Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Method for Updating Microphone Configuration in Audio Super-Resolution

Ryouichi Nishimura

Resilient ICT Research Center
National Institute of Information and Communications Technology
2-1-3 Katarahi, Aoba-ku, Sendai, 980-0812, Japan
ryou@nict.go.jp

Shuichi Sakamoto

Research Institute of Electrical Communication
Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, 980-8577, Japan
saka@ais.riec.tohoku.ac.jp

Yoshifumi Chisaki

Faculty of Advanced Engineering
Chiba Institute of Technology
2-17-1 Tsudanuma, Narashino, Chiba, 275-0016, Japan
yoshifumi.chisaki@p.chibakoudai.jp

Zhenglie Cui

Research Institute of Electrical Communication
Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, 980-8577, Japan
sai@ais.riec.tohoku.ac.jp

Received January 2017; revised June 2017

Abstract. *Audio super-resolution is a technique by which a high-resolution signal is reconstructed from a low-resolution input. The low-resolution input can be a set of signals captured by multiple microphones at a low sampling rate. In such a case, the microphone configuration might affect the resultant reconstruction performance. This study is related to a method for updating the microphone configuration to keep it from becoming ill-conditioned in a super-resolution problem. We specifically examine the condition number of the design matrix and attempt to reduce it by optimally updating the configuration in an empirical but effective manner. To serve as an alternative to the condition number, a simple measure is defined: the sum of the inverse of the distances between microphones. Based on this measure, the microphone and direction to move are selected. Computer simulations show that the method reduces the worst condition number in many cases. Moreover, results show that when it is applied to signal processing of audio super-resolution, it can improve the estimation performance by 90.4% on average.*

Keywords: Distributed microphones, Super-resolution, Design matrix, Condition number, Microphone configuration

1. **Introduction.** A distributed microphone array assumes neither a uniform nor a preliminarily fixed configuration. For that reason, many research works related to estimating the microphone array configuration can be found in the literature. For example, Chen *et*

al. used signal energy to estimate the microphone positions [1]. For this purpose, Ono *et al.* used temporal correlations between audio signals captured by each microphone [2]. In contrast, few studies have examined how one should update the distributed audio device configuration to achieve better performance for a specific purpose of signal processing. Enomoto *et al.* developed a sound reproduction system using multiple loudspeakers surrounding a listener [3]. To cancel out the reverberation existing in the reproduction environment, an inverse filter is used in calculating signals to replay at loudspeakers. Robustness of the filter against errors in numerical calculation varies depending on the loudspeaker positions. Therefore, they proposed to arrange a position of each loudspeaker one after another so that their positions become geometrically orthogonal using Gram–Schmidt orthonormalization. A similar problem arises also in the audio capturing with distributed multiple microphones. An important difficulty in applying the approach of Enomoto *et al.* to the problem of microphone configuration under consideration is that it is applicable only when all audio devices can be arranged arbitrarily with no prerequisite. In a practical situation, however, because of restrictions on where audio devices can be mounted, its constellation is likely to be far from the ideal one. For that reason, the resultant performance becomes unsatisfactory. A practical necessity exists for reconfiguration of the device locations to achieve better performance, by moving a few devices by as small displacement as possible.

2. **Audio super-resolution.**

2.1. **Formulation.**

Super-resolution is a technique by which an image with higher resolution is obtained from multiple images with lower resolution [4]. For audio signals, various aspects of super-resolution have been investigated, such as pitch frequency resolution [5] and direction of arrival [6]. As described herein, we specifically examine a super-resolution that estimates a signal from those captured by multiple microphones at a sampling rate lower than the Nyquist frequency of the signal to estimate, under the assumption of plane wave propagation [7].

