(7) Reality Speech/Audio Communications Technologies

Yasuyo Yasuda and Tomoyuki Ohya

The advent of mobile broadband communications provides a platform for super reality communication, even when applied in the mobile communications environment. This article describes the technologies that can help bring about a new approach to mobile communication by increasing the quality of the transferred speech/audio to the limit, and by the use of 3-D audio technology.

1. Introduction

Like two sets of automobile wheels, there is a mutual dependency between the development of high bitrate data communication infrastructures and the pervasion of applications taking advantage of the opportunities for high speed communication they provide. This mutual dependency supports the dynamics of today's Internet application development. In mobile communications as well, high bitrate data communication infrastructures such as the 3rd-Generation mobile communication (IMT-2000: International Mobile Telecommunications-2000) and Wireless Local Area Networks (WLAN) are being rapidly established. However, the applications that make full use of the advantages of mobile broadband communications are scarcer.

Transmitting speech was the starting point for communication via mobile phones. As the mobile communication infrastructure is advanced, leading to the expansion of inexpensive mobile broadband, it is expected that richer speech experiences will be made widely available. As shown in **Figure 1**, audio

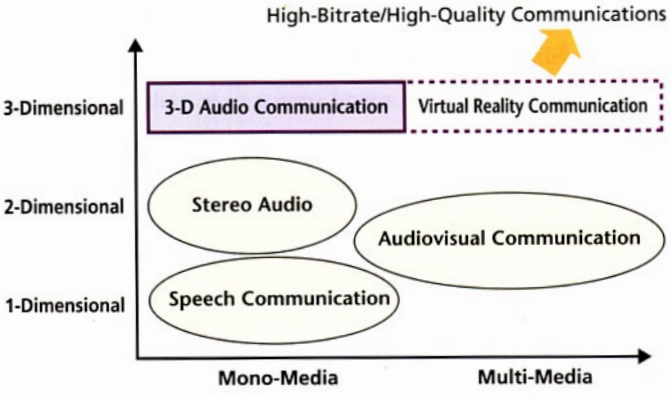

Figure 1 Multi-Dimensional/Multi-Media Communication

transmission in mobile communication systems used to be a monaural mono-medium for speech communication only; however, it has evolved gradually to cover multi-dimensional/multimedia content as well. This is typically seen in DoCoMo's M-Stage stereo music transmission and Freedom Of Mobile multimedia Access (FOMA) video phones. The form this multi-dimensionality will take when pursued to the limit can be referred to as reality speech/audio communication. Transmission of a three-dimensional (3-D) sound field to a mobile device can help establish a virtual sound space in which users can feel as if they are present in a remote auditory space. With advancement in 3-D audio technology, the user experience can approach the desired level of realism to achieve reality speech/audio communication. This level of realism is the ultimate goal of communication in the audiovisual multimedia environment.

DoCoMo considers that elevating the user experience to achieve reality speech/audio communication is the key to bringing about new applications, and to eventually creating new forms of mobile communication. Research is being conducted to open up the possibilities offered by mobile broadband, thus starting a new spiral of evolution in applications and in forms of communication.

This article focuses on 3-D speech/audio mobile communication technologies that improve the transmission quality of audio media to the highest degree possible. The article also explains the principles of those technologies. Specific service concepts and a prototype system are also introduced.

2. Virtual Audio Technology

2.1 Surround Audio Technology for Theaters

The most popular 3-D sound field technology in use is the multi-speaker systems used in movie theaters. Formats such as Dolby Surround [1], Dolby Digital [2] and DTS [3] are widely commercialized to reproduce a surround audio space by playing a recorded multi-channel sound source over multiple speakers installed in a theater. Similar systems have been commercialized for home theaters as well, and many surround speaker systems that support 5.1-channel formats (**Figure 2**) are being manufactured for consumer use in concurrence with the popularization of DVDs.

However, these systems are mainly aimed at static reproduction, such as presenting a sound source or sound effect from behind a listener. This limit on 5.1 systems means that they are not sufficiently advanced for uses that require precisely reproducing a complex sound space. An additional impediment to these systems in being directly applied to mobile communication is that it is not realistic to carry the multiple speakers required for playback.

