Speech and Audio Signal Processing for a User-friendly, Comfortable, Convenient Telecommunication Environment

Satoshi Takahashi[†], *Manabu Okamoto, and Hisashi Ohara*

Abstract

Telecommunication will be more effective if users can talk to each other in a free, comfortable, and natural style. It will also be more convenient if they can easily obtain information in a timely manner anytime and anywhere (e.g., at home, while working in an office, or while driving a car). NTT Cyber Space Laboratories is developing technologies to achieve a natural and comfortable next-generation voice communication environment. We also aim to create a man-machine interface that will let people access information and other people in an easy and natural manner. This article introduces some of our recent developments in speech and audio signal processing.

1. Diversification of speech communication environment

Human-to-human communication has been expanded by the popularity of the Internet and mobile devices. E-mail is the most representative example of the recent trend, with the videophone being another example. Advances in broadband IP (Internet protocol) networks now let us enjoy video telecommunication in our own home using personal computers. The personal computer lets us use many input/output devices such as keyboard, display, mouse, microphone, speaker, and touch panel. Multimodal telecommunication using those kinds of devices will make telecommunication more efficient and more enjoyable than that using the conventional fixed-line telephone, which provides just a single voice telecommunication channel. The broadband IP networks will lead to an advanced style of telecommunication via several media, such as still and moving images, sound, speech, and text. Furthermore, they can raise the quality of sound and images for telecommunication, enabling realistic telecommunication services that were not developed for narrowband networks. High-reality multimodal telecommunication is expected to become the most popular communication style in the near future.

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: takahashi.satoshi@lab.ntt.co.jp

2. Direction of media processing technologies

The Media Processing Project at NTT Cyber Space Laboratories has made significant advances in speech and audio signal processing and natural language processing. Considering the recent rapid growth of the broadband IP networks and globalization of the telecommunication network, the Media Processing Project is aiming at new technologies for a high-reality next-generation voice communication environment, which will require innovative techniques in speech and audio signal processing.

We have also been trying to create man-machine interface technologies that will let users access information and other people in an easy and natural manner. Speech recognition, speech synthesis, and natural language processing techniques will improve the man-machine interface and enhance the multimodality of telecommunication.

Figure 1 illustrates a high-reality videoconferencing service as an example of a next-generation telecommunication service. Participants are displayed on high-resolution screens with excellent color fidelity and 3D surround sound, giving them the feeling that they are sitting at the same table. The system transmits high-quality sound and voice while suppressing background noise, which would otherwise reduce the sense of reality. A scalable speech

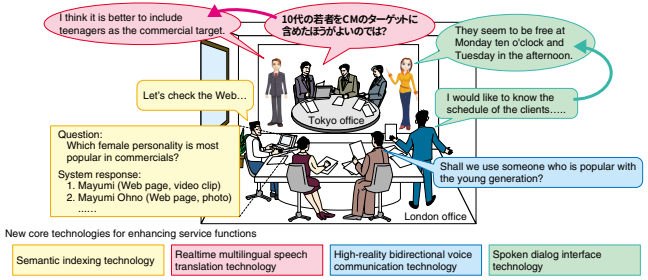

Fig. 1 Concept of future service using media processing technologies (speech processing, audio signal processing, and natural language processing).

coding technique designed for the best-effort network ensures that the transmitted sound is never interrupted, even when the network is congested. Intelligent interface agents appear on the screen to help the participants retrieve information. Participants can talk to the agent from any position without being conscious of the location of the microphone. The interface agents speak with natural-sounding voices and accomplish the information retrieval tasks set by users. A speech translation function permits communication among participants speaking different languages. This kind of network service will be widely used on broadband IP networks.

3. Restructuring the core technologies

To accelerate the development of the advanced functions of the network service mentioned above, several basic technologies were integrated to yield new core technologies. Namely, five conventional basic technologies—speech recognition, speech synthesis, natural language processing, speech coding, and audio signal processing—were integrated to generate four new core technologies.

1) Semantic indexing
Speech/text contents on the Internet will be recognized, transformed, and structured to generate a huge knowledge source for information retrieval that will satisfy the questions and demands of the users.
2) Realtime multilingual speech translation
Speech/text will be translated to enable real-time communication between people speaking different languages.
3) High-reality bidirectional voice communication
A high-reality sound environment will be reproduced as if the person at the remote site were right in front of the user.
4) Spoken dialogue interface for cyber attendants
Intelligent interface agents using multimodal communication will establish a natural and user-friendly man-machine interface.

The other feature articles in this issue introduce new techniques selected from among those developed for "high-reality bidirectional voice communication" and "spoken dialogue interface for cyber attendants".

4. High-reality bidirectional voice communication technology reproducing comfortable sound environment

High-reality bidirectional voice communication technology consists of two basic technologies: audio signal processing and speech/audio coding. Audio signal processing includes sound-pickup, sound-reproduction, and echo-cancellation. The next article introduces a new scalable speech coding technique, which is based on a flexible design policy that can control the quality of reproduced sound according to the conditions of the network and the user's environment while maintaining interconnectivity among var-

ious types of lines. In the third article, a new echo-canceling technique is introduced. This suppresses echoes as well as ambient noise to make the talker's speech clearer and easy to understand at the other site.

5. Spoken dialogue interface technology providing user-friendly man-machine interface

Spoken dialogue interface technology for the cyber attendant system consists of wide variety of basic technologies such as speech recognition, speech synthesis, natural language processing (dialog control and indexing), and audio signal processing (sound-pickup and echo-cancellation). The fourth article introduces an advanced spontaneous speech recognition technique, which can recognize spontaneous speech and extract key words. The fifth article describes a new text-to-speech synthesis technique using a corpus-based approach. This technique generates high-quality synthetic speech that is as natural as human speech from any text.

6. Conclusion

This article has introduced the research targets of the Media Processing Project in NTT Cyber Space Laboratories. We are convinced that the fusion of media processing technologies will lead to next-generation telecommunication and broadband IP networks will lead to media processing technologies being widely used to enrich human-to-human communication.

Satoshi Takahashi
Senior Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Waseda University, Tokyo in 1987, 1989, and 2002, respectively. Since joining NTT in 1989, he has been researching speech recognition and pattern recognition. He is a member of the Acoustical Society of Japan (ASJ) and the Institute of Electronics, Information and Communication Engineers (IEICE).

Manabu Okamoto
Senior Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He received a masters degree in design from Kyushu Institute of Design, Fukuoka in 1991. In 1991 he joined NTT Electrical Communication Laboratories. Since then, he has been researching the acoustic design of various kinds of teleconference systems. He is a member of ASJ and IEICE.

Hisashi Ohara
Executive Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.
He received the M.S. degree in electrical engineering from Keio University, Yokohama in 1979. Since joining NTT Laboratories in 1979, he has been researching natural language processing.He is a member of the Information Processing Society of Japan and IEICE.