Voice Processing Technology for Realistic-quality Voice and Rich Telecommunication Services

Hitoshi Ohmuro†*, Manabu Okamoto, Shoichiro Saito, Satoru Emura, and Sumitaka Sakauchi*

Abstract

This article describes the speech coding, acoustic design, and signal processing technologies for the high-quality telephone service that NTT provides over the Next Generation Network (NGN). Since the conventional telephone system uses a limited frequency range less than that of the human voice, voice individuality and clarity are sometimes insufficient. On the other hand, the high-quality NGN telephone service enables realistic-quality voice communication. A stereo conference phone under development for future services is also described.

1. Realistic-quality voice

In 2008, NTT launched a new IP (Internet protocol) telephone service called HIKARI DENWA[*1] [1] over FLET'S HIKARI NEXT [2], its optical fiber subscriber line service. This is a telephone service based on voice-over-IP (VoIP) technology, and it optionally provides multimedia communications that enables a user to make calls in the same way as normal telephone calls. One of the optional services is a high-quality telephone service. Conventional telephones use only the human voice components in the frequency range from 300 Hz to 3.4 kHz. This limitation sometimes leads to insufficient individuality and similar words may be misheard. High-quality telephones use components in the range from 50 Hz to 7 kHz: twice the width of conventional telephones. This enables telecommunication with realistic-quality voice.

2. Speech coding technology

To enable speech signals to be transmitted over digital or packet networks, speech coding technology is used. The encoder and decoder together are called a codec. The legacy analog telephone system uses ITU-T G.711 codecs at subscriber switchboards (ITU-T: International Telecommunication Union, Telecommunication Standardization Sector). By contrast, in the HIKARI DENWA system, the codec is implemented in each terminal, i.e., an IP phone terminal or home gateway. When a user makes a normal telephone call over HIKARI DENWA, the G.711 codec is used, as in the conventional telephone service; when a user makes a high-quality telephone call, a wideband codec is used. The codec to be used for a telephone call is determined automatically by negotiation between the calling and receiving terminals by means of session initiation protocol (SIP) signaling. One well-known wideband codec is ITU-T G.722, which has been used for video and audio conferencing systems over ISDN (integrated services digital network) and IP networks and has recently begun to be used for IP phones.

Even when a wideband codec is used for a telephone call, the same functions as in a conventional

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan

[*1] Hikari is the Japanese word for light; denwa is the Japanese word for telephone.

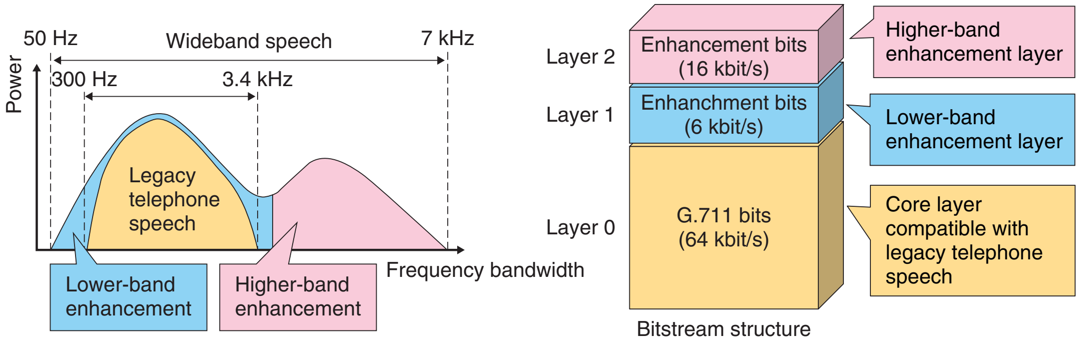

Fig. 1. Scalable coding.

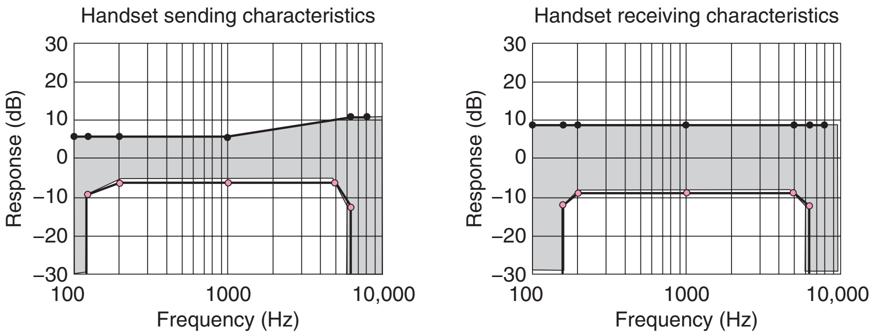

Fig. 2. Transmission characteristics for wideband telephone in CIAJ standards (CES-Q004).

telephone service, e.g., call waiting and call transfer, should be provided. However, codec negotiation is often problematic in this case. Codec negotiation was not considered before because there was only one codec: G.711. However, when G.711 and a wideband codec are used together, the codec should be switched appropriately.

