Speech Coding | Super-Wideband | VoIP

Technology Reports

High-quality Speech Coding

Kei Kikuiri, Nobuhiko Naka and Shinya Abe

The high-quality speech coding is a super-wideband speech coding technique for providing high-quality VoIP services in a high-speed wireless communication environment, and was developed in collaboration with DoCoMo Communications Laboratories USA, Inc. This technology encodes and transmits nearly all of the frequency components of the human voice to allow easily-understood communication with a sense of presence.

1. Introduction

Research topics in speech coding technologies have shifted from encoding algorithms for narrowband (300 Hz to 3.4 kHz) speech signals at around 10 kbit/s to encoding algorithms for speech signals having wider bandwidth at about 20 to 64 kbit/s. Current narrowband speech codecs, such as Adaptive Multi-Rate (AMR)[*1][1], provide speech quality that is almost equivalent to the legacy telephony networks, thus no further improvement can be expected. In addition, a high compression ratio for narrowband speech signals is less important in VoIP services where high-speed packet access lines are typically available. For perceivable improvement in speech quality, i.e., better than that of the legacy telephony services, it is necessary to extend the speech bandwidth. Transmitting the low- and high-frequency components that are not included in narrowband speech produces decoded speech signal that is close to the original and sounds more natural. Actually, a number of VoIP services adopt the wideband (50 Hz to 7 kHz) or super-wideband (upper frequency is more than 7 kHz) speech codecs that have previously been used mainly for video telephony.

In mobile communication, wireless VoIP services including PASSAGE DUPLE and Smartphones[*2] such as the hTc Z terminal have appeared on the market, and high-speed wireless packet access networks using Super 3G[*3] and Fourth-Generation mobile communication systems are also underway. These conditions show that speech communication with wideband or super-wideband codecs will be feasible in mobile environments.

We propose a high-quality speech coding technique that was developed in collaboration with DoCoMo Communications Laboratories USA, Inc. This is a super-wideband speech codec that is intended to provide better speech quality than wideband speech codecs for future VoIP services via high-speed wireless packet access networks. This codec handles speech signals with upper frequencies of from 10 to 16 kHz, which is three times higher than that of narrowband, and encodes the signal at 48 to 64 kbit/s. This codec also requires computational complexity comparable with a conventional narrowband speech codec.

This technology will enable realistic conversation over mobile terminals with speech quality close to natural human voice. This advantage will be especially helpful in mobile applications where voice quality is particularly critical, such as teleconferencing and remote education.

In this article, we present an overview of the developed high-quality speech coding algorithm and quality evaluation results. We also introduce mobile VoIP prototype software that implements this coding technology.

2. High-quality Speech Coding

2.1 Overview of Speech Coding for VoIP

The specifications of speech codecs used in VoIP services and applications are shown in **Table 1**. G.711[*4][2] is the speech coding technique used in the fixed

*1 **AMR**: A mandatory speech coding for Third-Generation mobile communication defined by 3GPP. It allows flexible variation of the transmission rate according the type and condition of networks.

*2 **Smartphone**: A mobile terminal equipped with the functions of a mobile data terminal.

*3 **Super 3G**: A high-speed wireless access system that is an extended Third-Generation mobile communication system. Standardization by 3GPP is in progress as Long Term Evolution (LTE).

*4 **G.711**: 64-kbit/s PCM telephone band speech coding recommended by International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) (See *5).

Table 1 VoIP codec specifications

	G.711	G.729 Annex A	G.722.1	G.722.1 Annex C	AAC-LD
Sampling frequency (kHz)	8	8	16	32	32 to 48
Bit rate (kbit/s)	64	8	24, 32	24, 32, 48	32 to 64
Frame length (ms)	—	10	20	20	10 to 16
Algorithmic delay (ms)	0.125	15	40	40	20 to 32
Speech bandwidth	Narrowband		Wideband	Super-wideband	
Coding algorithm	PCM	CELP	Transform coding		

