# Blind Spatial Sound Source Clustering and Activity Detection Using Uncalibrated Microphone Array

Keisuke Nakamura
Honda Research Institute Japan Co., Ltd.
8-1, Honcho, Wako-shi, Saitama, 351-0188, Japan
Email: keisuke@jp.honda-ri.com

Takeshi Mizumoto
Honda Research Institute Japan Co., Ltd.
8-1, Honcho, Wako-shi, Saitama, 351-0188, Japan
Email: t.mizumoto@jp.honda-ri.com

***Abstract*—This paper presents a method for estimating the number, as well as the activity periods of spatially distributed sound sources using an uncalibrated microphone array. This methodology is applied for the purposes of speaker diarization. In general, speaker diarization has difficulty with: 1) estimating the number of sound sources (speakers), and 2) activity detection of multiple sound sources including overlap of utterances. Several microphone array based techniques have already tackled these challenges. However, existing methods mainly assume that the steering vectors for the microphone array are calibrated in advance to identify sound sources, which is difficult to satisfy when ad-hoc or flexible microphone arrays are used. Thus our approach estimates the number of sound sources blindly in two steps. First, Time Delay of Arrival (TDOA) of the observed signal is clustered. Second, the sound source activity is detected by clustering the long-term spatial spectrum using the TDOA based steering vector for each cluster. The validity of the algorithm is confirmed by both synthesized signals and a real-world flexible microphone array application.**

## I. Introduction

As the number of speech-based home assistant devices increases, technologies estimating "who is talking when" (known technically as speaker diarization) in indoor environments has become more important. Speaker diarization research mainly tackles the simultaneous estimation of speaker segmentation (voice activity detection) and clustering (number of speaker estimation). Beside monaural signal based methods [1], [2], microphone array technologies tackles this by introducing spatial information about the speakers. However, most of the existing methods assume that the microphone location is given to estimate the direction of arrival of speakers [3]–[6]. Some methods using *Time Difference Of Arrival (TDOA)* have been proposed [7]–[9], which do not assume the known microphone location. These methods propose using HMM for speaker segmentation and clustering, as well as hierarchical agglomerative clustering using spacial information. However, the methods have difficulty with overlapping speech [7] and estimating the number of speakers deterministically [8], [9].

Estimating the number of speakers has also been studied separately from speaker diarization. However, these methods mainly assume that: the microphone location is known [10]–[12], the number of microphones is more than the number of sound sources (namely *underdetermined*) [13]–[15], and the sound follows the cylindrical harmonics model [16]. Voice activity detection is also studied separately, but the microphone array based methods assume: the space for detection is limited [17]–[19], microphone location is known [20]–[23], and there is only a single target source [24]–[26].

Recently, microphone array technologies that do not assume known microphone locations and synchronous microphones, so-called "ad-hoc microphone arrays and acoustic sensor networks", have been introduced [27]. Flexible microphone arrays [28] does not assume known microphone locations but synchronous microphones, which is useful since the normal microphone array device is often limited in physical size due to its portability, while flexible microphone arrays can be extended depending on the use case. Especially in the case with a robot-embedded microphone array, it is difficult to measure the location of microphones because a robot-embedded microphone array is attached to a complex robot surface. Moreover, the free space assumption in the above mentioned methods is not always satisfied since the sound arriving at a robot includes robot- and room-acoustics due to the diffraction and reflection properties of robot bodies and reverberant rooms [29].

This paper investigates the estimation of *Number of Sound Sources (NSS)* and *Sound Source Activity Periods (SSAP)* for multiple sound sources accepting overlaps using an uncalibrated microphone array which does not assume known microphone locations. We assume that: non-overlapped sounds are dominant compared to overlapped sounds considering conversation situations, sound sources are spatially distributed and do not dramatically move (for instance speakers sit on the same chairs with accepting the change of body/face orientations), the microphone array does not move, and microphones are synchronized. To estimate NSS, we first obtain TDOAs of framed observed signals based on *Generalized Cross Correlation with Phase Transform (GCC-PHAT)* [30]. Second, we propose to select major clusters of the TDOAs based on affinity propagation [31], which determines: the number of clusters (meaning NSS), TDOAs that belong to each cluster, and the exemplar of each cluster (similar to the cluster centroid). The affinity propagation is a clustering which does not require explicit number of clusters and can cluster outliers caused by noise, reverberation, and sound overlaps. Thus, it can robustly distinguish true sources and outliers. After the

