Estimating Number of Speakers via Density-based Clustering and Classification Decision

Junjie Yang Yi Guo,*Zuyang Yang, Liu Yang, and Shengli Xie †

1 Introduction

Audio source separation (ASS) targets at recovering multiple mixing speech sources recorded by multiple microphones [1–5]. Due to the existence of echoes in a real recording environment, the convolutive ASS is usually employed to depict the physical mixing mechanism of multiple speech source signals, where multiple speech sources are convolved from a sequence of delayed version of linear mixing system model [1, 3, 6]. In general, the mixing channels, including the system parameters such as the number of speakers (NoS), are unknown in advance. Therefore, it is essential to estimate NoS from a recorded mixture signals in the convolutive ASS [7, 8]. black The NoS estimation can be categorized as the model selection problem in machine learning, i.e., selecting the optimal model from a set of potential models as the best representation of data set [9, 10]. Some popular methods of model selection also can be found in the literatures of [11–13]. Here, the model selection of convolutive ASS is to find the best classification of speakers from the recorded mixtures, where multiple speech sources are convolved from a multiple delay mixing system. In this paper, we mainly focus on the NoS estimation problem based on the time-frequency (TF) domain.

In the area of NoS estimation, some works resort to developing various statistical methods to estimate NoS under the reverberant scenario. In [14], the independent component analysis (ICA) and a scaling technique are combined to estimate the power of speech component and noise component in order to distinguish speeches and noises. Then the correlation of component envelops are calculated to estimate the NoS. Authors in [15] exploit various time delays from the multi-speaker signals and count the NoS estimated from the cross-correlation of the Hilbert envelops of the linear prediction residuals of the mixtures. Based on statistical model, [16] provides a clustering method named DEMIX to exploit a local confidence measurement for the NoS detection. However, this method is restricted to the non-reverberant case because the room reverberation may affect the clustering performance and result in erroneous estimation of NoS. Authors in [17] transform the NoS estimation into a sparse recovery problem by fitting the direction of arrival histogram with von-Mises density functions. black In addition, various methods based on deep learning on counting the NoS have emerged, such as [18–21]. In [22], a new NoS estimation architecture is provided via combining the convolutional recurrent neural networks and adequate input features of speeches, which is designed to improve the performance of NoS estimation from the single channel mixtures.

Several works try to introduce additional assumption such as sparsity on the speeches in designing various NoS estimation algorithms. A common assumption, namely, approximately Window-Disjoint Orthogonal (WDO) [2, 3, 23, 24] plays an important role in such NoS estimation methods. The WDO condition assumes that only one speech component is active while other components are silent at each TF slot. Based on this assumption, [25] tries to cluster the mixtures via a validation index combining compactness and separation of cluster centers to determine the NoS. However, the WDO assumption may not hold in practical circumstance since a highly room reverberation may result in the internal interference problem as the mixtures are generated by overlapping the speech components. Several works try to solve this problem by relaxing WDO into a weaker version, e.g., local dominance assumption [16,26,27]. The local dominance assumption stems from the observation that the spectrum

*Y. Guo is with the Centre for Research in Mathematics, School of Computing, Engineering and Mathematics, Western Sydney University, Parramatta, NSW, 2150, Australia. (E-mail: Y.Guo@westernsydney.edu.au).

†J. Yang, Z. Yang, L. Yang and S. Xie are with the School of Automation, Guangdong University of Technology, and Guangdong Key Laboratory of IoT Information Processing, Guangzhou, 510006, China. (E-mail: yangjunjie1985@gmail.com, yangzuyuan@aliyun.com, willow_gao@126.com,shlxie@gdut.edu.cn).

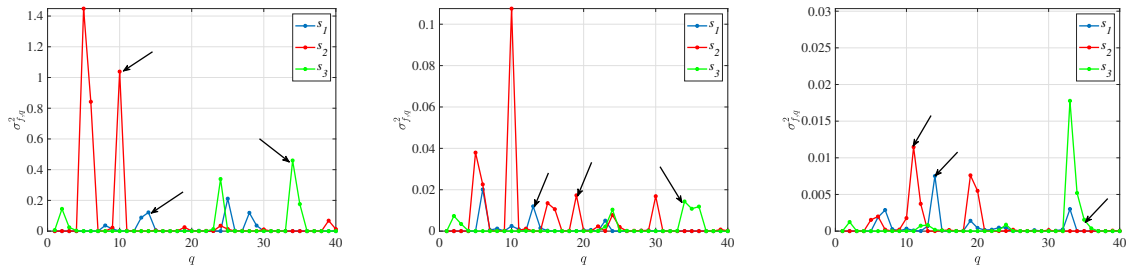


Figure 1: black Illustration of local dominance assumption ($N = 3$). (a) $f = 100$, (b) $f = 200$, (c) $f = 300$.

of each speech component is at least locally dominant once, i.e., one component is active while the others are silent, in one short period of successive TF slots. By exploiting the local dominance assumption, [28] converts the NoS estimation problem to the rank identification of the correlation matrix constructed from a probabilistic model, which means additional statistical assumption is required for this type of method. In general, the NoS estimation in a reverberant environment is still a tricky problem, especially when the room reverberation is high.

In this paper, we transform the estimate of NoS into a clustering problem with the consideration of robustness. Based on the assumption of local dominance, a combinational NoS detector is proposed and consists of three steps. First, the leading eigenvectors of local covariance matrices of mixtures (one eigenvector per covariance matrix) are extracted and ranked based on two factors, i.e., local density and minimum distance to other eigenvectors with higher density. Second, the cluster centers are detected from the ranked eigenvectors by a gap-based detector. Third, a criterion based on the averaged volume of cluster centers is used to select reliable clustering results in some frequency bins for the final estimation of NoS.

The main contributions of this paper are as follows:

A1. The proposed density-based clustering exploits the local dominance assumption by clustering the leading eigenvectors from the local covariance matrices. It has been demonstrated that this strategy is less sensitive to the interference of reverberation under various NoS estimation experiments.

A2. The NoS estimation strategy has extended our previous work in [29], which is further enhanced by combining some best clustering results in some frequency bins. This superiority of clustering is indicated by the averaged volume of cluster centers, which further improves the performance of NoS estimation.

The remainder of this paper is organized as follows. First, the system model and assumptions are discussed in Section II. Next, the proposed algorithm of NoS estimation is provided in Section III. Experimental results are presented in section IV. Finally, conclusions are drawn in Section V.

