Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 13, No. 1, March 2013, pp. 59-66
http://dx.doi.org/10.5391/IJFIS.2013.13.1.59

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Model-based Clustering of DOA Data Using von Mises Mixture Model for Sound Source Localization

Quang Nguyen Dinh and Chang-Hoon Lee
Department of Electronic Engineering, Pai Chai University, Daejeon, Korea

ijfis

Abstract

In this paper, we propose a probabilistic framework for model-based clustering of direction of arrival (DOA) data to obtain stable sound source localization (SSL) estimates. Model-based clustering has been shown capable of handling highly overlapped and noisy datasets, such as those involved in DOA detection. Although the Gaussian mixture model is commonly used for model-based clustering, we propose use of the von Mises mixture model as more befitting circular DOA data than a Gaussian distribution. The EM framework for the von Mises mixture model in a unit hyper sphere is degenerated for the 2D case and used as such in the proposed method. We also use a histogram of the dataset to initialize the number of clusters and the initial values of parameters, thereby saving calculation time and improving the efficiency. Experiments using simulated and real-world datasets demonstrate the performance of the proposed method.

Keywords: Sound source localization, Direction of arrival, Model-based clustering, von Mises distribution

1. Introduction

Sound source localization (SSL) is currently a widely researched topic in domains such as human–robot interfaces, diarization, and tracking systems. In this paper, we focus on the online task of multiple SSL.

Using a time-frequency method and a coordinate transformation on the signals received by an equilateral triangle microphone array, we can aggregate direction of arrival (DOA) information from all frequency bins in each time frame [1], giving us a complete set of DOA data for analysis. Because this data will be distributed around the true direction of sound sources, a clustering method can be applied to it to produce SSL information.

In the online task of multiple SSL, the number of sound sources is unknown and changes over time. The data may also contain many highly overlapping clusters and noises, making it difficult to return reliable results. Recognizing this, we suggest that a model-based clustering approach is the most suitable, since nearly all other clustering techniques require the number of clusters ahead of time, or are overly sensitive to noise.

On the other hand, DOA data tends to be distributed in a circular manner. Although mixtures of Gaussian distributions have been used to model DOA data [2], we will demonstrate that the circular distribution of the data makes von Mises (vM) distributions a more natural fit.

Received: Jun. 8. 2012
Revised : Mar. 14. 2013
Accepted: Mar. 15. 2013

Correspondence to: Chang-Hoon Lee
(naviro.lee@email.address)

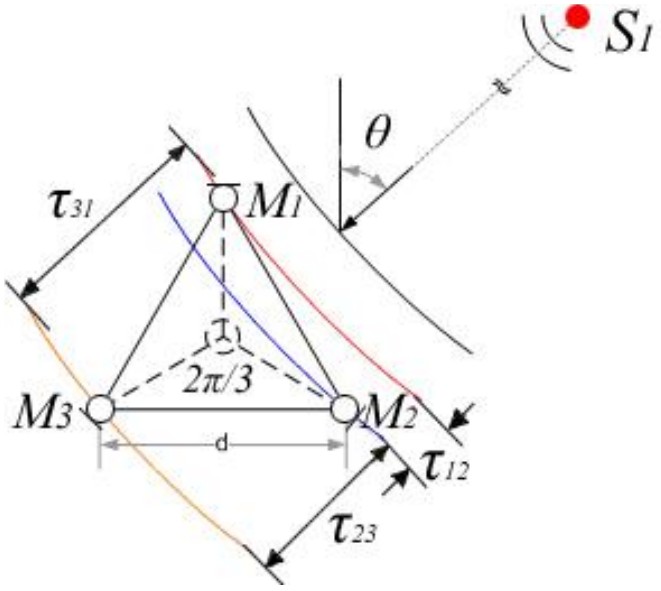

Figure 1. Equilateral triangle microphone array.

