Acoust. Sci. & Tech. **38**, 4 (2017)

INVITED REVIEW

Estimating power spectral density for spatial audio signal separation: An effective approach for practical applications

Yusuke Hioka[1,*] and Kenta Niwa[2,†]

[1]*Department of Mechanical Engineering, University of Auckland, 20 Symonds Street, Auckland, 1142 New Zealand*
[2]*NTT Media Intelligence Laboratories, NTT Corporation, 3–9–11 Midori-cho, Musashino, 180–8585 Japan*

Abstract: The audio signal separation has been extensively studied due to its wide variety of applications. Along with the advancement of processor calculation performance in the last a couple of decades, the research focus has also extended to realising effective and feasible signal separation in practical environments where the problem becomes more challenging. In addition to classical linear filtering techniques, manipulating the spectral amplitude of a signal by applying a postfilter such as the Wiener filter is known to be an effective approach for practical applications. However the postfilter calculation requires the power spectral densities (PSD) of the signals of interest to be estimated beforehand. This article overviews methods for estimating the PSD of signals using spatial characteristics of their sources. Several practical applications that utilise the estimated PSD are introduced with some experimental results demonstrating the potential of the approach for solving challenging problems in practical applications.

Keywords: Power spectral density, Audio signal separation, Microphone array, Wiener filter

PACS number: 43.60.Fg, 43.60.Ac, 43.55.Mc [doi:10.1250/ast.38.175]

1. INTRODUCTION

Audio signal separation is one of the areas in digital signal processing that has been extensively studied for several decades. Despite its wide breadth of potential applications, it is only the last decade that some of these technologies started to appear in commercial products and services for consumers. One reason for causing such a long gap would have been a lack of effective algorithms that did not require an extensive amount of processor speed and memory space. Thus, research on pursuing an effective and cost efficient algorithm in practical use is still an ongoing problem in the field.

Of the various approaches that have been taken to solve the problem, use of microphone arrays [1] for signal observation has been a popular approach because of its capability to utilise spatial properties of audio signals as clues for separating the signals. Beamforming [2] must be the most common technique used with a microphone array; it is able to "form" a directivity, which can be electronically steered towards an angle where targeted sound sources are located. There have been various beamforming techniques, but most of them can be classified either into fixed or adaptive beamforming. With fixed beamforming, the directivity of the beamforming is fixed once it is designed and its performance does not depend on the received signals so that it is robust to the changes of environment. In other words, performance of fixed beamforming may be limited in practice because it does not reflect the environmental changes to its design. On the other hand, the directivity of the beamforming "adapts" to the environment with the adaptive beamforming, which ideally outperforms the fixed beamforming in terms of the signal separation performance; however, in practice its performance is often degraded due to various deviation from the modelling. Thus, the performance of audio signal separation using beamforming often needs to be reinforced by some means in practical applications.

To this end applying a post-filter to the output of a beamforming has appeared to be a common approach in practical implementations. The idea of using a post-filter in combination with the beamforming was introduced by Zelinski [3], which originally aimed at reducing noise that could not be effectively removed by the beamforming. Later study revealed that applying a Wiener post-filter to

*e-mail: yusuke.hioka@ieee.org
†e-mail: niwa.kenta@lab.ntt.co.jp

the output of minimum variance distortion-less response (MVDR) beamforming [2] would provide an optimum solution for the signal separation problem with minimum mean square error (MMSE) sense [4]. Although this is a very effective approach for audio signal separation, the key challenge in this approach is that the power spectral density (PSD) of the target signal needs to be estimated in order to derive the Wiener post-filter. Thus, accurate estimation of the PSD would be the key contributor to the success of signal separation.

Some clues are needed in order to estimate the PSD of the target signal. A commonly used and probably the most straightforward approach would be the use of temporal properties of signals [5]. Many audio signals in real world have different temporal properties to some extent. For instance, speech is known as a non-stationary signal, i.e. its properties vary within milliseconds, whereas ambient noise such as fan and duct noise produced by HAVC systems does not change its properties for minutes or even hours. The approach utilises such differences in temporal properties of signals to separate the PSD of the target signal from that of others. However, such methods will not be effective when signals have the same or similar temporal properties.

Given that a microphone array has already been used for the signal observation, it is natural to utilise the spatial properties of the signals as a clue for the PSD estimation too. In fact, some early studies have already taken this approach and have succeeded in estimating the PSD of *incoherent* signals such as microphones' internal noise [3] and reverberation [6]. However, this approach had never been applied to estimate the PSD of *coherent* signals because theoretically beamformers were supposed to remove coherent signals completely so that there was no need for the post-filter to take it into account. Unfortunately, it is no longer the case in practical problems because the effect of beamformers is limited. This article introduces a framework for PSD estimation, also known as *PSD estimation in beamspace*, which is able to estimate the PSD of signals having different types of spatial properties. The method has been implemented to various applications, some of which are also briefly introduced in this article.

The rest of this article is organised as follows. Some fundamentals of audio signal separation using PSD of sound sources are reviewed in Sect. 2. In Sect. 3, the framework proposed by the authors known as PSD estimation in beamspace and its expansions are introduced. Various example applications using the PSD estimation method are presented in Sect. 4. Finally, the article is concluded with some comments and discussions in Sect. 5.

2. AUDIO SIGNAL SEPARATION USING PSD

Prior to having a detailed discussion about the PSD estimation methods, this section briefly reviews the definition of the PSD of a stochastic signal and the way it is used to separate audio signals.

2.1. PSD of Stochastic Signals

As covered in most signal processing textbooks, the PSD refers to the distribution of a signal's spectral energy appearing in a certain period of time. Let $x(t)$ be a real-valued stochastic signal observed at time t. The PSD of $x(t)$ is defined by the Fourier transform of its auto-correlation $\gamma_x(t') := E[x(t)x(t-t')]$ where t' is the lag and $E[\cdot]$ denotes an expectation operation, which can be replaced by simple time average provided that the signal follows an ergodic process:

$$\phi_x(\omega) = \int_{t'=-\infty}^{\infty} \gamma_x(t') e^{-j\omega t'} dt' \tag{1}$$

where ω denotes the angular frequency. According to the properties of Fourier transform, the PSD can also be derived from the signal's spectrum $X(\omega)$ (i.e. Fourier transform of $x(t)$) as

$$\phi_x(\omega) = E[X(\omega)X^*(\omega)] = E[|X(\omega)|^2] \tag{2}$$

where * represents the complex conjugate.

