2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Synchronization of Microphones Based on Rank Minimization of Warped Spectrum for Asynchronous Distributed Recording

Katsutoshi Itoyama[1] and Kazuhiro Nakadai[1,2]

***Abstract*— This paper describes a new method for synchronizing microphones based on spectral warping in an asynchronous microphone array. In an audio signal observed by an asynchronous microphone array, two factors are involved: the time lag caused by a mismatch of the sampling rate and offset between microphones, and the modulation caused by differences in spatial transfer function between the sound source and each microphone. A spectrum warping matrix representing a resampling effect in the frequency domain is formulated and an observation model of audio (spectrum) mixture in an asynchronous microphone array is constructed. The proposed synchronization method uses an iterative optimization algorithm based on gradient descent of a new objective function. The function is formulated as a logarithmic determinant of a spectrum correlation matrix that is derived from relaxation of a rank minimization problem. Experimental results showed that the proposed method effectively estimates modulated sampling rate and that the proposed method outperforms an existing synchronization method.**

I. Introduction

A microphone array is an audio recording device comprising multiple microphones, A/D converters, and an oscillator that provides a clock to synchronize the converters. In microphone array based signal processing, the recorded signals must be synchronized at the sample level in all channels. Microphone array signal processing can not be performed in signals that have been recorded by independent audio devices because the recorded signals are not synchronized at the sample level, even by devices of the same model and date of manufacture. Since a steering vector and a spatial correlation matrix determined by the relative position of the microphones used in many microphone array signal processing techniques are based on the premise that all the microphones are synchronized, these techniques can not be applied to asynchronous multichannel recordings even if the microphone positions are known.

Furthermore, microphones do not always operate at the exact sampling rate set in advance. Under a harsh environment, outdoors or mounted in a rescue robot, for example, the sampling rate is affected by internal factors such as power fluctuations due to motor driving and wireless communication and external factors such as temperature, humidity, and sunlight. To avoid such effects, a microphone may incorporate an oscillator that is hardy against ambient temperature such as an oven-controlled crystal oscillator (OCXO) and an atomic clock, and a large capacitor and independent power supply. However, in using an existing converter, some remodeling is indispensable but it could be unrealistic.

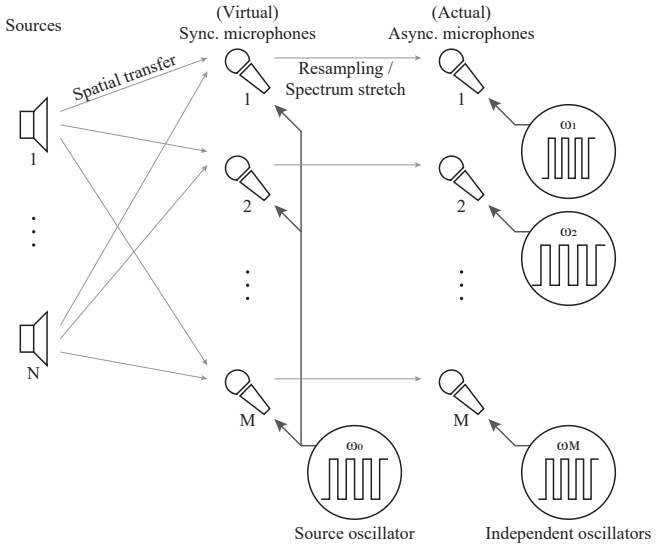

Fig. 1. Generative process of audio signals in an asynchronous microphone array. Audio signals or spectra from different sources are modulated due to their spatial properties, then resampled or warped according to differences in sampling rates and offsets.

This paper describes a new method for synchronization of asynchronously recorded audio signals that are obtained by multiple independent microphones. Initially audio sources provide source signals. The signal reaches each microphone under the influence of the spatial transfer function defined by a steering vector. Note that the steering vector is defined at a reference sampling rate. The signal is then resampled reflecting difference in sampling rate between the source and the microphones. If the same audio signal is recorded at different sampling rates by different microphones, the recorded signal can be regarded as resampled at a certain sampling rate (e.g., 16 kHz.) on the basis of sampling theorem. Since resampling is defined as a linear transformation equivalent to a warping (expansion and contraction) operation, the spectrum obtained by a Fourier transform of the signal is regarded as a warped form of the original spectrum. We developed a generative model of a warped spectrum from the different sampling rates of each microphone. A new objective function is defined as a logarithmic determinant of the observation spectrum whose warping is compensated.

