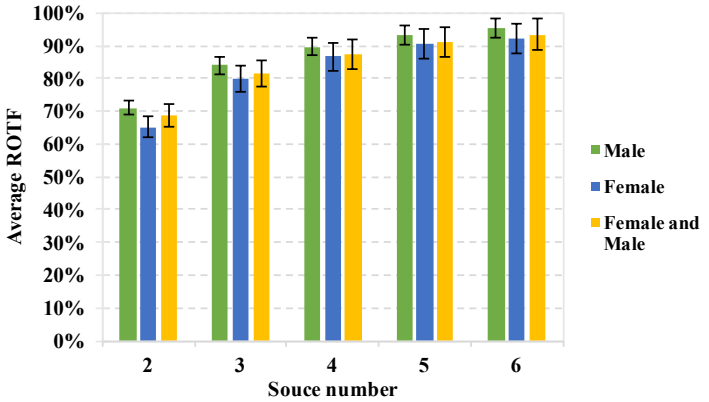

Figure 2. Average ratio of overlapped time-frequency (ROTF) for 2–6 sources.

It can be observed that more TF bins are overlapped when the number of simultaneously occurring sources is increasing. In particular, when $K \geq 5$, the TF bins are almost overlapped, with a high percentage of over 90%.

3. Proposed Method

Based on the investigation in Section 2, we can conclude that the W-DO has less accuracy as the number of simultaneously occurring sources increases. In order to eliminate the problem of poor

separation quality caused by this phenomenon, we propose a multiple source separation method based on self-reduction of dimensionality by using a B-format microphone. After proposing an effective detection method of the active sources in a non-sparse bin, we conduct a statistical analysis on the active source number that is involved in the non-sparse components and the corresponding possibility. It is found that when there exist K sound sources simultaneously and the non-sparse components are mostly caused by less than $K/2$ sound sources, it is very rare that the TF components of all sound sources are overlapped. Therefore, based on this phenomenon, we assume that when K sound sources simultaneously occur in a sound scene, the active source number corresponding to the non-sparse component does not exceed $K/2$, which we call the half-K assumption. In addition, due to the recording characteristics of B-format microphone, we can get three linear equations with K source signals as independent variables. If the number of sound sources is greater than three, the linear equations will have multiple solutions, so the B-format microphone can only be used to separate the source signals of the scene with three sound sources. Based on the half-K assumption proposed in this paper, the B-format microphone can be used to solve the separation problem in the sound scene with six sound sources.

The illustration of the proposed BSS scheme is shown in Figure 3. For the input mixture signals (four recording signals of the B-format microphone [21]), the DOA estimation can be obtained by a traditional localization procedure [21]. Thereafter, the sparse components recovery can be achieved by a clustering process of TF bins. The unprocessed non-sparse components are then obtained by masking the recovered sparse components from mixture signals. The half-K assumption provides a reduction of the dimensionality of the linear equations. By solving the linear equations, the non-sparse components will be effectively separated. Eventually, the final TF coefficients of each source will be recovered by the synthesis of sparse and non-sparse components.

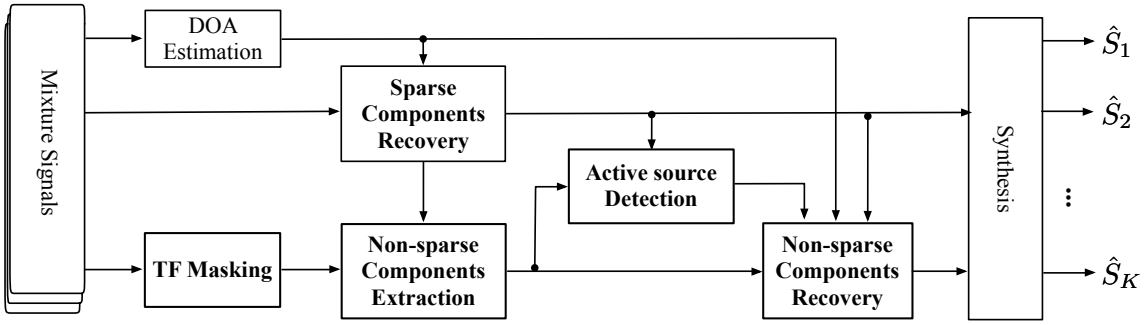

Figure 3. System block diagram of the proposed method. DOA: direction of arrival; TF: time-frequency.

3.1. Separation of Sparse Components

Under the sparse assumption discussed in [20], one source will have the same mixing parameter pairs A and μ. For a given TF instant of i_{th} source, this pair is approximated by:

$$[A_i(n,l), \mu(n,l)] = \left[\left| \frac{I_Y(n,l)}{I_X(n,l)} \right|, \angle \frac{I_Y(n,l)}{I_X(n,l)} \right] \tag{7}$$

The separated sources can be derived by grouping the TF instants using these parameter pairs. If the estimated DOA of i_{th} source is μ_i, the task is to determine a range around μ_i (i.e., $[\mu_i - \Delta\mu, \mu_i + \Delta\mu]$) such that the TF instants having the DOA estimates within this range are considered as the source i. It should be noted that if the threshold is set small enough, less interference from other sources may be contained in the separation. However, this may fail to derive many TF components whose DOA estimates are slightly different to the true source DOA due to the low fault tolerance of the estimation approach. If the threshold is larger, this may lead to the inclusion in the

separated source of TF components from other sources. Hence, an efficient clustering method is needed to dynamically achieve the separation of sparse components.

Based on the directional characteristic of B-format microphone, it can be seen that a strong correlation between the recording signals of adjacent channels in a certain TF zone implies that there is only one source active, while a weaker correlation means it is a region where multiple sources exist [21]. Hence, in this paper, we proposed a dynamic threshold clustering method of sparse components based on the inter-correlation of the raw recording (A-format) signals [21], and the threshold $\Delta\mu$ is dynamically set by:

$$\Delta\mu = \frac{1}{1 + e^{-\alpha(\overline{\gamma}-\beta)}} \cdot \mu_0 \tag{8}$$

where μ_0, α, and β are initial thresholds for the user to define; μ_0 is the threshold to control the dynamic range of the $\Delta\mu$, α is a threshold for controlling the $\Delta\mu$ change curve, and β is a symmetric point corresponding to the change curve. $\bar{\gamma}$ is the average of normalized cross-correlation coefficients [21] among four recording signals of the B-format microphone. More specifically, for any pair of soundfield microphone-recorded signals ($S_{chi}(n,l)$ and $S_chj(n,l)$), the function is defined as:

$$R_{i,j}(\mathcal{K}) = \sum_{(n,l)\in\mathcal{K}} |S_{chi}(n,l) \cdot S_{chj}(n,l)| \tag{9}$$

where $i \neq j$, $S_{chi}(n,l), S_{chj}(n,l) \in \{S_{ch1}(n,l), S_{ch2}(n,l), S_{ch3}(n,l), S_{ch4}(n,l),\}$. The normalized cross-correlation coefficient can be obtained by:

$$\gamma_{i,j}(\mathcal{K}) = \frac{R_{i,j}(\mathcal{K})}{\sqrt{R_{i,i}(\mathcal{K}) \cdot R_{j,j}(\mathcal{K})}} \tag{10}$$

It can be seen from Equation (8) that the value of $\Delta\mu$ is proportional to the correlation coefficient. Specifically, if the correlation coefficient is close to one, the threshold will be large to obtain more extractions of the corresponding source, while in other TF zones, it will get smaller with the average correlation coefficient being smaller in order to get rid of the interference of other sources. Considering this issue and the value range of the cross-correlation coefficient, Equation (8) should be a function whose independent and dependent variables are both with a value range in [0, 1]. By adjusting the value of α and β, we find $\alpha = 10$ and $\beta = 0.5$ can perfectly meet the requirement mentioned above. Future work will investigate alternative methods for optimizing the choice of these values and find whether there might be a more efficient function that can describe the relation between $\overline{\gamma}$ and $\Delta\mu$.

To make full use of the directional characteristic of B-format microphone, we can obtain the most appropriate vector as the mixed signal for source separation which can be obtained by $\boldsymbol{S_X}$ and $\boldsymbol{S_Y}$. It was found experimentally that the performance degrades when processing of S_W is employed for source separation, and a similar phenomenon was also found in other work [28]. In detail, the most appropriate vector $[\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_K]^T$ of K sources can be obtained by:

$$\begin{bmatrix} \boldsymbol{f_1} \\ \boldsymbol{f_2} \\ \vdots \\ \boldsymbol{f_K} \end{bmatrix} = \begin{bmatrix} \cos\mu_1 & \sin\mu_1 \\ \cos\mu_2 & \sin\mu_2 \\ \vdots & \vdots \\ \cos\mu_K & \sin\mu_K \end{bmatrix} \begin{bmatrix} \boldsymbol{S_X} \\ \boldsymbol{S_Y} \end{bmatrix} \tag{11}$$

The separation of sparse components in a certain frame proceeds as per Algorithm 1.

Algorithm 1 Sparse Component Separation

Input: $[\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_K}]$.
Initialize: Divide the frame into $\mathcal{J}$ sub-band regions with a equal width in the TF domain.
for $l = 0, 1, ..., L$ **do**
Calculate the average of normalized cross-correlation $\bar{\gamma}$, and obtain the $\Delta\mu$ by Equation (8).
Obtain the sparse components $\boldsymbol{S_i^C}$ of source i by clustering the TF components by parameter pair $\{\boldsymbol{f_i}, \mu_i\}$.
end for
Output: Eventually, we will get the sparse components $\boldsymbol{S_i^C}$ of each source.