2 System Model and Problem Description

We consider the mixing speeches problem in a reverberant scenario where N speech sources are recorded by M microphones. Let $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$, $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^T$ and $\mathbf{e}(t) = [e_1(t), \ldots, e_M(t)]^T$ denote the mixture signals, source signals and background noise signals, respectively. The speeches and noises are assumed to be uncorrelated in statistics. With the above notations, we consider the ASS problem based on a convolutive linear system model, i.e.,

$$\omega\mathbf{x}(t) = \mathbf{H} \star \mathbf{s}(t) = \sum_{\tau=0}^{L-1} \mathbf{H}(\tau)\mathbf{s}(t-\tau) + \mathbf{e}(t), \tag{1}$$

where $\star$ is linear convolutive operator, $\mathbf{H}(\tau) \in \mathbb{R}^{M\times N}$ is the mixing matrix at time lag τ, L is the channel order and large L indicates higher reverberation of the room. The elements of $\mathbf{H}(\tau)$, denoted by $h_{i,j}(\tau)$, are the Room Impulse Response coefficients (RIRs) from the jth source to the ith microphone.

In the short-time Fourier transformation (STFT) domain with a window length F, the mixing process of speeches can be approximately depicted by a multiplicative narrow-band model [16, 30] in TF domain, i.e.,

$$\mathbf{x}_{f,d} = \mathbf{H}_f \mathbf{s}_{f,d} + \mathbf{e}_{f,d}, \tag{2}$$

where $\mathbf{x}_{f,d} = [x_{f,1d},\ \dots, x_{f,Md}]^T \in \mathbb{C}^M$, $\mathbf{s}_{f,d} = [s_{f,1d},\ \dots, s_{f,Nd}]^T \in \mathbb{C}^N$ and $\mathbf{e}_{f,d} = [e_{f,1d},\ \dots, e_{f,Md}]^T \in \mathbb{C}^M$ are the vectors of mixture, speech and noise signal components at TF slot (f, d), respectively. Here, $f \in \{0, .., F-1\}$ is the frequency bin, and $d \in \{1, ..., D\}$ is the time frame index; $\mathbf{H}_f \triangleq [\mathbf{h}_{f,1}, ..., \mathbf{h}_{f,N}]$ is an $M \times N$ mixing matrix at frequency bin f, and its columns are called as steering vectors. The objective of this paper is to estimate the number of speakers (NoS) which is unknown in a priori by counting the number of steering vectors from the system model (2).

black In general, the speeches are assumed to be uncorrelated and wide-sense quasi-stationary in a short time, e.g., 40ms-80ms [31, 32]. Let P denote the number of frames in a short time period (namely sub-block) and let the mixture TF vectors be divided into Q sub-blocks, where $Q = \lfloor D/P \rfloor$ and $\lfloor \cdot \rfloor$ is round down operator. Here, P should be selected at a proper range to the wide-sense quasi-stationarity of speeches at each sub-block. The selection of P will be discussed in the experimental section. Calculate the local covariance matrix of size $M \times M$ by

$$\mathbf{R}^{\mathbf{x}}_{f,q} \triangleq \frac{1}{P}\Big(\sum_{d \in \Omega_q} \mathbf{x}_{f,d}\mathbf{x}^H_{f,d}\Big), q = 1, ..., Q, \tag{3}$$

where Ω_q defines the set of frame indices at qth sub-block , i.e., $\Omega_q = \{q(P-1)+1, .., qP\}$ and the cardinality of Ω_q is P; $(\cdot)^H$ denotes the Hermitian transpose. Assuming the speeches are independently distributed, $\mathbf{R}^{\mathbf{x}}_{f,q}$ can be approximately expanded as follows [33]:

$$\mathbf{R}^{\mathbf{x}}_{f,q} = \sum_{i=1}^{N} \sigma^2_{f,iq}\mathbf{h}_{f,i}\mathbf{h}^H_{f,i} + \varepsilon^2_{f,iq}, \tag{4}$$

where the local variance of source and noise are

$$\begin{cases} \sigma^2_{f,iq} \triangleq \sum\limits_{d \in \Omega_q} s_{f,id}s^*_{f,id}, \\ \varepsilon^2_{f,iq} \triangleq \sum\limits_{d \in \Omega_q} e_{f,id}e^*_{f,id}, \end{cases} \tag{5}$$

and $\sigma^2_{f,iq} \gg \varepsilon^2_{f,iq}, i = 1, ..., N$; $(\cdot)^*$ refers to complex conjugation.

The main assumption on the system model is as follows:

A1. For each speech component $s_i(t)$ at each frequency bin f, there exists at least one sub-block indexed by $\psi_i \in \{1, ..., Q\}$, such that $\sigma^2_{f,i\psi_i} > 0$ and $\sigma^2_{f,j\psi_i} = 0$ for all $j \neq i$, $i = 1, ..., N$.

black A1 is called as the local dominance assumption which stems from [16, 26, 27, 34]. Here, an example is shown in FIGURE 1 (a)-(c), to illustrate the local dominance assumption. Note that from the frames pointed out by the text arrows, it can be seen that only the local variance of one speech is non-zero while the other's local covariances are zeros in the TF domain. Such features can be observed at the majority of frequency bins. In addition, there may be more than one active speech component in other frames, which is not allowed in WDO. Thus, the local dominance assumption is much weaker than the WDO condition [2, 3] as it assumes that each speech component dominates in at least one sub-block (called as singular sub-block) within successive TF slots. This assumption provides a new perspective to estimate NoS via the clustering strategy.

Under such assumption, the covariance matrix at the ith singular sub-block has a rank-one structure, i.e.,

$$\begin{aligned} \mathbf{R}^{\mathbf{x}}_{f,\psi_i} &= \sigma^2_{f,\psi_i}\mathbf{h}_{f,i}\mathbf{h}^H_{f,i} + \varepsilon^2_{f,iq}, \\ &\approx \sigma^2_{f,\psi_i}\mathbf{h}_{f,i}\mathbf{h}^H_{f,i}, i = 1, ..., N. \end{aligned} \tag{6}$$

Hence, we can extract all of the leading eigenvectors from the local covariance matrices and clustering them into various groups, where the number of cluster centers indicates the NoS.

