Blind Estimation of Locations and Time Offsets for Real Distributed Recording Devices

Keisuke Hasegawa, Nobutaka Ono, Shigeki Miyabe, and Shigeki Sagayama

Department of Information Physics and Computing,
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, Japan
{khasegawa,onono,miyabe,sagayama}@hil.t.u-tokyo.ac.jp

Key words: Blind alignment, asynchronous recording, generalized cross correlation

Abstract. This paper presents a blind technique to estimate locations and recording time offsets of distributed recording devices from asynchronously recorded signals. In our method, locations of sound sources and recording devices, and the recording time offsets are estimated from observed time differences of arrivals (TDOAs) by decreasing the mean squared errors. The auxiliary-function-based updates guarantee the monotonic decrease of the objective function at each iteration. The TDOAs are estimated by the generalized cross correlation technique. The validity of our approach is shown by experiments in real environment, where locations of seven sound sources and eight microphones and eight time offsets were estimated from signals recorded by four stereo IC recorders in reverberant rooms.

1 Introduction

Microphone array processing is one of the powerful techniques for sound source localization and separation and it has been greatly developed in several decades. The framework has been recently extended from conventional fixed microphone array to *distributed* microphone array [1–4]. It is a new concept to exploit distributed recording devices such as internal microphones of personal computers, voice recorders, mobile phones as channels of array signal processing. Its wireless configuration, flexibility, large number of recording elements will enlarge application. The growth of mobile recording devices and the miniaturization of microphones such as silicon microphones would facilitate this direction.

While, unlike conventional microphone array, channels recorded by different devices are not synchronous in most cases due to their different recording time offsets, mismatch of sampling frequencies and so on. The location of their devices would be unknown. Therefore, in order to apply conventional array signal processing technique, it is necessary to estimate them, which is a new kind of blind estimation problem. In our previous work [6], we have proposed a method to estimate positions of sources and microphones and time offsets from observed

time differences of arrivals (TDOAs) for sources in a blind fashion and shown proof of concept by preliminary experiments.

In this paper, we challenge to the blind estimation of locations and time offsets of distributed recording devices in real environments. In order to obtain apparent time differences accurately from asynchronously recorded signals, we present the two-stage estimation method, where rough alignment of entire recorded signals is first performed, and then, the apparent TDOA for each source, (which still includes unknown time offsets,) is estimated by generalized cross correlation method (GCC method [7]). Finally, positions of sources and microphones and time offsets are estimated by minimizing square errors of observation model with auxiliary-function-based updates. We also show the experimental results in real reverberant rooms.

2 Problem Formulation

Suppose K sound sources are observed by L microphones. Let $\boldsymbol{s}_i$ $(i = 1 \ldots, K)$ and $\boldsymbol{r}_k$ $(k = 1, \ldots, L)$ be the locations of the sound sources and the microphones, respectively. In this paper, we assume that the recording devices have the same sampling frequency but different recording time offsets. Let t_k $(k = 1, \ldots, L)$ be the time when the kth microphone starts recording. The problem here is the estimation of all these parameters $\boldsymbol{s}_i$, $\boldsymbol{r}_k$, and t_k only with observed signals.

Generally, one of the most significant cues for localization of sources and microphones is the TDOA. However, it should be noted that the apparent TDOA include unknown time offsets in this problem. Figure 1 depicts the relationship between the apparent and the true TDOAs. The apparent TDOA between the mth and the nth channels for the ith source can be represented by

$$\tau_{imn} = \frac{||\boldsymbol{s}_i - \boldsymbol{r}_m|| - ||\boldsymbol{s}_i - \boldsymbol{r}_n||}{c} - (t_m - t_n), \tag{1}$$

where c is the speed of sound. Here $||\boldsymbol{x}||$ denotes the Euclidean norm of $\boldsymbol{x}$. In this equation, only τ_{imn} is observable. To estimate unknown variables, the number of the observable variables must exceed the number of the unknown variables. When τ_{imn} are given for all the combination of i,m and n, the necessary condition to solve the problem in 2D case is $(K - 3)(L - 3) \geq 5$, which is derived in the same manner as in [6].