The inverse discrete Fourier transform can be expressed as

$$y(n) = \frac{1}{N}\sum_{k=0}^{N-1} Y(k)e^{j2\pi kn/N}, \tag{1}$$

where N signifies the number of samples, $Y(k)$ denotes the discrete Fourier transform of a signal $y(k)$, and $j^2 = -1$. It is necessary for complete reconstruction that n be an integer. Therefore, to accept real numbers for n, an approximation of (1) is used as

$$y(n) \simeq \mathrm{Re}\left[\frac{1}{N}\sum_{k=0}^{N-1} Y(k)e^{j2\pi kn/N}\right]. \tag{2}$$

This approximation is required to estimate the signal at the position where the traveling time from an actual microphone to that position is not a multiple of the sampling period. In theory, the time difference can be represented by phase shift in the frequency domain. Therefore, depending on the time difference, n needs to be real. And by applying this theory to signals captured by multiple microphones at different positions, samples of the signal at various points of space and time can be obtained under the following assumptions:

1. We have an acoustic wave $x(s, t)$ where t is a continuous time index and s is the position, which is emitted to the space as a plain wave.
2. The signal $x(s, t)$ can be quantized at microphone positions s_i, which can be written as $y_i(n)$, where n is an integer.
3. $Y(k)$ is the discrete Fourier transform of $y(n)$.

4. From a temporal point of view, $x(s_0, t)$ can be approximated as $\mathrm{Re}[F^{-1}[Y_i(k)](n)]$ by giving a real number as n.
5. From a spatial point of view, let $y_m(n)$ be the sampled signal received at the m-th microphone. Then $y_m(n) = x(n + \Delta_t(m))$, where $\Delta_t(m)$ is the traveling time of sound between the m-th microphone and s_0 assuming the plain wave.

From (2), by integrating both the temporal and spatial view points, we can formulate the following equation

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{X} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{y}$ is a vector constructed by cascading the samples that are actually captured by multiple microphones for a specified period, $\boldsymbol{\epsilon}$ is an additive white Gaussian noise, $\boldsymbol{X}$ denotes the DFT of a signal x that would be obtained by observing the signal at the desired sampling rate for that period, and Φ represents the design matrix which controls the time difference caused by both space (traveling time) and time (sampling time). The design matrix is defined as

$$\boldsymbol{\Phi}(\boldsymbol{d}) = \begin{bmatrix} \phi_0(d_0) & \phi_1(d_0) & \cdots & \phi_{N-1}(d_0) \\ \phi_0(d_1) & \phi_1(d_1) & \cdots & \phi_{N-1}(d_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(d_{M-1}) & \phi_1(d_{M-1}) & \cdots & \phi_{N-1}(d_{M-1}) \end{bmatrix}, \tag{4}$$

where M represents the number of observations, which subsequently becomes equal to the product of the number of microphones and that of the samples captured by each microphone. Also, N is the number of samples of the signal to estimate; d_i denotes the time lag of the corresponding observation from the assumed sampling time and point where the super-resolution signal should be obtained. Therefore, each component of the design matrix is defined as

$$\phi_k(d) \equiv \frac{1}{N} e^{j2\pi kd/N}. \tag{5}$$

This time lag accounts for the time delay of a signal traveling among the microphones as well as the sampling time.

2.2. ML estimation.

Based on (3), the maximum likelihood (ML) estimation of the signal is obtainable as

$$\boldsymbol{X} = (\boldsymbol{\Phi}^*\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^*\boldsymbol{y}, \tag{6}$$

where $*$ denotes the Hermitian transpose of a matrix. For that reason, $\boldsymbol{P} = \boldsymbol{\Phi}^*\boldsymbol{\Phi}$ becomes a Hermitian matrix, which assures real-valued eigenvalues. We therefore assess the condition number of this matrix because it is related closely to the stability and sensitivity of the ML estimation of the signal by (6) [8]. Fig. 1 schematically presents a set of microphones that is useful for quadruple super-resolution. Whereas Fig. 1(a) depicts a good condition, Fig. 1(b) does not because one pair of microphones captures the same signal, and they cannot contribute to obtaining additional information for super-resolution. As presented in Fig. 1, a desirable microphone configuration generally varies depending on the incident direction of the sound under consideration.

