Scalable Speech Coding Technology for High-Quality Ubiquitous Communications

Yusuke Hiwasaki[†], *Takeshi Mori, Hitoshi Ohmuro, Jotaro Ikedo, Daisuke Tokumoto, and Akitoshi Kataoka*

Abstract

The rapid growth of broadband connections is providing an environment conducive to high-quality voice communications. Nevertheless, to achieve sophisticated voice services to replace conventional telephone calls, a variety of bandwidth capacities and sound-reproduction environments must be supported. Here, we introduce a scalable speech coding technique that enables flexible sound quality reproduction. Utilizing the hierarchical nature of scalable coding, we created a flexible-design policy that has a high affinity with network control technology and enables us to design and build applications that can be used in various scenarios.

1. Introduction

Scalable (embedded) coding is a core technology for achieving voice communications having superior quality and usability. Compared with conventional telephone-based communications capable of only narrowband monaural transmission, this technology basically allows us to transmit wideband, multi-channel "sound environments" over broadband links. It is based on a hierarchical coding policy that can control the quality of reproduced sound according to the user's environment. Scalable speech coding is a promising tool for creating next-generation ubiquitous communication environments.

2. The coming of high-quality VoIP services

Traditionally, telephone calls were transmitted on bandwidth-limited analog lines as monaural speech cut off at 3.4 kHz. The digitization of switches and transmission circuits, which began in the 1980s, simply replaced the analog data with digital data, so the speech quality did not change, for compatibility reasons. This quality specification was inherited even for new voice communication services such as voice over Internet protocol (VoIP). In addition, the use of compression coding over extremely band-limited transmission lines cannot avoid degrading speech quality.

At the same time, the use of optical fiber and broadband IP networks typified by ADSL (asynchronous digital subscriber line) has accelerated and, under the right conditions, data can be transmitted at a much higher rate. Furthermore, for realtime bidirectional voice communications as in telephone conferencing, the demand has grown for high-fidelity reproduction of remote acoustic environments. In short, there is a need for high-quality voice-transmission systems to replace the conventional 3.4-kHz monaural systems. **Figure 1** gives an overview of trends in speech quality for fixed-line telephones and the outlook for the future. Compared with conventional telephone communications, there are several requirements that must be fulfilled to achieve high-fidelity voice communications:

- Wideband frequency support
- Multi-channel capability
- Low distortion

Firstly, "wideband frequency support" means supporting AM-radio-level frequency bandwidth up to 7 kHz or CD-level bandwidth up to 20 kHz, in addition to the conventional telephone band up to 3.4 kHz. Secondly, "multi-channel capability" means support-

† NTT Cyber Space Laboratories
Musashino-shi, 180-8585 Japan
E-mail: hiwasaki.yusuke@lab.ntt.co.jp

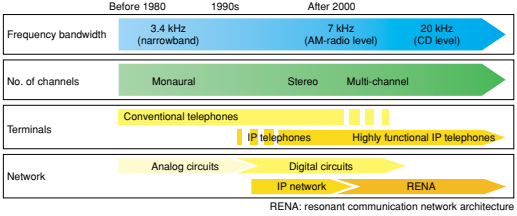

Fig. 1. Likely trends of speech quality in fixed-line telephony.

ing multiple channels including stereo transmission (two channels) as opposed to the conventional monaural (one channel) systems. Finally, "low distortion" means improving the signal-to-noise ratio (SNR) so that it approaches that of pulse code modulation, which has not been possible using conventional compression coding schemes.

3. Issues in high-quality voice communications

Although the IP network has grown dramatically and is expected to grow further, it is based on a "best-effort" design. This means that the data transfer rate, packet delay, packet-loss rate, and other factors that affect the quality of data transmission are bound to fluctuate. If quality-of-service (QoS) control technology is not used in applications such as realtime, bidirectional voice communications, then congestion at even one relaying node along the transferring path will delay packets, preventing voice data from arriving at the expected time. In short, it is impossible to avoid interruptions in the reproduced sound under such conditions.

In addition, a wide range of different sound-reproduction devices, such as telephones, personal computers, and specialized terminals are now being used. Each device differs in terms of the frequency bandwidth and number of channels that it supports. The access network has also diversified into a number of systems such as optical fiber, ADSL, and wireless LANs (local area networks), and the maximum transmission rate varies in each case. It has therefore become necessary to ensure interconnectivity between users in different environments.

In conventional coding schemes, the frequency bandwidth, bit-rate, and number of channels are all fixed. This means that different coding schemes must be applied to cope with the variety of transmission rates when generating encoded data (the bit-stream). Furthermore, to enable communications between terminals that do not support other coding schemes, transcoding technology that converts the bit-stream format is necessary. The disadvantage of transcoding, however, is that it involves considerable computation complexity, placing a load on gateways. Moreover, voice signals are usually encoded using lossy compression methods, and there is no way to compensate for the degradation in speech quality when decoding and re-encoding are performed in the transcoding process.