ISBN 978-0-9928626-7-1 

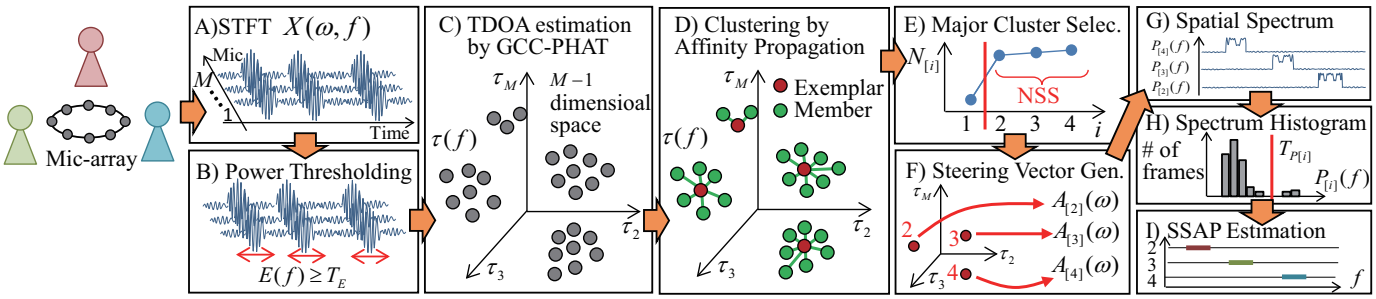


Fig. 1. Overview of the Proposed Algorithm

major cluster selection for true sources, we can estimate NSS more accurately compared to conventional hierarchical agglomerative clustering. For estimating SSAP, we first propose to compute the representative steering vector of each cluster using TDOA of the exemplar sample. Second, we propose to cluster the spatial spectrum histogram of *Multiple Signal Classification (MUSIC)* [32] using the steering vector, which is able to detect SSAP of overlapped sounds.

## II. Proposed Method

Fig. 1 shows the overview of the method. This section briefly describes each block.

### A. Estimation of Number of Sound Sources

*1) TDOA Estimation by GCC-PHAT:* Let $X_m(\omega, f)$ denote the input acoustic signal of the $m$-th channel ($1 \leq m \leq M$) after *Short Time Fourier Transform (STFT)* at the $f$-th frame, where $M$ is the number of microphones. We assume that the frame is sufficiently long with a sufficiently short period of interval. Let $\boldsymbol{X}(\omega, f) = [X_1(\omega, f), ..., X_M(\omega, f)]^T$ denotes $X_m(\omega, f)$ of all channels. First $\boldsymbol{X}(\omega, f)$ is averaged over $F$ frames as follows in order to make the spectrum robust against instant noise:

$$\bar{\boldsymbol{X}}(\omega, f) = \frac{1}{F} \sum_{i=0}^{F-1} \boldsymbol{X}(\omega, f+i). \quad (1)$$

This paper simply defines TDOA as the TDOA between the first channel $X_1(\omega, f)$ and others [1]. Finally, TDOA of the $m$-th channel in the $f$-th frame $\tau_m(f)$ is computed as follows:

$$\tau_m(f) = \underset{\tau}{\operatorname{argmax}} \int_{\omega_L}^{\omega_H} \frac{\bar{X}_1(\omega, f)\bar{X}_m^*(\omega, f)}{\left|\bar{X}_1(\omega, f)\bar{X}_m^*(\omega, f)\right|} \mathrm{e}^{j\omega\tau} d\omega \ , \quad (2)$$

where $()^*$ is a complex conjugate transpose operator, and $\omega_L$ and $\omega_H$ are the minimum and maximum frequency considered in TDOA estimation, and $\bar{X}_m(\omega, f)$ is $\bar{\boldsymbol{X}}(\omega, f)$ of the $m$-th channel. The range of $\tau$ is defined as $-D_m/c \leq \tau \leq D_m/c$, where $D_m$ and $c$ are the maximum array size candidate (which can be rough estimate) and the speed of sound, respectively. Finally, the TDOA vector for the clustering is obtained as $\boldsymbol{\tau}(f) = [\tau_2(f), ..., \tau_M(f)]^T$ whose size is $M-1$.

[1]Although $X_1(\omega, f)$ is being used as the reference channel, the selection is studied previously, for example in [9]. Therefore, the proposed method can be extended.

$\boldsymbol{\tau}(f)$ tends to be noisy when there is no spatially salient sound source. Thus, we simply eliminate silent frames based on the following thresholding before computing TDOA.

$$E(f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} ||\bar{\boldsymbol{X}}^*(\omega, f)\bar{\boldsymbol{X}}(\omega, f)||_2, \quad (3)$$

and the frames satisfying $E(f) < T_E$ are rejected from the GCC-PHAT computation, where $T_E$ is a threshold, which is described in Fig. 1-B).

