AUGMENTED BAYESIAN NONPARAMETRIC CLUSTERING FOR SOURCE COUNTING WITH A SMALL APERTURE MICROPHONE ARRAY

Kunkun SongGong[1], *Pufen Zhang*[1*], *Xiongwei Zhang*[1], *Wenwu Wang*[2], *Meng Sun*[1], *Chong Jia*[1], *and Yihao Li*[1]

[1]Army Engineering University, China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, U.K.
Email: [sgkk@nuaa.edu.cn, pufenzh@163.com, xwzhang9898@163.com, w.wang@surrey.ac.uk, sunmeng@aeu.edu.cn, jiachong@aeu.edu.cn, liyihao@aeu.edu.cn]

ABSTRACT

Source counting (SC) in an indoor environment is an important problem in computational auditory scene analysis. However, the problem is challenging, especially when reverberation and ambient noise are present in the environment. To address this problem, we propose an augmented Bayesian nonparametric (ABNP) clustering algorithm for source counting based on sound intensity (SI) captured by a small aperture microphone array. The core idea is to incorporate an infinite Gaussian mixture model (IGMM) and a time-frequency (TF) augmented weight selection and update scheme for sound intensity estimation. The use of IGMM enables the exemption of the maximum number of sources assumed in previous methods. Experiments on both simulated and real-world data show the improved performance by the proposed method as compared with the state of the art baseline methods.

Index Terms— Source counting (SC), Bayesian nonparametric (BNP), sound intensity (SI), microphone array

1. INTRODUCTION

Acoustic signal processing problems, such as multi-source localization [1], blind source separation [2], speaker recognition [3], and sound event detection [4], have important applications in many practical systems. A pivotal task in these domains is the estimation of the number of active sources, known as source counting (SC), which is often required by these systems in their operation [5, 6]. To address the SC problem, a variety of algorithms have been developed [5–12]. An early approach was based on the information theoretic criteria [8], such as the minimum description length (MDL). However, this method requires the assumption of spatially and temporally white noise, which might not be valid in practical situations due to the changing adverse environments and the non-ideal array configurations, which may lead to inaccurate estimation of the number of active sources [9, 10].

To overcome this limitation, several methods have been proposed by exploiting the directions of arrival (DOAs) of the sources. For example, Pavlidi *et al.* [11] developed three SC algorithms based on histogram of the DOA estimations, namely, a peak search (PS) approach, a linear predictive coding (LPC) approach, and a matching pursuit (MP) approach. Araki *et al.* [12] introduced a method to model the distribution of DOAs with Gaussian mixture models (GMMs), with their parameters learned by expectation maximisation (EM). With this method, the total number of sources can be estimated via counting the number of GMM components. A sound intensity (SI) based method was proposed in [13], which exploits first order differential microphone arrays (DMAs) for their compact size and high gain. This offers the potential to meet the increasing demand for portable and lightweight device in real world. However, it is prone to degradation caused by room reverberation and ambient noise. In addition, a common limitation of the aforementioned methods is that the maximum number of sources cannot exceed a value predefined by users.

To address the above limitations, we present a new augmented Bayesian non-parametric (ABNP) clustering algorithm for source counting based on sound intensity estimation via a small aperture microphone array. The proposed approach consists of two steps. First, it leverages the theory related to SI and the sparsity property of speech signals which enables the use of a small-sized array to estimate the DOAs at each time-frequency (TF) bin. This array is composed by two orthogonal first-order DMAs, offering potentials to satisfy the practical requirement for array miniaturization.

Second, we design an ABNP algorithm without the prior knowledge about the maximum number of sources, taking the estimated SI from the first step as input. This design of this step carries two novel aspects. On the one hand, the BNP clustering method, which assumes an infinite number of sources, is employed to estimate the number of sources using the DOAs derived from the array measurements. On the other hand, to mitigate the adverse impact of noise and reverberation, we design a TF augmented weight selection and update framework based on the infinite Gaussian mixture model (IGMM), thereby improving its performance in adverse environments.

In this way, sources can be successfully counted via compact arrays under diverse acoustic conditions. Experiments on public-domain artificial and real datasets show the superior performance of the proposed method as compared with baseline methods (e.g. BNP, MP, LPC, PS and EM).

