Realistic Audio Teleconferencing using Binaural and Auralization Techniques

Seong-Hoon Kang and Sung-Han Kim

CONTENTS

ABSTRACT

The goal of telecommunication may be to enable the participants in distant places to communicate with each other in an environment as if they were in the same room. This paper introduces the reason why realistic audio display is useful in telecommunication, reviews some approaches to its implementation, and proposes an audio teleconference model which realizes a two-way telecommunication with realistic sensations using binaural and auralization techniques.

I. INTRODUCTION

One of the goals of telecommunication is to provide face-to-face communication, overcoming the distance between persons. An ideal telecommunication service should allow participants in distant places to communicate with each other in a natural environment, as if they were in the same room. In the audio industry, multi-channel sound systems are becoming popular for enhancing the presence of music. The basic value of multi-channel sound is a spatial sound image including the sound field information. Some experiments with stereophonic teleconferencing have been introduced which gave good sound image localization of each participant at the other site [1], [2]. Spatial sound is increasingly exploited as a vital communication channel in telecommunication networks.

Recently, wideband-ISDN(Integrated Service Digital Network) is eagerly being developed by telecommunication engineers with the expectation of realizing higher communication capabilities. As telecommunication technology has advanced, there is more need and expectation for a new type of realistic teleconferencing service that produces a virtual environment where individuals can hear, see and feel a seemingly shared atmosphere [3]. An ideal teleconference system requires methods for producing effects enhancing the sensation of sharing a space with participants at distant sites.

From the viewpoint of acoustical engineering, it is essential to re-establish the realistic sound field. One of the most useful effects is to give listeners a good sound image localization of the remote participant. The most feasible solution to this problem may be the application of the binaural or transaural techniques, which reconstruct the desired sound field in the vicinity of each ear. This method may be appropriate for personal communication. Another solution is the acoustic space synthesis, which reconstructs the entire sound field of a particular room or environment. This method is appropriate for a group teleconferencing system. Auralization technique is an alternative to implement a realistic teleconference system for group teleconferencing. We expect that sharing a common acoustic environment by incorporating spatial sound will increase the efficiency of meetings by improving the speaker identification and thus enhancing speech intelligibility.

Figure 1 shows a proposed system for realistic teleconferencing. In this system, visual images of participants at distant sites are projected on large screens as three-dimensional pictures. Sound fields in conference rooms are controlled so that the participants have the same auditory events. This system is meant to provide the participants with impression as if they were in the same space. Consequently, intensive studies of methods for producing acoustical effects like a precise sound localization of the realistic impression of a spatial sound image are basically important for this system. This paper describes some techniques for audio teleconferencing which provides a communication with realistic sensation.

