Subjective evaluation of auralization using a directional sound source that simulates a trumpet

Ryoko Hara¹,*, Takahiro Iwami¹,† and Akira Omoto²,‡

¹Graduate School of Design, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan
²Faculty of Design, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan

(Received 30 June 2021, Accepted for publication 14 July 2021)

Keywords: Auralization, Musical instrument directivity, Subjective evaluation, Spherical harmonics

1. Introduction

Auralization is a tool for simulating and reproducing the listening environment of virtual or real spaces and making them audible. This technique is based on the convolution of room impulse responses (RIRs) with anechoically recorded music. In most cases, RIRs used in auralization are the ones modeled by geometric or wave-based acoustic simulation, or measured ones. Although various factors such as wall data and room volume are considered to improve the accuracy of auralization, sound sources are usually assumed to be omnidirectional throughout the entire process. However, sound sources in actual situations, such as musical instruments, have their own directivities [1]. Therefore, considering the sound source directivity may bring us one step closer to more realistic auralization.

It has been discovered that some changes in the directivity of the sound source can be perceivable in auralization using simulated RIRs [2]. In addition, there have been various attempts to construct a sound source whose directivity can be controlled [3–5]. However, previous studies face high hurdles to be introduced in RIR measurements in real space. There have been few cases of subjective evaluation of auralization using the measured RIRs, which focus on the sound source directivity.

The priority in our study is the balance between the accuracy of simulating the musical instrument directivities and the ease of implementation. We have thus attempted to synthesize musical instrument directivities by sequentially rotating a commonly available loudspeaker through various directions and combining directional characteristics with appropriate weightings [6]. In this Letter, we present the preliminary results of a small-scale subjective evaluation experiment to investigate the perceptual differences between realistic reproductions using signals recorded by a microphone array (referred to as “real performance” hereafter) and reproductions made by the convolution of a dry source with an impulse response obtained by taking into account the directivity of the musical instrument.

2. Experimental method

The subjective evaluation was performed by comparing the auralized sound that considers the sound source directivity with real performance as reference. Recently, we have developed a sound-field reproduction system with 24-channel narrow-directional microphones and 24-channel loudspeakers [7,8]. The same system was used in this experiment.

2.1. Measurement and recording

The measurement of RIRs and the recording of real performance were carried out in a concert hall (room volume of 10,400 m³ and seating capacity of 800). The fixed positions of the sound source, performer, and receiver are presented in Fig. 1(a). The microphone array shown in Fig. 1(b) was always placed at the receiver point with the center of the array at a height of 1.2 m above the floor. This microphone array consists of 24 narrow-directional microphones arranged at intervals of both the polar and azimuthal angles of 45°.

Firstly, we placed a commonly available three-way coaxial loudspeaker (GENELEC 8351) and its mounting that can rotate a loudspeaker manually in various directions (see Fig. 1(c)) at the sound source position with the center of the loudspeaker at a height of 1.2 m above the stage floor. The loudspeaker was rotated in 26 directions and the 26-pattern RIRs were measured. The 26 directions include 24 directions in 45° steps for both polar and azimuthal angles, the north pole, and the south pole. This combination of directions (hereinafter called “Step45”) was determined by considering the ease of rotations and the results of numerical simulations conducted to verify the accuracy of simulating musical instrument directivities [9].

Next, we asked the performer to sit on a chair placed at the same position as the sound source and play the trumpet. The performance was recorded simultaneously with a cardioid microphone placed immediately in front of the trumpet bell and the microphone array placed at the receiver point. The former was used as an anechoic signal that was subsequently convolved with RIRs for auralization, and the latter was used as a reference signal.

2.2. Synthesis of room impulse responses

The directivity of the musical instrument was simulated in the spherical harmonic domain. Although it is time-consuming to measure the musical instrument directivities accurately, detailed measured directivities of up to the fourth-order spherical harmonic coefficients are available as an open source [10]. We used them as the target directivity and denoted the coefficients as the vector $\phi_{des}$. $\phi_{des}$ is available at every frequency, but is explained for a single frequency below.
The polar angle and weights denoted as the vector \( \mathbf{w} \) obtained in this manner was denoted as measured at 2,522 points on a sphere [11,12]. The maximum obtained by spherical harmonic expansion using the data speaker is rotated in the direction of \( \mathbf{w} \). The angles of the loudspeaker was then converted into a vector when the loudspeaker rotated in each direction of \( \theta, \phi \). Also, \( \theta \) denotes the polar angle and \( \phi \) denotes the azimuthal angle. The weights denoted as the vector \( \mathbf{w} \) were calculated to satisfy [6]

\[
\mathbf{w} = \arg \min_{\mathbf{w} \in \mathbb{C}^2} \left\| \mathbf{c}_{des} - \sum_{i=1}^{26} w_i \mathbf{c}_{dp}(\theta_i, \phi_i) \right\|
\]

where \( \phi \) denotes the weight of the loudspeaker direction \( \theta, \phi \). Here, the inverse problem was solved with an appropriate predetermined regularization parameter.