To solve this problem, NTT Cyber Space Laboratories developed a scalable codec that extends the conventional G.711 and standardized it as ITU-T G.711.1. As shown in **Fig. 1**, this scalable codec has a multilayer structure consisting of a G.711 layer, lower-band enhancement layer, and higher-band enhancement layer. The G.711 layer is compatible with the conventional G.711 codec. This structure can be understood in terms of building blocks. G.711 and G.711.1 are easily converted to each other by extracting some blocks or by adding required blocks. Therefore, the use of G.711.1 enables wideband speech to be easily applied for conventional call functions, such as call waiting, call transfer, and multipoint teleconferencing [3].

3. Design strategy for high-quality handset

A handset is the most basic acoustic device for a telephone. A high-quality telephone cannot be made just by implementing wideband speech codec software. A normal handset is designed for narrowband (from 300 Hz to 3.4 kHz) speech, and it cannot pick up or reproduce over-3.4-kHz speech. We need a handset design for wideband speech and a standard for its characteristics.

The Communication and Information Network Association of Japan (CIAJ) standardized transmission characteristics for wideband (from 150 Hz to 7 kHz) digital handset telephones as CES-Q004 in 2007. CES-Q004 requires a wideband telephone to be able to send speech with a frequency range from 125 Hz to 6.3 kHz and receive speech ranging from 160 Hz to 6.3 kHz. The characteristics required for the handset are shown in **Fig. 2**.

We can use acoustic design knowledge for audio and broadcasting to design hands-free devices for communication using over-3.4-kHz speech. However,

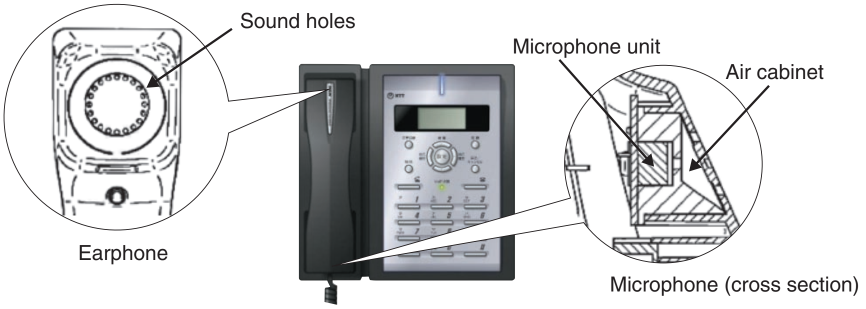

Fig. 3. High-quality telephone for NGN trial (NP-1).

this knowledge cannot be used for handsets: the design of a wideband handset is a new endeavor. The selection of a microphone and an earphone driver unit, the method of enclosing these units in a handset, noise suppression by circuit design, and equalization of frequency characteristics by signal processing are important for a high-quality handset design.

We developed the first high-quality telephone NP-1 for a trial of the Next Generation Network (NGN) in 2007. The handset uses microphone and earphone driver units intended for audio products rather than for telephones. The units were enclosed, and we made a small air cabinet in front of each unit. We adjusted the volume and shapes of the cabinets and the placement of sound holes for the earphone (**Fig. 3**). As a result, the handset achieved flat frequency characteristics up to 7 kHz. Generally, a flat-frequency earphone is more difficult to design than a flat-frequency microphone. The best design method differs according to the acoustic driver units that can be used. Therefore, when designing a handset, one must leave room for making sound holes and an air cabinet.

Noise suppression is not considered for the under-300-Hz or over-3.4-kHz frequency ranges in the circuits of a normal telephone. For wideband use, the circuits must be redesigned. If standard characteristics are not satisfied by the microphone and earphone designs, then the use of additional signal processing is an efficient measure.

We developed a high-quality Wi-Fi telephone as a prototype in 2009 (**Fig. 4**). It has a very small body for portability and does not have air cabinets. For a high-quality handset design, only large sound holes were added. The standard characteristics were satisfied by selecting an acoustic driver unit and by equalizing sound through signal processing.