3. Microphone configuration.

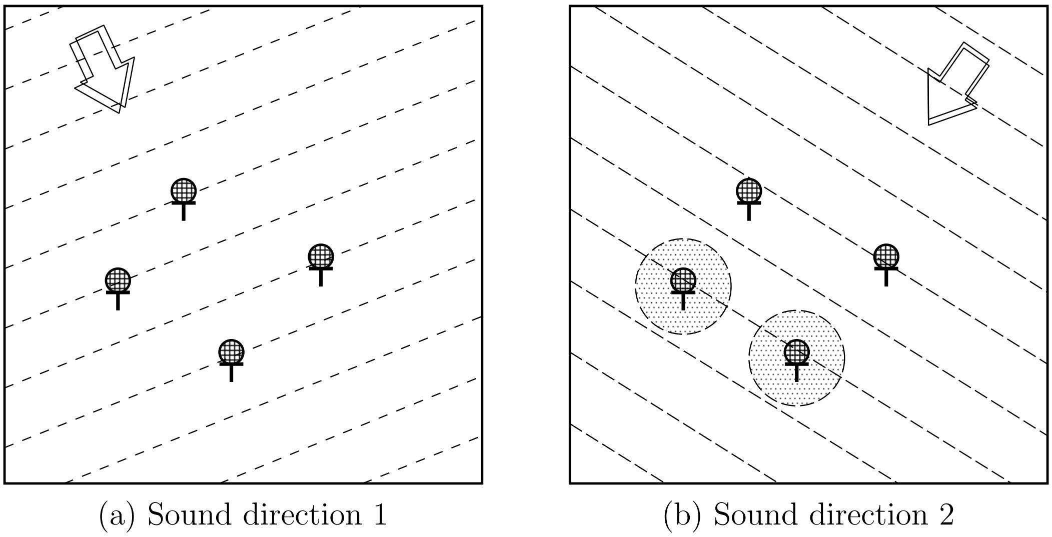

(a) Sound direction 1 (b) Sound direction 2

Figure 1. Microphone configuration examples for audio super-resolution. Dashed lines represent traveling time of sound equivalent to one sampling period for the signal to be estimated.

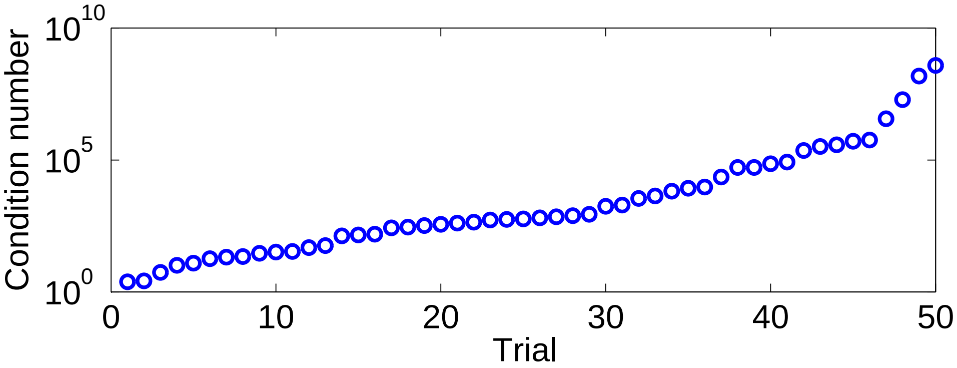


Figure 2. Condition number of matrix $\boldsymbol{P}$ for the random deployment of microphones. Results of 50 trials are shown in ascending order.

3.1. **Condition number.** A problem of the ML estimation given as (6) is that it includes inversion of a matrix. When the matrix to be inverted has a large condition number, a transformation using this inverted matrix becomes sensitive to added noise and configuration error, resulting in deterioration in the estimated signal.

Fig. 2 presents an example of condition numbers of $\boldsymbol{P}$ for 50 microphone configurations, permuted in ascending order. It is apparent that the condition number becomes considerably large for some cases. This observation indicates that, in these conditions, the ML estimation of the signal might result in markedly poor performance. Because no simple relation exists between the microphone configuration and the condition number of $\boldsymbol{P}$, we introduce a simple alternative measure.

3.2. **Alternative measure.** Ill-conditioning is likely to occur when multiple microphones capture closely resembling signals. To estimate the likelihood of its occurrence, a measure D is defined as the sum of inversion of distances between all combinations of microphones as

$$D = \sum_{p \neq q} \frac{1}{(d_{pq} + \lambda)^{\alpha}}, \tag{7}$$

where d_{pq} is the distance between a pair of microphones p and q. A small number λ is added to avoid infinity when two microphones are in proximity. In the following computer simulations, λ was set to EPS of Matlab, which is approximately 2.2×10^{-16}.