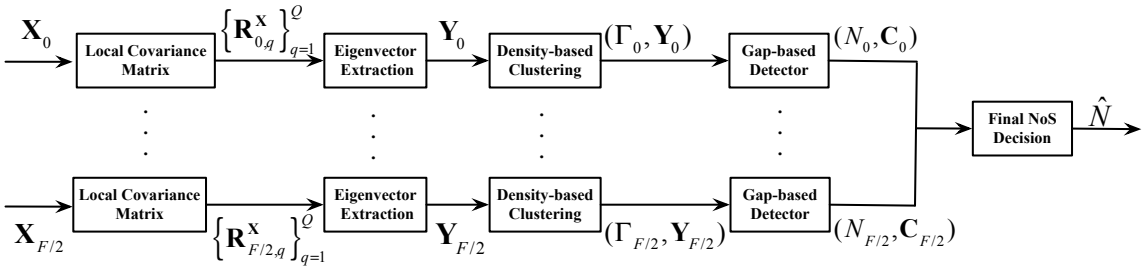
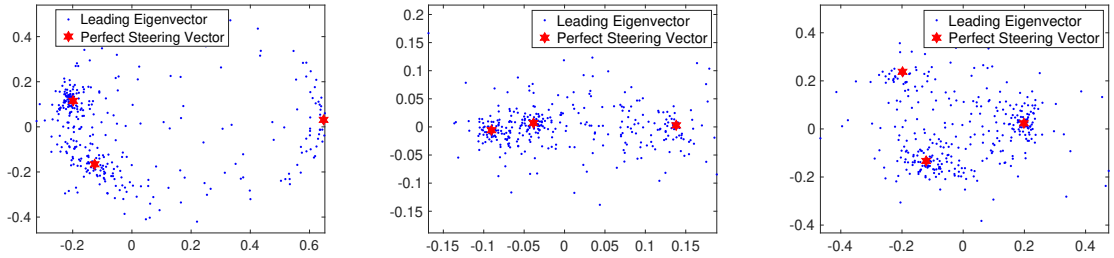


Figure 2: black Flow chart of proposed density-based clustering scheme.



(a) (b) (c)

Figure 3: black Scatter plot of leading eigenvectors for $Case(M, N) = (2, 3)$ based on Database A. (a) f=100, (b)f=200, (c)f=300.

3 Proposed Clustering and Decision Scheme

As shown in FIGURE 2, the proposed NoS estimation scheme will be performed at each frequency bin, which includes five steps as follows:

Step 1. Utilize the mixture TF components $\mathbf{X}_f \triangleq [\mathbf{x}_{f,1}, ..., \mathbf{x}_{f,D}]$ to calculate a sequence of local covariance matrices denoted by $\{\mathbf{R}^{\mathbf{x}}_{f,q}\}_{q=1}^{Q}$;

Step 2. Extract the leading eigenvectors wisely from $\{\mathbf{R}^{\mathbf{x}}_{f,q}\}_{q=1}^{Q}$ to give $\mathbf{Y}_f \triangleq [\mathbf{y}_{f,1}, ..., \mathbf{y}_{f,Q}]$;

Step 3. Cluster $\mathbf{Y}_f$ to give scores $\mathbf{\Gamma}_f \triangleq [\gamma_{f,1}, ..., \gamma_{f,Q}]$;

Step 4. Utilize $\mathbf{\Gamma}_f$ and $\mathbf{Y}_f$ to give the number of clusters N_f and cluster centers $\mathbf{C}_f \triangleq [\mathbf{c}_{f,1}, ..., \mathbf{c}_{f,N_f}]$, respectively;

Step 5. Determine the final NoS by integrating the clustering results from a selected frequency bins.

It is worth noting that the first 4 steps are repeated from $f = 0$ to $f = F/2$. The details of each step are provided in the following parts.

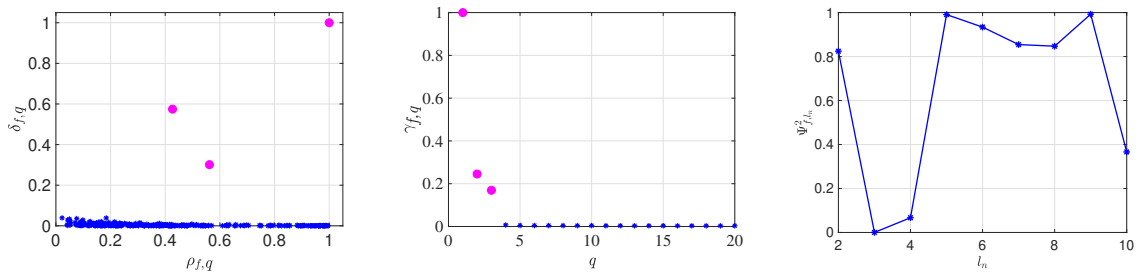
3.1 Leading Eigenvector Extraction

The eigenvector decomposition (EVD) is performed for each local covariance matrix $\mathbf{R}^{\mathbf{x}}_{f,q}$, i.e.,

$$\mathbf{R}^{\mathbf{x}}_{f,q} = \mathbf{U}_{f,q}\mathbf{\Lambda}_{f,q}\mathbf{U}^{H}_{f,q}. \tag{7}$$

The eigenvector of $\mathbf{U}_{f,q}$ corresponding to its largest eigenvalue is extracted as the leading eigenvector, i.e., $\mathbf{y}_{f,q} \triangleq \mathbf{U}_{f,q}(:, 1)$, $q = 1, ..., Q$. Based on the assumption of local dominance, the link between the leading eigenvectors at the singular sub-blocks of $\{\psi_i\}_{i=1}^{N}$ and the steering vectors of $\mathbf{H}_f$ is given by

$$\mathbf{y}_{f,\psi_i} = \frac{1}{\|\mathbf{h}_{f,i}\|_F^2}\mathbf{h}_{f,i}, \quad i = 1, ..., N. \tag{8}$$



(a) (b) (c)

Figure 4: Illustration of density-based clustering for $Case(M, N) = (2, 3)$, Database A. (a) Decision graph, (b) Score graph, (c) Curve of Ψ^2_{f,l_n}.

In (8), the leading eigenvector $\mathbf{y}_{f,\psi_i}$ represents a rescaling version of steering vector. If we can extract such local covariance vectors $\{\mathbf{y}_{f,\psi_i}\}_{i=1}^{N}$ from the set of $\{\mathbf{y}_{f,q}\}_{q=1}^{Q}$, we can classify the directions of steering vectors and count the number of clusters. Next, a density-based clustering scheme is employed to identify $\mathbf{y}_{f,\psi_i}$ from $\mathbf{Y}_f \triangleq [\mathbf{y}_{f,1}, ..., \mathbf{y}_{f,Q}]$.

