Voice Chat Multi-channel Coding Binaural Signal Processing

Technology Reports

Spatial Audio Transmission Technology for Multi-point Mobile Voice Chat

We have developed a spatial audio transmission technology for comfortable and smooth telecommunication in the mobile environment, which allows users participating in multi-point voice chat to assign a unique spatial position to each of the distant talkers' voice. This enables customization of the listening environment according to the individual user's preferences, and provides an intuitive sound interface for speaker identification as well as less tiring sound for voice chat.

Research Laboratories

Shinya Iizuka

Kei Kikuiri

Nobuhiko Naka

1. Introduction

Recently, communications services allowing multiple simultaneous participants, such as content shares[*1] and online games, have been receiving much attention. For these types of services, a multi-point voice chat function will play an important role in realizing rich communication because it is able to convey a sense of emotion and excitement in real-time.

At the same time, as bandwidth of access networks increases, there is much research and development toward more natural voice communication, providing a sense of presence, while also transmitting wider band speech signals. Intended for use in VoIP services over high-speed mobile packet access connections such as Long Term Evolution (LTE)[*2] and Fourth-Generation (4G) mobile communications, NTT DOCOMO has also developed high-quality speech coding technology able to transmit super-wideband speech (with frequency bandwidth over 10 kHz) at bit-rates from 48 to 64 kbit/s [1].

As one of our initiatives to improve Quality of Experience (QoE) for this sort of mobile voice communication service, we have also developed spatial audio transmission as an extension to the above mentioned high-quality speech coding technology. This provides a comfortable listening environment for conversation among several people, such as with multi-point voice chat.

In contrast to one-to-one conversation, mixed environments with several speakers involve new difficulties, including speaker identification, and following multiple simultaneous topics in the conversation. It is well known that applying spatial information such as direction and/or distance to each speaker's voice signal using binaural signal processing[*3] technology can be effective in reducing these types of difficulty [2].

Conventionally, spatial audio playback has been used mainly for reproducing a real or virtual acoustic space to create presence or share a sound space between participants[3]. On the other hand, the objective of the proposed spatial audio transmission technology is to allow each user to allocate a unique position to the voices of remote partici-

[*1] **Content share**: A service for sharing information such as video or images over a network.

[*2] **LTE**: An evolutional standard of the Third-Generation mobile communication system specified at 3GPP; LTE is synonymous with Super3G proposed by NTT DOCOMO.

[*3] **Binaural signal processing**: A type of signal processing which artificially adjusts the sound heard by each ear to create a spatial audio effect when playing back monaural sound.

pants for improving listening. There are three typical approaches for a voice chat system allowing listeners to determine the position of remote voices according to preference (**Table 1**). The first is client-side processing. Each client directly receives voice data from remote participants, and individually processes the voice data for spatial audio synthesis. Therefore all the functions required to generate spatial sound are implemented in the client, but the amount of data transmitted and the processing load on each client increases with the number of participants. The second is server-side processing. A server renders spatial audio signals from the received participants' voices, and multiplexes them for transmission. This reduces the volume of transmitted data and the amount of processing required in each client, but an additional back channel is required to send control information for generating spatial sound from the clients to the server. The last one is a hybrid approach. The server performs compression and multiplexing, while the client processes spatial audio synthesis. Compression on the server may degrade sound quality compared to other two approaches, but it reduces the volume of transmitted data, and allows distribution of the processing load and spatial processing at the client side.

Our spatial audio transmission technology is based on the hybrid approach, taking the limitations of wireless transmission and client processing capacity in a mobile environment into consideration. This consists of two major developments. One is using multi-channel coding[*4] on the server, which compresses multiple high-quality speech coding signals and generates a single stream at 48 to 96 kbit/s, while reducing sound quality degradation by taking advantage of human auditory characteristics. The other is spatial audio decoding, which reduces the complexity of client processing through efficient integration of decoding and spatial audio synthesis.

Practical implementation of this technology enables provision of smooth, multi-point voice chat communications with voices that are easy to distinguish intuitively in mobile environments.

This article describes the development of this spatial audio transmission technology, the results of a sound-quality evaluation and development of a mobile VoIP multi-point voice chat prototype using the technology.

