Improving Power Spectra Estimation in 2-Dimensional Areas Using Number of Active Sound Sources

Yusuke HIOKA, Member, Ken’ichi FURUYA, Yoichi HANEDA, and Akitoshi KATAOKA, Senior Members

SUMMARY An improvement of estimating sound power spectra located in a particular 2-dimensional area is proposed. We previously proposed a conventional method that estimates sound power spectra using multiple fixed beamformings in order to emphasize speech located in a particular 2-dimensional area. However, the method has one drawback that the number of areas where the active sound sources are located must be restricted. This restriction makes the method less effective when many noise sources located in different areas are simultaneously active. In this paper, we reveal the cause of this restriction and determine the maximum number of areas for which the method is able to simultaneously estimate sound power spectra. Then we also introduce a procedure for investigating areas that include active sound sources to reduce the number of unknown power spectra to be estimated. The effectiveness of the proposed method is examined by experimental evaluation applied to sounds recorded in a practical environment.

key words: microphone array, power spectrum estimation, speech enhancement, 2-dimensional area, estimating source position

1. Introduction

With the popularisation of teleconference and voice recognition systems, hands-free microphone systems have become indispensable devices due to their convenience. However, they often suffer from signal-to-noise ratio (SNR) deterioration due to the speaker’s mouth being positioned a few metres away from the microphone, which is much farther than the mouth’s position during use of a conventional handset or close-talking microphone. In particular, an intercom with a hands-free microphone may be used in a public space where various nonstationary noise sources are located around the speaker. In such an environment, the microphone might pick up confidential conversations that should not be sent to the other speaker. Thus, a hands-free microphone that picks up only the speech located in a particular 2-dimensional area, i.e., the area where the speakers are located, is required. Here, “2-dimension” corresponds to the direction and the distance.

For that purpose, using a microphone array [1] is a well-known strategy, and various conventional methods have been reported. However, only a few works [2]–[5] have dealt with the problem of discriminating sounds whose sources are located in the same direction but at different distances. One major approach reported is beamforming. Fixed beamforming [2] has the advantage of robustness against the change in noise source positions because its design is independent of the input signal; however, its noise suppression performance is limited. Adaptive beamforming is a different beamforming strategy that achieves better noise-suppression performance after sufficient adaptation calculation [3], [4]. However, its performance is sometimes severely degraded by noise source movement because the convergence speed of the adaptation lags behind the changes in noise position. There is also a method based on independent component analysis (ICA) [5], which is known to work similar to the adaptive beamforming [6]. But it also lacks the robustness against rapid change in noise positions because the ICA also includes an iterative calculation. Actually this goes for every conventional method referred above, the beamforming, which can be recognized as spatial filtering, is able to manipulate signals arriving from different directions up to the same number of its microphones due to the restriction on its degree of freedom. Due to this restriction, the beamforming could not achieve sufficient effect, otherwise it might be an ill-designed filter if it is forcedly optimised.

Another framework known as “fixed beamforming with nonlinear postfilter” [7] achieves sufficient noise suppression without suffering from limitation of the degree of freedom owing to the nonlinearity of the postfilter. Saruwatari et al. proposed a method within this framework that estimates the noise power spectrum for postfiltering using complementary beamforming [8]. However, even this method may also encounter the restriction due to its estimation of the noise power spectrum being performed in the fixed beamforming framework. Above all, this method deals with sound emphasis in a particular direction but not in a 2-dimensional area, therefore the method cannot be adopted to our problem directly.

The authors previously proposed a method [9] that estimates sound power spectra whose sources are located in the target sound areas using multiple fixed beamformings. Both the spectrum estimation and postfiltering are performed in the power spectrum domain, which belongs to nonlinear processing. Therefore, the method sufficiently emphasizes the speech without suffering from the restriction of degree of freedom. However, the method has one major drawback: the number of areas where the active sound sources are located must be restricted. Therefore, the method is less effective when many noise sources located in different areas are simultaneously active. Motivated by this situation, in this
paper, we aim to reveal the cause of this restriction and to determine the maximum number of areas for which the method is able to simultaneously estimate sound power spectra. We also present a method for investigating areas that include active sound sources in order to overcome the drawback of the conventional method.