The optimal parameter α was investigated using exhaustive search. Fig. 3 shows Spearman's rank correlation coefficients between the condition number and the D measure as a function of α, where 50 trials were performed with different microphone configurations to obtain the coefficients, and this was repeated 50 times for each α. It shows that there exists a certain correlation and that it reaches its highest values within $\alpha = 1$ and 2. We use $\alpha = 1.4$ hereinafter in this paper.

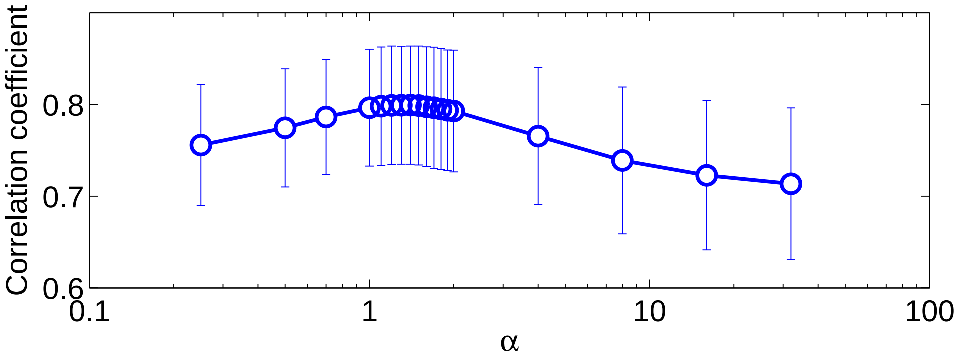


Figure 3. Spearman's rank correlation coefficient between D measure and condition number as a function of α. Error bars show the standard deviation.

Fig. 4 presents 50 trials of randomly distributed microphone arrays consisting of four microphones and their corresponding D measures. They are rearranged in ascending order of the condition number. The vertical axis is converted from microphone positions into sample delay, which is corresponding to the arrival time difference of sound. The condition of trial number 1 is a case where the condition number is the lowest. In this condition, the sample delay is distributed almost uniformly. In contrast, sample delays of multiple microphones become close as the condition number increases. Because of this reordering, the basic trend would be the same even if the number of trials increases.

3.3. **Update algorithm.** Our purpose is to reduce the worst condition number of $\boldsymbol{P}$. Fig. 4 and the definition of the D measure suggest that a microphone pair that is closest in terms of sample delay likely causes ill-conditioning. Our algorithm is to choose such a pair of microphones and to move one of them to the direction which will reduce the D measure most. The amount of movement is adjusted such that the increase/decrease of sample delay becomes a constant called a step parameter. This direction coincides with one from which the sound is assumed to come. Consequently, the algorithm is realized as described below.

1. Calculate condition numbers for all directions
2. Search for the direction which provides the worst condition number
3. At that direction, choose the pair of microphones which is closest in terms of the sample delay