2. SIGNAL MODEL

A small-sized array consisting of four omnidirectional sensors ($M_m, m = 1, \ldots, 4$) is adopted for sound source capture as shown in Fig. 1. This aperture array can be decomposed into two orthogonal first-order DMAs, namely M_1 and M_3 along the x-axis, and M_2 and M_4 along the y-axis. Both of the sizes of the two DMAs are the same and denoted by D.

Suppose that I far-field speech sources in a reverberant enclosure impinge on the array. Herein, the DOAs are defined with respect to the positive x-axis, which implies $\phi_i \in [-\pi, \pi), i = 1, \ldots, I$. Using the short-time Fourier transform (STFT), the source signals received at the mth sensor can be modeled as

This work was supported by the National Natural Science Foundation of China (62401624, 62071484, 62371469) and the Postdoctoral Fellowship Program of CPSF (2024M754248).

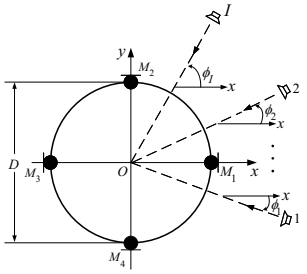


Fig. 1. Configuration of the sensor array constructed by two orthogonal first-order DMAs.

$$P_m(t,f) = \sum_{i=1}^{I} H_{mi}(t,f)S_i(t,f) + V_m(t,f), \quad (1)$$

where t is the time frame index and f is the frequency. $S_i(t,f)$ is the signal induced by one of the I speech sources at ϕ_i direction from the centre O of the sensor array, $H_{mi}(t,f)$ is the room impulse responses (RIRs) from the ith source to the mth sensor, and $V_m(t,f)$ is the additive background noise. Since speech signals are considered sparse in the time-frequency (TF) domain, at each TF bin, it could be assumed that only one source is dominant [1]. Thus, according to the output signal of DMAs and the theory on sound intensity (SI) [13], we can estimate the $\hat{\phi}(t,f)$ at TF bin dominated by each source as the observations of DOAs. The $\hat{\phi}(t,f)$ can be represented as

$$\begin{aligned}\hat{\phi}(t,f)&=\arctan\left\{\frac{\mathrm{Re}\left[I_{oy}(t,f)\right]}{\mathrm{Re}\left[I_{ox}(t,f)\right]}\right\}\\ &=\arctan\left\{\frac{\mathrm{Im}\left\{P_0(t,f)\left[P_4(t,f)-P_2(t,f)\right]^*\right\}}{\mathrm{Im}\left\{P_0(t,f)\left[P_3(t,f)-P_1(t,f)\right]^*\right\}}\right\},\end{aligned} \quad (2)$$

where $I_{ox}(t,f)$ and $I_{oy}(t,f)$ represent $x-$ and $y-$ components of the complex SI, respectively. $\mathrm{Re}(\cdot)$ denotes the real part of operation, the superscript $(\cdot)^*$ denotes the complex conjugate, $\mathrm{Im}(\cdot)$ denotes the imaginary part of operation. $P_0(t,f)$ is the the sound pressure at the coordinate origin, which can be estimated via the average of the sound pressures at all sensors [13],

$$P_0(t,f)=\frac{1}{4}[P_1(t,f)+P_2(t,f)+P_3(t,f)+P_4(t,f)]. \quad (3)$$

In practice, I is often unknown and needs to be estimated from the array measurements.

3. PROPOSED METHOD

To estimate I, we develop an augmented BNP [14] clustering algorithm based on the IGMM [15] for SC using DMAs.

3.1. IGMM Model

According to (2), we use the estimated DOAs, namely $\{\hat{\phi}(t,f)\} = \{\hat{\phi}_1, \cdots, \hat{\phi}_n, \cdots, \hat{\phi}_N\}$, as the observation inputs to the IGMM model. Thus, the probability density function (PDF) of $\hat{\phi}_n$, generated by the component l ($l \in (1, \cdots, \infty)$ with l being unknown) at the nth bin is given by

$$p(\hat{\phi}_n|\mu_l, \sigma_l^2, k_n) = \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{\frac{-(\hat{\phi}_n + 2\pi k_n - \mu_l)^2}{2\sigma_l^2}}, \quad (4)$$

where the parameter $\{\mu_l, \sigma_l^2\}$ obeys Gaussian Gamma distribution, which represents as

$$\{\mu_l, \sigma_l^2|\boldsymbol{\Theta}_l\} \sim \boldsymbol{N}(\mu_l|\chi_l, \sigma_l^2/\xi_l)\boldsymbol{Ga}(\sigma_l^{-2}|\eta_l, \gamma_l), \quad (5)$$

where $\boldsymbol{\Theta_l} = \{\chi_l, \xi_l, \eta_l, \gamma_l\}$ represents the hyperparameters of the IGMM, $\boldsymbol{N}$ and $\boldsymbol{Ga}$ represents the Gaussian and Gama distribution, respectively, and k_n is an integer parameter accounting for the shift of $\hat{\phi}_n$.