*This work was supported by JSPS KAKENHI Grant No. 16H02884, 17K00365, and 19K12017.

[1]Katsutoshi Itoyama is with Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, 152-8552 Tokyo, Japan itoyama@ra.sc.e.titech.ac.jp

[2]Kazuhiro Nakadai is with Honda Research Institute Japan, Co., Ltd., 351-0188 Saitama, Japan nakadai@jp.honda-ri.com

This objective function is derived from a relaxation of the rank minimization of that spectrum. The optimal sampling rate is estimated by a gradient descent method based on the gradient of the objective function. In summary, we claim the following contributions of this paper:

1) we demonstrated the approximate equivalence of resampling and spectrum warping, and defined an observation model of audio (spectrum) mixture in an asynchronous microphone array;
2) we have constructed a synchronization method using an iterative optimization algorithm based on matrix rank minimization and its relaxation to logarithmic determinant minimization; and
3) we conducted numerical evaluations to show the effectiveness of the proposed synchronization method.

II. Related Work

There are three main challenges in studying asynchronous distributed microphone arrays or ad hoc microphone arrays [1] are (1) position estimation of the microphones, (2) localization and position estimation of the sound sources, and (3) synchronization between the channels. Challenge (1) is considered as a microphone array calibration problem, and many solutions have been proposed including differences in observed sound energy [2], [3], local array pattern calibration and network geometry estimation [4], clustering of beam forming [5], a bilinear form in the sources and microphones about time of arrival delays [6], the low-rank property of the array distance matrix [7]. Several techniques have been reported to solve the array calibration problem by combining it with the synchronization problem: estimating the sampling rate of each microphone based on a simultaneous localization and mapping (SLAM) approach [8] and estimating the offset [9]. Regarding challenge (2), several methods using time difference of arrival (TDOA) for 2D [10] and 3D [11] localization have been reported. In this way, (1) and (2) may be amalgamated and, solved simultaneously [2], [3] by a SLAM-based approach [12], [13]. Use of independent A/D converters by distributed microphones results in the third challgenge, for which several solutions have been proposed including compensation of sampling rate mismatch [14] and measurement of TDOA [15]. In contrast, use of the wireless acoustic sensor networks allows synchronization by wireless communication among microphones (sensors) [16]–[20].

Most of the previous studies suggest that calibration may be achieved by clapping hands around the array while keeping the environment quiet. However, in a real environment where extraneous sounds cannot be controlled such a calibration process is impractical. In the present study we attempted to realize synchronization of microphones to facilitate array signal processing for source localization and separation of multiple simultaneous sound sources without such a calibration process.

III. Signal Model

This section describes the generative process in which audio signals emitted by multiple sound sources are observed as mixed sounds by an asynchronous microphone array.

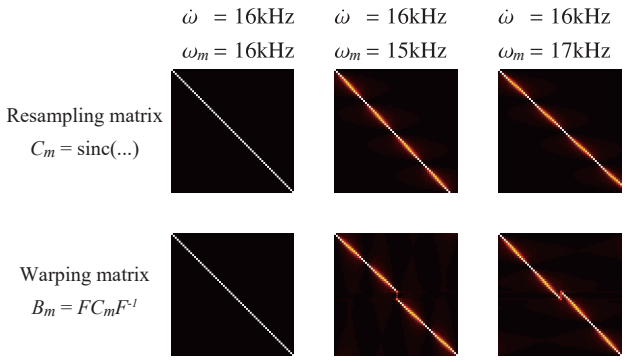

Fig. 2. Signal resampling and spectral warping matrices.

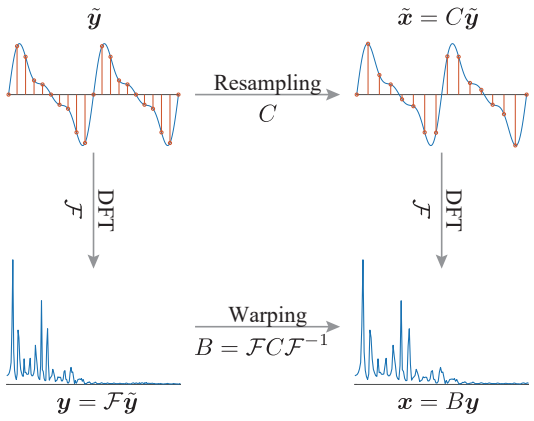

Fig. 3. Relationship of signal resampling and spectrum warping.