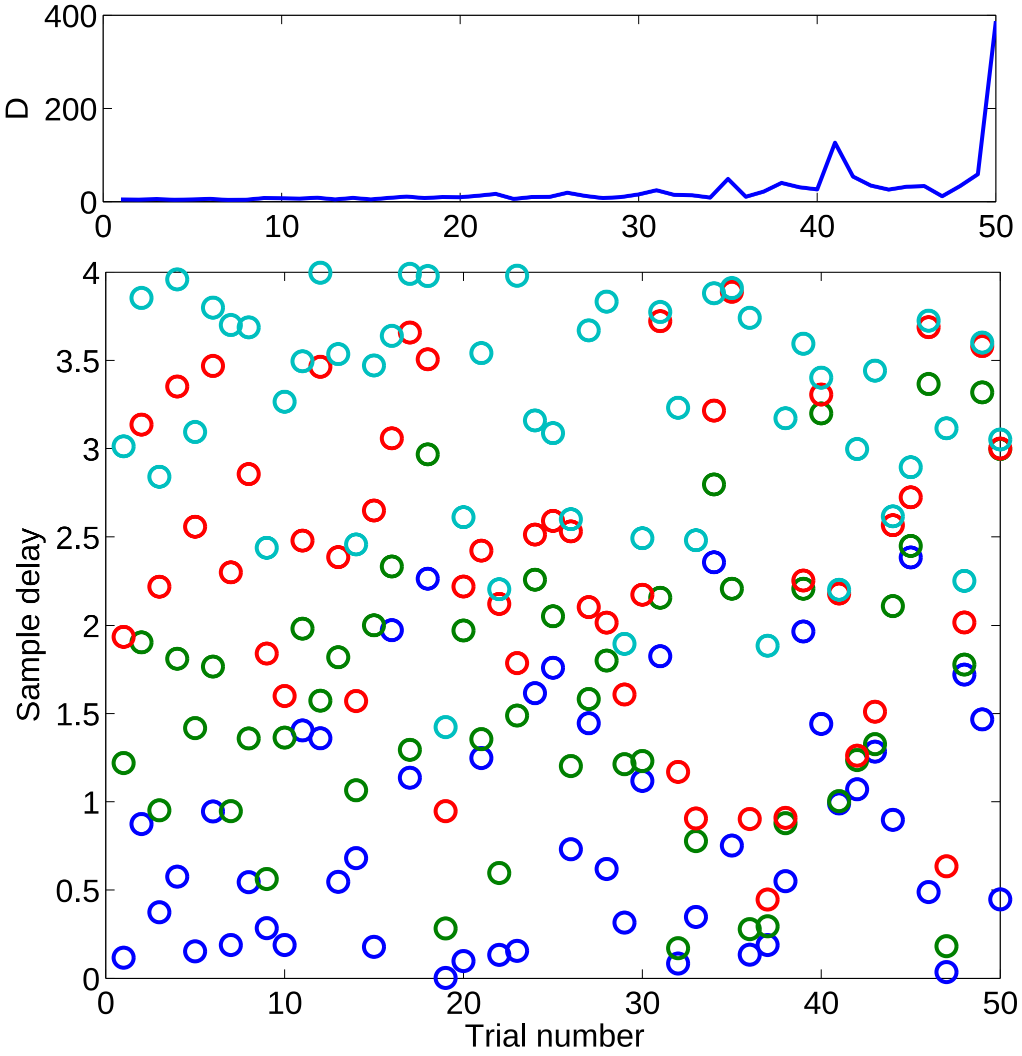


Figure 4. 50 trials of randomly distributed microphone configurations shown in ascending order of the condition number.

