Transmission of High-quality Sound via Networks Using Speech/Audio Codecs

Yutaka Kamamoto, Takehiro Moriya, and Noboru Harada

Abstract

This article describes two recent advances in speech and audio codecs. One is EVS (Enhanced Voice Service), the new standard by 3GPP (3rd Generation Partnership Project) for speech codecs, which is capable of transmitting speech signals, music, and even the ambient sound on the speaker's side. This codec has been adopted in a new VoLTE (voice over Long-Term Evolution) service with enhanced high-definition voice (HD+), which provides us with clearer and more natural conversations than conventional telephony services such as with fixed-line/land-line and 3G mobile phones. The other is MPEG-4 Audio Lossless Coding (ALS) standardized by the Moving Picture Experts Group (MPEG), which makes it possible to transmit studio-quality audio content to the home. ALS is expected to be used by some broadcasters, including IPTV (Internet protocol television) companies, in their broadcasts in the near future.

Keywords: audio/speech coding, data compression, international standards

1. Introduction

Many audio and speech codecs are available, and we can select the most suitable one for different usage scenarios ranging from those requiring reasonable quality with low bit rates to ones demanding original signal quality with high bit rates. With the increases in network capacity that have been achieved, content that requires high bit rates such as 4K television (TV) and high-resolution audio can also be transmitted. However, the first priority is to transmit speech signals in ordinary telephony without congestion. Therefore, speech codecs for telephony should use as low a bit rate as possible. In addition, they must have lower algorithmic delay because the longer the codec delay is, the more difficult it becomes for people to communicate with each other.

In contrast, one-way transmission such as broadcasting is less sensitive to delay. Most audio codecs utilize the advantages of longer delay and then efficiently compress audio signals by means of signal processing with sufficient frame length. Moreover, speech codecs use a human phonation model, so they are not suitable for music. When clean speech items are coded by audio codecs at low bit rates, we get the impression that a machine is talking. Speech and audio compression schemes have these kinds of trade-offs. To achieve the best quality of speech and music content with less delay, experts in speech and audio coding around the world have been working together to develop new codecs. Furthermore, lossless coding, which refers to compression without any loss, has been standardized to ensure that the quality of the original content is maintained.

This article presents an overview of the 3rd Generation Partnership Project (3GPP) Enhanced Voice Services (EVS) codec, which has been newly standardized for mobile communications, and MPEG-4 Audio Lossless Coding (ALS) for high-resolution audio, which was standardized by the Motion Picture Experts Group (MPEG).

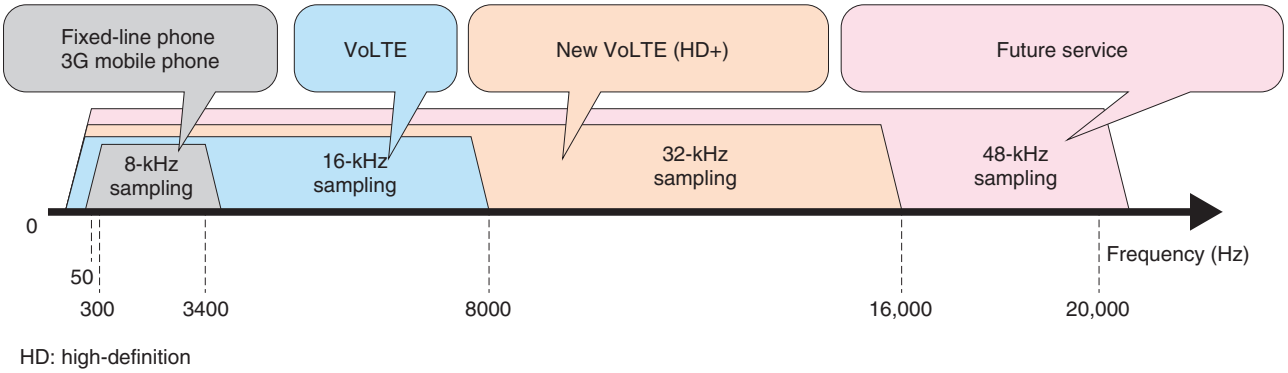


Fig. 1. Supported audio bandwidth in EVS.