A. Signal Resampling and Spectrum Warping

Let $(\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_{L-1})$ be a digital audio signal with sampling rate $\dot{\omega}$. The signal may be an entire observed signal, or a frame cut out by the window function. The sampling times are $0, 1/\dot{\omega}, \ldots, (L-1)/\dot{\omega}$, respectively. An analog audio signal $\tilde{z}(t)$ obtained by applying D/A conversion to this digital signal is represented by

$$\tilde{z}(t) \approx \sum_{l'=0}^{L-1} \mathrm{sinc}(\dot{\omega}t - l')\, \tilde{y}_{l'}. \tag{1}$$

In order to perform a complete sinc interpolation so that both sides of the above equation are equal, the range of the summation must be infinite, i.e., $-\infty$ to ∞. Actually in the above equation, both sides do not exactly match because the range of summation is finite. Therefore they are represented as an approximation relationship. Let $(\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_{L-1})$ be a digital audio signal with sampling rate ω and offset τ (time offset of the first sample point) which is obtained by resampling $\tilde{z}(t)$ at sampling times $\tau, 1/\omega + \tau, \ldots, (L-1)/\omega + \tau$. The relationship between $\tilde{y}_l$ and $\tilde{x}_l$ is represented as

$$\tilde{x}_l = \tilde{z}(l/\dot{\omega} + \tau) \approx \sum_{l'=0}^{L-1} \mathrm{sinc}\left(\frac{\dot{\omega}}{\omega} l + \dot{\omega}\tau - l'\right) \tilde{y}_{l'} \tag{2}$$

and it is summarized in a vectorized form as

$$\tilde{\boldsymbol{x}} \approx C\tilde{\boldsymbol{y}} \tag{3}$$

where

$$\tilde{\boldsymbol{x}} = (\tilde{x}_0, \ldots, \tilde{x}_{L-1})^{\mathsf{T}}, \tag{4}$$

$$\tilde{\boldsymbol{y}} = (\tilde{y}_0, \ldots, \tilde{y}_{L-1})^{\mathsf{T}}, \text{and} \tag{5}$$

$$C = \left(\operatorname{sinc}\left(\frac{\dot{\omega}}{\omega}l + \dot{\omega}\tau - l'\right)\right)_{\substack{l=0,\ldots,L-1 \\ l'=0,\ldots,L-1}}. \tag{6}$$

We call this coefficient matrix C as the *resampling matrix*.

On the other hand, $\tilde{z}(t)$ can be obtained by applying D/A conversion to $\tilde{\boldsymbol{x}}$ as

$$\tilde{z}(t) \approx \sum_{l'=0}^{L-1} \operatorname{sinc}(\omega t - l' - \omega\tau)\, \tilde{x}_{l'}. \tag{7}$$

$\tilde{\boldsymbol{y}}$ will be obtained by resampling this $\tilde{z}(t)$ with sampling rate $\dot{\omega}$ as

$$\tilde{y}_l = \tilde{z}(l/\dot{\omega}) \approx \sum_{l'=0}^{L-1} \operatorname{sinc}\left(\frac{\omega}{\dot{\omega}}l - l' - \omega\tau\right) \tilde{x}_{l'}. \tag{8}$$

This relationship also is summarized in a vectorized form as

$$\tilde{\boldsymbol{y}} \approx C^*\tilde{\boldsymbol{x}} \tag{9}$$

where

$$C^* = \left(\operatorname{sinc}\left(\frac{\omega}{\dot{\omega}}l - l' - \dot{\omega}\tau\right)\right)_{\substack{l=0,\ldots,L-1 \\ l'=0,\ldots,L-1}}. \tag{10}$$

We call this coefficient matrix C^* as the *resampling compensation matrix* because C and C^* are approximately inverse matrices.

