INDOOR MULTI-SOUND SOURCE LOCALIZATION BASED ON NONPARAMETRIC BAYESIAN CLUSTERING

Yao Guo[1], Hongyan Zhu[1] and Qi Cheng[2]

[1]School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China 710049
[2]School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK, USA 74078

ABSTRACT

This paper deals with sound source localization and number estimation in indoor environments using a circular microphone array. Multiple sound source localization is achieved by performing single source localization at each selected time-frequency (TF) point of received signals after short-time Fourier transform. A TF point selection method is proposed to reduce the computational time, which depends on a trained SVM with power and power ratio of TF points as its features. Nonparametric Bayesian clustering is applied on the obtained DOA estimates to identify the number of active sources. The algorithm is shown to outperform others through simulations.

Index Terms—DOA, short time Fourier Transform (STFT), SVM, source number estimation, DPMM.

1. INTRODUCTION

As an important research area in audio signal processing, sound source localization receives ample attention. It finds applications in teleconferencing [1], guiding a robot and in the next generation of hearing aids. As accurate localization algorithms emerge, indoor sound source localization may become an integral part in smart home applications.

Techniques such as Generalized Cross-Correlation Phase Transform (GCC-PHAT) [2] and Steered Response Power-Phase Transform (SRP-PHAT) [3], which use multiple microphone pairs, are relatively simple. They are designed for single source localization and the estimated results may be erroneous in the multi-source case. Another challenge for this type of methods is reverberation in indoor environments.

Subspace approaches, such as MUSIC and its wideband variations [4], are capable of estimating directions of arrivals (DOAs) of multiple sources. But it can only tackle the overdetermined case, i.e., the microphone number is more than the number of sound sources. Assuming sources are W-disjoint orthogonal, Jourjine *et al.* [5] proposed asolution for the underdetermined case, i.e., a blind separation of N sources from two mixtures.

Karbasi *et al.* proposed a uniform circular microphone array based sound source localization algorithm [6] which has an advantage over a linear array in overcoming the ambiguities. They assumed that the sources are sufficiently sparse, i.e., one source is dominant over others in certain time-frequency zones, and proposed the circular integrated cross spectrum (CICS) method to estimate DOA for a given frequency value.

In reverberant conditions, Pavlidi *et al.* imposed relaxed sparsity constraint on sound sources [1], i.e., in each time-frequency component, more than one source may be active. They set a threshold on the cross-correlation of a pair of microphone signals for detecting the single-source time-frequency (TF) zones. Peak locations of a DOA histogram indicate the DOAs of the multiple active sound sources. Based on the characteristics of speech signals, after selecting the TF points which represent the direct sound waves of individual sources, Sun *et al.* [7] improved the localization accuracy.

In practice, source localization is more difficult when the source number is unknown. Source number counting is needed. Pavlidi *et al.* used matching pursuit [1] and compared to peak search, minimum description length (MDL), linear predictive coding methods, but it has a relatively low detection rate. Algorithms in [8,9] using an infinite Gaussian mixture model can get a relative accurate number estimate, which outperforms the parametric approaches [10].

Existing DOAs estimation methods rely on prefixed and intuitive parameters for single source TF points selection. The main contribution of this paper is that we improve the method proposed in [7] and use pattern classification to select proper TF points, which not only significantly reduces the computational time, but also improves the accuracy of the subsequent nonparametric Bayesian clustering based source number counting and DOA estimation.

2. SOUND SOURCE LOCALIZATION

2.1 Uniform Circular Microphone Array Model

Research sponsored by the National Science Foundation (NSF) Grants CISE/IIS 1231671, National Natural Science Foundation of China (61673313), State Key Program for Basic Research of China (2013CB329405) and Fundamental Research Funds for the Central Universities.

The classic uniform circular microphone array model [7] is shown in Fig. 1. There are *M* microphones placed in a circle with radius *r*. Adjacent microphones have angular distance α which satisfies $\alpha=2\pi/M$. *N* sound sources are far around the microphone array. Sensor spacing is small enough to avoid the spatial aliasing problem.

