Reality Speech/Audio Communications Technologies

Yasuyo Yasuda† *and Tomoyuki Ohya*

Abstract

The advent of mobile broadband communications provides a platform for super reality communication even in a mobile communications environment. This article introduces NTT DoCoMo's initiatives for developing a new approach to mobile communication by maximizing the quality of the transferred speech/audio and using 3-D audio technology.

1. Introduction

The development of high-bitrate data communication infrastructures is driving the spread of applications that fully utilize this communication speed and vice versa. High-bitrate mobile communication infrastructures for data such as the third-generation (3G) mobile communication (IMT-2000: International Mobile Telecommunications 2000) and wireless local area networks (WLANs) are also being rapidly established. However, applications that make full use of the advantages of mobile broadband communications are relatively scarce.

Initially, communication via mobile phones meant transmitting speech. As the mobile communication infrastructure advances, and inexpensive mobile broadband becomes available, richer speech experiences are expected to become widely available. As shown in Fig. 1, audio transmission in mobile communication systems used to be monaural single-medium for speech communication only; however, it has gradually evolved to cover multi-dimensional/multimedia content as well. This is typically seen in NTT DoCoMo's FOMA (which stands for freedom of mobile multimedia access) videophones and M-stage stereo music transmission [1]. Transmission of a three-dimensional (3-D) sound field to a mobile device can help establish a virtual sound space in

† NTT DoCoMo Multimedia Laboratories
Yokosuka-shi, 239-8536 Japan
E-mail: yasuda@spg.yrp.nttdocomo.co.jp

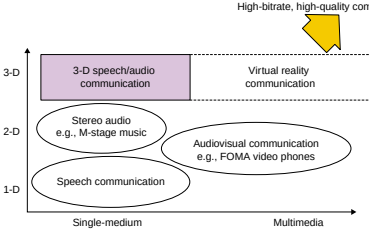

Fig. 1. Multi-dimensional/multimedia communication.

which users feel as if they are in a remote auditory space. Advances in 3-D audio technology will enable the user experience the desired level of realism to achieve "reality speech/audio communication", which is the ultimate goal of communication in the audiovisual multimedia environment.

NTT DoCoMo considers that raising 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 the 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. It also explains specific service concepts and introduces a prototype system.

2. Mobile 3-D audio communications

2.1 Service concepts

Many surround-sound speaker systems that support 5.1-channel formats are being manufactured for consumer use as DVDs grow in popularity. However, one obstacle preventing such systems from being applied directly to mobile communication is that it is impractical to carry the multiple speakers required for playback. One effective solution to this problem might be to apply binaural and transaural playback technologies [2] to mobile communications. Binaural playback reproduces 3-D sound fields via a pair of headphones, while transaural playback reproduces sound via a small number of loudspeakers.

Virtual audio technologies can be applied in two different ways: to exactly replicate the sound sources captured in an actual space and to create a virtual audio space that has no equivalent real space. Some examples of services utilizing these concepts are described below.

(1) Super reality service

The application of reality speech/audio communications technologies to basic speech and videophone services can provide users with the audio impression of a face-to-face conversation or allow them to hear a voice as if it were arriving from the videophone's screen, instead of hearing the voice in-head through a monaural headset. Users will be able to enjoy conversation in an environment that is closer to reality. Such benefits will reduce fatigue, even for long conversations.

In three-way calling, as shown in Fig. 2, it is easier to identify participants if their individual voices seem to come from different 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 (Fig. 3). With a large number of participants, the benefit of speaker separation may be more obvious, and even complicated discussions within a remote meeting can be understood more clearly.

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 an ordinary movie theater at any time and at any place. These technologies can also assist in the broadcasting of sport games and concerts: users feel the sensation of being present at the remote venue because the real atmosphere is conveyed.

(2) 3-D audio navigation

Virtual audio technologies are expected to support navigation services. Supplementing visual information, a 3-D sound field will give the user information about the direction and position of the target. This will require the use of position and head orientation tracking. This would for instance be useful when attempting to find someone at a crowded meeting point. With the help of location information from, for example, GPS the relative position of the other party could be recognized from the arrival direction of their voice in the listener's virtual sound field, even without visual information (Fig. 4).

3-D audio navigation is also applicable to virtual museums, galleries, and 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 through the museum by 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

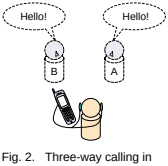

Fig. 2. Three-way calling in 3-D audio space.

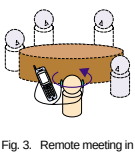

Fig. 3. Remote meeting in 3-D audio space.

