Separation of Vibration-Derived Sound Signals Based on Fusion Processing of Vibration Sensors and Microphones

Ryoichi Takashima, Yohei Kawaguchi, and Masahito Togami
Research and Development Group, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
Email: yohei.kawaguchi.xk@hitachi.com

***Abstract*—This paper proposes a sound source separation method for vibration-derived sound signals such as sounds derived from mechanical vibrations by using vibration sensors. The proposed method is based on two assumptions. First, a vibration signal and the sound derived from the vibration are assumed to have a linear correlation. This assumption enables us to model the vibration-derived sound as a linear convolution of a transfer function and a vibration signal recorded by a vibration sensor. Second, un-vibration-derived sound signals such that the sound source is not connected to vibration sensors via a solid medium are barely recorded by vibration sensors. This assumption leads to a constraint of the transfer function from the un-vibration-derived sound sources to the vibration sensors. The proposed framework is the same as a microphone-array-based blind source separation framework, except that the proposed method constructs arrays with microphones and vibration sensors, and the separation parameters are constrained by the prior knowledge gained from the above second assumption. Experimental results indicate that the separation performance of the proposed method is superior to that of a conventional microphone-array-based source separation method.**

***Index Terms*—blind source separation, vibration-derived sound, vibration sensor, microphone, local Gaussian model**

I. Introduction

Reduction of noise emanating from machines has high industrial value as a front-end of various interfaces used in noisy environments such as plants and car environments. Speech recognition system is a hands-free interface, and it has been utilized in the car navigation systems and is expected to help maintenance workers record the maintenance logs in plants. Sound logs of machines are also expected to be utilized for anomaly detections. However, the performances of those applications often degrade due to noises of various mechanical parts such as engines.

Various approaches to reduce the noise have been studied. On the one hand, conventional single-channel noise-reduction methods such as spectral subtraction [1], Wiener filtering [2], using minimum mean-square error short-term spectral amplitude [3], and using optimally-modified log-spectral amplitude [4] work well when the noise is temporally stationary (e.g., the sounds of engines and air-conditioners). However, when the noise is highly non-stationary, such as the sounds of printers and other complicated machines that have multiple driving parts, the noise reduction performances of these methods are greatly reduced. On the other hand, microphone-array-based approaches [5], [6], [7], [8], [9], [10], [11] have been studied for removing non-stationary noise. These methods utilize the difference in the direction-of-arrival for each sound source instead of the assumption of the stationary noise sources. Classical microphone-array-based approaches such as beamforming (BF) [6] and independent component analysis (ICA) [7] cannot separate more sound sources than sensors. Various multichannel methods such as multichannel nonnegative matrix factorization [8] and local Gaussian modeling (LGM) [9], [10] have also been studied for separating more sound sources than sensors. However, the separation is difficult for the microphone-array-based methods when the directions of the sound sources are close to each other, and when the noise spatially diffuses.

In this paper, we focus on the fact that the noises of machines are derived from the machines' vibrations, and we propose a sound source separation method for vibration-derived sound using vibration sensors. The vibration sensors are set such that they can record the vibrations from machines, and the inputs of the vibration sensors and microphones are recorded synchronously. The proposed method is based on two assumptions. First, when a sound is produced from a machine's vibration and arrives at a microphone through the air, we assume that the vibration-derived component observed by a microphone and the signal observed by a vibration sensor have a linear correlation. In accordance with this assumption, we can model the vibration-derived sound as a linear convolution of a transfer function and the vibration signal recorded by a vibration sensor, and we can apply the modeling used in the conventional microphone-array-based separation framework to fuse the vibration sensors and microphones. In this paper, we use a blind source separation (BSS) framework based on the LGM [9]. The second assumption is that, un-vibration-derived sound signals such that the sound source is not connected to vibration sensors via a solid medium (e.g., speech signals) are barely recorded by vibration sensors. This assumption leads to a constraint of the transfer function from the un-vibration-derived sound sources to the vibration sensors. In the proposed method, we implement this constraint on the LGM framework by using the prior of the covariance matrix of the multi-channel observed signal [10].