2.2. Approximation of PSD Using STFT

Although so far signals have been defined in continuous time, most signal processing algorithms are implemented to digital processors. In the rest of this article, signals are assumed to be in discrete time, namely t and ω should be read as the sample index and frequency bin, respectively.

There is a need to introduce some approximations for the PSD calculated from a discrete-time signal because the number of samples is limited. A commonly used method for the approximation is the use of the short-time Fourier transform (STFT). With this technique, the expectation operation in (2) is replaced by frame averaging of the STFT of $x(t)$, i.e. $X(\omega, \tau)$, as in (3), which is also denoted as (4) in the rest of this article.

$$\phi_x(\omega) \approx \frac{1}{T} \sum_{\tau=0}^{T-1} |X(\omega, \tau)|^2 \tag{3}$$

$$:= E[|X(\omega, \tau)|^2]_\tau. \tag{4}$$

Here τ is the frame index of the STFT and T is the number of frames to be averaged.

Further approximations may be necessary if the signal is non-stationary, i.e. the PSD of the signal varies as a function of τ. Such approximation can be realised by (5), which is also known as Welch's method [7]

$$\phi_x(\omega, \tau) \approx \alpha \phi_x(\omega, \tau - 1) + (1 - \alpha)|X(\omega, \tau)|^2, \tag{5}$$

where $0 < \alpha < 1$ is the forgetting factor. Discussions in the rest of this article will assume that signals are non-stationary unless otherwise specified.

2.3. Audio Signal Separation Using Wiener Filter

Although the post-filter may be designed by any means, the Wiener filter is the most commonly used design of the post-filter. Assume that a signal is modelled by a superposition of two signals in the time-frequency domain, i.e. $X(\omega,\tau) = S(\omega,\tau) + N(\omega,\tau)$. Note that $S(\omega,\tau)$ and $N(\omega,\tau)$ are often referred as a target signal and interfering noise, respectively. If one wishes to separate $S(\omega,\tau)$ from the mixture, the Wiener filter gain defined as

$$H(\omega,\tau) = \frac{\phi_S(\omega,\tau)}{\phi_S(\omega,\tau) + \phi_N(\omega,\tau)} \tag{6}$$

will be multiplied to $X(\omega)$, providing the output signal as

$$Z(\omega,\tau) = H(\omega,\tau)X(\omega,\tau). \tag{7}$$

Finally, the output signal in the time-domain is obtained by applying the inverse short-time Fourier transform to $Z(\omega,\tau)$.

When the target signal and interfering noise are mutually uncorrelated, i.e. $E[S^*(\omega,\tau)N(\omega,\tau)]_\tau = 0$, the denominator of (6) can be replaced by the PSD of the input signal, i.e. $\phi_X(\omega,\tau)$. By looking at the definition, it is obvious that estimating the PSD of the target signal ($\phi_S(\omega,\tau)$ in (6)) is the key challenge when the Wiener filter is utilised. An algorithm for estimating the PSD using spatial properties of the sound sources and their practical applications will be introduced in the following sections.

3. PSD ESTIMATION USING SPATIAL PROPERTY

Although there are several approaches for estimating the PSD of the target signal as referred to in Sect. 1, this article particularly focuses on the approach using microphone arrays in order to exploit spatial properties of signals. This section reviews the framework known as *beamforming with Wiener post-filter* with its problem setup, then introduces a PSD estimation method for designing the Wiener post-filter.

3.1. Problem Setup

Assume that a microphone array consisting of M microphones observes audio signals. The observed signal of the m-th microphone is modelled as

$$X_m(\omega,\tau) = \sum_{k=1}^{K} A_{m,k}(\omega)S_k(\omega,\tau) + V_m(\omega,\tau), \tag{8}$$

where $A_{m,k}(\omega)$ denotes the transfer function between the m-th ($m = 1,\ldots,M$) microphone and k-th ($k = 1,\ldots,K$) sound source. The model classifies the signals in the environment by their spatial properties; namely (spatially) *coherent* and *incoherent* signals [2], denoted in the STFT domain by $S_k(\omega,\tau)$ and $V_m(\omega,\tau)$, respectively. Coherent signals observed by microphones located at two close positions (e.g. any pairs of the microphones in the array) normally show a high coherence (correlation) between the microphones. Signals emitted from a sound source that propagate to a microphone directly usually fall into this category. On the other hand incoherent signals do not show such coherence between microphones. Room reverberation and microphones' internal noise are normally classified into this category.

It is also assumed that signals in the model are uncorrelated with each other, namely

$$E[S_k(\omega,\tau)S^*_{k'}(\omega,\tau)]_\tau = 0 \quad \forall k, k \neq k',$$
$$E[V_m(\omega,\tau)V^*_{m'}(\omega,\tau)]_\tau = 0 \quad \forall m, m \neq m',$$
$$E[S_k(\omega,\tau)V^*_m(\omega,\tau)]_\tau = 0 \quad \forall k, \forall m.$$

For brevity the observed signals defined in (8) for all microphones of the array are often represented in a vector form given by

$$\boldsymbol{x}(\omega,\tau) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega,\tau) + \boldsymbol{v}(\omega,\tau), \tag{9}$$

where the matrix and vectors in (9) are defined as follows:

$$\boldsymbol{A}(\omega) := [\boldsymbol{a}_1(\omega),\ldots,\boldsymbol{a}_K(\omega)], \tag{10}$$
$$\boldsymbol{a}_k(\omega) := [A_{1,k}(\omega),\ldots,A_{M,k}(\omega)]^{\mathrm{T}},$$
$$\boldsymbol{x}(\omega,\tau) := [X_1(\omega,\tau),\ldots,X_M(\omega,\tau)]^{\mathrm{T}}, \tag{11}$$
$$\boldsymbol{s}(\omega,\tau) := [S_1(\omega,\tau),\ldots,S_K(\omega,\tau)]^{\mathrm{T}}, \tag{12}$$
$$\boldsymbol{v}(\omega,\tau) := [V_1(\omega,\tau),\ldots,V_M(\omega,\tau)]^{\mathrm{T}}. \tag{13}$$

Note that $^{\mathrm{T}}$ denotes the transpose operator.