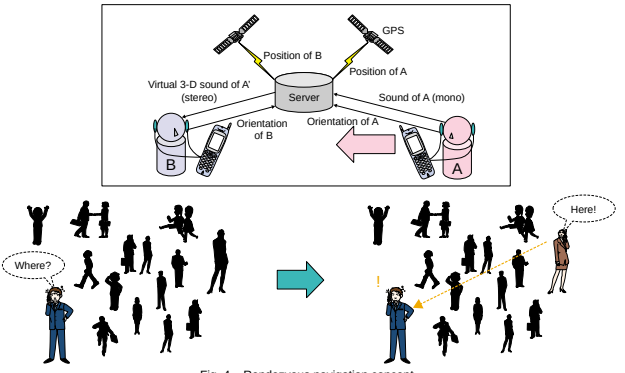

Fig. 4. Rendezvous navigation concept.

or product in a downtown area.

(3) 3-D audio in games

Communication and entertainment are becoming more integrated as evidenced by the emergence of online games for PCs and TVs and games for i-appli [3], a Java-based online software environment for mobile phones provided by NTT DoCoMo. Currently, "sound" is generally subordinate to the visual aspect of a game or used simply as background music, but it should be possible to provide more realistic and absorbing experiences by integrating 3-D audio into shooting games, role-playing games, and other action genres.

In Europe, there is a mobile multi-player shooting game that can be played against other mobile users [2]. By incorporating 3-D audio technologies, location-based games can add the element of synchronized 3-D sound fields to their traditional visual-centric tools. Utilizing the user's location and mobility in providing new experiences is a very good use of the strength of mobile communication.

2.2 Technical issues

The basic technologies to reproduce virtual audio have already been advanced in a variety of fields. Still, as shown in Fig. 5, there are many issues to be solved before these technologies are applicable to mobile communication. One is the signal processing technologies related to head related transfer functions (HRTFs) [5], which express how the sound reaches each ear of a human. In order to perform simple filtering with low complexity such that mobile communication devices can provide the processing performance, it is necessary to develop optimization methods that utilize the properties of human perception. The key technology for this optimization is perceptual weighted virtual audio reproduction based on analysis and modeling of the human auditory perception mechanism. Other strong requirements in the field of mobile communications are an examination of how to transmit this information via mobile communication paths, positioning technology optimized for the mobile environment, and improvements in the human interface.

3. Prototype system

In order to investigate the feasibility of mobile 3-D speech/audio communication, we have developed a system, shown in Fig. 6, that can be used to implement the above service concepts.

3.1 System configuration

The prototype system consists of location sensors, head trackers, PDAs (personal digital assistants) acting as client terminals, headphones, a virtual audio

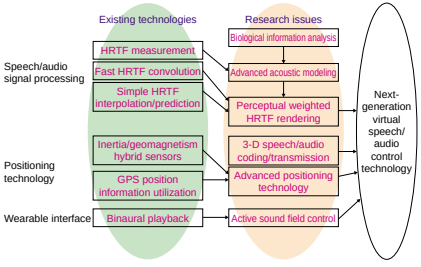

Fig. 5. Technical issues for advanced virtual speech/audio control.

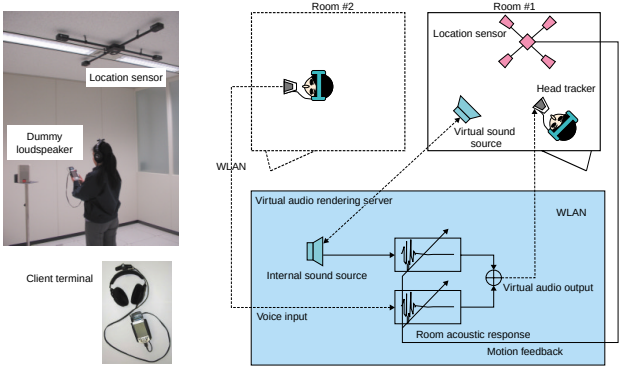

Fig. 6. Prototype 3-D speech/audio communication system.

rendering server, and a WLAN. User position is detected by location sensors on the ceiling. User head orientations (directions) are detected by the head tracker installed on the headphones. Based on information from user positions and head orientations (Fig. 7), the virtual audio rendering server produces virtual audio and transmits it to the PDAs via the WLAN. The PDAs and associated head tracking are completely wireless and provide a good approximation of mobile terminals. Communication in the virtual audio environment is reproduced by users communicating with each other via the microphones built into the PDAs.

