We propose a method for separating speech signals when sources outnumber the sensors. In this paper we mainly concentrate on the case of three sources and two sensors. Some existing methods employ binary masks to extract the signals, and therefore, the extracted signals contain loud musical noise. To overcome this problem, we propose the utilization of a directivity pattern based continuous mask, which removes a single source from the observations, and independent component analysis (ICA) to separate the remaining mixtures. Experimental results show that our proposed method can separate signals with little distortion even in a real reverberant environment of $T_R=130$ ms.

1. INTRODUCTION

Blind source separation (BSS) is an approach that estimates original source signals $s_i(n)$ only from observations $x_j(n)$ without source or mixing process information.

In this paper, we consider the BSS of speech signals observed in a real environment, i.e., the BSS of convolutive mixtures of speech. Recently, many methods have been proposed to solve the BSS problem of convolutive mixtures (e.g., [1]). However, most of these methods consider the determined or overdetermined case. In contrast, we focus on the underdetermined BSS problem where source signals outnumber the sensors, especially the case of three sources and two sensors.

There are two approaches with which to realize underdetermined BSS. Both approaches rely on the sparseness of source signals. One involves the clustering of time-frequency points with binary masks [2], and the other is based on ML estimation, where the sources are estimated after mixing matrix estimation [3–5]. Since separation in a real environment has already been tried with the former method, we decided to watch a binary mask approach [2]. If the signals are sufficiently sparse, that is, most of the samples of a signal are almost zero, we can assume that the sources rarely overlap. Rickard and Yilmaz [2] employ this assumption and extract each signal using a time-frequency binary mask (BM). However, the use of binary masks means that there is too much discontinuous zero-padding to the extracted signals, and they contain loud musical noise.

To overcome this, we have proposed combining binary masks and ICA (BMICA) to solve the underdetermined BSS problem [6]. In the paper, we estimate the time points when only one source is active using the sparseness assumption. Then, we remove this single source from the observations with a binary mask and apply ICA to the remaining mixtures to obtain separated signals. The single source removal with this method causes less discontinuous zero-padding than with the BM method, because the BMICA extracts more time-frequency points than the BM method. Therefore, we have been able to use the BMICA to obtain separated signals with little distortion. However, as the BMICA still employs a binary mask for one source removal, the zero-padding to the separated signals still remains. Moreover, heuristic parameters are used when designing the binary mask shapes not only for the BMICA but also for the BM method.

In this paper, we propose a new method for one source removal that employs a directivity pattern based continuous mask (DCmask). First, using the source sparseness, we estimate the direction of arrival (DOA) of each source. We then remove the single source from the observations with a DCmask, which has a small gain for the DOA of one source and a large gain for other directions. We next apply ICA to the remaining mixtures. Since this DCmask is not binary, discontinuous zero-padding to the extracted signals does not occur. Moreover, we do not need any parameters for designing the DCmask.

Experimental results show that our method can separate signals with little distortion even in a real reverberant environment of $T_R=130$ ms.
3. CONVENTIONAL METHODS

Several methods have been proposed (e.g., [2–7]) for solving the underdetermined BSS problem, and they all utilize source sparseness. If most of the samples of a signal are almost zero, we say that this signal is sparse. For a detailed experimental analysis of sparseness, see [7].

3.1 Conventional method 1: with only binary masks (BM)

Some methods use the sparseness assumption and extract each signal using time-frequency binary masks (e.g., [2]). When signals are sufficiently sparse, it can be assumed that sources do not overlap very often. This condition is closely discussed in [2]. Using this non-overlap assumption, they extract each source by selecting the time points at which there is only one signal. One way of estimating such time points is to use the level difference and the phase difference between the observations. In this paper, we utilize omni-directional microphones, therefore we use the phase difference \( \phi(\omega, m) = \frac{S_j(\omega, m)}{S_i(\omega, m)} \) between the observations.