Table 1 Comparison of chat systems with spatial audio playback

	Client-side processing approach	Server-side processing approach	Hybrid processing approach
Server	Not required	Required	Required
Back channel	Not required	Required	Not required
Downlink transmission data (transmission volume)	Encoded data from all participants (increases with number of participants)	Spatial audio synthesized data (fixed)	Multiplexed data (fixed)
Terminal processing (processing load)	Decoding and spatial audio synthesis of each data stream (increases with number of participants)	Decoding of spatial audio synthesis data (fixed)	Decoding and spatial audio synthesis of multiplexed data (fixed)

2. Spatial Audio Transmission Technology

2.1 Architecture of Spatial Audio Transmission Technology

The spatial audio transmission technology is composed of processes for speech encoding, multi-channel coding and the spatial audio decoding. The client performs the speech encoding and the spatial audio decoding, and the server performs the multi-channel coding (**Figure 1**).

A high-quality speech coding algorithm developed by NTT DOCOMO is used for the speech encoding. This transforms the input time-domain speech signal to its frequency-domain coefficients using a Modified Discrete Cosine Transform (MDCT)[*5] and quantizes each coefficient according to auditory significance. The method is able to encode a super-wideband speech signal with low latency of several tens of milliseconds and processing load comparable to conventional speech encoding methods.

The multi-channel coding process decodes the high-quality speech-coded stream from each client, determines the most important components by comparing frequency-domain coefficients, and

[*4] **Multi-channel coding**: A form of signal processing which takes input signals from multiple systems, and performs multiplexing and data compression onto a single system.

[*5] **MDCT**: A method for converting a time-series signal to its frequency components. It is able to reduce distortion at block boundaries without losing information by applying an overlapping transform with the preceding and following blocks, so it is widely used for audio encoding.

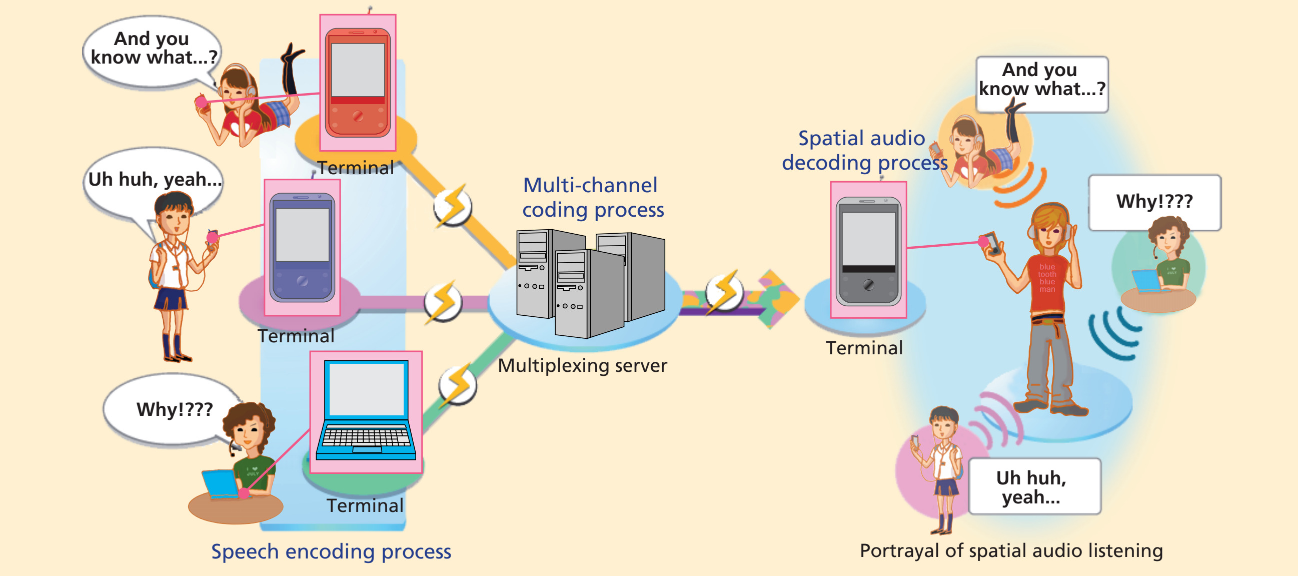

Figure 1 Architecture of spatial audio transmission technology

compresses and multiplexes them to create a single, compressed and encoded data stream (**Figure 2**).

