Sound Localization Using an Acoustical Telepresence Robot: TeleHead II

Abstract

TeleHead I is an acoustical telepresence robot that we built on the basis of the concept that remote sound localization could be best achieved by using a user-like dummy head whose movement synchronizes with the user's head movement in real time. We clarified the characteristics of the latest version of TeleHead I, TeleHead II, and verified the validity of this concept by sound localization experiments. TeleHead II can synchronize stably with the user's head movement with a 120-ms delay. The driving noise level measured through headphones is below 24 dB SPL from 1 to 4 kHz. The shape difference between the dummy head and the user is about 3% in head width and 5% in head length. An overall measurement metric indicated that the difference between the head-related transfer functions (HRTFs) of the dummy head and the modeled listener is about 5 dB. The results of the sound localization experiments using TeleHead II clarified that head movement improves horizontal-plane sound localization performance even when the dummy head shape differs from the user's head shape. In contrast, the results for head movement when the dummy head shape and user head shape are different were inconsistent in the median plane. The accuracy of sound localization when using the same-shape dummy head with movement tethered to the user's head movement was always good. These results show that the TeleHead concept is acceptable for building an acoustical telepresence robot. They also show that the physical characteristics of TeleHead II are sufficient for conducting sound localization experiments.

1 Introduction

Our goal is to develop an acoustical telepresence robot that can be operated from a remote place and accurately transmit the sound environment from its location to users. One of our ideas for accomplishing this goal is to equip the robot with a dummy head whose shape closely matches the shape of the user's head. The head and external ears (pinna) are large enough to change the characteristics of sound waves in the range of audible frequencies. Thus, the head and pinna shapes affect the acoustic transfer from a sound source to the ears. The mathematical functions describing the transformation of sound are known as head-related transfer functions (HRTFs). By convolving a pair of HRTFs with a sound source signal, binaural acoustic information is formed.
The brain computes the sound source position in a 3D space based on this information. Namely, it is the head shape that provides the binaural information that is crucial for reproducing 3D sound (Møller, 1992; Wenzel, 1996).

Our other idea is that the user-like dummy head should move synchronously with the user’s head movement to accurately transmit a sound environment. When one hears a sound, there is often some accompanying movement of the head and body. The position of the sound source relative to the head varies with head movement. The head movement alters the HRTFs, which are convolved with a sound source signal; thus, binaural acoustic information changes with head movement. The brain computes the sound source position in a 3D space on the basis of the head movement information and the binaural acoustic information. The relationships between head movement and the sound localization function have been widely discussed. In general, head movement improves sound localization accuracy (Wallach, 1940; Perrett & Noble, 1997; Wightman & Kistler, 1999; Begault, Wenzel, & Anderson, 2001). It has also been suggested that accurate individual HRTFs are not necessarily required in order to reproduce a virtual 3D sound when a subject is allowed to move the head while listening to the sound. Namely, head movement also provides binaural information that is crucial for reproducing 3D sound (Wightman & Kistler; Begault, Wenzel, & Anderson).

Virtual 3D sounds can be reproduced using HRTF-based signal processing technologies. HRTFs are generally measured using human subjects. Measuring a set of HRTFs for a human subject is very time-consuming. Even when an automatic 3D speaker positioning system is available, as is the case in our measurement system, it takes about 90 min to measure pairs of HRTFs for a source at 143 points on a spherical surface whose center is the subject’s head. An accurate and fast HRTF measurement system has been developed (Zotkin, Duraiswami, Grassi, & Gumerov, 2004). However, HRTFs change with distance, so the number of points that should be measured in each direction remains unclear. Human subjects must remain perfectly still during the measurements, which can be difficult and uncomfortable. A dummy head never complains, but humans do. Therefore, the robotic methods we propose here have merits compared with those using premeasured HRTFs for reproducing the sound environment at all 3D positions around a human subject. Although there are some recently developed methods that rapidly measure HRTFs (Zotkin, Duraiswami, & Davis, 2004), thereby reducing subject discomfort, such systems are not commonly available.

