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 also 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 beam-forming (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].
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 $c_{ui}(t)$ and the vibration-derived sound $c_{vj}(t)$ is recorded by one or more microphones. On the other hand, the mixed signal of vibration signals $v_{ji}(t)$ 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 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.

$$c_{vj}(t) \approx \sum_{t'=0}^{T} \alpha_{vj}(t') v_j(t - t')$$  \hspace{1cm} (1)
$a_{v_j}$ 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 time Fourier transform, the short term spectra $x^{mic,f}(t)$ and $x^{vib,f}(t)$ are obtained. $f$ and $t$ 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:

$$x^{mic,f}(t) = [x^{mic,1}(f,t) \ldots x^{mic,M_{mic}(f,t)}]^T$$
$$x^{vib,f}(t) = [x^{vib,1}(f,t) \ldots x^{vib,M_{vib}(f,t)}]^T$$

(2)

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

$$x(f, \tau) = \sum_i c_{ui}(f, \tau) + \sum_j c_{vj}(f, \tau)$$

(5)

$$= \sum_i s_{ui}(f, \tau)a_{ui}(f) + \sum_j s_{vj}(f, \tau)a_{vj}(f).$$

$c_{ui}(f, \tau)$, $c_{uj}(f, \tau)$ and $c_{vj}(f, \tau)$ are the spatial correlation matrices related to the un-vibration-derived sound and vibration-derived sound, respectively. $s_{ui}(f, \tau)$ and $s_{uj}(f, \tau)$ are the original clean signals from each un-vibration-derived sound source and each vibration-derived sound source, respectively. $a_{ui}(f)$ and $a_{uj}(f)$ 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.

$a_{ui}(f)$ and $a_{uj}(f)$ can be expressed as

$$a_{ui}(f) = [a_{ui}^{micT}, a_{ui}^{vibT}]^T$$

(6)

$$= [a_{ui}^{mic,1}(f), \ldots, a_{ui}^{mic,M_{mic}}(f), a_{ui}^{vib,1}(f), \ldots, a_{ui}^{vib,M_{vib}}(f)]^T$$

$$a_{uj}(f) = [a_{uj}^{micT}, a_{uj}^{vibT}]^T$$

(7)

$$= [a_{uj}^{mic,1}(f), \ldots, a_{uj}^{mic,M_{mic}}(f), a_{uj}^{vib,1}(f), \ldots, a_{uj}^{vib,M_{vib}}(f)]^T$$

where $a_{ui}^{mic,1}$ 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 covariance 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, $c_{ui}(f, \tau)$ and $c_{uj}(f, \tau)$ are modeled by Gaussian distributions with 0-mean and the time-variant covariant matrix $p_{ui}(f, \tau)R_{ui}(f) + p_{uj}(f, \tau)R_{uj}(f)$, respectively. Then, the multi-channel input signal is also modeled by a Gaussian distribution with 0-mean and the covariant matrix $R_{x}(f, \tau)$ and $R_{x}(f, \tau)$ as follows:

$$R_{x}(f, \tau) = \sum_i p_{ui}(f, \tau)R_{ui}(f) + \sum_j p_{uj}(f, \tau)R_{uj}(f).$$

(8)

where $R_{ui}(f)$ and $R_{uj}(f)$ denote the spatial correlation matrices, and $p_{ui}(f, \tau)$ and $p_{uj}(f, \tau)$ 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 $R_{ui}(f)$, $R_{uj}(f)$, $p_{ui}(f, \tau)$, $p_{uj}(f, \tau)$, $a_{ui}(f)$, and $a_{uj}(f)$. 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 $R_{ui}(f)$ 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 $R_{ui}(f)$ 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 $c_{ui}(f, \tau)$ and $c_{uj}(f, \tau)$ as latent variables. If no un-vibration-derived sounds are recorded by the vibration sensors, $R_{ui}(f)$ can be expressed as follows.

$$R_{ui}^{prior}(f) = \begin{bmatrix} R_{ui}^{M_{mic} \times M_{mic}}(f) & 0_{M_{mic} \times M_{vib}} \\ 0_{M_{vib} \times M_{mic}} & 0_{M_{vib} \times M_{vib}} \end{bmatrix}$$