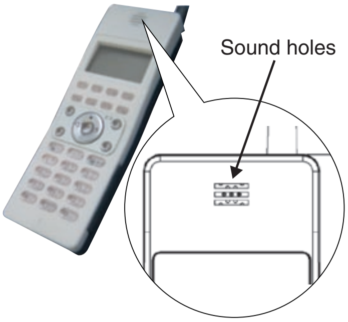

Fig. 4. High-quality Wi-Fi telephone.

4. Echo canceller

Hands-free speakerphones are useful for conversation and for sharing information among several people, while phone calls via a handset are suitable for one-to-one communications because the voice is heard only by the far-end speaker. Hands-free phones did not used to be widely utilized except for specific situations such as business teleconferences. In recent years, however, high-performance mobile terminals like netbooks or smartphones have become prevalent, and hands-free calls using those small devices have become more popular.

When you talk on a hands-free telephone (**Fig. 5**), the voice of the far-end speaker emitted from the loudspeaker is picked up by the microphone and sent back to the far-end speaker. In other words, the far-end speaker hears his or her own voice like an echo. To avoid this phenomenon, most hands-free speakerphones and conferencing systems are equipped with echo cancellers, which decrease or eliminate the echo through signal processing. If the environments at both ends are quiet enough and the echo is not too

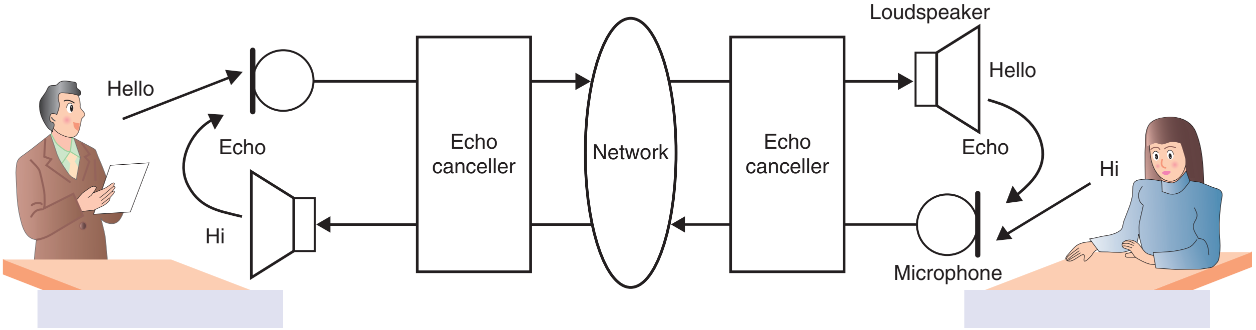

Fig. 5. Hands-free telephone.

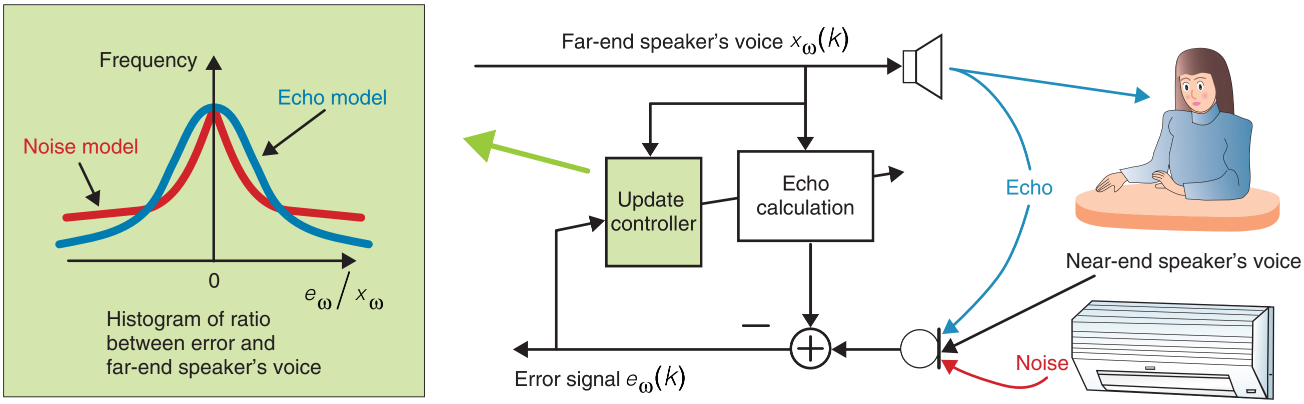
Frequency
Echo model
Noise model
0
e_ω / x_ω
Histogram of ratio between error and far-end speaker's voice
Far-end speaker's voice $x_\omega(k)$
Update controller
Echo calculation
Echo
Near-end speaker's voice
−
Noise
Error signal $e_\omega(k)$

Fig. 6. Noise-robust echo cancellation.

loud, most of these echo cancellers work properly. However, echo cancellation in small devices such as smartphones has two technical problems: the treatment of ambient noise and appropriate hardware design for the device.