ISBN 978-0-9928626-7-1

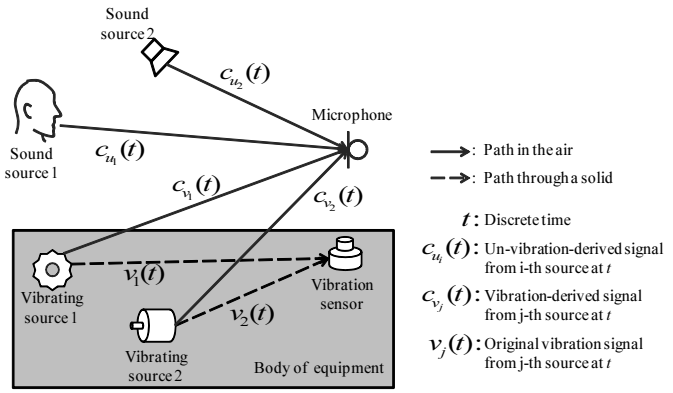

Fig. 1. Problem setting

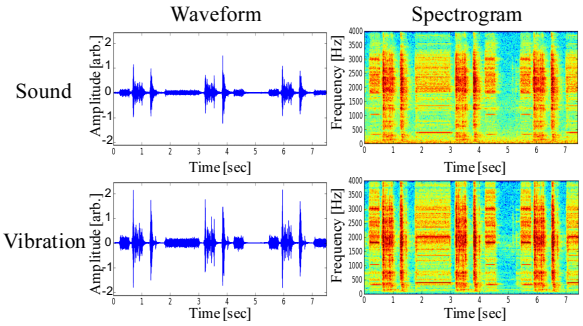

Fig. 2. Comparison between a sound signal and a vibration signal

The main contributions of this paper are twofold. First, we experimentally show the possibility of assuming a linear relationship between vibration and sound signals and applying them to the same models in the conventional microphone-array-based framework. Second, we demonstrate that fusing vibration sensors and microphones with the constraint described in the above paragraph provides better performances of the vibration-derived-sound separation than using microphones only.

II. Problem Statement

Figure 1 shows the problem statement in this paper. We define two kinds of sounds: '*vibration-derived sounds*' and '*un-vibration-derived sounds*'. The vibration-derived sound is produced from a machine's vibration and arrives at a microphone through the air, and the machine's vibration can be recorded by a vibration sensor. The un-vibration-derived sound is the sound of which the sound source is not connected to the vibration sensor via a solid medium (e.g., speech signals), and it is barely recorded by the vibration sensor. In Figure 1, as an example, there are two vibration sources and two un-vibration-derived sound sources. On the one hand, the mixed signal of the un-vibration-derived sound $c_{u_i}(t)$ and the vibration-derived sound $c_{v_j}(t)$ is recorded by one or more microphones. On the other hand, the mixed signal of vibration signals $v_i(t)$ is recorded by one or more vibration sensors. i and j are the indexes of the un-vibration-derived sound source and the vibration-derived sound source, respectively, and t denotes the discrete time index. Then, the goal of this paper is to separate the mixed sound signal for each un-vibration-derived sound source and each vibration-derived sound source.