3.2 Density-based Clustering

black To begin with, we show that the density-based clustering method [35] is suitable to estimate NoS from the leading eigenvectors. For a better observation on the distribution of leading eigenvectors, the leading eigenvectors are transformed from $\mathbf{Y}_f$ into a two-dimensional space while the Euclidean distance of pair-wise leading eigenvectors. The function 'mdscale' in Matlab is employed to illustrate the distribution of leading eigenvectors are maintained. As shown in FIGURE 3 (a)-(c), the blue point and red point refer to the relative spatial position of leading eigenvector and perfect steering vector. It can be observed that there are two factors of perfect steering vector in the scatter plot: 1) each perfect steering vector has a high local density of leading eigenvectors; 2) each perfect steering vector is far away from other steering vectors. In fact, such characteristics can be observed across the major part of frequency bins, which is determined by the sparsity of speech signals and the independence of steering vectors [27]. Specifically, the speech signals are usually sparse and local dominant at some singular sub-blocks. Thus, the leading eigenvectors might concentrate nearby the perfect steering vector, i.e., higher local density of perfect steering vector. Moreover, the distribution of perfect steering vector is determined by the RIRs function, which is usually assumed to be independent to all of steering vectors. Thus, the perfect steering vector should be far away from other steering vectors, i.e., large distance of any points with higher density. As described in [35], the cluster centers in a data set are: locally dense and far from other centers. Based on these two distinct features, we try to classify the leading eigenvectors via the density-based clustering to rank the clusters from the data samples.

First, for each eigenvector, its local density and minimal distance to other potential centers are computed in $\mathbf{Y}_f$. These two factors are calculated separately and then integrated into the classification decision of cluster centers. We denote a pair-wise distance matrix as Φ_f, whose component is calcuated by

$$\phi_{f,qk} \triangleq \| \mathbf{y}_{f,q} - (\mathbf{y}_{f,q}^H \mathbf{y}_{f,k}) \mathbf{y}_{f,k} \|_F^2, \quad q, k = 1, .., Q, \tag{9}$$

where $\| \cdot \|_F$ is Frobenius norm.

Second, the local density $\rho_{f,q}$ of $\mathbf{y}_{f,q}$ is computed by using the sum of Gaussian kernel functions,

$$\rho_{f,q} \triangleq \sum_{k \neq q} e^{-\frac{\phi_{f,qk}}{\tau_c^2}}, \tag{10}$$

where τ_c is the bandwidth or the cutoff distance. The local density in (10) is identical to the kernel density estimator used frequently in statistics. For the bandwidth of the Gaussian kernel, one use AMISE (asymptotic mean integrated squared error) to find its optimal value [36]. However, in the proposed algorithm, the parameter

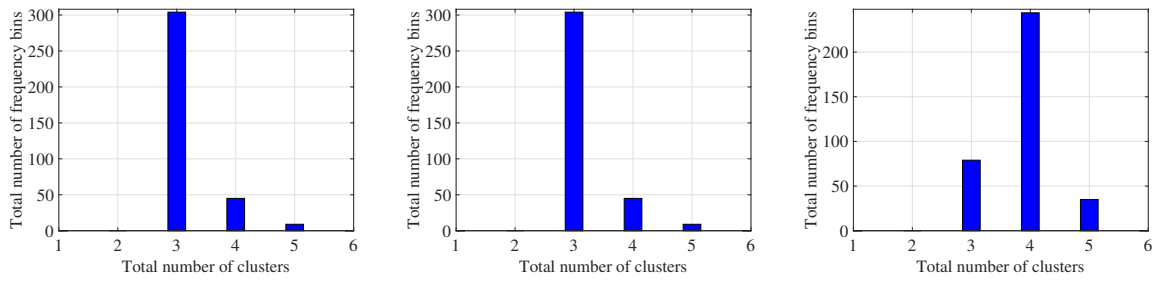


Figure 5: Histogram of NoS in selected frequency bins (frequency bins with top 35% of largest V_f), $\mathrm{RT}_{60} = 150ms$, Database A. (a) $Case(M,N) = (2,3)$, (b) $Case(M,N) = (3,3)$, (c) $Case(M,N) = (3,4)$.

τ_c is empirically chosen to account for around 6% to 8% of the distances in [35]. In our experiment, it works well to calculate the local density from leading eigenvectors of $\mathbf{Y}_f$.

Third, for each eigenvector $\mathbf{y}_{f,q}$, we obtain the minimal distance between $\mathbf{y}_{f,q}$ and other eigenvectors with higher local density by

$$\delta_{f,q} = \min_k \ (\phi_{f,qk}), \forall k : \rho_{f,k} > \rho_{f,q} \tag{11}$$

as the minimal distance to other potential centers. Note that the eigenvector with the highest local density, denoted by $\mathbf{y}_{f,q^*}$, has the assigned δ_{f,q^*} as the largest distance in Φ_f. Finally, $\rho_{f,q}$ and $\delta_{f,q}$ are combined to obtain a score for ranking the leading eigenvector of $\mathbf{y}_{f,q}$, i.e.,

$$\gamma_{f,q} = \rho_{f,q} \times \delta_{f,q}. \tag{12}$$

The score reflects the possibility of leading eigenvector being a cluster center, i.e., the higher the score of $\gamma_{f,q}$ is, the more likely the leading eigenvector being a cluster center for $\mathbf{y}_{f,q}$, and vice versa.

3.3 Gap-based Detector

Inspired by the work of [37], we utilize the scores (12) to identify the cluster centers via a gap-based detector. The scores are sorted in descending order as follows,

$$\Gamma_f \triangleq [\gamma_{f,l_1}, ..., \gamma_{f,l_Q}], \forall\, l_i \in \{1, ..., Q\}, i = 1, ..., Q. \tag{13}$$

It is assumed that the scores of true and fake cluster centers should satisfy that

$$\gamma_{f,l_1} \geq \cdots \geq \gamma_{f,l_N} > \gamma_{f,l_{N+1}} = \cdots = \gamma_{f,l_Q} = \eta, \tag{14}$$

where η is a small value and there exists a noticeable gap between γ_{f,l_N} and $\gamma_{f,l_{N+1}}$. Based on this observation, the number of clusters can be detected by searching the gap in the following manner. First, the difference of neighboring scores is caculated by

$$\triangle\gamma_{f,l_i} = \gamma_{f,l_i} - \gamma_{f,l_{i+1}}, i = 1, ..., Q-1. \tag{15}$$

Second, the variance of $\triangle\gamma_{f,l_n}$ is caculated by

$$\psi^2_{f,l_n} = \frac{1}{Q-n}\sum_{i=n}^{Q-1}(\triangle\gamma_{f,l_i} - \frac{1}{Q-n}\sum_{i=n}^{Q-1}\triangle\gamma_{f,l_i})^2. \tag{16}$$

Third, we define the ratio of neighboring variance of ψ^2_{f,l_n} as follows,

$$\Psi^2_{f,l_n} \triangleq \frac{\psi^2_{f,l_{n+1}}}{\psi^2_{f,l_n}}. \tag{17}$$

Then, the number of clusters at frequency bin f can be determined by

$$N_f = \underset{n=1,...,Q-2}{\arg\min} \Psi^2_{f,l_n}. \tag{18}$$

When N_f is determined, we further identify the clusters by extracting the leading eigenvectors with top N_f scores such that $\mathbf{C}_f \triangleq [\mathbf{y}_{f,l_1}, ..., \mathbf{y}_{f,l_{N_f}}]$.