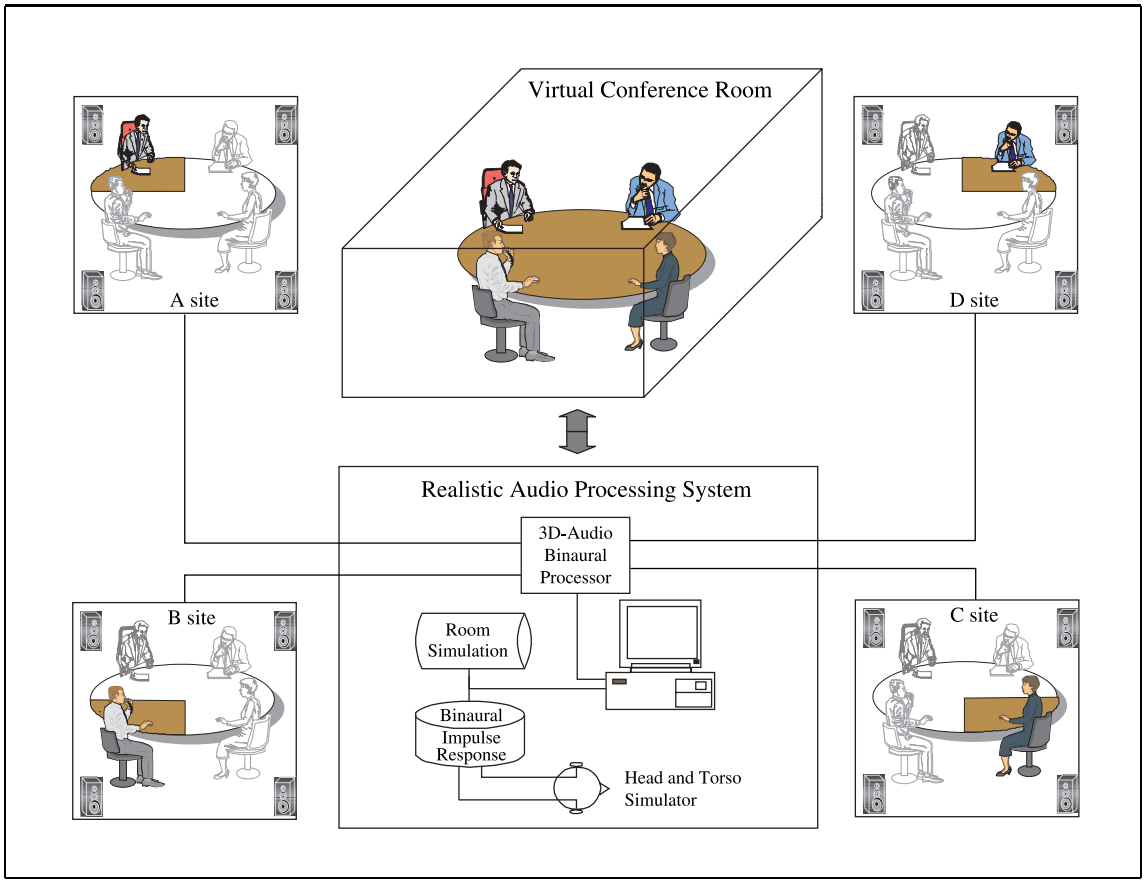

Fig. 1. Model of a proposed realistic teleconference system. The system produces a virtual conference room where participants can hear sounds in a seemingly shared atmosphere.

II. WHY REALISTIC AUDIO DISPLAY?

Normally, participants in meetings have no problem in listening to one particular speaker, even when several persons speak at the same time. This is due to the well-known cocktail party effect. However, if a meeting is recorded and transmitted with monaural technique, this effect is lost, and it is difficult to discriminate what a particular speaker says. But, transmission of spatial information can preserve the cocktail party effect and re-establishes our own selectivity tool. Even if, compared to conventional monaural listening, the total signal to noise ratio and the power spectrum in a communication environment are unchanged, speech intelligibility increases if noise and speech are separately localized.

To exploit the cocktail party effect in

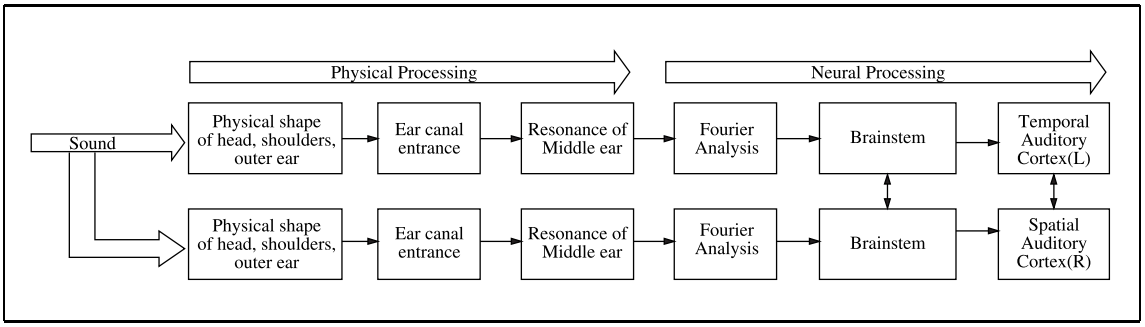

Fig. 2. Model of binaural hearing mechanism.

speech communication, voice signals should be accompanied by spatial information, which enables a reconstruction of the original or synthetic sound field near the listeners' ears. The first attempt to add a 'sense of space' to the stereophonic teleconferencing was made by Damaske who experimented with stereophonic recording, transmission, and reproduction techniques [1]. Stereophonic reproduction in teleconferencing helps a person to detect the positions of participants at an opposite site. This localization helps in identifying the speaker. Stereophonic teleconference is also useful for improving the articulation in a conversation between people located at a distance.