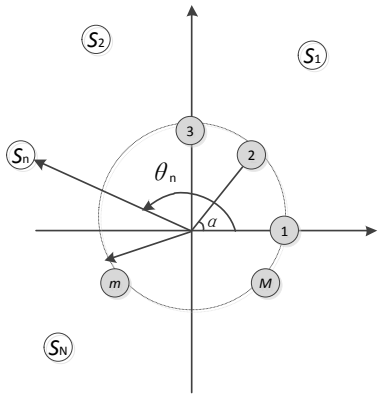


Fig.1 Geometry of a uniform circular microphone array and sound sources.

S_i is the sound source with DOA θ_i. In a reverberant environment, microphone *m* receives the following mixture signal,

$$x_{\mathrm{m}}(t)=\sum_{n=1}^{N}\sum_{l=1}^{L_{mn}}h_{mn}(\tau_l)s_n(t-\tau_l)+n_m(t),m=1,\dots,M \tag{1}$$

Here, $h_{mn}(\tau_l)$ is the room impulse response (RIR) from source *n* to microphone *m*, $\{\tau_l, l=1,\dots,L_{mn}\}$ are the time delays of the significant signal paths, and L_{mn} is the length of the RIR. Let $t_{mn}=min\{\tau_l\}$, which is the time delay of the direct path between source *n* and microphone *m*.

2.2 STFT and Sparse Analysis

For broadband signals, STFT is used to transform them to the time-frequency (TF) domain. STFT of the received signal for microphone *m* is as follows,

$$X_{\mathrm{m}}(k,f)=\sum_{n=1}^{N}H_{mn}(f)S_n(k,f)+N_m(k,f),m=1,\dots,M \tag{2}$$

where *k* is the time frame index and *f* is the frequency. Each microphone receives a mixture signal from all the sources. In the TF domain, we can utilize the sparsity of each TF point to detect the dominant source signal.

2.3 DOA Estimation

The circular integrated cross spectrum (CICS) approach [6] is proposed to locate a sound source by producing DOA for a certain frequency and frame. The calculated DOA is considered to contribute the most to this TF point. CICS is defined as

$$G_{\phi,\theta}^{(\omega)}\triangleq\sum_{i=1}^{n}G_{m_i\to m_2}^{(\omega)}(\phi)G_{m_im_{i-1}}^{(\omega)}(\theta) \tag{3}$$

where

$$G_{m_i\to m_2}^{(\omega)}(\phi)\triangleq e^{-j\omega\tau_{m_i\to m_2}(\phi)}\ (1\le i\le n) \tag{4}$$

$$G_{m_im_{i-1}}^{(\omega)}(\theta)=\frac{\Phi_{m_im_{i-1}}^{(\omega)}(\theta)}{\left|\Phi_{m_im_{i-1}}^{(w)}(\theta)\right|}\ (1\le i\le n) \tag{5}$$

$$\Phi_{m_im_j}^{(\omega)}=E\left[X_i(\omega)X_j^*(\omega)\right] \tag{6}$$

ϕ is the true direction of signal, ω is the radial frequency. For a TF point, θ which maximizes $G_{\phi,\theta}^{(\omega)}$ will be the estimated DOA.

Since DOA estimates of TF points may not always be accordance with true source directions in reverberant environments, Pavlidi *et al*. used the concept of dominant zone to restrict the TF point selection, and emphasized that only TF points whose cross-correlation over a pair of microphones is greater than a certain threshold can be used for DOA estimation.

Utilizing the power ratio of two adjacent frames and the continuity in frequency, Sun *et al*. [7] effectively extracted the sparse TF points based on some prespecified thresholds. These points correspond to the direct path of a single source, and are used to further improve the precision of DOA estimation.

Choosing appropriate threshold values has a significant effect on the localization performance and it is not straightforward. Next, we use a pattern classification based approach for TF point selection.

2.4 SVM based TF Point Selection

To determine if a TF point is sparse can be considered as a binary classification problem. The power ratio of adjacent frames and the power of the previous frame are important features, which are used to train a SVM classifier. To some extent, this method improves the accuracy of TF point selection. The diagram of the algorithm is shown in Fig. 2.