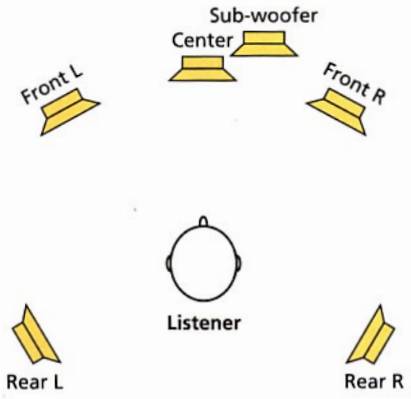

Figure 2 5.1-Channel Surround

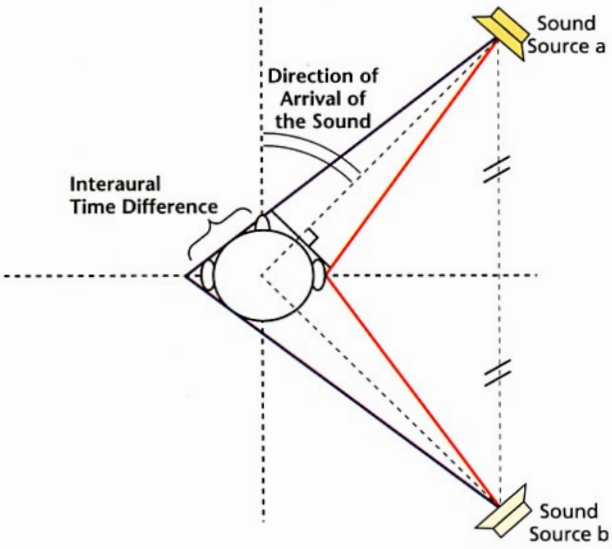

Figure 3 Assumption of Sound Source Arrival Direction

The following sections introduce the idea that virtual audio technologies can be [4], [5] effective solutions to this problem, by creating a 3-D sound field.

2.2 Binaural Playback Technology

Humans can estimate the direction from which a sound arrives with a very high level of accuracy, even with their eyes closed. As shown in **Figure 3**, the direction of arrival of a sound from source a can be estimated by perceiving the time lag between the respective moments at which the sound reaches the right and left ears and by the difference in sound pressure on arrival. Sound source b, yields the same interaural time lag as sound source a. In this case, humans can distinguish, by experi-

Photo 1 Dummy Head

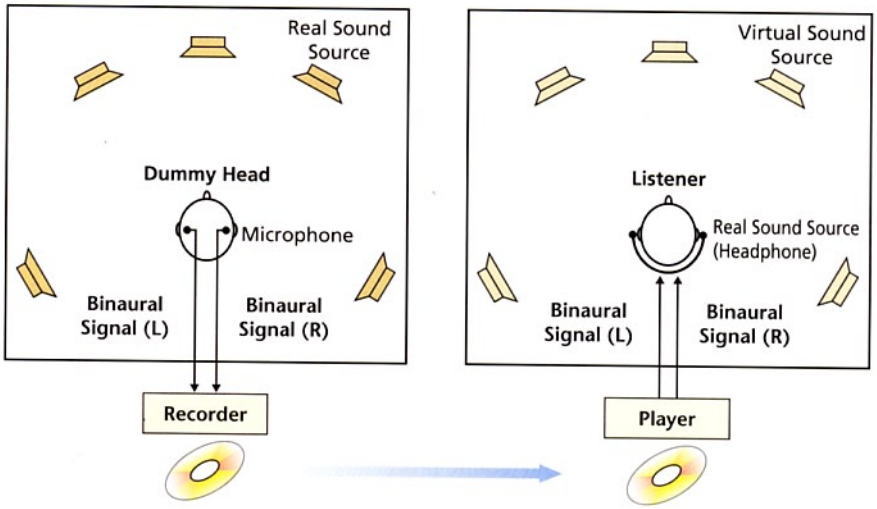

Figure 4 Binaural Playback System

ence, whether the sound is coming from the front or the back. This is because the pinna (earlobe) influences the frequency response of the audio transmission path from the sound source to the eardrum differently, depending on the direction of arrival of the sound. Binaural playback technology utilizes this principle to reproduce a sound environment, by providing different audio signals to the right and left ears via a set of headphones. The effect is to make users feel as though they are present in a remote auditory space.

When a normal music track recorded in stereo is played via headphones, the sounds appear as if localized inside the head of the listener, portraying the sound in a flat and lifeless manner. This phenomenon is called "in-head sound/image localization." In contrast to this, when a music track is played utilizing binaural playback technology via headphones, the listener experiences the sound as if coming from a natural position outside of his/her head, thus achieving "out-of-head sound/image localization". The sound appears as if it were coming from outside of the head, in spite of the fact that it is played back via a set of headphones.

(1) Binaural Recording

One approach to precise reproduction of the sound arriving at a listener's right and left ears in the actual environment involves Binaural Recording. In Binaural Recording, sound is recorded using microphones attached to the right and left ears of a dummy head (**Photo 1**) and subsequently played back via a set of headphones (**Figure 4**). A number of binaurally recorded CDs dedicated for headphone playback only have been released in several genres of music. The typical genres of these CDs are those which place high importance on lifelike sound reproduction, such as classics and audio dramas.

(2) Binaural Playback by Simulation

Another approach to binaural playback is to measure the Head Related Transfer Function (HRTF) that represents how the sound reaches the right and left ears of a human, from a sound source located in a particular direction. The measured HRTFs are then used to produce binaural signals.