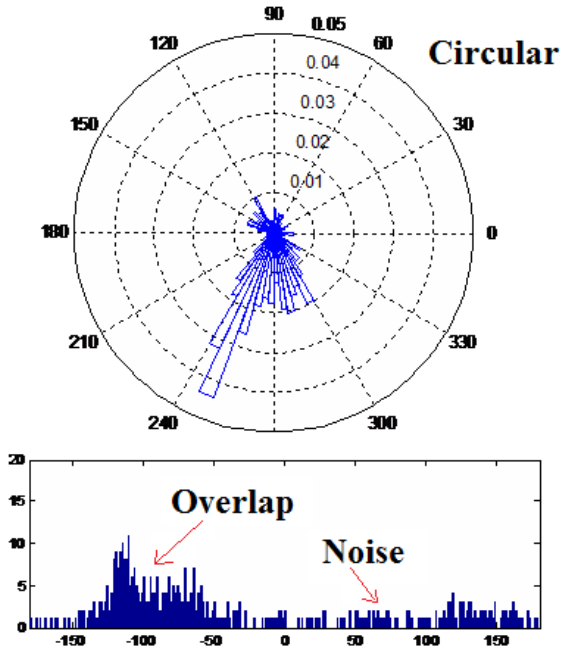

Figure 2. Typical direction of arrival (DOA) dataset.

2. Preliminary

DOA data is estimated using a time-frequency method and a coordinate transformation. The time-frequency method estimates the time delay of signals arriving at three pairs of microphones in a triangle microphone array. The layout of the microphone array is shown in Figure 1.

The time-frequency method is based on two assumptions [3]:

- The sources are disjoint in the time-frequency plane (in other words, at most one source is dominant at a time-frequency slot).
- The distance between microphones is very small compared to the distance between the sources and the microphone array (far-field approximation).

When these asumptions hold, the DOA estimates will be distributed around the true source locations and each cluster in the DOA data will represent a sound source. Note that employing a clustering method can yield an SSL estimation even in an underdetermined case (e.g., when the number of sources is larger than the number of microphones). However, DOA data does have specific characteristics that make clustering challenging:

- It is a kind of circular data, with several features distinguishing it from simpler, linear data.
- It may contain a lot of noise.
- It makes prior knowledge of the number of clusters infeasible.
- Its clusters tend to be highly overlapped.

A typical DOA dataset is shown in Figure 2.

Among various clustering algorithms, model-based clustering is the most common approach to clustering analysis because it deals with the problem of determining the intrinsic structure of clustered data when no information other than the observed values is available. In the family of model-based clustering algorithms, one generally selects certain models for clusters and then tries to optimize the fit between the data and the selected models. The EM framework is used to estimate the model with the objective of minimizing the likelihood function of the model.

In circular statistics, the vM distribution is the most popular and natural, much like the Gaussian distribution in linear statistics. Hence, although the Gaussian mixture model (GMM) for model-based clustering provided a classical and powerful approach to clustering analysis in most cases [4], we propose use of the von Mises mixture model (VMM) as the underlying model for DOA data, which are characteristically highly circular. Experiments show that VMM is more effective than GMM in SSL estimation. In [5], the authors provide a generalized EM framework for clustering multi-dimensional directional data on

the unit hyper-sphere using VMM. In this paper, DOA data is restricted to two dimensions, so that we need to only degenerate the method proposed in [5] to the 2D case before applying it to our dataset.

Although model-based clustering generally requires the number of clusters as *a priori* input, in the online task of multiple SSL, the number of active sound sources is unknown and may change over time. There are some model selection methods (e.g., BIC and AIC) used to select the best among several candidate models for different numbers of clusters [6]. Because this approach incurs the time cost of estimating model parameters for different candidate numbers of clusters, we make use of a data histogram to roughly estimate the number of clusters and their mean. This allows us to apply our clustering algorithm one time only, and helps the clustering process converge more quickly. It is imperative that the result after clustering based on these initial values may include some noise or clutter, so a threshold is used for filtering and estimate correction.

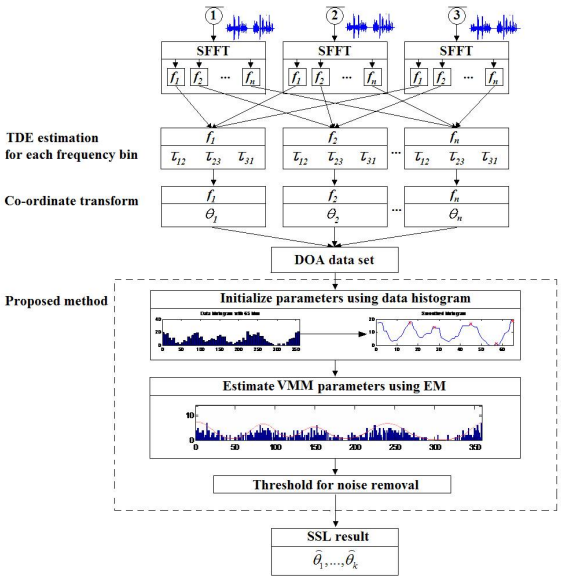

Figure 3. Proposed method workflow. DOA, direction of arrival; EM, expectation maximization; SFFT, short-time Fourier transform; SSL, sound source localization; TDE, time delay estimation; VMM, von Mises mixture model.