4. Scalable speech coding function and its advantages

Figure 2 illustrates the difference between conventional coding and our scalable coding, and **Fig. 3** shows the data configuration of the scalable-speech-coding scheme developed in our research group.

In conventional coding, the encoder generates only one type of bit-stream at a fixed rate, and the reproduced speech has a fixed frequency bandwidth and a fixed number of channels. On the other hand, the scalable encoder generates a bit-stream that can be decoded using either all or part of the bit-stream, offering a versatile choice of granularity for the bit-rate and quality of the reproduced output [1], [2]. This can be achieved by using a multi-layered bit-stream with enhancement data on top of the core data. For example, while using the entire bit-stream enables the reproduction of CD-level wideband stereo sound, using only portions of the bit-stream can provide AM-radio or telephone-level quality.

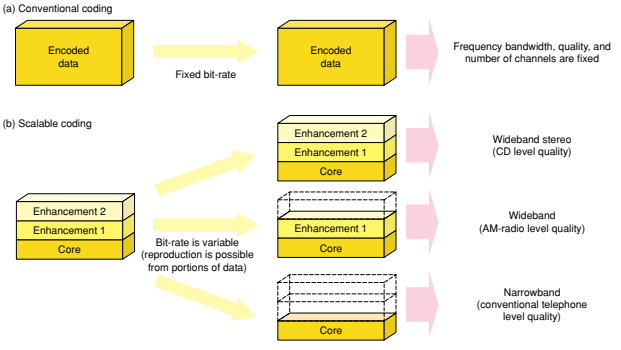

Fig. 2. Comparison of conventional coding and scalable coding.

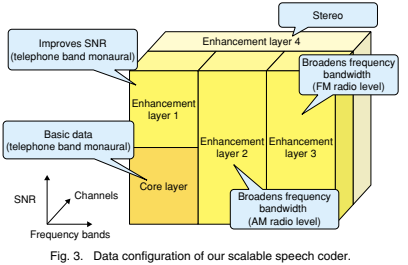

Fig. 3. Data configuration of our scalable speech coder.

The scalable coding scheme developed in our research group consists of multiple layers consisting of a core layer and enhancement layers extending the SNR, bandwidth, and number of channels. Here, the core layer is a bit-stream for narrowband speech of basic quality, and the enhancement layer 1 raises the SNR to PCM quality. Enhancement layer 2 corresponds to data extending speech up to the AM band (4–7 kHz), and enhancement layer 3 corresponds to an even broader frequency bandwidth (FM band: 8–15 kHz). Finally, layer-4 data includes stereo-signal information.

One advantage of this configuration is that bit-rate can be adjusted by transmitting only some of the layers (which must always include the core layer) when the transmission paths cannot provide sufficient throughput for any reason. In other words, the above configuration can support various access networks. In particular, it can easily adapt to best-effort IP networks in which throughput varies in real time.

Another advantage of using this configuration is that a bit-stream can be constructed in accordance with sender or receiver conditions. For example, for a portable terminal such as a PDA (personal digital

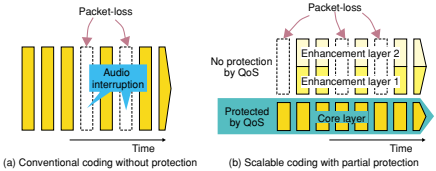

Fig. 4. Application of QoS-controlled packet protection to scalable coding.

assistant) that has limited computational ability (CPU power, memory, etc.), the overall processing power is reduced by limiting the number of layers to be encoded or decoded. In addition, if the original input speech signal was monaural and narrowband, it would be wasteful to construct all the layers, which means that the bit-rate and required processing power can be reduced.

In this way, scalable coding can provide flexible services for various scenarios, leading the way to a wide range of application fields. The scalable coding configuration presented here allows transmission of only part of the bit-stream according to the given data transfer rate and the sound capabilities of the sending and receiving terminals. It also assures the interconnectivity of voice communications between various types of lines without having to use transcoding.

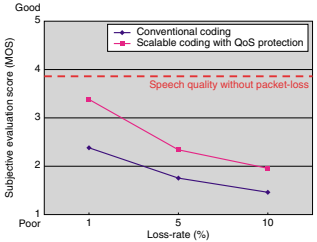

Fig. 5. Speech quality comparison under identical packet-loss conditions.

5. Application to QoS-controlled IP networks

In recent years, various QoS technologies have been developed to eliminate fluctuations in transmission quality in IP networks. While these technologies can generally be categorized into "bandwidth guaranteed" and "priority transfer" types, they both provide a mechanism for achieving a fixed level of transmission quality by protecting high-priority packets during times of congestion in the network. Since scalable speech coding can provide flexibility in design, it has a very high affinity with such network functions compared with the conventional speech coding schemes.