III. Proposed Method

A. Input signal modeling

The proposed method assumes that a vibration-derived sound and the vibration signal observed by a vibration sensor have a high linear correlation. Figure 2 shows a waveform of microphone input and its spectrogram and those of vibration sensor input when a vibration speaker on a table was operating. The vibration speaker is an actuator which can generate any

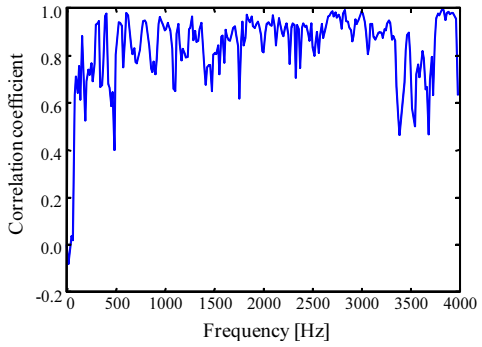

Fig. 3. Correlation coefficients between sound and vibration signals for each frequency bin. The average of correlation coefficients was 0.83.

signal given from an audio device on a solid medium. In this experiment, the vibration speaker generated a simulative vibration of an operating printer on the table. The vibration sensor was set on the table, and the microphone was set 19 cm above the table and the vibration sensor. The sampling frequency was 8 kHz. More experimental conditions are described in Section IV-A. As shown in Figure 2, waveforms and their spectrograms of sound and vibration are similar to each other. Figure 3 shows the correlation coefficients between the vibration-derived sound signal recorded by the microphone and the vibration signal recorded by the vibration sensor for each frequency bin. For measuring correlation, we calculated power spectrograms of signals, and then for each frequency bin, we calculated the correlation coefficients between time series of power of vibration sensor signal and that of microphone signal. Although there are some fluctuations, the result shows an average of correlation coefficients of 0.83. In accordance with this result, we assume that they have a linear correlation, and our proposed method approximates a vibration-derived sound signal as a linear convolution of a transfer function and a vibration signal.

$$c_{v_j}(t) \approx \sum_{t'=0}^{T} a_{v_j}(t') v_j(t - t') \tag{1}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

a_{v_j} denotes the transfer function with the length of T.

In accordance with the assumption of linear correlation, we can combine the multi-channel input signals of microphones and vibration sensors by applying the conventional microphone-array-based source separation frameworks. The proposed method constructs an array with one or more microphones and one or more vibration sensors. However, the conventional BF approaches [6] are difficult to be utilized because it is difficult to calculate the time difference between the acoustic signal passing in the air and the vibration signal passing through the solid medium. Therefore, we utilize the BSS framework which does not use any geometric information. We employ the LGM-based BSS framework [9]. The LGM has high separation performance compared to the conventional ICA-based approach [7], and we can easily add a constraint on the original LGM by using the maximum a posteriori (MAP) framework [10].

In this paper, we define the m-th microphone input and m-th vibration sensor input at time t as $x^{mic,m}(t)$ and $x^{vib,m}(t)$, respectively. The number of microphones and vibrations are defined as M_{mic} and M_{vib}, respectively. Applying the short term Fourier transform, the short term spectra $x^{mic,m}(f,\tau)$ and $x^{vib,m}(f,\tau)$ are obtained. f and τ are the index of the frequency bin and the frame index, respectively. We also define the short term spectra of a multi-channel input signal as follows:

$$\boldsymbol{x}^{mic}(f,\tau) = [\ x^{mic,1}(f,\tau) \quad \dots \quad x^{mic,M_{mic}}(f,\tau)\ ]^T \quad (2)$$

$$\boldsymbol{x}^{vib}(f,\tau) = [\ x^{vib,1}(f,\tau) \quad \dots \quad x^{vib,M_{vib}}(f,\tau)\ ]^T \quad (3)$$

$$\boldsymbol{x}(f,\tau) = [\ \boldsymbol{x}^{mic^T}(f,\tau) \quad \boldsymbol{x}^{vib^T}(f,\tau)\ ]^T. \quad (4)$$

Here, T denotes the transpose of the matrix. From the assumption of the linear correlation, $\boldsymbol{x}(f,\tau)$ can be expressed as

$$\begin{aligned}\boldsymbol{x}(f,\tau) &= \sum_i \boldsymbol{c}_{u_i}(f,\tau) + \sum_j \boldsymbol{c}_{v_j}(f,\tau) \quad (5)\\ &= \sum_i s_{u_i}(f,\tau)\boldsymbol{a}_{u_i}(f) + \sum_j s_{v_j}(f,\tau)\boldsymbol{a}_{v_j}(f).\end{aligned}$$

