EVS | Speech Coding | Standardization

Technology Reports

3GPP EVS Codec for Unrivaled Speech Quality and Future Audio Communication over VoLTE

NTT DOCOMO has been engaged in the standardization of the 3GPP EVS codec, which is designed specifically for VoLTE to further enhance speech quality, and has contributed to establishing a far-sighted strategy for making the EVS codec cover a variety of future communication services. NTT DOCOMO has also proposed technical solutions that provide speech quality as high as FM radio broadcasts and that achieve both high coding efficiency and high audio quality not possible with any of the state-of-the-art speech codecs. The EVS codec will drive the emergence of a new style of speech communication entertainment that will combine BGM, sound effects, and voice in novel ways for mobile users.

Research Laboratories *Kimitaka Tsutsumi*
Kei Kikuiri

1. Introduction

The launch of Voice over LTE (VoLTE) services and flat-rate voice service has demonstrated the importance of high-quality telephony service to mobile users. In line with this trend, the 3rd Generation Partnership Project (3GPP) completed the standardization of the speech codec for Enhanced Voice Services (EVS) [1] in September 2014.

The speech quality of existing telephony service has been as high as AM-radio quality*1 due to speech codecs such as Adaptive Multi-Rate - Wide-Band (AMR-WB)*2 [2] that is used in NTT DOCOMO's VoLTE and that support wideband speech with a sampling frequency*3 of 16 kHz. In contrast, EVS has been designed to support super-wideband*4 speech with a sampling frequency of 32 kHz thereby achieving speech of FM-radio quality*5. The introduction of EVS in VoLTE can therefore be expected to provide a quantum leap in the quality of telephony services. Additionally, considering the appearance of Moving Picture Experts Group Unified Speech and Audio Coding (MPEG USAC)*6 [3] [4], which in addition to speech can also encode music at high levels of quality and efficiency for non-real-time services, 3GPP experts agreed to adopt high requirements in the EVS codec for music despite its main target of real-time communication. Furthermore, considering that telephony services using AMR-WB are finding widespread use around the world, EVS is required to have a mode compatible with AMR-WB [5].

NTT DOCOMO has been participating in EVS standardization activities since 2010 emphasizing that the goal should be for early and widespread penetration. This would be accomplished

*1 **AM-radio quality:** The capability of expressing a speech band from 100 Hz to 7.5 kHz.
*2 **AMR-WB:** A speech codec used in, for example, telephony services, having a better quality than that of AMR-NB.
*3 **Sampling frequency:** A unit (in Hz) expressing the number of times per second that the acoustic pressure of a speech signal input from a microphone is recorded. A higher frequency enables a higher range of sounds to be recorded.

by establishing EVS design requirements that minimize the changes that would have to be made to the network when introducing EVS in VoLTE. In line with this argument, the EVS bit rate has been set so that data size transported over the radio interface in VoLTE could be the same as that of AMR-WB thereby enabling the design of the VoLTE radio network for AMR-WB to be unchanged. Moreover, envisioning the further evolution of VoLTE due to the future merging of voice and music services, emphasis has been placed on music quality at low bit rates. In the beginning, discussions proceeded in the direction that requirements for music should be set low relative to those for speech. In the end, however, high requirements were set beneficial to music coding specifying that audio quality should be on the same level as a codec having an algorithmic delay*7 longer than that of the EVS codec. In this regard, NTT DOCOMO's Melody Call®*8 (musical ring tone service) has been used widely in existing 3G and VoLTE telephony services, but since it uses the speech-specific AMR codec [6], it cannot necessarily provide high-quality music. It can therefore be expected that EVS satisfying the above high requirements should not only improve the quality of such existing music-content services but also facilitate the appearance of new services that use music as part of telephony services [7].

As a result of the above discussions at 3GPP, EVS has been standardized as a codec having three key features: (1) support of super-wideband speech having FM-radio quality, (2) easy implementation in VoLTE, and (3) high audio quality for both speech and music.

In this article, we first provide an overview of EVS. We then describe the main technologies used for improving quality and the technologies that NTT DOCOMO has contributed. Finally, we present the results of quality evaluation tests for super-wideband speech and music as a part of assessing EVS performance in the characterization phase.