*2) Affinity Propagation Based TDOA Clustering:* We cluster the estimated TDOAs $\boldsymbol{\tau}(f)$ to estimate NSS. This paper assumes speakers are spatially distributed, so TDOAs from the same speaker gather in a sufficiently small space. The difficulties for the sound source clustering are twofold: the number of clusters is unknown, and the data points are noisy due to noise, reverberation, and sound overlaps, etc. To tackle these difficulties, we introduce affinity propagation [31], which is a type of clustering that does not need to set the number of clusters. The method first defines *similarity* $S(i, j)$ between $i$-th and $j$-th data points and initializes candidate exemplars and updates two parameters, *responsibility* and *availability*, of each exemplar alternately and iteratively to decide which point should be an exemplar. Finally, we can obtain: the number of clusters $C$, the exemplar sample for each cluster $\hat{\boldsymbol{\tau}}_{[i]}$ ($1 \leq i \leq C$), the number of members in each cluster $N_{[i]}$ ($1 \leq i \leq C$), and the set of cluster members $\boldsymbol{\tau}_{[i]}$ ($1 \leq i \leq C$). In general, the advantages of the method are: the number of clusters is determined automatically, the performance is robust against initial states, and the similarity does not have to be symmetric and satisfy triangle inequality.

For the affinity propagation of TDOAs, we use the similarity definition as a negative squared euclidean distance of two data points as follows:

$$S(i, j) = -||\boldsymbol{\tau}(i) - \boldsymbol{\tau}(j)||_2^2 \ . \quad (4)$$

The set of non-clustered $\boldsymbol{\tau}(f)$, described in Fig. 1-C), becomes $C$ clusters with exemplar samples like Fig. 1-D).

*3) Major Cluster Selection:* After the clustering, some small clusters are organized due to the noise, reverberation, and sound overlaps. Based on the assumption mentioned in Section I, the cluster size between true sources and noise has a sufficiently salient gap. Therefore, we reject small clusters based on thresholding of the ratio of cluster size. For this thresholding, first, the clusters are sorted based on the number

of members $N_{[i]}$ $(1 \leq i \leq C)$ in the ascending order (Fig. 1-E)), and too small clusters satisfying $N_{[i]} < T_N$ is eliminated, which makes $C$ smaller to $\hat{C}$. $T_N$ is empirically derived as $T_N = 100$. Then we compute the ratio of the neighboring cluster size as follows:

$$\hat{N}_{[i]} = \frac{N_{[i]}}{N_{[i-1]}} (2 \leq i \leq \hat{C}) \ , \tag{5}$$

and the smallest $i$ that satisfies $\hat{N}_{[i]} > T_R$ is derived as $\hat{i}$. $T_R$ is empirically derived as $T_R = 1.5$. Finally, the clusters whose indices are $i < \hat{i}$ are rejected. If $\hat{N}_{[i]} \leq T_R$, all clusters are selected. Finally, $\hat{C} - \hat{i} + 1$ is the estimated NSS.

### B. Estimation of Sound Source Activity Periods

*1) Steering Vector Generation:* Each selected major cluster has the exemplar sample $\hat{\boldsymbol{\tau}}_{[i]}$ $(\hat{i} \leq i \leq \hat{C})$, which is the representative TDOA of the $i$-th cluster. Let $\hat{\boldsymbol{\tau}}_{[i]} = [\hat{\tau}_{[i]2}, ..., \hat{\tau}_{[i]M}]^T$ denotes TDOA of all channels. Therefore the candidate steering vector for the $i$-th cluster $\boldsymbol{A}_{[i]}(\omega)$ is described as

$$\boldsymbol{A}_{[i]}(\omega) = [\mathrm{e}^{j\omega 0}, \mathrm{e}^{j\omega\hat{\tau}_{[i]2}}, \cdots, \mathrm{e}^{j\omega\hat{\tau}_{[i]M}}]^T \ , \tag{6}$$

where the phase difference of the first channel is defined as zero since it is the TDOA between the same channels. In order to avoid all elements to have negative phase difference, we modified Eq. (6) to add $D_m/c$ to the time difference, which becomes as follows:

$$\boldsymbol{A}_{[i]}(\omega) = [\mathrm{e}^{j\omega\frac{D_m}{c}}, \mathrm{e}^{j\omega(\hat{\tau}_{[i]2}+\frac{D_m}{c})}, \cdots, \mathrm{e}^{j\omega(\hat{\tau}_{[i]M}+\frac{D_m}{c})}]^T \ , \tag{7}$$

which is schematically described in Fig. 1-F).