$\boldsymbol{c}_{u_i}(f,\tau) = s_{u_i}(f,\tau)\boldsymbol{a}_{u_i}(f)$ and $\boldsymbol{c}_{v_j}(f,\tau) = s_{v_j}(f,\tau)\boldsymbol{a}_{v_j}(f)$ represent multi-channel inputs for each un-vibration-derived sound source and each vibration-derived sound source, respectively. $s_{u_i}(f,\tau)$ and $s_{v_j}(f,\tau)$ are the original clean signals from each un-vibration-derived sound source and each vibration-derived sound source, respectively. $\boldsymbol{a}_{u_i}(f)$ and $\boldsymbol{a}_{v_j}(f)$ are the transfer functions from each un-vibration-derived sound source and each vibration-derived sound source to each microphone and each vibration sensor, respectively. $\boldsymbol{a}_{u_i}(f)$ and $\boldsymbol{a}_{v_j}(f)$ can be expressed as

$$\begin{aligned}&\boldsymbol{a}_{u_i}(f) = [\boldsymbol{a}_{u_i}^{mic^T}, \boldsymbol{a}_{u_i}^{vib^T}]^T \quad (6)\\ &= [a_{u_i}^{mic,1}(f), \dots, a_{u_i}^{mic,M_{mic}}(f), a_{u_i}^{vib,1}(f), \dots, a_{u_i}^{vib,M_{vib}}]^T(f)\\ &\boldsymbol{a}_{v_j}(f) = [\boldsymbol{a}_{v_j}^{mic^T}, \boldsymbol{a}_{v_j}^{vib^T}]^T \quad (7)\\ &= [a_{v_j}^{mic,1}(f), \dots, a_{v_j}^{mic,M_{mic}}(f), a_{v_j}^{vib,1}(f), \dots, a_{v_j}^{vib,M_{vib}}]^T(f)\end{aligned}$$

where $a_{u_i}^{mic,1}$ denotes the acoustic transfer function from the i-th un-vibration-derived sound source to the 1st microphone.

In the LGM framework, the probability density function of a multi-channel speech signal is modeled as a time-variant Gaussian with 0-mean and a time-variant multi-channel covariant matrix. Assuming the high linear correlation between the sound and the vibration, the multi-channel vibration signal can also be modeled by the 0-mean time-variant Gaussian. Therefore, $\boldsymbol{c}_{u_i}(f,\tau)$ and $\boldsymbol{c}_{v_j}(f,\tau)$ are modeled by Gaussian distributions with 0-mean and the time-variant covariant matrix $p_{u_i}(f,\tau)\boldsymbol{R}_{u_i}(f)$ and $p_{v_j}(f,\tau)\boldsymbol{R}_{v_j}(f)$, respectively. Then, the multi-channel input signal is also modeled by a Gaussian distribution with 0-mean and the covariant matrix $\boldsymbol{R}_x(f,\tau)$ and $\boldsymbol{R}_x(f,\tau)$ as follows:

$$\boldsymbol{R}_x(f,\tau) = \sum_i p_{u_i}(f,\tau)\boldsymbol{R}_{u_j}(f) + \sum_j p_{v_j}(f,\tau)\boldsymbol{R}_{v_i}(f). \quad (8)$$

where $\boldsymbol{R}_{u_i}(f)$ and $\boldsymbol{R}_{v_j}(f)$ denote the spatial correlation matrices, and $p_{u_i}(f,\tau)$ and $p_{v_j}(f,\tau)$ denote the activity of un-vibration-derived sound and vibration-derived sound, respectively.