3.2. Exploring Inter-Sparsity among Multiple Sources

Further, in order to investigate the inter-sparsity among multiple sources, and how many active sources are involved in the overlapped TF bins, we proposed a statistic algorithm of the number of the active source at each TF instant when there occur multiple sources. In detail, the source whose energy at a certain TF instant is dominant among all the sources is regarded as the active source at this instant, i.e., its energy occupies a significant proportion of the total energy of all sources. To find the proportion of active source number when different numbers of sources occur, we define a statistical measure which is reflected in Algorithm 2.

Algorithm 2 Statistic of Active Source Number

Input: $\boldsymbol{S_a^n} = (a_{li})^{L\times K} = [\boldsymbol{S_1^n}\ \boldsymbol{S_2^n}\ \cdots\ \boldsymbol{S_K^n}]$, where $\boldsymbol{S_i^n} = [|S_i(n,1)|, |S_i(n,2)|, ..., |S_i(n,L)|]^T$.
Initialize: $\boldsymbol{Counter} = [Count_1, Count_2, ..., Count_K] = \boldsymbol{0}$, c is used for counting the active source number within current TF bin, $c = 0$, loop frequency index: l = 1, loop source index: i = 0.
for $l = 1, ..., L$ **do**
for $i = 1, ..., K$ **do**
if $a_{li} > \eta \cdot \sum_{j=1}^{K} |S_j(n,l)|$
increment c.
end if
end for
increment $Counter_c$ (i.e., increment value of the element in **counter** whose index is c).
Reset: $c = 0$.
end for
Output:$\boldsymbol{Counter}$.

Then, we can calculate the probability of active source number ($PASN$) among K simultaneously occurring sources as:

$$PASN(c) = \frac{Counter_c}{L} \tag{12}$$

where $Counter_c$ denotes the number of TF instants when c sources are active simultaneously, L is the number of STFT points in a frame, and $PASN(c)$ represents the probability of TF bins which contain c active sources. In other words, it implies the probability of c sources active simultaneously over all TF bins. In order to analyze the $PASN$ among K simultaneously occurring sources, we calculate the $PASN$ of all groups mentioned above when $K = 6$. The results are shown in Figure 4 with 95% confidence intervals.

It can be seen that when there are six simultaneously occurring speech sources, most of the time only two or three sound sources are active at the same time over the non-sparse TF bins (occupying nearly 70%), i.e., most of the non-sparse components are involved with two or three sources. This implies that the proposed half-K assumption that no more than $K/2$ sources are active in a certain TF bin when there are K simultaneously occurring speech sources is reasonable.

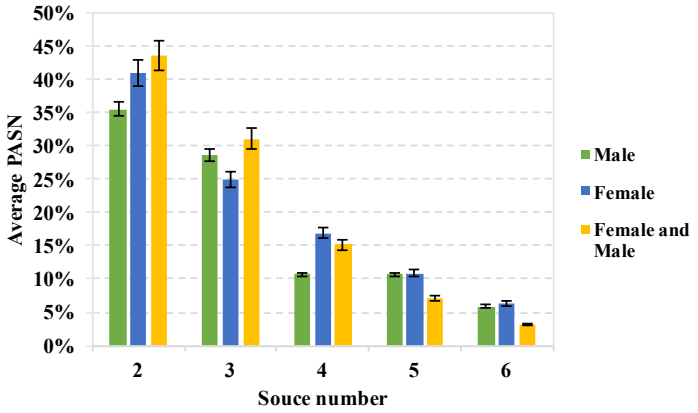

Figure 4. Average probability of active source number (PASN) among six simultaneously occurring sources.

3.3. Separation of Non-Sparse Components

Based on the statistical results in Section 3.2, we have validated the half-K assumption that the active source number in a certain TF bin does not exceed half the total number of simultaneously occurring sources. It means that for K sources, there are no more than $K/2$ sources active at a certain TF bin. We set $K \leq 6$ to ensure the localization accuracy in this work. Based on the proposed assumption, we can conclude that the number of active sources at a certain TF bin (K_a) is no more than three. It should be noted that $K_a = 1$ means that there is only one source active at a certain TF bin, i.e., the sparse bin. The set of these sparse bins are the above-mentioned sparse components.

Aiming to separate the corresponding non-sparse components of each source, we have to know all the active sources of the non-sparse components. Here, the problem is solved by first dividing the TF band into several zones with same width, and then calculating the similarity between the separated sparse components and the mixture signal in this TF zone. In detail, if the similarity exceeds a certain threshold, the frequency components of the regions are mostly derived from this source signal and the signal set that satisfies the threshold is named the active source in the current TF zone of non-sparse components. The normalized cross-correlation function is utilized for similarity calculation, and the cross-correlation coefficient between the mixture signal S_W and a sparse component signal S_i^C in a TF zone $\mathcal{Z}$ is calculated as follows:

$$R_{S_W,S_i^C}(\mathcal{Z}) = \frac{\sum\limits_{(n,l)\in\mathcal{Z}} |S_W(n,l)| \cdot |S_i^C(n,l)|}{\sqrt{\sum\limits_{(n,l)\in\mathcal{Z}} |S_W(n,l)|^2 \cdot \sum\limits_{(n,l)\in\mathcal{Z}} |S_i^C(n,l)|^2}} \tag{13}$$

where S_i^C represents the separated sparse component signal and $i = 1, 2, ..., K$. To obtain all the active sources of non-sparse components in a TF-analyzed one $\mathcal{Z}$, we define a active detecting vector $\mathfrak{D} = [D_1, D_2, ..., D_K]$; the ith element D_i can be obtained by:

$$D_i = \begin{cases} 1, \text{if } R_{S_W,S_{C_i}}(\mathcal{Z}) > \epsilon \\ 0, else \end{cases} \tag{14}$$

where ϵ is an experimental threshold, in order to ensure that the number detected active sound source is not larger than the real one. We set $\epsilon = 0.8$ in this paper (informal testing found this value generally led to satisfactory results but future work can explore the optimization of this value). The index value of the detected active source is recorded in a vector $\mathcal{I}$, which can be obtained by a $INDEX^0$ function as:

$$\boldsymbol{\mathcal{I}} = INDEX^0(\boldsymbol{\mathfrak{D}}) \tag{15}$$

where $INDEX^0(\cdot)$ is a non-zero index searching function; the output parameter is a vector that contains all the indexes of non-zero elements of the input vector.

$||\boldsymbol{\mathfrak{D}}||_0 = 3$, i.e., $\boldsymbol{\mathcal{I}} = [\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3]$, where $||\cdot||_0$ counts the number of non-zero components in its argument. This means there are three active sources at the current zone. Hence, the corresponding TF coefficient of these active sources is rewritten to find a solution to the linear equations:

$$\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \cos\mu_{\mathcal{I}_1} & \cos\mu_{\mathcal{I}_2} & \cos\mu_{\mathcal{I}_3} \\ \sin\mu_{\mathcal{I}_1} & \sin\mu_{\mathcal{I}_2} & \sin\mu_{\mathcal{I}_3} \end{bmatrix} \begin{bmatrix} S_{\mathcal{I}_1}(n,l) \\ S_{\mathcal{I}_2}(n,l) \\ S_{\mathcal{I}_3}(n,l) \end{bmatrix} = \begin{bmatrix} S_W(n,l) \\ S_X(n,l) \\ S_Y(n,l) \end{bmatrix} \tag{16}$$

where $\mu_{\mathcal{I}_1}$, $\mu_{\mathcal{I}_2}$, $\mu_{\mathcal{I}_3}$ denote the estimated DOA of the active source by applying the method in [21]. Equation (16) can be rewritten as a vector form as:

$$\boldsymbol{\mathfrak{K}}\boldsymbol{S}_{\boldsymbol{\mathcal{I}}} = \boldsymbol{S} \tag{17}$$

where $\boldsymbol{S}_{\boldsymbol{\mathcal{I}}} = [S_{\mathcal{I}_1(n,l)}, S_{\mathcal{I}_2(n,l)}, S_{\mathcal{I}_3(n,l)}]^T$ and $\boldsymbol{S} = [S_W(n,l), S_X(n,l), S_Y(n,l)]^T$.The separation problem is converted to solve the linear equations by regarding the TF coefficient of each active source as an independent variable. It should be mentioned that the process is based on the hypothesis that the mixing matrix of Equation (16) is a column full rank matrix. To jointly consider the case $||\boldsymbol{\mathfrak{D}}||_0 > 3$, the aim is converted to find a vector $\boldsymbol{S}^N_{\boldsymbol{\mathcal{I}}} = [S^N_{\mathcal{I}_1}(n,l), S^N_{\mathcal{I}_2}(n,l), S^N_{\mathcal{I}_3}(n,l)]$ to minimize $||\boldsymbol{S} - \boldsymbol{\mathfrak{K}}\boldsymbol{S}_{\boldsymbol{\mathcal{I}}}||^2_F$, i.e.,

$$\boldsymbol{S}^N_{\boldsymbol{\mathcal{I}}} = \arg\min_{\boldsymbol{S}_{\boldsymbol{\mathcal{I}}}} ||\boldsymbol{S} - \boldsymbol{\mathfrak{K}}\boldsymbol{S}_{\boldsymbol{\mathcal{I}}}||^2_F \tag{18}$$

where $||\cdot||_F$ represents the Frobenius norm. $\boldsymbol{S}^N_{\boldsymbol{\mathcal{I}}}$ represents the separated TF coefficients of the active sources at (n,l).