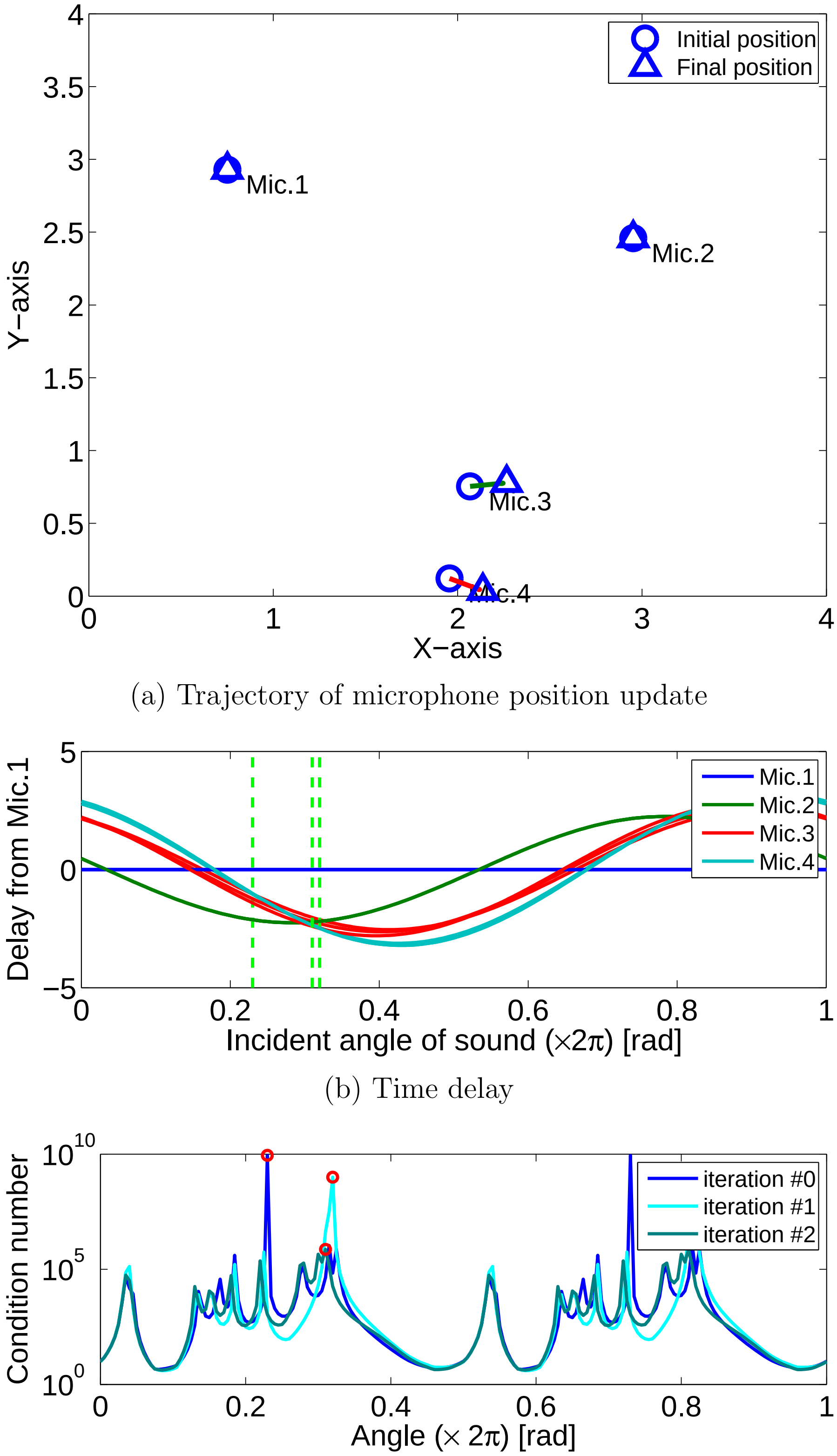
4. Calculate D measures for the four potential changes (two microphones × two directions)
5. Select one which reduces the D measure most
6. Go to Step 1, and stop the iteration if the condition number of new configurations for its worst case is greater than before.

An example of the evolution by this algorithm is presented in Fig. 5. In Figs. 5(a) and 5(b), results of each iteration are overlapped. Fig. 5(c) shows condition number for each iteration as a function of incident angle of the assumed sound. Circles indicate the maximum condition number for each iteration. This figure clearly shows that the maximum condition number dwindles as the iteration increases. Also, vertical dashed lines in Fig. 5(b) show the corresponding incident angle of sound, demonstrating that the maximum condition number is likely to occur when the time delays of multiple microphones are close.

4. Computer simulations.

4.1. Fundamental characteristics.

To elucidate the fundamental performance of the algorithm, condition numbers of initial and final microphone configurations are presented



(a) Trajectory of microphone position update

(b) Time delay

(c) Condition number

Figure 5. Update example: red circles in (c) show the maximum condition numbers for each iteration, and vertical dashed lines in (b) indicate the corresponding incident angle of sound.

as a scatter plot in Fig. 6. When the initial condition number is extremely large, the method almost always works. Results along the diagonal axis indicate that the method failed to reduce the condition number at the first update, which is apparent also in the histogram of the number of iterations depicted in Fig. 7. This problem might be mitigated by an adaptive step parameter.