As head size differs among individuals, so do HRTFs. The brain appears to be well adapted to using individual HRTFs in computing the location of a sound source. It is known that appropriate representation of individual HRTFs is necessary in order to reproduce accurate virtual 3D sound (Wenzel, Arruda, Kistler, & Wightman, 1993; Møller, Sorensen, Jensen, & Hammershoi, 1996; Spikofski & Fruhmann, 2001).

A sound rendered with a pair of HRTFs depends on the location of the sound source. Therefore, in computing a 3D sound, the position of the sound should be known beforehand. This does not matter in an experimental room, but it causes problems in attempting to reproduce real-world sound environments. This is because there are real and imaginary sound sources whose positions are not predefined and, moreover, because those sources are often moving in the real world.

We can avoid the drawbacks of HRTF-based 3D sound reproduction technologies by using a steerable user-like dummy head. The physical acoustics of the dummy head recreate those of the listener’s head automatically and thus always produce accurate binaural information. We have built an acoustical telexistence robot, TeleHead I, that has a user-like dummy head whose movement synchronizes with the user’s head movement in real time (Toshima, Uematsu, & Hirahara, 2003).

Some related work has been reported. Kock constructed a dummy head, named Oscar, that could rotate and showed that it reduced front-back localization errors made by listeners (Kock, 1950). Algazi, Duda, and Thompson (2004) reported a new binaural method, called motion-tracked binaural sound, that considers the user’s head rotation. The system can simulate realistic 3D sound for many users at the same time. Melick et al.
further developed motion-tracked binaural sound, which incorporated the effects of HRTFs and was used to clarify the effects of the pinna, torso, and ear location on the signals reaching the listeners’ ears (Melick, Algazi, Duda, & Thompson, 2004). All of these studies used a commercially available binaural dummy head system in which the movement of the dummy head was synchronized with the user’s head movement. These systems, however, are nonindividual and do not replicate the individual user’s head shape. On the other hand, there have been dozens of HRTF-based studies that take the user’s head movement into account (Wightman & Kistler, 1999; Begault, Wenzel, & Anderson, 2001). None of those studies, however, used measured individual HRTFs and accurate 3D head movement. Compared to these approaches, TeleHead I differs in two important ways: it uses accurate shapes that replicate the shapes of the individual listeners’ heads and reproduces the listener’s head movement by tethering the dummy head to that of the listener.

The next section describes HRTFs. Section 3 covers the mechanical structure of and control methods used in the latest version of TeleHead I, that is, TeleHead II. This section also describes the ability of TeleHead II to mechanically follow head movement, the effectiveness of replicating the user’s head in evaluating HRTFs, and the acoustical characteristics of TeleHead II. Section 4 describes sound localization experiments using TeleHead II. The results verify the effectiveness of replicating head shape and movement in an acoustical telepresence robot. This is followed by a brief discussion and concluding remarks.

2 Necessity of Head Reproduction

The role of the head in 3D sound field reproduction and the concept of the HRTF are illustrated in Figure 1. The two center graphs show impulse responses from a sound source to each ear. The sound reaches...
each ear at a different time, depending on the position of the sound source. This time difference is called the interaural time difference (ITD). The relative level of sound reaching the two ears also differs depending on the position of the sound source. This is called the interaural level difference (ILD). In addition, the effects of reflection and diffraction by the head changes the spectral characteristics of the sound that reaches each ear. This change also depends on the location of the sound source. The brain calculates the direction of a sound using these three cues. The HRTF includes these three cues and is defined as

\[ \text{HRTF}(\omega, \theta, \phi, r) = \frac{H_{sp - 1 \text{ ear}} (\omega, \theta, \phi, r)}{H_{sp - center} (\omega, \theta, \phi, r)} \]  

(1)

where \( H_{sp - center} \) is the transfer function from a far-field sound source point to the head centerpoint in a free field, and \( H_{sp - r} \) or \( H_{sp - l} \) is the transfer function from the sound source point to the ear canal entrance of a specific listener’s right or left ear. By convolving a pair of HRTFs of a certain azimuth and elevation with a sound source signal, the listener can localize a virtual sound source (Møller, 1992).