Figure 2 shows a model of the binaural hearing mechanism. When we hear a sound, it arrives at both of our ears and is modulated in various ways by the external auditory apparatus and the body. Binaural localization cues include interaural time delay and interaural intensity differences as a consequence of the acoustical diffraction at the body and the ear. Binaural localization cues are processed at the brainstem level, and then the brain recognizes what we hear and from also the direction. Ando has proven that the processing of spatial and temporal factors in a room is separately performed in the right and left cerebral hemispheres, respectively, as shown in Fig. 2 [4]. Until now, we have been focusing only on the temporal factors in the telecommunication research. Even though the information is a nonverbal signal, however, the spatial factor is very important in communication. Therefore, the achievement of a realistic audio display is very useful for telecommunication because it provides the listener with spatial information for localization and talker identification. Thus the quality of audio teleconferencing can be improved by introducing localization effects, and the efficiency of meetings will be remarkably increased.

III. REPRODUCTION OF REALISTIC AUDIO

As mentioned earlier, an ideal audio teleconference system requires methods for pro-

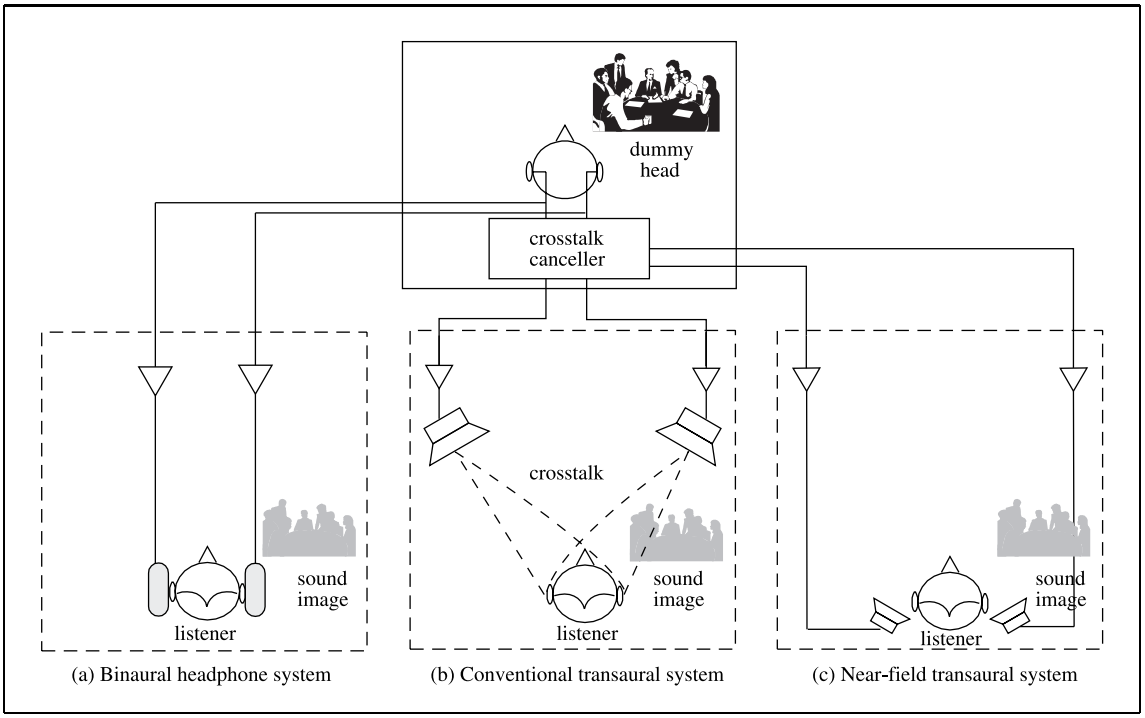

Fig. 3. Binaural and transaural system.

ducing effects to enhance the sensation of sharing a space with participants at distant sites. The most adequate solution to this problem lies applying binaural and transaural techniques. The aim of binaural transmission is to present a listener with the same sound that he/she would hear in the original sound field.