2. EVS Technical Features

2.1 EVS Overview

In addition to wideband and super-wideband audio bandwidths, EVS also supports narrowband speech having a sampling frequency of 8 kHz and full-band speech having a sampling frequency of 48 kHz higher than that of CDs. The EVS codec covers a wide range of bit-rates from 5.9 to 128 kbps.

A high-level overview of the EVS codec is shown in **Figure 1**. In low-bit-rate operational modes, signal classification is performed on a frame-by-frame*9 basis to decide which coding strategy to use: time-domain coding that provides high quality and efficiency for speech or frequency-domain coding for music. In high-bit-rate operational mode in which a substantial amount of information is available, only the frequency-domain coding is used regardless of the type of input signal.

Since it is impossible to avoid packet loss*10 in a packet-switched network such as VoLTE, Packet Loss Concealment (PLC) that reconstructs missing frames and ensures rapid and smooth recovery is important for providing high quality even under erroneous channel conditions. The PLC technique for the EVS codec has also been developed as part of the standard [8]. PLC for the EVS codec switches its strategy between time-domain and frequency-domain based on the coding strategy of the last frame normally decoded before packet loss.

The following provides an overview of time-domain coding, frequency-domain coding, and the PLC technique.

1) Time-domain Coding

An overview of time-domain coding is shown in **Figure 2**. Human hearing is more sensitive to lower-frequency components than higher-frequency components. Therefore, the total amount of information can be efficiently reduced while maintaining speech quality by coding low-frequency and high-frequency components separately with uneven bit allocation, which assigns a lot less bits to high-frequency components.

(1) In the EVS codec, Code Excited Linear Prediction (CELP) is used to encode lower-frequency com-

*4 **Super-wideband:** A speech band with lower and upper frequency limits of 50 Hz and 16 kHz, respectively.

*5 **FM-radio quality:** The capability of expressing a speech band from 50 Hz to 15 kHz.

*6 **MPEG USAC:** MPEG unified speech and audio codec. MPEG is a set of standards specifying coding and transmission systems for digital audio and video. It was formed by a ISO/IEC joint working group.

*7 **Algorithmic delay:** An index indicating the delay in outputting decoded sound with respect to the original sound. It is determined by codec specifications, and in the case of a codec in the frequency domain, a longer delay can generally improve coding efficiency.

*8 **Melody Call®:** A NTT DOCOMO service that enables the user to change the ring tone on the mobile phone to his/her favorite tunes. A registered trademark of NTT DOCOMO, Inc.

*9 **Frame:** The period in which an encoder/decoder operates or a speech signal of a length corresponding to that period. In the EVS codec, the frame length is 20 ms, which means that encoding/decoding is performed once every 20 ms.

*10 **Packet loss:** The failure of a speech packet to be delivered as far as the decoding stage due to congestion or other problems.

ponents to obtain linear prediction*11 coefficients and a linear prediction residual signal*12 after quantization*13. The encoding process for the linear prediction residual signal exploits the property of speech signals that similar waveforms (each being a pitch waveform*14) are repeated along