For the case $||\boldsymbol{\mathfrak{D}}||_0 = 2$, i.e., $\boldsymbol{\mathcal{I}} = [\mathcal{I}_1, \mathcal{I}_2]$, we can omit the first equation in Equation (16) i.e.,

$$\begin{bmatrix} \cos\mu_{\mathcal{I}_1} & \cos\mu_{\mathcal{I}_2} \\ \sin\mu_{\mathcal{I}_1} & \sin\mu_{\mathcal{I}_2} \end{bmatrix} \begin{bmatrix} S_{\mathcal{I}_1}(n,l) \\ S_{\mathcal{I}_2}(n,l) \end{bmatrix} = \begin{bmatrix} S_X(n,l) \\ S_Y(n,l) \end{bmatrix} \tag{19}$$

We can still get the solution by solving Equation (18) for this case. Then, we can get the actual TF coefficients of $S_{\mathcal{I}_1}(n,l)$ and $S_{\mathcal{I}_2}(n,l)$ by:

$$\begin{cases} S^N_{\mathcal{I}_1}(n,l) = \frac{sin(\mu_{\mathcal{I}_2})\cdot S_X(n,l) - cos(\mu_{\mathcal{I}_2})\cdot S_Y(n,l)}{sin(\mu_{\mathcal{I}_2} - \mu_{\mathcal{I}_1})} \\ S^N_{\mathcal{I}_2}(n,l) = \frac{sin(\mu_{\mathcal{I}_1})\cdot S_X(n,l) - cos(\mu_{\mathcal{I}_1})\cdot S_Y(n,l)}{sin(\mu_{\mathcal{I}_1} - \mu_{\mathcal{I}_2})} \end{cases} \tag{20}$$

Finally, the final recovered signal can be obtained by a synthesis as:

$$\hat{S}_i(n,l) = S^C_i(n,l) + S^N_i(n,l) \tag{21}$$

Figure 5a illustrates an example of the TF component extraction of the proposed method. Figure 5b is an example of the recovered signals from the mixture signal.

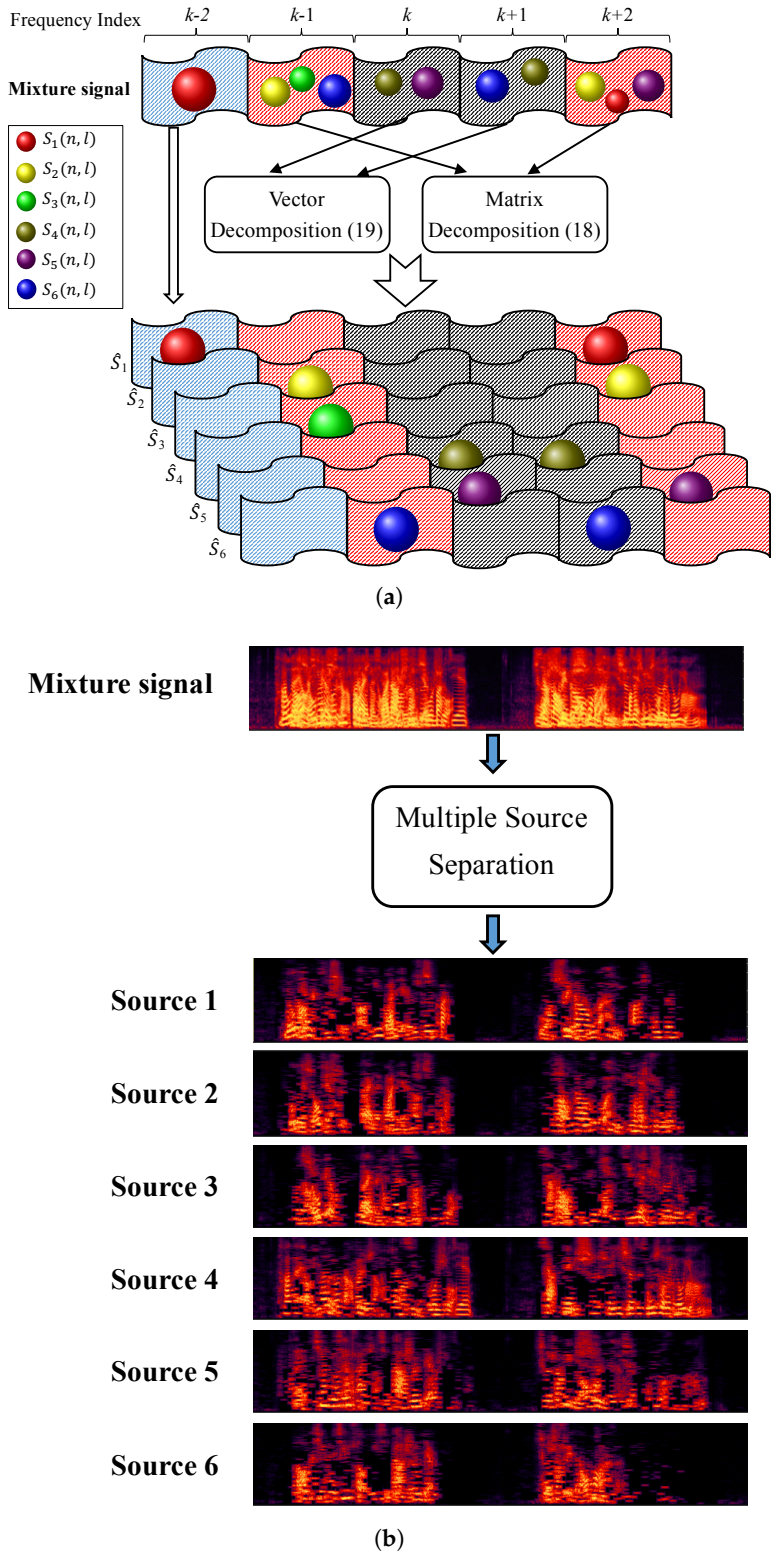

Figure 5. (**a**) Example of a TF component extraction of the proposed framework. Both the frame length and number of STFT points are 2048. Each square represents a TF instant. Blue-shadowed squares denote the sparse component, while the squares in other colors denote the non-sparse component. The respective TF instants for each source (six sources for example) are indicated by a ball; (**b**) Example of six recovered signals from the mixture signal in the frequency domain.

4. Evaluation

In this section, to verify the effectiveness of the proposed method, a series of objective and subjective evaluation tests are presented. Several aspects are considered in the tests to assess the separation quality including source number, the angle between sources, and the environment.

NTT [27], as a speech database including various speakers from different countries, has been chosen as the testing database. In addition, all the data are monorecordings and the energy of all speech in NTT database is the same. Thus, this database is suitable for evaluating the quality of multiple speech object separation methods. For the evaluation in simulated scenarios, all of the test segments are derived from the database. Each test segment representing a speech source is created with a length of 8 s. In order to evaluate the separation quality when different types of multiple speech sources are active simultaneously, a complicated situation where different proportions of male and female speakers are simultaneously talking is considered in this work.

To evaluate the proposed method in different environments, we used Roomsim [29] to simulate a room measuring $6.25 \times 4.75 \times 2.5$ m^3 with different reverberation conditions. The main parameters of the simulated rooms are illustrated in Table 1. The B-format microphone was placed in the center of the room parallel with the z-axis, and the power of sound sources from different directions was equal in each simulation. It should be noted that the B-format microphone was simulated via Roomsim, and was completed by simulating the recording condition of each channel. In addition, the radius of the cube (radius of the circumscribed circle) is 12 mm. The azimuth and elevation pairs of the four channels (i.e., FLU, FRD, BLD, and BRU) are $\{(45^\circ, 45^\circ), (-45^\circ, -45^\circ), (135^\circ, -45^\circ), (-135^\circ, 45^\circ)\}$. For the objective evaluation in real environments, all the test data was recorded in a room measuring $6.25 \times 4.75 \times 2.5$ m^3, SNR = 20 dB, $RT60$ = 0.5 s.

In addition, the width of TF-analyzed region mentioned in Section 2 is determined by applications. In this work, in order to obtain a most efficient width of the TF-analyzed zone, we took a number of tests and found that the score keeps stable for five different widths: $\{128, 64, 32,$ and $16\}$. We chose 64 as the width in [21] for "single-source" zone detecting, so the width of analyzed TF zone was also set by a constant of 64 considering both efficiency and low computational complexity. Other allocation strategies might improve the quality of individual speech sources or balance the quality amongst all sources, which will be investigated as the future work.

Table 1. Parameters of the testing room.

Simulated Room	Absorption of the Wall	RT_{60} (ms)	Reflection Order
Anechoic Room	1	0	0
Room1	0.75	250	18
Room2	0.75	500	18

4.1. Objective Evaluation

For objective evaluation, three measurements, i.e., perceptual evaluation of speech quality (PESQ) [30], signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) were adopted to evaluate the perceptual quality of extracted speech signals. Specifically, the PESQ generated by the evaluation software [30] was used to evaluate the perceptual similarity between the separated signal and the original signal. The score interval of PESQ is [0, 5], where smaller values imply a degradation of the quality of separated speech signal. The SDR and SIR were obtained by using the BSS EVAL Toolbox [31]; the SDR measures the overall performance (quality) of the algorithm, and the SIR focuses on the interference rejection. The tests were conducted in both simulated scenarios and real environments.