Let $\boldsymbol{y} = (y_0, \ldots, y_{L-1})^{\mathsf{T}}$ and $\boldsymbol{x} = (x_0, \ldots, x_{L-1})^{\mathsf{T}}$ be the spectra of the signal $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{x}}$ obtained by performing the discrete Fourier transform (DFT) respectively,

$$\boldsymbol{y} = \mathcal{F}\tilde{\boldsymbol{y}} \text{ and } \boldsymbol{x} = \mathcal{F}\tilde{\boldsymbol{x}}, \tag{11}$$

where $\mathcal{F}$ is the coefficient matrix of the DFT,

$$\mathcal{F} = \left(\frac{e^{-2\pi jkl/L}}{\sqrt{L}}\right)_{\substack{k=0,\ldots,L-1 \\ l=0,\ldots,L-1}}. \tag{12}$$

Equations (3), (9), and (11) give the relationship between these spectra as

$$\boldsymbol{x} \approx B\boldsymbol{y} \text{ s.t. } B = \mathcal{F}C\mathcal{F}^{-1}. \tag{13}$$

We name this B as the *spectrum warping matrix*. Similar to C^*, we name $B^* = \mathcal{F}C^*\mathcal{F}^{-1}$ that is an approximately inverse matrix of B as the *spectrum warping compensation matrix*.

B. Observation Model of Spectrum Mixture

Let s_{nft} be a source spectrum of the n-th source at the (f,t)-th frequency-time slot. Let $\boldsymbol{y}_{ft}^{(m)} = (y_{1ft}, \ldots, y_{Mft})^{\mathsf{T}}$ be the virtual synchronous spectrum that is recorded by the virtual synchronous microphones while being modulated by the spatial transfer function defined by the steering vector $\boldsymbol{v}_{nf} = (v_{nf1}, \ldots, v_{nfM})^{\mathsf{T}}$:

$$\boldsymbol{y}_{ft}^{(m)} = \sum_{n=1}^{N} \boldsymbol{v}_{nf} s_{nft}. \tag{14}$$

Then the synchronous spectrum $\boldsymbol{y}_{ft}^{(m)}$ is modulated by the spectrum warping matrix B_m and it is observed by the asynchronous microphone as the asynchronous spectrum $\boldsymbol{x}_{mt}^{(f)}$,

$$\boldsymbol{x}_{mt}^{(f)} = B_m \boldsymbol{y}_{mt}^{(f)} \tag{15}$$

$$\text{s.t. } \boldsymbol{y}_{mt}^{(f)} = (y_{m1t}, \ldots, y_{mFt})^{\mathsf{T}} \text{ and} \tag{16}$$

$$\boldsymbol{x}_{mt}^{(f)} = (x_{m1t}, \ldots, x_{mFt})^{\mathsf{T}}. \tag{17}$$

Once the asynchronous spectrum $\boldsymbol{x}_{mt}^{(f)}$ was observed, the estimation of the synchronous spectrum $\boldsymbol{\psi}_{mt}^{(f)} = (\psi_{m1t}, \ldots, \psi_{mFt})^{\mathsf{T}}$ is obtained by applying the spectrum warping compensation matrix B_m^* as

$$\boldsymbol{\psi}_{mt}^{(f)} = B_m^* \boldsymbol{x}_{mt}^{(f)}. \tag{18}$$

Since the spectrum warping compensation matrix B_m^* is determined by the sampling rate ω_m and the offset τ_m, a good estimation of the synchronous spectrum $\boldsymbol{\psi}_{mt}^{(f)}$ will be obtained if the appropriate ω_m and τ_m can be estimated from the observed spectrum $\boldsymbol{x}_{mt}^{(f)}$.

IV. Synchronization Algorithm

This section describes a new algorithm for synchronization based on the signal model written in the previous section. Let N, M, T, and F be the number of sources, microphones, time frames, and frequency bins, respectively. We suppose that $N < M$, $N < T$, and $M < F$. In the field of microphone array signal processing, because it is common to use more microphones than the number of sound sources, $N < M$ holds well. If the audio signal of a sufficiently long duration is used, the number of frames T that is larger than N can be easily obtained, so $N < T$ holds well. $M < F$ also holds well by extracting frames using a sufficiently long window function. Furthermore, we assume that the source spectra are mutually independent.

