RealSound Interaction: A Novel Interaction Method with Mixed Reality Space by Localizing Sound Events in Real World

Mai Otsuki, Asako Kimura, Takanobu Nishiura, Fumihisa Shibata, and Hideyuki Tamura

Graduate School of Science and Engineering, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, 525-8577, Shiga, Japan

Abstract. We developed a mixed reality (MR) system which merges the real and the virtual worlds in both audio and visual senses. Our new approach "RealSound Interaction" is based on the idea that the sound events in the real world can work as interaction devices with an MR space. Firstly, we developed a sound detection system which localizes a sound source. The system consisted of two types of microphone arrays, fixed type and wearable type. Secondly, we evaluated the accuracy of the system, and proposed three practical usages of the sound events as interactive devices for MR attractions.

Keywords: Mixed Reality, Sound Input, Microphone Array, Sound Source Localization, and Interactive Device.

1 Introduction and Objectives

In this paper, we present a very unique and novel method for interacting with a mixed reality (MR) space that merges the real and the virtual worlds. Our approach "RealSound Interaction" is based on the idea that the sound events occur in the real world can work as input or interaction devices into/with an MR space.

Until now, the various devices like sensors, keyboard and mouse have mainly been used for input into virtual reality (VR) or MR space, and these devices needed to change their shape to keep proper mental model of users. It means that we had to prepare many differently shaped sensors with different functions.

On the other hand, it is not difficult to prepare some sound sources when we use a sound for the input. In addition, it would be simple and easy to customize sounds for each user. Our approaches to use the sound event as interaction device for MR systems were able to realize the intuitive interface. Though, the use of this function is not necessarily limited to MR, it could be expected to be used in a general system as new interaction device.

This paper describes the method of detecting sound events in the real world and its actual implementations as inputs to the virtual or mixed world.

J. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2007, LNCS 4551, pp. 653–662, 2007.

2 Related Work

There are some studies on sound inputs into the VR space with one microphone 1. An ON/OFF switching function could be easily implemented only by detecting input sounds. Mihara et al. 2 developed "the migratory cursor" system which operates a cursor by a certain nonverbal vocalization as well as voice commands. However, some of the testers felt unnatural to use the nonverbal vocalization as commands.

On the other hand, as a study example using microphone array, Nagai et al. proposed an accurate speech recognition method in noisy environments, which estimates sound source direction and enhances only sounds of this direction 3.

Our method can detect not only ON/OFF but also the direction and location of sound events occurred in the real world by using microphone arrays. Compared to the studies described above, we aim at a real world oriented and intuitive interactive device used in MR space.

3 Key Component and Its Function

3.1 Wearable Microphone Array

Fixed type of linear microphone arrays have been investigated in the field of acoustics. One of its drawbacks is that it can work well only in a limited range of the front direction because of the low angular resolution in the crosswise direction. In this research, we newly use a microphone array in a wearable fashion by attaching it onto a head mounted display. Since the user moves freely, the array can capture sound constantly in his/her front direction and near the sound source in an effective range of magnetic sensor. Consequently it can estimate angular resolution with higher accuracy. Fig 1 shows both fixed and wearable microphone arrays used in this research.

3.2 Direction Estimation of Sound Events

A sound source direction can be estimated by one microphone array. CSP (Cross-power Spectrum Phase analysis) method is used for sound source direction estimation algorithm 4. This method gives a direction of sound source in a horizontal plane. It is assumed that the environment using this system includes background noise, so that additional estimation errors are expected. Therefore, using Nishiura's method adding CSP coefficient of several microphone pairs 5, we implement the noise robust system.

3.3 Localizing Sound Events

Two or more microphone arrays can localize a sound source. In this research, the position and the orientation of the fixed type microphone array are measured beforehand. On the other hand, the position and the orientation of the wearable type microphone array are determined by a magnetic sensor. Besides, this system using two microphone arrays can estimate the location only on the horizontal plane. Estimable area and accuracy depend on the interval of microphones, the number of

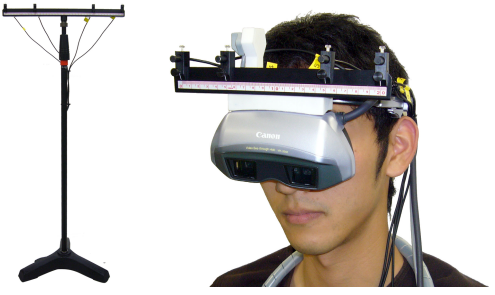


Fig. 1. Microphone arrays (Left: fixed type, Right: wearable type)

microphones, the sampling frequency, the positions of two microphone arrays, and the angle between two microphone array's lines.