As an example, it can be seen in FIGURE 4 (a)-(c) that the number of cluster centers can be automatically detected in $Case(M, N) = (2, 3)$ when $f = 52$. FIGURE 4 (a) provides a decision graph of $\rho_{f,q}$ and $\delta_{f,q}$ given by (10) and (11). It is observed that three singular leading eigenvectors are distinctive from other leading eigenvectors with higher ρ and δ. FIGURE 4 (b) illustrates the score of leading eigenvector given by (12) and it can be seen that three singular leading eigenvectors occupy the top scores. FIGURE 4 (c) shows the curve of Ψ^2_{f,l_n} given by (17) and it can be seen that the index with the lowest value correctly indicates NoS at frequency bin f.

Algorithm 1 Implementation of NoS Estimation

Input: $\mathbf{X}_f = [\mathbf{x}_{f,1}, \ldots, \mathbf{x}_{f,D}]$,$f = 0, ..., F/2$, $d = 1, ..., D$.
for $f = 0$ to $F/2$ **do**
  for $q = 1$ to Q **do**
    Calculate $\hat{\mathbf{R}}^{\mathbf{x}}_{f,q}$ by (3).
    Calculate $\mathbf{y}_{f,q}$ by (7).
  end for
  Calculate similarity matrix $\mathbf{\Phi}_f$ by (9).
  for $q = 1$ to Q **do**
    Calculate $\rho_{f,q}$ by (10).
    Calculate $\delta_{f,q}$ by (11).
    Calculate $\gamma_{f,q}$ by (12).
  end for
  Calculate $\triangle\gamma_{f,l_i}$ by (15).
  Calculate ψ^2_{f,l_n} by (16).
  Calculate Ψ^2_{f,l_n} by (17).
  Calculate N_f by (18).
  Calculate V_f by (19).
end for
Output: Select most frequent occurrence of $\{N_f\}_{f=0}^{F/2}$ from a preset portion (e.g. 35%) of top $\{V_f\}_{f=0}^{F/2}$ as the final decision of $\hat{N}$.

3.4 Final NoS Decision

It is worth noting that the energy of speech TF components is usually strong in certain frequency bins, e.g., lower frequency bins, while the energy is relative weak in other frequency bins. Hence, in the final NoS decision, it is necessary to select reliable results from some frequency bins. It is observed that the distribution of cluster centers can substantially impact the performance of gap-based detection. For example, if the cluster centers are far from each other, the gap between the true cluster centers and fake ones is much more apparent, which also means that the energy of speech components are more concentrated at this frequency bin. Based on this observation, we define the following confidence measurement [38],

$$V_f = \frac{\det|\mathbf{C}_f \mathbf{C}_f^H|}{N_f}, \tag{19}$$

where $\det|\cdot|$ refers the determinant operation. V_f can be interpreted as the average volume of cluster centers of $\mathbf{C}_f$. The larger the V_f is, the further apart the centers, and therefore the more reliable NoS decision. By using the measurement V_f, we select a preset portion of frequency bins (e.g. 35%) with the highest confidence and the most frequent estimate of NoS in these frequency bins is the final NoS.

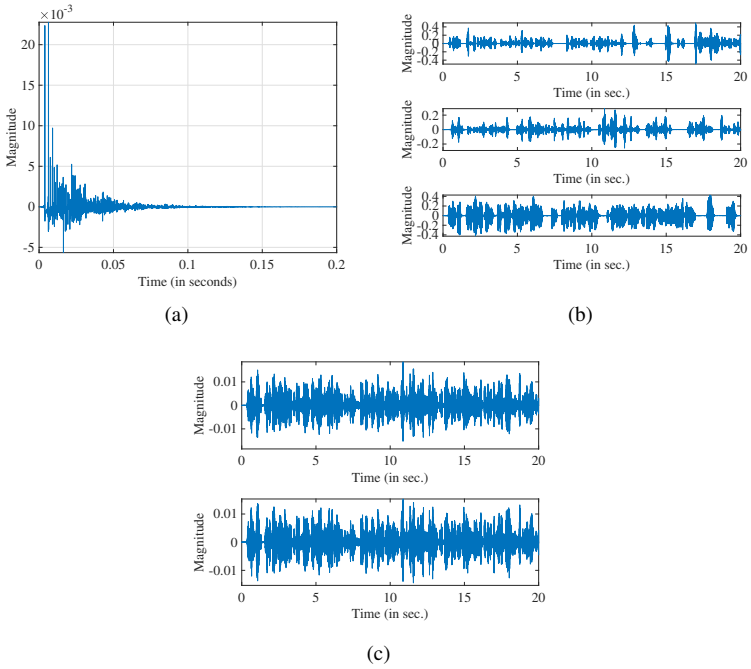


Figure 6: Illustration of Database A. (a) RIRs from the first speaker to the first microphone (200ms), (b) Speeches, (c) Mixtures.

FIGURE 5 (a)-(c) illustrate an example in terms of the histograms of estimated number of clusters in selected frequency bins (top 35% with highest V_f) for $Case(M, N) = (2, 3)$, $(3, 3)$, $(3, 4)$, respectively. We see that the index corresponding to the highest rate of occurrence corresponds to the correct NoS in all cases. The proposed confidence measurement of NoS provides an effective evaluation to enhance the performance of NoS estimation. This criterion enforces the final NoS decision performance based on the reliable clustering results from those frequency bins with higher averaged volume. black Finally, the implementation of NoS estimation is concluded in Algorithm 1.

4 Experiment Results

4.1 Experiment Settings

We briefly introduce the experiment settings used for the proposed algorithm. First, two public databases are introduced as follows.