Since $\boldsymbol{y}_{ft}^{(m)}$ is a linear combination of N steering vectors $\boldsymbol{v}_{nf}$ as shown in equation (14), an $M \times T$ matrix $\left(\boldsymbol{y}_{f1}^{(m)}, \ldots, \boldsymbol{y}_{fT}^{(m)}\right)$ has a rank at most N. From equations (15) and (18), the following relation is derived as

$$\boldsymbol{\psi}_{mt}^{(f)} = B_m^* B_m \boldsymbol{y}_{mt}^{(f)} \approx \boldsymbol{y}_{mt}^{(f)}. \tag{19}$$

By changing the index grouping direction, an equivalent relation is derived as

$$\boldsymbol{\psi}_{ft}^{(m)} \approx \boldsymbol{y}_{ft}^{(m)}. \tag{20}$$

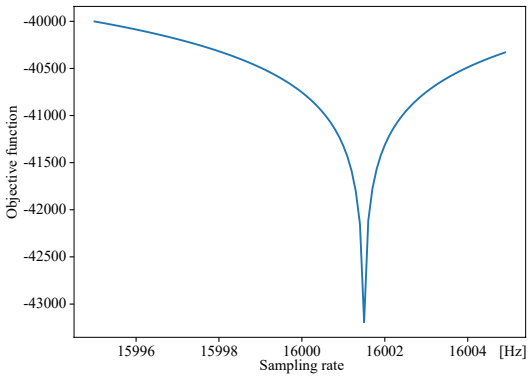

Fig. 4. Objective function $J'(\boldsymbol{\omega}, \boldsymbol{\tau})$ with respect to ω_m when the actual sampling rate is 16001.5 Hz.

Therefore, the rank of a matrix $\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right)$ is approximately the same as that of $\left(\boldsymbol{y}_{f1}^{(m)}, \ldots, \boldsymbol{y}_{fT}^{(m)}\right)$:

$$\operatorname{rank}\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right) \approx \operatorname{rank}\left(\boldsymbol{y}_{f1}^{(m)}, \ldots, \boldsymbol{y}_{fT}^{(m)}\right), \quad (21)$$

if the appropriate parameters ω_m and τ_m were used.

On the other hand, when ω_m or τ_m is an unsuitable value, the rank of $\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right)$ should be larger than the above matrices. B_m and B_m^* are no longer approximate inverse matrices, and non-zero values appear in the off-diagonal elements of their product $B_m^* B_m$. These non-zero values add multiple $\boldsymbol{y}_{f't}^{(m)}$ $(f' \neq f)$ to the components of $\boldsymbol{\psi}_{ft}^{(m)}$ in addition to $\boldsymbol{y}_{ft}^{(m)}$. As a result, $\boldsymbol{\psi}_{ft}^{(m)}$ will be represented by a linear combination of more elements than n, therefore the rank increases.

The relationship of the rank of the matrices is summarized as

$$\operatorname{rank}\left(\boldsymbol{y}_{f1}^{(m)}, \ldots, \boldsymbol{y}_{fT}^{(m)}\right) \lessapprox \operatorname{rank}\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right). \quad (22)$$

If appropriate parameters ω_m and τ_m are used, $\operatorname{rank}\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right)$ should be approximately equal to $\operatorname{rank}\left(\boldsymbol{y}_{f1}^{(m)}, \ldots, \boldsymbol{y}_{fT}^{(m)}\right)$, and the rank will increases as the parameters become incorrect. Thus $\operatorname{rank}\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right)$ can be used as an objective function $J(\boldsymbol{\omega}, \boldsymbol{\tau})$ for the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$,

$$J(\boldsymbol{\omega}, \boldsymbol{\tau}) = \sum_{f=1}^{F} \operatorname{rank}\left(\boldsymbol{\psi}_{f1}^{(m)}, \ldots, \boldsymbol{\psi}_{fT}^{(m)}\right). \quad (23)$$

The appropriate parameters $\hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\tau}}$ can be given by minimizing the objective function $J(\boldsymbol{\omega}, \boldsymbol{\tau})$ as

$$\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\tau}} = \underset{\boldsymbol{\omega}, \boldsymbol{\tau}}{\operatorname{argmin}} J(\boldsymbol{\omega}, \boldsymbol{\tau}). \quad (24)$$