Ambient noise deteriorates the echo canceller's performance, especially the adaptive filter's performance. An echo canceller usually has an adaptive filter for estimating the acoustic echo path and calculating the pseudo-echo signal. This estimation is done only when the loudspeaker emits sounds and the near-end speaker does not talk. The microphone picks up only the echo signal, and the adaptive filter estimates the acoustic echo path so that the calculated pseudo-echo can be made equal to the observed signal. If the acoustic echo path is estimated precisely, the echo canceller can eliminate the echo from the sending voice even when the speakers at both ends talk simultaneously (called the double-talk state). However, if the observed signal is contaminated by ambient noise, the estimated echo path has a large amount of error. As a result, the echo canceller cannot eliminate the echo and the remaining echo interferes with communication.

Our new technology, noise-robust echo cancellation (**Fig. 6**), enables echo path estimation by means of an adaptive filter that is less sensitive to ambient noise. This technology uses the ratio of the noise in the microphone signal and controls the update speed of the acoustic echo path. Specifically, the update controller in Fig. 6 calculates the ratio between the far-end speaker signal and the error signal. It then determines whether the noise is dominant within the error signal on the basis of the ratio's statistical distribution. The more noise included in the error signal, the more the update controller decreases the update speed. An echo canceller using this technology cancels echoes three times as much as the conventional echo canceller in a noisy environment, and this performance is suitable for most conditions for hands-free

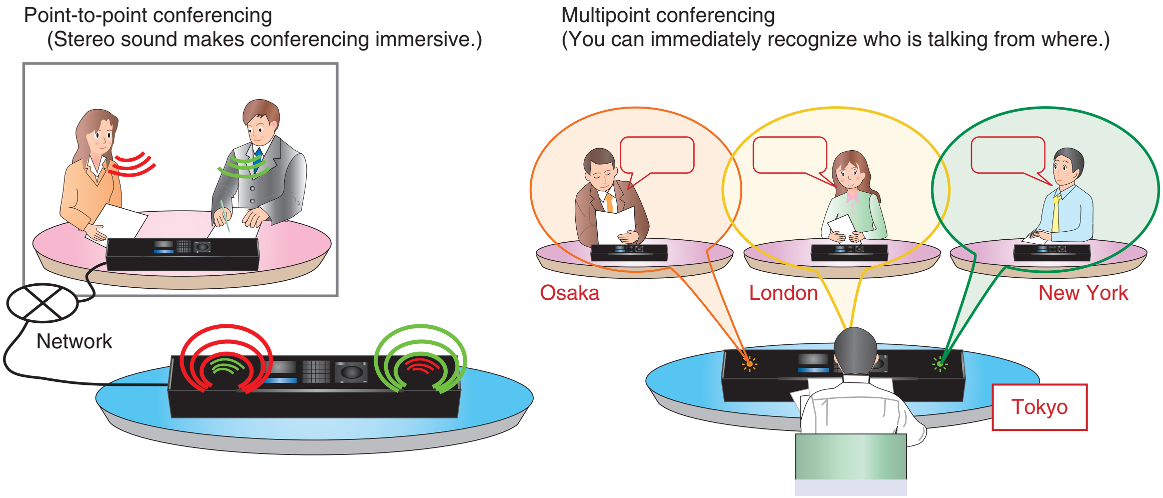
Point-to-point conferencing
(Stereo sound makes conferencing immersive.)
Network
Multipoint conferencing
(You can immediately recognize who is talking from where.)
Osaka
London
New York
Tokyo

Fig. 7. Typical usages of stereo conference phone.

telephone calls.

The other problem is hardware design. In a small device, the loudspeaker and microphone are located much closer together than in ordinary videoconference systems, and the echo signal is too large to eliminate. Therefore, some ingenuity is needed to make an echo canceller that works well. One solution is to use a directional microphone and locate its low-sensitivity direction toward the loudspeaker in order to avoid picking up echoes. In addition, sound propagating directly along the device's housing to the microphone must be reduced. Adding a rib to strengthen the housing structure, dividing the space between the loudspeaker and microphone, and attaching the microphone to the housing by rubber or sponge mountings, for example, are effective ways to reduce the propagation of vibration from the loudspeaker. It is very difficult to know what the optimal echo cancellation structure is or how to design the hardware; the current design is mainly based on experience. We are considering calculating the optimal structure by computer simulation in the future.