3.2. Beamforming with Wiener Post-filter

Beamforming is the most commonly used technique for audio signal separation when microphone arrays are utilised for signal observation. In order to boost the performance of audio signal separation, beamforming is often combined with a Wiener post-filter discussed in Sect. 2, which is known as *Beamforming with Wiener post-filtering* [3]. In this framework, beamforming is first applied to the microphone observation, the output signal of which is given by

$$Y_l(\omega,\tau) = \boldsymbol{w}_l^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,\tau), \tag{14}$$

where

$$\boldsymbol{w}_l(\omega) = [W_{l,1}(\omega),\ldots,W_{l,M}(\omega)]^{\mathrm{T}}, \tag{15}$$

and $^{\mathrm{H}}$ denotes the Hermitian transpose. The weights of the beamforming $\boldsymbol{w}_l(\omega)$ are designed to emphasise sound arriving from θ_l with, e.g. the minimum variance distortionless response (MVDR) method [8].

To boost noise-reduction performance, the Wiener filter gain $H(\omega,\tau)$ given by (6) is then applied to the output of the beamforming as

$$Z(\omega,\tau) = H(\omega,\tau)Y_l(\omega,\tau). \tag{16}$$

3.3. PSD Estimation in Beamspace

To calculate the Wiener filter gain, the PSDs of the target signal and interfering noise need to be estimated from the microphone array observation. As modelled in (8), the interfering noise consists of $K-1$ spatially *coherent* signals and a spatially *incoherent* signal. *PSD estimation in beamspace* [9] was originally invented to separate the coherent signals and the target signal, however, it was later extended to cope with incoherent signals, too. The rest of this section discusses the details of the PSD estimation method.

Let $L\ (\geq K)$ beamformers, which focus their directivity on different angles, be applied for microphone array observation. As signals are assumed to be uncorrelated with each other, the PSD of the l-th beamforming output $\phi_{Y_l}(\omega,\tau)$ can be approximated by an affine transformation of the PSDs of each source $\phi_{S_k}(\omega,\tau)$ with the directivity gain of the l-th beamformer to the k-th source direction $|D_{l,k}(\omega)|^2$ and noise PSD, as defined by

$$\underbrace{\begin{bmatrix}\phi_{Y_1}\\ \vdots \\ \phi_{Y_L}\end{bmatrix}}_{\boldsymbol{\Phi}_Y(\omega,\tau)} \approx \underbrace{\begin{bmatrix}|D_{1,1}|^2 & \dots & |D_{1,K}|^2\\ \vdots & \ddots & \vdots \\ |D_{L,1}|^2 & \dots & |D_{L,K}|^2\end{bmatrix}}_{\boldsymbol{D}(\omega)} \underbrace{\begin{bmatrix}\phi_{S_1}\\ \vdots \\ \phi_{S_K}\end{bmatrix}}_{\boldsymbol{\Phi}_S(\omega,\tau)} + \underbrace{\begin{bmatrix}\phi_{\tilde{V}_1}\\ \vdots \\ \phi_{\tilde{V}_L}\end{bmatrix}}_{\boldsymbol{\Phi}_{\tilde{V}}(\omega,\tau)}, \quad (17)$$

where $\phi_{\tilde{V}_l}(\omega,\tau)$ denotes the PSD of incoherent signals in the output of the l-th beamformer, i.e. $\phi_{\tilde{V}_l}(\omega,\tau) = E[|\sum_{m=1}^{M} W_{l,m}^*(\omega)V_m(\omega,l)|^2]_\tau$. Note that the indices of ω and τ are omitted for brevity in (17).

Assuming that the source positions are known *a priori*, if the transfer function $A_{m,k}(\omega,\tau)$ can be modelled by any means such as using the array manifold vectors [1,2] or pre-measuring an impulse response between the source positions and the microphone array in a practical setup, the directivity gain is given by

$$D_{l,k}(\omega) = \sum_{m=1}^{M} W_{l,m}^*(\omega)A_{m,k}(\omega). \quad (18)$$

Thus, the PSD of each coherent sound signal can be separated by

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_{S+\tilde{V}}(\omega,\tau) &= [\hat{\phi}_{S_1+\tilde{V}}(\omega,\tau),\dots,\hat{\phi}_{S_K+\tilde{V}}(\omega,\tau)]^{\mathrm{T}}\\ &:= \boldsymbol{G}(\omega)\boldsymbol{\Phi}_Y(\omega,\tau)\\ &\approx \boldsymbol{\Phi}_S(\omega,\tau) + \boldsymbol{G}(\omega)\boldsymbol{\Phi}_{\tilde{V}}(\omega), \end{aligned} \quad (19)$$

where

$$\boldsymbol{G}(\omega) = \begin{cases}\boldsymbol{D}^{-1}(\omega) & L = K\\ \boldsymbol{D}^{+}(\omega) & L > K,\end{cases} \quad (20)$$

and $\hat{}$ and $^+$ denote an estimated value and the pseudo inverse, respectively.

Now let us simplify the problem by assuming the amount of incoherent signal be negligibly small, i.e. $\phi_{\tilde{V}_l}(\omega,\tau) = 0$ for all l. The relationship in (17) may be rearranged as

$$\boldsymbol{\Phi}_Y(\omega,\tau) \approx \boldsymbol{D}(\omega)\boldsymbol{\Phi}_S(\omega,\tau). \quad (21)$$

In such a scenario, the PSD of each coherent signal and the target signal can be separately estimated as

$$\hat{\boldsymbol{\Phi}}_{S+\tilde{V}}(\omega,\tau) \approx \boldsymbol{\Phi}_S(\omega,\tau), \quad (22)$$

which can be directly utilised for calculating the Wiener post-filter by setting $\phi_S(\omega,\tau)$ and $\phi_N(\omega,\tau)$ in (6) as $\phi_S(\omega,\tau) = \hat{\phi}_{S_1}(\omega,\tau)$ and $\phi_N(\omega,\tau) = \sum_{k=2}^{K}\hat{\phi}_{S_k}(\omega,\tau)$, respectively[1]. Details of the performance of the method can be found in [9]. Figure 1 shows the block diagram of audio signal separation algorithm using PSD estimation in beamspace.

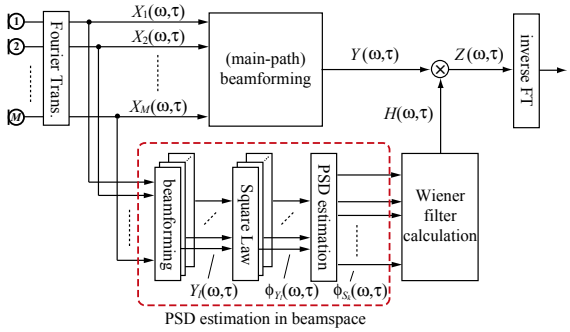

Fig. 1 Block diagram of audio signal separation algorithm using Wiener post-filter calculated using PSD estimation in beamspace.