Note that, here, we use Chinese Restaurant Process (CRP) [15] as the prior information to restrict the IGMM. Then, the probability of the class label z_n of the nth observation belonging to an existing class or a new class can be expressed as,

$$p(z_n = l|\boldsymbol{z}_{\backslash n}) = \begin{cases} \dfrac{n_l}{n+\alpha-1}, & l = 1, \cdots, L \\ \dfrac{\alpha}{i+\alpha-1}, & l = l_{new} = L+1 \end{cases} \quad (6)$$

where n_l is the number of DOA estimates assigned to the nth component, and $\boldsymbol{z}_{\backslash n}$ is set of class labels without the nth DOA estimate, and the parameter α is the concentration parameter of the Dirichlet process [16].

Then, the likelihood function is obtained by the marginalized integral of the product of (4) and (5),

$$\begin{aligned}&p(\hat{\phi}_n|\hat{\boldsymbol{\phi}}_{\backslash i}, z_n = l, \boldsymbol{k}_{\backslash n}, \boldsymbol{z}_{\backslash n}, \boldsymbol{\Theta}_l^a)\\ &\quad\propto \sum_{k_n=-K}^{K} T_{2\eta_l^a}\left[(\hat{\phi}_n + 2\pi k_n)\middle|\chi_l^a, \frac{\gamma_l^a(\xi_l^a+1)}{\eta_l^a\xi_l^a}\right],\end{aligned} \quad (7)$$

where $\hat{\boldsymbol{\phi}}_{\backslash i}$ denotes the set of all the DOA measurements without θ_i, $\boldsymbol{k}_{\backslash n}$ is the set of all the shifts without k_n, $\boldsymbol{z}_{\backslash n}$ denotes the set of class labels without the nth label z_n, and $\boldsymbol{\Theta}_l^a = \{\xi_l^a, \eta_l^a, \chi_l^a, \gamma_l^a\}$ is the set of hyperparameters for component l at the current iteration a.

Next, we use Gibbs sampling [16] to approximate the posterior probability of $z_n = l$, i.e. the class label z_n belonging to the mixture component l, given all $\hat{\phi}_i$ and $\boldsymbol{\Theta}_l^a$, which can be calculated as follows

$$\begin{aligned}&p(z_n = l|\hat{\phi}_n, \hat{\boldsymbol{\phi}}_{\backslash n}, \boldsymbol{k}_{\backslash n}, \boldsymbol{z}_{\backslash n}, \boldsymbol{\Theta}_l^a)\\ &\quad\propto p(z_n = l|\boldsymbol{z}_{\backslash n})p(\hat{\phi}_n|\hat{\boldsymbol{\phi}}_{\backslash n}, z_n = l, \boldsymbol{k}_{\backslash n}, \boldsymbol{z}_{\backslash n}, \boldsymbol{\Theta}_l^a).\end{aligned} \quad (8)$$

3.2. Augmented BNP Clustering

The IGMM model can adaptively adjust the number of classes and parameters required for model construction based on the observation values (namely, the estimated $\hat{\phi}$). To be specific, the first observation is established as the first class, and subsequent observations are either established as new classes or are assigned to existing classes. As the class becomes larger, subsequent observations are more likely to be assigned to that class. Clearly, the previous observations are more likely to become the main components of the mixture model, while subsequent observations can be assigned based on the model established by previous observations. The sorting of the observation sequence can have a certain impact on the clustering results based on this model.

However, the sorting of observation values is usually random, and not every observation is valid. The reason is that the estimated $\hat{\phi}$ is prone to the corruptions by the ambient noise and reverberation, especially using the DMAs. This means that if there are invalid or erroneous observations at

the beginning, the established mixture model is likely to deviate from the overall distribution of the observations, which may degrade the performance of source counting. To tackle this issue, we propose a novel augmented BNP clustering approach, which can be divided into two steps.