*2) Spatial Spectrum Computation by MUSIC:* We used MUSIC [32] to compute the spatial spectrum. We first compute a correlation matrix of $\bar{\boldsymbol{X}}(\omega, f)$ and take its Eigen value decomposition. Let $\boldsymbol{V}(\omega, f) = [\boldsymbol{v}_1(\omega, f), \ldots, \boldsymbol{v}_M(\omega, f)]$ denotes the Eigen vectors. Finally, the spatial spectrum is computed as:

$$P_{[i]}(f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} \frac{|\boldsymbol{A}^*_{[i]}(\omega)\boldsymbol{A}_{[i]}(\omega)|}{\sum_{m=L+1}^{M} |\boldsymbol{A}^*_{[i]}(\omega)\boldsymbol{v}_m(\omega, f)|}, \tag{8}$$

where $L$ is the number of sound sources. Here we defined $L = 1$ since the purpose is mainly to detect one sound source. The example of the sequence of $P_{[i]}(f)$ is shown in Fig. 1-G).

*3) Estimation of SSAP by Spectrum Histogram:* The histogram based SSAP estimation is inspired by long-term signal variability with adaptive thresholding [33]. We take the long-term histogram of $P_{[i]}(f)$ for each $i$ after elimination of invalid frames when $P_{[i]}(f) = 0$. The threshold $T_{P[i]}$ is determined for each $i$ based on k-means clustering of the histogram with $k = 2$, meaning "active" and "inactive" clusters. $T_{P[i]}$ is obtained as the minimum value of $P_{[i]}(f)$ that is classified as "active" cluster, namely higher value of minimum values of 2 clusters. The intuitive diagram is shown in Fig. 1-H). The general limitation of k-means is that it has to tune the number of clusters, but in this case the number of clusters is automatically determined as two. Finally, SSAP of the $i$-th sound source is determined by the frames satisfying $P_{[i]}(f) \geq T_{P[i]}$ (Fig. 1-H)).

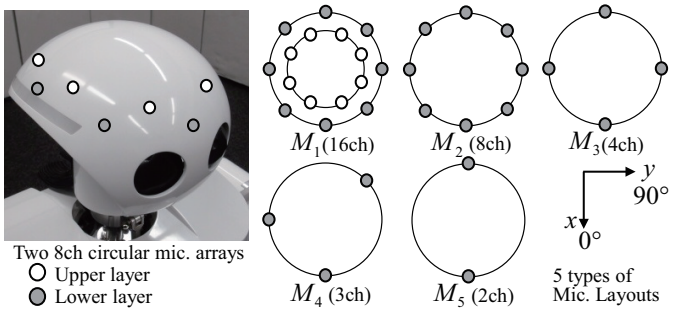


Fig. 2. Robot-embedded Microphone Array Setup and Layouts

## III. Evaluation

This section shows two types of evaluations as follows:

- Estimation accuracy of NSS and SSAP to see the effectiveness of the proposed algorithm using synthetic data using recorded directional white noise (Section III-A)
- Application to normal conversation using a flexible microphone array (Section III-B)

Both evaluations used a normal room whose reverberation time and size were 0.2s (RT20) and 4m×7m, respectively. The signal was sampled with 16kHz and 16bits while frame size and shift length were 512 and 160 samples, respectively. $F = 10$ in Eq. (1). $\omega_L$ = 500Hz and $\omega_H$ = 2800Hz.

### A. Estimation Accuracy of NSS and SSAP

This section evaluates the estimation accuracy of NSS and SSAP with the variation of number of microphones, number of sound sources/locations, and overlapping periods. We used a robot-embedded microphone array shown in Fig. 2 where the free space assumption does not hold. The robot has two 8ch circular microphone arrays (in total 16ch), and we selected 5 types of microphone layouts as $M_i(1 \leq i \leq 5)$, shown in Fig. 2, so as to see the robustness against the change of the number of microphones. We considered 10 types of sound source layouts as $S_i(1 \leq i \leq 10)$, shown in Table I. We first recorded 4.0s white noise on the same horizontal plane as the mirophone array from each direction with the distance of 1.0m and synthesized each white noise one by one with the following three kinds of intervals: 1) 0.5s interval, *Non-overlapped* shown in Table II, 2) 0.8s [20%] overlap, *Overlapped* shown in Table III, 3) $O_i = i$×0.8s [$i$×20%] overlap $(0 \leq i \leq 5)$, shown in Table IV.

