Original Paper

Virtual Microphone Technique for Binauralization for Multiple Sound Images on 2–Channel Stereo Signals Detected by Microphones Mounted Closely

R. Jinzai¹, K. Yamaoka¹,², S. Makino¹,³, N. Ono², T. Yamada¹ and M. Matsumoto¹*

¹University of Tsukuba, Japan
²Tokyo Metropolitan University, Japan
³Waseda University, Japan

ABSTRACT

Because inter–channel time differences (ICTDs) between signals detected by real microphones mounted close to each other are much smaller than inter–aural time differences (ITDs) for sound image localization, sound images are localized at azimuths different from those of sound sources. In this paper, we propose a virtual microphone technique, which simulates binaural signals by equalizing ICTDs to ITDs, to localize sound images at azimuths of the sound sources with reference to the real microphones. Binaural signals simulated by the proposed method were examined objectively and subjectively by tests on two-sound-image localization. The tests revealed that the two sound images were localized at azimuths of the sound sources with reference to the real microphones.

Keywords: Virtual microphone, extrapolation, binaural signal, 2–ch stereo signals, array signal processing.

*Corresponding author: M. Matsumoto, matsumotojazz814@gmail.com.

Received 29 November 2022; Revised 27 May 2023
ISSN 2048-7703; DOI 10.1561/116.00000079
© 2023 R. Jinzai, K. Yamaoka, S. Makino, N. Ono and T. Yamada and M. Matsumoto
1 Introduction

A head-related transfer function, HRTF, is a well-known transfer function between a sound source and a pair of ears. HRTFs are applied to simulate binaural signals for localizing a sound image by convolving with a source signal [3]. A wide variety of methods for simulating HRTFs and binaural signals with signals detected by a microphone array have been studied. As an example, a method for localizing a sound image with signals picked up by a tetrahedral microphone array is proposed [5]. The method is composed of analysis and synthesis steps. In the analysis step, a number of sound sources is identified and directional information of the sound sources is extracted. In the synthesis step, binaural signals are synthesized with pre-measured HRTFs assigned with the directional information. In the paper, sparsity between signals of the sound sources is assumed.

In another example, a sound field captured by a spherical microphone array is represented with spherical wave functions or plane wave expansions. Furthermore, binaural signals are synthesized with HRTFs and signals reconstructed with the functions [7]. In addition, binaural signals are simulated by an inverse wave propagation method with signals detected by linear and circular microphone arrays [26, 28]. In these studies, binaural signals synthesized are evaluated objectively and subjectively. Within a few years, a variety of microphone arrays and microphone arrangements, including dummy head microphone, are proposed and are examined objectively and subjectively [18, 19]. Six 3D microphone techniques for binaural listening are evaluated subjectively [6]. Furthermore, an improved binaural signal matching with arbitrary array is proposed [9].

These methods exhibit good performance. Array signal processing, however, requires many microphones [27]. For example, four microphones are mounted at the four vertices of a tetrahedron and 60 microphones on a linear array are applied. The availability of such a variety of microphone arrays in addition to so many microphones is quite questionable.

For the human auditory system, binaural listening enables us to perceive spatial impressions. In an exemplary arrangement, a loudspeaker as a sound source is located in the horizontal plane. In this case, an arrival time difference of sound waves reaching to the right and left ears is called the inter-aural time difference “ITD”. ITD depends on an azimuth of a sound source relative to both ears of a listener and has a maximum value of the time difference corresponding to a sound source located laterally. Additionally, a difference in sound intensity between both ears is called the inter-aural level difference “ILD”. According to Duplex Theory [3, 22], ITD and ILD are dominant cues for localizing sound images in the horizontal plane. In particular, ITD is dominant for localization at frequencies below 1.5 kHz. ITD has individuality and remains a hot research topic [12, 17].
In recent years, IC recorders and smartphones having a recording capability have been quite popular. On such products, microphones for stereo recording are mounted quite close to each other. Therefore, an arrival time difference between signals detected by the microphones is necessarily small in value. As microphones on a small recording device are mounted closely, a distance between the microphones is different from that between both ears, that is, inter-channel time difference, ICTD between signals detected by the microphones of a sound source is smaller than ITD between binaural signals of the sound source at both ears. Therefore, an azimuth of the sound source relative to the microphones is different from that of a sound image perceived from the detected signals by the microphones. Related to this issue, methods for sound source separation, re-panning, and up-mixing are proposed [1, 2, 4]. These methods are applied to signals of two-channel stereophonic and multichannel audio formats with amplitude panning.