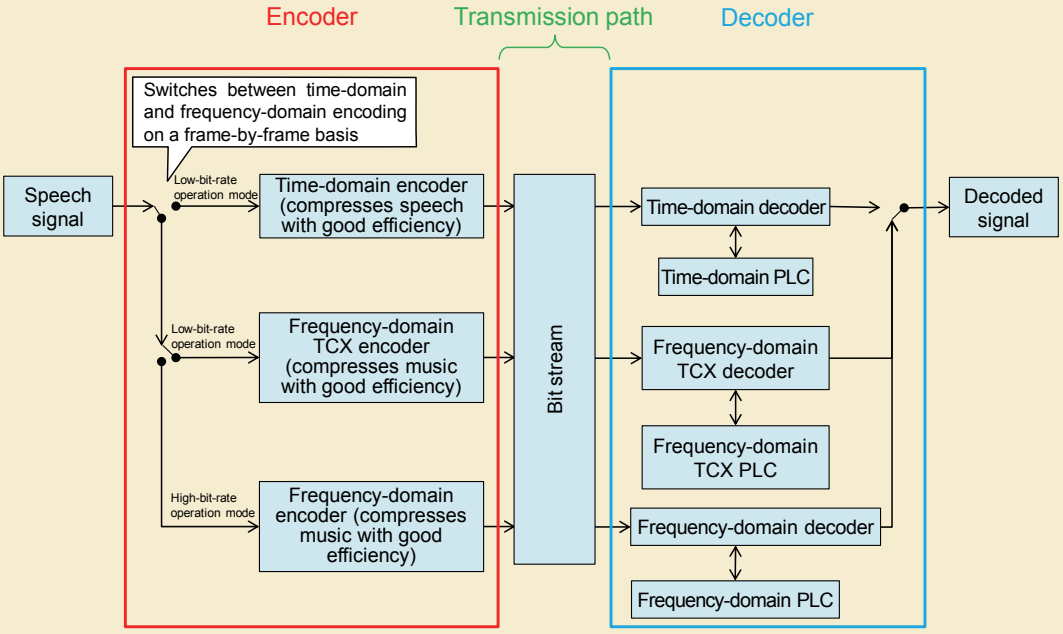


Figure 1 Basic configuration of EVS encoder/decoder

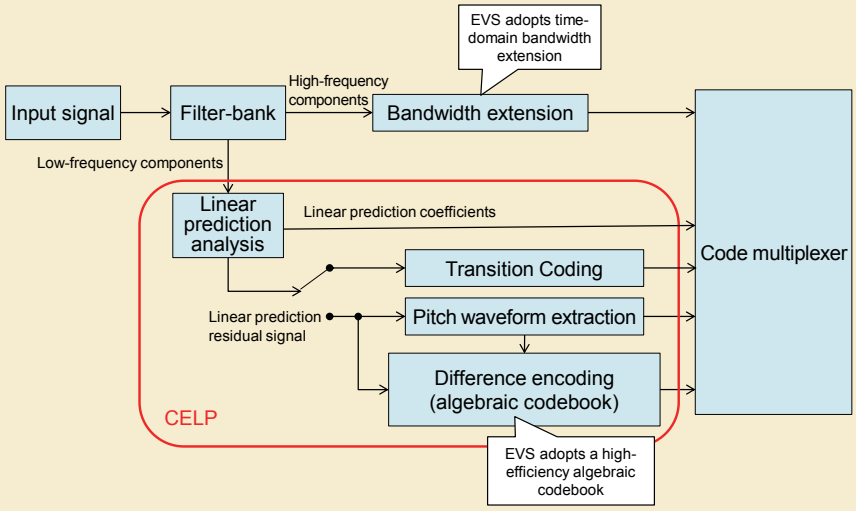


Figure 2 Configuration of time-domain encoding

*11 **Linear prediction:** A technique that approximates the speech signal at a certain point in time by taking a linear sum of previous speech signals.

*12 **Linear prediction residual signal:** The signal representing the prediction error when applying linear prediction to the input signal.

*13 **Quantization:** A process of mapping input values to a smaller set of predetermined discrete values. While resulting in some distortion, quantization can significantly reduce the amount of information.

*14 **Pitch waveform:** One period's worth of similar repeating waveforms that characterize speech.

the time axis. In the encoding, the difference with the immediately preceding pitch waveform together with the length of the pitch waveform is encoded. The difference is encoded by an algebraic codebook*15, which is a technique incorporated in many recent speech codec standards including AMR and AMR-WB. In the EVS codec, coding-efficiency of the algebraic codebook is significantly improved to provide higher sound quality even at low bit rates.

(2) Higher-frequency components encoding is based on a band extension technique that produces higher-frequency components by replicating and shaping lower-frequency components. Only the parameter for the shaping is transmitted at a low bit rate by the encoder, and higher-frequency components are reconstructed based on this parameter by the decoder. Therefore, high-frequency components can be obtained even at low bit rates. In the EVS codec, time-domain band extension encoding is also used to achieve high quality with low delay.

2) Frequency-domain Coding

An overview of frequency-domain coding is shown in **Figure 3**. The input signal is first transformed into a frequency-domain representation using the Modified Discrete Cosine Transform (MDCT)*16 and then encoded.

There are two methods for encoding these MDCT coefficients. The first method divides the MDCT coefficients into sub-bands*17, and obtains the scale factor*18 of each sub-band and the MDCT coefficients normalized by those scale factors. The normalized MDCT coefficients are encoded with vector quantization*19. The second method is Transformed Code Excitation (TCX) that encodes the linear prediction residual signal in the frequency domain.