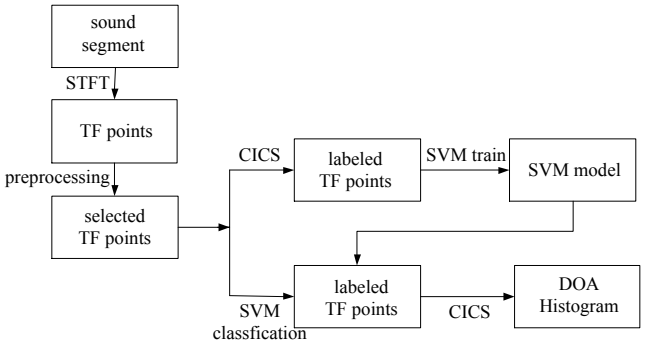


Fig.2 Diagram of proposed algorithm.

2.4.1. SVM Training

After STFT, preprocessing is conducted by choosing the TF points with top 5% power ratio with respect to their previous frames. Among the chosen TF points, the ones whose previous frame power is greater than 0.05 are eliminated. The CICS approach is then used to estimate DOAs from the selected TF points. If the DOA error is less than 5°, the TF point is labeled as 1, otherwise as 0. Finally, a SVM classifier is trained with the labeled TF points, using power ratio and previous frame power as its features.

2.4.2. SVM Classification and DOA Estimation

Similar to training, preprocessing is conducted to reduce the number of TF points of interest. The SVM is used to classify these TF points and only those classified as 1 are kept. In order to further improve the accuracy of DOA estimation, only those points that form a vertical pattern are ultimately selected [7]. Finally, the CICS approach is used and a DOA histogram is formed.

3. SOURCE NUMBER ESTIMATION

Nonparametric Bayesian methods rely on data to determine the complexity of a model [12]. The basic idea here is to use a Bayesian unsupervised learning technique to cluster the data [16].

A Dirichlet process (DP) is a distribution over distributions. Dirichlet process mixture models (DPMM) allow adapting the number of active clusters as we feed more data over time.

Let x_i be a data point and z_i be the discrete cluster label of x_i, where $i = 1, 2, \ldots, n$. For each x_i, the Chinese restaurant process (CRP) can be used to generate the corresponding cluster label z_i [13], $z_i \sim CRP(\alpha)$. In the CRP, the parameter α controls the total number of clusters generated. The data points x_i of cluster k are assumed to follow a Gaussian distribution, $x_i \sim N(\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean and covariance of cluster k. The cluster model parameters μ_k and Σ_k are drawn from a Dirichlet process, $G \sim DP(\alpha, G_0)$. Note that the base distribution G_0 acts as a prior over the model parameters μ_k and Σ_k.

Given a dataset, the cluster assignment is performed by posterior inference [14]. Let $x_{1:n}$ be the complete dataset, z_{-i} the set of cluster assignments except the one of the ith observation, x_{-i} the complete dataset excluding the ith observation, $c_{k,-i}$ the total number of observations assigned to cluster k excluding the ith observation while $\mu_{k,-i}$ and $\Sigma_{k,-i}$ are the mean and covariance matrix of cluster k excluding the ith observation, respectively. The probability of x_i in cluster k, given the dataset, all the hyperparameters α and λ of DP and G_0 is given below:

$$P(z_i = k | x_{1:n}, \alpha, \lambda) \propto P(z_i = k | z_{-i}, \alpha) P(x_i | x_{-i}, z_i = k, z_{-i}, \lambda) \quad (7)$$

$$P(z_i = k | z_{-i}, \alpha) = \begin{cases} \frac{c_{k,-i}}{\alpha + n - 1} & if\ k\ is\ an\ existing\ cluster \\ \frac{\alpha}{\alpha + n - 1} & if\ k\ is\ a\ new\ cluster \end{cases} \quad (8)$$

$$P(x_i | x_{-i}, z_i = k, z_{-i}, \lambda) \propto N(\mu_{k,-i}, \Sigma_{k,-i}) \quad (9)$$

4. SIMULATION RESULTS

4.1 Simulation Setting

The Image method [11] is used in our simulations to generate RIRs imitating indoor circumstances. A visualized instance of microphones and sound sources positions is shown in Fig. 3, where blue circle represents a circular array and red triangles are sound sources. Other parameters are listed in Table I.