The spatial audio decoding process receives the compressed and multiplexed encoded data from the multi-channel coding process, separates out and decodes the frequency-domain components of each participant's voice, and performs spatial audio synthesis.

Figure 3 shows the mechanism by which humans recognize the location of a sound source. Sound generated by a sound source propagates to both ears through different paths. The direction from which it arrives is recognized based on the Inter-aural Intensity Difference (IID) and the Inter-aural Time Difference (ITD), both resulting from the difference in distances from each ear to the sound source. Thus, if signal processing is used to simulate IID and ITD for a monaural sound signal and the resulting signals are presented separately to the left ear and the right ear using headphones, the listener perceives the sound signal with a spatial audio effect.

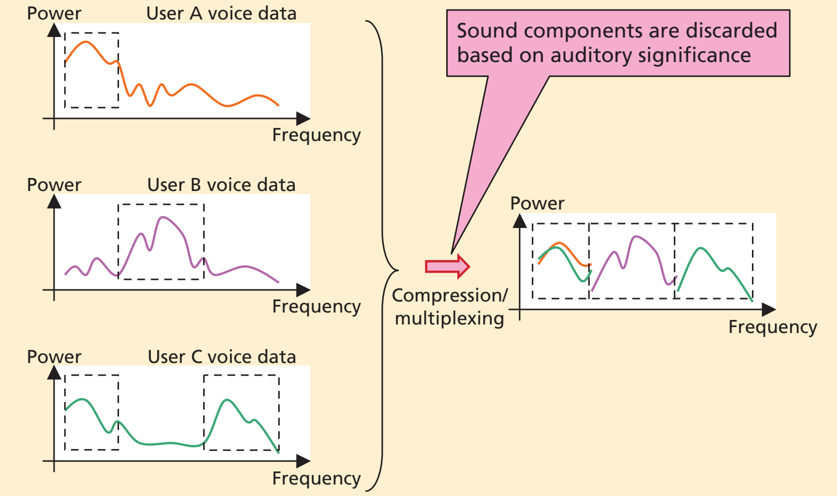

Figure 2 Multiplex processing for multi-channel coding

Conventionally, spatial audio synthesis processing is applied to the decoded time-domain sound signal, but we developed a method of spatial audio synthesis operating directly on the frequency domain coefficients (i.e., decoded MDCT coefficients) while decoding the encoded data for this technology (**Figure 4**). By combining the process of decoding and spatial audio synthesis processing, we achieved to reduce the processing required for spatial audio playback by approximately 30% to 50% relative to conventional methods.

2.2 Verification of Sound Quality

To verify the quality of sound transmitted by the spatial audio transmission technology, we conducted subjective evaluation tests. Conditions for the test are shown in **Table 2**. We used the Multi-Stimulus test with Hidden Reference and Anchor (MUSHRA) method [4], which evaluates test stimuli (including the original sound) on a range from 0 to 100 points.

Figure 5 shows the test results. The error bar in the figure shows a 95% confidence interval[*6] for the averaged scores. Conversation A contains momentary instances of simultaneous utterances, while conversation B contains continuous periods of two or more participants speaking. Results of conversation A at 64 kbit/s and conversation B at 96 kbit/s show that our technology achieves equivalent quality to that using multiple high-quality encoded sound signals encoded at 64 kbit/s per channel. In other words, the spatial audio transmission technology offers a 20% to 25% reduction in downlink data transmission for each of the conversations through the multi-channel coding.