B. Parameter estimation by using MAP-EM algorithm

In the proposed method, the unknown parameters are $\boldsymbol{R}_{u_i}(f)$, $\boldsymbol{R}_{v_j}(f)$, $p_{u_i}(f,\tau)$, $p_{v_j}(f,\tau)$, $\boldsymbol{c}_{u_i}(f,\tau)$, and $\boldsymbol{c}_{v_j}(f,\tau)$. In the same way as in the original LGM framework, these parameters are estimated by using the EM algorithm [12], [8]. Moreover, the proposed method uses a constraint for spatial correlation matrices related to the un-vibration-derived sound sources $\boldsymbol{R}_{u_i}(f)$ in accordance with the assumption that un-vibration-derived sound signals are barely recorded by vibration sensors. The constraint of the spatial correlation matrices can be implemented to the original LGM by using the MAP estimation approach [10].

In accordance with [10], the proposed method models the probability density function of $\boldsymbol{R}_{u_i}(f)$ as the inverse-Wishart distribution and estimates the unknown parameters in the framework of the MAP estimation with the EM algorithm. In the EM algorithm, we define $\boldsymbol{c}_{u_i}(f,\tau)$ and $\boldsymbol{c}_{v_j}(f,\tau)$ as latent variables. If no un-vibration-derived sounds are recorded by the vibration sensors, $\boldsymbol{R}_{u_i}(f)$ can be expressed as follows.

$$\boldsymbol{R}_{u_i}^{prior}(f) = \begin{bmatrix} \boldsymbol{R}_{u_i}^{M_{mic}\times M_{mic}}(f) & \boldsymbol{0}^{M_{mic}\times M_{vib}} \\ \boldsymbol{0}^{M_{vib}\times M_{mic}} & \boldsymbol{0}^{M_{vib}\times M_{vib}} \end{bmatrix} \quad (9)$$

$$\boldsymbol{R}_{u_i}^{M_{mic}\times M_{mic}}(f) = \boldsymbol{a}_{u_i}^{mic}(f)\boldsymbol{a}_{u_i}^{mic^H}(f) \quad (10)$$

Since a few un-vibration-derived sounds are recorded by the vibration sensors in real situations, the proposed method models the probability density function of $\boldsymbol{R}_{u_i}(f)$ as the inverse-Wishart distribution as follows:

$$\begin{aligned}&\boldsymbol{Pr}(\boldsymbol{R}_{u_i}(f)|\boldsymbol{R}_{u_i}^{prior}(f), d)\\ &= \frac{|\boldsymbol{R}_{u_i}^{prior}(f)|^d |\boldsymbol{R}_{u_i}(f)|^{-(d+M_{mic}+M_{vib})} e^{-tr(\boldsymbol{R}_{u_i}^{prior}(f)\boldsymbol{R}_{u_i}^{-1}(f))}}{\pi^{(M_{mic}+M_{vib})(M_{mic}+M_{vib}-1)/2}\Pi_{m=1}^{M_{mic}+M_{vib}}\Gamma(d-m+1)} \quad (11)\end{aligned}$$

where d $(d \geq M_{mic}+M_{vib})$ is a parameter of the inverse-Wishart distribution that determines the degrees of freedom [13], and Γ denotes the gamma function.

ISBN 978-0-9928626-7-1

The n-th step of the estimation algorithm based on the MAP-EM algorithm is written as follows.

E step

The sufficient statistics of the separated signals are computed as the following steps. At first, the filters $\boldsymbol{W}_{u_i}(f,\tau)$ and $\boldsymbol{W}_{v_j}(f,\tau)$ for separating i-th un-vibration-derived sound signal $\boldsymbol{c}_{u_i}(f,\tau)$ and j-th vibration-derived sound signal $\boldsymbol{c}_{v_j}(f,\tau)$, respectively, are obtained by the following equations.