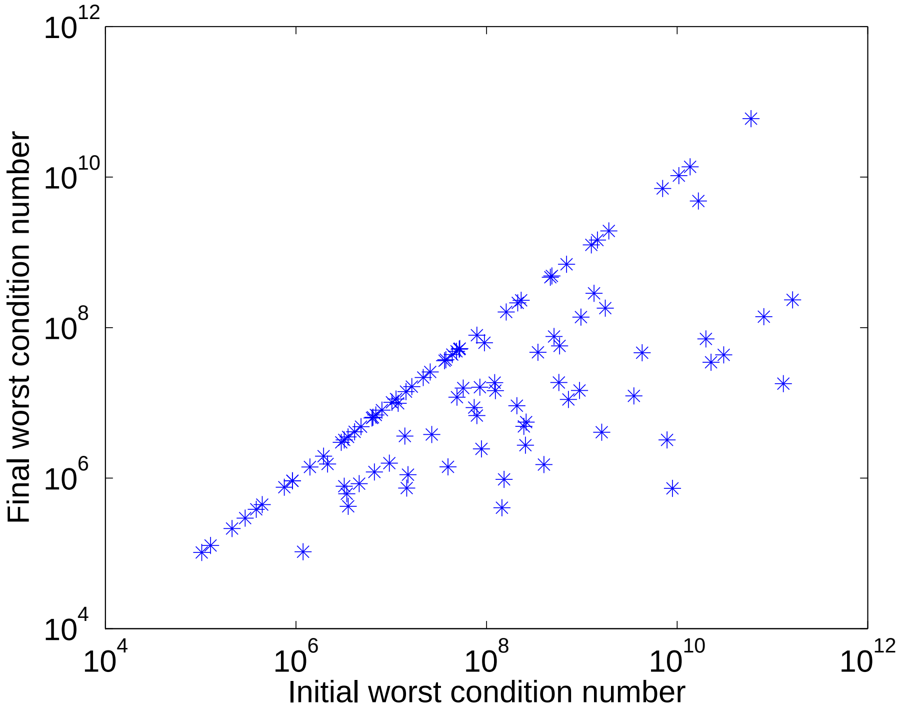


FIGURE 6. Relation between condition numbers for the initial and final microphone configurations.

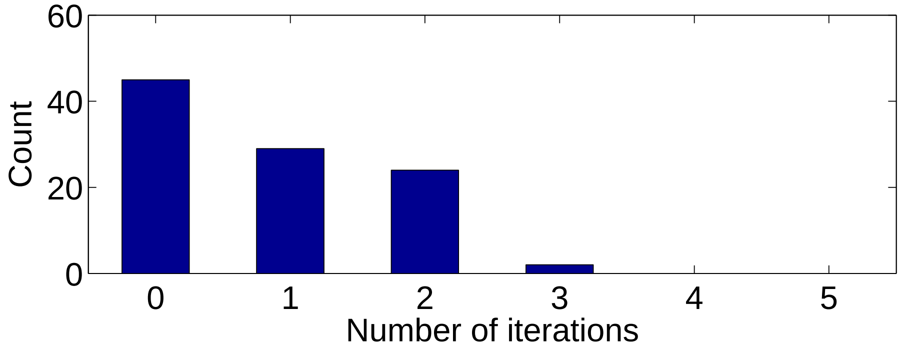


FIGURE 7. Histogram of the number of iterations.

4.2. **Performance in super-resolution.** To evaluate the effectiveness of the proposed method for signal processing, computer simulations of super-resolution were conducted using the setup shown in Table 1. The frequency of the signal to estimate, 600 Hz, is higher than the Nyquist frequency of the actual sampling rate, 1 kHz. In contrast, the desired sampling rate, 4 kHz, is high enough to satisfy the Nyquist sampling theorem for measuring a pure tone of 600 Hz. Therefore, if super-resolution by the proposed method works well, the signal should be reconstructed correctly. The super-resolution performance

TABLE 1. Setup of computer simulations.

Actual sampling rate	1 kHz
Desired sampling rate	4 kHz
Signal to estimate	Pure tone (600 Hz)
Signal-to-noise ratio	0 dB
Number of microphones	4
Step parameter	0.2
Size of design matrix	256 × 256

was measured using Pearson's correlation coefficient r between the original signal and the estimated one. Improvement by the proposed method was measured quantitatively as

$$q = (r_f - r_i)/r_i \times 100, \tag{8}$$

where r_i and r_f respectively denote performance measures of super-resolution under the initial and final microphone conditions. Fig. 8 shows r_i (Initial) and r_f (Final) for the 100 trials after rearranging it in ascending order of r_i. It should be noted that these results correspond to the worst case, namely the lower limit, for the given microphone position. Additionally, Fig. 9 shows signal to noise ratios after super-resolution for the initial and final microphone positions. These figures illustrate that the proposed method successfully improve the super-resolution performance for the worst condition especially when it is severely bad. However, it is also shown that the proposed method may worsen the performance in some cases and needs further improvement.