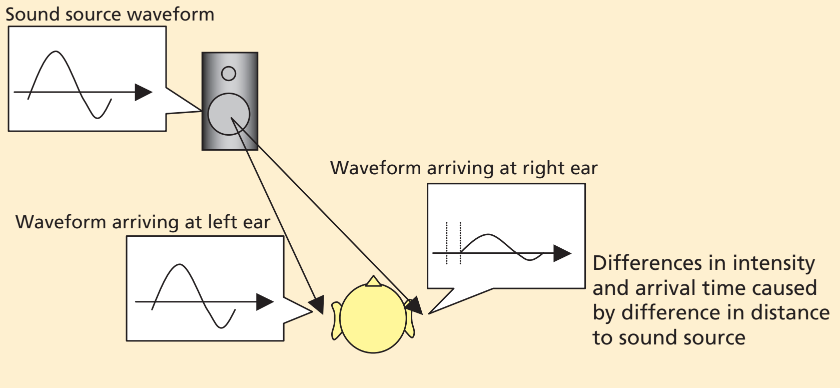
Sound source waveform
Waveform arriving at right ear
Waveform arriving at left ear
Differences in intensity and arrival time caused by difference in distance to sound source

Figure 3 Spatial sound recognition mechanism

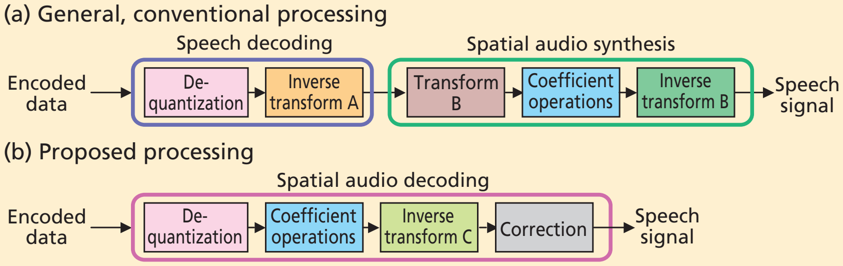

Figure 4 Spatial audio decoding architecture

Table 2 Subjective evaluation test conditions

Methodology	MUSHRA
Number of subjects	10
Test items	Conversation A (five participants, few concurrent utterances) Conversation B (six participants, many concurrent utterances)
Reference sound (sampling frequency)	Binaural playback with sources reconstructed separately (22.05 kHz)
Encoded sound (bit-rate/ sampling frequency)	Binaural playback with spatial audio transmission (48, 64, 96 kbit/s / 22.05 kHz)
Uncompressed, multiplex encoded sound	High-quality encoded sound (64 kbit/s / 22.05 kHz), spatial audio synthesis with separately reconstructed sources.
Band-limited sound	7 kHz bandwidth, 3.5 kHz bandwidth
Listening method	Headphones (both ears)

3. Prototype

This technology was implemented in a VoIP-based, multi-point voice chat system using the Session Initiation Protocol (SIP)[*7]. The server and client functions were implemented as Windows®[*8] and Windows Mobile®[*9] applications respectively. We confirmed execution of the client software on FOMA PRO Series HT-01A termi-

[*6] **95% confidence interval**: Assuming the sample has a particular distribution, an interval containing 95% of the sample.

[*7] **SIP**: A call control protocol defined by the Internet Engineering Task Force (IETF) and used for IP-phone with VoIP, etc.

[*8] **Windows®**: A trademark or registered trademark of Microsoft Corp. in the United States and other countries.

[*9] **Windows Mobile®**: A trademark or registered trademark of Microsoft Corp. in the United States and other Countries.