Figure 5 shows an example of measured HRTFs. The coefficient sequence of a transfer function represents the response, measured at the entry to the ear canals, to a sound source transmitting an impulse waveform. This figure shows that the Interaural Time Difference (ITD) and the Interaural Intensity

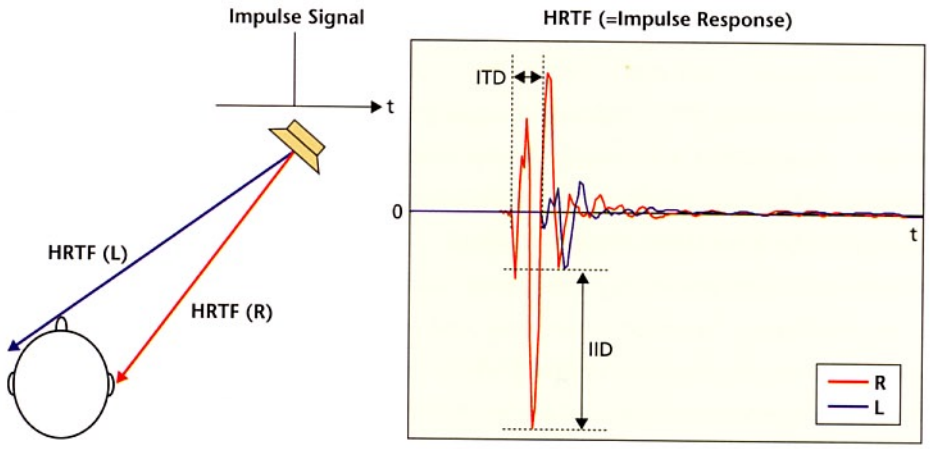

Figure 5 Head Related Transfer Function

Difference (IID), the two primary clues for human localization of a sound source in an actual environment, are present in the measured HRTF. Once this impulse response filter is measured, binaural signals can be reproduced by convolving this HRTF with any sound source waveform.

If time domain convolution processing is used, calculations of $O(N^2)$ are required for a HRTF with a filter length of N taps. Impulse responses in actual environments may be several seconds long, depending on the audio properties of the room. At CD quality sampling rate (44.1 kHz), these HRTF filters can be tens of thousands of taps long each. In the time domain, convolution processing of these filters would require processing of the order of several Giga operations per second, per HRTF filter. This amount exceeds the calculation capacity of practical Digital Signal Processors (DSP), and even fast PC CPU's. For this reason, a method of reducing the calculation amount to $O(N\log N)$ by the Fast Fourier Transform (FFT) is applied in some cases. However, the side effect of this method is to introduce a processing time delay, because calculations are performed block by block. Recently, solutions to this trade-off have been achieved. Efficient convolution processing with low delay is now possible through advancement in signal processing technologies, such as those implemented in Lake Technology's [7] Convolution methods. HRTFs, in addition to ITD and IID, incorporate the effects of sound wave reflection and diffraction of the head, the pinna, and other factors. It is known that individual differences in the shape of the head and the pinna may have a significant impact on the reproduction of localization. Recently, a commercial product that has mitigated these problems, by reducing the negative impact of typical HRTF filtering properties, has been introduced.

The Dolby Headphone [8] technology developed by Lake Technology and licensed by Dolby is a system that reproduces the 5.1 format audio format over ordinary stereo headphones using the aforementioned HRTF convolution signal processing. The system is installed not only in home theater products, but also in PC based DVD players and portable MD players.

(3) Binaural Playback with Head Tracking Function

With binaural playback technology, "out-of-head" sound image localization can be achieved for headphone playback. However, it causes an unnatural situation where the sound image moves along with the motion of the listener's head. In contrast, implementing a head tracking function, that follows the motion of the head, can reproduce a more natural and realistic sound field. Theoretically, a sound space identical to a certain space can be produced by feeding back the listener's head orientation and sequentially updating the HRTF filter such that it reflects the correct angle measured in the space.

Humans dynamically perceive sound images based on the changes in information obtained by both ears. The localization accuracy is thus improved by the use of head motion feedback in a binaural simulation system. For this reason, the addition of head motion feedback can be expected to make a significant contribution to the establishment of virtual audio applications.