Step 1: In this step, we design a TF augmented weight selection scheme for obtaining reliable TF bins for source counting, based on the property that speech signals are often sparse in the TF domain. A higher value indicates a stronger reliability of the TF bins, while a lower value is less reliable and could be from noise.

With the instantaneous DOA estimates ($\hat{\phi}$) and the instantaneous power of signals received by the sensors, the augmented weights $AW(t,f)$ can be calculated as

$$AW(t,f) = 3\left[Pow(t,f)\cdot Pr(t,f)\cdot Var(t,f)\right]^2, \quad (9)$$

where $Pow(t,f)$ denotes the weight of power at the TF bin, $Pr(t,f)$ denotes the weight of power ratio [17], and $Var(t,f)$ denotes the weight of local DOAs variance [18], determined via the sigmoid compression, respectively

$$Pow(t,f) = 1/\left[1+e^{-\alpha_1[\log E(t,f)-\beta_1]}\right], \quad (10)$$

$$Pr(t,f) = 1/\left[1+e^{-\alpha_2[\log E(t,f)/E(t,f-1)-\beta_2]}\right], \quad (11)$$

$$Var(t,f) = 1/\left[1+e^{-\alpha_3[\log \sigma^2_{\phi}(t,f)-\beta_3]}\right]. \quad (12)$$

where $E(t,f)=|P_0(t,f)|^2$, $\sigma^2_{\phi}(t,f)$ is the local variance of $\phi(t,f)$, and the $\alpha_1,\alpha_2,\alpha_3$ and the β_1,β_2,β_3 are the sigmoid slope and center parameters, respectively [19].

Step 2: To obtain the new sequence for SC, we select and reorder the reliable observations with the largest augmented weights, as described below.

First, we choose Q observations $\hat{\phi}^{'}_q(q=1,\cdots,Q)$ with the largest $AW(t,f)$. Second, we perform peak search of the histogram constituted by all the $\hat{\phi}^{'}_q$ and obtain the pre-estimation $\hat{\phi}_p$. Third, we calculate the difference between the $\hat{\phi}^{'}_q$ and $\hat{\phi}_p$ and obtain the minimum difference $diff^q_{mim}$. Fourth, we rearrange the observations $\hat{\phi}^{'}_q$ in ascending order of $diff^q_{mim}$ and get the new sequence $\hat{\phi}^{''}_q$.

In the following, with the new $\hat{\phi}^{''}_q$, we can obtain the posterior probability by (7) and estimate the class label of the current observation by the maximum posterior probability. Meanwhile, in order to increase the dominant role of reliable TF bins in the hyper-parameters $\boldsymbol{\Theta}^{\boldsymbol{a}}_{\boldsymbol{l}}$ updating process (a is the iteration number with initial value $a=0$), the $AW(t,f)$ is brought into and the new update formulas are expressed as,

$$\xi_l^{(a+1)} = \xi_l^{(a)} + AW_q, \quad (13)$$

$$\eta_l^{(a+1)} = \eta_l^{(a)} + \frac{1}{2}AW_q, \quad (14)$$

$$\chi_l^{(a+1)} = \frac{1}{\xi_l^{(a+1)}}\left[\xi_l^{(a)}\chi_l^{(a)} + AW_q(\hat{\phi}^{''}_q + 2\pi k_q)\right], \quad (15)$$

$$\gamma_l^{(a+1)} = \gamma_l^{(a)} + \frac{1}{2}\Big[AW_q(\hat{\phi}^{''}_q + 2\pi k_q)^2 + \xi_l^{(a)}(\chi_l^{(a)})^2 - \xi_l^{(a+1)}(\chi_l^{(a+1)})^2\Big], \quad (16)$$

In addition, AW_l^{sum} denotes the sum of the augmented weights for the lth class of the DOA estimation. The number of AW_l^{sum} that is greater than the threshold th_{AW} is taken as the estimated number of speakers, namely, I. The threshold th_{AW} can be calculated as

$$th_{AW} = 0.5[mean(AW_l^{sum}) + \sqrt{var(AW_l^{sum})}], \quad (17)$$

where $mean(\cdot)$ and $var(\cdot)$ represent taking the mean and variance over its argument, respectively.

By using this scheme, reliable TF bins can be selected and used for clustering and estimating the number of sources. The specific analysis of this framework is given in Section 4.2. The proposed ABNP-DMAs is summarized in Algorithm 1.