This paper is organized as follows. First, in Sect. 2, we explain the basic principle of the conventional method that estimates the power spectra of a 2-dimensional area using fixed beamformings. Then in Sect. 3, the cause of the drawbacks in the conventional method is discussed and the key strategy of the proposed method to overcome the drawback is mentioned. A detailed explanation about the procedure of the proposed method is given in Sect. 4. Section 5 shows some experimental results to confirm the improvement achieved by the proposed method, and we conclude this paper with some comments in Sect. 6.

2. Power Spectra Estimation in 2-Dimensional Area Using Multiple Fixed Beamformers

Many of the conventional speech enhancement procedures based on nonlinear postfilter designs necessitate the power spectra of the target speech and noise to be estimated a priori [7]. For example, a well-known Wiener filter [10] is estimated by solving this simultaneous equation as long as the columns of the matrix are linearly independent. Because the experimental directivity shape is not like the ideal design assumed above, the average gains of actual directivity (depicted by thick solid line in the figure of beamformers’ directivity pattern shown in Fig. 1) for each angular sector can be approximated by

$$\begin{bmatrix} E[E[Y_1(\omega)^2]] \\ \vdots \\ E[E[Y_M(\omega)^2]] \end{bmatrix} \approx \begin{bmatrix} |a_{11}(\omega)|^2 \cdots |a_{1M}(\omega)|^2 & E[E[S_1(\omega)^2]] \\ \vdots & \vdots & \vdots \\ |a_{N1}(\omega)|^2 \cdots |a_{NM}(\omega)|^2 & E[E[S_M(\omega)^2]] \end{bmatrix},$$

where $E[\cdot]$ denotes the expectation, which is normally substituted by averaging the data of multiple frames. The power spectrum of the signals in each angular sector can be estimated by solving this simultaneous equation as long as the columns of the matrix are linearly independent. Because the experimental directivity shape is not like the ideal design assumed above, the average gains of actual directivity (depicted by thick solid line in the figure of beamformers’ directivity pattern shown in Fig. 1) for each angular sector can be approximated by $|a_{nm}(\omega)|^2$. Therefore, the left side of Eq. (1) is an approximation of the right side. In the following, we explain the expansion of this power spectrum estimation for sound sources located in a 2-dimensional area.

2.1 Power Spectrum Estimation Using Fixed Beamformers

We briefly explain here the basic structure of power spectrum estimation. When we have $N$ fixed beamformers whose complex directivity gains are assumed to be constant within each $M$ angular sector, depicted by dotted line in the figure of beamformers’ directivity pattern shown in Fig. 1, the output of the $n$-th beamformer is given by $Y_n(\omega) = \sum_m a_{nm}(\omega)S_m(\omega)$. Here, $a_{nm}(\omega)$ is the constant complex directivity gain of the $n$-th beamformer in the $m$-th angular sector, and $S_m(\omega)$ is the spectrum of the signal located in the $m$-th angular sector. If only the directivity gain to the $m$-th angular sector is not zero while the other gains are all zero, the beamformer is able to estimate the spectrum of each sound source separately. However, this is often impossible because the beamformer whose directivity gain satisfies such a condition cannot be designed in practice. This is because the beamforming could suffer from the degree of freedom in its design.

In contrast, our method does not estimates the signals themselves but it estimates the power spectrum using multiple beamformer outputs. If the signals are assumed to be mutually uncorrelated, the relation between the sound sources and outputs of beamformers are approximated by

$$\begin{bmatrix} E[E[Y_1(\omega)^2]] \\ \vdots \\ E[E[Y_M(\omega)^2]] \end{bmatrix} \approx \begin{bmatrix} |a_{11}(\omega)|^2 \cdots |a_{1M}(\omega)|^2 & E[E[S_1(\omega)^2]] \\ \vdots & \vdots & \vdots \\ |a_{N1}(\omega)|^2 \cdots |a_{NM}(\omega)|^2 & E[E[S_M(\omega)^2]] \end{bmatrix},$$

where $E[\cdot]$ denotes the expectation, which is normally substituted by averaging the data of multiple frames. The power spectrum of the signals in each angular sector can be estimated by solving this simultaneous equation as long as the columns of the matrix are linearly independent. Because the experimental directivity shape is not like the ideal design assumed above, the average gains of actual directivity (depicted by thick solid line in the figure of beamformers’ directivity pattern shown in Fig. 1) for each angular sector are used for $a_{nm}(\omega)$. Therefore, the left side of Eq. (1) is an approximation of the right side. In the following, we explain the expansion of this power spectrum estimation for sound sources located in a 2-dimensional area.