We have now achieved echo cancellation in a small device with performance equivalent to that of an ordinary (large) hands-free telephone. The HIKARI FLET'S phone VP3000 [4], which uses the technologies described above, is now available from NTT EAST and NTT WEST.

5. Stereo conference phone

In recent years, audioconferencing and videoconferencing have been in strong demand in order to cut the cost of business trips and reduce the environmental load. Though audioconferencing (using telephones) is inexpensive, easy to use, and more popular than videoconferencing, it has two drawbacks: it is hard to judge who is talking at the remote site and it is hard to sense the atmosphere at the remote site.

To address these drawbacks, we are developing a new stereo conference phone with 14-kHz stereo sound that corresponds to the quality of FM (frequency modulation) radio. The use of wideband stereo sound makes it easier for people to understand the situation at the remote site.

The stereo conference phone has three features: a 14-kHz audio codec, automatic gain control, and a multimodal display. We extended the 7-kHz codec G.711.1 to 14 kHz. This codec is already standardized as G.711.1 Annex D (G.711.1 SWB). Automatic gain control improves listenability by adjusting voice levels moderately depending on whether the talker is far from or close to the conference phone. The multimodal display uses light emitting diodes and changes their lighting pattern according to the balance between the left and right channels and the levels of stereo sound. This helps users determine who is speaking where at the remote site.

Typical uses of the stereo conference phone are shown in **Fig. 7**. In point-to-point conferencing, FM-quality stereo sound can deliver not only voices but also ambient sound in the remote site vividly. The multimodal display visually assists users in determining who is talking at the remote site. This makes phone conferencing increasingly immersive. In multipoint conferencing, our new stereo conference phone can reproduce sound from each site at a separate position. This overcomes the problem of ordinary multipoint conference phones, which mix voices into monaural sound with the result that you cannot immediately distinguish who is talking from where. Our stereo conference phone makes conferencing more efficient and productive. It can also be used as external hands-free audio equipment with videoconferencing software for a personal computer (PC). Since this audio equipment is self-contained and no tuning is necessary, it will make PC-based high-definition videoconferencing tools more convenient.

References

[1] HIKARI DENWA.
http://flets.com/english/hikaridenwa/index.html

[2] FLET'S HIKARI NEXT
http://flets.com/english/next/index.html

[3] S. Sasaki, T. Mori, Y. Hiwasaki, and H. Ohmuro, "Global Standard for Wideband Speech Coding: ITU-T G.711.1 (G.711 wideband extension)," NTT Technical Review, Vol. 6, No. 8, 2008.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200808le1.html

[4] HIKARI FLET'S Phone (in Japanese).
http://flets.com/fletsphone/VP3000/

Hitoshi Ohmuro

Senior Research Engineer, Supervisor, Promotion Project 1, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Aichi, in 1988 and 1990, respectively. Since joining NTT Human Interface Laboratories in 1990, he has been researching speech coding. He contributed to ITU-T G.711.1 standardization. He is currently developing VoIP applications and terminals. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), and the Acoustical Society of Japan (ASJ).

Manabu Okamoto

Senior Research Engineer, Supervisor, Promotion Project 1, NTT Cyber Space Laboratories.

He received a masters degree in design from Kyushu Institute of Design, Fukuoka, in 1991 and a doctor's degree in design from Kyushu University, Fukuoka, in 2007. Since joining NTT Electrical Communication Laboratories in 1991, he has been researching the acoustic design of various kinds of teleconferencing systems. He is a member of IEICE and ASJ.

Shoichiro Saito

Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees from the University of Tokyo in 2005 and 2007, respectively. Since joining NTT Cyber Space Laboratories in 2007, he has been engaged in research on echo cancellers and the development of hands-free telephone terminals. He is a member of IEEE, IEICE, ASJ, and the Information Processing Society of Japan.

Satoru Emura

Senior Research Engineer, Promotion Project 1, NTT Cyber Space Laboratories.

He received the B.E., M.E., and D.E. degrees from the University of Tokyo in 1992, 1994, and 1997, respectively. Since joining NTT Human Interface Laboratories in 1997, he has been researching adaptive signal processing. He is a member of IEEE, IEICE, and ASJ.

Sumitaka Sakauchi

Senior Research Engineer, Promotion Project 1, NTT Cyber Space Laboratories.

He received the B.S. degree in physics from Yamagata University in 1993, the M.S. degree in physics from Tohoku University, Miyagi, in 1995, and the Ph.D. degree in engineering from Tsukuba University, Ibaraki, in 2005. Since joining NTT Human Interface Laboratories in 1995, he has been researching acoustic echo cancellers and noise reduction. He received the Best Paper Award from IEICE in 2001. He is a member of IEICE and ASJ.