Figure 6 shows the results of a localization accuracy test with the listener's head fixed. Using the multi-speaker system shown in **Photo 2**, subjects were asked to guess the direction of arrival of sound from a sound source. The direction 0° is in front of the subject. The stimuli were played back at random with different playback methods (loudspeaker playback/binaural playback) and different angles (twelve angles with 30° intervals). Two tests were conducted with different randomizing pat-

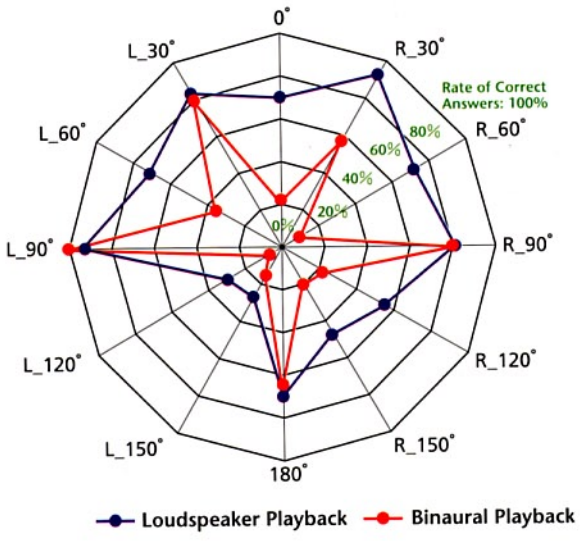

Figure 6 Results of Localization Accuracy Test with The Listener's Head Fixed

Photo 2 Localization Accuracy Test Facilities

terns by 21 persons. In the graph showing the rate of correct answers, the blue line is the percentage of correct answers for loudspeaker playback and the red line is correct answers for binaural playback by simulation. The closer to the circle's center the plotted line is, the lower the level of accuracy is. The results show that, when the head is fixed, the level of accuracy deteriorates, even in a real sound field where the sound is actually played back over the loudspeaker. The level of accuracy was 80 to 90% even at angles considered to be easier for localization (L_90° and R_90°) and it falls to about 30% in case of oblique directions. The results also show that the level of accuracy falls further in the case where the virtual sound field is created by binaural playback. However, when the subjects were allowed to move their head freely, the accuracy level improved to 100%, not only in the case of loudspeaker playback, but also in the case of binaural playback with a head tracking function implemented.

An example of commercial products incorporating applications using this head tracking technology is the digital surround headphone system [9] by Sony. The surround sound field is fixed irrespective of the head motion. Several problems still remain to be solved, such as impairment of HRTF accuracy caused by drift/accuracy of the sensors and the limited calculation capacity available. For this reason, this technology has not been widely applied yet.

2.3 Transaural Playback Technology

As described above, having users wear headphones allows the reproduction of virtual spaces using binaural playback technology. Transaural playback technology attempts to achieve the same effect, using a limited number of loudspeakers (for example, two loudspeakers), instead of headphones. When binaural signals are played back over two loudspeakers, the desired sound image is not reproduced correctly because the signal intended only for the right ear (X_R) also reaches the left ear. This phenomenon is called cross-talk. However, if the transfer function from the loudspeaker to the opposite ear can be obtained by employing the same approach as for the HRTF measurement, then cross-talk can be cancelled by applying the appropriate inverse filter as illustrated in **Figure 7**. The use of cross-talk calculation can help provide the correct signal to the correct ear, enabling the loudspeakers to reproduce a sound image equivalent to binaural playback. This technology also has commercial examples, such as the following:

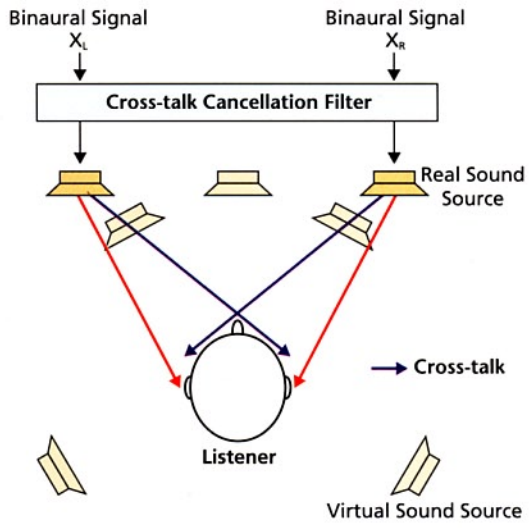

Figure 7 Transaural Playback System

① Dolby Virtual Speaker [10]: This product has been developed by Lake Technology, licensed by Dolby and commercialized as a product for PCs. Transaural technology provides 5.1-channel surround experience with only two speakers installed in the front.

② 3-D Sound Speaker System P2DiPOLE [11]: This system supports 5.1-channel audio with only one body and two speakers. It is characterized by two closely located cylindrical speakers to facilitate cancellation of cross-talk.

3. Mobile 3-D Audio Communications

Although there is a difference in whether headphone playback or loudspeaker playback is used, all the commercialization examples introduced in the previous section are designed for home theater use. These systems provide a useful experience of 5.1-channel surround audio with DVDs, etc. at home. As for application of virtual audio technologies to mobile communications, sufficient examination has not been conducted and it can be said to be an unexplored field.