2.2 Definition of Area

In general, a microphone array requires a sufficiently large aperture size to discriminate the distances of the sources. To satisfy this requirement, we introduce a pair of separated microphone arrays, “array-L” and “array-R.” When $M_L$ and $M_R$ angular sectors are defined for respective arrays, in this placement, a 2-dimensional Area is defined by the combination of angular sector $\Theta_{xa}$ for both array-L and array-R, where $x$ and $a \in \{L, R\}$ represent the angular sectors and arrays, respectively. For example, when three angular sectors ($M_L = M_R = 3$) that are symbolized by $x \in \{C, L, R\}$ are assumed as shown in Fig. 2, the target Area $T$ and eight noise Areas $RR$, $R$, $C$, $L$, $LL$, $NC$, $NR$, and $NL$ are defined. Note that in the following we denote the Areas in an italic font, e.g., “T,” while the arrays are denoted in a Roman font, e.g., “L,” to avoid confusion. Regarding the Area, the following discussion uses the definition in Fig. 2 unless otherwise stated.

The positions of the arrays and Areas are arbitrarily determined under the following restrictions: (a) the distance
between the target Area and each microphone array is nearly the same \( (d_L = d_R) \), (b) the distance between two microphone arrays is unknown but longer than that between the target Area and each microphone array \( (d_{\text{inter}} \geq d_L \text{ or } d_R) \), and (c) the target Area is located in front of each microphone array. Furthermore, the received signal can be recognized as a plane wave because the aperture length of each microphone array is sufficiently short. In addition, we assume that each sound signal is uncorrelated.

2.3 Power Spectrum Estimation of Areas

In the same manner as that for the single array case mentioned in Sect. 2.1, we introduce fixed beamformers applied to both array-L and array-R. Each beamformer is designed to have a different directivity shape and has a null in its directivity pattern (directivity null hereafter) pointing toward one of the angular sectors \( \Theta_{\alpha} \), as depicted in Fig. 3. Now, the relationship between the input and output power spectra of the beamformers is described by

\[
\mathbf{Y}(\omega) = \mathbf{T}(\omega) \cdot \mathbf{Z}(\omega),
\]

where \( \mathbf{Y}(\omega, l) \) and \( \mathbf{Z}(\omega, l) \) are column vectors that consist of the power spectra of fixed beamformers’ outputs and the unknown power spectra of each Area, respectively. They are given by

\[
\begin{align*}
\mathbf{Y}(\omega) := & \begin{bmatrix} E \left| Y_{LL}(\omega) \right|^2 \\ E \left| Y_{LC}(\omega) \right|^2 \\ E \left| Y_{LR}(\omega) \right|^2 \\ E \left| Y_{RC}(\omega) \right|^2 \\ E \left| Y_{RR}(\omega) \right|^2 \end{bmatrix} \\
\mathbf{Z}(\omega) := & \begin{bmatrix} P_T(\omega) \\ P_{LL}(\omega) \\ P_L(\omega) \\ P_{NL}(\omega) \\ P_C(\omega) \\ P_{NC}(\omega) \\ P_R(\omega) \\ P_{NR}(\omega) \end{bmatrix}.
\end{align*}
\]

Here, \( Y_{SB}(\omega) \) denotes the output signal of beamformer \( \text{xSB} \) of array-a. Furthermore, the sum of the power spectra of signals whose sources are located in the Area \( b \) \( (b \in \{T, LL, L, C, R, RR, NR, NC, NL\}) \) are defined by

\[
P_b(\omega) := \sum_i E \left| S_{b_i}(\omega) \right|^2.
\]

where \( S_{b_i}(\omega) \) is the \( i \)-th sound source located in Area \( b \). On the other hand, the matrix \( \mathbf{T}(\omega) \) consists of column vectors given by Eq. (6) and Eq. (7), where each vector is composed of a set of squared directivity gains for each Area. As \( \mathbf{Y}(\omega) \) and \( \mathbf{T}(\omega) \) consist of known parameters, we have power spectra estimated using \( \mathbf{Z}(\omega) \), which is given by solving the simultaneous equation Eq. (2).