Using \( \phi(\omega, m) \), we estimate the direction of arrival (DOA) for each time point \( m \) by calculating \( \theta(\omega, m) = \cos^{-1} \phi(\omega, m) \), where \( c \) is the speed of sound and \( d \) is the microphone spacing, and draw a histogram of the DOA \( \theta(\omega, m) \) (Fig. 2). Each peak corresponds to each source in the histogram for each frequency. Let these peaks be \( \theta_1, \theta_2, \) and \( \theta_3 \), where \( \theta_1 \leq \theta_2 \leq \theta_3 \) (Fig. 2), and the signal from \( \theta_i \) be \( S_i \) (\( r = 1, 2, 3 \)).

We can extract each signal with a binary mask

\[
M_{\text{BM}}(\omega, m) = \begin{cases} 
1 & \theta_i - \Delta \leq \theta(\omega, m) \leq \theta_i + \Delta \\
0 & \text{otherwise}
\end{cases}
\]

by calculating \( Y_j(\omega, m) = M_{\text{BM}}(\omega, m)X_j(\omega, m) \) where \( j=1 \) or 2. Here, \( \Delta \) is an extraction range parameter.

Although we can extract each signal using this binary mask (1), such extracted signals are discontinuously zero-padded by the binary masks, and therefore, we hear musical noise in the outputs. Moreover, the performance depends on the heuristic parameter, \( \Delta \).

3.2 Conventional method 2: with binary masks and ICA (BMICA)

To overcome the musical noise problem, we have proposed using both a binary mask and ICA (BMICA) [6]. The BMICA has two stages (Fig. 3).

In the first stage, unlike the conventional method 1 where single source is extracted, we remove one source from the mixtures using a binary mask

\[
M_{\text{BMICA}}(\omega, m) = \begin{cases} 
1 & \theta_{\text{min}} \leq \theta(\omega, m) \leq \theta_{\text{max}} \\
0 & \text{otherwise}
\end{cases}
\]

by calculating \( \hat{X}^{\text{BMICA}}(\omega, m) = M_{\text{BMICA}}(\omega, m)X(\omega, m) \).

In (2), \( \theta_{\text{min}} \) and \( \theta_{\text{max}} \) are extraction range parameters that cover the direction range of two sources, and in (3), \( \hat{X}^{\text{BMICA}}(\omega, m) = \begin{bmatrix} X_1^{\text{BMICA}}(\omega, m), X_2^{\text{BMICA}}(\omega, m) \end{bmatrix} \) are expected to be mixtures of \( S_1 \) and \( S_2 \). Then, as we can expect the remaining mixtures \( \hat{X}^{\text{BMICA}} \) to consist of only two signals, we apply a standard ICA to these remaining mixtures in the second stage.

In the 1st stage, we expect that the zero-padding of the extracted signals to cause less trouble because we extract more time-frequency points than conventional method 1. Therefore we obtained a result with less musical noise [6].

However, as BMICA still employed a binary mask for one source removal, the zero-padding of the separated signals still remained. Moreover, we have to find a reasonable \( \theta_{\text{min}} \) and \( \theta_{\text{max}} \) in (2). This is not an easy problem and we relied on manual setting.

4. PROPOSED METHOD

[1st stage] One source removal with new DC mask:

In this paper, we propose a new method (Fig. 4) which employs a directivity pattern based continuous mask (DCmask) in the 1st stage. The DCmask has a small gain for the DOA of one source and preserves signals from other directions. As shown in Fig. 2 and Sec. 3, DOA is a powerful clue with which to estimate the active source in each frame and to extract or remove one signal. Therefore the DCmask is a natural alternative. Because this new mask is not binary, it completely avoids any zero-padding to the extracted signals.

One way to obtain such a mask is to utilize the gain of the directivity pattern of a null beamformer (NBF) which makes a null toward the estimated direction \( \theta_i \) of one source.

