Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Three-party sound field sharing system based on the boundary surface control principle

-Subjective assessment of voice reproduction with speaker's facing angle-

Yusuke Ikeda (1), Seigo Enomoto (1), Shiro Ise (2) and Satoshi Nakamura (1)

(1) National Insitutute of Information and Communications Technology, Kyoto, Japan
(2) Graduate School of Engineering, Kyoto University, Kyoto, Japan

PACS: 43.38.-p, 43.38.Vk

ABSTRACT

A telecommunication system makes communicating more comfortable if it ensures that parties involved in distant communication feel as if they are located in the same space during their conversation. By applying physically accurate sound field reproduction, we aim to develop a telecommunication system which enables us to feel the presence of a conversational partner. In pursuit of physically accurate sound field reproduction, we have developed a sound field reproduction system based on the boundary surface control principle. We have also developed a two-party sound field sharing telecommunication system using that reproduction system. In this paper, we describe an extension of that system to three-party system and conduct the subjective assessment of its voice reproduction. In pursuit of decreasing the amount of real-time convolution calculations, we applied Gram-Schmidt orthogonalization to reduce the number of secondary sound sources. In a three-party conversation, it is important to know "who talks to whom". Accordingly, when one of conversational partners turns towards another partner in three-party conversation, we reproduce natural changes in voice directivity caused by head rotation by detecting facing angle through image recognition and by adjusting the voice filter to suit that angle. However, this requires the voice reproduction with accuracy enough to acoustically perceive "who talks to whom". Thus, we conducted subjective assessments of the speaker's facing angle both in real environment and in sound reproduction environment. As a result of average angle error in sound reproduction environment, we found out that the system reproduced voice with accuracy enough to perceive who talks to whom. And we also found that there was little difference in the voice facing angle between perception in the real environment and in the sound field reproduction environment for a half of the subjects.

INTRODUCTION

Communication lies at the root of human activities. Telecommunication technology such as telephone has made the distant communication possible and has become an essential tool in our daily life. On the other hand, the face-to-face communication still remains important. We think that the telecommunication system makes communicating more comfortable if it ensures that parties involved in distant communication feel as if they are located in the same space during their conversation.

The idea of transmitting feelings of "being there" goes back to the concept of "telepresence"suggested by Minsky long time ago. Minsky used the technology of remote control for space development and other dangerous operations to explain the importance of telepresence (Minsky 1979). Since then considerable research has been done on transmission of various senses including vision and hearing (Bly, Harrison, and Irwin 1993; Buxton 1992). As for hearing sense, it lacks the accuracy needed for the feeling of "being there".

In pursuit of physically accurate sound field reproduction, we have developed a sound field reproduction system based on the boundary surface control principle (Ise 1993). By conducting subjective assessments we have confirmed that it is possible to reproduce with high accuracy the sense of sound localization, the sense of distance, and the speaker's facing angle(Enomoto et al. 2008; Ikeda et al. 2009). We used two such systems to develop the sound field sharing system, so that two remotely positioned persons can have a distant communication feeling the same sound field (Ise et al. 2007).

Perceiving the spatial information such as speaker's position and distance through a voice sound becomes even more important as the number of communicating persons increases. Now we are trying to extend the existing two-party sound field sharing system to three-party system.

While developing the three-party system, we encountered two problems. The first problem is an increase in the number of calculations required for a voice sound reproduction. However, in the previous experiment we applied the Gram-Schmidt orthogonalization and confirmed that it is possible to reduce the number of loudspeakers from 62 to 24 without degrading the sound localization (Enomoto et al. 2010).

Although it is possible to perceive a speaker's facing angle with the existing 62ch sound field reproduction system (Ikeda et al. 2009), the other problem is that the inter-system conversation uses the voice sound recorded with a close-talking microphone and does not include the head rotation information. Especially in a communication which involves more than three parties, a more natural-sounding conversation can be achieved by acoustically reproducing "who talks to whom". Therefore, we install a camera in front of the each conversation party, and detect the speaker's head rotation angle through the camera image recognition, and then reproduce the voice directivity in accordance with that angle.

In this paper, we conduct two subjective assessments: we compare the accuracy of the voice facing angle obtained in a reduced-to -24 ch environment with the facing angle in a real environment, and thus aim to examine the accuracy of the voice facing angle needed to reproduce "who talks to whom" in a three-party conversation. We especially take into consideration the individual differences in ability to perceive the facing angle, and compare the results obtained from the same subject in both the real environment and the reproduced environment.