3. Restriction on Number of Areas to Estimate Power Spectra

The problem in solving Eq. (2) is that the simultaneous equation is under-determined because the rank of \( \mathbf{T}(\omega) \) is smaller than the number of unknown variables in \( \mathbf{Z}(\omega) \). This is caused by the lack of degrees of freedom in specifying a particular Area, which is determined by the number of angular sectors defined at each array. In fact, the maximum number of angular sectors that can be defined at each array is limited by the number of sound sources using this array, but we lose one degree of freedom in the directivity design because the distance between array-L and array-R is unknown.

\footnote{For this reason, we call them as “Suppression Beamformers”\textsuperscript{1}(SBs).}
sources. For example, there is no Area selected to estimate by the Area selection. We found that this occurs depending on the combination of Areas, as in Fig. 4(c). Eventually, such a problem will be avoided if the number of Areas is reduced to the smaller of $M_L$ and $M_R$.

Thus, the rank of $T(\omega)$ is reduced to $M_L + M_R - 1 (= 5)$. Consequently, to modify Eq. (2) so that it is determined, we need to reduce the number of unknown variables in $Z(\omega)$ to 5. In the proposed method, we achieve this by restricting the number of Areas for which to estimate the power spectra by solving Eq. (2). This idea is based on the fact that a nonstationary signal such as speech satisfies the sparseness assumption on the time-frequency plane [11]. This means the spectrum of input signals at an arbitrary frequency bin and frame is composed of a few sound sources. Thus, we concluded that the number of Areas could be reduced without causing severe degradation in the estimation accuracy as long as the method is used for speech enhancement.

Figure 4 shows examples of this Area restriction where the power spectrum of sound sources in the shaded Area is estimated. In Fig. 4(a), the Areas are located in all three angular sectors for both arrays. In this case, the rank of $T(\omega)$ is 5, which means we can solve Eq. (2) after reducing the number of unknown variables in $Z(\omega)$ to 5. However, in some cases, the rank of $T(\omega)$ becomes smaller than 5, although the number of variables in $Z(\omega)$ is reduced to 5 because the definition of $T(\omega)$ itself is changed by the Area selection. In other words, some vectors defined in Eq. (7) that are indispensable to span the original space of $T(\omega)$ are lost by the Area selection. We found that this occurs depending on the combination of Areas selected, as that includes sound sources. For example, there is no Area selected to estimate power spectra in the angular sector $\Theta_{RL}$ in Fig. 4(b). Because every vector that spans the subspace of $\Theta_{RL}$ is lost in this case, the rank of $T(\omega)$ is reduced to 4. The same problem occurs even if the number of Areas is restricted to 4, as in Fig. 4(c). Eventually, such a problem will be avoided if the number of Areas is reduced to the smaller of $M_L$ and $M_R$.

The process flow of the proposed method is shown in Fig. 5. The method mainly consists of the following steps: (a) investigating the possibility of an active sound source existing, (b) estimating the power spectrum, and (c) postfiltering. These steps are explained in the following.

Note that because of the nonstationarity of the sound sources, we practically deal with signals by frames whose index is denoted by $l$ hereafter. In addition to this, we also approximate the expectations appeared above by the average of frames, for example

$$E\left[|Y_{xa}(\omega, l)|^2\right] = \frac{1}{\Lambda} \sum_{\lambda=0}^{\Lambda-1} |Y_{xa}(\omega, l-\lambda)|^2,$$

where the symbol $\Lambda$ shows the number of frames used for the average.

### 4.1 Investigating Possibility of Existence of Active Sound Source

To specify the Area to be excluded from spectrum estimation, the proposed method preliminarily estimates the noise positions by investigating the possibility of each Area including active sound sources. We then omit Areas that do not include active sound sources from the power spectra estimation. This enables reformulation of the simultaneous equation to be over-determined, where the equation can be solved by the least squares method, by reducing the number of unknown variables.

The possibility is investigated by using the amount of input/output (I/O) level differences of the beamformers. As each beamformer in Fig. 3 is designed to point its directivity null towards a particular angular sector, the I/O level difference decreases if an active sound source is located in the corresponding angular sector. Because each Area is defined by the combination of the angular sectors of array-L and array-R, we can specify the Area b where an active sound source exists by examining the product of the I/O level differences given by

$$IO_b(\omega, l) =$$
HIOKA et al.: IMPROVING POWER SPECTRA ESTIMATION IN 2-DIMENSIONAL AREAS

Fig. 5 Process flow of proposed method.