This section explains specific service concepts of mobile 3-D speech/audio communication, which can be achieved by applying the virtual audio technologies described in Chapter 2 to mobile communication. Technical issues in the realization of mobile 3-D speech/audio communication are also mentioned.

3.1 Service Concepts

Virtual audio technologies can be applied in two different ways. One way is to exactly replicate sound sources in an actual space; the other method is to create a virtual audio space that

has no equivalent real space.

Described below are some examples of services utilizing these different ways of applying virtual audio technologies.

(1) Super Reality Service

In basic speech and video phone services, virtual audio technologies can provide users with the simulated audio of a face to face conversation, or allow a voice to be heard as if it were arriving from a video phone's screen, instead of hearing the voice in-head. Users will be able to enjoy conversation in an environment that is closer to the reality. Thus, it can be expected that fatigue will be reduced, even for long conversations.

In three-way calling, as shown in **Figure 8**, it becomes easier to identify participants by locating their individual voices at separate positions in a virtual audio space. Adding head orientation tracking allows the experience to approach the naturalness of an audio conversation in a real meeting space when a listener joins a meeting from a remote location (**Figure 9**). With a large number of participants, the benefit of separating of a speaker may be more obvious, and even complicated discussions during a remote meeting can be recognized more naturally.

In addition, multimedia distribution and broadcasting services can utilize these technologies to achieve super reality—for example, a mobile theater providing surround audio equivalent to that experienced in a movie theater at any time and at any place, The technologies can also assist in broadcasting of sport games and concerts, conveying the real atmosphere, by making users feel the sensation of being present at the remote venue.

(2) 3-D Audio Navigation

It is envisaged that virtual audio technologies can be applied to navigation services. In addition to visual information, a 3-D sound field could give the user information on the direction and position of a target. This would require the use of position and head orientation tracking. This could for instance be useful when attempting to find a rendezvous point. With the help of location information from, for example, the Global Positioning System (GPS), the relative position of the other party could be recognized from the direction of arrival of their voice in the listener's virtual sound field, even with no visual information present (**Figure 10**).

3-D audio navigation is also applicable to virtual museums, galleries or town guides. In the case of a gallery, an audio track explaining a particular exhibit could be co-located in a 3-D sound field with the actual exhibit. Users could be guided