3.4. Number of Separable Signals and Appropriate Design of Beamformers

One of the benefits of using the post-filtering is its potential for maintaining the signal separation performance even if the problem is *under-determined*, namely when the number of coherent signals exceeds the number of microphones available ($K > M$). Unlike other techniques such as beamforming, the post-filtering itself is a non-linear process; therefore, its performance would not be restricted by the number of microphones as long as the PSD of each signal were obtained. A study in [9] has revealed that the PSD of up to $M(M-1)+1$ coherent signals may be accurately estimated by the PSD estimation in beamspace. This maximum number of separable signals (MNSS) is subject to the beamformers used to calculate $D_{l,k}(\omega)$ in (17) being appropriately designed. Another study found that the MNSS would be reduced to $2M-1$ given that the cylindrical mode beamforming with a circular micro-

[1]Strictly speaking, gains of the l-th beamfomer have to be multiplied as $\phi_S = |D_{l,1}|^2\hat{\phi}_{S_1}$ and $\phi_N = \sum_{k=2}^{K}|D_{l,k}|^2\hat{\phi}_{S_k}$ when the Wiener post-filter is applied to the output of the l-th beamfoming as defined in (16).

phone array on a rigid cylinder is used for the beamformers [10].

As these studies show, appropriate selection of the beamformers is a key to guarantee the accuracy of the estimated PSDs. Another study pursued an appropriate design of beamformers for determined or over-determined cases, and suggests that the beamformers should be selected so as to make the inverse of $\mathbf{D}(\omega)$ be an M-matrix [11].

3.5. PSD Estimation of Incoherent Background Noise

When the simplified problem in (22) was introduced, the second term of (19), namely the components originating from the spatially incoherent signals were ignored. This naturally causes some errors in the estimated PSDs when the observed signal contains such incoherent signals. Thus, attempts should be made to extend the PSD estimation algorithm so as to take into account the incoherent signals. Two approaches have been taken to realise this: utilising spatial properties, and temporal properties, of incoherent signals.

3.5.1. Spatial properties

Although spatially incoherent signals observed at different locations show no correlation, in some cases they still have a pattern in their spatial properties. One case is the isotropical propagation, which could be seen in reverberation in an indoor environment [12]. If an isotropical propagation can be assumed, the incoherent signals can also be separated by modifying the model in (17) by adding one more column to $\boldsymbol{D}(\omega)$ as the $(K+1)$-th column, rearranging the simultaneous equation as

$$\underbrace{\begin{bmatrix}\phi_{Y_1}\\ \vdots \\ \phi_{Y_L}\end{bmatrix}}_{\boldsymbol{\Phi}_Y(\omega,\tau)} \approx \underbrace{\begin{bmatrix}|D_{1,1}|^2 & \dots & |D_{1,K}|^2 & \int_\theta |D_{1,\theta}|^2 d\theta \\ \vdots & \ddots & \vdots & \vdots \\ |D_{L,1}|^2 & \dots & |D_{L,K}|^2 & \int_\theta |D_{L,\theta}|^2 d\theta\end{bmatrix}}_{\boldsymbol{D}'(\omega)} \underbrace{\begin{bmatrix}\phi_{S_1}\\ \vdots \\ \phi_{S_K} \\ \phi_{\bar{V}}\end{bmatrix}}_{\boldsymbol{\Phi}_{S+\bar{V}}(\omega,\tau)}, \tag{23}$$

where $|D_{l,\theta}|^2$ is the directivity gain of the l-th beamformer to the angle and θ, $\phi_{\bar{V}}(\omega,\tau)$ is the PSD of the incoherent signals, and $L > K$ [13]. Solving this modified simultaneous equation will provide the PSD of incoherent signals separately from other signal components.

3.5.2. Temporal properties

Another available approach is using the differences in the temporal energy fluctuations of signals. In many practical applications the incoherent signals are stationary whereas the coherent signals such as speech and music are non-stationary. In such a scenario, the PSD of incoherent signals can be estimated by measuring the level of stationary components in the estimated PSD [14]. The stationary components in $\phi_{S_k+\bar{V}}(\omega,\tau)$ can be estimated by taking the minimum value of a temporary smoothed PSD in a given time interval Υ, as in

$$\hat{\phi}_{V_k}(\omega,\tau) = \min_{\tau\in\Upsilon}\{\bar{\phi}_{S_k+\bar{V}}(\omega,\tau)\}. \tag{24}$$

Here, $\bar{\phi}_{S_k+\bar{V}}(\omega,\tau)$ is calculated by applying a recursive update algorithm, as in

$$\bar{\phi}_{S_k+\bar{V}}(\omega,\tau) = \beta\hat{\phi}_{S_k+\bar{V}}(\omega,\tau) + (1-\beta)\bar{\phi}_{S_k+\bar{V}}(\omega,\tau-1), \tag{25}$$

where β denotes the forgetting factor, which is set so that its time constant is around 150 ms [14].

Given that the directivity of the first beamformer ($l=1$) points to the angle of the target source, the PSD of target source $\phi_S(\omega,\tau)$ is calculated by

$$\phi_S(\omega,\tau) = \hat{\phi}_{S_1+\bar{V}}(\omega,\tau) - \hat{\phi}_{V_1}(\omega,\tau). \tag{26}$$

Likewise the PSD of the interfering noise can be calculated using the PSD of other coherent signals and incoherent signal, as in

$$\phi_N(\omega,\tau) = \underbrace{\gamma(\omega)\left\{\sum_{k=2}^{K}(\hat{\phi}_{S_k+\bar{V}}(\omega,\tau) - \hat{\phi}_{V_k}(\omega,\tau))\right\}}_{\text{PSD of coherent signals}} + \underbrace{\hat{\phi}_{V_1}(\omega,\tau)}_{\text{PSD of incoherent signal}}, \tag{27}$$

where $\gamma(\omega)$ is a weighting parameter.

4. PRACTICAL APPLICATIONS

This section introduces several recent studies that applied the audio signal separation algorithm using the PSD estimation techniques discussed in Sect. 3 to various practical problems.

4.1. Sound Source Enhancement and Noise Reduction

The most common application area of audio signal separation would be sound source enhancement and noise reduction, which have been extensively studied in the last few decades. The problem aims to emphasise a target sound sources contaminated by various interfering noise.