For comparison, one of the most efficient sparse component separation (SCS) methods was selected as the reference method [28] to indicate the effect of the non-sparse component recovery by using the proposed method. Then, five outstanding existing methods were chosen for further evaluation.

Specifically, for the determined case ($K = 3$ in this paper), we compared the PESQ score of our proposed method with BSS methods, which belongs to other categories under simulated condition.

4.1.1. Simulated Environment

First, by evaluating the same test data in s environment, the average PESQ scores of fixed threshold method and our proposed approach virus different source numbers and separations are shown in Figure 6. Condition SCS is the result extracted by the fixed threshold (i.e., $\Delta\mu = \mu_0 = 8°$) sparse component separation method, while Pro-SCS is the proposed approach of sparse components with a dynamic threshold, and condition Pro-BSS is the proposed sparse and non-sparse component separation method. Figure 6a–d represents the result for separations $\{30°, 40°, 50°, 60°\}$. It can be seen that the two proposed separation methods reach a higher score, especially for *source number* $\geqslant 3$. In addition, Pro-BSS reaches the highest score, and the average scores are all above 2 for all source numbers, which indicates a better perceptual quality of the proposed method compared to the fixed threshold separation approach.

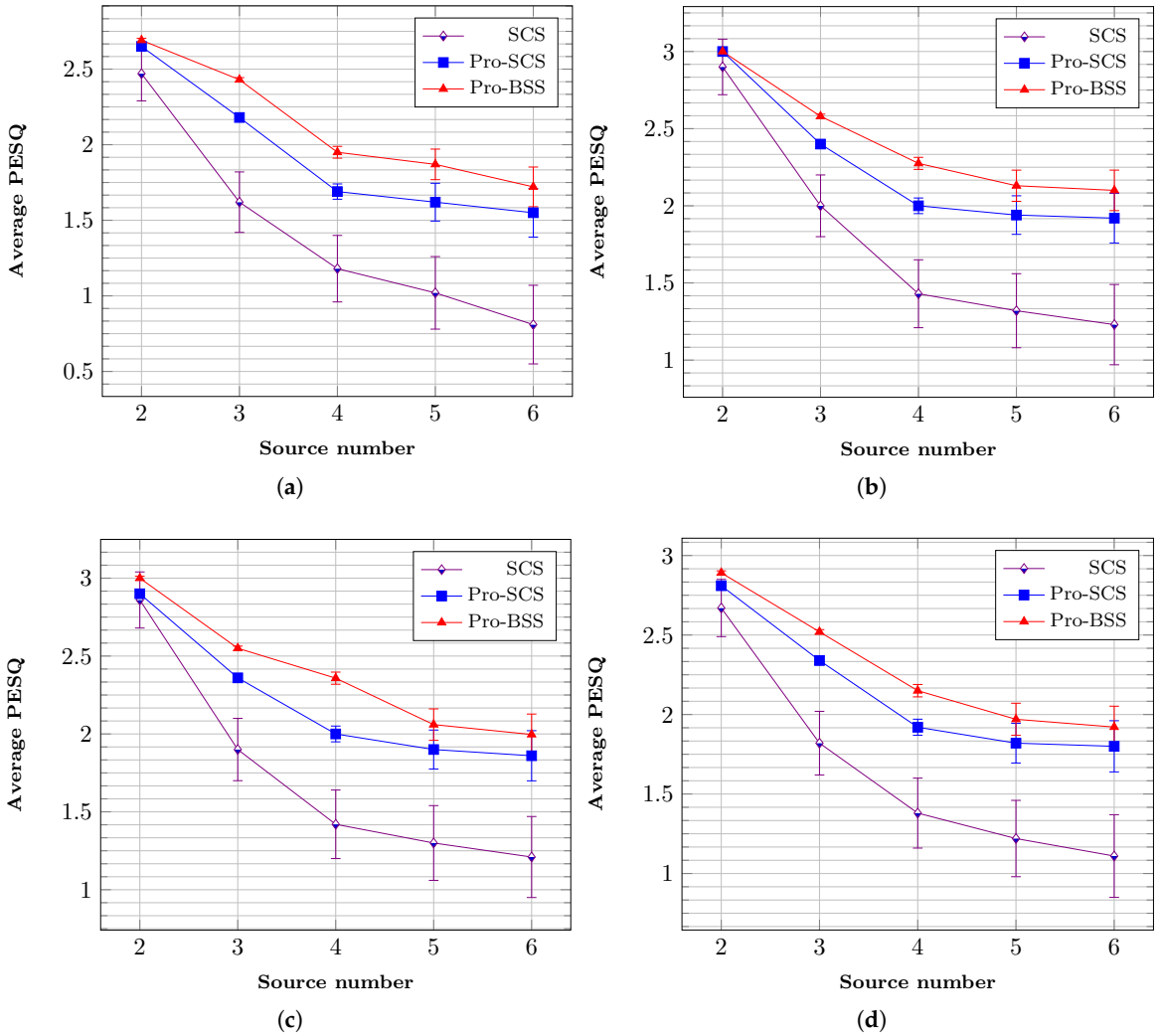

Figure 6. Perceptual evaluation of speech quality (PESQ) results. Error bars represent 95% confidence intervals; (**a**–**d**) represent the results for separations {30°, 40°, 50°, 60°}. SCS: sparse component separation; Pro-SCS: the proposed approach of sparse components with a dynamic threshold; Pro-BSS: the proposed sparse and non-sparse component separation method.

The corresponding results are shown in Figures 7 and 8, respectively. Condition mixture (W) represents the B-format input signal S_W. The SDR and SIR results follow a similar trend to the PESQ results. Overall, it can be concluded that the proposed method obtains a great improvement in extracted sources.

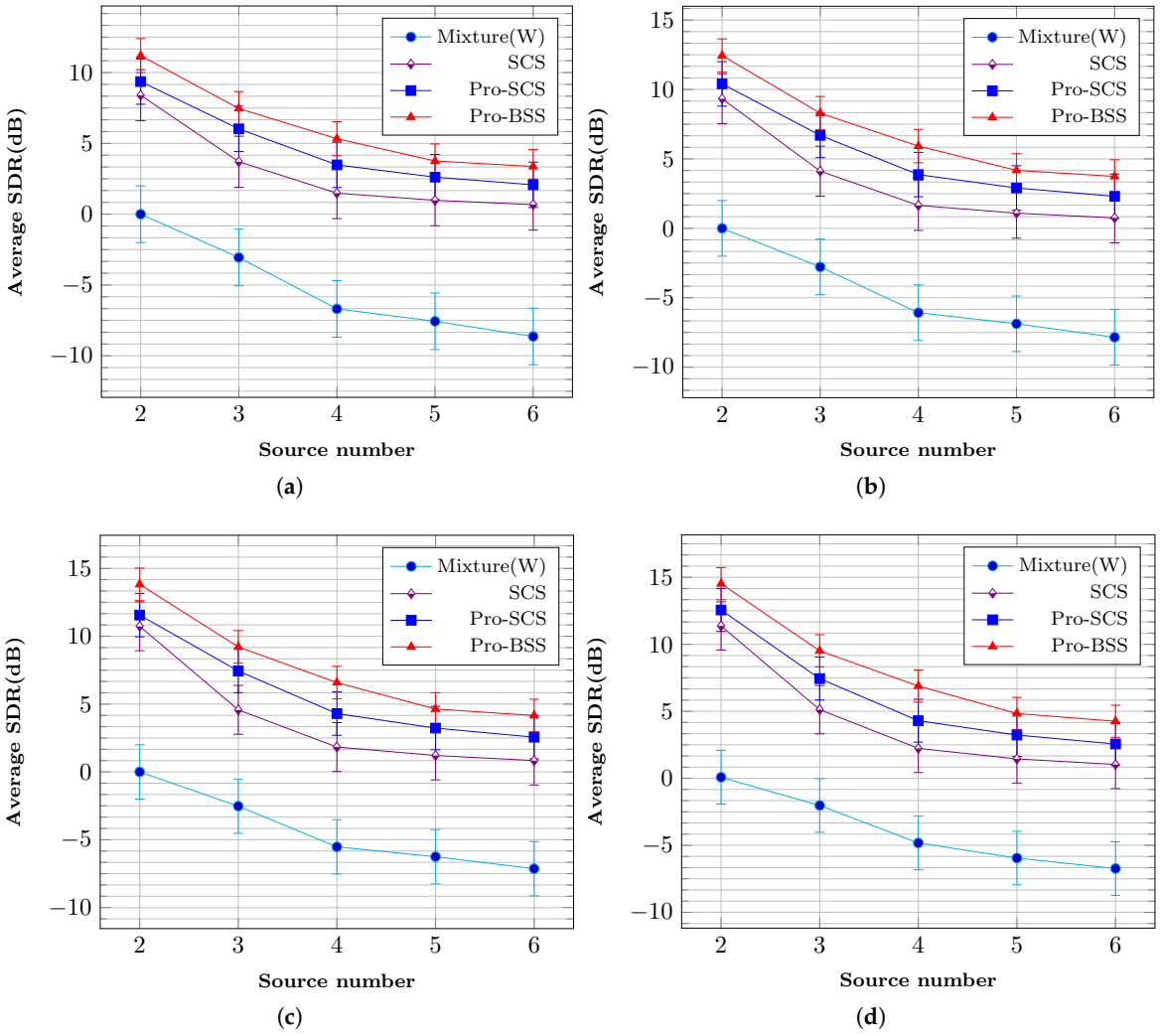

Figure 7. Signal-to-distortion ratio (SDR) results. Error bars represent 95% confidence intervals; (**a**–**d**) represent the results for separations {30°, 40°, 50°, 60°}.