In the case of a single sound source, a sound image of the sound source is basically perceived by a pair of signals of the sound source detected by two microphones. Further, the sound image is re-panned by simple time-delaying or phase-shifting the signal detected by the microphone contralateral to the sound source [20]. For multiple sound sources, however, these methods are partly applicable to re-panning multiple sound images. Suppose that two sound sources are located at opposite azimuths referred to two microphones, that is, one sound source is located relatively close to the left microphone and another is relatively close to the right microphone. It is difficult to apply time delaying or phase shifting for re-panning multiple sound images for such multiple sound sources at opposite azimuths referred to the two microphones.

The technical term “virtual microphone VM” describes a signal processing method for signal interpolation using the signals detected by real microphones. For example, a method is proposed to improve the performance of beam forming in under-determined conditions (the number of microphone is fewer than that of sound sources) by increasing the number of virtual microphones [14]. The technique has been applied to noise reduction and speech enhancement [13–16]. As another example, a method is proposed for enhancing the spatial resolution of a microphone array with VM [26].

Previously, we proposed a method for estimating a signal of a virtual microphone placed at an arbitrary position by extrapolating signals of real microphones [11]. We applied the method to locate a virtual microphone at the position acoustically equivalent to the inter-aural distance [10]. Through this method, ICTD between signals of a real microphone mounted closely are equalized to ITD between signals at both ears, and binaural signals are simulated to localize a sound image to be perceived at an azimuth of a sound source with reference to the real microphone. The simulated binaural signals are examined objectively. In this paper, the signals are evaluated by a subjective test. Furthermore, VM applied for binaural listening is summarized.
This paper contains 5 sections. Following the introduction in Section 1, a method for interpolation with VM is reviewed. In addition, a new method for extrapolation with VM is described in Section 2. In Section 3, the method is applied for simulating binaural signals with signals detected by real microphones mounted closely and the simulated binaural signals are objectively examined in terms of the phase difference, the arrival time difference between waveforms, and the inter-aural cross correlation, IACC [3]. Furthermore, in Section 4, the simulated binaural signals are psycho-acoustically evaluated by a subjective test. This paper is concluded in Section 5.

2 Virtual Microphone, VM, Technique

2.1 W-DO Assumption

In cases in which multiple sound sources are active in a real acoustic environment, signals detected by microphones are complicated. In such cases, we assume that the sources show W-disjoint orthogonality, W-DO [24, 25, 29]. W-DO is the sparsity of a signal in the time-frequency domain, which is any time-frequency slot of a short-time Fourier transform, STFT is regarded as being occupied by a signal of one sound source only. Therefore, methods for interpolation and extrapolation described later are applicable to a signal in a time-frequency slot.

2.2 Interpolation with VM Technique

In this section, a method for interpolating a virtual microphone by the VM technique is reviewed [14]. An arrangement of two real microphones $M_1, M_2$ and a virtual microphone $M_v$ is shown in Figure 1. Suppose that $x_i(\omega, t)(i = 1, 2)$ are signals detected by the real microphones $M_i(i = 1, 2)$ at an angular frequency $\omega$ in a time frame in the time-frequency domain. $x_v(\omega, t)$ denotes a signal of the virtual microphone. In the VM technique, the phase and amplitude of a signal of the virtual microphone are interpolated independently. The phase $\phi_i$ and amplitude $A_i$ of a signal $x_i(\omega, t)(i = 1, 2)$ detected by a real microphone are expressed as

$$\phi_i = \angle x_i(\omega, t) = \tan^{-1} \frac{\text{Im}(x_i(\omega, t))}{\text{Re}(x_i(\omega, t))}$$ (1)

$$A_i = |x_i(\omega, t)|.$$ (2)

Furthermore, the phase $\phi_v$ of a signal of the virtual microphone is interpolated linearly as

$$\phi_v = \phi_1 + \alpha (\phi_2 - \phi_1)$$ (3)

$$= (1 - \alpha) \phi_1 + \alpha \phi_2.$$ (4)
the phase of the signal of the virtual microphone is interpolated on the 
assumption that
\[ |\phi_2 - \phi_1| \leq \pi \] 

(5)
because the phase is periodic for a natural number \(n\) in \(\phi_v \pm 2n\pi\).