3D SOUND REPRODUCTION SYSTEM BASED ON THE BOUNDARY SURFACE CONTROL PRINCIPLE

In 1993, Ise proposed the boundary surface control principle which is 3D sound reproduction method based on Kirchhoff-Helmholtz integral equation and inverse system (Ise 1993; Ise 1997; Ise 1999). Figure 1 shows its basic concept.

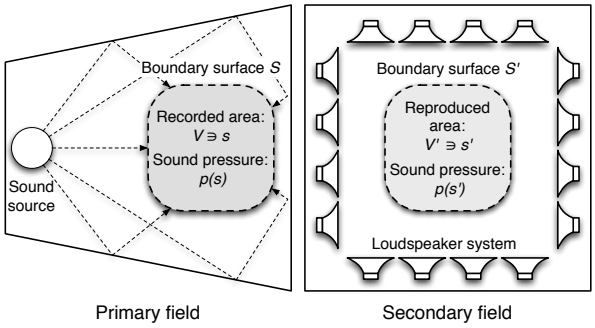

Figure 1: Concept of boundary surface control principle

We are considering reproduction of a sound field at recorded area V in the primary field into reproduced area V' in the secondary field. Given that V is congruent with V', the following equations hold.

$$|r'-s'| = |r-s| \quad (s \in V, r \in S, s' \in V', r \in S') \tag{1}$$

where let S and S' denote a boundary of the recorded area and a boundary of the reproduced area respectively. If we denote sound pressure in V and V' as $p(s)$ and $p(s')$ respectively, $p(s)$ and p(s') are denoted by following equations.

$$p(s) = \iint_S G(r|s)\frac{\partial p(r)}{\partial n} - p(r)\frac{\partial G(r|s)}{\partial n}dS,\ s \in V \tag{2}$$

$$p(s') = \iint_{S'} G(r'|s')\frac{\partial p(r')}{\partial n} - p(r')\frac{\partial G(r'|s')}{\partial n}dS',\ s' \in V' \tag{3}$$

where let n and n' denote normal vectors on S and S' respectively. By applying the equation 1, we obtain the following relationship of Green's function and its gradient in equations 2 and 3.

$$G(r|s) = G(r'|s') \tag{4}$$

$$\frac{\partial G(r|s)}{\partial n} = \frac{\partial G(r'|s')}{\partial n'} \tag{5}$$

Hence, it follows that if the sound pressure and its gradient on each boundary are equal to each other, then the sound pressures in each area are also equal to each other from equations 2 and 3. This is expressed as

$$\begin{aligned} &\forall r \in S, \forall r' \in S', \\ &p(r) = p(r'), \frac{\partial p(r)}{\partial n} = \frac{\partial p(r')}{\partial n'} \\ &\Longrightarrow \forall s \in V, \forall s' \in V', \quad p(s) = p(s'). \end{aligned} \tag{6}$$

Considering this as boundary value problem, uniqueness of the solution follows that either sound pressure value or its gradient value are sufficient to determine the value for both (Kleinman and Roach 1974).

Another sound reproduction method using Kirchhoff-Helmholtz integral equation is wave field synthesis(Bourkhout, Vries, and Vogel 1993). However, a characteristic of boundary surface control principle is that a configuration of closed surface is not restricted because of the inverse system.

Figure 2 shows 70 ch fullerene-shaped microphone array and 62.8 ch multi-channel loudspeaker system which we have developed. 70ch fullerene-shaped microphone array is designed based on 70 elements of the C_{80} fullerene which has its lower part cut short by 10 elements. Diameter of the microphone array is about 46 cm which is large enough to enclose a human head. Each microphone is an omnidirectional microphone (DPA 4060). The loudspeaker system is composed of 4 layers of loudspeaker array and 4 columns. All four layers, placed in vertical order, have 6, 16, 24, 16 full-range loudspeakers (FOSTEX FE83E) respectively. The sound reproduction based on boundary surface control principle only uses 62 layer-installed loudspeakers.We assume that the head height of a listener is almost the same as the height of the third layer. An elevating machine is used for moving the listener's head to a suitable position.