First we estimate the source directions with a histogram such as in Fig. 2. Then, we form a matrix

\[
H_{\text{NBF}}(\omega) = \begin{bmatrix} 
\exp(i\omega\varphi_{11}) & \exp(i\omega\varphi_{12}) \\
\exp(i\omega\varphi_{21}) & \exp(i\omega\varphi_{22})
\end{bmatrix},
\]

Figure 2: Example histogram. A male-male-female combination with STFT frame size \( T = 512 \), \( T_R = 0 \) ms.
where $\tau_{ji} = \frac{d_j}{c} \cos \theta_i$, $d_j$ is the position of the $j$-th microphone and $c$ is the speed of sound. Where, one of the $\theta_i$ values should be the estimated DOA $\hat{\theta}_i$ of signal to be removed, and the other should be a different direction from which the signal’s gain and phase are constrained at 1.

The directivity pattern of the NBF, $W(\omega) = H_{\text{DC}}^{-1}(\omega)$, is obtained by

$$F_r(\omega, \theta) = \frac{1}{M} \sum_{m=1}^{M} W_{ir}(\omega) \exp(j o \omega d_i \cos \theta / c).$$

Here, we use the gain of the directivity pattern of NBF as our mask,

$$M_{\text{DC}}(\omega, m) = |F_r(\omega, \theta(m, o))| \quad (p \neq r, q \neq r).$$

This is our new mask, the DCmask. Figure 5 shows an example of the shape of a DCmask.

Finally, the one source removal is achieved by

$$\hat{X}^{\text{PN}}(\omega, m) = M_{\text{DC}}^{\text{PN}}(\omega, m)X(\omega, m).$$

It should be noted that the DCmask is applied to both channels (Fig. 4), because ICA in the 2nd stage needs two inputs that maintain the mixing matrix information. Here, our new mask $M_{\text{DC}}^{\text{PN}}(\omega, m)$ changes only the gains of inputs $X_1$ and $X_2$ and preserves their phases. Therefore, we can preserve to some extent the spatial information on the sources, and a standard ICA can work in the 2nd stage.

**[2nd stage] Separation of remaining sources by ICA:**

Because the remaining signals $\hat{X}^{\text{PN}}$ are expected to be mixtures of two signals, we can use 2×2 ICA to separate $\hat{X}^{\text{PN}}$. The separation process can be formulated as

$$Y^{\text{PN}}(\omega, m) = W^{\text{PN}}(\omega) \hat{X}^{\text{PN}}(\omega, m),$$

where $Y^{\text{PN}}(\omega, m)$ is the masked observed signal obtained by (7), $Y^{\text{PN}}(\omega, m)$ represents a (2×2) separation matrix, $W^{\text{PN}}(\omega)$ is determined so that $Y_p(\omega, m)$ and $Y_q(\omega, m)$ become mutually independent by ICA.

Note that we need two paths to obtain three separated signals (see Fig. 3) because our system has only two outputs.

Although we consider the case of two sensors ($M = 2$) and three sources ($N = 3$) here, we can expand our DCmask method to the underdetermined case of $N \leq (\text{number of nulls formed by $M$ sensors}) + (\text{number of outputs of a standard ICA}) = (M - 1) + M$.

5. EXPERIMENTS

5.1 Experimental conditions

We conducted anechoic tests and reverberant tests. For the anechoic tests ($T_R = 0$ ms), we simulated the recording in an anechoic room using the mixing matrix $H_{ji}(\omega) = \exp(j o \tau_{ji})$, where

$$\tau_{ji} = \frac{d_j}{c} \cos \theta_i, d_j \text{ is the position of the } j \text{-th microphone, and } \theta_i \text{ is the direction of the } i \text{-th source. The source directions were } 45^\circ, 90^\circ \text{ and } 135^\circ.$$

For the reverberant tests, we used speech data convolved with impulse responses recorded in a real room (Fig. 6) whose reverberation time was $T_R = 130$ ms.

As the original speech, we used three Japanese sentences spoken by three male and three female speakers. We investigated three combinations of speakers: male-male-female (m-m-f), male-male-male (m-m-m), and female-female-female (f-f-f).