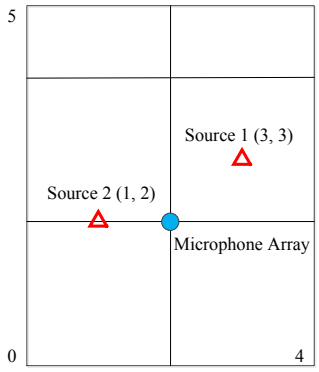


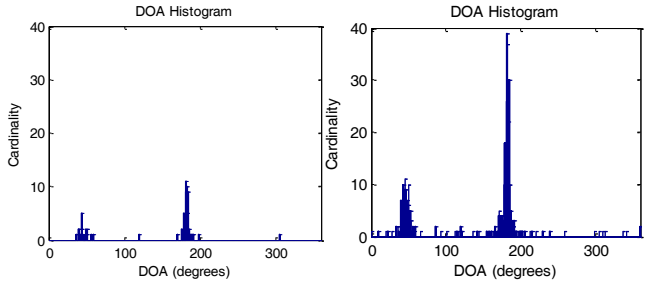
Fig. 3 Layout of sound sources and the microphone array.

TABLE I Simulation setting

Parameters	value
Room dimension	5m ×4m × 3m
Speech length	2s
Sampling rate	16kHz
Microphone array center	(2, 2, 3)
Microphone number	9
Inter-microphone distance	0.025m
STFT frame length	320
STFT frame shift	160
Reverberation time RT_{60}	0.3s

4.2 DOA Estimation

In the experiments, we use 2-second sections of the mixture of two different speech signals to train SVM with Gaussian Radial Basis Function kernel and sigma of 1. Then use 2-second sound signal mixture for classification. When source positions are set at 45° and 180° respectively, the result is shown in Fig. 4(a). While using only prefixed thresholds, the DOA histogram is shown in Fig. 4(b).



(a) (b)

Fig. 4 (a) DOA histogram using only prefixed thresholds method. (b) DOA histogram using SVM based method.

The proposed method reduces the amount of computation by selecting much fewer TF points. Excluding preprocessing, the computational time is shown in Table II, compared to the method of using only prefixed thresholds.

TABLE II Computational time (in second)

Section no.	Proposed method	Method using only prefixed thresholds
1	*3.63*	*23.52*
2	*4.59*	*21.89*
3	*2.82*	*20.49*
4	*5.95*	*27.22*
5	*5.65*	*25.31*
6	*6.71*	*26.23*
7	*4.84*	*24.80*
8	*2.47*	*20.44*

4.3 Sound Source Number Estimation

Here, we simulate three sources, which are placed at 45, 90 and 180 degrees respectively. The histogram of DOA estimates from the selected TF points is shown in Fig. 5(a). To improve clustering performance, we remove those DOA estimates whose cardinality is less than 2 in the histogram. The clustering result using DPMM is shown in Fig. 5(b)[1], which gives the correct source number. The number of clusters converges to three after 12 iterations. When two souces are close to each other (45° and 60°), it is generally harder to separate (see Fig. 6(a)). For comparison, the MDL [7] based souce number counting method outputs three clusters while the DPMM approach provides the correct cluster number (Figs. 6(b)). This demonstrates that DPMM is more effective in separating close sources, and the correct source number can help improve source localization accuracy.

[1] Note that for better visualization, in Fig. 5(b) and Fig. 6(b), the x-value of each point is the DOA estimate and the y-value follows a standard normal distribution with mean the corresponding x-value.