We evaluated the following criteria: NSS, *Recall Rate (RR)* and *Precision Rate (PR)* of SSAP estimation. RR and PR are defined as follows:

$$\text{RR} = \frac{\text{\# of correct frames}}{\text{\# of active frames}}, \text{PR} = \frac{\text{\# of correct frames}}{\text{\# of frames estimated as active}}.$$

The results are shown in Table II for the non-overlapped case and Table III for the overlapped case. The NSS which was correctly estimated is shown as bold. For more than 2 microphones, NSS was correctly estimated. In the case of $M_5$, NSS was not correctly estimated when $S_1$, $S_2$, $S_4$, and $S_6$ due to the spatial ambiguity, so-called front and back confusion. In most of the cases, RR and PR are around 90% or more, which validates the proposed algorithm. NSS was correctly

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE I
10 TYPES OF SOUND SOURCE LAYOUT ($S_1, \cdots, S_{10}$)

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ | $S_8$ | $S_9$ | $S_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Initial [deg] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Interval [deg] | 30 | 45 | 30 | 60 | 45 | 90 | 60 | 120 | 90 | 180 |
| Final [deg] | 360 | 360 | 180 | 360 | 180 | 360 | 180 | 360 | 180 | 360 |
| NSS | 12 | 8 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 2 |

TABLE II
EVALUATION FOR NON-OVERLAPPED SOUNDS. NOTATION : NSS[RR/PR]

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
|---|---|---|---|---|---|
| $S_1$ | **12** [1.00/0.89] | **12** [1.00/0.89] | **12** [1.00/0.76] | **12** [1.00/0.71] | 4 [1.00/0.15] |
| $S_2$ | **8** [1.00/0.89] | **8** [1.00/0.89] | **8** [1.00/0.89] | **8** [1.00/0.34] | 5 [1.00/0.18] |
| $S_3$ | **7** [1.00/0.89] | **7** [1.00/0.89] | **7** [1.00/0.89] | **7** [1.00/0.89] | **7** [1.00/0.22] |
| $S_4$ | **6** [1.00/0.89] | **6** [1.00/0.89] | **6** [1.00/0.89] | **6** [1.00/0.89] | 4 [1.00/0.27] |
| $S_5$ | **5** [1.00/0.89] | **5** [1.00/0.89] | **5** [1.00/0.89] | **5** [1.00/0.64] | **5** [1.00/0.25] |
| $S_6$ | **4** [1.00/0.89] | **4** [1.00/0.89] | **4** [1.00/0.89] | **4** [1.00/0.89] | 2 [1.00/0.30] |
| $S_7$ | **4** [1.00/0.89] | **4** [1.00/0.89] | **4** [1.00/0.89] | **4** [1.00/0.89] | **4** [1.00/0.35] |
| $S_8$ | **3** [1.00/0.89] | **3** [1.00/0.89] | **3** [1.00/0.89] | **3** [1.00/0.89] | **3** [1.00/0.53] |
| $S_9$ | **3** [1.00/0.89] | **3** [1.00/0.89] | **3** [1.00/0.89] | **3** [1.00/0.89] | **3** [1.00/0.53] |
| $S_{10}$ | **2** [1.00/0.89] | **2** [1.00/0.89] | **2** [1.00/0.89] | **2** [1.00/0.89] | **2** [1.00/0.86] |

TABLE III
EVALUATION OF OVERLAPPED SOUNDS. NOTATION : NSS[RR/PR]

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
|---|---|---|---|---|---|
| $S_1$ | **12** [0.92/0.95] | **12** [0.90/0.97] | **12** [0.89/0.89] | **12** [0.85/0.71] | 5 [1.00/0.11] |
| $S_2$ | **8** [0.90/0.97] | **8** [0.90/0.97] | **8** [0.88/0.97] | **8** [0.90/0.65] | 5 [1.00/0.17] |
| $S_3$ | **7** [0.91/0.97] | **7** [0.90/0.96] | **7** [0.89/0.96] | **7** [0.88/0.97] | **7** [0.82/0.46] |
| $S_4$ | **6** [0.90/0.97] | **6** [0.90/0.97] | **6** [0.90/0.97] | **6** [0.87/0.97] | 3 [0.91/0.55] |
| $S_5$ | **5** [0.90/0.96] | **5** [0.91/0.97] | **5** [0.89/0.97] | **5** [0.92/0.62] | **5** [1.00/0.26] |
| $S_6$ | **4** [0.91/0.96] | **4** [0.90/0.96] | **4** [0.87/0.96] | **4** [0.91/0.65] | 2 [0.78/0.76] |
| $S_7$ | **4** [0.92/0.96] | **4** [0.92/0.96] | **4** [0.92/0.96] | **4** [0.89/0.96] | **4** [0.88/0.88] |
| $S_8$ | **3** [0.92/0.95] | **3** [0.92/0.95] | **3** [0.88/0.95] | **3** [0.90/0.93] | **3** [0.99/0.59] |
| $S_9$ | **3** [0.88/0.95] | **3** [0.92/0.96] | **3** [0.92/0.95] | **3** [0.94/0.93] | **3** [0.92/0.58] |
| $S_{10}$ | **2** [0.93/0.94] | **2** [0.93/0.94] | **2** [0.94/0.90] | **2** [0.93/0.94] | **2** [0.99/0.91] |