We calculate the inverse system by using the impulse responses which are preliminarily measured with the microphone array inside the loudspeaker system. The system does not lead to the problem of head rotation followed by sound because we only reproduce the sound field at fixed area. We have used this system to conduct subjective assessments on the accuracy of such factors as horizontal localization, front-directed sense of distance, and speaker's facing angle. We have also conducted assessment on comprehensive feeling of reproduced sound field by applying recorded data of natural sound and orchestra (Enomoto et al. 2008; Ikeda et al. 2008; Ikeda et al. 2009).

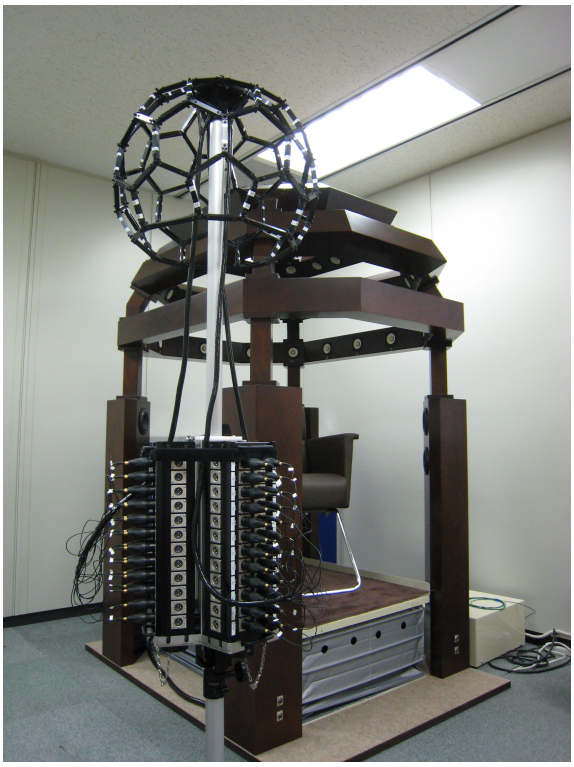

Figure 2: 3D sound reproduction system

EXTENTION OF 2-PARTY SOUND FIELD SHARING SYSTEM TO 3-PARTY SYSTEM

Sound field sharing system

We have been conducting research and development of the sound field sharing system so that conversation parties can have distant communication with a feeling of "being there". The sound field sharing system which is based on boundary surface control principle allows its multiple users to talk to each other and to listen simultaneously to the same sound field despite mutually distant location. For example, two distantly located persons can enjoy a concert as if they are seated side by side.

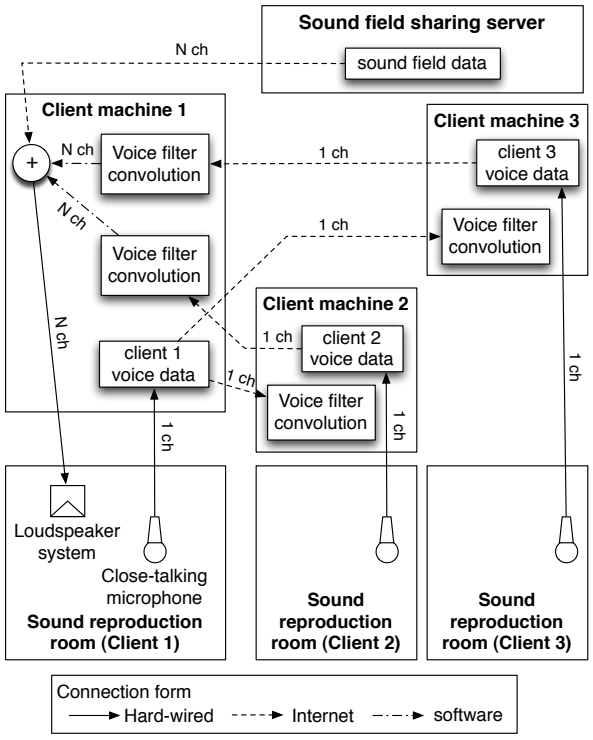

Figure 3: Data flow for a client in three-party sound field sharing system. Client 2 and 3 have the same data flow as Client 1. Voice filter: convolution of the voice transfer function and the inverse system; N: Number of loudspeakers.

A sound sharing system has a client-server relationship. Figure 3 shows data flow in a three-party sound field sharing system. The server sends the sound field data through the internet to each client. There are two types of sound field data.The first type is accumulated sound field data. The sound field data is recorded beforehand using a fullerene-shaped microphone array and is convolved with the inverse system. In this case, the server only sends prepared data. The second type is real-time sound field data. The server records the sound field data with the microphone array and convolves it with the inverse system in real time. Then, the server sends the data to each client. In both cases, the client machine's only task for sound field data is playing the received data.