Eventually, for the determined case (i.e., $K = 3$), we conducted a comparison with four other existing approaches: (a) spatio-temporal ICA [32] applied using a single (recording from M_i using channel W_i, X_i, Y_i) B-format speech mixture (S-ICA); (b) spatio-temporal ICA applied using a dual (recording from M_i using channel W_i, X_i, Y_i and corresponding channels of M_j) B-format speech mixture (D-ICA); (c) source DOA-based BSS using single coincident microphone recording (S-BSS) [28]; and DOA-based collaborative BSS (CBSS) [20] using a pair of coincident spatial microphones. Note that the reference mixture (W) is an unprocessed speech mixture (W channel of the B-format recording) used for indicating the worst quality. It should be noted that there are still many good algorithms, like the independent vector analysis (IVA)-based method [33,34], the independent low-rank matrix analysis (ILRMA)-based method, and so on [35]. Their methods focus on audio source separation, while we prefer the case with all speech sources, so only a few algorithms are chosen for comparison.

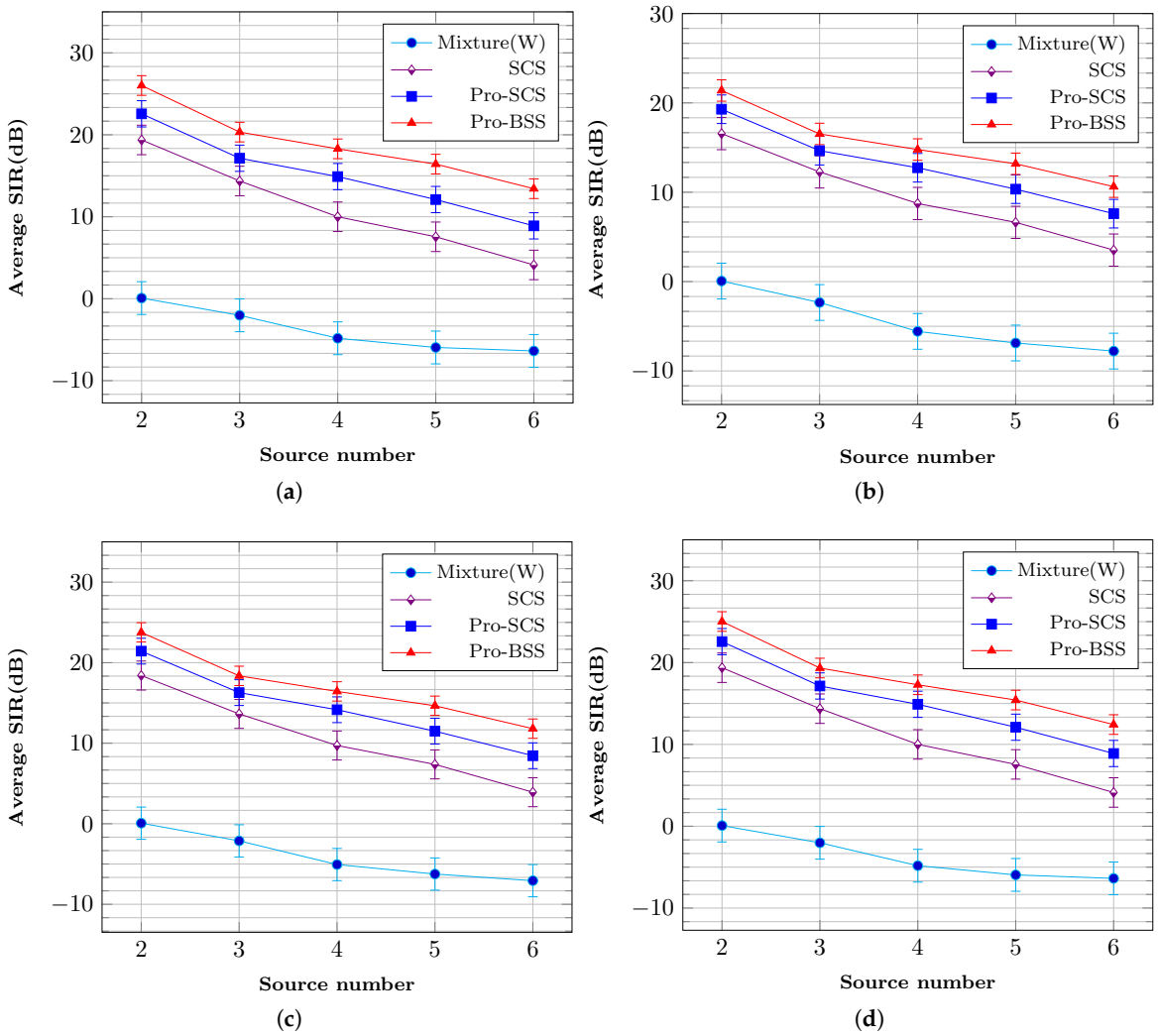

Figure 8. Signal-to-interference ratio (SIR) results. Error bars represent 95% confidence intervals; (**a**–**d**) represent the results for separations {30°, 40°, 50°, 60°}.

From Figure 9, the proposed BSS approach outperforms the other BSS techniques based on the PESQ measure. It should be noted that Figure 9a–c are calculated by different references in order to compare the separated speech using different methods under the same acoustic condition. In detail, for the reverberant conditions, the reference is selected as the clean speech with the same level of reverberation rather than anechoic clean speech. The major improvement (approximately 1 against the third best) is achieved by the proposed dynamic threshold-based sparse components and stability-based non-sparse components separation. Specifically, compared with C-BSS, we achieve a better perceptual quality by using only a B-format microphone, while C-BSS adopts a pair.

4.1.2. Real Environment

In total, 36 sentences (sampling frequency 16 kHz) recorded in a room measuring $10 \times 5 \times 3$ m^3 (SNR = 20 dB, $RT60$ = 0.5 s) were utilized for the evaluation in a real environment. The average length of each recording is also about 8 s same as NTT [27] database. Based on the aforementioned conditions, a statistical analysis of $PESQ$ is taken. Statistical results are shown in Figure 10 with 95% confidence intervals.

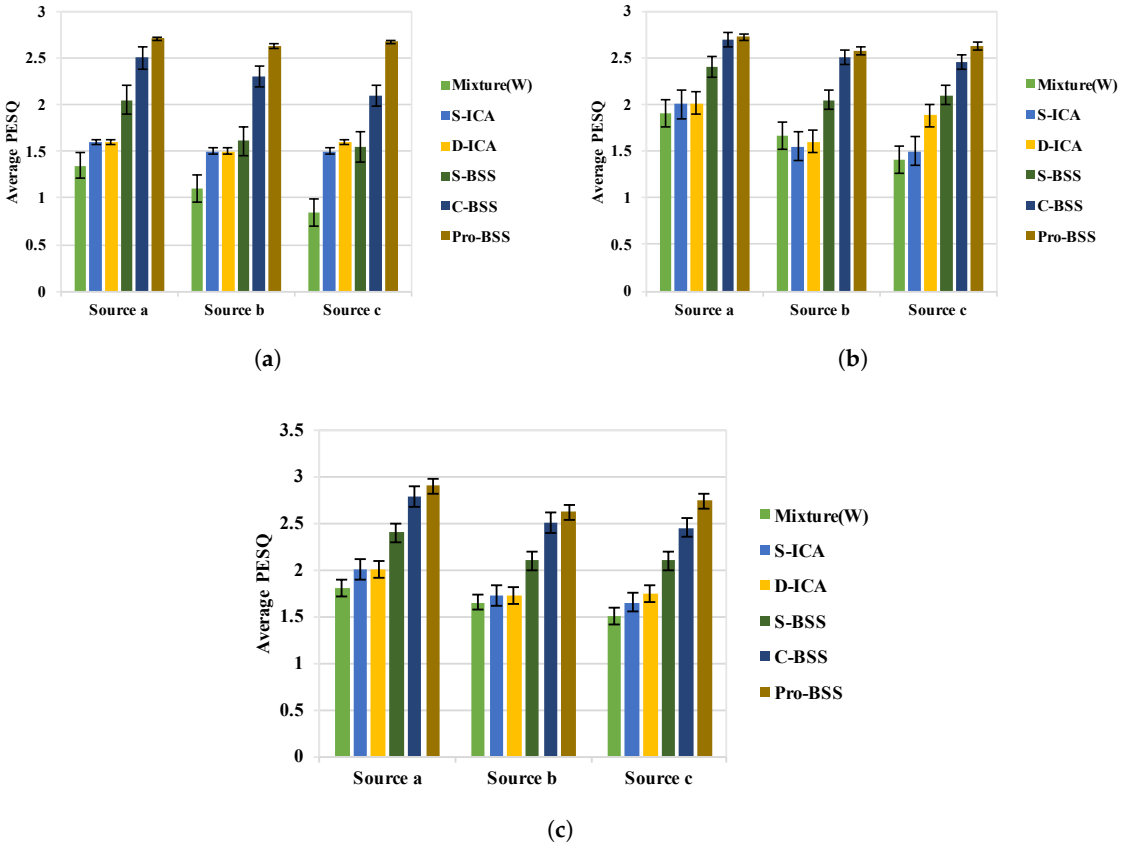

Figure 9. PESQ results (**a**) Anechoic Room; (**b**) Room1; (**c**) Room2. Error bars represent 95% confidence intervals.