TABLE IV
EVALUATION OF OVERLAP PERIODS. NOTATION : NSS[RR/PR]

| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
|---|---|---|---|---|---|
| $O_0$ | **7** [0.96/0.92] | **7** [0.95/0.93] | **7** [0.94/0.93] | **7** [0.94/0.94] | **7** [0.85/0.43] |
| $O_1$ | **7** [0.91/0.97] | **7** [0.90/0.96] | **7** [0.89/0.96] | **7** [0.88/0.97] | **7** [0.82/0.46] |
| $O_2$ | **7** [0.91/0.94] | **7** [0.78/0.95] | **7** [0.79/0.96] | **7** [0.85/0.84] | **7** [0.75/0.46] |
| $O_3$ | **7** [0.88/0.84] | **7** [0.79/0.83] | **7** [0.85/0.94] | **7** [0.86/0.75] | **7** [0.71/0.48] |
| $O_4$ | 8 [0.86/0.90] | 8 [0.97/0.91] | 3 [0.96/0.91] | 3 [0.94/0.91] | **7** [0.70/0.52] |
| $O_5$ | 4 [0.97/0.75] | 2 [0.96/0.91] | 3 [0.94/0.93] | 3 [0.94/0.55] | 4 [0.94/0.69] |

estimated even when the number of microphones is less than the number of sound sources. Since the sound sources are not simultaneous, this is not an underdetermined condition, however this shows the method effectively utilizes temporal sparseness to handle high number of sound sources. Table IV shows the result with the variation of overlapped periods when $S_3$. As shown in the table, the proposed algorithm could estimate NSS up to 60% overlap. The robustness improvement for more overlapped sounds is the future work.

Fig. 3 shows a result of each proposed step using an overlapped sounds when $S_3$, $M_1$, and $O_3$. Fig. 3-B) shows the sequence of $\boldsymbol{\tau}(f)$, and rejected frames based on $E(f)$ are shown as white. Fig. 3-C) shows $N_{[i]}$ $(1 \leq i \leq C)$, and the clusters above the red line are the selected major clusters. Fig. 3-D) shows the sequence of $\boldsymbol{\tau}_{[i]}$ $(\hat{i} \leq i \leq \hat{C})$, which does not accept overlapped sounds. Fig. 3-E) shows the sequence of $P_{[i]}(f)$. Fig. 3-F) shows the histogram of $P_{[\hat{i}]}(f)$ of all frames and $T_{P[\hat{i}]}$, which shows k-means clustering successfully determined the threshold. Fig. 3-G) shows SSAP estimation results using $T_{P[i]}$, which accepts overlapped sounds and improves SSAP estimation.

### B. Flexible Microphone Array Application

The proposed method was applied to the flexible microphone array shown in Fig. 4 where we randomly put each