2. 3GPP EVS codec for mobile communications

3GPP is the international standardization consortium for mobile communications. It has newly defined EVS speech and audio coding standards for voice over Long-Term Evolution (VoLTE) [1, 2]. Conventional speech coding schemes for mobile phones have been based on code excited linear prediction (CELP). These schemes have utilized a human voice production model and achieved high-quality speech transmission with very low bit rates. EVS consists of newly developed low-delay and low-bit-rate audio coding modules in addition to CELP, and it achieves high-quality transmission of various types of input signals, including speech, audio, background noise, and background music [3, 4].

EVS uses new bandwidth extension technologies to support signals with higher sampling rates up to 48 kHz, in contrast to the narrowband signal (8-kHz sampling rate) of conventional fixed-line/land-line telephones and 3G mobile phones and the wideband signal (16-kHz sampling rate) of VoLTE. Note that the wideband signal is used for AM (amplitude modulation) radio, the super-wideband signal (32-kHz sampling rate) is used for FM (frequency modulation) radio, and the full-band signal (48-kHz sampling rate) is used for digital broadcasting, as shown in **Fig. 1**.

EVS has been optimized for VoLTE with a frame length of 20 ms and algorithmic delay of 32 ms. It has been designed to minimize perceptual distortion against packet loss, whereas coding schemes for conventional 3G mobile phones were optimized for robustness against bit errors. In addition, EVS covers a wide range of bit rates from 5.9 kbit/s to 128 kbit/s and enables frame-by-frame selection of bit rates. This enables smooth migration from the conventional VoLTE system since EVS has inter-operability with AMR-WB (Adaptive Multi-Rate Wideband).

During the standardization process, a huge number of subjective quality evaluations were conducted for various coding conditions, input items, and languages. The results of the evaluations indicated that EVS outperformed conventional speech and audio coding schemes in terms of quality [5]. NTT used a similar procedure to conduct listening tests on Japanese materials [6]. The results of the tests confirmed the superiority of EVS over the coding schemes for conventional mobile communications systems.

Note that all EVS development has been carried out by 12 organizations[*1] based mainly in Europe, North America, and East Asia, including Japanese companies. EVS has been deployed in commercial services such as VoLTE (HD+) by NTT DOCOMO since the summer of 2016 [7] and by some operators in the USA and Europe as well [8]. We believe that EVS will allow billions of people around the world to enjoy high-quality communication in the near future.

3. MPEG-4 ALS for high-resolution audio transmission

MPEG has standardized many useful audio and video codecs that are commonly used in daily life. The MPEG audio subgroup standardized the lossless compression scheme MPEG-4 ALS, which can perfectly reconstruct original signals. NTT is one of the

*1 In alphabetical order, Fraunhofer IIS, Huawei Technologies Co. Ltd, Nokia Corporation, NTT, NTT DOCOMO, INC., ORANGE, Panasonic Corporation, Qualcomm Incorporated, Samsung Electronics Co., Ltd., Telefonaktiebolaget LM Ericsson, VoiceAge Corporation, and ZTE Corporation.