4 System Overview

shows the system configuration detecting sound events in the real environment and reflecting it into the MR space. We use Canon MR Platform system for managing and displaying the MR space. Users watch the MR space through HMD (Canon VH-2002). Magnetic sensor 3SPACE FASTRAK (Polhemus) detects position and orientation of the HMD.

We use microphone arrays Fig. 3) for detecting sound events in the real environment. Signal from microphone array is amplified to a line level by Microphone Amplifier (PAVEC MA-2016C 16ch Microphone Amplifier). After going through AD converter (Thinknet DF-2X16-2) with a sampling frequency of 32kHz, the signal is input into PC for detecting the direction and location of sound events.

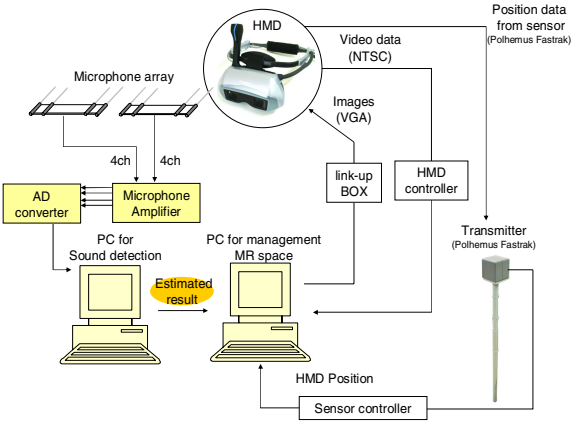


Fig. 2. System configuration

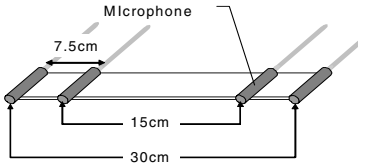


Fig. 3. Microphone array configuration

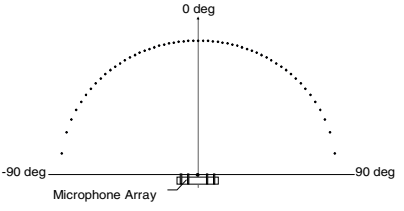


Fig. 4. Angle resolution of fixed type microphone array (sampling frequency: 32 kHz)

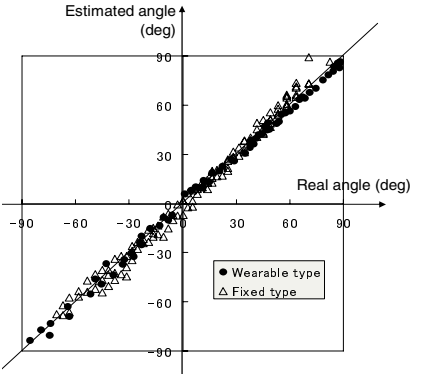


Fig. 5. Results of direction estimation for fixed type and wearable type microphone array

5 Evaluation

5.1 Accuracy of Direction Estimation

Fig. 4 shows the angle resolution of the fixed type microphone array on the horizontal plane. It can detect the direction in the range between -90 to 90 degrees with 58 points. The resolution becomes lower near the -90 and 90 degree (crosswise) directions. In contrast, since the wearable type microphone array can constantly track the sound source in front of the array, the accuracy becomes higher.

Fig. 5 shows the relation between the estimated angle (vertical axis) and the real angle (horizontal axis) of the fixed type and the wearable type microphone array. The 45 degree slope line in this graph indicates an ideal case that the estimated angle is the same as the real angle. The results of the wearable type are overlapped on the ideal line much better than the results of the fixed type. Especially, when looking at the results of estimation in the crosswise direction, the wearable type microphone array shows a better accuracy than the fixed type's.

5.2 Accuracy of Localization

We evaluated the localization accuracy depending upon the layout of two microphone arrays in the cases of 90, 120, 180 degrees of the angle between the two arrays.