From Figure 10, we can concluded that the proposed method greatly improved the perceptual quality of extracted sources. The corresponding SDR and SIR results are shown in Figures 11 and 12, respectively. Condition mixture (W) represents the B-format input signal S_W. Similar to the results in the simulated environment, the SDR and SIR results follow a similar trend to the PESQ results. Overall, it can be concluded that the proposed method obtain a great improvement of extracted sources.

4.2. Subjective Evaluation

Subjective evaluation consists of two major listening tests. For all cases (the overdetermined case, determined case, and underdetermined case), the perceptual quality of speech sources generated in section IV-A-1 corresponds to the case where $K = \{2, 3, 4, 5, 6\}$. The separation is $\{30°, 40°, 50°, 60°\}$, and the source radius is 1 m. Note that each separated speech source is evaluated separately by using headphones for playback. A MUSHRA [36] listening test which contains 16 listeners is employed to measure the subjective perceptual quality with four conditions, namely, Ref, Pro-BSS, Pro-SCS, SCS, and Anchor. Condition Ref refers to the original speech sources in each test, which are also served as the hidden references of this MUSHRA test. Condition SCS, Pro-SCS, and Pro-BSS are the same as in the objective evaluation. Condition Anchor is the unprocessed (W channel of the B-format recording) mixed signal. In total, 16 listeners participated in the test.

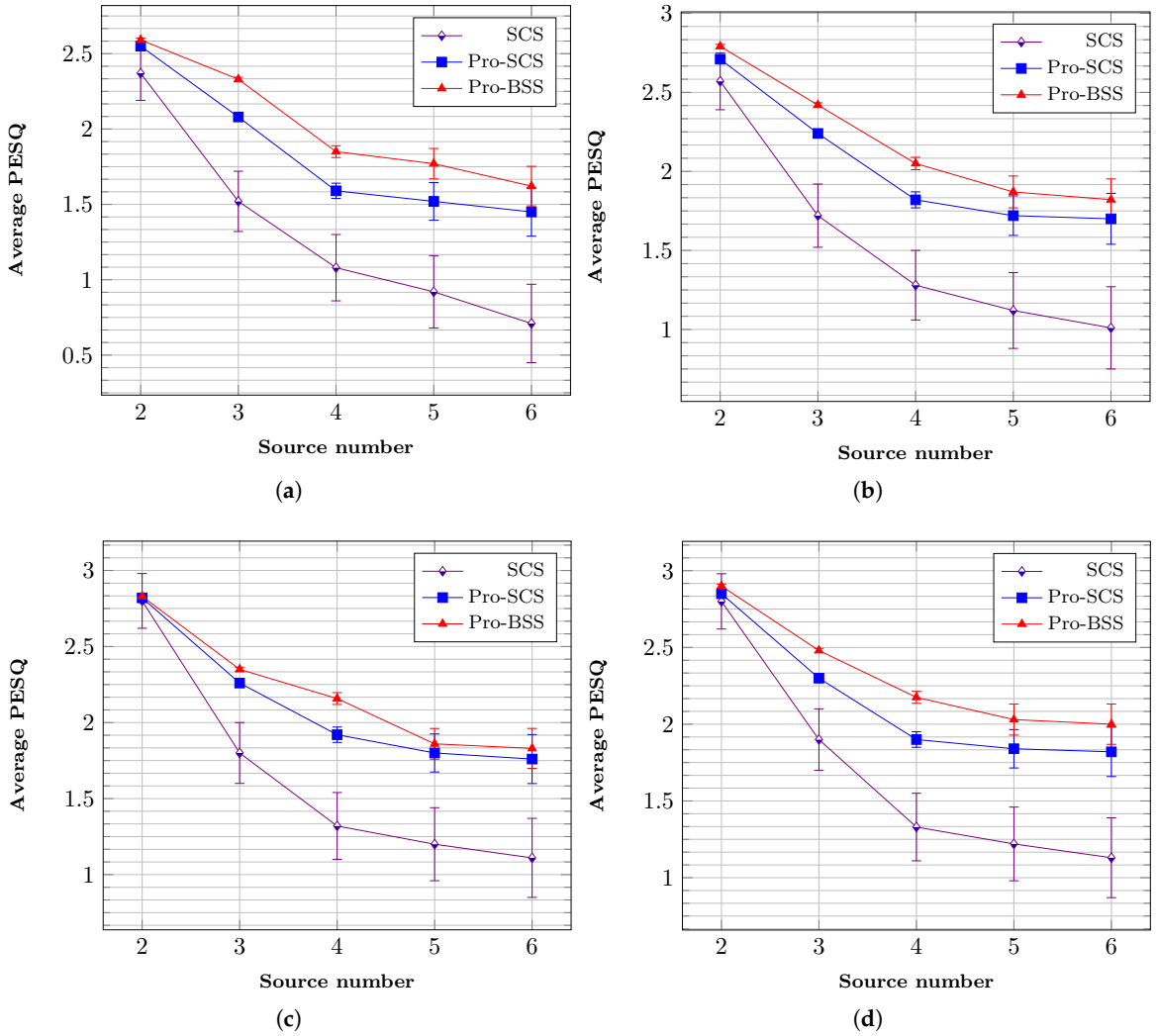

Figure 10. PESQ results. Error bars represent 95% confidence intervals, (**a**–**d**) represent the results for separations {30°, 40°, 50°, 60°}, SNR = 20 dB, $RT60$ = 0.5 s.

For each source number and separation, we calculated the average of all tested speeches and results are shown in Figures 13–16. It can be observed that our proposed method achieves significantly higher scores compared to the fixed threshold BSS approach, which uses the same number of microphones as our proposed method. In addition, the PSM scores decreases as the source number increases from 2 to 6, and rises as the separation between the two adjacent sources gets larger. For the cases K = 2, 3, the scores are always about 0.8, which reaches a nearly excellent quality. For the undermined case, the MUSHRA scores are about 0.7 when the source number is four, while for the cases K = 5, 6, the quality of the extracted speech is below 0.6 but still over 0.4. This means that the extracted speech is not quite euphonious but can still represent clear and understandable speech.

To compare the extracted speech quality with the reference method in objective evaluation further, a MUSHRA test was also employed to measure the subjective quality of the separated speech. Six middle sources from each test group were selected for the listening test. Similarly, the unprocessed (W channel of the B-format recording) mixed signal was used as the anchor and the original speech was used as the hidden reference. Note that each separated speech source is evaluated separately by using loudspeakers for playback in the Anechoic Room, Room 1, and Room 2. Average MUSHRA scores are presented in Figure 17.

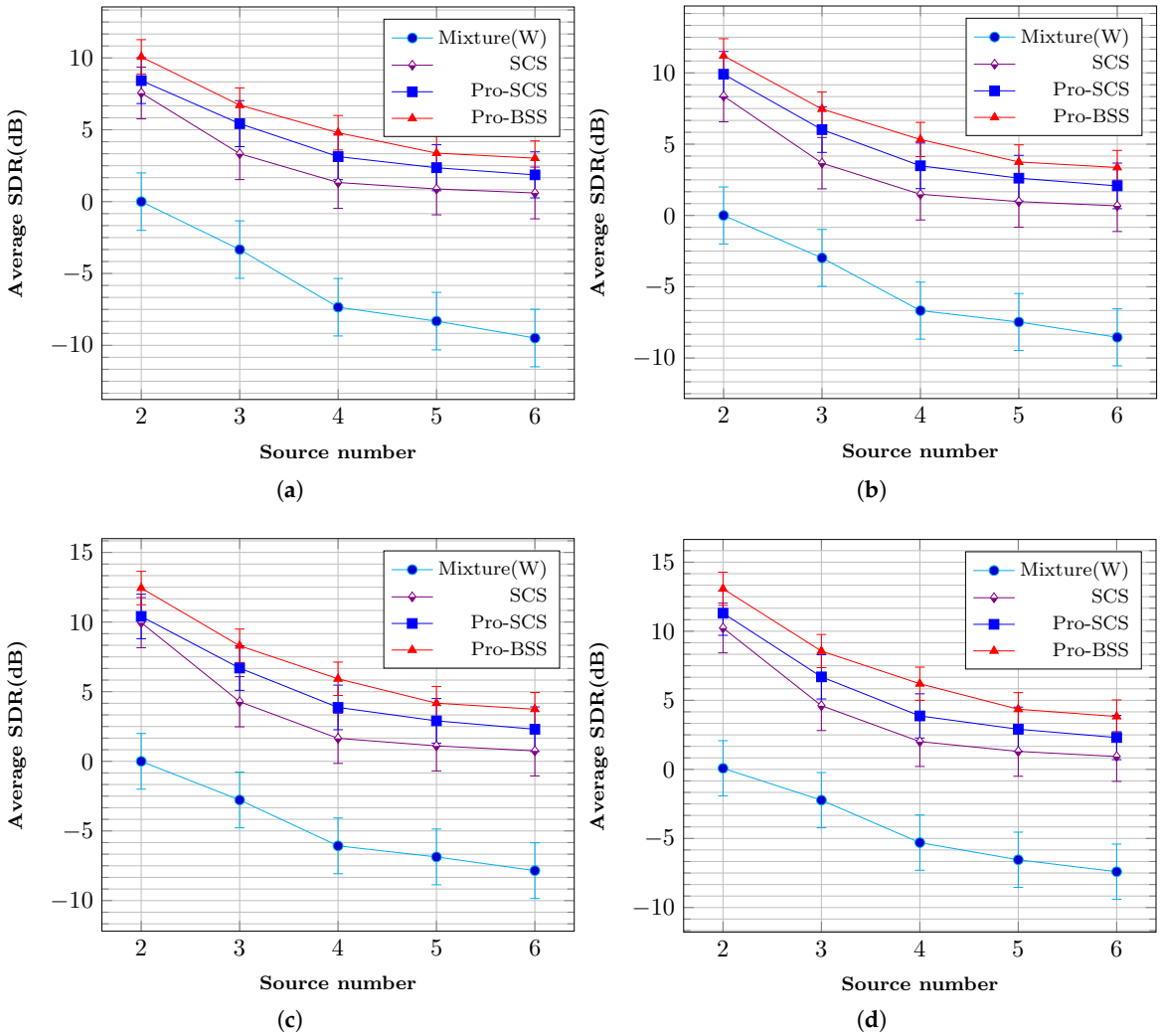

Figure 11. SDR results. Error bars represent 95% confidence intervals, (**a**–**d**) represent the results for separations {30°, 40°, 50°, 60°}, SNR = 20 dB, $RT60$ = 0.5 s.