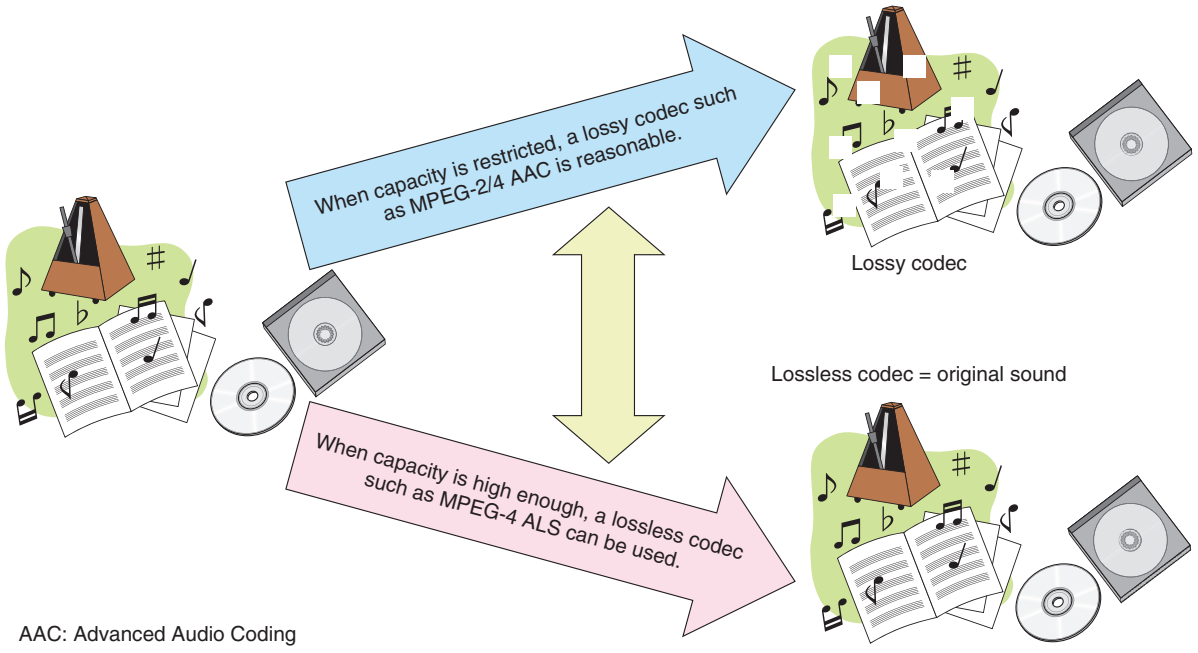


Fig. 2. Selection of audio codec according to the allowed bit rate.

contributors to this standard [9, 10]. Although the compression ratio depends on the input signal, a file is normally compressed to around 30% to 70% of its original size [11, 12]. High-resolution audio*2 has become more popular recently and enables precisely digitized music. Lossy codecs often cannot convey the fidelity of high-resolution audio, so a lossless codec such as MPEG-4 ALS is necessary.

For example, the audio signal in TV broadcasting is usually produced in a studio or broadcasting station in a 48-kHz, 24-bit format, which is referred to as high-resolution audio. Since access to radio waves is limited, we cannot assign a high bit rate to the content, and it is necessary to compress the audio signal at a loss of quality. This lossy codec enables us to enjoy broadcasting content because the codec reduces the bit rate remarkably without any noticeable difference from the original signal.

The ultrahigh-definition (or super-high-definition) TV service called 4K/8K TV has now started, and it uses very high bit rates. The audio signal is also expected to use higher bit rates. The Association of Radio Industries and Businesses (ARIB), which defines the standards for radio-wave systems in Japan, standardized ARIB STD-B32 for the 4K/8K TV system. This standard enables the use of MPEG-4 ALS as one of the audio codecs [14]. MPEG-4 ALS can reconstruct in the home music content that was produced in a broadcasting studio—with the quality of the original signal—because the lossless codec guarantees bit-exactness over the entire transmission. We can enjoy high-resolution audio in our living rooms when a sufficient bit rate is assigned to the audio signal (**Fig. 2**). IPTV (Internet protocol TV) services, which use optical-fiber lines, may introduce MPEG-4 ALS before the radio-wave services do.