$$\boldsymbol{R}_{\boldsymbol{c}_{u_i}}(f,\tau) = p_{u_i}^{(n-1)}(f,\tau)\boldsymbol{R}_{u_i}^{(n-1)}(f) \quad (12)$$

$$\boldsymbol{R}_{\boldsymbol{c}_{v_j}}(f,\tau) = p_{v_j}^{(n-1)}(f,\tau)\boldsymbol{R}_{v_j}^{(n-1)}(f) \quad (13)$$

$$\boldsymbol{R}_x(f,\tau) = \sum_i \boldsymbol{R}_{\boldsymbol{c}_{u_i}}(f,\tau) + \sum_j \boldsymbol{R}_{\boldsymbol{c}_{v_j}}(f,\tau) \quad (14)$$

$$\boldsymbol{W}_{u_i}(f,\tau) = \boldsymbol{R}_{\boldsymbol{c}_{u_i}}(f,\tau)\boldsymbol{R}_x^{-1}(f,\tau) \quad (15)$$

$$\boldsymbol{W}_{v_j}(f,\tau) = \boldsymbol{R}_{\boldsymbol{c}_{v_j}}(f,\tau)\boldsymbol{R}_x^{-1}(f,\tau) \quad (16)$$

By using the estimated filters $\boldsymbol{W}_{u_i}(f,\tau)$ and $\boldsymbol{W}_{v_j}(f,\tau)$, the separated signals $\boldsymbol{c}_{u_i}(f,\tau)$ and $\boldsymbol{c}_{v_j}(f,\tau)$ are estimated as follows.

$$\hat{\boldsymbol{c}}_{u_i}(f,\tau) = \boldsymbol{W}_{u_i}(f,\tau)\boldsymbol{x}(f,\tau) \quad (17)$$

$$\hat{\boldsymbol{c}}_{v_j}(f,\tau) = \boldsymbol{W}_{v_j}(f,\tau)\boldsymbol{x}(f,\tau) \quad (18)$$

Then, the sufficient statistics that are used in the M step are computed by the following equations.

$$\hat{\boldsymbol{R}}_{\boldsymbol{c}_{u_i}}(f,\tau) = \hat{\boldsymbol{c}}_{u_i}(f,\tau)\hat{\boldsymbol{c}}_{u_i}^H(f,\tau) + (\boldsymbol{I} - \boldsymbol{W}_{u_i}(f,\tau))\boldsymbol{R}_{\boldsymbol{c}_{u_i}}(f,\tau) \quad (19)$$

$$\hat{\boldsymbol{R}}_{\boldsymbol{c}_{v_j}}(f,\tau) = \hat{\boldsymbol{c}}_{v_j}(f,\tau)\hat{\boldsymbol{c}}_{v_j}^H(f,\tau) + (\boldsymbol{I} - \boldsymbol{W}_{v_j}(f,\tau))\boldsymbol{R}_{\boldsymbol{c}_{v_j}}(f,\tau) \quad (20)$$

Here, $\boldsymbol{I}$ denotes the identity matrix, and H means the conjugate transpose of the matrix.

M step

By using the sufficient statistics of the separated signals computed in the E step, the activities and spatial correlation matrices are updated by the following equations.

$$p_{u_i}^{(n)}(f,\tau) = \frac{1}{M_{mic}+M_{vib}} tr(\boldsymbol{R}_{u_i}^{(n-1)^{-1}}(f)\hat{\boldsymbol{R}}_{\boldsymbol{c}_{u_i}}(f,\tau)) \quad (21)$$

$$p_{v_j}^{(n)}(f,\tau) = \frac{1}{M_{mic}+M_{vib}} tr(\boldsymbol{R}_{v_j}^{(n-1)^{-1}}(f)\hat{\boldsymbol{R}}_{\boldsymbol{c}_{v_j}}(f,\tau)) \quad (22)$$