Fig. 6 Example of investigating possibility of existence of active sound source.

\[
\left\{ \sum_{\lambda=0}^{\hat{\lambda}} |X_{L\lambda}(\omega, l - \lambda)|^2 - \sum_{\lambda=0}^{\hat{\lambda}} |Y_{2\lambda}(\omega, l - \lambda)|^2 \right\} \\
\times \left\{ \sum_{\lambda=0}^{\hat{\lambda}} |X_{R\lambda}(\omega, l - \lambda)|^2 - \sum_{\lambda=0}^{\hat{\lambda}} |Y_{2\lambda}(\omega, l - \lambda)|^2 \right\}, \tag{9}
\]

where \( k \) is an arbitrary microphone of each array. The smaller this value, the higher the possibility that a sound source exists.

Figure 6 shows an example of this investigation using the I/O level differences of the beamformers. As Fig. 6(a) shows, sound sources are located at each of Areas \( R \) and \( T \). When the outputs of beamformers applied to array-\( R \) are mutually compared, as described in Fig. 6(b), the output of \( \text{CSBR} \) is significantly reduced because both \( R \) and \( T \) are located in the direction of \( \Theta_{CR} \) to which \( \text{CSBR} \) points its null. In the same manner, in the case of array-\( L \) as shown in Fig. 6(c), both the outputs of \( \text{LSBL} \) and \( \text{CSBL} \) are reduced due to \( R \) and \( T \) being located in the direction of \( \Theta_{RL} \) and \( \Theta_{CL} \), respectively. The I/O level differences defined by Eq. (9) are the combinations of results measuring how much the signal was reduced by the beamformers of array-\( R \) and array-\( L \). Therefore, as described in Fig. 6(d), the I/O level differences of both \( R \) and \( T \) become bigger as they are defined by the combinations of \( \text{CSBR}-\text{LSBL} \) and of \( \text{CSBR}-\text{CSBL} \), respectively. Thus, we can specify the Areas that have high possibility of sound source existence by searching the Areas whose I/O level differences are bigger than those of the other Areas.

Note that there are some specific cases that we cannot determine the sound source existence by using I/O level differences. However, this restriction could be ignored because the chances of occurrence of such cases are extremely small. The details of this restriction is stated in Appendix.

4.2 Power Spectrum Estimation under Area Restriction

The number of unknown variables in \( \mathbf{Z}(\omega, l) \) is reduced by using the investigated possibilities of active sound source existence. Because the rank of \( \mathbf{T}(\omega) \) is not more than 5, we first discard 4 out of 9 Areas whose product of I/O level differences is large. Thus, we newly introduce a vector \( \mathbf{T}_5(\omega) := [t_{c1} \ t_{c2} \ t_{c3} \ t_{c4} \ t_{c5}] \), where the set \( \mathcal{C} := \{c_1, \ldots, c_5\} \) consists of the names of remaining (not discarded) Areas listed in descending order of the product of the I/O level differences, i.e., \( IO_{c1} > IO_{c2} > \cdots > IO_{c9} \). Then, the power spectra of the remaining Areas are estimated by solving the modified simultaneous equation

\[
\mathbf{T}_5(\omega) \mathbf{Z}_5(\omega, l) \simeq \mathbf{Y}(\omega, l), \tag{10}
\]

\[
\mathbf{Z}_5(\omega, l) := \begin{bmatrix}
P_{c1}(\omega, l) \\
P_{c2}(\omega, l) \\
P_{c3}(\omega, l) \\
P_{c4}(\omega, l) \\
P_{c5}(\omega, l)
\end{bmatrix}. \tag{11}
\]

The equation is now over-determined, so it can be solved by the least squares method as \( \hat{\mathbf{Z}}_5(\omega, l) = \mathbf{T}_5^\dagger(\omega) \mathbf{Y}(\omega, l) \), where \( \dagger \) denote the Moore-Penrose pseudo inverse and estimated value, respectively. If \( \mathbf{T}_5(\omega) \) is still rank deficient, we further discard the Area with the least possibility of containing a sound source from the estimation until the rank of the matrix corresponds to the number of Areas to be estimated.
4.3 Modified Postfilter Based on Selected Areas