TCX is also used in MPEG USAC. In the EVS codec, arithmetic coding is improved to obtain enhanced error resistance while maintaining coding efficiency.

In TCX, all MDCT coefficients cannot be encoded due to limited bits for coding and those that cannot are quantized to zero. This results in frequency bands with

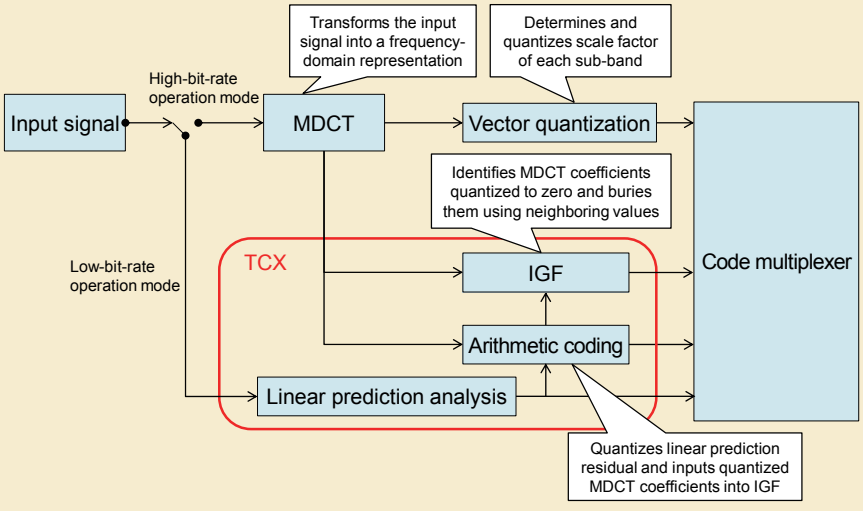


Figure 3 Configuration of frequency-domain encoding

*15 **Codebook:** A set of previously determined candidate vectors for quantizing input vectors.

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

*17 **Sub-band:** One of the bands that result from splitting an entire frequency band into multiple parts.

*18 **Scale factor:** A power of a sub-band or its quantized amplitude.

*19 **Vector quantization:** A quantization technique that maps a numerical sequence of length two or more to the closest value of predetermined numerical sequences of the same length.

no signal components and introduces sound quality degradation. To mitigate this degradation, conventional codecs have used a noise filling technique that buries noise in such frequency bands. In EVS, Intelligent Gap Filling (IGF) has been adopted to fill those bands with nearby MDCT coefficients.

3) Packet Loss Concealment

(1) Time-domain PLC

CELP is based on inter-frame prediction*20, which results in error propagation even after receiving packets followed by a packet loss. For smooth and fast recovery of linear prediction coefficients and a linear prediction residual signal, Transition Coding mode*21 (fig. 2) is used in the EVS codec. In this mode, the linear prediction coefficients and the linear prediction residual signal are encoded independently from those in the previous frame thereby improving packet loss resilience.

(2) Frequency-domain PLC

The conventional frequency-domain PLC technique basically copies the MDCT coefficients in the last frame before the packet loss as a substitute for the parameter in the lost frame. However, repetition of the MDCT coefficients in the past frame sometimes introduces discontinuities in the waveform. In the EVS codec, waveform adjustment based on phase control is used for smooth connection between a concealed frame and its adjacent frames.

2.2 Quality-improvement Technologies in EVS

A variety of technologies have been introduced in EVS for speech quality improvement. The following provides an overview of those that are particularly important.

1) Time-domain Bandwidth Extension

Bandwidth extension is a technology that generates higher-frequency components at low bit rates. Higher-frequency components are first generated using lower-frequency components and then shaped so as to have a power distribution indicated by the encoder.