Because the amplitude of a signal of a virtual microphone depends on 
various acoustical conditions, for example, arrival directions of sound waves
and reverberation, it is a hard task to model the amplitude faithfully. Therefore,
\(\beta\)-divergence as a substitute for a physical model is applied for interpolating
the amplitude in the VM technique. The amplitude is interpolated as
\[
A_v = \begin{cases} 
exp \left( (1 - \alpha) \log A_1 + \log A_2 \right) & (\beta = 1) \\
\left( (1 - \alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1} \right)^{\frac{1}{\beta-1}} & \text{(otherwise)}
\end{cases}
\] 

(6)

With the parameter \(\beta\), it is possible to non-linearly interpolate the amplitude
of a signal of the virtual microphone from the amplitudes of signals detected
by the two real microphones. From the above, a signal \(x_v(\omega, t)\) interpolated
for the virtual microphone is represented as
\[
x_v(\omega, t) = A_v \exp(j\phi_v). 
\] 

(7)

2.3 Extrapolation with VM Technique

An arrangement of two real microphones and one virtual microphone for 
extrapolation is shown in Figure 2. For phase extrapolation, (4) is applied, 
which is the equation previously applied in the phase interpolation. As
described in Section 1, because ITD related to the inter-aural phase difference
is a dominant cue at frequencies below 1.5 kHz [3, 22] and the difference in
amplitude owing to the distance between the two real microphones is small, the
amplitude of the signal of the real microphone closer to the virtual microphone
is considered to be the amplitude of a signal of the virtual microphone as described below.

\[ A_v = \begin{cases} \frac{A_1}{\alpha} & \alpha < 0 \\ \frac{A_2}{0} & 0 < \alpha. \end{cases} \] (8)

An extrapolated signal of the virtual microphone is represented below similarly to the interpolated signal as follows,

\[ x_v(\omega, t) = A_v \exp(j\phi_v). \] (9)

In this paper, the VM technique is applied to equalize ICTDs between microphones mounted closely to ITDs between both ears for localizing multiple sound images by binaural listening.

3 Objective Evaluation

In this section, the method for extrapolating a signal of the virtual microphones proposed in the previous section is applied to simulate binaural signals, and the proposed method is examined objectively.

3.1 Experimental Conditions

The layout of sound sources and real microphones in the objective evaluation is shown in Figure 3. Other experimental conditions are shown in Table 1. In this experiment, two sound sources denoted \( S_1 \) and \( S_2 \), which are respectively located at the azimuths of 10° and 170°, are assumed. In addition, two real microphones denoted \( M_1 \) and \( M_2 \) are mounted \( d = 2.83 \) cm apart. Furthermore, a virtual microphone \( M_v \) is mounted virtually at a distance from \( M_1 \) that is \( \alpha \) times as long as the distance \( d \). Impulse responses \( h_{1L}(n) \), \( h_{1R}(n) \), \( h_{2L}(n) \) and \( h_{2R}(n) \) in the time domain shown in Figures 3 and 6 are described in the next subsection.

As an example, the power spectrum of a speech of Female Japanese \( s_1(n) \) at \( S_1 \) and that of Male English \( s_2(n) \) at \( S_2 \) in a time frame are shown in
Virtual Microphone Technique for Binauralization

Figure 3: Arrangement of real and virtual microphones in objective evaluation.

Table 1: Experimental conditions.

<table>
<thead>
<tr>
<th>Sampling rate</th>
<th>8 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance between</td>
<td>2.83 cm</td>
</tr>
<tr>
<td>real microphones</td>
<td></td>
</tr>
<tr>
<td>Reverberation time</td>
<td>300 ms</td>
</tr>
<tr>
<td>FFT frame length</td>
<td>1024 samples</td>
</tr>
<tr>
<td>FFT hop size</td>
<td>256 samples</td>
</tr>
<tr>
<td>Speech for $S_1$</td>
<td>Female Japanese</td>
</tr>
<tr>
<td>Speech for $S_2$</td>
<td>Male English</td>
</tr>
</tbody>
</table>

Figure 4. In each spectrum, multiple peaks, whose power exceeds 0 dB in vertical axis, can be seen. The signal $s_1(n)$ has peaks at frequencies close to 220, 440, 720, 960 and 1200 Hz. The signal is dominant at these frequencies in the frame. On the other hand, the signal $s_2(n)$ is dominant at frequencies close to 120, 250, 400 and 550 Hz. Therefore, there are no overlaps (i.e., “sparsity”) between these peaks and W-DO between the two speeches is satisfied. Under the condition that sparsity between signals of sound sources is unsatisfied, sound images based on “summing localization” will be perceived [3].