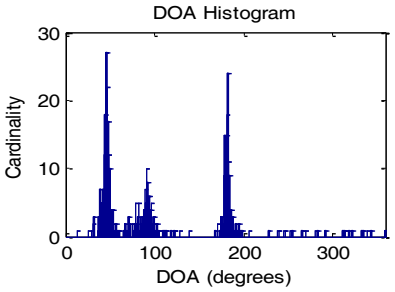


Fig. 5 (a) DOA histogram of three sources (45°, 90° and 180°);

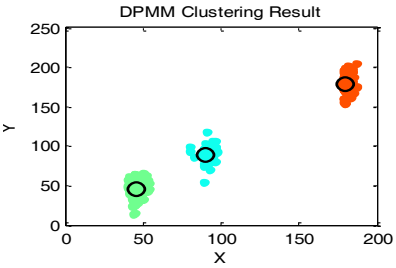


Fig. 5 (b) DPMM clustering result.

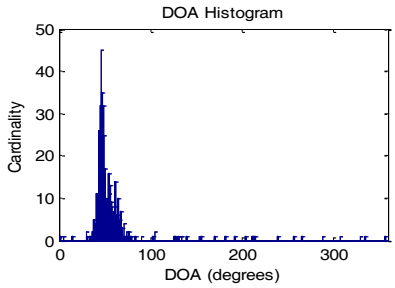


Fig. 6 (a) DOA histogram of two sources (45° and 60°);

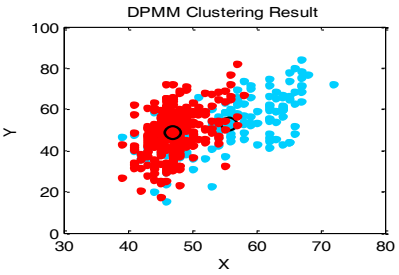


Fig. 6 (b) DPMM clustering result.

5. CONCLUSIONS

In this paper, we use pattern classification for TF point selection, which helps achieve better localization performance. This approach is feasible for real-time implementation. Besides, we use the nonparametric Bayesian clustering method to obtain an accurate source number, even when two sources are close to each other. In future work, we will conduct more experiments in both simulations and real environments, and compare with other existing methods to evaluate our proposed framework for indoor real-time multi-sound source localization.

6. REFERENCES

[1] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2293-2206, 2013.

[2] J. Chen, J. Benesty, and Y. Huang, "Time Delay Estimation in Room Acoustic Environments: An Overview." Eurasip Journal on Advances in Signal Processing , vol. 1, pp. 1-19, 2006.

[3] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays." European Journal of Biochemistry , vol. 216, no.1, pp. 281-291, 2000.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation." IEEE Transactions on Antennas & Propagation, vol. 34, no. 3, pp. 276-280, 1986

[5] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures." IEEE International Conference on Acoustics, vol. 5, pp. 2985-2988, 2000.

[6] A. Karbasi, and A. Sugiyama, "A new DOA estimation method using a circular microphone array." European Signal Processing Conference, pp. 778-782, 2007.

[7] L. Sun and Q. Cheng, "Indoor multiple sound source localization using a novel data selection scheme." 48th Conf. on Information Sciences and Systems, Princeton, NJ, pp. 1-6, 2014.

[8] O. Walter, L. Drude, and R. U. Haeb, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite Gaussian mixture model." IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 459-463, 2015.

[9] L. Sun, Q and Cheng, "Indoor sound source localization and number estimation using infinite Gaussian mixture models." Asilomar Conference on Signals, Systems & Computers. pp. 1189-1193, 2015.

[10] S. Araki, T. Nakatani and H. Sawada, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior." IEEE International Conference on Acoustics. pp. 33-36, 2009.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," The Journal of the Acoustical Society of America, vol. 65, no.4, pp. 943-950, 1976.

[12] S. J. Gershman, and D. M. Blei, "A tutorial on Bayesian nonparametric models." Journal of Mathematical Psychology, vol. 56, no.1, pp. 1-12, 2012.

[13] V. Vryniotis, http://blog.datumbox.com/overview-of-cluster-analysis-and-dirichlet-process-mixture-models/

[14] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models." Journal of Computational & Graphical Statistics, vol. 9, no.9, pp. 249-265, 2010.

[15] F. Caron, http://www.stats.ox.ac.uk/~caron/

[16] X. Yu, http://yuxiaodong.files.wordpress.com/2009/09/technical-details-in-gibbs-sampling-for-dp-mixture-model.pdf