The STFT frame size $T$ was 512 and the frame shift was 256 at a sampling rate of 8 kHz. We used $\theta_{\text{min}} = \hat{\theta}_2 - \Delta$ and $\theta_{\text{max}} = 180^\circ$ for $M_{\text{max}}^{\text{BMICA}}$ (BMICA 23), and $\theta_{\text{min}} = 0^\circ$ and $\theta_{\text{max}} = \hat{\theta}_2 + \Delta$ for $M_{\text{max}}^{\text{BMICA}}$ (BMICA 12). The $\Lambda$ value in the conventional methods’ binary masks was 15° in DOA.

The adaptation rule of ICA that we used in our experiments was $W_{i+1}(\omega) = W_i(\omega) + \eta [\text{diag}((\Phi Y)^H Y) - (\Phi Y)^H] \cdot W_i(\omega)$, where $\Phi(y) = \phi(y) \cdot e^{i \angle y}$, $\phi(x) = \tanh(g x)$ and $g = 100$. To solve the permutation problem of frequency domain ICA, we employed the DOA and correlation approach [8], and to solve the scaling problem of frequency domain ICA, we used the minimum distortion principle [9].

5.2 Performance measures

We used the signal to interference ratio (SIR) as a measure of separation performance, and the signal to distortion ratio (SDR) as a measure of sound quality.

$$SIR_i = 10 \log \left( \frac{\sum_{n=6}^{N} y_i^2(n)}{\sum_{n=6}^{N}(\sum_{j \neq i} y_{ij}(n))^2} \right)$$

Figure 4: Block diagram of proposed method. Only one path is drawn for visibility.
5.3 Experimental results

5.3.1 Applicability of ICA at the 2nd stage

Before trying to separate signals with our method, we investigated the applicability of ICA at the 2nd stage, because if the contribution of the third signal is large in $\hat{x}$, we cannot use a standard ICA algorithm at the 2nd stage. Table 1 shows the percentage of each source power extracted by $M_2^I$ and $M_3^I$. Three signals are dominant and the contributions of the third signal are small. Therefore, we can say that we can use ICA in the 2nd stage.

Table 2 shows the experimental results we obtained for $T_R = 0$ ms. The SIR and SDR values were averaged over three speaker combinations. The first row shows the results obtained by conventional method 1; the rows labeled ‘BMICA’ show the results obtained by conventional method 2, and the rows labeled ‘DCICA’ show the results obtained with our proposed method.

With conventional method 1, the SDR values were unsatisfactory, and a large musical noise was heard. In contrast, with our proposed method (DCICA), we were able to obtain high SDR values without any serious deterioration in the separation performance SIR. Compared with conventional method 2 (BMICA), although SDR values were a bit degraded, we hear no musical noise with DCICA. Some sound samples can be found at our web site [10].

Table 3 and 4 show the results of reverberant tests for $T_R = 130$ ms. In the reverberant case, due to the decline of sparseness, the performance with all methods was worse than when $T_R = 0$ ms. However, we were able to obtain higher SDR values with DCICA than with conventional method 1 even in a reverberant environment without musical noise.

It should be noted that it remains difficult to separate signals at the center position with any method.

Table 1: Percentage of each source power. $T_R = 0$ ms.

<table>
<thead>
<tr>
<th>comb.s</th>
<th>DCmask</th>
<th>BMICA 12</th>
<th>BMICA 23</th>
<th>DCICA 12</th>
<th>DCICA 23</th>
</tr>
</thead>
</table>

Table 2: Results of $T_R = 0$ ms simulations. ‘Conv. 1’: with conventional method 1, ‘BMICA 12’ and ‘BMICA 23’: with conventional method 2, ‘DCICA 12’ and ‘DCICA 23’: with proposed method.