On the other hand, the method of reproducing each client's voice for inter-system communication is described below. Each client's voice is recorded using a close-talk microphone. Each client machine sends the recorded voice data to all other client machines. The client machine convolves each received voice data with a voice filter which corresponds to the client's positional relationship. The voice filter is a signal which convolutes the above-mentioned inverse system and impulse responses performing the assumed voice transfer function. The impulse responses contain information on relationship between a speaker and a listener, a speaker's facing angle and room sound reflections. After each voice is convoluted, the client machine puts together all other clients' voices and the sound field data received from the server machine and plays it with the loudspeaker system.

Voice filter for three-party system

As shown in Figure 3 , because conversational partners are positioned in different places, it is necessary to convolve the same number of voice filters as the number of conversational partners.Consequently, when we extend the two-party system to a three-party system, the amount of convolution calculations only for the voice filter becomes twice as much in each client machine.

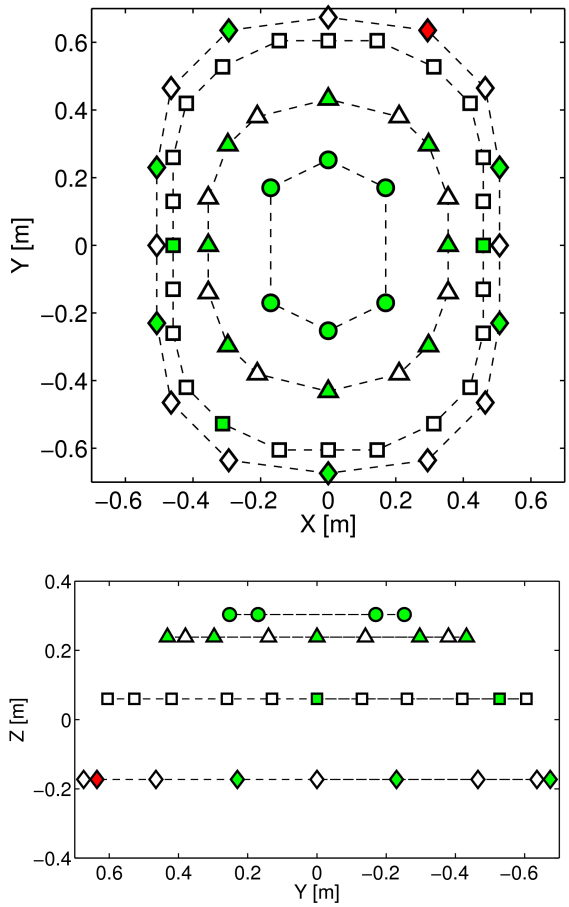

Figure 4: Selected 24 loudspeakers' configuration by Gram-Schmidt orthogonalization.Upper and lower figures respectively show top and side view. Markers: original 62 loudspeakers' position; Colored Markers: Selected loudspeakers' position; Red Marker: the loudspeaker's position given as an initial value; Dot-line: connection of the loudspeakers in the same layer; Origin: position of listener's head; Front direction: the negative direction of Y-axis.

In this paper, we suppressed the amount of voice filter-related calculations by effectively reducing the number of loudspeakers from 62 to 24 using the Gram-Schmidt orthogonalization(Enomoto et al. 2010). Sound source selection using Gram-Schmid orthogonalization is a method of selecting microphones (i.e. control points) and loudspeakers (i.e. secondary sound source) so that their linear independency becomes highest in terms of geopemtrical relationship (Asano, Suzuki, and Swanson 1999). As this method sequentially selects the loudspeakers, we need

to select a criterion loudspeaker and give to it an initial value. We examined all loudspeakers by giving the initial value to each, and selected the combination which shows highest linear independency in case of using 24ch. Figure 4 shows loudspeakers selected using Gram-Schmidt orthogonalization. The results of previous subjective assessment show that the sound localization does not deteriorate in this combination of selected 24 loudspeakers and almost equals the all-62ch combination (Enomoto et al. 2010).

In addition to such information as speaker's location and room sound reflection, the voice filter also contains data on speaker's facing angle. We have found that the 62 ch loudspeaker system based on boundary surface control principle reproduces a voice with accuracy enough to perceive a voice facing angle (Ikeda et al. 2009). As the number of parties in the system increases to three, it becomes more important to know"who talks to whom" for natural-sounding conversation. We measured the impulse responses for voice filter by rotating a directional loudspeaker. We detect the facing angles using a camera which is installed in front of each party. The system reproduces a voice directivity by selecting a voice filter which corresponds to the detected speaker's facing angle.