Figure 4 shows an example of applying QoS control to scalable coding. In this example, only the core layer is protected and the other enhancement layers are sent on a normal-priority basis. In a conventional coding scheme, losing packets results in interruptions in the reproduced speech, which can easily be noticed. However, in this example, only packets representing enhanced layers are lost, so the packet loss is less noticeable. **Figure 5** shows simulation results comparing the speech quality of this method against the conventional methods for identical packet loss rates (1%, 5%, 10%). As can be seen, speech quality is improved by using scalable coding in all three cases.

In addition to the above example, we can consider another method of reducing the average throughput of transmitted data. By detecting the state of each speech segment, core layer data can be omitted from the protected packets for non-speech segments. Conversely, we can consider a method that improves the minimum guaranteed speech quality by protecting certain enhancement layers in addition to the core layer. Similarly, many applications for controlling the quality of speech can be considered.

6. Conclusions and future outlook

We presented the basic target of next-generation high-quality voice communications, gave an

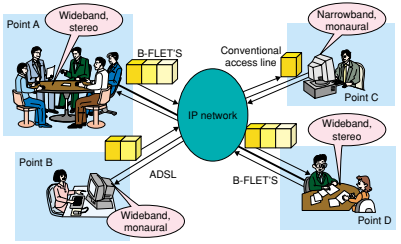

Fig. 6. Service image of future multipoint telephone conferencing over IP networks.

overview of our scalable speech coder, and described a possible application to QoS-controlled IP networks. We plan to continue our studies on various techniques for applying scalable speech coding to next-generation networks and achieving high-quality VoIP services with no audio interruptions. To conclude this article, **Fig. 6** schematically shows the concept of a high-quality VoIP service using scalable speech-coding technology. In this service, two sets of B-FLET'S users are enjoying high-quality communications by AM-radio-level stereo speech, and users having relatively low-speed connections such as ADSL and ISDN (integrated services digital network) are communicating at a level of speech quality appropriate to the wire-speed. This multipoint audio conference connects remote points using several wire speeds, enabling each point to communicate at an optimal level of quality.

References

[1] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," IEEE Transactions on Information theory, Vol. 37, No. 2, pp. 269-275, 1991.

[2] A. Jin, T. Moriya, N. Iwakami, and S. Miki, "Scalable audio coding based on hierarchical transform coding modules," Trans. IEICE, Vol. J83-A, No. 3, pp. 241-252, 2000.

Yusuke Hiwasaki

Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E. degree in instrumentation engineering and M.E. degree in computer science from Keio University, Yokohama in 1993 and 1995, respectively. Since joining NTT Human Interface Laboratories, Tokyo, Japan in 1995, he has been engaged in research on low-bit-rate speech coding. From 2001 to 2002, he was a guest researcher at KTH (Royal Institute of Technology) in Stockholm, Sweden. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), and the Acoustical Society of Japan (ASJ).

Takeshi Mori

Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in computer science from Tokyo Institute of Technology, Tokyo in 1994 and 1996, respectively. Since joining NTT Laboratories in 1996, he has been engaged in research on speech and audio coding. He is a member of IEEE, IEICE, and ASJ.

Hitoshi Ohmuro

Senior Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Nagoya in 1988 and 1990, respectively. In 1990, he joined NTT Human Interface Laboratories, Tokyo, Japan. He is researching highly efficient speech coding and developing VoIP applications. He is a member of IEEE, IEICE, and ASJ.

Jotaro Ikedo

Senior Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E. degree in electronic engineering and M.E. degree in electrical engineering from Kogakuin University, Tokyo in 1989 and 1991, respectively. Since joining NTT in 1991, he has been engaged in R&D of low-bit-rate speech coding and wireless transmission. He contributed to establishing the ARIB STD-27 and ITU-T G.729 standards. He is now developing VoIP systems. He is a member of IEEE, IEICE, and ASJ.

Daisuke Tokumoto

Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E. and M.E. in mechano-infomatics and systems from Nagoya University, Nagoya in 1998 and 2000, respectively. Since joining NTT Cyber Space Laboratories, Tokyo, Japan in 2000, he has been researching speech signal processing. He is a member of IEICE.

Akitoshi Kataoka

Manager, Acoustic Information Group, Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Doshisha University, Kyoto in 1984, 1986, and 1999 respectively. Since joining NTT Laboratories in 1986, he has been engaged in research on noise reduction, acoustic arrays, speech-signal processing, and medium-bit-rate speech and wideband coding algorithms for ITU-T standards. He is a member of ASJ, IEICE, and IEEE. He received the Technology Development Award from ASJ in 1996 and the Prize of the Commissioner of the Japan Patent Office from the Japan Institute of Invention and Innovation in 2003.