4.1.1. Region-wise sound source enhancement

The majority of studies in sound source enhancement using microphone arrays attempt to emphasise a sound arriving from a particular angle while suppressing noise arriving from other angles. The study in [15] tailored the PSD estimation in beamspace to be implemented to a hands-free speakerphone system as shown in Fig. 2. As Fig. 3 shows, the method is able to reduce sounds arriving from a certain range of angles (called *angular region*) in order to enable the speakerphone to mute unwanted sounds in particular angular regions. For the sake of reducing computational complexity the algorithm rearranges (17) to be a 2×2 problem given by

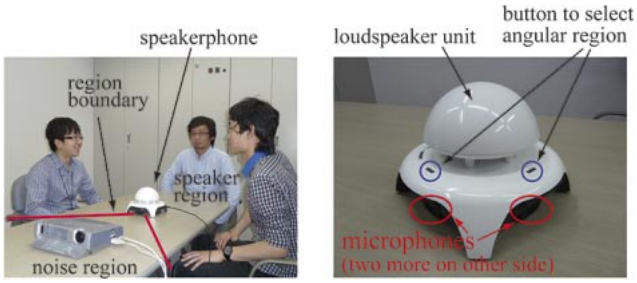

Fig. 2 Left: Angle-wise sound source enhancement implemented to a speakerphone system; Right: prototyped speakerphone [15].

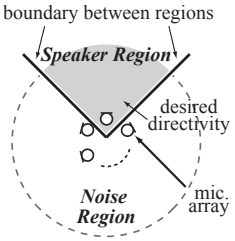

Fig. 3 Defined regions and desired shape of directivity [15].

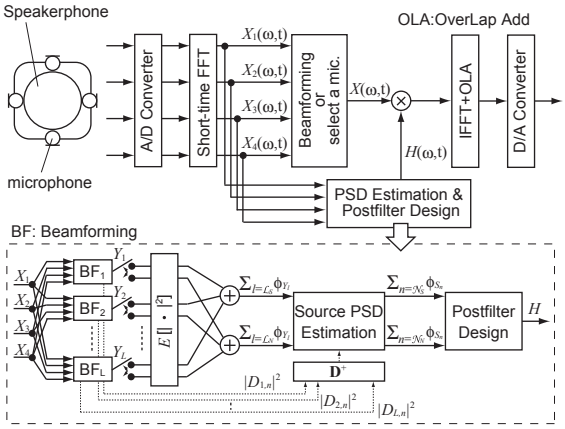

Fig. 4 Block diagram of the sound source enhancement algorithm implemented to the speakerphone system [15].

$$\underbrace{\begin{bmatrix} \sum_{l\in\mathcal{L}_S}\phi_{Y_l} \\ \sum_{l\in\mathcal{L}_N}\phi_{Y_l} \end{bmatrix}}_{\overline{\boldsymbol{\Phi}}_Y(\omega)} \approx \underbrace{\begin{bmatrix} |\overline{D}_{\mathcal{L}_S,\mathcal{N}_S}|^2 & |\overline{D}_{\mathcal{L}_S,\mathcal{N}_N}|^2 \\ |\overline{D}_{\mathcal{L}_N,\mathcal{N}_S}|^2 & |\overline{D}_{\mathcal{L}_N,\mathcal{N}_N}|^2 \end{bmatrix}}_{\overline{\boldsymbol{D}}(\omega)} \underbrace{\begin{bmatrix} \sum_{k\in\mathcal{N}_S}\phi_{S_k} \\ \sum_{k\in\mathcal{N}_N}\phi_{S_k} \end{bmatrix}}_{\overline{\boldsymbol{\Phi}}_S(\omega)}, \tag{28}$$

where $\mathcal{L}$ and $\mathcal{N}$ denote the sets of indices of angular regions to which each respective beamformer and sound source belong. By solving (28), the approximated sum of PSDs of sound signals arriving from each angular region is estimated. Figure 4 shows the block diagram of the sound source enhancement algorithm implemented to the speakerphone system.

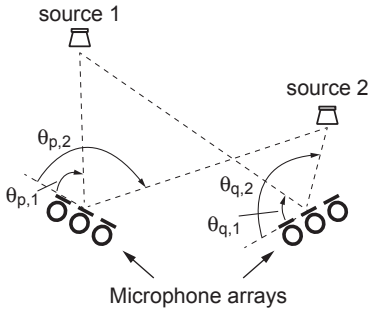

Fig. 5 An example case for 2-dimensional audio signal separation.

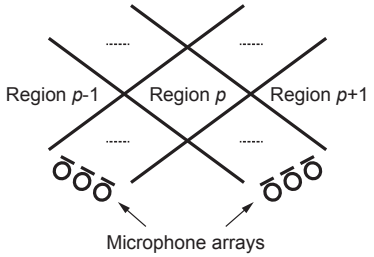

Fig. 6 Definition of 2-dimensional regions using two microphone arrays.

PSD estimation in beamspace may also be extended for solving 2-dimensional audio signal separation problems, i.e. separating signals originating from different *positions*. A study achieved this by using two sets of microphone arrays placed apart from each other [16]. With this setup a position can be defined by the combination of angles from each microphone array; similar to the idea of triangulation. Figure 5 shows the simplest case as an example; a beamforming is applied to the observation of each microphone array. Given that the PSD of a sound source located at the r-th position is denoted by $\phi_{S_r}(\omega,\tau)$, the PSD of the beamformers' outputs are described as

$$\begin{bmatrix} \phi_{Y_p} \\ \phi_{Y_q} \end{bmatrix} \approx \begin{bmatrix} |D_{p,\theta_{p,1}}|^2 & |D_{p,\theta_{p,2}}|^2 \\ |D_{q,\theta_{q,1}}|^2 & |D_{q,\theta_{q,2}}|^2 \end{bmatrix} \begin{bmatrix} \phi_{S_1} \\ \phi_{S_2} \end{bmatrix}, \tag{29}$$

where $D_{p,\theta_{p,r}}(\omega)$ is the directivity gain of the beamformer applied to the p-th microphone array to the angle of the r-th sound source $\theta_{p,r}$. By solving (29) in the same manner as that for the 1-dimensional case, the PSD of sound sources located at each position will be estimated. The number of positions may be increased and treated as 2-dimensional regions, as shown in Fig. 6, when a greater number of microphone arrays are utilised. More details, including the maximum number of regions that the method is able to estimate PSDs simultaneously, are presented in [16].