In time-domain bandwidth extension in the EVS encoder, a linear prediction spectrum*22 is calculated based on higher-frequency components from a filter-bank*23 and coded to indicate a power distribution of higher-frequency components. In the decoder, lower-frequency components from the CELP decoder are modified to obtain an excitation signal, and then the excitation signal is fed to a synthesis filter having the decoded linear prediction spectrum (**Figure 4**) to obtain

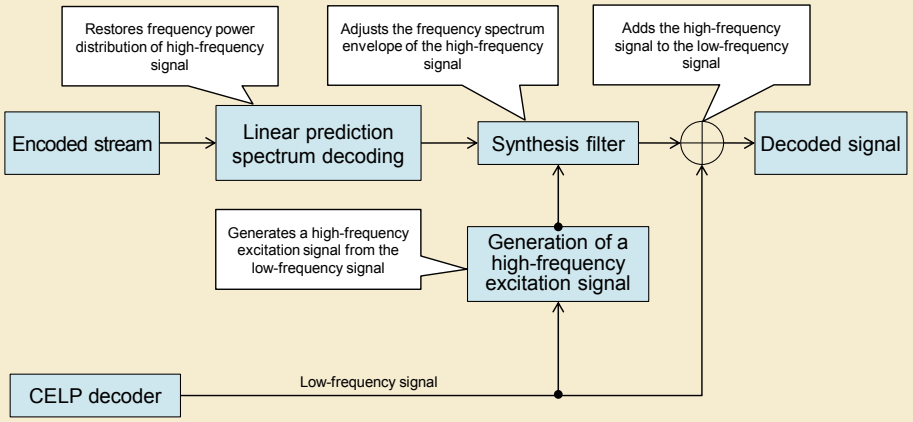


Figure 4 Time-domain bandwidth extension decoding

*20 **Inter-frame prediction:** A technique for improving coding efficiency by quantizing the difference between the values of the current frame and that of the previous frame.

*21 **Transition Coding mode:** A coding mode of Algebraic CELP (ACELP) designed to eliminate inter-frame dependency as much as possible and to control error propagation. The encoder is designed to select this mode in the frame following the frame that includes the onset of speech.

*22 **Linear prediction spectrum:** The frequency spectrum of an IIR filter determined by linear prediction coefficients.

*23 **Filter-bank:** A series of digital filters that split an input signal into multiple frequency bands.

higher-frequency components having a power distribution indicated by the encoder. This scheme enables the encoding of high-frequency components using only a limited amount of information with low computational complexity.

2) Arithmetic Coding

As described in section 2.1 (2), arithmetic coding is used for the quantization of the linear prediction residual in the frequency domain.

As shown in **Figure 5**, a bit plane is first created based on two adjacent MDCT coefficients in binary, and then higher-order bits are encoded with a codebook that is selected based on the immediately preceding results of quantization. Finally, lower-order bits are encoded according to the number of remaining bits available for use.

For higher bit rates in which a sufficient number of linear prediction residual signals can be encoded, the codebook is determined based on the linear prediction residual obtained by decoding in the previous frame. For lower bit rates in which a sufficient number of linear prediction residual signals cannot be obtained, the codebook is determined based on the linear prediction spectrum.

3) IGF

As shown in **Figure 6**, the IGF technique copies adjacent decoded MDCT coefficients to a missing frequency band that could not be encoded. However, if the MDCT coefficients at the copy source have strong peaks, those peaks will also be generated at upper frequencies at the copy destination resulting in degraded sound quality. To suppress such unnecessary strong peaks in the high-frequency band, whitening*24 is performed in

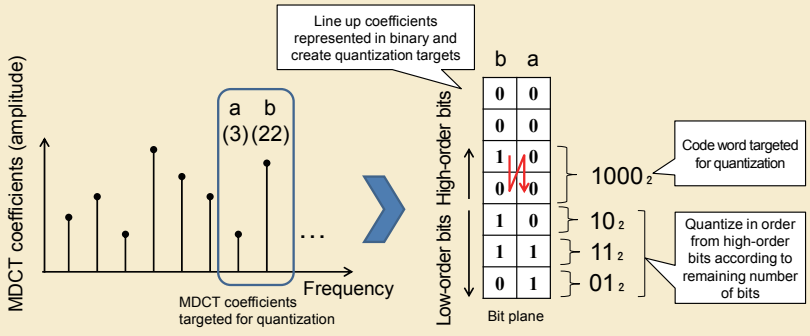


Figure 5 Arithmetic coding

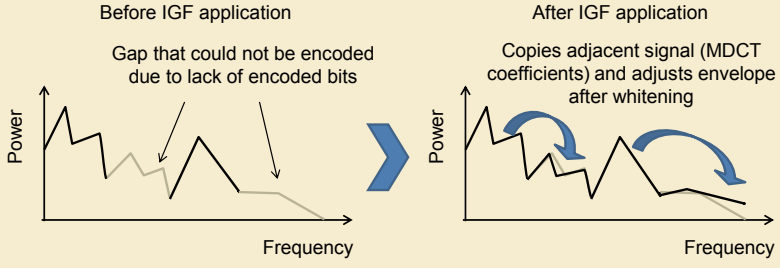


Figure 6 IGF

*24 **Whitening:** Processing for making the frequency distribution of signal power uniform.