SUBJECTIVE ASSESSMENT OF VOICE REPRODUCTION WITH SPEAKER'S FACING ANGLE

Experimental condition

Considerable research has been done on perception of voice facing angle using a loudspeaker or a real voice (Neuhoff, Rodstrom, and Vaidya 2001; Kato, Takemoto, and Nishimura 2008; Takano et al. 2005). Previously, we recorded a head-rotating person's voice in a real environment using a fullerene-shaped microphone array. Then, we conducted a subjective assessment of a voice facing angle perception in a sound field reproduction environment with 62ch loudspeakers (Ikeda et al. 2009).

Comparing with the previous 62ch sound reproduction, two notable changes were made to the voice filter of the three-party sound field sharing system. The first change is a reduced number of loudspeakers from 62ch to 24ch by using Gram-Schmidt orthogonalization. The second change is a limitation of the voice filter length or the length of impulse responses used for the voice filter. This limitation is required for real-time convolution of the voice filter. Consequently, we should examine whether the voice filter for three-party system reproduces a voice with accuracy enough to acoustically perceive who talks to whom.

In this paper, we assess the accuracy of voice reproduction with speaker's facing angle. We compare the accuracy of facing angle perception both in real environment and in sound reproduction environment which is based on the real environment. Figure 5 shows measurement of impulse responses aimed at the voice filters. Figure 6 shows configuration of a loudspeaker and the microphone array. The positional relationship is determined by one of the loudspeakers and the microphone array. Consequently, we need to assume the positional relationship of three parties before measuring the impulse responses. Here, we assume a three-party conversation in which the parties are positioned on the apexes of 2m-sided equilateral triangle.

Position of the loudspeaker corresponds to the position of a conversational partner. As Figure 6 shows, loudspeaker is positioned 2m away from the center of microphone array at 30 degree angle from its front on the left side. Similarly, by setting the loudspeaker at 30 degrees from the front of microphone array on its right side and using the impulse responses on both right and lefts sides, it becomes possible to realize the three-party conversation in the above-assumed positional relationship.

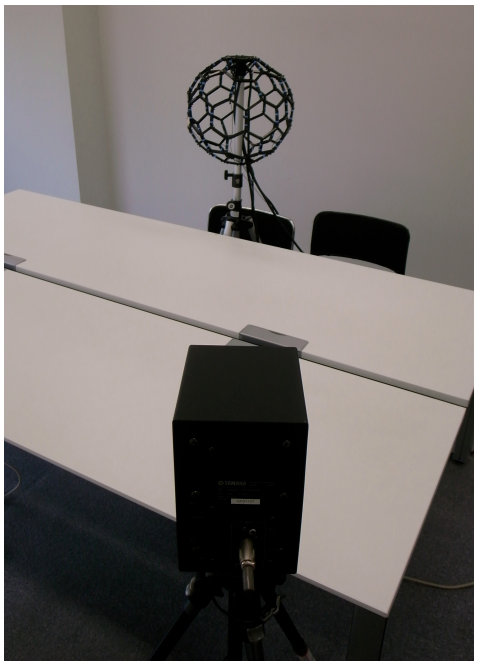

Figure 5: Picture of the implulse response measurement for voice filter

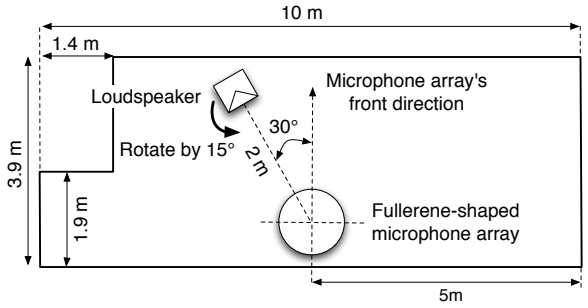

Figure 6: Top view of the loudspeaker and microphone array configuration in the impulse response measurement for voice filter.

The sound reflection in the same room differs depending on the listening position. Therefore, we have to repeat the same measurement by moving the microphone array to each apex where the parties are positioned. However, if we take only the positional relationship into consideration, equilateral triangle's apex-based positioning becomes possible by applying these two sets of impulse responses to the whole client system. In this subjective assessment, we only used the impulse responses measured by the loudspeaker on the left side.