Since rank minimization is known to be NP-hard, this can be relaxed by minimizing the logarithmic determinant of the

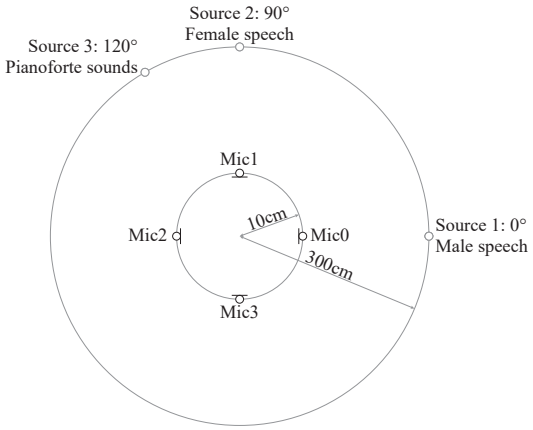

Fig. 5. Evaluation environment.

TABLE I
ESTIMATION ERRORS (RMSE) OF SAMPLING RATE FOR DIFFERENT MODIFICATIONS.

Modifications [Hz]	RMSE [Hz]	RMSE [ppm]
± 0.1	$< 10^{-4}$	$< 10^{-3}$
± 0.2	$< 10^{-4}$	$< 10^{-3}$
± 0.5	0.0001	0.006
± 1.0	0.0002	0.012
± 1.5	0.0005	0.031
± 2.0	0.0009	0.056
± 5.0	0.0071	0.44

correlation matrix [21]:

$$J'(\boldsymbol{\omega}, \boldsymbol{\tau}) = \sum_{f=1}^{F} \log\det \Psi_f \quad (25)$$

$$\text{and} \quad \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\tau}} = \underset{\boldsymbol{\omega}, \boldsymbol{\tau}}{\operatorname{argmin}} J'(\boldsymbol{\omega}, \boldsymbol{\tau}) \quad (26)$$

where $\Psi_f = \left(\sum_{t=1}^{T} \psi_{m_1 f t} \psi_{m_2 f t}{}^*\right)_{\substack{m_1=1,\ldots,M \\ m_2=1,\ldots,M}}$. The rank of a matrix is equal to the number of non-zero eigenvalues, and the log-determinant of a matrix is equal to the summation of the logarithmic eigenvalues. Finding parameters that minimizes the log-determinant can be interpreted as trying to increase the number of zero eigenvalues of the matrix, i.e., deficienting rank of the matrix.

The optimal parameters ω_m and τ_m are obtained by minimizing the log-det objective function by assuming that the objective function is convex. The objective function is minimized by the gradient descent method using the numerically determined gradient

$$(\boldsymbol{\omega}, \boldsymbol{\tau})^{\mathsf{T}} \leftarrow (\boldsymbol{\omega}, \boldsymbol{\tau})^{\mathsf{T}} - \alpha \nabla J'(\boldsymbol{\omega}, \boldsymbol{\tau}) \quad (27)$$

where α is a step size parameter. An illustration of the objective function is shown in Figure 4.

V. EVALUATION

We conducted three numerical evaluations of the proposed method by simulation. In the first experiment, the estimation accuracy of the modified sampling rate was evaluated. In the second experiment, the estimation accuracy for the different number of sound sources was evaluated. In the third

TABLE II

ESTIMATION ERRORS (RMSE) OF SAMPLING RATE FOR DIFFERENT NUMBER OF SOUND SOURCES.

# of sources	Source direction	RMSE [Hz]	RMSE [ppm]
1	[0°]	0.0005	0.031
2	[0°, 90°]	0.0005	0.031
3	[0°, 90°, 120°]	0.0008	0.050

TABLE III

ESTIMATION ERRORS (RMSE) OF SAMPLING RATE FOR DIFFERENT SIGNAL LENGTH.

Signal length [s]	RMSE [Hz]	RMSE [ppm]
0.1	0.0014	0.088
0.2	0.0015	0.094
0.5	0.0010	0.063
1.0	0.0006	0.038
3.0	0.0005	0.031
5.0	0.0005	0.031

experiment, the estimation accuracy for the different signal length was evaluated. As the first step of the research, this paper only evaluated the sampling rate estimation. The offset was set randomly and assumed to be known. The root mean squared errors (RMSEs) of the estimated sampling rate and its parts per million (ppm, 10^{-6}) representation were used for evaluation criteria.