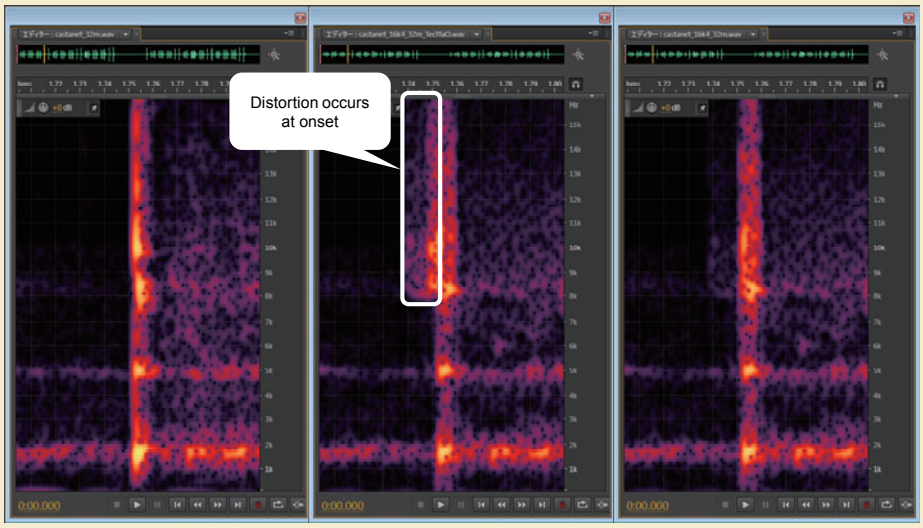
the frequency domain as needed.

On the encoder side, IGF encodes and transmits the rough shape of the frequency spectrum and ON/OFF information for whitening, and on the decoder side, it adjusts the MDCT coefficients to reconstruct the frequency spectrum. In this way, IGF can perform high-quality audio encoding especially for music at low bit rates.

2.3 Technical Contribution from NTT DOCOMO

1) Technology for Improving Higher-frequency Components

The encoding of high-frequency components in EVS is often done by bandwidth extension using low-frequency components for time-domain encoding and by IGF using components of another frequency band for frequency-domain encoding. Thus, a mismatch may occur in the power distribution in the temporal direction between the higher-frequency components of the input signal and the signal components of the frequency band that becomes the basis for encoding high-frequency components, or distortion may occur in the power distribution in the temporal direction when encoding high-frequency components. These issues are addressed by the following method. In the encoder, the method detects whether the power distribution in the temporal direction of the high-frequency components of the input signal is flat and transmits the result, and in the decoder, the method uses that result as a basis for performing applicable processing such as flattening of the high-frequency components. As a result, a match can be achieved with the power distribution of the input signal in the temporal direction. This flattening process is performed either in the time domain or frequency domain according to which coding strategy is selected. If time-domain coding is selected, the encoder transmits the result of detecting a sudden increase in power distribution in the temporal direction, and the decoder adjusts the increase accordingly to reconstruct that power fluctuation. The spectrograms in **Figure 7** (b) and (c) depict the temporal change in a signal's frequency spectrum before and after adjusting the power increase shown in fig. 7 (a) by this technology.



(a) Original sound (b) Before application (c) After application

Figure 7 Effect of technology for improving high-frequency components

In the figure, the vertical and horizontal axes represent frequency and time, respectively, and a change from dark to bright colors represents an increase in power. The results shown in the figure reveal that distortion occurs at high frequencies prior to the signal's sudden power increase before applying this technology and that this distortion is suppressed after applying the technology.