3.2 Technical overview

The system utilizes only horizontal (x-y) plane information even though the location sensor can also detect height. Since the human sense of direction in

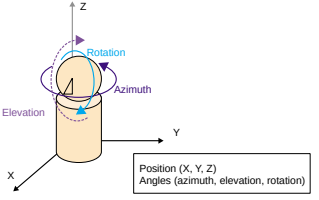

Fig. 7. Position and direction information.

the vertical direction is said to be weaker than that in the horizontal plane, the amount of data processing is reduced by not using tracking information from the z-axis (height), elevation, or rotation.

The system produces virtual audio by using advanced rendering techniques incorporating low-latency long convolution [6]. 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 7000 taps at 22.05-kHz sampling are used, providing a latency of around 15 ms. Distance perception and HRTF interpolation are achieved using advanced panning techniques with eight pairs of pre-stored HRTFs. The use of these advanced techniques reduces the memory and processing power required to achieve the desired result. The reduction 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 users at the same time.

3.3 Functions

The system offers two functions. One reproduces a 3-D virtual audio space. By feeding back user motion as part of real-time processing, it can localize a virtual sound source at a particular location in a demonstration room. In our demonstration, a dummy loudspeaker box is placed in a demonstration room, and the sound source is localized at the box's position. In reality, the dummy loudspeaker does not make any sound but in the virtual audio space created via the headphones, the user perceives the sound as if it came from the dummy loudspeaker. Even if the user moves around in the demonstration room, the sound continues to seem to originate from the position of the dummy loudspeaker, with no deterioration in sound quality. This is a demonstration of a communication scenario that provides a more natural sound field.

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

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

4. Conclusion

This article discussed the application of 3-D audio technologies to mobile communications to provide a platform for attractive services for wireless broadband communications. We have developed a prototype system that makes it possible to experience the vision of 3-D speech/audio communications in future mobile networks.

In order to apply these technologies to actual networks, we will conduct further research using this prototype system to solve technical issues, such as the trade-off between the orientation accuracy and processing cost, and to verify the network delay and speech/audio coding factors. We also intend to research and develop sound field control and call control technologies, which will make better use of the user's location and the benefits of mobile communication.

The current style of mobile speech communication involving monaural audio conversation by placing a mobile telephone terminal directly next to the users' ear may be replaced by technology that offers communication in a 3-D audio space using headphones or earphones. Headphones may even become unnecessary if the mobile phone has small loudspeakers that can generate a 3-D audio space. Mobile phones will evolve from devices conveying just sound and information to ones conveying virtual reality—even including the atmosphere and perceptions in a remote space.

References

[1] http://www.nttdocomo.co.jp/p_s/mstage/music/index.html (in Japanese).
[2] N. Kitawaki, N. Sugamura, and N. Koizumi, "Sound Communication Engineering," First Edition, pp. 172-178, 1996, Corona Publishing Co., Ltd., Japan (in Japanese).
[3] http://www.nttdocomo.com/corebiz/imode/services/iappli.html
[4] http://www.wired.com/news/wireless/0,1382,50205,00.html
[5] C. I. Cheng and G. 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.
[6] Y. Yasuda, T. Ohya, D. McGrath, and P. Flanagan, "3-D Audio Communications Services for Future Mobile Networks," AES 23rd International Conference, Helsingør, Denmark, pp. 158-163, May, 2003.

Yasuyo Yasuda
NTT DoCoMo Multimedia Laboratories.

She received the B.E. and M.E. degrees in communication engineering from Kyushu University, Fukuoka in 1997 and 1999 respectively. Since joining NTT DoCoMo in 1999, she has been engaged in R&D on audio systems for mobile communications. Currently, she is focusing on 3-D audio communications systems. She is a member of the Acoustical Society of Japan (ASJ) and the Institute of Electronics, Information and Communication Engineers (IEICE).

Tomoyuki Ohya
Executive Research Engineer, NTT DoCoMo Multimedia Laboratories.

He received the B.E. and M.E. degrees in electronic engineering from Kyoto University, Kyoto in 1986 and 1988, respectively, and the M.S. degree in management of technology from Massachusetts Institute of Technology, U.S.A., in 2000. He joined NTT in 1989, and has been working at NTT DoCoMo, Inc since 1992 engaged in R&D of digital speech coding technologies for PDC and IMT-2000. Currently, his main research interest is multimedia signal processing and the QoS architecture for 4th generation mobile communications networks. He is a member of ASJ, IEICE, and IEEE. He received the Young Engineer Award from IEICE in 1995.