In order to support practical deployment of MPEG-4 ALS, NTT has prepared related standards such as MPEG-4 ALS Simple Profile and IEC 61937-10 Edition 2 by the International Electrotechnical Commission (IEC) [15]. MPEG-4 ALS Simple Profile restricts the parameters of the input signal such as the sampling frequency, number of channels, bit depth, and frame size, and also restricts some processing tools and functionalities that require higher computational complexity such as compression for floating-point format signals. ARIB STD-B32 recommends the use of MPEG-4 ALS Simple Profile with LATM/LOAS (Low-overhead Audio Transport Multiplex/Low Overhead Audio Stream)*3 capsuling. To facilitate

*2 High-resolution audio: The Japan Electronics and Information Technology Industries Association (JEITA) defines high resolution audio as an audio signal with a sampling frequency higher than 48 kHz and a bit depth greater than 16 bits [13].

*3 LATM/LOAS: A transmission scheme of the header and stream defined in MPEG-4 Audio.

the connection of a TV and a digital audio device, IEC 61937-10 Edition 2 newly supports the bitstream of MPEG-4 ALS Simple Profile with LATM/LOAS, which can be transmitted via radio waves with 4K/8K video. Then, with 4K/8K TV, we can listen to high-quality music through a high-resolution TV or by using a digital amplifier that can decode MPEG-4 ALS.

4. Future work

The EVS codec enables high-quality communications by means of speech and music even when delay and bit rates are low. MPEG-4 ALS can transmit original audio content losslessly. The delivery of high-quality music has now been achieved, so we will start to consider ways to achieve even more realistic audio transmission. Basic research on interactive communication may achieve a synergistic effect between live venues and reception sites. We will continue to develop speech and audio codec schemes to make timely contributions to new services.

References

[1] 3GPP TS26.441: Codec for Enhanced Voice Services (EVS); General overview, 2015.

[2] 3GPP TS26.445: Codec for Enhanced Voice Services (EVS); Detailed algorithmic description, 2015.

[3] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, H. Sung, E. Oh, H. Yuan, and C. Zhu, "Overview of the EVS Codec Architecture," Proc. of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), pp. 5698–5702, Brisbane, Australia, Apr. 2015.

[4] S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Järvinen, L. Laaksonen, N. Harada, N. Naka, S. Ragot, S. Proust, T. Sanda, I. Varga, C. Greer, M. Jelínek, M. Xie, and P. Usai, "Standardization of the New 3GPP EVS Codec," Proc. of ICASSP 2015, pp. 5703–5707, Brisbane, Australia, Apr. 2015.

[5] 3GPP TR26.952: Codec for Enhanced Voice Services (EVS); Performance characterization, 2015.

[6] Y. Kamamoto, T. Moriya, and N. Harada, "Subjective Evaluation of 3GPP EVS Codec for Japanese Speech Items," The Acoustical Society of Japan 2016 Spring Meeting, 3-2-9, 2016 (in Japanese).

[7] Website of NTT DOCOMO, VoLTE/VoLTE (HD+) service (in Japanese),
https://www.nttdocomo.co.jp/support/area/volte/

[8] Press release issued by Ericsson on June 14, 2016.
https://www.ericsson.com/news/160614-evolved-hd0-voice-for-volte_244039855_c

[9] ISO/IEC14496-3: Information Technology—Coding of Audio-visual Objects—Part 3: Audio, fourth edition, 2009.

[10] T. Liebchen, T. Moriya, N. Harada, Y. Kamamoto, and Y. A. Reznik, "The MPEG-4 Audio Lossless Coding (ALS) Standard - Technology and Applications," Proc. of the 119th Audio Engineering Society Convention, New York, USA, Oct. 2005.