3.2 Determination of Extrapolation Coefficient $\alpha$

Figure 5 shows changes in ICTD between the signals $x_1(n)$ and $x_2(n)$ detected at the real microphones $M_1$ and $M_2$ attributable to changes in the azimuth of the sound source $S_1$ at 20° intervals from 10° to 170° in the setup shown in Figure 3. Additionally, the changes in ITD of HRTFs of the Kemar dummy head, are also shown in Figure 5 [8]. Both ICTD and ITD change proportionally to the azimuth of the sound source, and the ratio of ITD to ICTD is close to 8 at every azimuth of the sound source. Therefore, the coefficient for extrapolation shown in Figure 2 is assumed to be 8. The coefficient is acoustically equivalent
Figure 4: Power spectra of speeches $s_1(n)$ and $s_2(n)$.

Figure 5: Changes in ICTD for microphones mounted closely and ITD of the Kemar dummy head as a function of azimuth of sound source.
to the ratio of the distance between both ears of the dummy head to that between the real microphones in Figure 3. The coefficient $\alpha$ depends on individuality because the binaural distance is individual. Localization error due to non-individuality, however, appears in the median plane significantly [21].

Signal processing on extrapolation in the objective and subjective evaluations described in the next section is shown in Figure 6. For example, the signals $s_1(n)$ and $s_2(n)$ denote speeches in Japanese and English, by female and male speakers. $h_{1L}(n)$ denotes an impulse response between the microphone $M_1$ and the sound source $S_1$, and $h_{1R}(n)$ denotes that between $M_2$ and $S_1$. Similarly, $h_{2L}(n)$ and $h_{2R}(n)$ denote the impulse responses of $M_1$ and $M_2$ with respect to $S_2$, respectively, as shown in Figure 3. In this paper, a set of impulse responses measured with adjacent microphones in a line array in RWCP Sound Scene Database was adopted [23]. The responses were measured in a room at a reverberation time of 300 ms, as shown in Table 1. The signals $x_1(n)$ and $x_v(n)$ in Figure 6 as the simulated binaural signals are evaluated objectively in this section and subjectively in the next section.

### 3.3 Numerical Evaluation

Here, the simulated signals $x_1(n)$ and $x_v(n)$ shown in Figure 6 are examined in terms of the phase difference in the frequency domain, the arrival time difference between waveforms in the time domain, and inter-aural cross correlation, IACC from the binaural viewpoint.

Firstly, the signals $x_1(n)$ and $x_v(n)$ in Figure 6 are examined with regard to the phase difference. Figure 7(a) shows phase differences between the signals $x_1(n)$ and $x_v(n)$ for the extrapolation coefficient $\alpha = 1$, that is, without the VM technique. As shown in Figure 3, because $M_1$ is a microphone ipsilateral
Figure 7: Objective evaluation without VM technique ($\alpha = 1$).
to $S_1$, the phase of the signal $x_1(n)$ for $s_1(n)$ at $M_1$ is advanced relative to that of $x_v(n)$ at $M_v$ (solid line). In contrast, for $S_2$, because $M_1$ is contralateral to $S_2$, the phase of the signal $x_1(n)$ for $s_2(n)$ at $M_1$ is delayed relative to that of $x_v(n)$ at $M_v$ (dotted line).

Waveforms of the signals $x_1(n)$ and $x_v(n)$ are shown in Figure 7(b). In the figure, the $x$–axis denotes the relative time in the frame. Because $S_1$ is closer to $M_1$ than to $M_v$ the sound wave of $s_1(n)$ arrives at $M_1(x_1(n)$, solid line) slightly earlier than at $M_v$, ($x_v(n)$, dotted line). In contrast, in Figure 7(c), because $S_2$ is located farther from $M_1$ than from $M_v$, the arrival of the sound wave $s_2(n)$ at $M_1(x_1(n)$, solid line) is slightly delayed relative to that at $M_2(x_v(n)$, dotted line).

Furthermore, the simulated signals $x_1(n)$ and $x_v(n)$ are evaluated by IACC, which is a well-known measure for evaluating binaural signals, and the time delay $\tau$ at which IACC yields its maximum is considered to be ITD. In Figures 7(d) and 7(e), IACCs are close to 1 at $\tau = 0$. This implies that a sound image to be perceived with the signals $x_1(n)$ and $x_v(n)$ as binaural signals in Figures 7(b) and 7(c) will be localized in the median plane because of the small ICTD as the ITD.