$$\boldsymbol{R}_{u_i}^{(\alpha,\beta)^{(n)}}(f) = \begin{cases} \frac{1}{\gamma(d+M_{mic}+M_{vib})+L}\sum_\tau \frac{1}{p_{u_i}^{(n)}(f,\tau)}\hat{\boldsymbol{R}}_{\boldsymbol{c}_{u_i}}(f,\tau) & ((\alpha,\beta) < M_{mic}) \\ \frac{1}{L}\sum_\tau \frac{1}{p_{u_i}^{(n)}(f,\tau)}\hat{\boldsymbol{R}}_{\boldsymbol{c}_{u_i}}(f,\tau) & (otherwise) \end{cases} \quad (23)$$

$$\boldsymbol{R}_{v_j}^{(n)}(f) = \frac{1}{L}\sum_\tau \frac{1}{p_{v_j}^{(n)}(f,\tau)}\hat{\boldsymbol{R}}_{\boldsymbol{c}_{v_j}}(f,\tau) \quad (24)$$

L is the number of frames. $\boldsymbol{R}_{u_i}^{(\alpha,\beta)}(f)$ denotes (α,β) elements of $\boldsymbol{R}_{u_i}(f)$. γ is a hyper parameter which controls the strength of the constraint for $\boldsymbol{R}_{u_i}(f)$. In this paper, we set hyper parameters as $d = M_{mic} + M_{vib}$ and $\gamma = 10.0$, respectively.

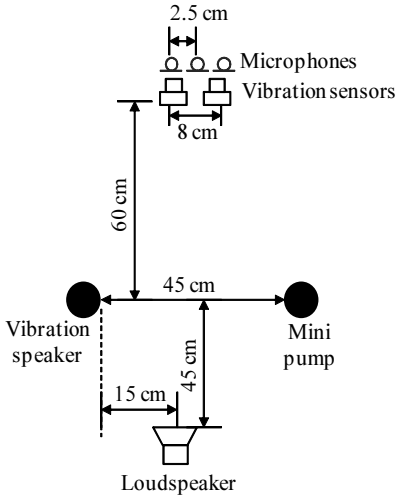

Fig. 4. Experimental environment (top view)

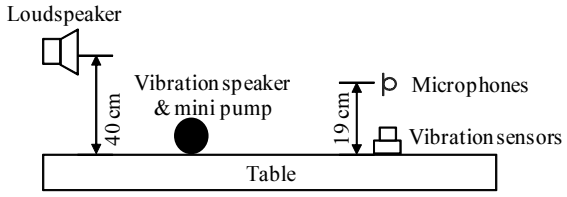

Fig. 5. Experimental environment (side view)

IV. Experiments

A. Experimental conditions

We carried out experiments for separating an un-vibration-derived sound and two vibration-derived sound signals. Figure 4 and Figure 5 show the top view and the side view of the experimental environment, respectively. A loudspeaker was set 40 cm above the table, and the impulse response was recorded by using the time stretched pulse (TSP) method [14]. The speech uttered from the loudspeaker's position was simulated by convolution of the impulse response and clean speech. The original clean speech was extracted from the TIMIT database [15] for 34 speakers (one utterance each).

For vibration-derived sound sources, we set a mini pump and a vibration speaker on the table. The vibration speaker is a kind of actuator that can generate any vibration given from an audio device on a solid medium. In this experiment, we recorded the vibration of the operating printer in advance. Then, by outputting the recorded vibration signal from the vibration speaker, we artificially generated the printer's vibration on the table and its vibration-derived sound. The mini pump generates stationary vibration and sound. We recorded the mixed sound and mixed vibration of those from the vibration speaker and the mini pump by using two vibration sensors and three microphones. The mixed observed signal was simulated

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE I
SDR [dB] FOR EACH NUMBER OF MICROPHONES AND VIBRATION SENSORS. SDR OF MICROPHONE INPUT WITHOUT SEPARATION WAS -2.1 dB. "NUM. OF VIBRATION SENSORS = 0" MEANS THAT THE CONVENTIONAL MICROPHONE-ARRAY-BASED METHOD IS APPLIED.