The experimental environment is shown in Fig. 5. A circular 4-channel microphone array with a radius of 10 cm was used for recording. One microphone is randomly chosen from the four microphones then the sampling rate of the chosen microphone is slightly modified from the reference sampling rate of $16\,\mathrm{kHz}$. The sampling rate of the remaining three microphones remained at $16\,\mathrm{kHz}$. The amount of change is chosen from $\pm 0.1, 0.2, 0.5, 1.0, 1.5, 2.0, 5.0\,\mathrm{Hz}$. Those correspond to the change of $\pm 6.25, 12.5, 31.25, 62.5, 93.75, 125, 312.5\,\mathrm{ppm}$, which are realistic as practical bias [14]. Audio signals are recorded, i.e., created by numerical calculation, using such an asynchronous microphone array. The signal was recorded once in a $16\,\mathrm{kHz}$ synchronized manner, then each channel of the recorded signal was de-synchronized (resampled) based on the actual sampling rate and offset. The sampling rate of each microphone is estimated from the recorded signals, and the accuracy of sampling rate estimation is evaluated.

The number of sources are chosen from one to three. The first, the second, and the third sources were oriented at $0°$, $90°$, and $120°$ respevtively. Each sound source was set at a distance of $300\,\mathrm{cm}$ from the center of the microphone array. When the number of sources was one, two, and three, only the first source, the first and the second sources, and all of the three sources were used respectively. The first and second sources were male and female speech signals chosen from ATR Japanese speech database [22]. The third source was a pianoforte sound chosen from RWC Music Database: Instrument sounds [23]. The steering vector in each direction was generated based on geometric calculations. The length of the STFT frames and its shift were 512 and 256 samples. The step size for gradient descent was set to 10^{-4}.

A. Results

The RMSEs of the estimation of sampling rate for different modification of sampling rate are shown in Table I. The number of sound sources was fixed at two and the signal length was fixed at five seconds in this experiment. The RMSEs are less than $10^{-2}\,\mathrm{Hz}$ and less than $0.5\,\mathrm{ppm}$ for any modifications. Therefore this result shows high estimation accuracy of the proposed method. When the modification is ± 0.1 and $\pm 0.2\,\mathrm{Hz}$, the RMESs are less than $10^{-4}\,\mathrm{Hz}$, which is considered to contain only the error derived from the numerical calculation error of the gradient. When modification changed from $\pm 2.0\,\mathrm{Hz}$ to $\pm 5.0\,\mathrm{Hz}$, i.e., 2.5 times, the RMSE increased about 8 times. This result implies that some improvement of the proposed method is necessary to achieve a robust estimation for modifications exceeding $\pm 5.0\,\mathrm{Hz}$.

The RMSEs for different number of sources are shown in Table II. The modification was fixed at $\pm 1.5\,\mathrm{Hz}$ and the signal length was fixed at five seconds in this experiment. The estimation accuracy did not change when the number of sources was one and two, and the RMSE increased 1.6 times when the number of sources was three. This result suggests that the estimation accuracy degrades as the number of sources increases. However, it is thought that the adverse effect of this degradation is small in a realistic situation such as meeting. This is because it is rare that three or more people always speak at the same time and it is common that only one person speaks in many sections.

The RMSEs for different signal lengths are shown in Table III. The modification and the number of sources were fixed at $\pm 1.5\,\mathrm{Hz}$ and two respectively. The RMSEs are less than $0.0015\,\mathrm{Hz}$ and less than $0.1\,\mathrm{ppm}$ for any signal lengths. This result implies that one second of the signal length is sufficient long for the proposed method because the RMSEs are almost the same when the signal length is between one to five seconds. Even in a dynamic environment where the steering vector fluctuates, it is considered that the proposed method can be modified into the online version by introducing blockwise processing. For example, by setting the block length to five seconds, the environment in that block can be considered sufficiently static. When the signal length is $0.1\,\mathrm{s}$, the number of frames T is 5, so the proposed method achieves robust estimation accuracy for a small number of frames. According to the method of Miyabe *et al.* [14], they have been reported that the RMSEs are 2.2 and $1.4\,\mathrm{ppm}$ when the signal length is three and five seconds respectively. Although the experimental environment and conditions are not the same, the RMSE of the proposed method in each case is $0.031\,\mathrm{ppm}$ which is better than an existing method.