It can be seen that a significant improvement in the separation quality is achieved by applying the proposed scheme. The MUSHRA score for the proposed method is of nearly 'excellent' quality, the second best score is about 'good'. It should be noted that we just use one B-format microphone, while C-BSS adopts a pair. The majority of listeners indicated that their choice for the closest match to the reference was based on files which contained the minimal amount of crosstalk and musical distortion. For other conditions, listeners reported that while the target speech is significantly separated from the mixture, there is audible crosstalk from other talkers with higher musical distortion.

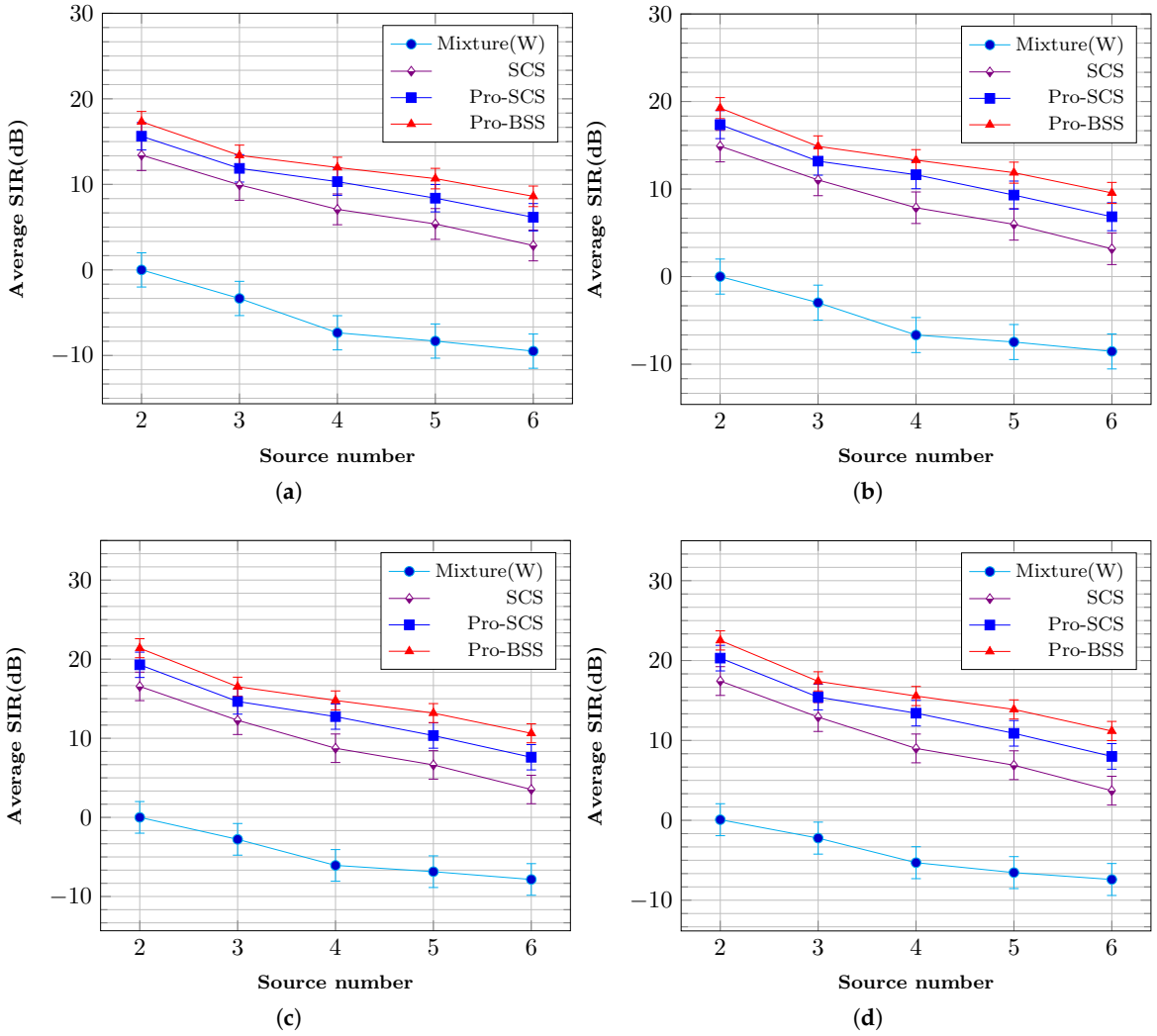

Figure 12. SIR results. Error bars represent 95% confidence intervals, (**a**–**d**) represent the results for separations {30°, 40°, 50°, 60°}, SNR = 20 dB, $RT60$ = 0.5 s.

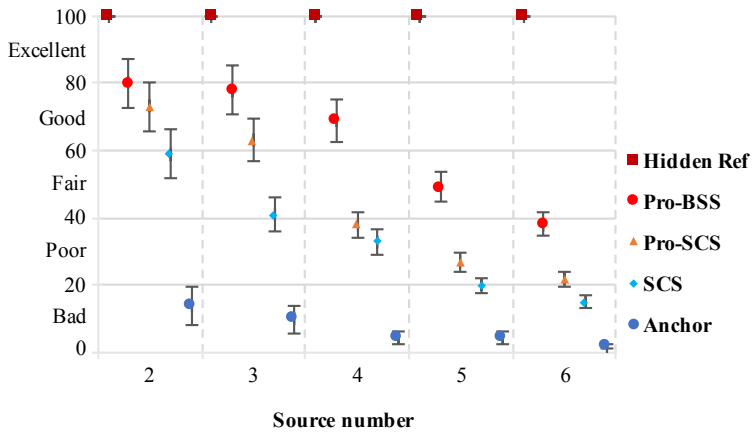

Figure 13. MUSHRA test results of *separation* = 30°. Error bars with 95% confidence intervals.

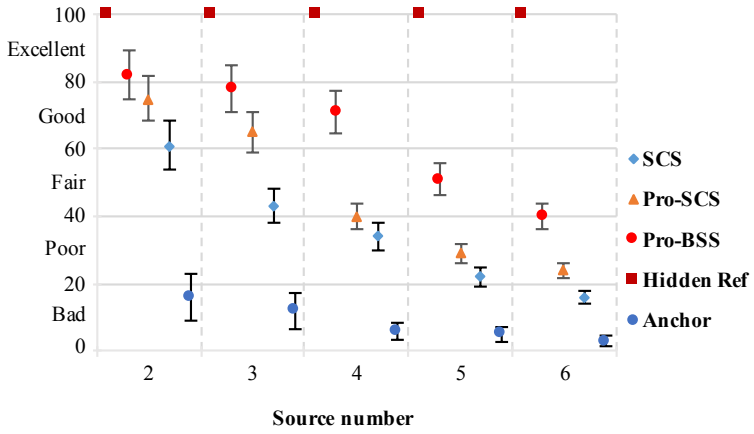

Figure 14. MUSHRA test results of $separation = 40°$. Error bars with 95% confidence intervals.

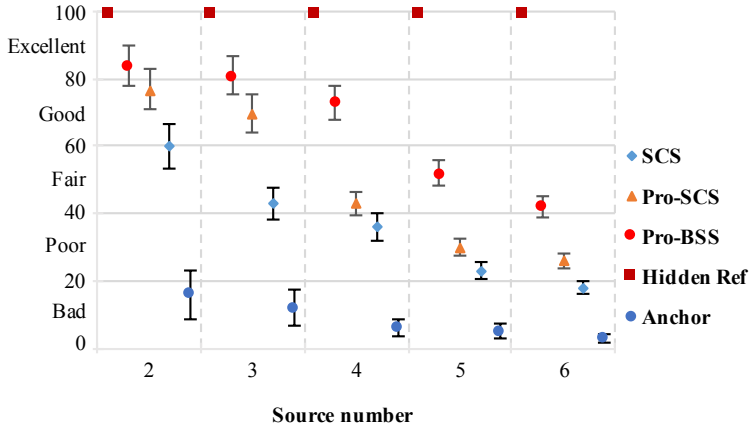

Figure 15. MUSHRA test results of $separation = 50°$. Error bars with 95% confidence intervals.