Let \( h_i = h(\theta, \phi) \) be the RIR obtained when the loudspeaker is rotated in the direction of \( \theta, \phi \). \( h_i \) was transformed into the frequency domain using the discrete Fourier transform, and added together with each weight. Next, it was converted back into the time domain using the inverse discrete Fourier transform, which produced the RIR with the sound source that simulated the target musical instrument directivity.

### 3. Listening test

#### 3.1. Tested stimuli

A preliminary listening test was conducted to investigate how realistic our auralization sounded. The test was based on MUSHRA (MultStimuli with Hidden Reference and Anchor) [13].

A flowchart for preparing stimuli is shown in Fig. 2. There were two stimuli, each of which was a passage of around 20 s. The test contained four signals, each of which was a convolution of an anechoic signal and the RIR, as follows: (Step45) the RIR synthesized as described in Sect. 2.2; (Front) the RIR measured only with the loudspeaker direction of \( 90^\circ \) (= facing the audience); (Back) the RIR measured only with the loudspeaker direction of \( 90^\circ, 180^\circ \) (= facing the audience); and (Side) the RIR measured only with the loudspeaker direction of \( 90^\circ, 225^\circ \) (= turned to the left when viewed from the audience). The hidden anchor was prepared by passing a reference signal through a low-pass filter with a cutoff frequency of 1.5 kHz.

All the signals were processed through the signal processing of “beamforming + direct,” which was the combination of beamforming in the low-frequency range and direct reproduction in the other ranges [7]. Direct is the method in which the sounds captured by the 24-channel microphones are reproduced directly without signal processing by the 24-channel loudspeakers, whose arrangement is almost the same as the microphones. Beamforming + direct is one of the basic methods of achieving better performance with our sound-field reproduction system [8]. By comparing the measured wavefronts, it has become clear that this method can reproduce the original sound field with high accuracy in the frequency range of most musical instruments [7].

#### 3.2. Tested variables

Subjects were asked to rate the samples by comparison with the reference signal, regarding the five variables described below.

- **Spaciousness:** spread of sound
- **Envelopment:** sound reflected from side or rear
- **Depth:** distance between sound source and listener
- **Localization:** location of sound source
- **Timbre:** sound quality of trumpet

The grading scale was continuous from “very different” to “very similar” on a scale of 0 to 100, a score of 100 being “very similar.” The subjects were instructed to give a score of 100 to the signal they felt was the most similar to the reference. In addition to the evaluation work, the subjects were asked to provide descriptive answers as to how different signals were from the reference. This work was intended to...
capture the trend of the difference, for example, if the
difference of spaciousness was perceived as “stronger,”
“weaker” or other.

3.3. Test environment

The test was conducted by setting 24-channel loud-
speakers (GENELEC 8020) in an anechoic chamber as shown
in Fig. 3(a), and the detailed layout is shown in Fig. 3(b).
Each loudspeaker faced the center of the subject’s head.

The subjects were 10 students, all of whom reported
normal hearing and had received technical ear training for
more than one year. The number of subjects was chosen for
the convenience of asking for both ratings and descriptions.
Each subject was allowed to switch instantly among seven
signals (1 reference + 4 auralized signals + 1 hidden refer-
ence + 1 hidden anchor) by using our original graphical user
interface. Each signal was adjusted so that the equivalent
continuous A-weighted sound pressure level at the receiving
point was the same. The order of presenting signals and
variables was randomized.

3.4. Results

Although the answers by those who could not distinguish
the hidden reference in the training session, which was
conducted in advance of the test, were to be excluded through
a prescreening process, none were ruled out. Figure 4 shows
average of all the scores for each signal evaluated. Error bars
indicate 95% confidence intervals. The results shown in Fig. 4
contain the scores for all stimuli. Since common expressions
were seen in the descriptive answers, the individual dif-
fferences of the stimuli were considered to be small.