- **Database A** is a public benchmark audio database provided in [39], where the pure speeches are recorded from a number of female and male speakers with a sampling rate $F_s = 16$ kHz. In this database, there are 16 clean speeches independently recorded by 8 male speakers and 8 female speakers. The duration of each speech signal is set as 20 seconds. black The room size is set as 4.45m×3.55m×2.5m. The location of microphones are (2m, 2.5m, 1.155m), (2m, 2.55m, 1.155m) and (2m, 2.6m, 1.155m), respectively. The location of speakers are (3.2m, 2.0m, 1.6m), (3.2m, 2.4m, 1.6m), (3.2m, 2.8m, 1.6m) and (3.2m, 3.2m, 1.6m), respectively. The artificial function [40] are utilized to simulate Room Impulse Response coefficients (RIRs) by setting various parameter of RT_{60}, e.g., 100ms, 150ms, 200ms, 250ms, respectively. (RT_{60} is defined as the transmission time of signal decay by 60 dB, which is crucial to reflect the reverberation of the room.) These clean speeches are convolved with the generated RIRs to give mixtures by (1). The RIRs function, three clean speeches and two-channel mixtures are illustrated in FIGURE 6 (a)-(c), respectively. black Three typical cases with various speakers and microphones are tested in this database, i.e., $Case(M, N) = (2, 3)$, $(3, 3)$ and $(3, 4)$, respectively. In each case, N speeches are

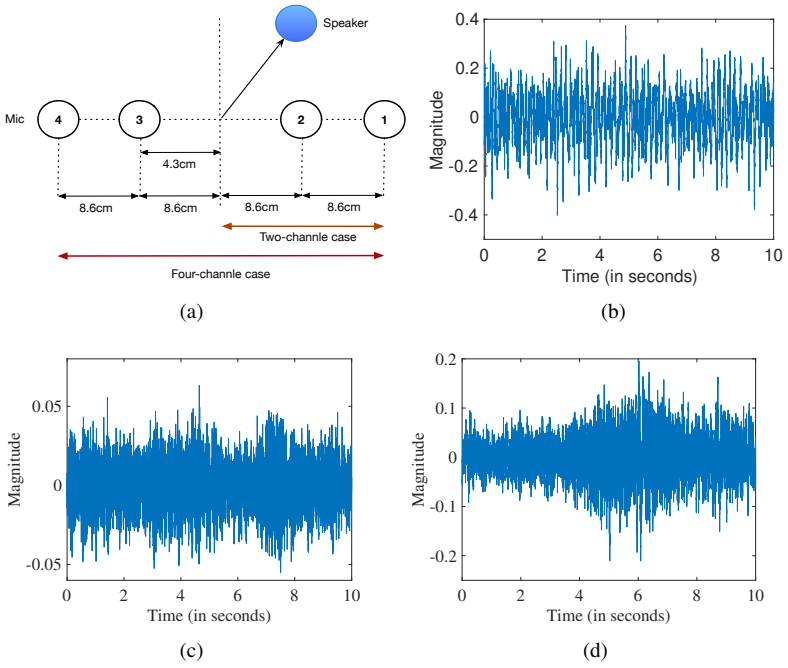


Figure 7: Illustration of Database B. (a) Setting of microphone arrays, (b) Subway car noises, (c) Cafeteria noises, (d) Square noises.

completely selected from either 8 male speakers or 8 female speakers. All combination of speeches, i.e., $2C_8^N$, are tested for each case.

- **Database B** is a data collection recorded by a group of speeches involving various background noises [41]. black In this database, the clean speeches are selected from 3 male speakers and 3 female speakers. In addition, various real noises, e.g., subway car noise, cafeteria noise and square noise, are provided. The speeches and noises are separately recorded via omni-directional microphones spaced by 8.6cm. The RIRs with RT_{60} is set as 200ms. The speeches and noises are recorded with a sampling rate $F_s = 16$ kHz, and the duration of recorded mixtures is truncated to 10 seconds. The mixtures include two separated parts, i.e., one is the convolved speeches and the other is the recorded real noises. These two parts are added in a linear superposition as described in (1). Two cases with various speakers and microphones are tested in these experiments, i.e., $Case(M, N) = (2, 3)$ and $Case(M, N) = (4, 3)$. As illustrated in FIGURE 7 (a), the microphone arrays are located in a line with identical length of 8.6cm. In $Case(M, N) = (2, 3)$, microphones of No. 1 and 2 are employed to record the mixtures. In $Case(M, N) = (4, 3)$, all microphones are employed to record the mixtures. Three type of background noises of subway car, cafeteria and square are illustrated in FIGURE 7 (b)-(d), respectively. In order to make the database large enough, we have 40 mixture files by combining various locations, sources, microphones and speech samples. Particularly, in the scenarios of cafeteria and square, we have 2 locations $\times$ 2 type of speeches $\times$ 2 type of microphones $\times$ 2 samples = 16 mixtures; in the scenario of subway, we have 1 location $\times$ 2 speeches $\times$ 2 microphones $\times$ 2 samples=8 mixtures.

black Overall, database A and B provide the NoS estimation experiments under noise-free condition and noise condition. In addition, the Signal-to-Noise Ratio (SNR) is introduced to evaluate the level of noises as follows,

$$\mathrm{SNR} = 10\log_{10}\frac{\|\mathbf{H} \star \mathbf{s}(t)\|_F^2}{\|\mathbf{e}(t)\|_F^2}(dB). \tag{20}$$

black Second, the implementation settings of the proposed algorithm are given as follows. The window function is selected as Hanning window, the length of STFT is fixed at 2048 and the STFT frame shift is set as 128