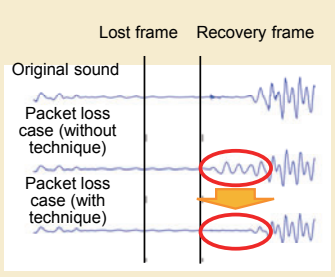
2) PLC Technique

In CELP, pitch lag estimation error caused by packet loss can result in audio discontinuities. To obtain linear prediction coefficients that minimize coding distortion, a portion of the frame following the frame targeted for encoding is exploited for linear predictive analysis as a look-ahead signal. Transmitting the pitch lag calculated for this look-ahead signal provides a pitch estimate for the next frame without adding extra delay and improves the accuracy of pitch-lag estimation under packet loss conditions.

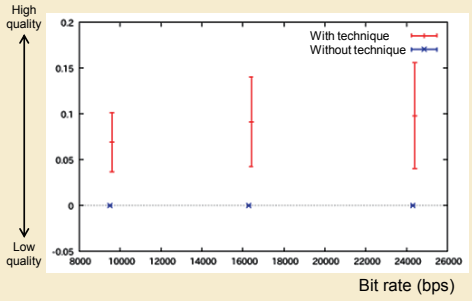
Additionally, since linear prediction coefficients are encoded based on inter-frame prediction in the EVS codec, the linear prediction filter can be unstable at the recovery frame after a packet loss especially at the onset where spectral energy of speech increases from zero. This sometimes causes a large ripple in the decoded signal. To deal with this problem, the linear prediction filter on the decoder side is modified based on auxiliary information from the encoder to prevent it from having excessive gain. In the encoder, frames at which the filter can be unstable are first detected by simulating a linear prediction filter at the lost frame, and the result is transmitted as auxiliary information. In the decoder, the linear prediction filter is modified based on those detection results transmitted as auxiliary information. This technique prevents perceptual quality degradation caused by ripples as shown in **Figure 8** (a). We evaluated the effectiveness of the technique by Perceptual Evaluation of Speech Quality (PESQ)*25 [9]. For this evaluation, we created an error pattern in which the onset of speech is lost. The difference in the scores between with and without the proposed technique is shown in Fig. 8 (b). The error bar*26 represents a 95% confidence interval. These results indicate a significant improvement in sound quality.

3) Technology for Reducing Computational Complexity

When switching between low-bit-rate and high-bit-rate operational modes, linear prediction coefficients need to be recomputed since the internal sampling rate changes. However, if the conventional method is employed for this purpose, a large amount of calculations will be needed. For this reason, linear prediction coefficients are computed here by performing resampling*27 in the frequency domain on the linear prediction spectrum. This approach reduces computa-



(a) Speech waveform on applying technique



(b) Difference in evaluation scores with/without applying the proposed technique

Figure 8 Effect of PLC technique proposed by NTT DOCOMO

*25 **PESQ:** A method of subjective evaluation that estimates speech quality from the difference between a reference signal and the signal being tested.

*26 **Error bar:** A graphical bar that indicates the range of error.

*27 **Resampling:** Performing sampling a second time using a different sampling frequency after returning the digital signal to an analog signal.

tional complexity to around one-third that of conventional values.

3. EVS Performance

3.1 EVS Selection Tests

In the EVS selection phase, subjective quality assessments [10] consisting of a total of 24 tests were conducted at three evaluation institutions with 32 subjects participating in each test. For super-wideband input, a Degradation Category Rating (DCR) test*28 that evaluates degradation from the original sound on five levels was performed multiple times. In the tests, clean speech, noisy speech, music, and clean speech under packet loss conditions were tested [11].

3.2 Evaluation Results

The results of these selection tests are partially shown in **Figure 9**. In the figure, each error bar represents a 95% confidence interval. First, for the tests targeting clean speech, EVS achieved a level of quality equivalent to or greater than reference codec G.722.1 Annex C*29 [12] even at half the bit rate.

For the tests targeting noisy speech, which simulates actual use case in real life, EVS at 13.2 kbps achieved a level of quality significantly higher than AMR-WB at its maximum bit rate of 23.85 kbps. This result indicates that sound quality significantly improves when migrating from AMR-WB to EVS. Similarly, test results revealed an improvement in performance when using EVS for music and under packet-loss conditions. In this regard, AMR-WB+*30 is a codec having a long delay of 80 ms, but EVS could nevertheless achieve an equivalent level of quality with a relatively short delay of 32 ms.