3. Model-based clustering of DOA data

In this section, we will describe the proposed method in detail. The overall workflow of the method is show in Figure 3. The content will concentrated in model-based clustering of DOA data.

3.1 Von Mises Distribution

In circular statistics, certain basic terms (e.g., mean value, distance) differ from those used in linear statistics. Hence, we cannot apply the distribution functions in linear statistics in the circular domain.

The vM distribution function is commonly used in circular statistics because it has the same merits as Normal Distribution in linear statistics. The vM probability distribution function (pdf) has the form [7]

$$f\left(\theta|\mu,\kappa\right)=\frac{1}{2\pi I_0(\kappa)}\exp\left(\kappa\cos\left(\theta-\mu\right)\right), \quad -\pi\le\theta\le\pi,\ -\pi\le\mu\le\pi,\ 0\le\kappa \tag{1}$$

where $I_0(\kappa)$ is the modified Bessel function of zero order, μ is the mean direction, and $kappa$ is the concentration coefficient. Figure 4 shows examples of the vM pdf with mean direction $\mu = 0$ and concentration coefficient $\kappa = 0.5$, 1, 2, and 4.

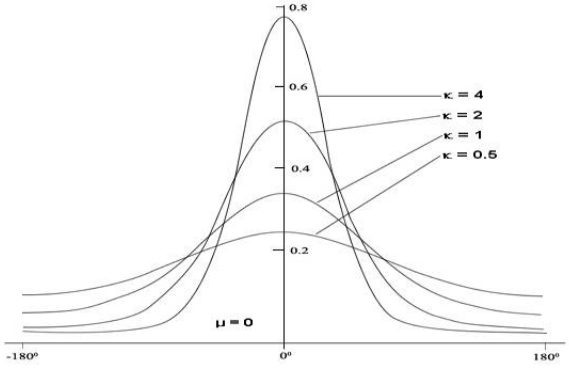

Figure 4. von Mises distribution with $\mu = 0$ and $\kappa = 0.5$, 1, 2, and 4.

3.2 Von Mises Mixture Model with EM Framework

We assume that DOA data can be modeled using a mixture of vM distributions. In [5], the authors proposed a method for modeling directional data on a unit hyper sphere using vM distributions. In our case, the dimensions are only two, so the model can be expressed as

$$f\left(\theta|\Theta\right)=\sum_{h=1}^{k}\alpha_h f_h\left(\theta|\mu_h,\kappa_h\right) \quad \text{with } \alpha_h\ge 0 \text{ and } \sum_{h=1}^{k}\alpha_h=1 \tag{2}$$

where $\Theta = \{\alpha_1, \ldots, \alpha_k, \mu_1, \ldots, \mu_k, \kappa_1, \ldots, \kappa_k\}$ is the set of all of model parameters. Assume that we have a set $X = \{\theta_1, \ldots, \theta_n\}$ of DOA data that is modeled by Eq. (2); let $Z = \{z_1, \ldots, z_n\}$ be the set of hidden random variables that indicate the vM distribution sample at the corresponding sample point, such that if the point x_i is sampled by h-th vM distribution, then $z_i = h$. The complete data log-likelihood function of the model is then expressed as

$$\ln P(X, Z|\Theta) = \sum_{i=1}^{n} \ln\left(\alpha_{z_i} f_{z_i}(\theta_i|\mu_{z_i}, \kappa_{z_i})\right). \quad (3)$$

In this case, Z is unknown, so we cannot calculate the value of likelihood function directly. However, from the given (X, Θ), it is possible to estimate the most likely conditional distribution of $Z|(X, \Theta)$ using the standard EM framework, so that

$$p(h|\theta_i, \Theta) \leftarrow \frac{\alpha_h f_h(\theta_i|\Theta)}{\sum_{l=1}^{k} \alpha_l f(\theta_i|\Theta)} \quad (4)$$