In Fig. 4, the difference between each signals and the
reference was perceived to some extent for every evaluated
variable. The results of the multiple comparison analysis
showed that there was a significant difference between each
auralized signal and the hidden reference ($p < 0.05$). There-
fore, it is clear that there is a perceivable difference between
the auralized sounds and the real performance.

On the other hand, there was no significant difference
between the four auralized signals for the four evaluated
variables of spaciousness, envelopment, depth, and local-
ization ($p > 0.05$). The only variable that could be judged
to have a significant difference was timbre. There was a
significant difference for two pairs: the pair of Step45 and
Front and the pair of Step45 and Back. As for timbre, the
perceptual characteristics of each signal described were
relatively common among the subjects. For example, in the
case of Front, the sound was perceived to have a hard sound
impression owing to the stronger direct sound, with fewer
low-frequency components. Additionally, Back was per-
ceived to have much reverberation and sounded like a french
horn, and Step45 had a muffled sound with strong low-
frequency components.

The muffling of Step45 sounds was considered to depend
on the regularization parameters set when calculating the
weights. Although there are some methods used to determine
the parameters mathematically, there is no perfect solution for
practical implementation that is suitable for obtaining the
desired sound quality. Therefore, the empirical determina-
tion of parameters is often required. In fact, we determined the
parameters for each 1/3-octave band by listening with our
own ears while changing the values, but changing them for
each frequency bin is preferable. Thus, the timbre for
Step45 may be improved by finding a better combination
of parameters.

The same tendency was observed for depth, and local-
ization as for timbre. In the case of depth, most of the subjects
answered that the sense of depth “decreased” in Front
compared with the reference, while they answered “increas-
ed” in Side, Back and Step45. As for localization, Front
and Back presented a clearer sense of localization, while Side
was slightly blurred. In addition, there were some responses
that accurately captured the direction of the sound source,
such as the lateral blurring of Side and the blurring of Back
along the depth direction. Although they did not appear in the
scores, there seemed to be some perceivable differences.

However, the differences for the variables related to
spatial impressions, such as spaciousness and envelop-
ment,
were smaller than those for the other three variables, and the descriptive answers showed a mixture of different opinions. Therefore, the perceptual difference was considered to be insignificant.

4. Discussion

There are some issues to be discussed. Firstly, in this listening test, we used sound-field reproduction through loudspeakers instead of binaural reproduction so that the accuracy of reproduction would not be affected by individual auditory characteristics. However, the accuracy of the reproduction of the reference is another issue that must be addressed.

Secondly, the trumpet used in the recording was not the same as that used in the directivity open data. Moreover, we have only considered the directivity up to the fourth-order components, but we must rethink how much the fifth- and higher-order components actually contribute.

Thirdly, the auralized signals might be near the threshold since many subjects commented that it was difficult to distinguish all of the signals. In particular, the primary sound field used in this study was a relatively reverberant space with a reverberation time of about 2.3 s at 500 Hz. If the primary field is less reverberant, the perceptual difference in auralized signals resulting from the difference in sound source directivity can become more pronounced. More experiments should be conducted in the future to determine whether or not the results depend on the characteristics of the primary sound field.

Finally, the trumpet has a directivity that is relatively close to that of the loudspeaker and it is simple among the many musical instruments. The results may indicate that it is sufficient to use the RIR when the loudspeaker is rotated in the single most dominant direction for the auralization of such musical instruments, considering the difficulty in determining the regularization parameters. Thus, auralization for various musical instruments should be performed in the future.

5. Conclusion and future work

In this Letter, we presented a subjective evaluation of auralization using a directional sound source that simulates trumpet directivity. The results showed that there was a perceivable difference between the auralized signals and the reference signal. As for the change in sound source directivity in auralization, there was a significant difference for timbre between some of the signals. Similar results can be obtained by increasing the number of subjects for depth and localization. Note that the auralized sound that considers trumpet directivity has room for improvement by finding a better combination of regularization parameters during the auralization process.

However, the tests conducted in this study were only preliminary and it was still too early to draw any final conclusions. As a future work, we would like to design a subjective evaluation on a larger scale to investigate how much detail of the sound source directivity should be simulated for various musical instruments and sound fields.

Acknowledgement

This research was supported by JSPS KAKENHI under grant number JP 17H00811, 21H03764.

References