Fig 6 shows the error map for the layout of 180degrees, 120 degrees and 90 degrees respectively. Each bubble size shows the error size. The dotted lines show the front regions of the microphone arrays. In this paper, the front region of a microphone array is defined as the range (from - 34.5 to +34.5 degrees) where the angle resolution of each microphone array is smaller than 2.5 degrees from Fig. 4. According to this

figure, localization accuracy is higher in the intersection region of the microphone array's front regions, and the error becomes larger for the sound source being further away from the microphone array, even though it is in the intersection region.

Fig 7 shows the error average of localization for three cases of the layout of two fixed microphone layout. From this figure, the error average of the 120 degrees case is found to be smallest.

We also tried the case where one of the fixed type microphone arrays was replaced by a wearable type for the best array layout of 120 degree angle as is also shown in Fig 7. The result of the latter case is shown in Fig **8** where A is the wearable microphone array's position in the real environment, and B is the position detected by magnetic sensor on the HMD.

The tendency of the result from one fixed and one wearable type microphone array (Fig **8**) seems to be almost the same as the former result. However, in the case of using wearable microphone array, the location and direction error of the magnetic sensor causes the sound localization errors becoming larger. If the magnetic sensor on the HMD is far from the transmitter, its error becomes also larger. Consequently, in figure 7, the error average seems to become larger for a combination of wearable type and fixed type, than those using two fixed types.

However, there are some advantages for wearable type microphone array. Firstly, sound source would be always taken in front of the microphone array because the user

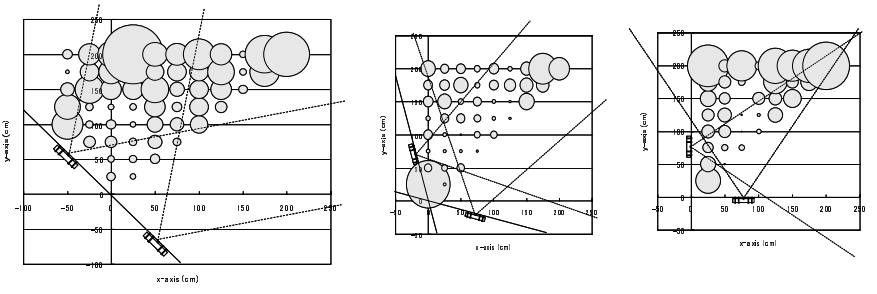


Fig. 6. Error maps (Left: 180deg, Center: 120deg, Right: 90deg)

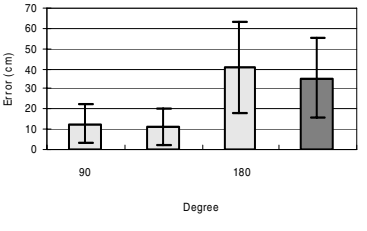


Fig. 7. Error averages of localization

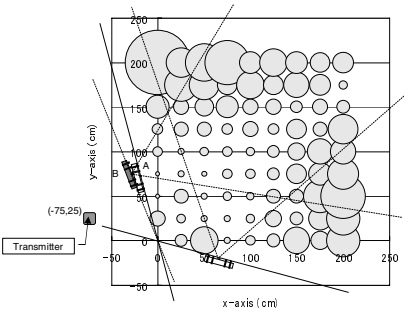


Fig. 8. Error map (Combination of fixed type and wearable type microphone array)

can turn to a sound source in the effective area of magnetic sensor. Secondly, the distance between the microphone array and the sound source would be shorter because user can get closer to the sound source.

6 Implementation of New Interaction Devices

6.1 Discrete Menu Selection Interface

As one example of the new interaction device using sound source direction, we tried to implement a menu selection interface using one wearable type microphone array. Some CG menu items aligned cylindrically are superimposed onto the real scene around the user. Upon recognizing the menu items, the user generated a sound signal in front of the item which he/she wants to select. In other words, the sound source direction detected by the wearable type microphone array is used as "pointing," and occurrence of the sound signal is used as "input".

The menu items are aligned with a radius of 60cm from the user within a semicircle of front 180 degrees (Fig 9). When the direction of sound source is recognized, the menu item which is located in the estimated direction is "Selected," (Fig 10) and the colour of the menu item is changed from green to pink (Fig 11).

In this system, a handclap is chosen as a sound source, since it is simple and intuitive for people. Other sound sources satisfying the following conditions could also be available, for example a buzzer and a castanet.