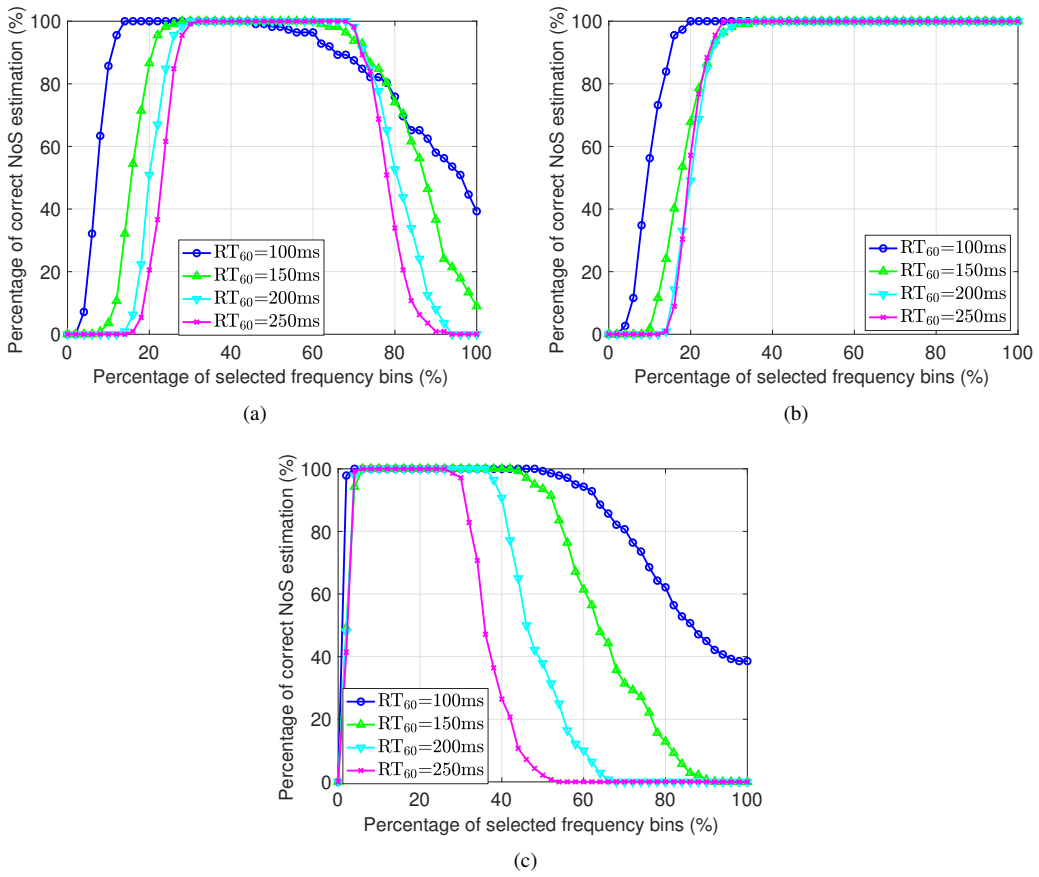


Figure 8: Precision of NoS estimation comparing to the percentage of selected frequency bins. (a) $Case(M,N) = (2,3)$, (b) $Case(M,N) = (3,3)$, (c) $Case(M,N) = (3,4)$.

samples (8ms). Suppose the duration time of recorded mixtures is 10 seconds, then the number of frames D is calculated as 1234. The number of frames P at each sub-block is set as 9, then the number of local covariance matrices Q is calculated as 137. In this case, the time duration of each sub-block is 73 ms, which is at the range of 40ms-80ms of wide-sense quasi-stationary assumption [31, 32]. All the experiments are carried out by a MacBook Air laptop equipped with Intel Core i5, CPU 1.8 GHz and macOS 10.13.6 system, and the programs are coded by Matlab R2018b. The code of proposed algorithm can be found in the following website: https://staff.scem.uws.edu.au/~yiguo/code/sourcenumber.zip.

Third, two state-of-the-art methods are employed as a fair comparison of proposed algorithm, i.e., DEMIX and Simplex analysis based method [16,28]. The proposed algorithm and DEMIX can be categorized as clustering based method while the simplex analysis based method is a rank detection method from a constructed correlation matrix. The set up of these two baseline methods are similar to the proposed algorithm. It is worh noting that, in the Simplex based analysis method, the threshold of selecting the rank of constructed correlation matrix is set empirically as recommended in [28].

Table 1: Running time (Sec.) of provided algorithms under Database A

$Case\ (M,N)$	$(2,3)$	$(3,3)$	$(3,4)$
DEMIX	24.38	27.93	35.63
Simplex based method	42.4	88.81	129.59
proposed algorithm	52.77	62.5	63.6

Table 2: Percentage (%) of correct estimation of NoS under Database A

$Case\ (M,N)$	$(2,3)$				$(3,3)$				$(3,4)$			
RT_{60}	100ms	150ms	200ms	250ms	100ms	150ms	200ms	250ms	100ms	150ms	200ms	250ms
DEMIX	58.82	50.0	47.32	34.83	76.79	40.0	37.5	33.04	50.71	50.0	42.86	29.63
Simplex based method	91.07	79.46	67.86	53.57	87.5	80.36	70.54	52.68	53.57	47.86	45.0	35.71
proposed algorithm	**99.77**	**98.71**	**93.52**	**87.38**	**100**	**94.97**	**93.13**	**94.86**	**99.95**	**99.06**	**84.95**	**56.51**

4.2 NoS estimation Results under noise-free condition

In this experiment, various NoS estimation tasks are tackled based on database A. First, the final NoS decision performance is tested by varying the percentage of selected frequency bins from 0% to 100%. FIGURE 8. (a)-(c) illustrate the percentage of correct NoS estimation along with the percentage of selected frequency bins in $Case(M,N) = (2,3)$, $(3,3)$ and $(3,4)$, respectively. It can be observed that the proposed algorithm achieves the best NoS estimation accuracy when retaining top 20% to 50% frequency bins with the largest V_f's, and this range is our recommendation for the final NoS decision. The reason is that the speech components concentrate only on some frequencie bins where the local dominance assumption holds and cluster centers are far apart, while in other frequency bins, the local dominance assumption breaks down and the speech components are entangled with each other, generating many fake cluster centers. In addition, FIGURE 8 (a) and (b) show that increasing the number of microphones can substantially improve the performance of NoS estimation when N is fixed.

Table 3: Percentage (%) of correct estimation of NoS under Database B

$Case\ (M,N)$	$(2,3)$			$(4,3)$		
Noise Environment	Subway car	Cafeteria	Square	Subway car	Cafeteria	Square
SNR (dB)	6.01	29.13	3.18	10.87	33.99	8.05
DEMIX	33.33	41.67	58.33	50.0	50.0	58.33
Simplex based method	33.33	58.33	58.33	33.33	83.33	41.67
proposed algorithm	**66.67**	**83.33**	**66.67**	**100**	**100**	**100**

The running time of the proposed algorithm, DEMIX and Simplex based method are listed in Table 1. It can be seen that the complexity of these algorithms are not very high from the perspective of running time. The results of averaged accuracy of NoS estimation are listed in Table 2 with various RT_{60} for $Case(M,N) = (2,3)$, $(3,3)$ and $(3,4)$, respectively. It is quite clear that the accuracy of NoS estimation of all algorithms decreases when RT_{60} increases in all cases. DEMIX does not work well in all cases, especially when the NoS is greater than the number of microphones. Simplex analysis based method achieves better results in all test cases comparing to DEMIX. The NoS estimation performance of the proposed algorithm is consistently better than that of other methods, especially when RT_{60} is high. Table 2 shows the robustness of the proposed algorithm in a high reverberant environment.