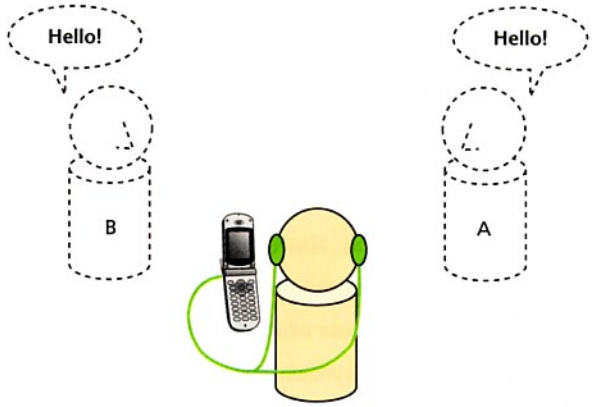

Figure 8 Three-way Calling in 3-D Audio Space

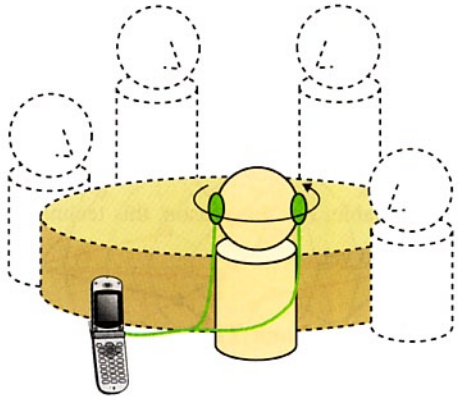

Figure 9 Remote Meeting in 3-D Audio Space

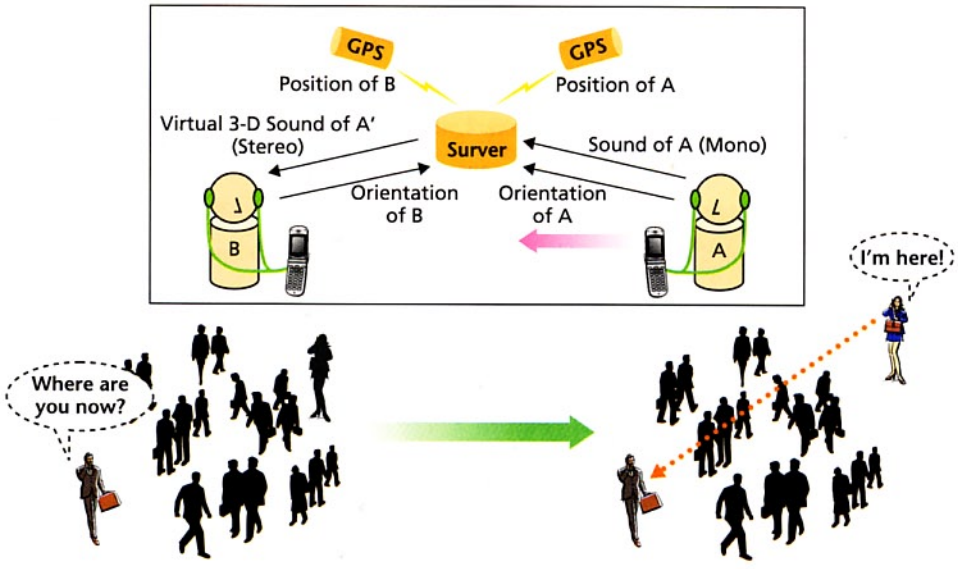

Figure 10 Rendezvous Navigation Concept

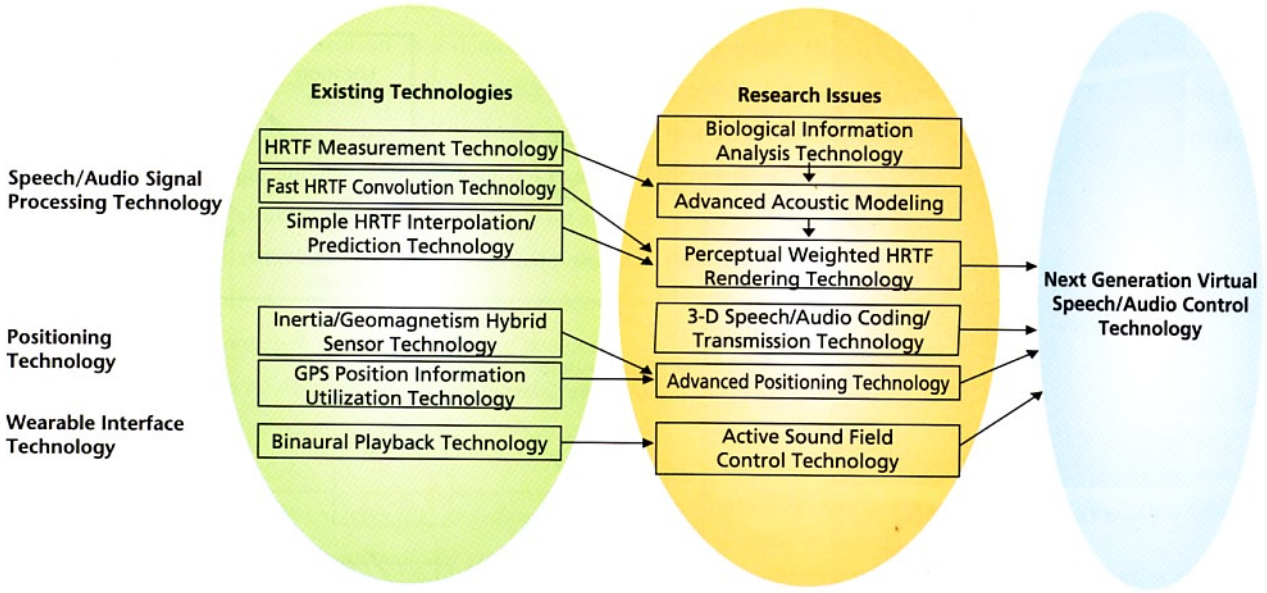

Figure 11 Technical Issues on Advanced Virtual Speech/Audio Control

through the museum using the 3-D sound field. In the case of a town guide, 3-D audio could be used to provide tourists with audio information advertising a particular shop or product in a downtown area.

(3) 3-D Audio in Games

Communication and entertainment are becoming more integrated as symbolized by the emergence of on-line games for PCs and TVs and games for i-appli, a mobile phone service provided by DoCoMo. Currently, "sound" is generally subordinate to the visual aspect of a game or used simply as background music. It would, however, be possible to provide more realistic and absorbing experiences by integrating 3-D audio into shooting games, RPGs and action games.

In Europe, there is a mobile multi-player shooting game which can be played against other mobile users [12]. By incorporating 3-D audio technologies, location-based games can add the element of synchronized 3-D sound fields to their traditional visual-centric nature. Utilizing the user's location and mobility in providing new experiences, is a very good use of the strength of mobile communication.

3.2 Technical Issues

As discussed above, basic technologies to reproduce virtual audio have already been advanced in a variety of fields. Still, as shown in **Figure 11**, there are many issues to be solved before these technologies are applicable to mobile communication. One is the signal processing technologies related to the HRTFs. In order to perform simple filtering with low requirements to processing amount suitable for mobile communication, it is necessary to develop methods for optimization that proactively utilizes the properties of human perception. The key technology for this optimization is the perceptual weighted virtual audio reproduction technology based on analysis and modeling of the human auditory perception mechanism. Furthermore, examination of how to transmit this information via mobile communication paths, positioning technology optimized for the mobile environment, and advancement of the human interface are particularly required in the field of mobile communications.