- Point sound source
- Easy to grasp
- Easy to generate a sound
- Able to make a sound only when users need

An assessment experiment with 10 subjects was made in order to evaluate the utility of this interface. In the experiment, almost all subjects directed their wearable type microphone array on the HMD toward the target menu item and generated a sound in front of the menu item. At the beginning of the experiment, some selecting error has occurred when the sound is small or pointed to the boundary area. Finally, all subjects could select menu items reliably with all devices (hands clap, a buzzer, and a castanet) in a few trials. Fig 12 shows the rate of successful input into each menu item by the castanet in the case of two, four, and six items. This result means that our system constantly estimated sound source direction in a high accuracy range of microphone array. In addition, the sound generated by themselves seemed natural as auditory feedback of input confirmation. Consequently, all these feedbacks seemed to help user's learning.

We developed, after all, a general-purpose interface like menu selection. If sound sources and CG objects are changed, a sound event can be used for a command input in various cases, and extendible for various applications.

6.2 Non-step Direction Selection Interface

In the previous section, we achieved the discrete menu selection interface. As the novel step, we developed the non-step (in other words, non-discrete) direction

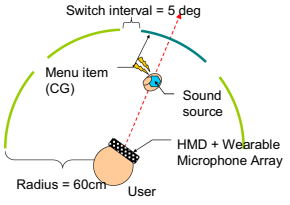


Fig. 9. Layout of menu items (in case of four menu items)

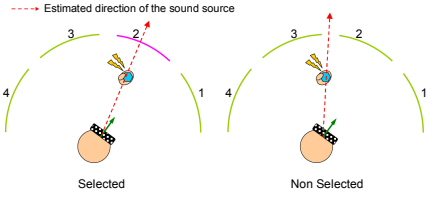
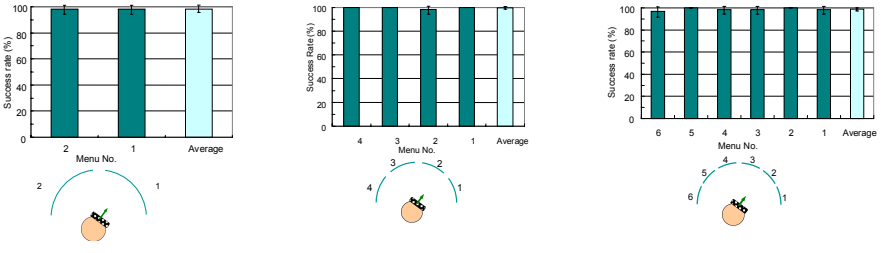


Fig. 10. Menu selection method

Fig. 11. A scene selecting a menu item using a sound device



(a) Two items (b) Four items (c) Six items

Fig. 12. Success rate of menu item selection (by the castanet)

selection interface. Firstly, we implemented the system which estimates a sound source direction, and superimposes a CG object in the direction as shown in Fig 13. This interface seems to be enough to be used as interactive device

This function was implemented in the MR attraction named "Watch the Birdie!" which is demonstrated in VRSJ 11th conference in Sendai, Japan in September, 2006 as technical exhibit. In this attraction, users can watch many birds (CG) flying in the air (Fig 14). The birds fly from the direction indicated by a sound source "birdcall[1]" (Fig 15). The result means that a very intuitive sound device can be used as an interaction device in the MR space.

However, unlike in fundamental experiment in adjusted environment, estimation error would be expected to occur more often by user and audience voice, and ambient noise in the exhibition hall. Therefore, we gave a weight in high frequency range

[1] Birdcall is a device imitating the sound of a bird. It is used for bird watching.

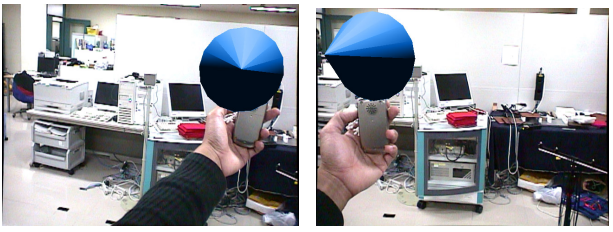


Fig. 13. Superimposing a cone-shaped CG object to the direction of sound source (A mobile phone speaker) at a distance of 50cm from the user

Fig. 14. User's view of "Watch the birdie!"