4.1.2. Combination with optimum microphone array design

Although the sound source enhancement using PSD estimation in beamspace is already practically effective, its

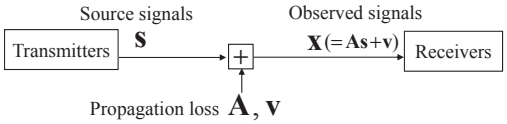

Fig. 7 Problem definition of array structure optimization.

performance may be further improved when it is used with microphone array hardware that is designed in an optimum way [17–19]. An example study that incorporated the sound source enhancement algorithm into such a microphone array and a brief summary of its idea are introduced.

The study regards the model in (8) as information transmission as shown in Fig. 7, which is an analogy of a model used in wireless communication. It is based on a hypothesis that obtaining more information about the sources in the microphone array observation by devising the array structure would help achieving better sound source enhancement performance with any algorithms. To this end the microphone array structure is optimised in such a way that the mutual information between the sound sources and microphone array observation defined below in (30) is maximised

$$I(\boldsymbol{s};\boldsymbol{x}) = H(\boldsymbol{s}) - H(\boldsymbol{s}|\boldsymbol{x}), \tag{30}$$

where $H(\boldsymbol{s})$ and $H(\boldsymbol{s}|\boldsymbol{x})$ denote the marginal entropy of $\boldsymbol{s}(\omega,\tau)$ and the conditional entropy when $\boldsymbol{x}(\omega,\tau)$ is known, respectively. The information loss $H(\boldsymbol{s}|\boldsymbol{x})$ is increased when $\boldsymbol{A}(\omega)$ is not regularised or the background noise level is substantially large. When the microphone array structure is optimised so that $\boldsymbol{A}(\omega)$ is modified to maximise $I(\boldsymbol{s};\boldsymbol{x})$, effective spatial cues for estimating source signals would be included in $\boldsymbol{x}(\omega,\tau)$. Based on this theory, a large microphone array shown in Fig. 8 was constructed by sequentially placing 96 (8 for each reflector) microphones around the focal point of parabolic reflectors so as to increase $I(\boldsymbol{s};\boldsymbol{x})$ [17].

The proposed audio signal separation algorithm was applied to $\boldsymbol{x}(\omega,\tau)$ observed by this microphone array. Various positions in a room where sound sources are likely to be placed were determined. The impulse responses from these positions to the microphones were then premeasured in order to design the beamformers used in the algorithm. The system succeeded in separating two sound sources located 16.5 m away from the microphone array but were separated only 50 cm apart from each other (i.e. spatial resolution of the directivity was approximately 3.0 degrees) [17].

4.1.3. Audio recording using UAV

An example of the PSD estimation methods discussed in this paper being implemented to a unmanned aerial vehicles (UAV) [20] is introduced. The study was aimed to

(a) Microphone array structure

(b) Arrangement of sound sources and array

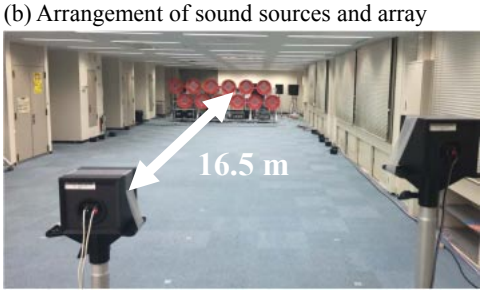

Fig. 8 Constructed microphone array [4.0 m (W) × 1.5 m (H) × 1.0 m (D)] [17].

Fig. 9 Prototyped UAV system carrying a microphone array [20].

record high quality audio from a UAV by minimising the level of rotor noise and other interfering sounds. Figure 9 shows a prototyped UAV carrying a microphone array consisting of six microphones, four of which are utilised to focus on a target signal while the others mainly extract the noise generated by the rotors of the UAV. By applying several beamformers using these microphones, the voice of a target speaker was able to be clearly recorded [20]. A video clip that demonstrates the quality of audio signals recorded by the prototyped UAV is available from the link [21].

4.2. Source Separation for Virtual Reality Audio Rendering

As another application, the audio signal separation algorithm has been used for audio rendering in virtual reality (VR) systems. VR enables users to experience as if they were "immersed" in a remote or virtual place by projecting a 360° video through head mount displays (HMDs) [22,23]. In many existing VR systems/services, audio signals are presented through a headphone without being processed so that the signals are perceived as their sources being located in fixed positions even though the

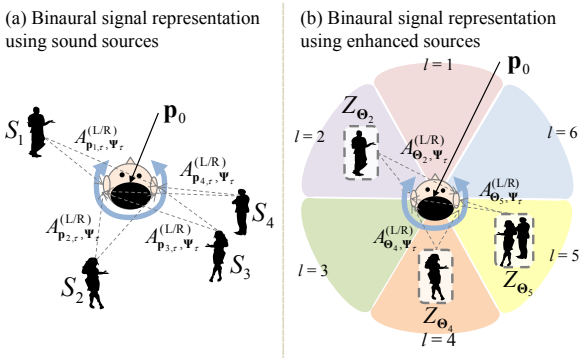

Fig. 10 Binaural signal representation using (a) K (= 4) sound sources and (b) J (= 6) region-enhanced sources [22].

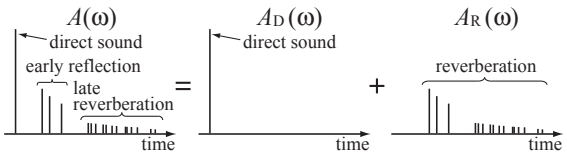

Fig. 11 Calculation of DRR from a room impulse response.

user's view seamlessly varies corresponding to the user's looking direction. Some recent studies attempted to match the visual and auditory locations of an object by projecting semi-binaural signals [22]. For generating the semi-binaural signals, the audio signals arriving from several different angular regions have to be acquired separately, which was achieved by utilising the region-wise source enhancement discussed in Sect. 4.1.1. but with several angular regions.