The most simple binaural system uses dummy head recordings to produce binaural signals via a headphone, as shown in Fig. 3(a). This is the easiest way to feed each of the two-channel signals separately into each ear canal of the listener. In this case, no further processing is required except an optimal equalization. The binaural headphone system may be appropriate for person-to-person communication, which will be used for multipoint teleconferences, termed binaural telecommunication [5]. The problems in binaural headphone listening are in-head-localization and front-back confusion of sound images. Another problem in binaural reproduction using headphones is the static situation. When a listener turns his head, the apparent source direction is not changed accordingly. The head tracking system in the 'convolvotron' can give a considerable improvement in realism [6].

The headphone reproduction of binaural signals can be replaced by using loudspeakers. In this case, it is necessary to eliminate

the acoustical crosstalk components from the right loudspeaker to the left ear and vice versa, as shown in the transaural system of Fig. 3(b) [7]. It is also important to realize the individual head-related transfer functions (HRTFs) needed for the acoustical crosstalk cancellation. This processing is easy if the loudspeakers are placed in a small and anechoic chamber. A transaural system will improve localization tremendously in comparison with headphone listening, even with nonindividualized HRTF. This method, however, has several problems when used for teleconferencing. That is, precise sound reproduction can only be achieved in a space where sound is almost completely absorbed and head-movement is very critical. Miyoshi introduced a new transaural system based on the inverse-filtering theorem, which gave the conditions for precisely controlling sound waveforms at multiple points in a conventional room [8].

A more simple method can be considered for binaural reproduction, the so-called near-field transaural system, as shown in Fig. 3(c). The sound localization experiments for two types of a conventional transaural system and a near-field transaural system were conducted. In the conventional transaural system, the two loudspeakers were located in front of the listener at angles of $\pm 60^\circ$, measured from the median plane. In the near-field transaural system, the two loudspeakers were located at a position 10 cm away from the head at angles of $\pm 120^\circ$ measured from the median plane, so that the acoustical crosstalk components were at a minimum. In the teleconference system, it is difficult to consider the effect of HRTFs for every participant. Therefore, the HRTFs were measured using a universal dummy head (B&K Type 4128).

A five second sentence was recorded and used as test material for the sound localization experiment in the anechoic chamber. It is obtained from an utterance of a female person at twelve discrete positions of 30° separation at the median plane. Thirteen subjects (4 male, 9 female) participated in the sound localization experiments. The subject was seated on an adjustable stool such that her/his head was at the center of the arc in the anechoic chamber.

Figure 4 shows the perceived sound directions against the real source directions. The size of circles shows the answer rates for the perceived sound directions in the horizontal plane. From these results, 74 % and 94 % correct responses for the conventional and near-field system, respectively, was obtained. That is, the correct response for the near field transaural system was obtained about 20 % higher than that for the conventional transaural system [9]. After subjective tests, the listeners reported his impression about sound localization.

For 11 of the 13 subjects, the virtual source are judged to have the same spatial localization as sources presented in the free-field. Furthermore, head-movement is not so critical and no in-head localization occurred. Therefore this is expected to yield suitable method for present-