VI. CONCLUSIONS

This paper describes a new method for synchronizing microphones from asynchronously recorded audio signals obtained by an asynchronous microphone array. We demonstrated two facts: both resampling and spectrum warping are represented as an equivalent linear transform, and a linear transform which compensates the difference of the sampling rate and offset can be defined. An observation

model of spectrum mixture recorded by an asynchronous microphone array using the spectrum warping matrix is constructed. A logarithmic determinant of the correlation matrix of the compensated observation spectrum is derived as a new objective function to relax a rank minimization problem. An iterative estimation algorithm based on gradient descent efficiently estimated the sampling rate through the numerical evaluations. In future work, we will incorporate an accomplished spectrum source model such as low-rank spectral model and a neural-network based model as reported in [24], [25], and extend the model to include sound source separation.

REFERENCES

[1] A. Bertrand, S. Doclo, S. Gannot, N. Ono, and T. Waterschoot, "Special issue on wireless acoustic sensor networks and ad hoc microphone arrays," *Signal Processing*, vol. 107, pp. 1–3, 2015.

[2] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," in *ICASSP*, vol. 2, 2007, pp. 761–764.

[3] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *WASPAA*, 2007, pp. 22–25.

[4] M. Hennecke, T. Plötz, G. A. Fink, J. Schmalenströer, and R. Häb-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *SSP*, 2009, pp. 257–260.

[5] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 661–676, 2011.

[6] M. Crocco, A. D. Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, 2012.

[7] M. J. Taghizadeh, R. Parhizkar, P. N. Garner, H. Bourlard, and A. Asaei, "Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees," *Signal Processing*, vol. 107, pp. 123–140, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168414003508

[8] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *IROS*, 2011, pp. 524–529.

[9] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama, "Blind estimation of locations and time offsets for distributed recording devices," in *LVA/ICA*, 2010, pp. 57–64.

[10] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 2, pp. 439–443, 2013.

[11] ——, "Distributed 3D source localization from 2D DOA measurements using multiple linear arrays," *Wirel. Commun. Mob. Com.*, vol. 2017, p. 11, 2017.

[12] D. Su, T. Vidal-Calleja, and J. V. Miro, "Simultaneous asynchronous microphone array calibration and sound source localisation," in *IROS*, 2015, pp. 5561–5567.

[13] K. Sekiguchi, Y. Bando, K. Itoyama, , and K. Yoshii, "Layout optimization of cooperative distributed microphone arrays based on estimation of source separation performance," *J. Robot. Mechatron.*, vol. 29, no. 1, pp. 83–93, 2017.

[14] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," in *ICASSP*, 2013, pp. 674–678.

[15] F. Jiang, Y. Kuang, and K. Åström, "Time delay estimation for TDOA self-calibration using truncated nuclear norm regularization," in *ICASSP*, 2013, pp. 3885–3889.

[16] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 25, no. 3, pp. 674–686, 2017.

[17] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *SCVT*, 2011, pp. 1–6.

[18] Y. Chisaki, D. Murakami, and T. Usagaway, "Network-based multi-channel signal processing using the precision time protocol," in *APSIPA*, 2012, pp. 1–6.

[19] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *ICASSP*, vol. 4, 2003, pp. 840–843.

[20] J. Schmalenstroeer and R. Haeb-Umbach, "Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model," in *EUSIPCO*, 2013, pp. 1–5.

[21] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," in *Proceedings of the 2003 American Control Conference, 2003.*, vol. 3, June 2003, pp. 2156–2162 vol.3.

[22] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[23] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 2003, pp. 229–230.

[24] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel audio source separation based on integrated source and spatial models," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 26, no. 4, pp. 831–846, 2018.

[25] J. J. Carabias-Orti, J. Nikunen, T. Virtanen, and P. Vera-Candeas, "Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 26, no. 9, pp. 1512–1527, 2018.