4. Prototype System

In order to investigate the feasibility of mobile 3-D speech/audio communication, we have developed a system (**Figure 12**) that can be used to prototype the service concepts described in Section 3.1.

This section introduces the system.

4.1 System Configuration

The system consists of location sensors, head trackers, Personal Digital Assistants (PDAs), headphones, a virtual audio rendering server, and a Wireless LAN (WLAN). User positions are detected with location sensors installed on the ceiling. User head orientations (direction) are detected with the head tracker installed on the headphones. Based on information from user positions and head orientations (**Figure 13**), the virtual audio rendering server produces virtual audio and transmits it to the PDA clients via the WLAN. The PDA clients and associated

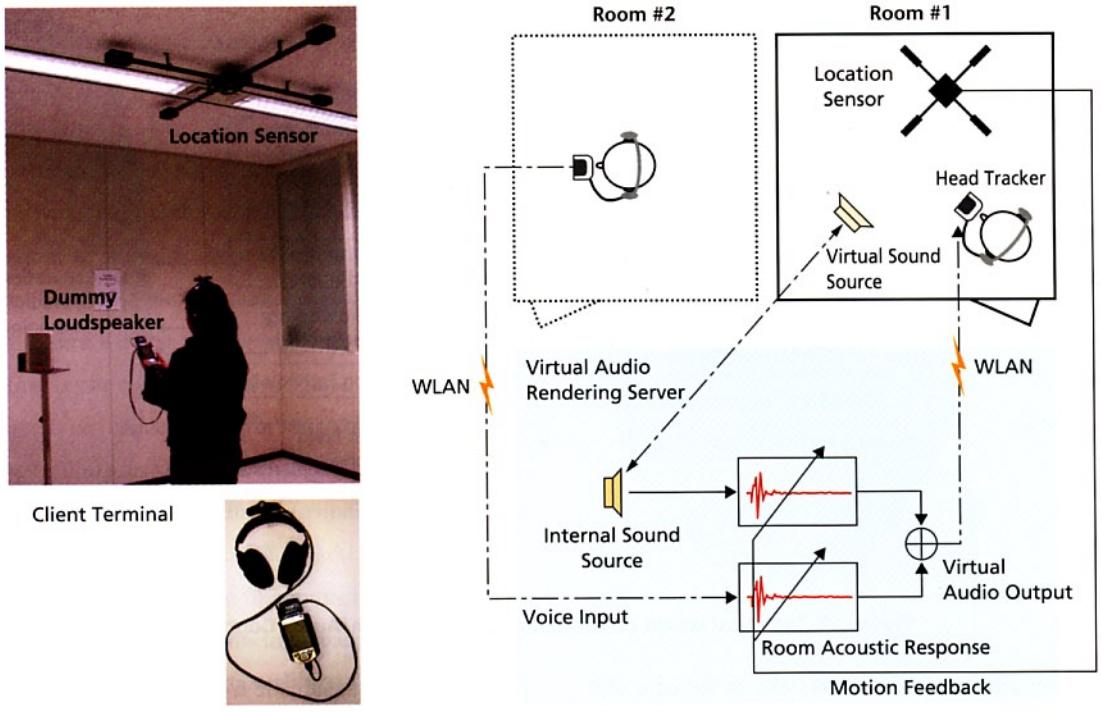

Figure 12 Prototype 3-D Speech/Audio Communication System

head tracking are completely wireless, and give a good approximation of a mobile terminal. Reproduction of communication in the virtual audio environment is achieved by allowing users to communicate with each other using the microphones built into the PDAs.

4.2 Technical Overview

Though the location sensor can detect a position along each of the x-, y- and z-axes with an accuracy of 7 mm RMS (Root Mean Square), the system utilizes the information in the horizontal x-y plane only. Since the human sense of direction in the vertical direction is said to be weak compared to that of the horizontal plane, the amount of data processing is reduced by not using the z-axis (height information). For the same reason, the head tracker also utilizes information along the horizontal azimuth only. The detection accuracy of the head tracker is 0.25° RMS.

Production of virtual audio in the system is enabled by incorporating processing technology owned by Lake Technology. Advanced rendering techniques incorporating low latency long convolution are used. Non-individualized HRTFs are combined with impulse response convolution and distance and direction panning to provide 3-D audio simulation. Finite Impulse Response (FIR) filters with impulse response lengths of approximately 7,000 FIR filter taps are used, providing a latency of around 15ms. Eight pairs of HRTFs are used. Distance perceptions and HRTF interpolation are achieved using advanced panning techniques. The use of these advanced techniques allows for reductions in both memory and processing power required to achieve the desired result.