Convolving a sound source with a listener’s HRTFs is equivalent to putting a user-like dummy head in the sound space and recording the signals at the dummy head’s ears. We use a real robot because it offers the possibility of interactions with a remote environment as well as advantages in transmitting the sound environment. When methods based on signal processing are used to choose a convolved HRTF, we need information about the sound’s location and the time when the sound reaches the ears. If we were to try to simulate a complete sound environment, we would need both the HRTF and the frequency characteristics of the space. In addition, if one wants to use one’s own HRTFs, they would have to be measured beforehand. Accurate HRTF measurement is difficult because the targets of the measurement are humans (Riederer, 1998). Moreover, the HRTF of any direction that is not measured must be created by interpolating or extrapolating from the measured HRTFs.

In contrast, because convolving the HRTFs is equivalent to putting the user-like dummy head in the sound space to be transmitted to the user, methods using the dummy head do not require any prior information. However, methods like the one we are pursuing require an accurate dummy head. Technological progress will make it very easy to construct accurate dummy heads in the near future. Progress in 3D measurement technology, such as MRI and optical measurement, makes precise measurement of forms fast and easy. In addition, producing 3D forms will also become easy with progress in 3D molding technology, such as light molding and 3D plotting. The progress in these areas will also contribute to HRTF calculation, but, because there is a trade-off between calculation time and accuracy, there is still no practical technology. At present, making a dummy head such as TeleHead II is time-consuming and costly compared with fast measurement and signal processing methods (Zotkin, Duraiswami, & Davis, 2004), but considering the overall cost of making physical interaction at a remote place possible, the dummy head method holds promise.

3 Architecture of TeleHead II

3.1 Outline

The acoustical telepresence robot TeleHead II consists of four functional components: a human head posture data detector, dummy head, motor and controller, and sound transmission hardware. TeleHead II is diagrammed in Figure 2. The human head posture data are sampled at 120 Hz and motors are driven based on the data. There are three motors, one each for yaw, pitch, and roll motions of the dummy head. The motor for yaw is a direct-drive brushless AC servomotor (0.47 Nm torque at 60 rpm), which reduces driving noise. The motors for pitch and roll are 200 W AC servomotors. There are no gears, so little noise is generated from movement. Forces generated by the motors for pitch and roll are transmitted via driving rods and belts (Figure 2).

The dummy head was made by molding a human head using impression material and modified using 3D shape measurement. The complicated pinna part was
created separately and modified using MRI data. The face was also modified using data obtained with a 3D range finder because MRI data have a large degree of error due to the effects of gravity. These processes make it possible to construct a highly accurate dummy head.

Table 1 shows the ranges of movement. Because of mechanical structure constraints, the ranges for TeleHead II are narrower than those of the human head (Kapandji, 1974) for roll and pitch motion. The center of the range of the movement can be set arbitrarily. Two small microphones (Sony, ECM77B) are placed at the entrance of the dummy head’s outer ears. Sound received by the microphones is fed to headphones through an audio amplifier without any signal processing.

### Table 1. Range of Head Movement

<table>
<thead>
<tr>
<th>Direction (°)</th>
<th>Yaw</th>
<th>Roll</th>
<th>Pitch</th>
</tr>
</thead>
<tbody>
<tr>
<td>TeleHead</td>
<td>180</td>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td>Human</td>
<td>160</td>
<td>90</td>
<td>130</td>
</tr>
</tbody>
</table>

**3.2 Characteristics of TeleHead II**

#### 3.2.1 Mechanical Characteristics.** The gain and phase delay in the frequency response for TeleHead II are shown in Figure 3. Frequency responses were measured presenting sine waves of varying frequency to drive each of the axes. The frequency ranges of the input sine waves are shown on the horizontal axis of Figure 3. The ranges of the sine inputs were 60° in yaw, 20° in roll, and 30° in pitch. Only the results for yaw are shown, because the results for each axis were nearly the same. The speed of human head motion is generally below 1 Hz (see the range shown by the arrows in Figure 3); measurements with TeleHead II show that continuous motion at over 2 Hz (720 deg/s) is impossible. Figure 3 shows that TeleHead II moves stably in the frequency range subjects employ. Time delays are about
120 ms below 1 Hz. A long delay and low gain can be read from this measurement. The reason for the low gain is to suppress vibration, which is the main noise in the system. According to the results of previous studies (Sandvad, 1996; Wenzel, 2001; Brungart et al., 2004), this 120 ms delay makes sound localization accuracy worse.