<table>
<thead>
<tr>
<th>SIR1</th>
<th>SIR2</th>
<th>SIR3</th>
<th>SDR1</th>
<th>SDR2</th>
<th>SDR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv. 1</td>
<td>18.0</td>
<td>8.9</td>
<td>18.4</td>
<td>7.9</td>
<td>11.5</td>
</tr>
<tr>
<td>BMICA 12</td>
<td>12.6</td>
<td>5.9</td>
<td>-</td>
<td>18.1</td>
<td>15.2</td>
</tr>
<tr>
<td>BMICA 23</td>
<td>-</td>
<td>6.1</td>
<td>13.0</td>
<td>-</td>
<td>13.6</td>
</tr>
<tr>
<td>DCICA 12</td>
<td>16.1</td>
<td>4.8</td>
<td>-</td>
<td>15.2</td>
<td>12.5</td>
</tr>
<tr>
<td>DCICA 23</td>
<td>-</td>
<td>4.6</td>
<td>16.2</td>
<td>-</td>
<td>12.8</td>
</tr>
</tbody>
</table>

$\text{SDR}_i = 10 \log \frac{\sum_n y_{is_i}^2 (n)}{\sum_n (x_{is_i} (n) - \alpha y_{is_i} (n - D))^2}$, (10)

where $y_i$ is the estimation of $s_i$, and $y_{is_i}$ is the output of the whole separating system at $y_i$ when only $s_i$ is active, and $x_{is_i} = h_{is_i} * s_i$ (+ is a convolution operator). $\alpha$ and $D$ are parameters to compensate for the amplitude and phase difference between $x_{is_i}$ and $y_{is_i}$.

5.3.2 Separation results

Table 2 shows the experimental results we obtained for $T_R = 0$ ms. The SIR and SDR values were averaged over three speaker combinations. The first row shows the results obtained by conventional method 1; the rows labeled ‘BMICA’ show the results obtained by conventional method 2, and the rows labeled ‘DCICA’ show the results obtained with our proposed method.

With conventional method 1, the SDR values were unsatisfactory, and a large musical noise was heard. In contrast, with our proposed method (DCICA), we were able to obtain high SDR values without any serious deterioration in the separation performance SIR. Compared with conventional method 2 (BMICA), although SDR values were a bit degraded, we hear no musical noise with DCICA. Some sound samples can be found at our web site [10].

Table 3 and 4 show the results of reverberant tests for $T_R = 130$ ms. In the reverberant case, due to the decline of sparseness, the performance with all methods was worse than when $T_R = 0$ ms. However, we were able to obtain higher SDR values with DCICA than with conventional method 1 even in a reverberant environment without musical noise.

It should be noted that it remains difficult to separate signals at the center position with any method.

Table 3: Percentage of each source power. $T_R = 130$ ms.

<table>
<thead>
<tr>
<th>comb.s</th>
<th>DCmask</th>
<th>BMICA 12</th>
<th>BMICA 23</th>
<th>DCICA 12</th>
<th>DCICA 23</th>
</tr>
</thead>
</table>

Table 4: Results of reverberant tests. $T_R = 130$ ms.

<table>
<thead>
<tr>
<th>SIR1</th>
<th>SIR2</th>
<th>SIR3</th>
<th>SDR1</th>
<th>SDR2</th>
<th>SDR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv. 1</td>
<td>12.3</td>
<td>6.3</td>
<td>11.0</td>
<td>5.0</td>
<td>13.9</td>
</tr>
<tr>
<td>BMICA 12</td>
<td>9.8</td>
<td>5.5</td>
<td>-</td>
<td>7.8</td>
<td>15.9</td>
</tr>
<tr>
<td>BMICA 23</td>
<td>-</td>
<td>5.5</td>
<td>9.2</td>
<td>-</td>
<td>14.5</td>
</tr>
<tr>
<td>DCICA 12</td>
<td>11.4</td>
<td>3.2</td>
<td>-</td>
<td>6.7</td>
<td>6.3</td>
</tr>
<tr>
<td>DCICA 23</td>
<td>-</td>
<td>4.0</td>
<td>11.7</td>
<td>-</td>
<td>12.3</td>
</tr>
</tbody>
</table>

6. CONCLUSION

We proposed utilizing a directivity pattern based continuous mask and ICA for BSS when speech signals outnumber the sensors. Our method avoids discontinuous zero-padding, and therefore, can separate the signals with no musical noise. Moreover, our new mask does not involve any heuristic parameters.

REFERENCES