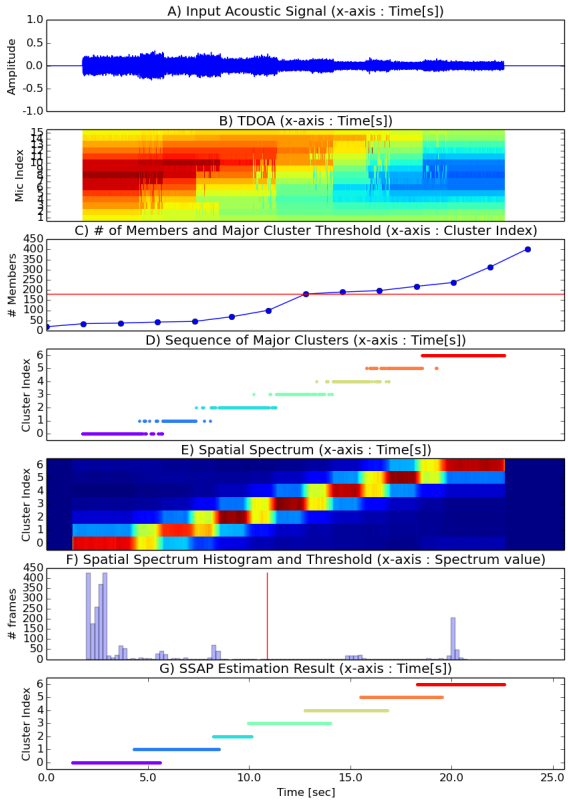


Fig. 3. Results of Each Step When $S_3$, $M_1$, and $O_3$

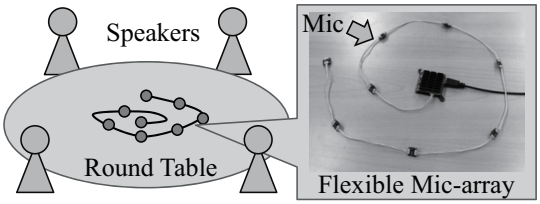


Fig. 4. Flexible Microphone Array Setup

microphone on a circular table. As shown in Fig. 4, there were 4 speakers seated around the table (approximately 1.5m from the array with different heights), and the azimuth difference between two speakers was approximately 90 degrees. Fig. 5 shows the result of 30s free conversation. Compared to Fig. 3, Fig. 5 changed Fig. 5-F) from the histogram to the hand labeled SSAP. Fig. 5-G) shows considerable similarity with Fig. 5-F), and RR = 0.59, PR = 0.81. We recorded 15 minutes conversation and divided it into 30 of 30s conversation, and the average and standard deviation of NSS estimation is $3.77 \pm 0.62$, which shows the validity of the proposed algorithm with natural conversation.

## IV. CONCLUSION

This paper investigated the estimation of NSS and SSAP using an uncalibrated microphone array. We proposed the major cluster selection of affinity propagation of TDOA to estimate NSS robust against noise, reverberation, sound overlaps, etc. To estimate SSAP of overlapped sounds, we proposed to cluster the long-term spatial spectrum into *active* and *inactive* using the steering vector estimated by representative TDOA.

ISBN 978-0-9928626-7-1 © EURASIP 2017

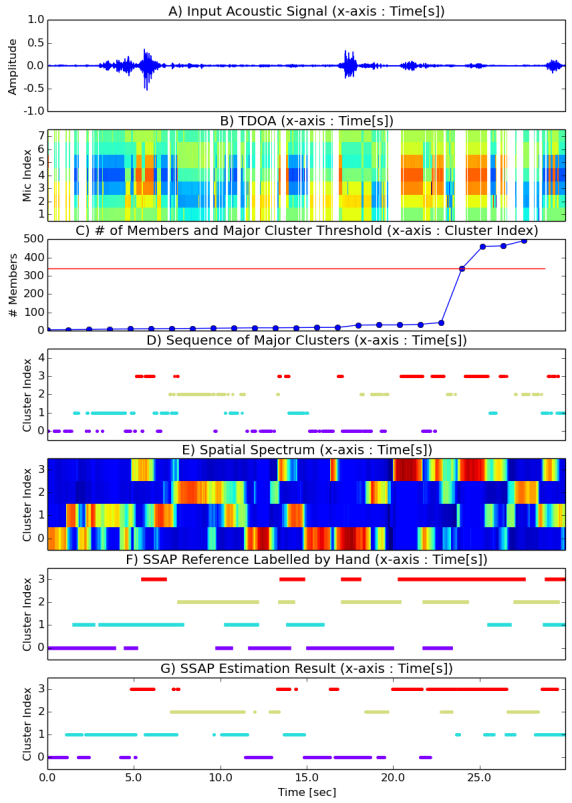


Fig. 5. Results of Each Step Using a Flexible Microphone Array

The evaluation showed: 1) NSS was correctly estimated when the microphone apparatus did not have spatial ambiguity and the overlap is sufficiently short, 2) SSAP were estimated with high performance for both synthesized and real-world data, which proved the effectiveness of the proposed algorithm.

We have a variety of future works planned. As mentioned above, NSS estimation had some error when overlap period was long, so robustness against overlap period should be improved. Additionally, the extension of the proposed method to speaker diarization by introducing sound source separation and automatic speech recognition should be investigated.

## REFERENCES


[1] S. E. Tranter *et al.,* "An overview of automatic speaker diarization systems," in *IEEE TASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," in *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, 2012.

[3] K. Ishiguro *et al.*, "Probabilistic Speaker Diarization With Bag-of-Words Representations of Speaker Angle Information," in *IEEE TASLP*, vol. 20, no. 2, pp. 447–460, 2012.

[4] X. Anguera *et al.*, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. of IEEE ASRU*, pp. 426–431, 2005.

[5] J. Schmalenstroeer and R. Haeb-Umbach, "Online Diarization of Streaming Audio-Visual Data for Smart Environments," in *IEEE J-STSP*, vol. 4, no. 5, pp. 845–856, 2010.

[6] D. Korchagin, "Audio spatio-temporal fingerprints for cloudless real-time hands-free diarization on mobile devices," in *Proc. of HSCMA*, pp. 25–30, 2011.