The room has a table in its centre and is used for small conferences. Reverberation time of the room is about 0.6 s. We measured impulse responces by rotating the loudspeaker (YAMAHA MSP-3) around its front surface by 15 degrees in 360 degrees. Considering the real-time convolution of the voice filter, measured impulse responses are too long to get attenuated within the voice filter time frame. We attenuated the impulse responses to the length of 2048 points by using Hanning window. The Hanning window has at its centre the direct sound of impulse responses recorded by each microphone. We also designed the inverse system to have the length of 2048 points.

For reference, Figure 7 shows the original impulse response and the attenuated one which are measured with the microphone

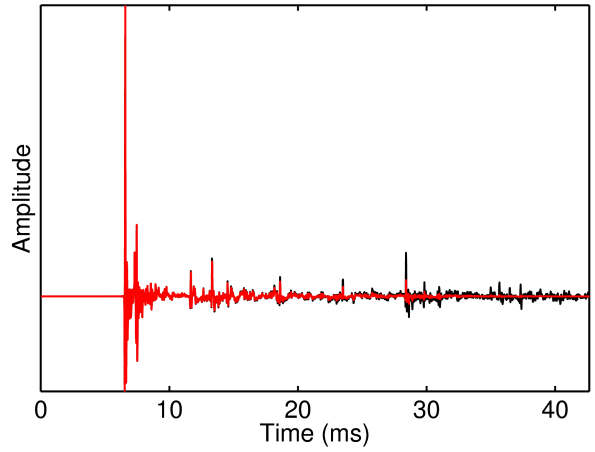

Figure 7: Original and attenuated impulse responces (Black line: the original impulse responce; Red line: attenuated impulse response using the Hanning window.)

positioned in front of the microphone array and with the loudspeaker facing the microphone array. By visually examining the attenuated impulse responses, we see that they sufficiently contain the early reflections and do not contain any of the late reflections.

For a stimulus in this experiment, we used a voice of a male in his 30s saying a general greeting “Konnichiwa” in Japanese. The subjects in the experiment are ten Japanese people in their 20s or 30s, of which 5 persons are women and 5 persons are men. The angle used in this experiment ranges from 0 degrees to 90 degrees moving counterclockwise by 15 degrees. Zero degree position implies that the loudspeaker faces the microphone array at that point. By using this angle range, we can determine whether or not the listener acoustically perceives to whom the speaker is talking in the assumed three-party relationship.

We conducted two types of subjective assessments.

1. Voice reproduction with a loudspeaker rotating in real environment, and
2. Voice reproduction in an environment of 3D sound field reproduction system based on boundary surface control principle using the above-mentioned voice filter.

The first subjective assessment was held at the same place where impulse responses were measured and with similar conditions so that the loudspeaker was randomly rotated in real environment The loudspeaker which was used for measuring voice filter-aimed impulse responses was also used for voice reproduction in real environment. We positioned the subjects on the places where impulse response-measuring microphones were previously put. Then we used a curtain in front of the rotating loudspeaker in order to block the view to the subjects. The results obtained from the sound-level meter show that the curtain’s effect on the sound field is insignificant. We adjusted the loudspeaker’s power output so that the sound volume is not affected by the facing angle or the two above-mentioned environments.

Before getting the answers for our questionnaire, we inform the subjects about speaker’s location. We also let the subject listen to the changing voice direction by rotating the loudspeaker from 0 degrees to 90 degrees by 15 degrees and backwards from 90 to 0. According to the questionnaire, the subject first listens to the voice at zero degree point, and then listens to two more stimuli voices. The subject has to choose from seven possible voice directions ranging up to 90 degrees by 15 degrees. Seven voice directions were tested at random order on each subject. In total, the subjects answered to 14 questions related to both real environment and sound field reproduction environment.