3.2.2 Characteristics of the System Noise. Whenever a robot moves, noise is generated. This is called driving noise, and it cannot be avoided. We measured two kinds of driving noise. One is mechanical noise, which is radiated from the robot and was measured at a point 0.5 m in front of the TeleHead I and II. We called this radiation noise. The other is the same noise transferred from the vibration to the acoustical signal by the microphones of TeleHead I and II and transmitted to the listener on an acoustical line. This was measured through headphones using IEC couplers, and we call it line noise. The former makes people around the robot feel uncomfortable. The latter is important for the listener. Therefore, reducing line noise is very important for acoustical telepresence robots. Figure 4 shows the characteristics of these two kinds of driving noise. The noise levels were calibrated to a 1 kHz tone. The gray lines show the noise of the original TeleHead I and the black lines show that of TeleHead II. The broken gray lines show the background noise level, that is, the noise level when the robot is deactivated.

The line noise level is important for evaluating an acoustical telepresence robot. The maximum measured noise level corresponds to about 24 dB SPL from 1 to 4 kHz at the amplifier settings used, to which humans are quite sensitive. This level is equal to the ambient noise in a quiet library. We were able to reduce the line noise by more than 20 dB SPL from what it was for TeleHead I by using low-sound-impedance material to connect the microphone with the dummy head and by reducing the feedback gain of the servo motors.

3.2.3 Accuracy of the Dummy Head. We measured the accuracy of dummy head reproduction for two subjects. Figure 5 shows subject one’s real head (RH1) and dummy head (DH1), and subject two’s real head (RH2) and dummy head (DH2), as measured using a 3D range finder (NEC, Danae-R). Table 2 summarizes the results of the measurements. The dimensions in Table 2 are shown in Figure 6. Differences between DH1 and RH1 are 1–7 mm; head-width error/  head-width and head-length error/  head-length are about 3–5%. The differences between DH2 and RH2 are about the same. The differences between DH1 and RH2 and between DH2 and RH1 are considerable. The dimensions used for the measurement are defined in Burkhard and Sach (1975).
3.3 HRTF Measurement

The HRTFs of the real heads and dummy heads were measured in an anechoic room. The distance from the center of the head to the sound source was 1.2 m. The range of the measurement was 0°–360° in azimuth and −40° to 90° in elevation. For the real heads, measurements were performed at 143 points. Each measurement point in the median plane and the horizontal plane was set at intervals of 10°. Other measurement points were set so their HRTFs could be interpolated from neighboring measurement points located less than 20° in the vertical or horizontal directions. If the interval between HRTF data is less than 20° in the vertical or horizontal direction, interpolation is adequate in sound localization experiments (Nishino, Ikeda, Takeda, & Itakura, 2000). Here, we used 143 measurement points to keep the measuring time below 90 min and thereby reduce the burden on the subjects.

Small condenser microphones (Panasonic, WM62-AT102) embedded in earplugs made of silicon impression material were placed in the vicinity of the left and right external auditory canal entrance of the subjects. The earplug closed the external auditory canal. The earplugs were custom-made for each individual. The HRTFs were measured three times. Before each measurement, we confirmed that the positions of the tragus (the prominence in front of the external opening of the ear) and tip of the nose were accurate using laser pointers. The sound source signals were time-stretched pulses (TSPi Suzuki, Asano, Kim, & Sone, 1995) with a sampling frequency of 48 kHz. Each HRTF was obtained by averaging the measurement value 10 times. The HRTFs were calculated using a 512-point fast Fourier transform.