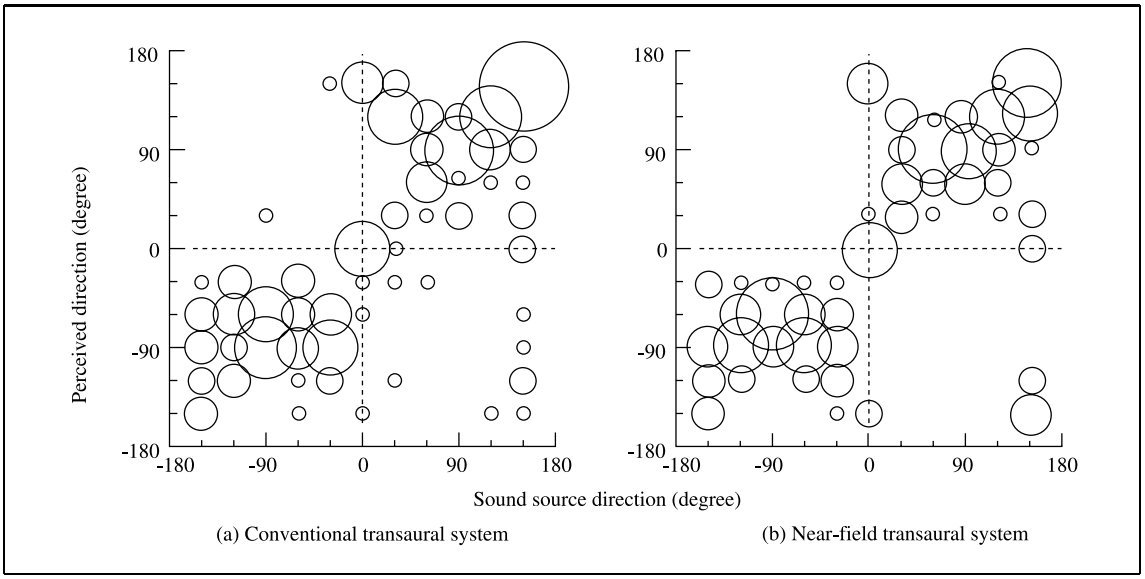

Fig. 4. Sound localization for a conventional and a near-field transaural system.

ing realistic sound to participants in conference rooms whether or not they are reverberant rooms, and also can be used for multipoint teleconferences. In the near-field transaural system, even without insertion of an acoustical crosstalk canceller, the crosstalk components are suppressed by about 17 dB or more because of relative independence of surroundings [9].

IV. REALISTIC AUDIO TELECONFERENCING

Auralization is a term introduced to be used in analogy with visualization to describe rendering imaginary sound fields [10]-[14]. Naylor has described auralization as the 'making audible of the imaginary sound fields', but other authors use the word 'telepresence', or virtual audio reality. Kleiner has defined it as follows: *Auralization is the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space* [11].

Auralization is one of the advanced tools that are becoming available to sound system designers and computer-aided design (CAD) developers. The aim is to auralize music or speech signals with data either from a computer-simulated room impulse response, or from measured impulse responses in an actual room or in a model room. Auralization is important to the audio engineering community. It promises to provide tools for an ade-

quate modeling of room acoustic qualities in an immediate and obvious way. Auralization is also important in the rapid growing field of virtual reality. Auralization is currently one of the state-of-the-art functions in the acoustic simulation world. It allows us to validate the results of an acoustic design with aural assistance instead of the visual assistance.

Auralization technology is an alternative for implementing a realistic teleconference system, which may be appropriate for group teleconferencing. The author's prime objective is to achieve a realistic teleconference system using auralization techniques. Fig. 1 shows the blockdiagram of a teleconference system using auralization software and binaural techniques. In order to achieve teleconferencing in a sense of realism, this system creates a virtual conference room using auralization software. In this case, the synthesized sound field needs not to be an exact reproduction of an existing room or a virtual room. For example, the design of a sound field for video teleconference is quite artificial. In this teleconferencing system, users will share a virtually created sound field. That is, the goal of the realistic teleconference is to distribute a common sound field among many people.