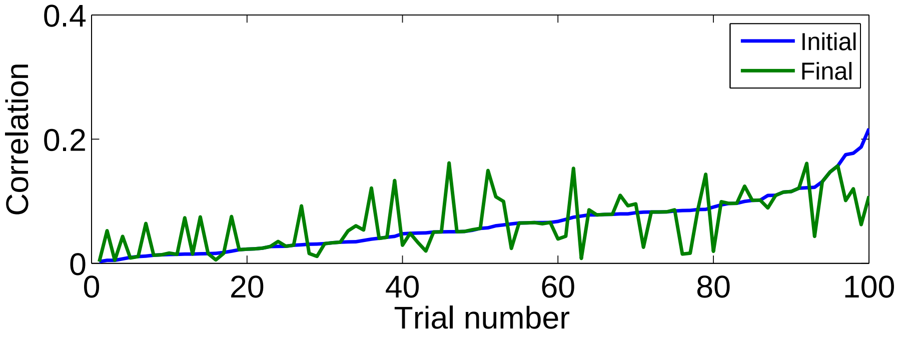


FIGURE 8. Correlation coefficients for the worst condition of the initial and final microphone positions.

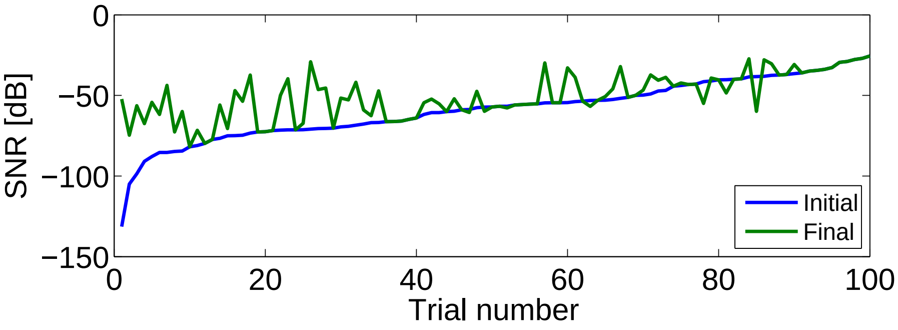


FIGURE 9. SNR for the worst condition of the initial and final microphone positions.

5. **Conclusion.** Using a measure defined to serve as an alternative to the condition number, microphone positions are updated iteratively such that the performance in the worst case will be relaxed. Computer simulations revealed that the proposed method certainly helps to improve the worst performance when applied to super-resolution. For better performance, an optimal selection of the step parameter must be investigated. Another matter to consider in the future is the case in which the number of microphones differs from the minimum number that is necessary.

Acknowledgment. This work is partially supported by JSPS KAKENHI (16KT0100).

REFERENCES

[1] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Chang, Energy-based position estimation of microphones and speakers for ad hoc microphone arrays, *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE*, pp. 22-25, 2007.

[2] N. Ono, H. Kohno, N. Ito, and S. Sagayama, Blind alignment of asynchronously recorded signals for distributed microphone array, *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE*, pp. 161-164, 2009.

[3] S. Enomoto, Y. Ikeda, S. Ise, and S. Nakamura, Optimization of loudspeaker and microphone configurations for sound reproduction system based on boundary surface control principal, *Proceedings of the 20th International Congress on Acoustics, ICA*, pp. 1-7, 2010.

[4] S. C. Park, M. K. Park, and M. G. Kang, Super-resolution image reconstruction: a technical overview, *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21-36, 2003.

[5] Y. Medan, E. Yair, and D. Chazan, Super resolution pitch determination of speech signals, *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 40-48, 1991.

[6] B. D. Rao and K. V. S. Hari, Performance analysis of root-music, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1939-1949, 1989.

[7] R. Nishimura, Y. Suzuki, Reconstruction of a high-sampling audio signal from low-sampling audio signals using super-resolution, *Proc. of Western Pacific Acoustics Conference (WESPAC)*, pp. 369-375, 2006.

[8] G. Strang, Linear algebra and its applications, Academic Press, 1976.