For emphasizing the sound sources in the target Area, we multiply the sum of fixed beamformer outputs with the post-filter $G(\omega, l)$. In the summation, only the beamformers whose directivity nulls are not pointing towards the target Area, i.e., LSB and RSB, are applied.

$$Y(\omega, l) = G(\omega, l) \cdot Y_B(\omega, l)$$  \hspace{1cm} (12)

$$Y_B(\omega, l) := \frac{1}{4} (Y_{LL}(\omega, l) + Y_{RL}(\omega, l) + Y_{LR}(\omega, l) + Y_{RR}(\omega, l))$$  \hspace{1cm} (13)

The postfilter is calculated using the estimated power spectra of the target and noise Area by well-known conventional methods, such as Wiener filtering [10], i.e.,

$$G(\omega, l) := \frac{\hat{P}_T(\omega, l)}{\sum_{i \in C} \hat{P}_i(\omega, l)}$$  \hspace{1cm} (14)

except for the case where the target Area $T$ was not included in the set $C$. In that case, we recognize that the target sound source is not active; thus, the $G(\omega, l)$ is set to a small constant. Finally, the speech-emphasized signal is given by applying the inverse Fourier transform to $Y(\omega, l)$.

5. Experimental Evaluation

5.1 Situation and Parameters

Experiments were performed in a room whose reverberation time was approximately 250 ms. The parameters used in the experiment are shown in Table 1. The positions of the microphone arrays and sound sources are shown in Fig. 7. The target speaker was positioned at $T$, which was located inside the target Area, and the noise sources were positioned at $N_1$ to $N_6$. Each microphone array was composed of four microphones in a linear and equi-spaced configuration with an inter-microphone distance of 3 cm. For comparison, we also evaluated the conventional method [9], which assumes that no noise exists in the Areas $NL$, $NR$, $L$, and $NC$ unless otherwise stated.

5.2 Noise Suppression Performance

First, we examined the performance for a single noise source. The improvements in the signal-to-interference ratio (SIR) [12] were measured when the noise source was located at one of the noise positions defined in Fig. 7. In the results shown in Fig. 8(a), the conventional method failed to suppress any noise signal whose source was located in Areas $NR$, $NL$, $L$, or $NC$, which the conventional method assumed that there were no active sound sources. In contrast, the proposed method, whose results are shown in Fig. 8(b), achieved 6 to 11 dB SIR improvement for most of the Areas, which was nearly the same performance as that of the conventional method when the noise sources were located in the supposed Areas.

Table 1 Parameters applied to experimental evaluation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>$d$</td>
<td>3 cm</td>
</tr>
<tr>
<td>Frame Length</td>
<td>32 ms</td>
</tr>
<tr>
<td>Frame Shift</td>
<td>16 ms</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>3</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>1</td>
</tr>
<tr>
<td>Number of microphones in array-L</td>
<td>4</td>
</tr>
<tr>
<td>Number of microphones in array-R</td>
<td>4</td>
</tr>
<tr>
<td>$\Theta_{NL}$</td>
<td>$[10^\circ, 70^\circ]$</td>
</tr>
<tr>
<td>$\Theta_{LR}$</td>
<td>$[-70^\circ, -10^\circ]$</td>
</tr>
<tr>
<td>$\Theta_{CL}$</td>
<td>$[-10^\circ, 10^\circ]$</td>
</tr>
<tr>
<td>$\Theta_{CR}$</td>
<td>$[-10^\circ, 10^\circ]$</td>
</tr>
<tr>
<td>$\Theta_{CL}$</td>
<td>$[-50^\circ, -10^\circ]$</td>
</tr>
<tr>
<td>$\Theta_{CR}$</td>
<td>$[10^\circ, 50^\circ]$</td>
</tr>
</tbody>
</table>

Fig. 7 Positions of microphone array and sound sources.

Fig. 8 Noise suppression performance measured by SIR improvements: (a) conventional and (b) proposed methods.
Second, the noise suppression performance with regard to the number of Areas including simultaneously active sound sources was evaluated. Figure 9 shows the improvements in SIR, which are the averages of that calculated for every case of Area combination, when the number of active noise Areas was changed. The results showed that the conventional method loses its ability to suppress noise as the number of Areas increases, while the proposed method maintains its effectiveness even if noise sources are located at every noise Area. For the conventional method, we also showed the improvements in SIR averaged for only the combinations of the Areas where the conventional method assumed, i.e., C, R, LL, and RR. In comparison, the results of the proposed method were not much worse than those of the conventional method for assumed Areas.