In contrast, for the extrapolation coefficient $\alpha = 8$, that is, with the VM technique, Figure 8(a) shows that the phase difference between the signals $x_1(n)$ and $x_v(n)$ detected at $M_1$ and $M_v$, respectively, is eight times that shown in Figure 7(a).

Similarly, the arrival time difference between the waveforms of the signals $x_1(n)$ and $x_v(n)$ shown in Figure 8(b) (i.e., the difference between the straight line and the dotted line) is also eight times that shown in Figure 7(b). This shows that the acoustical distance between $M_1$ and $M_v$ is eight times that between $M_1$ and $M_2$ in accordance with the extrapolation coefficient $\alpha$.

Furthermore, as shown in Figure 8(d), the IACC between the signals $x_1(n)$ and $x_v(n)$ for $S_1$ is maximum at $\tau = -0.58$ ms. This means that a listener will perceive a sound image on the left because of the sufficient ICTD as the ITD. Similarly, as shown in Figure 8(e), when the IACC for $S_2$ is maximum at $\tau = 0.52$ ms, a sound image will be localized on the right of the listener. Sound images localized with the VM technique are evaluated subjectively in the next section.

4 Subjective Evaluation

The binaural signals examined objectively in the previous section are evaluated subjectively in this section. Because none of the subjects who participated in this test had experienced such a listening test, a listening test for localizing one sound image with one sound source as a preliminary test was conducted before evaluation of two sound images with simultaneous two sound sources.
Figure 8: Objective evaluation with VM technique ($\alpha = 8$).
4.1 Equipment and Subjects

Signal processing for binaural signals simulated with the VM technique for two sound sources is shown in Figure 6. In the preliminary test with one sound source, signal processing with $s_1(n)$, $h_{1L}(n)$ and $h_{1R}(n)$ with the VM technique (upper half in Figure 6) was performed.

Binaural signals as stimuli were reproduced over AKG K240 headphones at a normal listening level with the AT-HA50 headphone amplifier. The tests were conducted in a soundproof listening room, whose background noise level is 27 dB (A).

Four subjects, postgraduate students with ages ranging from 23 to 27, participated in the test. None of them had hearing loss. The test was performed individually. The subjects were instructed to determine the azimuth of a sound image by referring to Figure 9. The subjects were allowed to listen to the stimuli repeatedly until they were completely confident of their answers, which were collected in a PC.

![Figure 9: Azimuths of sound sources in subjective test.](image)

4.2 Conditions of Preliminary Test with One Sound Source

As shown in Figure 9, 20° that is deference in azimuth between adjacent sound-sources at the front and 40° at the lateral exceed minimum audible angles, MAAs, respectively.

In the preliminary test, the subjects were instructed to determine an azimuth of a sound image for one sound source. A stimulus in the preliminary test was composed of preceding pips as a marker and one speech for evaluation of its azimuth. A subject evaluated a total of 56 stimuli ($7$ azimuths $\times 2$ coefficients ($\alpha = 1, 8$) $\times 4$ repetitions). Combinations of an azimuth of the sound source and a coefficient were determined randomly for each subject.

One Japanese sentence and one English sentence spoken by two native male and two native female speakers were chosen as speeches for $s_1(n)$ and $s_2(n)$ in Figure 6. Each speech was nearly 2.7–second long. A combination of the two sentences and the four speakers was randomly assigned to each
stimulus. All replies of the subjects are shown in figures 10 through 13 because the number of the subjects is too small to analyze statistically.

### 4.3 Conditions of Subjective Test with Two Sound Sources

Two sound sources were used in the test. As shown in Figure 9, one was located at one of the four azimuths of 10°, 50°, 70° and 90° in the left quadrant for a listener and the other was located at 90°, 110°, 130° and 170° in the right quadrant. Combinations of the four azimuths in the left quadrant and those in the right quadrant were randomized for each stimulus.