3.4 HRTF Comparison

Figure 7 shows the results of HRTF measurement in the horizontal plane for RH1, DH1, RH2, and DH2. The vertical axis shows the direction of the sound source. The changing depth of the color shows the relative gain of the HRTFs. The complex gain patterns around 10 kHz contribute to sound localization, particularly in the case of source elevation (Middlebrooks, 1999a). Areas around 10 kHz (circled areas) are especially important and show the differences between the two subjects as well as between the matching dummy.
heads and subjects. In the circled areas in the top and second panels, the dark parts, which represent dips, appear as two lines extending from the upper left side to the lower right side. In contrast, in the lower two panels, the dark lines extend from the upper right side to the lower left side. It is clear that the dummy head HRTFs are more similar to the matching subjects than the subjects are to one another. This suggests that the realism achieved with the matching dummy head will be much greater than with an unmatched dummy head.

We estimated the spectral difference for all directions $D_{\text{HRTF}}^{\text{FFT}}$ quantitatively as

$$D_{\text{HRTF}}^{\text{FFT}} = \frac{\sum_{\omega, d} \sum_{i, j} (|H_i(\omega, d)|^2 - |H_j(\omega, d)|^2)^2 / N_{\omega}}{N_d}$$  \hspace{1cm} (2)

where $d$ is the measurement direction, $\omega$ the angular frequency, $N_{\omega}$ the number of points of FFT, and $N_d$ the total number of directions. The magnitude of HRTF in direction $d$, where head $i$ and $j$ are assumed to be the objects of measurement, is denoted as $H_i(\omega, d)$ and $H_j(\omega, d)$, abbreviated $H_i$ and $H_j$. The frequency range for the calculation was between 20 Hz and 20 kHz, which is the human audible range. The average gains of HRTFs were equalized before we calculated the difference between the subjects and the dummy heads.

Figure 8 shows the HRTFs’ spectral differences. The HRTF measurement was repeated three times. Measurement errors were calculated as the average difference between the three measurements. All 143 values were calculated as means of the results of Equation (2) with all possible pairs. All values are significantly different from each other (from an ANOVA of all measurement points, $p < .001$). The difference between RH1 and DH1 (RH1 – DH1) is 4.6 dB; that between RH2 and DH2 (RH2 – DH2) is 5.3 dB; that between RH2 and DH1 (RH2 – DH1) is 6.6 dB; and that between RH1 and DH2 (RH1 – DH2) is 5.8 dB. The smaller difference between the subjects and their dummy heads (about 5 dB on average) than that between the subjects and the dummy head that is not theirs (about 6 dB on average) indicates that each dummy head is acoustically closer to its model than to the other subject. This comparison has a problem in that the ANOVA for 143 points easily becomes significant. More measurements are needed to verify the points raised in this discussion. However, in our study we think the dummy head is accurate from the viewpoint of HRTFs. The errors of the measurements were about 4 dB for the real head and 3 dB for the dummy heads. We did not use an apparatus to fix subjects’ heads in order to avoid the acoustical effects of such tools.
A previous study (Middlebrooks, 1999a) found an average difference of directional transfer functions (DTFs) between subjects of about 2.88 dB. The measurement errors were about 0.6 to 1.4 dB and the average was about 1.1 dB. In addition, another previous study (Middlebrooks, 1999b), which attempted to reduce the individual difference between HRTFs by scaling HRTFs using physical parameters, succeeded in reducing it to 2.5 dB. The results of sound localization experiments using scaled HRTFs showed more accurate localization than those using unscaled HRTFs. Our results of HRTF comparison cannot be compared directly with previous works. Moreover, dummy heads cannot be modified by such signal processing methods, but the existence of the dummy head avoids problems associated with choosing an HRTF and convolving it. If the dummy heads move, they can collect the sound continuously with the effect of head movement.

4 Sound Localization Experiments

4.1 Method

TeleHead II was set in an anechoic room, and the same two subjects listened to sound stimuli in a sound-proof room. Loudspeakers (VIFA, MG10SD0908) were placed around TeleHead II at intervals of 30° in the horizontal plane or at intervals of 15° in the median plane from −45° to 75°. The distance from the loudspeakers to TeleHead II was 1.2 m. The arrangements of the loudspeakers are shown in Figure 9. Sound stimuli were generated at the loudspeakers, collected by microphones in the dummy head, and transmitted from TeleHead II to the subjects through closed-type headphones (Sennheiser, HDA200). In a preliminary trial, we tested open-type headphones (AKG, K-1000), insert-type headphones (Etymotic Research, ER-4S), and closed-type headphones (Sennheiser, HDA200), and found that sound localization performance in the median plane using each type of headphones was similar. We chose the closed-type headphones because their acoustical characteristics have been studied in detail and because the difference in acoustical gain from session to session is small (Hirahara, 2004). The stimulus for all experiments was white Gaussian noise generated independently by a computer and D/A converter (sampling frequency 48 kHz, 16 bits) for each trial. The noise had a duration of 10 s and its level was roughly 70 dB SPL. The levels of the stimuli were little changed in each trial.