3 Applying Auxiliary Function Method to Parameter Estimation

Let $\boldsymbol{\Theta}$ be the unknown parameter set: $\boldsymbol{\Theta} := \{\boldsymbol{s}_i, \boldsymbol{r}_m, t_m | i = 1, \ldots, K,\ m = 1, \ldots, L\}$.In order to find $\boldsymbol{\Theta}$, the squared error of Eq. (1) is considered as the objective function:

$$J(\boldsymbol{\Theta}) := \frac{1}{c^2 K L^2} \sum_{i=1}^{K} \sum_{m=1}^{L} \sum_{n=1}^{L} \varepsilon_{imn}^2, \tag{2}$$

$$\varepsilon_{imn} := ||\boldsymbol{s}_i - \boldsymbol{r}_m|| - ||\boldsymbol{s}_i - \boldsymbol{r}_n|| - c(\tau_{imn} + t_m - t_n). \tag{3}$$

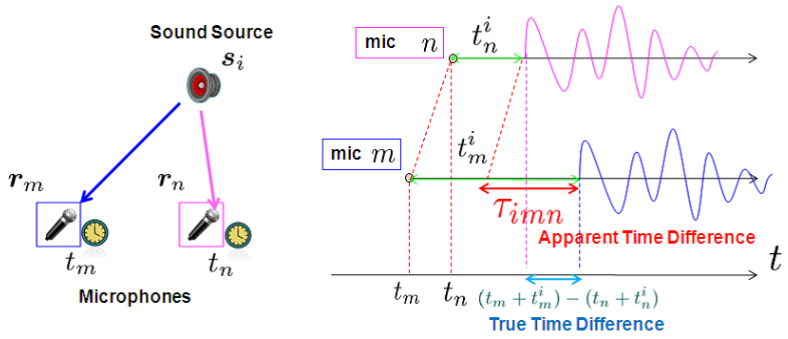

Fig. 1. The relationship between the time offsets t_m, t_n and apparent TDOA τ_{imn}.

The objective function derived in Eq. (2) contains some cross terms of the norms of vectors, which make it infeasible to calculate the parameter set of optimization analytically. Here we introduce *auxiliary function method* [5] for solving this problem. Consider the function $J^+(\boldsymbol{\Theta}, \boldsymbol{\Theta}^+)$ which holds

$$J(\boldsymbol{\Theta}) = \min_{\boldsymbol{\Theta}^+} J^+(\boldsymbol{\Theta}, \boldsymbol{\Theta}^+). \tag{4}$$

Here, $J^+(\boldsymbol{\Theta}, \boldsymbol{\Theta}^+)$ is refereed to as an auxiliary function of an objective function $J(\boldsymbol{\Theta})$ and $\boldsymbol{\Theta}^+$ an *auxiliary variable*. $J^+(\boldsymbol{\Theta}, \boldsymbol{\Theta}^+)$ is non-increasing during following update procedure of variables $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^+$:

$$\boldsymbol{\Theta}_{l+1}^+ = \mathrm{argmin}_{\boldsymbol{\Theta}^+} J^+(\boldsymbol{\Theta}_l, \boldsymbol{\Theta}^+), \tag{5}$$

$$\boldsymbol{\Theta}_{l+1} = \mathrm{argmin}_{\boldsymbol{\Theta}} J^+(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{l+1}^+), \tag{6}$$

where l indicates the index of iterations. A brief proof follows:

1. $J(\boldsymbol{\Theta}_l) = J^+(\boldsymbol{\Theta}_l, \boldsymbol{\Theta}_{l+1}^+)$ is held due to Eq. (4) and Eq. (5),
2. $J^+(\boldsymbol{\Theta}_l, \boldsymbol{\Theta}_{l+1}^+) \geq J^+(\boldsymbol{\Theta}_{l+1}, \boldsymbol{\Theta}_{l+1}^+)$ from Eq. (6),
3. $J^+(\boldsymbol{\Theta}_{l+1}, \boldsymbol{\Theta}_{l+1}^+) \geq J(\boldsymbol{\Theta}_{l+1})$ from Eq. (4),

then,

$$J(\boldsymbol{\Theta}_l) \geq J(\boldsymbol{\Theta}_{l+1}). \tag{7}$$

The auxiliary function method does not require tuning of parameters such as step size required in gradient descent and many other iterative optimization algorithms. However it is not always guaranteed that an adequate auxiliary function can be easily designed.