Algorithm 1 ABNP-DMAs

Input: $\boldsymbol{\Theta}^{\boldsymbol{0}}_{\boldsymbol{l}}$, $\alpha_1,\alpha_2,\alpha_3,\beta_1,\beta_2,\beta_3$.

```
for P_m(t,f) do using (2) and (3) to estimate the φ̂(t,f);
    for φ̂(t,f) do using (9), (10), (11) and (12) to
        calculate the AW(t,f);
        choosing φ̂'_q with the largest AW(t,f);
        calculating the diff^q_mim and rearranging the φ̂'_q
        in ascending order get new sequence φ̂''_q;
        for φ̂''_q do using the IGMM and calculating the
            posterior probability by (7) and (8);
            update the hyper-parameters Θ^a_l with
            AW(t,f) by (13), (14), (15) and (16);
            Input the updated hyper-parameters Θ^a_l to
            the IGMM. Cluster and obtain the estimated
            number of speakers I.
        end for
    end for
end for
```

4. EXPERIMENTAL EVALUATIONS

4.1. Datasets and Set up

The performance of the proposed ABNP-DMAs is evaluated and compared with the traditional BNP method [1, 14] and several baselines including the MP [11, 20], LPC [11, 21], PS [11] and EM [12, 22] methods in both simulated and real room environments. Note that, we consider 1 to 4 sound sources, and the maximum and minimum interval between speakers is 25° and 120°, respectively.

The dimension of the simulated rectangular room is 6 m × 6 m × 4 m. To generate the RIRs [23] from speaker sources to sensors, we use a software that is based on the well-known image method for simulating a reverberant room [24]. The DMAs with M = 4 equidistant omnidirectional sensors and the radius of r = 0.02 m are placed in the center of the room at $(3,3,1)$ m, coinciding with the origin of the x and y axes. The speakers are located at the same height as the microphone array with distance from the speaker to the center of the array being 2 m. The additive noises on the sensors are mutually uncorrelated white Gaussian, and also are uncorrelated with the speech signal. The sound speed is 340 m/s. Speech signals of 1 s length, sampled at 16 kHz, are chosen randomly from the well-known TIMIT speech database [25]. For all the evaluated algorithms, the STFT is calculated using a Hamming window of 1024 samples with 50% overlap between consecutive frames.

The dimensions of the real rectangular conference room is approximately 9.7 m × 7.05 m × 3 m with a reverberation time of 350 ms. A DMAs was placed horizontally around the

center of the room, and the other conditions resembled those in the above simulations.

The corresponding parameters of the proposed method are set empirically to $\xi_l^{(0)} = 0.01$, $\eta_l^{(0)} = 0.01$, $\chi_0 = \frac{\sum_{q=1}^{Q} AW_q \hat{\phi}_q''}{\sum_{q=1}^{Q} AW_q}$, and $\gamma_0 = \frac{\sum_{q=1}^{Q} AW_q (\hat{\phi}_q'' - \chi_l^{(0)})^2}{\sum_{q=1}^{Q} AW_q}$. We set Q to be equal to 20% of the total number of TF bins. $\alpha_1 = -1$, $\alpha_2 = -6$, $\alpha_3 = 3$, β_1 is the $\log E(t,f)$ value of 640 TF bins in descending order, $\beta_2 = -0.5$, β_3 is the $\log \sigma_\phi^2(t,f)$ value of 320 TF bins in descending order.

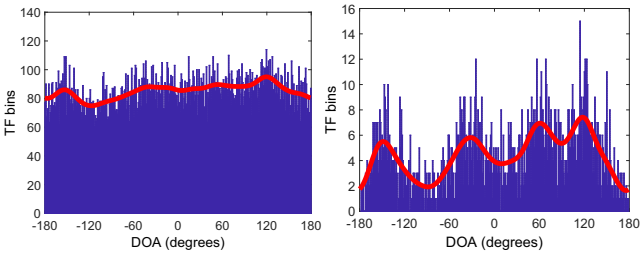
To facilitate evaluations, we use source counting success rate (SR) as performance metrics, which is defined as:

$$SR = \hat{C}_s / C_s \times 100\%, \tag{18}$$

where $\hat{C}_s$ is the number of experiments with counting success and C_s is the number of simulated or real-world experiments.