5.3 Evaluation on Sound Distortion

The nonlinear postfilter as to what the proposed method applied is well-known to cause signal distortion that degrades the sound quality in return for the noise suppression effect. From the sound quality point of view, we performed a subjective evaluation for the output signals of the proposed methods. In the test, each of seven subjects (five males and two females aged from twenties to forties) rated the sound quality of the outputs of the proposed method comparing to that of the conventional method by using a scale whose assessment words are defined in Table 2. The test was performed only for the conditions where the conventional method correctly worked, i.e., the noise sources were located only in the assumed Areas mentioned above. Furthermore, in order to avoid evaluating the effect of noise suppression, each set of data consisted of results of the conventional and proposed methods that gave nearly the same amount of SIR improvements. As the number of results which met such condition was limited, the subjective test was performed by using only the results given under the situation where either two or three of noise Areas were active.

Figure 10 shows the results of the subjective evaluation. Each “x” plot shows the average score and the error bar denotes 95% confidence interval. From the results, we cannot find any significant degradation of sound quality because the score “0,” which means “the sound quality of the proposed method is the same as that of the conventional method” is included within the range of confidence interval in every case.

Added to this, for the purpose to know the deviation of sound quality depending on the noise source positions, we also evaluated the amount of distortion in the extracted target sounds using the signal-to-distortion ratio (SDR) [12]. The evaluation was only performed to the output of proposed method at each position of noise source given in Fig. 8(b). As a result, average SDR for every noise source position was 7.11 dB, and there was no large dispersion depending on the noise source position where the standard deviation of SDR was 0.75 dB. In conclusion, the proposed method is superior to the conventional method in their noise suppressing effect while there is no significant difference in the sound quality.

5.4 Example Result of Proposed Method

Finally, Fig. 11 shows the waveforms of the target speech, input signal, and output signals when the noise source is located at the positions N3, N4, N6, N7, and N8. From these results, we can see that the proposed method was successful to retrieve the target signal because the waveform of the output signal resembles the target speech much more than the contaminated input signal. Furthermore, during the non-target interval, e.g., 0 to 4.7 sec or 4.7 to 6.2 sec, the level of the interfering sound was sufficiently reduced even though five noise sources were active in this interval.
An improved method for estimating the power spectra of a sound source located in a particular 2-dimensional area has been proposed. We first introduced the conventional method of estimating power spectra using multiple fixed beamformers, and then we extended the method to a 2-dimensional case and discussed the maximum number of Areas for which the method is able to simultaneously estimate power spectra. Then we presented a way to overcome this limitation by investigating the source location. The effectiveness of the proposed method was examined by experimental evaluation applied to sound recorded in a practical environment.

### References


### Appendix: Restrictions on Estimating Sound Source Existence by Using I/O Level Difference

As we noted in Sect. 4.1, there are some specific cases where we cannot determine the sound source existence by using the I/O Level Difference. Here we explain about this restriction in detail. First of all, let us assume that each suppression beamformer defined in Fig. 3 ideally suppresses the signals arriving from each angular sector while it passes the signals from other sectors without any distortion. For example, the gains of $\text{LSB}$ in such situation are given by $\alpha_{\text{LSB}}(\omega) = \alpha_{\text{CL}}(\omega) = 1$ and $\alpha_{\text{SR}}(\omega) \approx 0$. Based on this assumption, the I/O level differences $\text{IO}_b(\omega, l)$ expressed by a matrix form is given by

$$
\text{IO}_b(\omega, l) = \begin{bmatrix}
\Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} \\
\Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} \\
\Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} \\
\Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}} & \Pi_{\text{LSB}} & \Pi_{\text{RSB}}
\end{bmatrix},
$$