4. Conclusion

This article provided an overview of EVS, described the major technologies making it possible, and presented the results of subjective quality assessment performed in the EVS selection phase. EVS achieves a level of quality as high as FM-radio in telephony and is easy to implement in VoLTE. It is a codec that can achieve a high level of quality for both speech and music at low bit rates. These features can, of course, be leveraged to raise the sound quality of telephony and existing services that make use of music content as in Melody Call®, but they are also expected to give

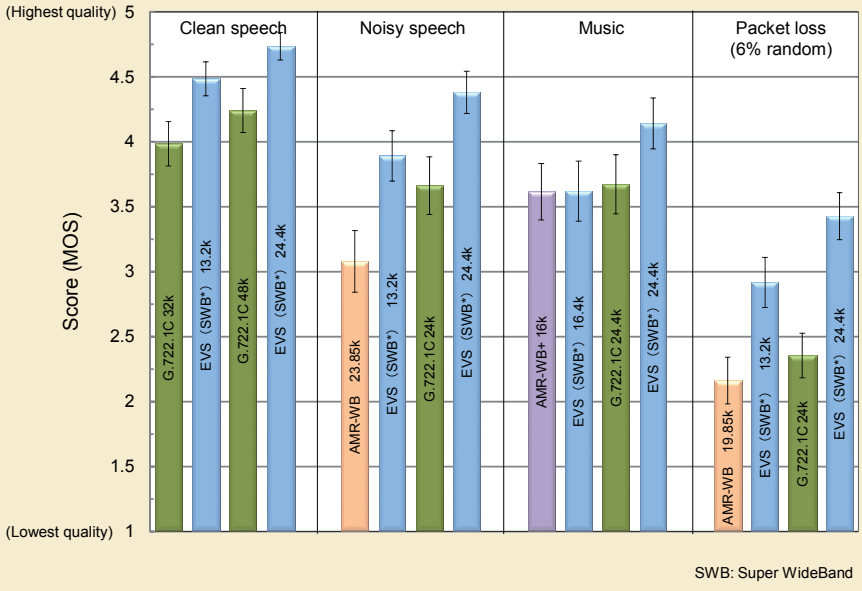


Figure 9 Results of subjective evaluation tests

*28 **DCR test:** A method of subjective evaluation that measures the extent to which the target signal is degraded with respect to a reference signal that represents base quality. A subject listens to both the reference signal and the target signal. This method is specified in ITU-T P.800.

*29 **G.722.1 Annex C:** A speech codec supporting super-wideband signals standardized by ITU-T and used in voice conferencing equipment from Polycom, Inc.

*30 **AMR-WB+:** An extended coding scheme of AMR-WB, the speech coding scheme standardized by 3GPP, which enables it to be used for general audio signals such as music.

birth to a new style of mobile audio communications.

References

[1] 3GPP TS26.441 V12.0.0: "Speech codec speech processing functions; Enhanced Voice Service (EVS) speech codec; General description," 2014.

[2] 3GPP TS26.171 V12.0.0: "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description," 2014.

[3] ISO/IEC 23003-3: "Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding," 2012.

[4] K. Kikuiri et al.: "MPEG Unified Speech and Audio Coding Enabling Efficient Coding of both Speech and Music," NTT DOCOMO Technical Journal, Vol.13, No. 3, pp. 17-22, Dec. 2011.

[5] 3GPP TR22.813 V10.0.0: "Study of Use Cases and Requirements for Enhanced Voice Codecs for the Evolved Packet System (EPS)," 2010.

[6] 3GPP TS26.071 V12.0.0: "Mandatory speech CODEC speech processing functions; AMR speech Codec; General description," 2014.

[7] Kaneko et al.: "VoLTE-compatible High-quality Melody Call," NTT DOCOMO Technical Journal, Vol. 22, No. 4, pp. 29-33, Jan. 2014. (in Japanese)

[8] 3GPP TS26.447 V12.0.0: "Codec for Enhanced Voice Services (EVS); Error Concealment of Lost Packets," 2014.

[9] ITU-T P.862.2: "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2007.

[10] ITU-T P.800: "Methods for subjective determination of transmission quality," 1996.

[11] 3GPP AHEVS-311: "EVS Permanent Document EVS-8b: Test plans for selection phase including lab task specification," 2014.

[12] ITU-T G.722.1: "Low complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," 2005.