4.2. Results in Simulated Experiments

Analysis of TF Augmented Scheme: Fig. 2 shows the histogram comparison of TF bins selection with $RT_{60} = 0.4\ s$, SNR=15 dB. Herein, we use four sources located at $[-150°, -30°, 60°, 125°]$ as an example. Comparing Fig. 2(a) and Fig. 2 (b), it is clear that the latter can provide more reliable TF bins for the following algorithm.



(a) Before (b) After

Fig. 2. The histogram before and after the TF selection.

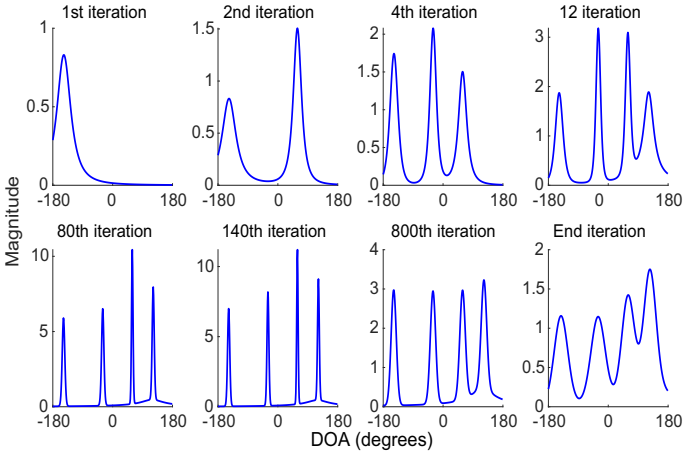
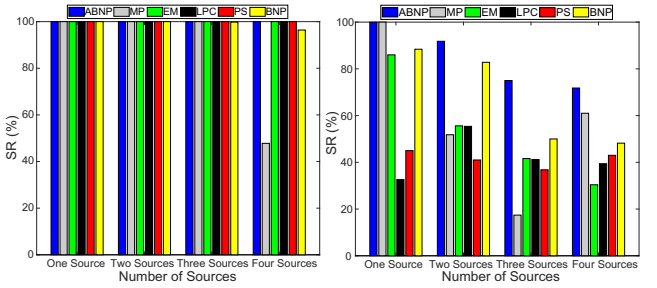


Fig. 3. The IGMM after various iterations.

Fig. 3 shows the diagram of IGMM obtained after different iterations by the new sequence $\hat{\phi}_q''$. From the results, we can see that in the 1st iteration, the first class is established. Then, the second, third and fourth classes are established in the 2nd, 4th and 12th iterations, respectively. An interesting phenomenon can also be noticed in the figure, that is the magnitude gradually increases until the 140th iteration and after that it decreases until the end of iterations. The reason is that the observed data in the cluster are close to each other and are more concentrated, so the variance of the components is smaller, and the posterior probability is larger. When the observations increase, the distance between the observations within the classes increases, and the posterior probability decreases.

Effect of Noise and Reverberation: Fig. 4 shows the performance of each method for different number of sources under noise and reverberation environments. The proposed ABNP method offers better source counting success rate as compared with BNP, MP, EM, LPC and PS methods, especially in adverse environments. This is because the proposed approach can select the reliable TF bins which are less affected by noise and reverberation, and improve the source counting success rate with the help of augmented BNP processing. Note that, except for the traditional BNP and our proposed method, other the baseline methods need the maximum number of speakers as prior knowledge, which manifests the superiority of the proposed method.



(a)$\mathrm{RT}_{60}=0.15$s, SNR$=25$dB (b)$\mathrm{RT}_{60}=0.55$s, SNR$=10$dB

Fig. 4. The performance for different number of sources.

4.3. Results in Real-World Experiments

Fig. 5 shows the source counting performance achieved on the real dataset. As can be seen, the results behave in a similar manner to those found in the simulation results above. The results show good performance of our proposed ABNP method, which indicates the effectiveness of using TF augmented weight selection and update framework based on the BNP clustering approach in a practical environment.

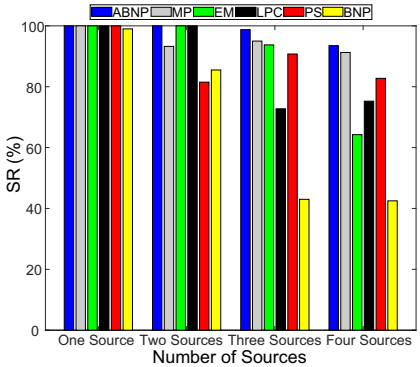


Fig. 5. The results for real-world data.