where

$$
\Pi_{\text{LSB}}(\omega, l) = P_{\text{C}}(\omega, l) + P_{\text{L}}(\omega, l) + P_{\text{LL}}(\omega, l),
\Pi_{\text{LSB}}(\omega, l) = P_{\text{R}}(\omega, l) + P_{\text{P}}(\omega, l) + P_{\text{NL}}(\omega, l),
\Pi_{\text{LSB}}(\omega, l) = P_{\text{RR}}(\omega, l) + P_{\text{NR}}(\omega, l) + P_{\text{NC}}(\omega, l),
\Pi_{\text{RSB}}(\omega, l) = P_{\text{C}}(\omega, l) + P_{\text{R}}(\omega, l) + P_{\text{RR}}(\omega, l),
\Pi_{\text{RSB}}(\omega, l) = P_{\text{L}}(\omega, l) + P_{\text{P}}(\omega, l) + P_{\text{NR}}(\omega, l),
\Pi_{\text{RSB}}(\omega, l) = P_{\text{LL}}(\omega, l) + P_{\text{NL}}(\omega, l) + P_{\text{NC}}(\omega, l).
$$

As we search the Areas whose I/O level differences are bigger than those of the others, we cannot specify the Area if some I/O level differences have an equal value. From Eq. (A-1), such cases occur when a part/all of the following equations are satisfied.

$$
\Pi_{\text{LSB}}(\omega, l) = \Pi_{\text{LSB}}(\omega, l) = \Pi_{\text{LSB}}(\omega, l),
\Pi_{\text{RSB}}(\omega, l) = \Pi_{\text{RSB}}(\omega, l) = \Pi_{\text{RSB}}(\omega, l).
$$

If a part of these equations exactly holds, the proposed I/O level difference cannot be used to determine the Areas. However, in practical situation, it is an extremely rare case because most of the noise sources are nonstationary and their level change so quickly. Therefore, we could ignore the influence of this restriction on practical implementation of the proposed method.

### Fig. 11

Observed waveforms in experiment: (a) target speech, (b) input signal, (c) output of proposed method.
Yusuke Hioka received B.E., M.E., and Ph.D. degrees in electrical engineering in 2000, 2002, and 2005 respectively from Keio University, Yokohama, Japan. From 2002 to 2005, he was a Research Associate of 21th Century COE program at Keio University. In 2005, he joined Nippon Telegraph and Telephone Corporation (NTT), and currently he is a Research Engineer in NTT Cyber Space Laboratories. Since June 2010, Dr. Hioka has also been a Visiting Research Fellow at Massey University, Wellington, New Zealand. His research interests include array signal processing and acoustic signal processing. He is a member of IEEE and the Acoustical Society of Japan.

Ken’ichi Furuya received the B.E. and M.E. degrees in acoustic design from Kyushu Institute of Design, Fukuoka, Japan, in 1985 and 1987, respectively, and the Ph.D. degree from Kyushu University, Japan, in 2005. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1987. He is now a Senior Research Engineer at NTT Cyber Space Laboratories, Tokyo, Japan. His current research interests include signal processing in acoustic engineering. Dr. Furuya was awarded the Sato Prize by the Acoustical Society of Japan (ASJ) in 1991. He is a member of the Acoustical Society of Japan, the Acoustical Society of America, and IEEE.

Yoichi Haneda received the B.S., M.S., and Ph.D. degrees from Tohoku University, Sendai, in 1987, 1989, and 1999, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 1989, he has been investigating the modelling of acoustic transfer functions, acoustic signal processing, and acoustic echo cancellers. He is now a Senior Research Engineer, Group Leader at NTT Cyber Space Laboratories. He received the paper awards from the Acoustical Society of Japan, and from the Institute of Electronics, Information, and Communication Engineers of Japan in 2002. Dr. Haneda is a senior member of the IEEE, the Acoustical Society of Japan.

Akitoshi Kataoka received the B.E., M.E., and Ph.D. degrees in Electrical Engineering from Doshisha University of Kyoto in 1984, 1986, and 1999 respectively. Since joining NTT Laboratories in 1986, he has been engaged in research on noise and reverberation reduction, acoustic microphone arrays, and medium bit-rate speech and wideband coding algorithms for the ITU-T standard. He contributed to establishing ITU-T G.729 standards. He is currently Professor of the Faculty of Science and Technology at Ryukoku University. He received the Technology Development Award from ASJ, the Telecommunication Systems Technology Prize awarded by The Telecommunications Advancement Foundation in 1996, and the Prize of the Commissioner of the Japan Patent Office from the Japan Institute of Invention and Innovation in 2003. Dr. Kataoka is a member of the Acoustical Society of Japan.