As depicted in Fig. 10(a), a 360° camera is assumed to be located at a fixed position surrounded by sound sources which may move from time to time. A user located at the camera's position will hear the audio signal of each sound source from its direction by projecting binaural signals consisting of the user's head related transfer function (HRTF) with respect to the sound source position. However, the signal of each sound source needs to be acquired separately for generating the binaural signals. The study simplified the problem by quantising the acoustic field into J angular regions as defined in Fig. 10(b). Since the HRTF does not drastically change when the source position slightly moves, the method only uses the HRTF of a specific position which "represents" the angular region it belongs to. Assuming that the acoustic field is divided into J angular regions and their representative angles are defined by $\boldsymbol{\Theta}_j$ $(j = 1, \ldots, J)$, K sound sources are grouped into J *virtual* sources. Given that the HRTFs with respect to the angle $\boldsymbol{\Theta}_j$ for the user's left and right ears are denoted by $A^{(\mathrm{L})}_{\boldsymbol{\Theta}_j,\boldsymbol{\Psi}_\tau}(\omega)$ and $A^{(\mathrm{R})}_{\boldsymbol{\Theta}_j,\boldsymbol{\Psi}_\tau}(\omega)$, respectively, the semi-binaural signals are calculated by

$$B^{(\mathrm{L})}(\omega,\tau) \approx \sum_{j=1}^{J} A^{(\mathrm{L})}_{\boldsymbol{\Theta}_j,\boldsymbol{\Psi}_\tau}(\omega)\, Z_{\boldsymbol{\Theta}_j}(\omega,\tau), \tag{31}$$

$$B^{(\mathrm{R})}(\omega,\tau) \approx \sum_{j=1}^{J} A^{(\mathrm{R})}_{\boldsymbol{\Theta}_j,\boldsymbol{\Psi}_\tau}(\omega)\, Z_{\boldsymbol{\Theta}_j}(\omega,\tau), \tag{32}$$

where the user's looking angle at frame τ is represented by $\boldsymbol{\Psi}_\tau$ and $Z_{\boldsymbol{\Theta}_j}(\omega,\tau)$ denotes the audio signal arriving from $\boldsymbol{\Theta}_j$ separated by the audio signal separation algorithm. Subjective evaluations confirmed that this technique succeeded in matching auditory and vidual localization [22]. A further study also realised a real-time 360° video streaming/rendering system using smartphones [24].

4.3. Blind Estimation of DRR

All applications introduced so far somehow use the estimated PSD for sound source separation or enhancement, however, they are not the only application of the estimated PSDs. The method could be applied to any applications that need to separate audio signals where the spatial properties of their propagations can be utilised. Attempts have been made to estimate direct-to-reverberant ratio (DRR), which is an important parameter in room acoustics, in a blind manner by estimating the PSD of direct sound and reverberation [25–29]. DRR is defined as the energy ratio between direct sound and reverberation observed in a reverberant environment. As Fig. 11 shows, an impulse response measured in a reverberant room can be split into its direct sound and reverberation components. The DRR is defined by the ratio of these two components formulated as

$$\mathrm{DRR\ (dB)} = 10\log_{10}\frac{\sum_\omega |A_{\mathrm{D}}(\omega)|^2}{\sum_\omega |A_{\mathrm{R}}(\omega)|^2}. \tag{33}$$

Blind estimation of DRR is a problem that aims to estimate the DRR directly from an arbitrary recording in the room but without measuring the room impulse response.

This can actually be achieved by estimating the PSD of the direct sound and reverberation separately, which is realised by modifying the methods discussed in Sect. 3. Let us revisit the model introduced in (23) where the propagation of incoherent signals were assumed to be isotropical. The same assumption could actually be imposed for the reverberation components if the room is diffuse [12] as described in Fig. 12. Thus a DRR estimation method reported in [28] estimates the PSD of direct sound and reverberation by utilising two beamformers as shown in Fig. 13, which provides the following equation:

$$\begin{bmatrix}\phi_{Y_1}(\omega)\\ \phi_{Y_2}(\omega)\end{bmatrix} \approx \begin{bmatrix}|D_{1,\theta_{\mathrm{D}}}(\omega)|^2 & \int_\theta |D_{1,\theta}(\omega)|^2 d\theta\\ |D_{2,\theta_{\mathrm{D}}}(\omega)|^2 & \int_\theta |D_{2,\theta}(\omega)|^2 d\theta\end{bmatrix}\begin{bmatrix}\phi_{\mathrm{D}}(\omega)\\ \phi_{\mathrm{R}}(\omega)\end{bmatrix}. \tag{34}$$

Here $D_{p,\theta_{\mathrm{D}}}(\omega)$ is the directivity gain of the p-th $(p = 1, 2)$

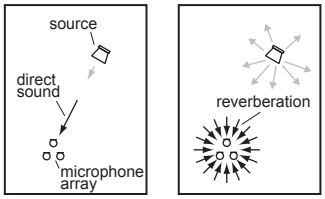

Fig. 12 Propagation path from sound source to microphone array in reverberant room: (a) direct sound, (b) reverberation [28].

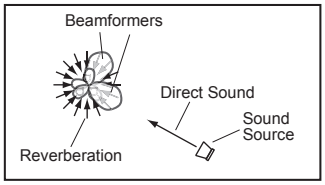

Fig. 13 PSD estimation in beamspace for estimating DRR. Two different beamformers are applied to the microphone array observation to create the beamspace for estimating the DRR [28].

beamformer towards the sound source located in the angle θ_{D}, and $\phi_{\mathrm{D}}(\omega)$ and $\phi_{\mathrm{R}}(\omega)$ denote the PSD of the direct sound and reverberation, respectively. Once these PSDs are estimated the DRR is calculated by

$$\mathrm{DRR\ (dB)} = 10\log_{10}\frac{\sum_{\omega}\phi_{\mathrm{D}}(\omega)}{\sum_{\omega}\phi_{\mathrm{R}}(\omega)}. \tag{35}$$

More details of this method and its performance examined by the ACE (Acoustic Characterisation of Environment) Challenge corpus [29] can be found in [28,29].

5. CONCLUSION

This article has overviewed signal processing techniques for spatial audio signal separation using microphone arrays that are effective in practical applications. The techniques are based on the framework of beamforming with post-filtering, which requires the PSD of sound signals to be estimated. PSD estimation in beamspace, a method for estimating the PSDs used to calculate Wiener post-filter, and its various extensions were introduced. Examples that applied the method to various practical applications have also been presented.

Some problems are still open to future studies. Pursuing an optimum design of beamformers for accurate PSD estimation will add extra values to the framework. Reconfiguring the model between the PSD of beamformers' output and source signals using a non-linear neural network presented e.g. in [30] would be another approach to improve the PSD estimation accuracy using the proposed framework.

REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, Berlin/Heidelberg, 2001).