		Num. of mic.		
		1ch	2ch	3ch
Num. of vibration sensors	0ch	none	3.2	4.0
	1ch	5.6	4.7	4.1
	2ch	7.2	8.2	7.6

TABLE II
SDRS [dB] OF THE PROPOSED METHOD WITHOUT CONSTRAINT FOR $\boldsymbol{R}_{u_i}(f)$.

		Num. of mic.		
		1ch	2ch	3ch
Num. of vibration sensors	1ch	4.4	4.8	5.6
	2ch	5.9	5.5	5.4

by adding the convoluted speech signal and the vibration-derived sound. The sampling frequency was 8 kHz, and the frame size and frame shift were 1024 point and 64 point, respectively.

B. Experimental results

In order to compare the performance of the proposed method with the conventional method using microphones only, we also evaluated the performance of the original LGM method. Table I shows the signal-to-distortion-ratio (SDR) of signals separated by using the proposed method and the original LGM method. "Num. of vibration sensors = 0" means that the original LGM method using a microphone array is applied. The average SDR of the observed signal without separation was -2.1 dB. As shown in this table, the proposed method using at least one vibration sensor outperforms the original LGM method, even when the total number of sensors in the proposed method is smaller than that of original LGM (See that the proposed method with one microphone and one vibration sensor outperforms the original LGM with three microphones). This result indicates that the use of vibration sensors improves the separation performances compared with using microphones only. One reason may be that the vibration sensors can obtain the information of the vibration-derived sound signal with higher quality than microphones because few un-vibration-derived sound signals such as speech signals are recorded by vibration sensors. Increasing the number of vibration sensors increases the performance of the proposed method.

In our proposed method, the spatial correlation matrix is constrained by using the inverse-Wishart distribution. Table II shows the SDR without the constraint. By comparing with Table I, the constraint works better when the number of vibration sensors is larger.

V. CONCLUSION

This paper has described a sound source separation method for vibration-derived sound signals such as sounds derived from mechanical vibrations by using vibration sensors. In accordance with the assumption of a high linear relationship between a vibration signal and the sound derived from the vibration, the proposed method constructs arrays with microphones and vibration sensors and separates the vibration-derived sound signal in the similar framework of the conventional microphone-array-based BSS framework. In addition, in accordance with the assumption that few un-vibration-derived sound signals such as speech signals are recorded by vibration sensors, the estimation of the spatial correlation matrices are constrained. The experimental results indicates that the proposed method outperforms the conventional separation method using microphones only. In future work, we will evaluate the performance of the proposed method with a greater variety in the number of vibration sensors and microphones. We will also evaluate the performance in more real situations using real complicated machines instead of the vibration speaker.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[2] P. C. Loizou, *Speech Enhancement: Theory And Practice*, CRC Press, NY, USA, 2007.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[5] D. Johnson and D. Dudgeon, *Array Signal Processing*, Prentice Hall, Upper Saddle River, NJ, USA, 1996.

[6] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," in *Proc. IEEE*, 1972, vol. 60(8), pp. 926–935.

[7] E. Oja A. Hyvärinen, J. Karhunen, *Independent component analysis*, John Wiley & Sons, 2001.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[9] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[10] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 149, pp. 11 pages, 2013.

[11] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 268–272.

[12] A.P. Dempster et al., "Maximum likelihood from incomplete data via the em algorithm," *J. of the Royal Statistic Society*, vol. 39, no. 1, pp. 1–38, 1977.

[13] D. Maiwald and D. Kraus, "Calculation of moments of complex wishart and complex inverse wishart distributed matrices," *IEE Proc. Radar, Sonar and Navigation*, vol. 147, no. 4, pp. 162–168, 2000.

[14] Y. Suzuki, F. Asano, H. Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1995.

[15] *TIMIT corpus*, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogID=LDC93S1.

ISBN 978-0-9928626-7-1