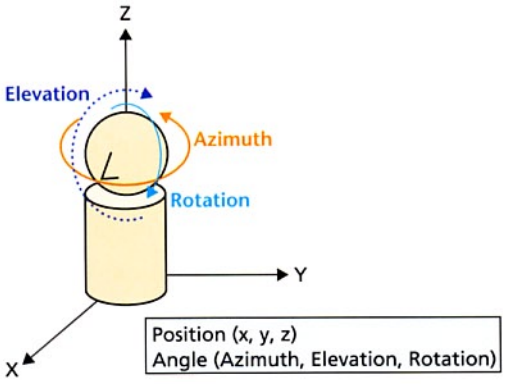

Figure 13 Position and Direction Information

The reduction of the processing amount and memory consumption allows the system to be executed comfortably on a PC with a standard Intel® Pentium 4 processor. It can accommodate real-time processing of motion feedback and HRTF convolution processing for up to three simultaneous users.

4.3 Functions

Two functions are achieved in the system. One is the reproduction of a 3-D virtual audio space. By feeding back the

motion of a user for real-time processing, a virtual sound source can be localized at a particular location in a demonstration room. As a demonstration, a dummy speaker box is placed in a demonstration room, and a sound source is localized at the box's position. In reality, the dummy speaker does not make any sound but in the virtual audio space created via a set of headphones, a user perceives the sound as if it came from the dummy speaker. Even if the user moves around in the demonstration room, the sound continues to be located at the position of the dummy speaker, with no deterioration in the sound quality. This is a demonstration of a communication scenario that provides a more natural sound field.

The second function is a 3-D speech/audio communication function achieved by virtual reproduction of the relative location of users. When two users are located in demonstration rooms 1 and 2, the voice of the other user heard via the headphones is localized (through the wall) at the relative position the user in the other room. This is a demonstration of 3-D audio navigation services, in which users can identify each other's position by perceiving the direction of arrival of the other's voice even though they cannot see each other.

Both of these two demonstrations help to conceptualize some basic functions of services achieved by the virtual audio technologies described so far. They provide a practical experience of advanced speech/audio services.

5. Conclusion

This article explained the value of reality speech/audio communication technology in providing a platform for attractive services for wireless broadband communication and introduced a prototype system for this technology. In order to apply the technologies to actual networks, further research will be conducted with this prototype system to solve technical issues, such as trade-off between the orientation accuracy improvement and processing amount reduction, and verification of network delay and speech/audio coding influences. We also intend to conduct research and development of sound field control and call control technology, which take advantage of the user's location and the mobility of mobile communication.

The current style of mobile speech communication involving monaural audio conversation by placing a mobile telephone terminal directly to the users' ear, may be replaced by technology to communicate in a 3-D audio space with headphones or earphones. Headphones may even become unnecessary if a small speaker in a mobile phone can be used to control the 3-D audio space. The role of mobile phones will evolve from being a device that conveys sound and information into being a device that conveys virtual reality—even including atmosphere and perceptions in a remote space.

REFERENCES

[1] http://www.dolby.com/tech/
[2] http://www.dolby.com/digital/
[3] http://www.dtsonline.com/aboutdts/index.shtml
[4] Ohya, et al.: "Audio Coding Technology", NTT DoCoMo Technical Journal, Vol. 8, No. 4, pp. 17–24, Jan. 2001.[Japanese Edition]
[5] N. Kitawaki, N. Sugamura, N. Koizumi: "Sound Communication Engineering", First Edition, pp. 172–178, 1996, Corona Publishing Co., Ltd., Japan. [In Japanese]
[6] Corey I. Cheng, Gregory H. Wakefield: "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space", J. Audio Eng. Soc., Vol. 49, No. 4, 2001.
[7] http://www.laketechnology.com/
[8] http://www.dolby.com/dolbyheadphone/
[9] http://www.sony.jp/products/Models/Library/MDR-DS8000.html [In Japanese]
[10] http://www.dolby.com/dvs/white.paper.html
[11] http://www.iodata.co.jp/products/sounds/p2dp/index.htm [In Japanese]
[12] http://www.wired.com/news/wireless/0,1382,50205,00.html

GLOSSARY

DSP: Digital Signal Processor
FFT: Fast Fourier Transform
FIR: Finite Impulse Response
FOMA: Freedom Of Mobile multimedia Access
GPS: Global Positioning System
HRTF: Head Related Transfer Function
IID: Interaural Intensity Difference
IMT-2000: International Mobile Telecommunications-2000
ITD: Interaural Time Difference
LAN: Local Area Network
PDA: Personal Digital Assistant
RMS: Root Mean Square