[2] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques* (Simon & Schuster, Englewood Cliffs, NJ, 1993).

[3] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," *IEEE ICASSP 1988*, Vol. 5, pp. 2578–2581 (1988).

[4] K. Simmer, J. Bitzer and C. Marro, *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. (Springer, Berlin/Heidelberg/New York, 2001), Chap. 3, pp. 39–60.

[5] R. Martin, *Statistical Methods for the Enhancement of Noisy Speech* (Springer, Berlin/Heidelberg, 2005), pp. 43–65.

[6] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, **11**, 709–716 (2003).

[7] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, **15**, 70–73 (1967).

[8] H. Cox, R. M. Zeskind and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust. Speech Signal Process.*, **35**, 1365–1376 (1987).

[9] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. Audio Speech Lang. Process.*, **21**, 1240–1250 (2013).

[10] Y. Hioka and T. Betlehem, "Under-determined source separation based on power spectral density estimated using cylindrical mode beamforming," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2013* (*WASPAA 2013*), Oct. (2013).

[11] K. Niwa, T. Kawase, K. Kobayashi and Y. Hioka, "PSD estimation in beamspace using property of M-matrix," *IEEE Int. Workshop on Acoustic Signal Enhancement 2016* (*IWAENC 2016*), Sep. (2016).

[12] H. Kuttruff, *Room Acoustics*, 5th ed. (Applied Science Publishers Ltd., London, 1973), Chap. 2.

[13] Y. Hioka and K. Niwa, "PSD estimation in beamspace for source separation in a diffuse noise field," *IEEE Int. Workshop on Acoustic Signal Enhancement 2014* (*IWAENC 2014*), Sep. (2014).

[14] K. Niwa, Y. Hioka and K. Kobayashi, "Post-filter design for speech enhancement in various noisy environments," *IEEE Int. Workshop on Acoustic Signal Enhancement 2014* (*IWAENC 2014*), Sep., pp. 35–39 (2014).

[15] Y. Hioka, K. Furuya, K. Kobayashi, S. Sakauchi and Y. Haneda, "Angular region-wise speech enhancement for hands-free speakerphone," *IEEE Trans. Consum. Electron.*, **58**, 1403–1410 (2012).

[16] Y. Hioka, K. Kobayashi, K. Furuya and A. Kataoka, "Enhancement of sound sources located within a particular area using a pair of small microphone arrays," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, **E91-A**, 561–574 (2008).

[17] K. Niwa, Y. Hioka and K. Kobayashi, "Optimal microphone array observation for clear recording of distant sound sources," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **24**, 1785–1795 (2016).

[18] K. Niwa, Y. Hioka, K. Furuya and Y. Haneda, "Diffused sensing for sharp directive beamforming," *IEEE Trans. Audio*

Speech Lang. Process., **21**, 2346–2355 (2013).
[19] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi and Y. Hioka, "Pinpoint extraction of distant sound source based on dnn mapping from multiple beamforming outputs to prior snr," *IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP 2016*), pp. 435–439 (2016).
[20] Y. Hioka, M. Kingan, G. Schmid and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," *IEEE Int. Workshop on Acoustic Signal Enhancement 2016* (*IWAENC 2016*), Sep., pp. 1–5 (2016).
[21] https://mediastore.auckland.ac.nz/library/public/2017/QuietUAV-hi.wide.preview (accessed 2017-05-24).
[22] K. Niwa, Y. Koizumi, K. Kobayashi and H. Uematsu, "Binaural sound generation corresponding to omnidirectional video view using angular region-wise source enhancement," *IEEE Int. Conf. Acoust. Speech Signal Process. 2016* (*ICASSP 2016*), pp. 2852–2856 (2016).
[23] D. Ochi, K. Niwa, A. Kameda, Y. Kunita and A. Kojima, "Dive into remote events: Omnidirectional video streaming with acoustic immersion," *23rd ACM Multimedia*, pp. 737–738 (2015).
[24] K. Niwa, D. Ochi, A. Kameda, Y. Kamamoto and T. Moriya, "Smartphone-based 360° video streaming/viewing system including acoustic immersion," *141-st Audio Eng. Soc. Conv.* (2016).
[25] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Trans. Audio Speech Lang. Process.*, **19**, 2374–2384 (2011).
[26] Y. Hioka, K. Furuya, K. Niwa and Y. Haneda, "Estimation of direct-to-reverberation energy ratio based on isotropic and homogeneous propagation model," *IEEE Int. Workshop on Acoustic Signal Enhancement 2012* (*IWAENC 2012*), Sep., pp. 1–4 (2012).
[27] O. Thiergart, T. Ascherl and E. Habets, "Power-based signal-to-diffuse ratio estimation using noisy directional microphones," *IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP 2014*), pp. 7440–7444 (2014).
[28] Y. Hioka and K. Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," *arXiv preprint arXiv:1510.08963* (2015).
[29] J. Eaton, N. D. Gaubitch, A. H. Moore and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **24**, 1681–1693 (2016).
[30] T. Kawase, K. Niwa, K. Kobayashi and Y. Hioka, "Application of neural network to source PSD estimation for wiener filter based array sound source enhancement," *IEEE Int. Workshop on Acoustic Signal Enhancement 2016* (*IWAENC 2016*), Sep. (2016).

Yusuke Hioka received his B.E., M.E., and Ph.D. degrees in engineering in 2000, 2002, and 2005 from Keio University, Yokohama, Japan. From 2005 to 2012, he was with the NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories), Nippon Telegraph and Telephone Corporation (NTT). From 2010 to 2011, he was also a visiting researcher at Victoria University of Wellington, New Zealand. In 2013 he moved to New Zealand and was appointed as a Lecturer at the University of Canterbury, Christchurch. Then in 2014, he joined the Department of Mechanical Engineering, the University of Auckland, Auckland, where he is currently a Senior Lecturer. He has wide interests in audio and acoustics, especially in microphone array signal processing and room acoustics. He is a Senior Member of IEEE and a Member of ASJ and IEICE.

Kenta Niwa received his B.E., M.E., and Ph.D. degrees from Nagoya University in 2006, 2008, and 2014. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2008, he has been engaged in research on microphone array signal processing. He is now a Research Engineer at NTT Media Intelligence Laboratories. He was awarded the Awaya Prize by the Acoustical Society of Japan (ASJ) in 2010. He is a member of IEEE, ASJ and the Institute of Electronics, Information and Communication Engineers (IEICE).