4.3 NoS estimation Results under various noise condition

In this experiment, a more challenge task of NoS estimation with consideration of various noises are tackled based on database B. black As shown in the third row of Table 3, the SNR level of subway car noise and square noise are similar while cafeteria noise is relatively lower. In this case, the top 50% frequency bins with the largest V_f's are retained to strengthen the performance of NoS estimation. The results of averaged accuracy of NoS estimation are listed in Table 3 with various noise condition. It is obvious that the NoS estimation accuracy of all algorithms deteriorate slightly comparing to the noise-free condition in the experiment A. Moreover, increasing the number of microphones can substantially improve the NoS estimation performance comparing to $Case(M,N) = (2,3)$ and $(4,3)$. DEMIX and Simplex analysis based method achieve better results in the cases of cafeteria noise and square noise while the NoS performance are not well under the case of Subway car noise. On the contrary, the NoS estimation performance of the proposed algorithm is less sensitive to the interference of noises, especially when the number of microphones increases to 4. It is worth noting that the proposed algorithm correctly identifies the NoS at each test, which shows the robustness of the proposed algorithm in a noisy environment.

5 Conclusion

A new NoS detector in reverberant environment has been proposed in this paper. Based on the local dominance assumption, the NoS estimation is transformed into a density-based clustering problem by exploiting the leading eigenvectors from the local covariance matrices of mixtures in TF domain. A frequency bins selection procedure is also proposed to improve the final NoS estimation so that the most reliable NoS estimation results are retained from the frequencie bins where the local dominance assumption holds. The experiment results demonstrate the superiority of the proposed algorithm to the state-of-the-art methods in various cases. black In the future, we will extend the study of NoS estimation in TF domain from a linear narrowband system to a convolutive narrowband system, which is more suitable to depict a highly reverberant scenario [42].

References

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2003.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[3] S. Makino, T. W. Lee, and H. Sawada, *Blind Speech Separation*. Berlin: Springer-Verlag, 2007.

[4] Z. Yang, G. Zhou, S. Xie, S. Ding, J. Yang, and J. Zhang, "Blind spectral unmixing based on sparse nonnegative matrix factorization," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1112–1125, 2011.

[5] G. Zhou, Q. Zhao, Y. Zhang, T. Adal, S. Xie, and A. Cichocki, "Linked component analysis from matrices to high-order tensors: Applications to biomedical data," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 310–331, Feb. 2016.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, *A Survey of Convolutive Blind Source Separation Methods*. New York:Springer, 2007.

[7] L.Rabiner and R.Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[8] K. Han and A. Nehorai, "Improved source number detection and direction estimation with nested arrays and ulas using jackknifing," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 6118–6128, Dec. 2013.

[9] L. Xu, A. Krzyzak, and E. Oja, "Rival penalized competitive learning for clustering analysis, RBF net, and curve detection," *IEEE Trans. Neural Networks*, vol. 4, no. 4, pp. 636–649, 1993.

[10] S. Tu and L. Xu, "Learning binary factor analysis with automatic model selection," *Neurocomputing*, vol. 134, pp. 149–158, 2014.

[11] C.-D. Wang and J.-H. Lai, "Energy based competitive learning," *Neurocomputing*, vol. 74, no. 12-13, pp. 2265–2275, 2011.

[12] C.-D. Wang, J.-H. Lai, C. Y. Suen, and J.-Y. Zhu, "Multi-exemplar affinity propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2223–2237, 2013.

[13] C.-D. Wang, J.-H. Lai, and P. S. Yu, "Multi-view clustering based on belief propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1007–1021, 2016.

[14] H. Sawada, S. Winter, R. Mukai, S. Araki, and S. Makino, "Estimating the number of sources for frequency-domain blind source separation," in *Inter. Conf. Ind. Comp. Analy. and Signal Sep.*, 2004, pp. 610–617.

[15] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 481–484, Jul. 2007.

[16] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121–133, Jan. 2010.

[17] Y. Chen, W. Wang, Z. Wang, and B. Xia, "A source counting method using acoustic vector sensor based on sparse modeling of DOA histogram," *IEEE Trans. Signal Process. Lett.*, vol. 26, no. 1, pp. 69–73, Jan. 2019.

[18] C.Wang, H.Zhang, L.Yang, S.Liu, and X.Cao, "Deep people counting in extremely dense crowds," in *Proc. ACM Int. Conf. Multimedia*, 2015, p. 12991302.

[19] S. G. A. Khan and M. Salzmann, "Deep convolutional neural networks for human embryonic cell counting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, p. 339348.

[20] X. W. C. Zhang, H. Li and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. Int. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, p. 833841.

[21] S. S. K. L. Boominathan and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proc. ACM Int. Conf. Multimedia*, 2016, p. 640644.

[22] F. Stöter, S. Chakrabarty, B. Edler, and E. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Trans. on Audio, Speech and Lang. Proc.*, vol. 27, no. 2, pp. 268–282, Feb. 2019.

[23] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. on Audio, Speech and Lang. Proc.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.

[24] X. Zhong and J. R. Hopgood, "A time-frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2356–2370, Dec. 2015.

[25] V. Reju, S. N. Koh, and I. Soon, "Underdetermined convolutive blind source separation via time-frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 101–116, Jan. 2010.

[26] T. Chan, W. Ma, C. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5120–5134, Oct. 2008.

[27] X. Fu, W. K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May. 2015.

[28] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6458–6473, Dec. 2018.

[29] J. Yang, Z. Yang, Y. Guo, and S. Xie, "Blind source separation: Detecting unknown sources number in covariance domain," in *9th Inter. Conf. Comp. and Auto. Engin. (ICCAE)*, 2017.

[30] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time-frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 101–116, Jan. 2010.

[31] L.Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[32] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[33] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1193–1207, Aug. 2010.

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

[35] A. Rodriguez and L. Alessandro, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[36] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics.*, vol. 33, no. 3, p. 1065, 1962.

[37] Z. He, A. Cichocki, S. Xie, and K. Choi, "Detecting the number of clusters in n-way probabilistic clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2006–2021, 2010.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambriadge Press, 2004.

[39] OpenSLR,http://cn-mirror.openslr.org/resources/12/dev-clean.tar.gz, [Online].

[40] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, Apr. 1979.

[41] http://sisec2010.wiki.irisa.fr/tiki-indexa8a1.html?page=Source+separation
+in+the+presence+of+real-world+background+noise, [Online].

[42] F. Feng and M. Kowalski, "Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrow-band approximation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 442–456, Feb. 2019.