The subjects had normal hearing, as determined with an audiometer. We made two dummy heads, DH1 and DH2. DH1 is a user-like dummy head for subject 1 and DH2 is a user-like dummy head for subject 2. Sound sources were positioned in the horizontal plane and in the median plane. Each subject listened directly in an anechoic room with speakers, using the subject’s own dummy head, and using the other subject’s dummy head. When the subjects listened directly, they used a head-mounted display on which the response form was projected, and responded with a trackball. One head-movement condition was that subjects could move their heads and TeleHead II was synchronized with the movement. The other was that they could move their heads and TeleHead II was kept stationary. In real-head listening conditions, that is, when subjects listen to sound with their own head and ears, subjects were either asked to keep their head as still as possible or were allowed to move their head freely. A session consisted of five trials for each direction. Therefore, each session consisted of 60 trials for the horizontal plane and 45 trials for the median plane. Each experiment consisted of three sessions. The stimuli were generated in a different random order within each session. The experiments in the horizontal plane and in the median plane were performed separately. At the start of a trial using Tele-
Head II, we made sure that the subject’s head and TeleHead II were facing in the same direction. Subjects were in a soundproof room located 10 m from the anechoic room and listened to the sound through headphones mounted with a head tracker. The subjects knew where the loudspeakers were. The subjects’ eyes were open, but they could not see the loudspeakers because they were in another room. The method was forced choice of one of 12 directions in the horizontal plane and one of nine in the median plane. The subjects responded by picking one of the directions on the response form. Subjects could move their heads freely in all trials.

4.2 Results

Figures 10 and 11 show results for subject 2 in the horizontal and median planes, respectively. The results for subject 1 showed the same tendencies as those in previous works (Wightman & Kistler, 1999). However, the tendencies of our results for subject 1 were not so clear, so we do not discuss them here. The results for subject 2 showed the same tendencies in the horizontal plane, but not in the median plane when the subject used a non-user-like dummy head with head movement. The sound stimulus angle was set at 0° at the front of the subject. The degree of the angle increases clockwise in the horizontal plane and upwardly in the median plane. The sizes of the filled circle are proportional to the number of subjects’ responses. Correct answers are on the diagonal line. The broken lines in Figure 10 show answers with front-back confusion. The lower two panels show the results when TeleHead II was synchronized with the subjects’ head movement, and the upper two show results when it was static. The left two panels show the results for subject 2’s dummy head (DH2), and the right two those for the other subject’s dummy head (DH1). In both planes, the results were most accurate when using the synchronized user-like dummy head (left lower panel).

Average errors of sound localizations are shown in Figures 12 (horizontal plane) and 13 (median plane). The data indicate that sound localization performance of subject 1 with the user-like dummy head and syn-
vision was equivalent to that for direct listening in both planes. However, in the case of subject 2, the results using the user-like dummy head are not equivalent to those for direct listening but are better than for the non-user-like dummy head. For both subjects, performance for the non-user-like dummy head was worse than that for the user-like dummy head in both dimensions. In general, synchronous results were not systematically better than stationary results. The reproduction of head shape improved the accuracy of sound localization in both planes for subject 2. In contrast, the accuracy in the case of synchronized movement was not always better than that when the dummy heads were stationary. For subject 1, accuracy was improved in both planes with synchronization, whereas for subject 2, it was improved in the horizontal plane but was not improved in some cases in the median plane.