The derivative of Eq. (2) cannot be calculated. Here we intend to design an auxiliary function whose derivative is calculable instead. We have the auxiliary

function below, which is derived from [6] in detail.

$$\begin{aligned} J_2(\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{e}) = & \frac{2}{c^2 K L^2} \sum_{i,m,n} \{(\boldsymbol{s}_i - \boldsymbol{r}_m - \boldsymbol{e}_{im}\mu^m_{imn})^2 \\ & +(\boldsymbol{s}_i - \boldsymbol{r}_n - \boldsymbol{e}_{in}\mu^n_{imn})^2\}, \\ \text{where } \boldsymbol{\mu} := & \{\mu^m_{imn}, \mu^n_{imn} \ | i = 1, \dots K, \ m, n = 1, \dots L\} \\ \boldsymbol{e} := & \{\boldsymbol{e}_{im} | i = 1, \dots, K, \ m = 1, \dots, L\} \end{aligned} \tag{8}$$

Here we have $\boldsymbol{\mu}, \boldsymbol{e}$ as auxiliary parameters. The whole update rule is written as follows [6]:

$$\varepsilon_{imn} \leftarrow ||\boldsymbol{s}_i - \boldsymbol{r}_m|| - ||\boldsymbol{s}_i - \boldsymbol{r}_n|| - c(\tau_{imn} + t_m - t_n), \tag{9}$$

$$\mu^m_{imn} \leftarrow ||\boldsymbol{s}_i - \boldsymbol{r}_m|| - \frac{1}{2}\varepsilon_{imn}, \tag{10}$$

$$\mu^n_{imn} \leftarrow ||\boldsymbol{s}_i - \boldsymbol{r}_n|| + \frac{1}{2}\varepsilon_{imn}, \tag{11}$$

$$\boldsymbol{e}_{im} \leftarrow (\boldsymbol{s}_i - \boldsymbol{r}_n)/||\boldsymbol{s}_i - \boldsymbol{r}_i||, \tag{12}$$

$$\boldsymbol{s}_i \leftarrow \frac{1}{L^2} \sum_{m=1}^{L} \left(L\boldsymbol{r}_m + \boldsymbol{e}_{im} \sum_{n=1}^{L} \mu^m_{imn} \right), \tag{13}$$

$$\boldsymbol{r}_n \leftarrow \frac{1}{KL} \sum_{i=1}^{K} \left(L\boldsymbol{s}_i - \boldsymbol{e}_{in} \sum_{m=1}^{L} \mu^n_{imn} \right), \tag{14}$$

$$t_n \leftarrow t_n + \frac{1}{cKL} \sum_{i=1}^{K} \left(L||\boldsymbol{s}_i - \boldsymbol{r}_n|| - \sum_{m=1}^{L} \mu^n_{imn} \right). \tag{15}$$

4 Estimation of Time Difference of Arrival Based on Generalized Cross Correlation

In this section we introduce two-stage TDOA analysis to obtain the estimation of the apparent time differences in real room environment, which were assumed to be given in the discussions above. The apparent time differences τ_{imn} are the time differences of the sources in the observation channels, and have to be estimated only with the observed signals. A key to accurate TDOA analysis is the efficient use of the limited number of observation samples, and one approach is frame analysis with many frames. However, we have to use short window to increase the number of the frames, and that disable us to deal with long time differences: Frame-based TDOA analysis can not estimate longer time difference than the frame length. To overcome this problem, the first stage aligns the time differences among the channels roughly in the time domain, and the second stage estimates more detailed TDOA for each source using frame analysis and GCC.

While the detailed TDOA estimation in the second stage is conducted for each of the source-channel pairs, the rough alignment in the first stage is applied

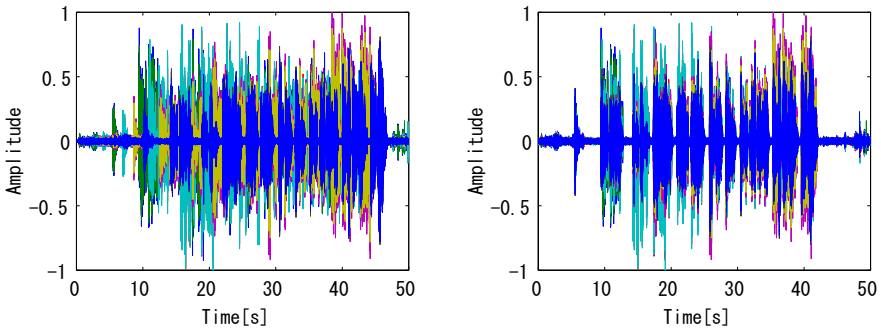

Fig. 2. Asynchronously recorded signals (left) and the same signals after preprocessing, rough alignment (right).