network as well as many VoIP services. Because it uses the Pulse-Code Modulation (PCM)[*5] algorithm, the algorithmic delay[*6] is only 0.125 ms (1 sample). A packet loss concealment[*7] algorithm has also been recommended together with G.711, which is robust against errors such as packet loss. The G.729 Annex A[*8] (G.729A)[3] codec is a low-complexity version of the G.729[*8][4] used in NTT DoCoMo's Hypertalk®[*9] service. Like other speech codecs for mobile communication, it adopts the Code Excited Linear Prediction (CELP)[*10] algorithm to model the human voice production mechanism, and achieves speech quality close to that of the fixed network at 8 kbit/s. NTT DoCoMo's PASSAGE DUPLE uses G.711 and G.729A. G.722.1[*11][5] is a wideband speech codec that is mainly used in video telephony, but has recently come into use in VoIP applications as well. Its extended version, G.722.1 Annex C[*11](G.722.1C)[5] is designed for 14-kHz bandwidth speech signals. In comparison with other super-wideband speech coding, it allows a lower bit rate and lower computational complexity, although it has a longer algorithmic delay of 40 ms.

Advanced Audio Coding-Low Delay (AAC-LD)[*12][6], which is a low-delay version of the AAC[*12][6] codec used in music distribution, is applicable for two-way communication. It requires a high computational complexity for a precise modeling of the auditory property[*13] used in audio codecs. G.722.1 and AAC-LD adopt a transform coding algorithm that converts the time domain signal into a frequency domain representation. This algorithm does not use a voice production model, so it handles audio signals other than speech.

2.2 High-quality Speech Coding Specifications

The specifications of the high-quality speech coding we developed are shown in **Table 2**. By setting a sampling frequency of 22.05 kHz or higher, this technique extends the upper frequency limit for speech coding from 10 to 16 kHz, which covers nearly all of the frequency components of human voice. In addition, the bit rate can be set freely between 48 and 64 kbit/s corresponding to the types of networks and applications. The values of frame length and algorithmic delay in the table are the respective values for operation at 22.05 kHz and 32 kHz.

This technique is based on a transform coding algorithm like G.722.1 and AAC-LD (**Figure 1**). In the encoder, the Modified Discrete Cosine Transform (MDCT) is used to convert the time domain speech signal into frequency spectrum coefficients (transform coefficients). The MDCT is a lapped transform over-lapping with the adjacent transform blocks and prevents distortion at the block boundary without any redundant data. This technique adopts 256-sample MDCT (8 ms sampled at 32 kHz) to achieve comparable algorithmic delay to G.729A and other

Table 2 High-quality speech coding specifications

Sampling frequency (kHz)	22.05	32
Bit rate (kbit/s)	48 to 64	
Frame length (ms)	11.61	8
Algorithmic delay (ms)	23.22	16
Speech bandwidth	Super-wideband	
Coding algorithm	Transform coding	

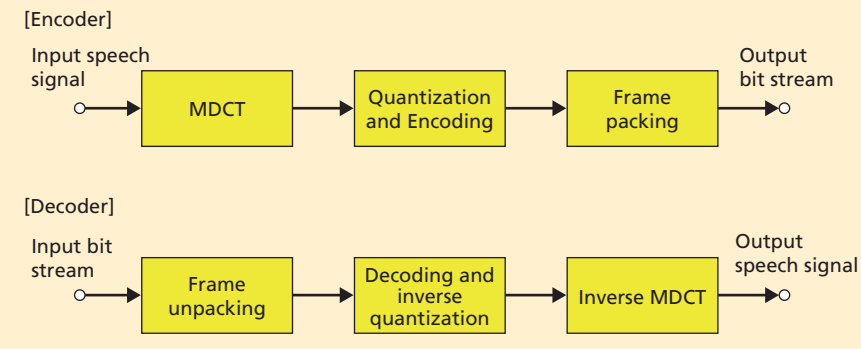

Figure 1 Basic structure of the high quality speech coding

*5 **PCM**: A coding algorithm in which the signal amplitude sampled at the sampling frequency is expressed as a binary number.
*6 **Algorithmic delay**: The time required for processing by the coding algorithm regardless of hardware performance of computation and transmission.
*7 **Packet loss concealment**: A function for concealing the speech distortion caused by packet loss.
*8 **G.729****: The 8-kbit/s narrowband CELP speech coding recommended by ITU-T (See *10). A compatible low-complexity version is defined as Annex A.
*9 **Hypertalk®**: Registered trademark of NTT DoCoMo.
*10 **CELP**: A speech coding algorithm that compares the input speech to a signal synthesized by standard patterns and transmits the indices of the pattern generating the closest synthetic signal.

such conventional speech coding techniques and to suppress pre-echo[*14], a distortion that is specific to the transform coding algorithm. Next, it quantizes and encodes the transform coefficients weighted according to the human auditory property at low computational complexity. Finally, the encoded data are packed frame by frame (frame packing) and sent to the transmission channel. It is also possible to introduce a super-frame structure which improves coding efficiency by using the correlation between frames, and packet loss concealment that utilizes information from the preceding and succeeding frames.