Discussion

The angle error can be defined in each environment in the following way: as the absolute difference between the loudspeaker’s facing angle and the answered angle in real environment, and the absolute difference between the reproduced voice’s facing angle and the answered angle in the sound reproduction environment. Figure 8 shows a box plot of average

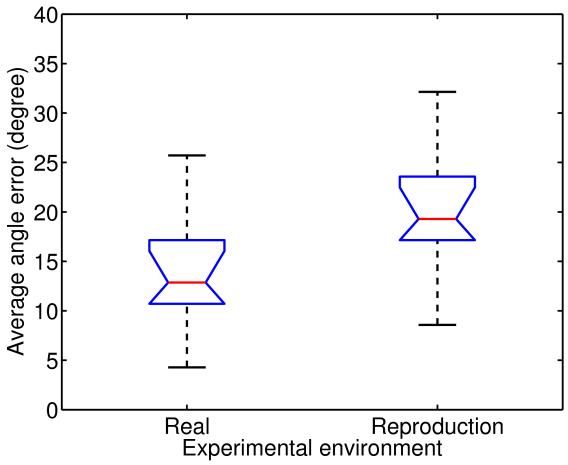

Figure 8: Box plot of average angle error by subjects

angle error by subjects in each environment. Each average angle error in real environment and in sound reproduction environment is respectively 13.7 degrees and 20.8 degrees. Given the assumed relationship of three parties positioned on the apexes of equilateral triangle, the average angle error in sound reproduction environment is small enough to perceive who talks to whom. However, there is 7.1 degrees difference between the average angle error in two environments. Two-tailed t test result shows that the difference of average angle error have a statistical significance ($p < 0.05$). Consequently, we found out that it is more difficult for subjects to perceive the voice facing angle in sound reproduction environment than in real environment. Majority of the subjects also commented that perception of voice facing angle in sound reproduction environment is more difficult than in real environment. Most of the comments referred to the reverberation length as the difference between two environments. An interaural level difference especially in anechoic chamber is cited among physical factors of facing angle perception (Neuhoff, Rodstrom, and Vaidya 2001; Takano et al. 2005). However, we infer that in a real environment with reverberations which we used this time, the change in reverberation also largely affects the facing angle perception. Figure 9 shows average angle error by each speaker’s facing angles. When the facing angle is 90 degrees, there is a significant difference between two environments. Some subjects comment that they can’t perceive the voice facing angle rotating up to 90 degrees.

Figure 10 shows a scatter plot of average angle error by each subject. Figure 10 shows that there was little difference in the voice facing angle between perception in two environments for the first half of the subjects. As for the second half of the subjects, the facing angle perception in the real environment showed higher accuracy. One of the subjects whose questionnaire results showed little difference between the two environments commented that she clearly perceives the voice facing angle rotating from 0 degrees to 90 degrees in the sound field reproduction environment. These results indicate that the per-

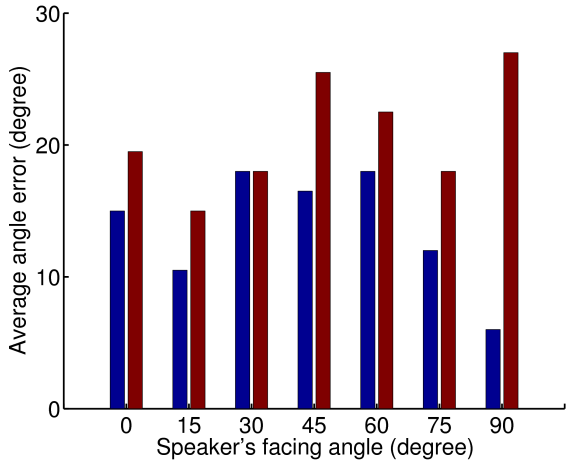

Figure 9: Average angle error by speaker's facing angle. Blue bar: in real environment; Read bar: in sound reproduction environment.

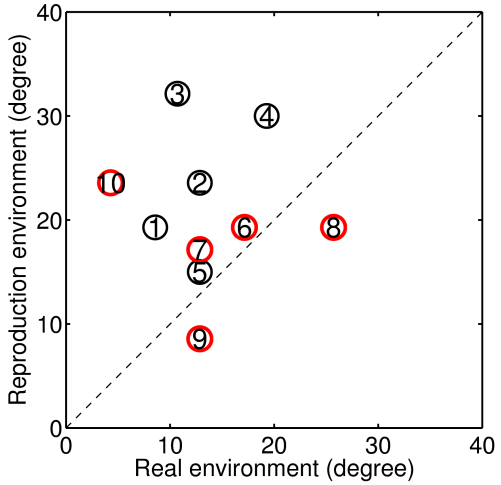

Figure 10: scatter plot of average angle error by subjects between real and sound reproduction environment. Number in circles: subject number; Red circle: woman; Black circle: man.

ception of voice facing angle differs among individuals depending on their abilities. And also Figure 10 shows that especially women subjects have little difference between two environments. In the subjective assessment, we controlled the output power of loudspeaker in order to keep the voice sound volume in each angle constant. However, perception of voice facing angle in real communication and in three-party sound field sharing system is easier than in the subjective assessment. Because the sound volume of a voice directed at the listener spontaneously changes in accordance with the speaker's facing angle.