for each channel without distinguishing the sources. Here we justify this strategy. As we discussed, each apparent time difference is the sum of the true TDOA and the recording time offset. While the length of the former is limited by the size of the room, the latter can have any length and is often much longer than the former. To cancel such large time differences caused by the different starting times of recording, we obtain average time difference of all the sources for each channel. The rough alignment is obtained by the following procedure. First, we select one channel as a reference channel. Second, we calculate cross correlations between the reference and each rest of the channels and detect its peak. Finally, we can roughly synchronize the channels by shifting the other channels keeping the reference fixed so that the cross correlations have peaks at the zeroth sample. An example of the rough alignment is shown in Fig. 2. It can be seen that the rough alignment successfully cancels the average time difference of all the sources for each channel.

For the TDOA analysis in the second stage, we use GCC method, which is known to be robust against interference. Note that we assume the time durations when only one source is active for each source is given (referred to as double-talk detection), and we analyze the signal in those time durations. GCC gives the maximum likelihood estimate of the TDOA under an observation model, which assumes that only the direct wave reaches the microphones from the desired source, and the desired source and the observation noises are uncorrelated Gaussian. As a result, the estimated TDOA is obtained by detecting a peak of the GCC function, which is a filtered version of the cross correlation function between the analyzed signals. The filter of GCC enhances the peak of the cross correlation by effective whitening and noise suppression.

Although we omit the discussion of probabilistic modeling of the observed signal here, we review the filtering of GCC and its property. Let $X(\omega)$ and $Y(\omega)$ be the Fourier transform of the two observed signals $x(t)$ and $y(t)$ to be analyzed.

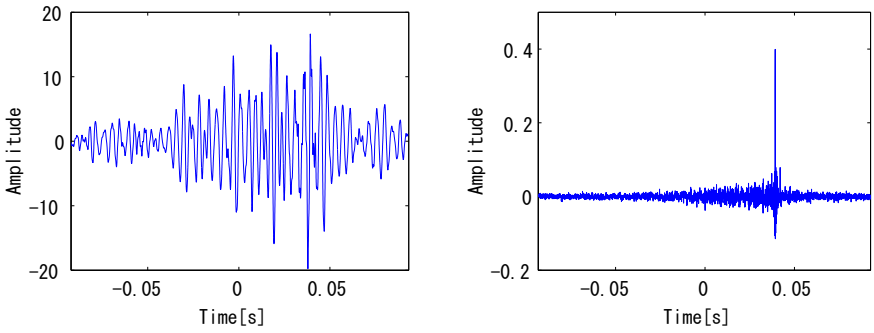

Fig. 3. Conventional cross correlation(left) and generalized cross correlation(right).

According to [7], the filter $U(\omega)$ in the frequency domain is designed as follows:

$$U(\omega) = \frac{1}{|E\left[X(\omega)Y^*(\omega)\right]|} \cdot \frac{|\gamma(\omega)|^2}{1-|\gamma(\omega)|^2}, \tag{16}$$

where

$$\gamma(\omega) = \frac{E\left[X(\omega)Y^*(\omega)\right]}{\sqrt{|E\left[X(\omega)\right]|^2}\sqrt{|E\left[Y(\omega)\right]|^2}} \tag{17}$$

is the coherence of each frequency. According to the coherence, reliable frequency components in the frequencies with low SNR is weighted lightly, and the weighting results as noise suppression. Also, the normalization of the amplitude whitens the cross correlation and enhances its peak.

Figure 3 shows examples of conventional cross correlation and GCC of observed signals after the preprocessing of the rough alignment in the first step. Successful enhancement of the peak by GCC can be seen.

5 Experimental Results

5.1 Experimental Condition

We executed an experiment in real room environment to evaluate the performance of blind alignment. We placed seven loudspeakers and four stereo IC recorders of the same model number (SANYO ICR-PS603EM) in the identical plane in the room with $6\times7\times2.7\mathrm{m}^3$ (See Fig. 4). As source signals, recorded speech signals whose bit rate is 44100 Hz are used. The reverberation time of the room is approximately 300 ms. There were no temporal overlaps from different sound sources in every observed signal and we manually gave the segmentation in the observed signals corresponding to each sound source. The relative error of sampling rate among devices were up to 2.6×10^{-6}. We set flame length to

Fig. 4. IC recorder used in the experiment (left) and The recording room (right).