In the test, extrapolation for binaural signals simulated with the VM technique for the two sound sources is shown in Figure 6. A stimulus was composed of preceding pips as a marker and two speeches for evaluation of their azimuths. A subject evaluated a total of 32 stimuli (4 azimuths × 4 azimuths × 2 coefficients α). Two Japanese and two English sentences spoken by two native male and two native female speakers were chosen as the speeches for $s_1(n)$ and $s_2(n)$ in Figure 6. The speeches for $s_1(n)$ and $s_2(n)$ were always spoken

![Figure 10: Subjective evaluation for one sound source without VM technique ($\alpha = 1$).](image-url)
simultaneously. Each speech was nearly 2.7–second long. A combination of the four sentences and the four speakers was randomly assigned to each stimulus. The subjects were instructed to identify two azimuths of sound images, one in the left quadrant and the other in the right quadrant, referring to Figure 9.

### 4.4 Results and Discussion

The results for one sound source in the preliminary test are shown in Figure 10 for the subjective evaluation without the VM technique ($\alpha = 1$) and are shown in Figure 11 for the subjective evaluation with the VM technique ($\alpha = 8$). In every figure, the x-axis and y-axis denote the azimuth of the sound source and the perceived azimuth of the sound image, respectively. Furthermore, gray-scale shows the statistical frequency of the subjects’ answers.

In Figure 10, for all subjects, the sound images perceived are mostly localized at azimuths from $70^\circ$ to $130^\circ$, which are close to the center (in the front). Furthermore, even when sound sources are located laterally at azimuths of $10^\circ$ and $170^\circ$, sound images are less frequently perceived at azimuths of $10^\circ$,
50°, 130°, and 170°. The perceived azimuths of the sound images are different from those of the sound sources.

Regarding the answers of subjects C and D, some confusion, left–right confusion for subject C and front–side confusion for subject D was seen in their answers. According to their introspections, under the condition without VM technique ($\alpha = 1$), they perceived sound images in their heads, and it was a hard task for them to perceive azimuths of the sound images.

Results under the condition with the VM technique ($\alpha = 8$) are shown in Figure 11. In contrast to those shown in Figure 10, every subject perceived sound images to be close to the azimuths of the sound source. For example, for sound sources at the azimuths of 10° and 170°, sound images are mostly localized at the azimuths of 10° and 170°, respectively. Furthermore, a sound image for the sound source at the azimuth of 90°, that is, in the front, was perceived to be close to the azimuth and was not localized laterally.
Results for two sound sources are shown in Figures 12 and 13, for the subjective evaluation without and with the VM technique, respectively. In the subjective evaluation without the VM technique ($\alpha = 1$) shown in Figure 12, for the sound sources located at azimuths from $10^\circ$ to $90^\circ$, that is, in the left quadrant, and at azimuths from $90^\circ$ to $170^\circ$ in the right quadrant, two sound images were perceived to be mostly within azimuths from $70^\circ$ to $130^\circ$. That is, sound images were less localized laterally. Azimuths of the sound images were different from those of the sound sources.

In the subjective evaluation with the VM technique ($\alpha = 8$) shown in Figure 13, azimuths of sound images are mostly close to those of the sound sources, which is in contrast to the subjective evaluation of without the VM technique. Sound images for sound sources located laterally are localized laterally. With the VM technique, sound images are localized at the sound source azimuths with reference to the two real microphones shown in Figure 3.
5 Conclusion

Because microphones on a small device are mounted close to each other, ICTDs between signals detected by the microphones are too small in value to localize sound images with the signals at the azimuth of the sound source with reference to the microphones.

This paper describes a newly developed technique that simulates binaural signals derived from outputs of two real microphones placed close to each other. ICTDs of the real microphone outputs are made equalized to ITDs caused from two sound sources, thereby two sound images are perceived respectively as if the signals of the two sound sources to be listened to are detected by two microphones placed at binaural distance.

The simulated binaural signals were objectively examined in terms of the phase difference, and also examined in the arrival time difference and IACC. Moreover, the simulated binaural signals are evaluated by a subjective test performed on two sound image localization. The test revealed that two sound images are localized at azimuths of the sound sources with reference to the microphones. The proposed method enables binaural reproduction of 2 channel signals recorded by microphones mounted on rather small devices. This method enables localization of multiple sound images without given information on sound sources or HRTFs.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

Ryoga Jinzai performed the experiments and wrote the majority of the manuscript, and other authors reviewed and revised the manuscript. All authors made contributions to the conception and design of the work, analyzed the data, and interpreted the results. All authors read and approved the final manuscript.

Funding

This research work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI through a Grant-in-Aid for Scientific Research under Grants 16H01735, 19H04131 and 19J20420, and the SECOM
Science and Technology Foundation. In addition, the authors are grateful to the subjects for participating in the subjective test.

References