This is the expectation step (E step) in the EM framework. The maximization step (M step) will then re-estimate model parameters Θ to maximize the model likelihood function. There are two strategies for assigning samples to the clusters, suggesting two approaches to re-estimating the parameters in the M step. The two sample assignment strategies are:

- Hard assignment: a sample can be assigned only to a single cluster with the highest conditional distribution (winner takes all). The distribution of the hidden variables is given by

$$p(h|x_i, \Theta) = \begin{cases} 1, & \text{if } h = \underset{1 \le l \le k}{\arg\max} \frac{\alpha_l f_l(\theta_i|\Theta)}{\sum_{c=1}^{k} \alpha_c f(\theta_i|\Theta)} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

- Soft assignment: a sample can be assigned to many clusters with the probability given by Eq. (4).

The advantage of hard assignment is low computation cost, whereas that of soft assignment is a better fit of model to data. It should be noted that soft assignment can lead to over-fitting problems. In our research, we chose to use the hard assignment strategy due to the time constraints of real-time applications involving SSL detection. Experiments show that even hard assignment can produce good SSL estimations.

In the M step, parameters are re-estimated based on current estimates of hidden variables, according to the following equations:

$$\alpha_h = \frac{1}{n} \sum_{i=1}^{n} p(h|\theta_i, \Theta) \quad (6)$$

$$r_h = \sum_{i=1}^{n} \theta_i p(h|\theta_i, \Theta) \quad (7)$$

$$\mu_h = r_h / \|r_h\| \quad (8)$$

$$\overline{r} = \|r_h\| / n\alpha_h \quad (9)$$

$$\kappa_h = \left(2\overline{r} - \overline{r}^3\right) / \left(1 - \overline{r}^2\right) \quad (10)$$

In the above equations, $\overline{r}$ denotes the sample mean resultant vector. It is used for approximating a concentration coefficient. Note that these estimation equations are 2D simplifications of the equations in [5].

3.3 Initial Parameters

Initial parameters are crucial for model-based clustering. Good initialization of parameters can help the algorithm converge more quickly as well as avoid bad estimations. Since a histogram is a quantized version of the true pdf, we can use it to roughly choose initial parameters for our clustering algorithm. Because the mean of a component pdf strongly corresponds to the location of a histogram peak, we can use the locations of peaks in a histogram as our initial mean directions for the data model. If the dataset histogram is very complicated and yields too many peaks, reducing the number of bins will solve the problem. With DOA data, the degree value of one sample in this dataset can only lie within the range of [0, 360], so a histogram with a number of bins from 36 to 72 is recommended. In our experiment, we used a histogram with 65 bins. The histogram should also be smoothed by a smoothing function such as median filter. After smoothing, we can identify the number of peaks as the number of clusters, and the location of these peaks as the initial values for means directions. Incorrect selections of initial values can be removed using an appropriate threshold after clustering. The steps for obtaining initial values from histogram are shown in Figure 5.

3.4 Noise Threshold

The clustering results of DOA data can contain a lot of clutter caused by noise or reverberation. In estimated models, this type of clutter often appears as low priority (meaning the number of samples belonging to the clusters is small) and low concentration clusters. Thus, we proposed two thresholds for removing such clutter from our clustering results: the κ threshold ($thre_\kappa$ =

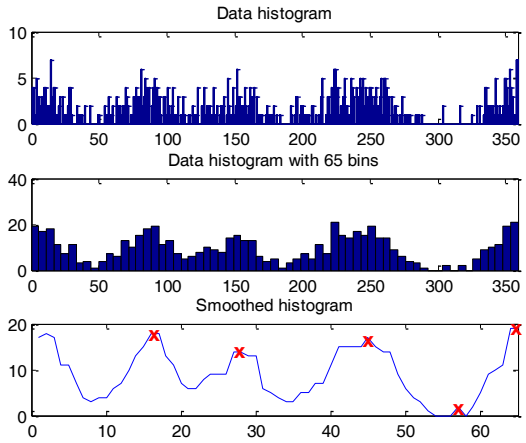

Figure 5. Obtain initial values from histogram of direction of arrival data.

0.5) and the α_h threshold ($thre_\alpha = 0.17$). Experiments showed that these thresholds effectively removed clutter without biasing estimations.