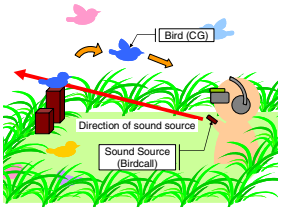


Fig. 15. In "Watch the Birdie!," user can select a CG bird using the sound device

because the frequency of the birdcall sound is higher than noise, and also discriminated the input sound from noise by "weighted CSP analysis" method which Denda et al. proposed 6. The visitors of this exhibition commented that this device was intuitive and easy to understand.

6.3 Localization of a Sound Event and Its Response

Localization of a sound event could be used to superimpose a CG object at the sound source position. Fig. 16 shows that an octahedral crystal (CG) is superimposed at "the location" of sound source (handclap) estimated by one fixed type and one wearable type microphone array in horizontal plane at a height of user's eyes (note that in Fig 13, the CG object was superimposed just onto "the direction" looked from the user).

This function was also applied to "Watch the Birdie!" This system determines the location of a mother duck which is a real toy with a speaker squawking, and then many ducklings (CG) move toward their mother (Fig. 17). In addition, a user can call

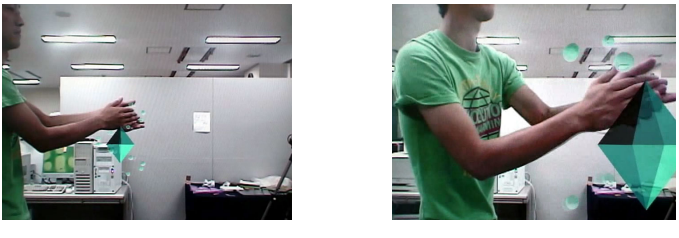


Fig. 16. Superimposed CG object in the location of a sound source (handclap)

Fig. 17. Ducklings (CG) gather toward mother duck (real toy object with speaker) by using sound source localization in "Watch the Birdie!

the ducklings around him/her using the birdcall. This is the designation of gathering spot by localization of sound source, so that it is different from indicating direction by a sound source described in the preceding section.

By tracking the estimated location of a sound generated continuously, it is possible to guide a CG object. It could also be used as a novel representation method such as a paint tool which draws a trajectory of a sound in the air.

7 Conclusion

We have developed novel interfaces to reflect sound events in the real environment into the MR space, and these are unique and useful interactive interface.

In these interfaces, we implemented not a single microphone customarily utilized in the field of VR, but the microphone arrays which are remarkable in the field of acoustics. Accordingly, we can use them not only for ON/OFF of sound events but also for direction and location of sound events as inputs into the MR space. We have also proposed a new usage of microphone arrays by implementing both traditional fixed type and a new wearable type microphone array attached onto a HMD.

The intuitive interface that was not provided by other methods became available by implementing our proposed approaches to the MR system. This function does not necessarily be used only in MR, but is expected to be widely used in a general system.

Acknowledgements. This research is supported by the Japan Society for the Promotion of Science through Grants-in-aid for Scientific Research (A), "A Mixed Reality system that merges real and virtual worlds with three senses."

References

1. Irawati, S., Calderon, D., Ko, H.: Spatial ontology for semantic integration in 3D multimodal interaction framework. In: Proc. of VRCIA 2006, pp. 129–135 (2006)
2. Mihara, Y., Shibayama, E., Takahashi, S.: The Migratory Cursor: Accurate Speech Based Cursor Movement by moving Multiple Ghost Cursors using Non-Verbal Vocalization. In: Proc. of ASSETS 2005, pp. 76–83 (2005)
3. Nagai, T., Kondo, K., Kaneko, M., Kurematsu, A.: Estimation of Source Location Based on 2-D MUSIC and Its Application to Speech Recognition in Cars. In: IEEE Proc of ICASSP 2001, vol. 5, pp. 3041–3044 (2001)

4. Omologo, M., Svaizer, P.: Acoustic Event Location Using a Crosspower -Spectrum Phase Based Technique. In: IEEE Proc. of ICASSP 94, Adelaide vol. 2, pp. 273–276 (1994)
5. Nishiura, T., Yamada, T., Nakamura, S., Shikano, K.: Localization of multiple sound sources based on a CSP analysis with a microphone array. In: IEEE Proc. of ICASSP 2000, vol. 2, pp. 1053–1056 (2000)
6. Denda, Y., Nishiura, T., Yamasita, Y.: A Study of Weighted CSP Analysis with Average Speech Spectrum for Noise Robust Talker Localization. In: Proc. of 9th EUROSPEECH 2005, Lisbon pp. 2321–2324 (2005)