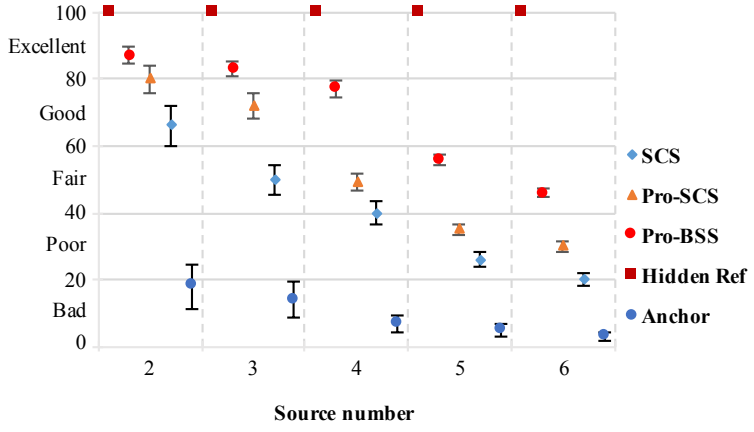

Figure 16. MUSHRA test results of $separation = 60°$. Error bars with 95% confidence intervals.

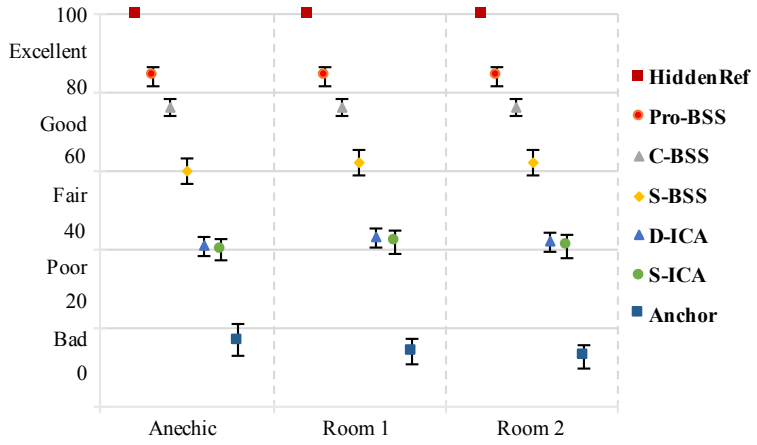

Figure 17. MUSHRA test results. Error bars with 95% confidence intervals.

5. Conclusions

A multiple speech source separation method using inter-channel correlation and the half-K assumption was proposed in this paper. To recover the sparse components, we proposed a dynamic threshold-based clustering algorithm where the threshold was determined by the inter-channel correlation among the recording signals of B-format microphone. Thereafter, a half-K assumption was proposed after conducting a statistical analysis of the number of the active source at each TF instant versus different number of sources. By applying this assumption, the non-sparse components were separated by regarding the extracted sparse components as a guide, jointly combined with vector decomposition and matrix factorization. Ultimately, the final TF coefficients of each source were recovered by the synthesis of sparse and non-sparse components. The approach has been evaluated via objective and subjective tests for both the anechoic and reverberant condition. Compared with the fixed threshold sparse components separation method, the proposed approach achieved significant improvement in the perceptual quality of separated sources. In addition, the comparison was also conducted with other BSS approaches. According to both objective and subjective evaluation, the proposed method achieved a better perceptual quality of separated sources than others.

Acknowledgments: This work has been supported by China Postdoctoral Science Foundation(2017M610731), the Project supported by Beijing Postdoctoral Research Foundation and "Ri xin" Training Programme Foundation for the Talents by Beijing University of Technology.

Author Contributions: Maoshen Jia and Jundai Sun contributed equally in conceiving the overall proposal, and critically reviewed and implemented the final revisions. Maoshen Jia and Jundai Sun supervised all aspects of this separation architecture, design and realization of the experiments, collection and analysis of the data, and writing of the manuscript. Xiguang Zheng critically reviewed and implemented the final revisions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zheng, X. *Soundfield Navigation: Separation, Compression and Transmission, Doctoral Dissertation*; University of Wollongong: Wollongong, Austral, 2013.
2. Asaei, A.; Taghizadeh, M.J.; Haghighatshoar, S.; Raj, B.; Bourlard, H.; Cevher, V. Binary Sparse Coding of Convolutive Mixtures for Sound Localization and Separation via Spatialization. *IEEE Trans. Signal Process.* **2016**, *64*, 567–579.
3. Jia, M.; Yang, Z.; Bao, C.; Zheng, X.; Ritz, C. Encoding Multiple Audio Objects Using Intra-Object Sparsity. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1082–1095.
4. Zheng, X.; Ritz, C.; Xi, J. Encoding and communicating navigable speech soundfields. *Multimed. Tools Appl.* **2016**, *75*, 5183–5204.

5. Hilpert, J.; Disch, S. The MPEG Surround Audio Coding Standard [Standards in a Nutshell]. *IEEE Signal Process. Mag.* **2009**, *26*, 148–152.
6. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949.
7. Petridis, S.; Pantic, M. Deep complementary bottleneck features for visual speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2304–2308.
8. Abou-Zleikha, M.; Tan, Z.H.; Christensen, M.G.; Jensen, S.H. A discriminative approach for speaker selection in speaker de-identification systems. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31August–4 September 2015; pp. 2102–2106.
9. Hu, Y.; Wu, D.; Nucci, A. Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 762–774.
10. Hyvärinen, A.; Hurri, J.; Hoyer, P.O. *Independent Component Analysis*; Cambridge University Press: Cambridge, UK, 2001; pp. 60–83.
11. Shashanka, M.V.S.; Raj, B.; Smaragdis, P. Sparse Overcomplete Latent Variable Decomposition of Counts Data. In Proceedings of the 17th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 12–15 December 2007; pp. 13–15.
12. Nesta, F.; Fakhry, M. Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; Volume 19, pp. 86–90.
13. Wang, L.; Ding, H.; Yin, F. A Region-Growing Permutation Alignment Approach in Frequency-Domain Blind Source Separation of Speech Mixtures. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 549–557.
14. Schölkopf, B.; Platt, J.; Hofmann, T. An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments. In Proceedings of the Twentieth Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems 19, Vancouver, BC, Canada, 3–6 December 2006; pp. 953–960.
15. Asaei, A.; Taghizadeh, M.J.; Bahrololum, M.; Ghanbari, M. Verified speaker localization utilizing voicing level in split-bands. *Signal Process.* **2009**, *89*, 1038–1049.
16. Taghizadeh, M.J.; Garner, P.N.; Bourlard, H.; Abutalebi, H.R. An integrated framework for multi-channel multi-source localization and voice activity detection. In Proceedings of the 2011 Joint Workshop on Hands-Free Speech Communication & Microphone Arrays, Edinburgh, UK, 31 May–1 June 2011; pp. 92–97.
17. Dmour, M.A.; Davies, M. A New Framework for Underdetermined Speech Extraction Using Mixture of Beamformers. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 445–457.
18. Itu-R Broadcasting Service Multichannel Stereophonic Sound System With and Without Accompanying Picture. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf (accessed on 4 December 2017).
19. Bosi, M.; Goldberg, R.E. *Introduction to Digital Audio Coding and Standards*; Kluwer Academic Publishers: Amsterdam, Holland 2003; pp. 399–400.
20. Zheng, X.; Ritz, C.; Xi, J. Collaborative Blind Source Separation Using Location Informed Spatial Microphones. *IEEE Signal Process. Lett.* **2013**, *20*, 83–86.
21. Jia, M.; Sun, J.; Bao, C. Real-time multiple sound source localization and counting using a soundfield microphone. *J. Ambient Intell. Hum. Comput.* **2017**, *8*, 829–844.
22. Eargle, J. *The Microphone Book: From Mono to Stereo to Surround–A Guide to Microphone Design and Application*; Taylor & Francis, London, UK, 2012.
23. Yılmaz, O.; Rickard, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.* **2004**, *52*, 1830–1847.
24. Gunel, B.; Hacihabiboglu, H.; Kondoz, A.M. Acoustic Source Separation of Convolutive Mixtures Based on Intensity Vector Statistics. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 748–756.
25. Pulkki, V. Spatial Sound Reproduction with Directional Audio Coding. *J. Audio Eng. Soc.* **2007**, *55*, 503–516.
26. Jia, M.; Sun, J.; Deng, F.; Sun, J. Single source bins detection-based localisation scheme for multiple speech sources. *Electron. Lett.* **2017**, *53*, 430–432.
27. NTT Database. Available online: http://www.ntt-at.com/product (accessed on 4 December 2017).

28. Shujau, M.; Ritz, C.H.; Burnett, I.S. Separation of speech sources using an Acoustic Vector Sensor. In Proceedings of the 2011 IEEE 13th International Workshop on Multimedia Signal Processing (MMSP), Hangzhou, China, 17–19 October 2011; pp. 1–6.
29. Campbell, D.; Palomaki, K.; Brown, G. A Matlab simulation of "shoebox" room acoustics for use in research and teaching. *Comput. Inf. Syst.* **2005**, *9*, 48.
30. Huber, R.; Kollmeier, B. PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1902–1911.
31. Vincent, E.; Gribonval, R.; Fevotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1462–1469.
32. Douglas, S.C.; Gupta, M.; Sawada, H.; Makino, S. Spatio-Temporal FastICA Algorithms for the Blind Separation of Convolutive Mixtures. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1511–1520.
33. Kim, T.; Attias, H.T.; Lee, S.Y.; Lee, T.W. Blind Source Separation Exploiting Higher-Order Frequency Dependencies. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 70–79.
34. Ono, N. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 16–19 October 2011; pp. 189–192.
35. Kitamura, D.; Ono, N.; Sawada, H.; Kameoka, H.; Saruwatari, H. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1626–1641.
36. ITU. Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf (accessed on 4 December 2017).