In summary, our method can be expressed by the following pseudo-code:

Algorithm: clustering DOA data using VMM

Input: A set X of DOA data from STFT and coordinate transformation.

Output: Clustering of X over a mixture of k vM distributions.

1. Initialize parameters:

a. Get initial number of clusters k and mean directions from smoothed histogram.

b. Initialize the concentration coefficient and prior probability for every cluster.

2. Estimate model parameters using EM:

repeat

E-step: evaluate conditional probability of hidden variables using given data and current parameter values according to (5).

M-step: re-estimate the parameters using the current estimate of hidden variables according to (6), (8), and (10).

until convergence

3. Apply threshold to select final result.

---

4. Experiments

Table 1. Parameters for generating the simulated dataset

	h = 1	h = 2	h = 3
α_h	0.333	0.5	0.167
μ_h	0	90	270
κ_h	10	6	8

Table 2. Result from model-based clustering of simulated data

	h = 1	h = 2	h = 3
α_h	0.339	0.490	0.172
μ_h	0	91	272
κ_h	11.6742	6.3420	6.7678

4.1 Evaluation of Clustering Algorithm

In order to test the correctness of the proposed algorithm, we tested it against simulated datasets. These datasets were composed by randomization based on the vM distribution.

The simulated dataset is composed of highly overlapped clusters randomly generated by the parameters in Table 1.

The dataset comprised 600 samples: dataset 1 had 200 samples (33.3%), dataset 2 had 300 samples (50%), and dataset 3 had 100 samples (16.7%). The histogram of input data is shown in the first diagram of Figure 2. The clustering result is shown in Table 2.

In Figure 6, the diagram below the data histogram is the reconstructed model using the parameters estimated by clustering with vMM.

From the result, we can see that the estimation is only slightly different from the parameters used to generate the model. The reconstructed model matches the data histogram closely.

4.2 Experiment on Simulated Dataset

We evaluated the utterances from a TIMIT database, which includes many individual corpora of speech data with durations of approximately 2–3 s. Speech data were randomly selected from the TIMIT database to form a single channel of data. Several of these channels were then passed into a simulated room using the Audio Systems Array Processing Toolbox [8] to obtain mixture signals. We experimented with various numbers of sound sources using both VMM and GMM and compared

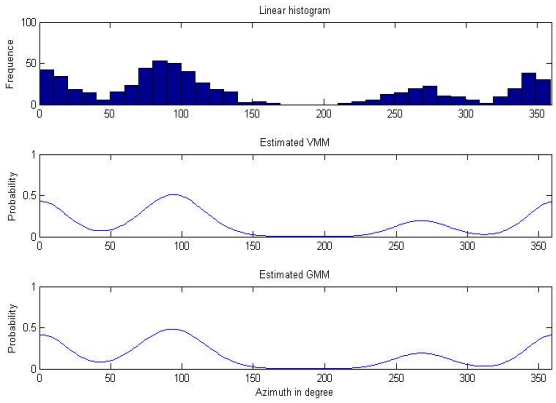

Figure 6. Histogram of dataset and estimated model by von Mises mixture model (VMM) and Gaussian mixture model (GMM).

Table 3. Parameters for the simulation

Sampling frequency	8 kHz
Number of data bits	16 bit
Wave velocity	344 m/s
Microphone array aperture	4 cm
Window	Hamming
Frame length	1,024 samples (64 ms)
Frame shift	240 samples (15 ms)
FFT point	1,024
Frequency band	120 Hz – 4,300 Hz

the performances using root-mean-square error (RMSE):

$$RMSE = \sqrt{\sum (Estimated\_DOA - True\_DOA)^2 / N}, \quad (11)$$

where N is the total number of estimates. The parameters for experiments are set as shown in Table 3. Results showed that VMM worked better than GMM in most of cases, as shown in Figure 7.

4.3 Experiment on Real World Dataset

We also performed an experiment to test the performance of the proposed algorithm in the real world. The experimental data for this test was originally used to demonstrate a BSS algorithm in [9]. Four speakers located at directions Θ = 50^o, 170^o, 250^o, and 295^o speak for approximately 6 seconds at a time.