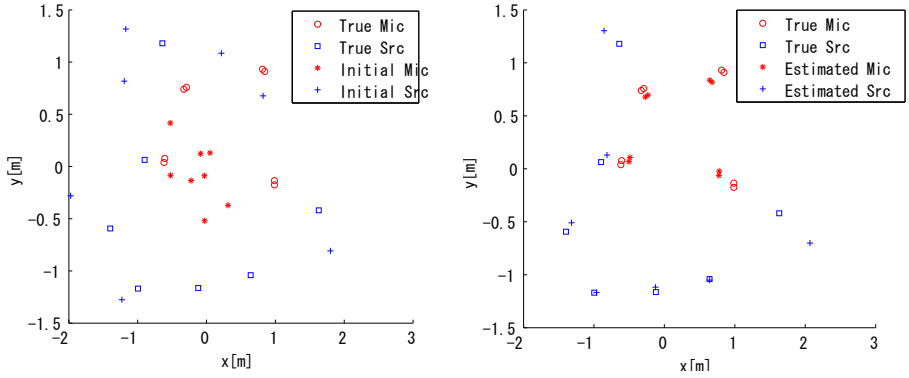

Fig. 5. The initial position (left) and the estimated position (right).

8192 samples. The frame analysis with half-overlap was performed to execute the GCC method. Since the objective function in our proposed method has many local minima, initialization of the parameters is an important problem. However, appropriate initialization can be easily designed with some rough assumption on the layout. We gave a true but rough assumption that the microphones are surrounded by the sources. Specifically, the initial $\boldsymbol{r}_m$ were randomly given in a circle whose radius was 1 m. The initial $\boldsymbol{s}_i$ were also randomly given by the rule that the norm of each $\boldsymbol{s}_i$ be from 1 m up to 1.5 m. The estimation of the positions were performed in a two-dimensional plane.

5.2 Experimental Result of a Blind Alignment Task

Figure 5 shows the estimation result of positions of microphones and sources. After 50000 iterations, the objective function converged toward a satisfactorily small value. The microphone pairs of indices $\{1, 2\}$, $\{3, 4\}$, $\{5, 6\}$ and $\{7, 8\}$ are the stereo microphone pairs mounted on the same recording devices. Thus the

Table 1. The estimated time offsets t_m[ms]

m	1	2	3	4	5	6	7	8
t_m	0.052	0.075	-18.33	-18.24	23.27	23.29	20.74	20.74

corresponding time offsets of each pair should be identical, e.g., $t_m = t_n$ for $m = 1, 3, 5, 7$ and $n = m + 1$. In table 1, it is confirmed that the time offsets of each pairs are estimated to be very close. The average error of position was 15.33 cm. It is ascertained that the proposed method works properly for the blind alignment task in real environment.

6 Conclusion

We presented an auxiliary-function based technique for estimating the locations of microphones and sound sources and time offsets of each microphones by observed signals alone. For the estimation of the apparant time differences used in the auxiliary function method, we used two-stage TDOA analysis; the first stage for rough alignment of the channels with large time differences, and the second stage for the estimation of each apparant time difference using frame-based GCC method. The blind alignment experiment in real reverberant room environment was ascertained to be performed well. To attack the remaining problem that the our current framework reqires the time durations when only a single source is active for each source, our ongoing work is development of a new framework to perform both blind alignment and source separation simultaneously.

References

1. R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio processing," Proc. ICASSP, pp. 840–843, 2003.
2. P. Aarabi, "The fusion of distributed microphone arrays for sound localization," EURASIP Journal of Applied Signal Processing, vol. 2003, no. 4, pp. 338–347, 2003.
3. A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," Proc. Interspeech, pp. 2337–2340, 2005.
4. K. Kobayashi, K. Furuya, A. Kataoka, "A blind source localization by using freely positioned microphones," Trans. IEICE, vol. J86-A, no.6, pp. 619–627, 2003. (in Japanese)
5. H. Kameoka, N. Ono, and S. Sagayama, "Auxiliary Functional Approach to Parameter Estimation of Constrained Sinusoidal Model for Monaural Speech Separation," Proc. ICASSP, pp. 29–32, Mar. 2008.
6. N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array," Proc. WASPAA, pp.161–164, Oct. 2009.
7. C. H. Knapp and G. C. Cartar, "The Generalized Correlation Method for Estimation of Time Delay," IEEE Trans. ASSP, vol. 24, no. 4, pp. 320–327, Aug. 1976.