5. CONCLUSION

We have presented an ABNP clustering algorithm for source counting via a small aperture array, where we use a TF augmented weight selection and update scheme based on IGMM for the source counting problem. Experiments in both simulated and real environments demonstrated the effectiveness of the proposed method compared with baseline methods.

6. REFERENCES

[1] K. SongGong, H. Chen and W. Wang, "Indoor multi-speaker localization based on bayesian nonparametrics in the circular harmonic domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1864-1880, May, 2021.

[2] A. Alinaghi, P. J. Jackson, Q. Liu and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1434-1448, Sep. 2014.

[3] R. Tao, K. A. Lee, Z. Shi and H. Li, "Speaker recognition with two-step multi-modal deep cleansing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5.

[4] D. Krause, A. Politis and K. Kowalczyk, "Feature overview for joint modeling of sound event detection and localization using a microphone array" in *Proc. European Signal Process. Conf. (EUSIPCO)*, Amsterdam, Netherlands, 2021, pp. 31-35.

[5] L. Wang, T. -K. Hon, J. D. Reiss and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079-1093, June 2016.

[6] 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 Signal Process. Lett.*, vol. 26, no. 1, pp. 69-73, Jan. 2019.

[7] T. Sgouros and N. Mitianoudis, "A novel directional framework for source counting and source separation in instantaneous underdetermined audio mixtures,"*IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2025-2035, 2020.

[8] E. Fishler, H. V. Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Trans. Signal Process.*, vol. 53, no. 9, pp. 3543-3553, 2005.

[9] L. Huang, S. Wu, X. Li, "Reduced-rank MDL method for source enumeration in high-resolution array processing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5658-5667, 2007.

[10] L. Huang, T. Long, E. Mao and H. C. So, "MMSE-based MDL method for robust estimation of number of sources without eigendecomposition," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 4135-4142, 2009.

[11] D. Pavlidi, A. Griffin, M. Puigt and A. Mouchtaris, "Real-Time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193-2206, Oct. 2013.

[12] S. Araki, Tomohiro Nakatani, H. Sawada and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, 2009, pp. 33-36.

[13] Y. Li and H. Chen, "Reverberation robust feature extraction for sound source localization using a small-sized microphone array," *IEEE Sensors J.*, vol. 17, no. 19, pp. 6331-6339, Oct., 2017.

[14] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian nonparametrics for microphone array processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 493-504, Feb. 2014.

[15] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *J. Maths. Psy.*, vol. 56, no. 1, pp. 1-12, Feb. 2012.

[16] X. Yu, "Gibbs Sampling Methods for Dirichlet Process Mixture Model: Technical Details," pp. 1-18, Sep. 2014. [Online]. Available: http//www.mendeley.com/catalogue/0be87fb1-00a9-328b-b9b2-0a1b454e3f4f/

[17] L. Sun and Q. Cheng, "Indoor multiple sound source localization using a novel data selection scheme," *Proc. IEEE 48th Ann. Conf. Info. Sci. Sys.*, Princeton, NJ, USA, Mar. 2014, pp. 1-6.

[18] S. He and H. Chen, "Closed-form DOA estimation using first-order differential microphone arrays via joint temporal-spectral-spatial processing," *IEEE Sensors J.*, vol. 17, no. 4, pp. 1046-1060, Feb. 2017.

[19] M. Kuhne, R. Togneri, and S. Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering," *IEEE Signal, Process. Lett.*, vol. 16, no. 2, pp. 85-88, Feb. 2009.

[20] Z. Zhao, H. Kan, J. Lin and Z. Xu, "DOA estimation for multiple speech sources based on flexible single-source zones and concentration weighting," *IEEE Sensors J.*, vol. 23, no. 10, pp. 10683-10693, May, 2023.

[21] M. Kim and J. Skoglund, "Neural Speech and Audio Coding," *arXiv:2408.06954v1*, 2024.

[22] M. Jia, Y. Wu, C. Bao and C. Ritz, "Multi-source DOA estimation in reverberant environments by jointing detection and modeling of time-frequency points," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 379-392, Dec. 2021.

[23] E. A. P. Habets, "RIR Generator," 2016. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, Apr. 1979.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N.L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST Interagency/Internal Rep. 4930, Feb. 1993.