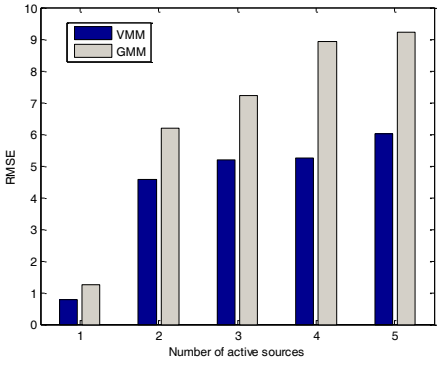

Figure 7. RMSE of DOA estimations for multiple active sources. DOA, direction of arrival; GMM, Gaussian mixture model; RMSE, root-mean-square error; VMM, von Mises mixture model.

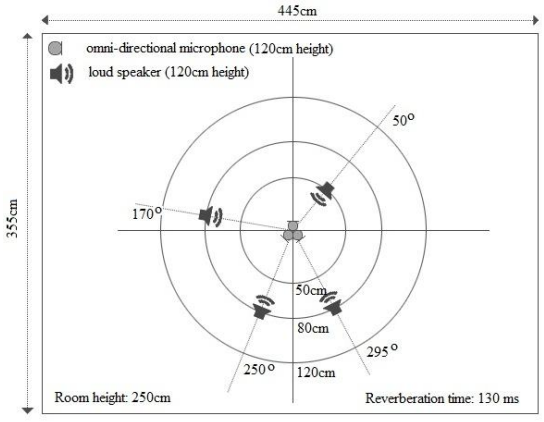

Figure 8. Experimental setup.

A mixture signal is sampled at 8 kHz. A STFT is performed for 1,024 points with a frame of 64 ms (512 samples) and shifted for each 15 ms (120 samples). The bandwidth of the signal is from 60 Hz to 3.96 kHz, resulting in 500 DOA data points in each time frame. The setup of the experiment is shown in Figure 8.

We chose some specific time frames to obtain the detailed results of clustering data from those frames, as shown in Figures 9 and 10. From these figures, we can see that the models estimated using model-based clustering are acceptable for input data, with the number of peaks as our number of clusters and the form of the graph as the initial parameters.

In Figure 11, we see a comparison of the results produced by GMM and VMM for the given dataset at each time frame. It is clear that VMM yielded better results than GMM with respect to both the validity and accuracy of estimates.

Table 4 shows the RMSE of VMM and GMM for the given

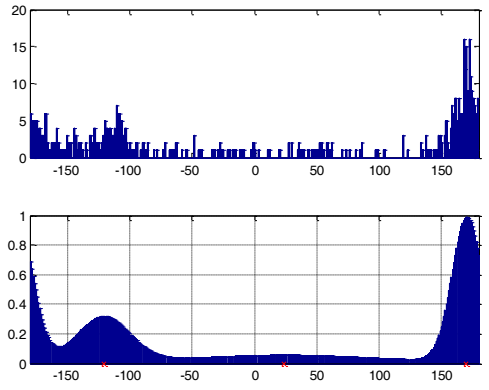

Figure 9. Clustering result at time frame #50.

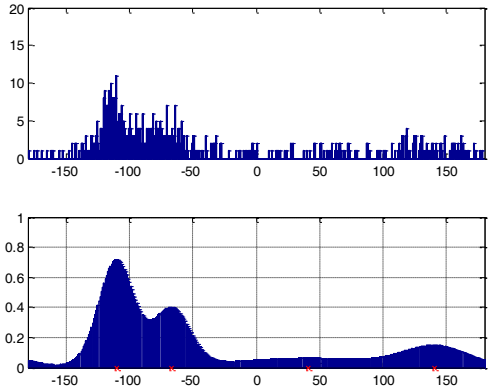

Figure 10. Clustering result at time frame #200.

dataset. Note that the proposed method yielded an RMSE of 6.1175 (in degrees) compared to 8.2218 for GMM method.

The distribution of results for the entire input signal after applying our noise threshold is shown in Figure 12. From this graph, we can see that there are four sound sources located around the real set-up directions 50^o, 170^o, 250^o, and 295^o.

5. Conclusion

The results of the experiments clearly show that model-based clustering using the von Mises mixture model can estimate the number of sources and their directions with higher stability than the Gaussian mixture model.

The proposed noise threshold effectively removed clutter from the estimated model and added further stability to the estimate. In future work, we will consider a more consistent noise threshold selection method to further reduce estimation errors.