5 Discussion

Improvements in localization with head movement in the horizontal plane were found in previous studies (Perrett & Noble, 1997; Wightman & Kistler, 1999). However, the results in the median plane for subject 2 with DH1 did not indicate improvement. The biggest reason for this is that subject 2 could not localize the sounds at all in the static condition: the subject was confused and thought all the sounds came mostly from above. If subjects had answered that the sound came from around the front point, the error would be 30°–40°. In addition, in the case of subject 2 using the non-user-like dummy head, the 95% confidence intervals are very large, nearly 30°. Therefore, the result for subject 2 using DH1 is little better than the chance level without head movement and the same as the chance level with head movement. There may be two additional reasons for subject 2’s results. One might be related to the strategy of the subject’s head movement. If subject 2 had used head movements with as much rolling as possible, the results in the median plane might have been closer to those in the horizontal plane. This strategy of head movement would have been better for subject 2, but head rolling is not natural for that subject. If we had continued the experiments, subject 2 may have found a new head-movement strategy and the results would have been different. Another reason might be the use of a virtual sound source. Still another might be the accuracy of head movement reproduction. TeleHead II has 120 ms delay and 24 dB SPL noise, which would reduce the utility of head movement. We have reported
the relationship between delay and sound localization for the case of using TeleHead II with the user-like dummy head only (Toshima & Aoki, 2005). That report did not discuss general human auditory perception characteristics or tendencies, but it showed that the user could localize sound accurately after 540 sound localization experiments in the case of using a sound stimulus of 1 s delay and 5 s duration time. A few reports have addressed the relationship between delay and sound localization. A delay of more than 96 ms was found to reduce the accuracy of sound localization (Sandvad, 1996); even a 250.4 ms delay has little effect in the case of using 8 s sound stimuli (Wenzel, 2001). In addition, it has been found that latencies of 70 ms or less are unlikely to adversely impact localization ability in virtual audio display systems and that those exceeding 90 ms do impair localization ability (Brungart et al., 2004).

The results using the user-like dummy head and being synchronized with the users’ head movement were good. This indicates that the concept of TeleHead, a user-like dummy head and synchronizing with user’s head movement in real time, is an acceptable guideline for developing an acoustical telesence robot. Cost and time issues make it very difficult to build the many dummy heads that we would need in order to clarify these things exactly. However, the present results indicate the possibility of transmitting remote acoustical environments using a telerobotics system and that a telerobotics system also has the possibility of interaction at a remote place in the future. Therefore, clarifying the problems and possibilities of telerobotics for acoustical telesence is important and the results of this paper are basic examples of perceptual evaluations with an acoustical telesence robot.

6 Conclusion

We built an acoustical telesence robot, TeleHead I, based on the concept that remote sound localization can be best achieved by using a user-like dummy head whose movement synchronizes with the user’s head movement in real time. We compared and contrasted the physical characteristics of the latest version, TeleHead II, and verified the validity of this concept using sound localization experiments.

TeleHead II can synchronize stably with the user’s head movement with a 120 ms delay. Its driving noise is on the order of 24 dB SPL from 1–4 kHz. The shape difference between the dummy head and the user is about 3% in head width and 5% in head length. The difference between HRTFs of the dummy head and the modeled listener’s head is about 5 dB in the range of the whole audible frequency, under 3–4 dB in measurement error.

The results of the sound localization experiments using TeleHead II showed that head movement improves horizontal-plane sound localization performance even when the dummy head shape differs from the user’s head shape. In contrast, head movement was not found to improve sound localization performance in the median plane when dummy head shape and user head shape were different for one of the two subjects tested here. The accuracy of sound localization in the case of using the same-shape dummy head whose movement synchronizes with the user’s head movement was always better or equal to results with a nonindividual head or no movement.

These results show that the TeleHead concept works for building an acoustical telesence robot. Namely, there is a possibility that the use of user-like dummy head and synchronization with user’s head movement improves horizontal sound localization accuracy even when reproducibility of head shape and head movement are not perfect and additional mechanical delay and noise exist. Therefore, though acoustical telerobotics technology such as the TeleHead II still has some problems, it offers an alternative to the widely used digital-signal-processing-based acoustical telesence technologies.

References


Begault, D. R., Wenzel, E. M., & Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverbation and individualized head-related transfer functions on...
the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society, 49*, 904–916.