CONCLUSION

We have been conducting research and development of a sound field sharing telecommunication system aiming to realize a distant communication with the feeling of being in the same space. In this paper, we extended the existing two-party system to a three-party system. We examined the accuracy of the voice sound reproduction by conducting a subjective assessment of the speaker's facing angle. As a result of the experiment, we found out that it is possible to acoustically reproduce "who talks to whom" in a three-party conversation. We also figured out the influence of the unreproducible late reflected sound caused by the voice filter length limitation on the accuracy of the speaker's facing angle. Also, there are individual differences in the voice facing angle perception. There was little difference in the voice facing angle between perception in the real environment and in the sound field reproduction environment for the first half of the subjects. In the future, we are going to conduct a comprehensive subjective assessment of a three-party conversation in the sound field sharing system using the voice filter, and to examine the effectiveness of this system.

REFERENCES

Asano, F., Y. Suzuki, and D.C. Swanson (1999). "Optimization of control source configuration in active control systems using Gram-Schmidt orthogonalization". *IEEE transaction on Speech and Audio Processing* 7.2, pp. 213–220.

Bly, Sera A., Steve R. Harrison, and Susan Irwin (1993). "Media spaces: bringing people together in a video, audio, and computing environment". *Communications of the ACM* 36.1, pp. 28–46.

Bourkhout, A. J., D. de Vries, and P. Vogel (1993). "Acoustic control by wave field synthesis". *Journal of the Acoustical Society of America* 93.5, pp. 2764–2778.

Buxton, William A. S. (1992). "Telepresence: Integrating shared task and person spaces". *Proc. of graphics interface*, pp. 123–129.

Enomoto, Seigo et al. (June 2008). "Three-dimensional sound field reproduction and recording systems based on boundary surface control principle". *Proc. of 14th International Conference on Auditory Display*, Presentation o 16.

— (Aug. 2010). "Optimization of loudspeaker configuration using Gram-Schmidt orthogonalization for the sound reproduction system based on the boundary surface control principle". *Proc. of International Congress on Acoustics*.

Ikeda, Yusuke et al. (Nov. 2008). "Evaluation of 3D sound field reproduction system using multi-channel loudspeakers and the boundary surface control principle (in Japanese)". *Technical report of IEICE. EA*, pp. 77–82.

— (Sept. 2009). "Subjective assessment of speaker's facing angle in 3D sound field reproduction system based on the boundary surface control priciple". *Proc. of meeting of acoustical society of Japan*.

Ise, Shiro (Oct. 1993). "A study on the sound field reproduction in a wide area(1) -based on krchhoff-helmholtz integral equation- (in Japanese)". *Proc. of meeting of acoustical society of Japan*, pp. 479–480.

— (1997). "A principle of active control of sound based on the Kirchhoff-Helmholtz integral equation and the inverse system theory (in Japanese)". *The Journal of Acoustical Society of Japan* 53.9, pp. 706–713.

— (1999). "A principle of sound field control based on the kirchhoff-helmholtz integral equation and the theory of inverse systems". *Acustica* 85, pp. 78–87.

Ise, Shiro et al. (Sept. 2007). "The development of the sound field sharing system based on the boundary surface control principle". *Proc. of International Congress on Acoustics*.

Kato, Hiroaki, Hironori Takemoto, and Ryouichi Nishimura (Mar. 2008). "Auditory perception of speaker's facing angle (in Japanese)". *Proc. of meeting of acoustical society of Japan*, pp. 583–584.

Kleinman, R. and G. Roach (1974). "Boundary integral equations for the three dimensional Helmholtz equation". *SIAM Review* 16, pp. 214–236.

Minsky, Marbin (1979). "Toward a remotely-manned energy and production economy". *A.I. Memo. MIT* 544.

Neuhoff, John G., Mary-Alice Rodstrom, and Tanaya Vaidya (2001). "The audible facing angle". *ARLO* 2, pp. 109–114.

Takano, Hiroki et al. (2005). "A study on a perception of the speech-direction (in Japanese)". *Technical report of IEICE. EA* 348.105, pp. 37–42.