Figure 5 shows the detailed configuration of the model of a two-way audio teleconference system using auralization. A virtual conference room was computer-simulated with the help of a CAD program. Several CAD programs are commercially available, such as electro-acoustic simulator for engineers (EASE) [15], computer aided theater technique (CATT) [16], etc. The transfer functions between participants are calculated in the virtual conference room. Then, this transfer function is multiplied by the left and right HRTFs measured for various incidence angles using a dummy head. By the use of inverse discrete Fourier transformation, the binaural room impulse response (BRIR) is calculated. The BRIRs are convolved with the speech signals from a microphone. The convolution is performed using a programmable finite impulse response (FIR) filter, working at real time. In this experimental system, real-time convolution DSP hardware was utilized [17]. The digital filter can convolve a two-channel impulse response of a maximum length of 2.1 s at 48 kHz sampling frequency in real time. After the signal processing, the manipulated signal will be converted to analog form and transmitted to either headphones or loudspeakers with acoustical crosstalk compensation. One could then enjoy a communication with a distant person as if they were in the same artificial room or environment.

V. CONCLUSION AND DISCUSSION

Our goal is to implement a realistic teleconference system using the transaural and auralization techniques. This paper has described the importance of the realistic audio display in the telecommunication based on the binau-

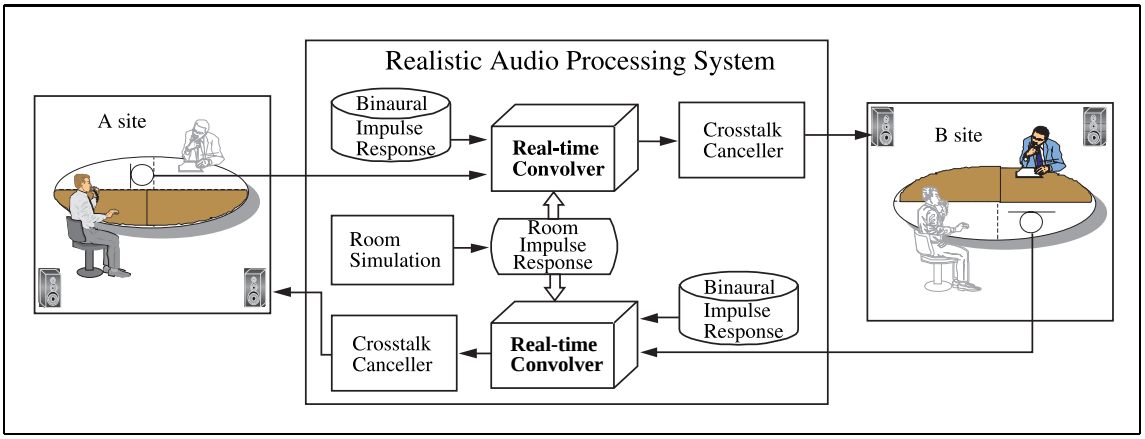

Fig. 5. Configuration of a model for realistic two-way teleconference system.

ral hearing model. This paper has described a near-field reproduction technique for presenting realistic sound in teleconference system. The correct response in the sound localization experiments for the near-field transaural system increased about 20 % higher than that for the conventional transaural system. A realistic teleconference system using auralization technique and near-field reproduction was suggested, which provides the participants with impression as if they were in the same room, thus, the system is useful for the efficiency of meetings.

Unfortunately, these techniques need to be further developed in order to be applied to teleconferencing. The further research areas for realistic sound field control are as follows. 1) Teleconferencing system requires dynamic changes in the head-related-transfer functions. Dynamic auditory environments, where the sound sources and the listener can move around independently, is extremely important for teleconferencing. Therefore, more advanced transaural systems need, for example, an algorithm to maintain natural sound images even when the head of a listener moves. 2) Computation time also becomes important in dynamic environments, because of the real time operation demanded. Real time calculation of a new full-length detailed BRIR for each new head direction or movement is currently not feasible.

ACKNOWLEDGMENT

The authors wish to express their gratitude to the following persons for their invaluable comments of this paper: Prof. Yochi Ando of Kobe University in Japan, Dr. Peter Damaske in Germany, Dr. David McGrath of Lake DSP Ltd. in Australia, Dr. Shigeaki Aoki of NTT Human Interface Laboratories in Japan, and Prof. Mendel Kleiner of Chalmers University

of Techniques in Sweden.