[7] D. Vijayasenan *et al.*, "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization," in *IEEE TASLP*, vol. 19, no. 2, pp. 431–438, 2011.

[8] M. Zelenak *et al.*, "Simultaneous Speech Detection With Spatial Features for Speaker Diarization," in *IEEE TASLP*, vol. 20, no. 2, pp. 436–446, 2012.

[9] X. Anguera *et al.*, "Acoustic Beamforming for Speaker Diarization of Meetings," in *IEEE TASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[10] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. of IWAENC*, 2008.

[11] Jwu-Sheng Hu and Chia-Hsin Yang, "Estimation of Sound Source Number and Directions under a Multisource Reverberant Environment," in *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 63, 2010.

[12] Radu Balan, "Estimator for Number of Sources Using Minimum Description Length Criterion for Blind Sparse Source Mixtures," in *Independent Component Analysis and Signal Separation*, vol. 4666, pp. 333–340, 2007.

[13] K. Yamamoto *et al.*, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Proc. of IEEE ICASSP*, vol. 5, pp. 485–488, 2003.

[14] H. Sawada *et al.*, "Estimating the number of sources using independent component analysis," in *Acoust. Sci. & Tech Letter*, vol. 5, pp. 450-452, 2005.

[15] V. Choqueuse *et al.*, "Blind detection of the number of communication signals under spatially correlated noise by ICA and K-S tests," in *Proc. of IEEE ICASSP*, pp. 2397–2400, 2008.

[16] H. Teutsch and W. Kellerman, "Estimation of the number of wideband sources in an acoustic wave field using eigen-beam processing for circular apertures," in *IEEE WASPAA*, pp. 110–113, 2005.

[17] X. Wang *et al.*, "A reverberation robust target speech detection method using dual-microphone in distant-talking scene," in *Speech Communication*, vol. 72, pp. 47–58, 2015.

[18] J. H. Choi and J. H. Chang, "Dual-Microphone Voice Activity Detection Technique Based on Two-Step Power Level Difference Ratio," in *IEEE TASLP*, vol. 22, no. 6, pp. 1069–1081, 2014.

[19] J. Park *et al.*, "Dual Microphone Voice Activity Detection Exploiting Interchannel Time and Level Differences," in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1335–1339, 2016.

[20] E. Nemer and A. Pandey, "A dual-microphone subband-based Voice Activity Detector using higher-order cumulants," in *Proc. of IEEE ICASSP*, pp. 5954–5958, 2014.

[21] K. Ishizuka *et al.*, "Speech Activity Detection for Multi-Party Conversation Analyses Based on Likelihood Ratio Test on Spatial Magnitude," in *IEEE TASLP*, vol. 18, no. 6, pp. 1354–1365, 2010.

[22] I. Potamitis and E. Fishler, "Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays," in *J. Acoust. Soc. Amer.*, vol. 116, pp. 2406–2415, 2004.

[23] A. Davis *et al.*, "A subband space constrained beamformer incorporating voice activity detection," in *Proc. of IEEE ICASSP*, vol. 3, pp.65–68, 2005.

[24] M. W. Hoffman *et al.*, "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," in *IEEE TSAP*, vol. 9, no. 2, pp. 175–179, 2001.

[25] Y. Hioka and N. Hamada, "Voice activity detection with array signal processing in the wavelet domain," in *IEICE Trans. Fundamentals*, vol. E86-A, pp. 2802–2811, 2003.

[26] L. Armani *et al.*, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. of Interspeech*, pp. 501–504, 2003.

[27] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. of 18th IEEE SCVT*, pp. 1–6, 2011.

[28] H. Saruwatari *et al.*, "Flexible microphone array based on multichannel nonnegative matrix factorization and statistical signal estimation, " in *Proc. of International Congress on Acoustics*, pp. 1–10, 2016.

[29] K. Nakamura *et al.*, S. Ambrose and K. Nakadai, "On-the-spot calibration of microphone array Transfer Functions for robot audition," in *Proc. of IEEE ICRA*, pp. 3354–3359, 2015.

[30] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Trans. on ASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[31] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," in *Science*, no. 315, pp. 972–976, 2007.

[32] R. Schmidt, "Multiple emitter location and signal parameter estimation", in *IEEE Trans. Ant. Prop.*, vol. 34, no. 3, pp. 276–280, 1986.

[33] P Ghosh *et al.*, "Robust voice activity detection using long-term signal variability," in *IEEE TASLP*, vol. 19, no. 3, pp. 600–613, 2011.


ISBN 978-0-9928626-7-1 © EURASIP 2017