[11] Y. Kamamoto, T. Moriya, N. Harada, and C. Kos, "Enhancement of MPEG-4 ALS Lossless Audio Coding," NTT Technical Review, Vol. 5, No. 12, 2007.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200712sp2.html

[12] N. Harada, T. Moriya, and Y. Kamamoto, "MPEG-4 ALS: Performance, Applications, and Related Standardization Activities," NTT Technical Review, Vol. 5, No. 12, 2007.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200712sp3.html

[13] JEITA, "Definition of High-resolution Audio," Mar. 2014 (in Japanese).
http://home.jeita.or.jp/page_file/20140328095728_rhsiN0Pz8x.pdf

[14] ARIB STD-B32: Video Coding, Audio Coding, and Multiplexing Specifications for Digital Broadcasting, 2011.

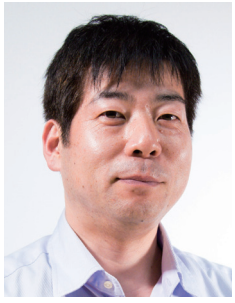
[15] IEC 61937-10: Digital Audio - Interface for Non-linear PCM Encoded Audio Bitstreams Applying IEC60958 – Part 10: Non-linear PCM Bitstreams According to the MPEG-4 Audio Lossless Coding (ALS) Format, 2011.

Yutaka Kamamoto
Senior Research Scientist, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received a B.S. in applied physics and physico-informatics from Keio University, Kanagawa, in 2003 and an M.S. and Ph.D. in information physics and computing from the University of Tokyo in 2005 and 2012. Since joining NTT Communication Science Laboratories in 2005, he has been studying signal processing and information theory, particularly speech/audio coding and lossless compression of time-domain signals. He was also with NTT Network Innovation Laboratories from 2009 to 2011, where he developed the audio-visual codec for ODS (Other Digital Stuff/Online Digital Source). He has contributed to the standardization of coding schemes for MPEG-4 ALS, the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Recommendation G.711.0: Lossless compression of G.711 pulse code modulation, and 3GPP EVS. He received the Telecom System Student Award from the Telecommunications Advancement Foundation (TAF) in 2006, the IPSJ Best Paper Award from the Information Processing Society of Japan (IPSJ) in 2006, the Telecom System Encouragement Award from TAF in 2007, the Awaya Prize Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2011, and the Technical Development Award from ASJ in 2016. He is a member of IPSJ, ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and the Institute of Electrical and Electronics Engineers (IEEE).

Takehiro Moriya
NTT Fellow, Head of Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his B.S., M.S., and Ph.D. in mathematical engineering and instrumentation physics from the University of Tokyo in 1978, 1980, and 1989. Since joining NTT laboratories in 1980, he has been conducting research on medium- to low-bit-rate speech and audio coding. In 1989, he worked at AT&T Bell Laboratories as a visiting researcher. He has contributed to the standardization of coding schemes for the Japanese Public Digital Cellular System, ITU-T, ISO/IEC MPEG, and 3GPP. He has received a number of awards including the IEEE James L. Flanagan Speech and Audio Processing Award in 2016. He is an IEEE Fellow, IEICE Fellow, and a member of IPSJ and ASJ. He is also a member of the IEEE Speech Technical Committee of the Signal Processing Society (SPS) and a chair of the IEEE SPS Tokyo Joint Chapter.



Noboru Harada
Senior Research Scientist, Supervisor, Moriya Research Laboratory, NTT Communication Science Laboratories.

He received his B.S. and M.S. in computer science and systems engineering from Kyushu Institute of Technology, Fukuoka, in 1995 and 1997. He joined NTT in 1997. His main research area is lossless audio coding, high-efficiency coding of speech and audio, and their applications. He was also with NTT Network Innovation Laboratories from 2009 to 2011, where he developed the audio-visual codec for ODS. He is an editor of numerous international standards specifications including ISO/IEC 23000-6:2009 Professional Archival Application Format and ITU-T G.711.0 and has contributed to the standardization of MPEG-4 ALS and 3GPP EVS. He is a member of IEICE, ASJ, Audio Engineering Society (AES), and IEEE.