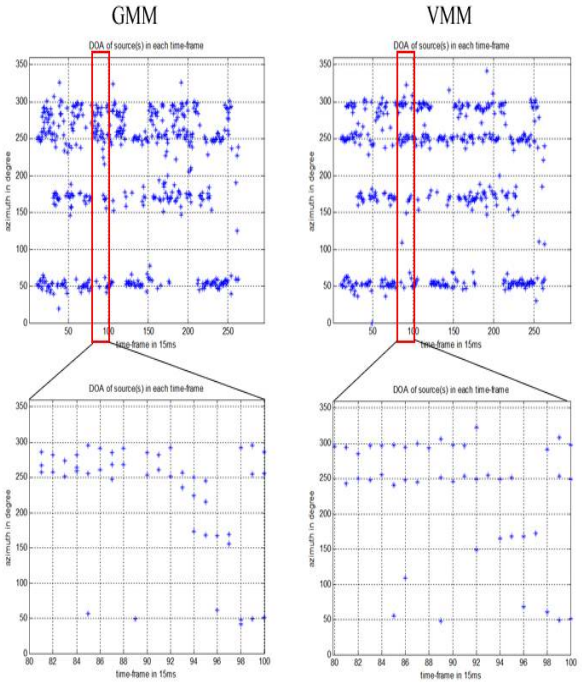

Figure 11. Direction of arrival of source(s) for each time frame. GMM, Gaussian mixture model; VMM, von Mises mixture model.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This work was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy (MOCIE).

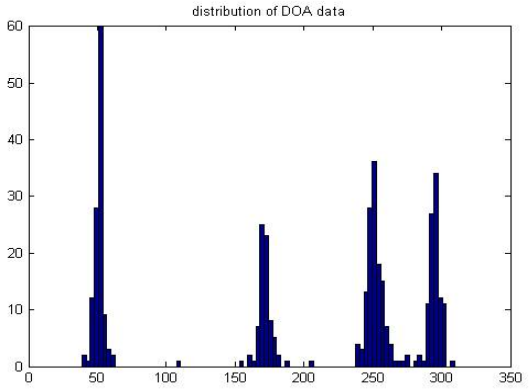

Figure 12. Distribution of estimated direction of arrival.

Table 4. RMSE for VMM and GMM on real-world dataset

	VMM	GMM
RMSE	6.0114	8.2218

GMM, Gaussian mixture model; RMSE, root-mean-square error
VMM, von Mises mixture model.

References

[1] Y. Hioka, M. Matsuo, and N. Hamada, "Multiple-speech-source localization using advanced histogram mapping method," *Acoustical Science and Technology*, vol. 30, no. 2, pp. 143-146, Mar. 2009. http://dx.doi.org/10.1250/ast.30.143

[2] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," in *Proceedings of the 9th International Conference on Digital Audio Effects*, pp. 233-238, Montreal, 2006.

[3] S. Rickard, "The DUET blind source separation algorithm. blind speech separation," in *Signals and Communication Technology*, S. Makino, T. W. Lee, and H. Sawada, Eds. Dordrecht: Springer, 2007.

[4] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, Philadelphia: Society for Industrial and Applied Mathematics, 2007.

[5] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises-fisher distributions," *Journal of Machine Learning Research*, vol. 6, no. 9, pp. 1345-1382, Sep. 2005.

[6] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261-304, Nov. 2004. http://dx.doi.org/10.1177/0049124104268644

[7] S. R. Jammalamadaka and A. SenGupta, *Topics in Circular Statistics*, River Edge: World Scientific, 2001.

[8] K. D. Donohue, "Audio systems array processing toolbox," Available http://www.engr.uky.edu/~donohue/au-dio/Arrays/MAToolbox.htm

[9] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, Eindhoven, 2005, pp. 117-120.

Quang Nguyen Dinh received the M.E. degree in electronic engineering from Pai Chai University, Korea, in 2012.
Research Area: source localization, noise reduction, embedded system
E-mail : ndquangr@gmail.com

Chang-Hoon Lee received the Ph.D. degree in system science from Tokyo Institute of Technology, Japan, in 1999. He is currently an associate professor in the Department of Electronic Engineering at the Pai Chai University.
Research Area: robot auditory system, human-robot interaction, soft computing, embedded system
E-mail: naviro.lee@email.address