REFERENCES

[1] R. Botros, O. Alim, and P. Damaske, "Stereophonic speech teleconferencing," *ICASSP '86*, Tokyo, pp. 1321-1324, 1986.

[2] S. Aoki, H. Miyata, and K. Sugiyama, "Stereo reproduction with good localization over a wide listening area," *J. Audio Eng, Soc.*, vol. 38, pp. 433-439, 1990.

[3] J. Ohya, "Real-time reconstruction of 3D face images in teleconference with realistic sensation," *Technical Report of IEICE,* HC 92-61, 1993.

[4] Y. Ando, *Concert Hall Acoustics.* New York: Springer-Verlag, 1985.

[5] M. Cohen, N. Koizumi, and S. Aoki, "Design and control of shared conferencing environments for audio telecommunication," *ISMCR '92,* pp. 241-248, 1992.

[6] E. Wenzel, S. H. Foster, and F. Wightman, "Realtime digital synthesis of localized auditory cues over headphones," in Proc. ICASSP, 1989.

[7] P. Damaske, "Head related two channel stereophony with loudspeaker reproduction," *J. Acoust. Soc. Am.*, vol. 50, pp. 1109-1115, 1971.

[8] M. Miyoshi and N. Koizumi, "Transaural system using multiple loudspeakers," in Proc. DAGA '91, Bochum, Germany, pp. 781-784, 1991.

[9] S. Kang, "Acoustic signal processing for sound field communication," *J. Acoust. Soc. Kr.,* vol. 11, pp. 72-82, 1992, in Korean.

[10] M. Klein, P. Svensson and B. Dalenbck, "Auralization: Experiments in acoustical CAD," presented at the 89th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.,* vol. 41, p. 874, 1991.

[11] Mendel Klein, B. Dalenbck, and P. Svensson, "Auralization-An overview," *J. Audio Eng. Soc.,* vol. 41, no. 11, pp. 861-875, 1993.

[12] W. Ahnert and R. Feistel, "Binaural auralization from a sound system simulation program," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc.*, vol. 39, p. 996, 1991.

[13] B. Dalenbck and D. Mcgrath, "Narrowing the gap between virtual reality and auralization," *15th ICA Trondheim,* Norway, pp. 429-432, 1995.

[14] J. Sandvad and D. Hammershoi, "Binaural auralization comparison of FIR and IIR representation of HIRS," presented at the 94th Convention of the Audio Engineering Society, Preprint 3862, 1994.

[15] W. Ahnert, "EARS auralization software," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 894-904, 1993.

[16] CATT, S-41471 Gothenburg, Sweden.

[17] P. S. Single and D. S. MacGrath, "Implementation of a 32768-Tap FIR filter using real-time fast convolution," presented at the 87th Convention of the Audio Engineering Society, Preprint 2830, 1989.

Seong-Hoon Kang received the B.S. degree in electronics engineering from Kwangwoon University in 1981, the M.S. degree in electronics engineering from Yonsei University in 1983, and the Ph.D. degree in acoustical engineering from the Kobe University, Kobe, Japan in 1987. Since 1988, Dr. Kang was with Acoustic Communication Section of Human Interface Department of ETRI. His research areas of interest are acoustic signal processing, acoustic transmission quality evaluation, MPEG audio for HDTV, especially 3-D audio reproduction for realistic telecommunication. He is a board of director of the Acoustical Society of Korea, from which he received the Acoustical Academic Award in 1995. Since 1996 Dr. Kang is a professor of the Departmenment of Broadcast Production and Technolgy in Taejon Medical College, where he teaches and conducts research on acoustical engineering for broadcasting.

Sung-Han Kim received the B.S. degree in computer engineering from Kwangwoon University in 1989, the M.S. degree in 1991, respectively. Since 1991, Mr. Kim has been with Acoustic Communication Section of Human Interface Department in ETRI. His recent research is the MPEG-2 Audio software implementation for HDTV and have many interest in applications of digital signal processing like sound reproduction techniques for realistic telecommunication, MPEG-4 audio and MSDL. He is a member of the Acoustical Society of Korea.