(9)

$$R_{uj}^{M_{mic} \times M_{mic}}(f) = a_{ui}^{mic}(f)a_{ui}^{micH}(f)$$

(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 $R_{ui}(f)$ as the inverse-Wishart distribution as follows:

$$P_{R_{ui}^{prior}(f)}(f) = \frac{\left|R_{ui}^{prior}(f)\right|^d}{\pi^{\frac{d(d+M_{mic}+M_{vib})}{2}}} e^{-\frac{1}{2}tr(R_{ui}^{prior}(f)R_{ui}^{-1}(f))}$$

(11)

where $d$ is a parameter of the inverse-Wishart distribution that determines the degrees of freedom [13], and $\Gamma$ denotes the gamma function.
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 $W_{u_i}(f, \tau)$ and $W_{v_j}(f, \tau)$ for separating $i$-th un-vibration-derived sound signal $c_{u_i}(f, \tau)$ and $j$-th vibration-derived sound signal $c_{v_j}(f, \tau)$, respectively, are obtained by the following equations.

$$R_{c_{u_i}}(f, \tau) = p_{u_i}^{(n-1)}(f, \tau)R_{u_i}^{(n-1)}(f)$$  \hspace{1cm} (12)

$$R_{c_{v_j}}(f, \tau) = p_{v_j}^{(n-1)}(f, \tau)R_{v_j}^{(n-1)}(f)$$  \hspace{1cm} (13)

$$R_{x}(f, \tau) = \sum_{i} R_{c_{u_i}}(f, \tau) + \sum_{j} R_{c_{v_j}}(f, \tau)$$  \hspace{1cm} (14)

$$W_{u_i}(f, \tau) = R_{c_{u_i}}(f, \tau)R_{x}^{-1}(f, \tau)$$  \hspace{1cm} (15)

$$W_{v_j}(f, \tau) = R_{c_{v_j}}(f, \tau)R_{x}^{-1}(f, \tau)$$  \hspace{1cm} (16)

By using the estimated filters $W_{u_i}(f, \tau)$ and $W_{v_j}(f, \tau)$, the separated signals $c_{u_i}(f, \tau)$ and $c_{v_j}(f, \tau)$ are estimated as follows.

$$\hat{c}_{u_i}(f, \tau) = W_{u_i}(f, \tau)x(f, \tau)$$  \hspace{1cm} (17)

$$\hat{c}_{v_j}(f, \tau) = W_{v_j}(f, \tau)x(f, \tau)$$  \hspace{1cm} (18)

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

$$\hat{R}_{c_{u_i}}(f, \tau) = \hat{c}_{u_i}(f, \tau)\hat{c}_{u_i}^{H}(f, \tau) + (I - W_{u_i}(f, \tau))R_{c_{u_i}}(f, \tau)$$  \hspace{1cm} (19)

$$\hat{R}_{c_{v_j}}(f, \tau) = \hat{c}_{v_j}(f, \tau)\hat{c}_{v_j}^{H}(f, \tau) + (I - W_{v_j}(f, \tau))R_{c_{v_j}}(f, \tau)$$  \hspace{1cm} (20)

Here, $I$ 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.

$$p_{u_i}^{(n)}(f, \tau) = \frac{1}{\gamma_{d+M_{mic}+M_{ vib}}}tr(R_{c_{u_i}}(f, \tau))$$  \hspace{1cm} (21)

$$p_{v_j}^{(n)}(f, \tau) = \frac{1}{\gamma_{d+M_{mic}+M_{ vib}}}tr(R_{c_{v_j}}(f, \tau))$$  \hspace{1cm} (22)

$$R_{u_i}^{(\alpha, \beta)}(f, \tau) = \begin{cases} \frac{1}{L} \sum_{\tau} R_{c_{u_i}}(f, \tau) & (\alpha, \beta) < M_{mic} \\ \frac{1}{L} \sum_{\tau} R_{c_{u_i}}(f, \tau) & (\alpha, \beta) \geq M_{mic} \end{cases}$$  \hspace{1cm} (23)

$$R_{v_j}^{(\alpha, \beta)}(f, \tau) = \frac{1}{L} \sum_{\tau} R_{c_{v_j}}(f, \tau)$$  \hspace{1cm} (24)

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