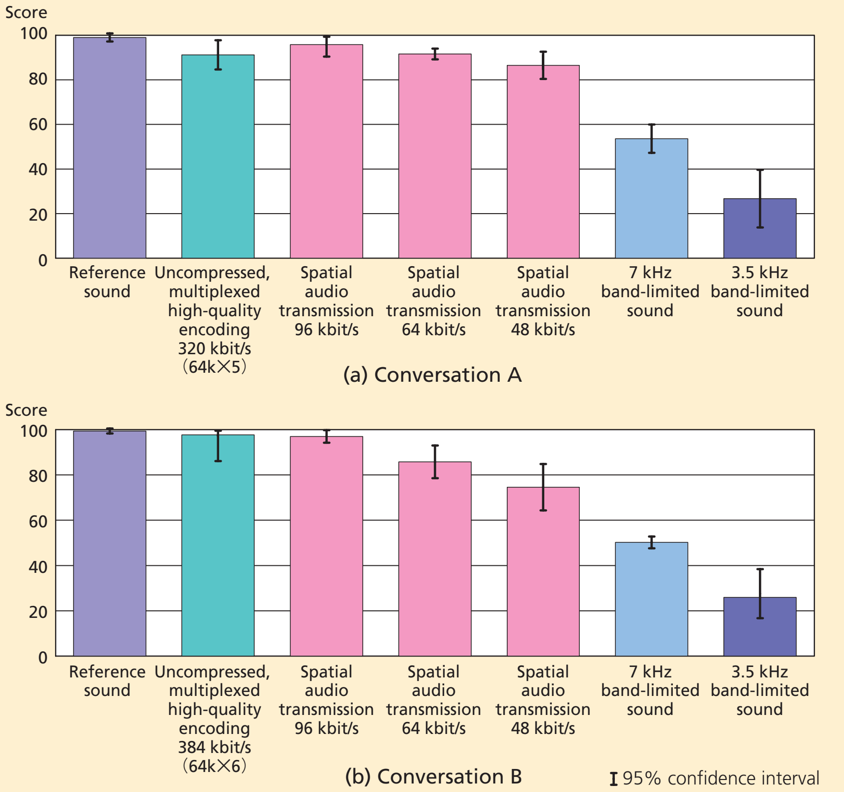

Figure 5 Subjective evaluation test results

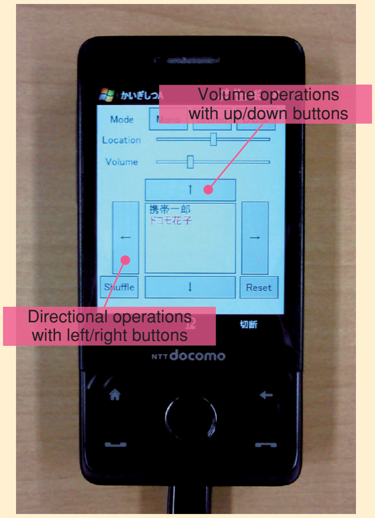
Volume operations with up/down buttons
Directional operations with left/right buttons
Mode
Location
Volume
Shuffle
Reset
携帯一郎
ドコモ花子
切断
NTT docomo

Photo 1 Prototype software display example

nals (**Photo 1**). Clients participate in a voice chat session by placing calls to meeting rooms configured on the server. The client screen shows a participant list and whether participants are speaking, and after selecting a participant, the left and right buttons can be used to adjust the speaker position while the up and down buttons adjust the volume.

4. Conclusion

In this article, we have described a spatial audio transmission technology used in a multi-point voice chat application for mobile environments. The technology provides spatial audio synthesis that allows participants to adjust the positions of the other participants' voices according to their preferences, and was developed to provide comfortable, smooth, multi-user voice communication. The subjective listening test results indicated that the proposed multi-channel coding method reduced transmitted data volume by 20% to 25%, while maintaining sound quality. We also described a prototype of this technology, implemented in the form of a VoIP-based, multi-point voice-chat system.

In addition to improving the experience of voice-chat participants, the spatial audio transmission technology, which allows the direction and volume of individual participants' voices to be adjusted, is promising for applications attempting to improve a sense of shared space or presence. In the future we plan to continue study of improvements to the technology's binaural signal processing, such as personalizing spatial audio effects to user preferences.

References

[1] K. Kikuiri et. al: "High-quality Speech Coding," NTT DoCoMo Technical Journal, Vol.9, No.2, pp.38-41, Sep. 2007.

[2] R. Drullman and A. W. Bronkhorst: "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," J. Acoust. Soc. Am., 107, pp.2224-2235, 2000.

[3] Y. Yasuda et. al: "Reality Speech/Audio Communications Technologies," NTT DoCoMo Technical Journal, Vol.5, No.1, pp.61-69, Jun. 2003.

[4] ITU-R Recommendation BS.1534-1: "Method for the subjective assessment of intermediate quality level of coding systems," 2003.