The decoder on the receiving side unpacks the received frame information to encoded data (frame unpacking), and then, after decoding and inverse quantization of the transform coefficients, the time domain speech signal is reproduced by the inverse MDCT. This technology does not involve the predictive processing across multiple packets used in the CELP algorithms, thus one of its features is to suppress the degradation of speech quality instantaneously due to packet loss.

2.3 Verification of the Subjective Speech Quality of the High-quality Speech Coding

To verify the subjective speech quality of the high-quality speech coding, we conducted subjective evaluation experiments. The test conditions are shown in **Table 3**. In the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)[7] method, the subjects are presented with a reference signal (original), the signals to be evaluated, the hidden version of the reference signal, and the band-limited signals (anchors). And they are asked to assign a score from 0 to 100 to the signals to be evaluated, the hidden reference and the anchors, in comparison with 100 points for the reference signal. Here, it can be said that the results of 7 kHz band-limited and 3.5 kHz band-limited speech show the respective upper limits of subjective quality for the wideband speech coding and the narrowband speech coding.

The subjective evaluation test results are shown in **Figure 2**. The error ranges in the figure are for the 95% confidence interval. The statistical tests of significance[*15] show that all of the encoded speech including the high-quality speech coding are not good compared to the original speech for the female speech, male speech and speech with BGM samples, but the subjective quality of the high-quality speech coding is equivalent to that of AAC-LC (Low Complexity)[*12][6] at 64 kbit/s and that of G.722.1C at 48 kbit/s. The AAC-LC generally has higher coding efficiency than AAC-LD, thus can be said to achieve the equivalent or better subjective quality as that of AAC-LD at the same bit rate. Therefore, the high-quality speech coding can be said to offer the equivalent subjective quality to AAC-LD at 64 kbit/s. Furthermore, the high-quality speech coding has significantly better subjective quality than the 7 kHz and the 3.5 kHz band-limited speech at 64 kbit/s and 48 kbit/s respectively.

The experiment results show that the high-quality speech coding achieves comparable speech quality to the international standard super-wideband speech codec and audio codec, and it outperforms conventional narrowband and wideband speech codecs.

3. High-quality VoIP Prototype Software

Targeting high-quality VoIP services via a mobile terminal, we implemented a high-quality speech coding module that is capable of real-time encoding and decoding on Windows Mobile®[*16] 5.0. We also created high-quality VoIP prototype software with that module running on the hTc Z terminal (**Photo 1**). For comparison with the speech quality of ordinary telephony, this software is also equipped with G.711. Real-time Transport Protocol (RTP)[*17][8] is used for transmission of the speech data

Table 3 Experimental conditions

Test method	MUSHRA
Number of subjects	19
Reference speech (sampling frequency)	Original speech (32 kHz)
Encoded speech (bit rate/sampling frequency)	High-quality speech coding (48, 64 kbit/s, 22.05 kHz) AAC-LC (64 kbit/s, 48 kHz) G.722.1 Annex C (48 kbit/s, 32 kHz)
Band-limited speech	7-kHz bandwidth, 3.5-kHz bandwidth
Listening system	Headphones (both ears)

*11 **G.722.1****: An audio coding recommended by ITU-T. A fixed-point implementation of the 7-kHz bandwidth and the 14-kHz bandwidth modes is defined as Annex C.

*12 **AAC****: An audio coding specified by International Organization for Standardization/International Electrotechnical Commission (ISO/IEC). AAC-LC is a profile that reduces the computational complexity of the main profile. A low-delay extension that allows two-way communication is also specified as AAC-LD.

*13 **Auditory property**: The property that it is difficult to perceive the noise in a frequency region near the region in which the input signal power spectrum has large power.

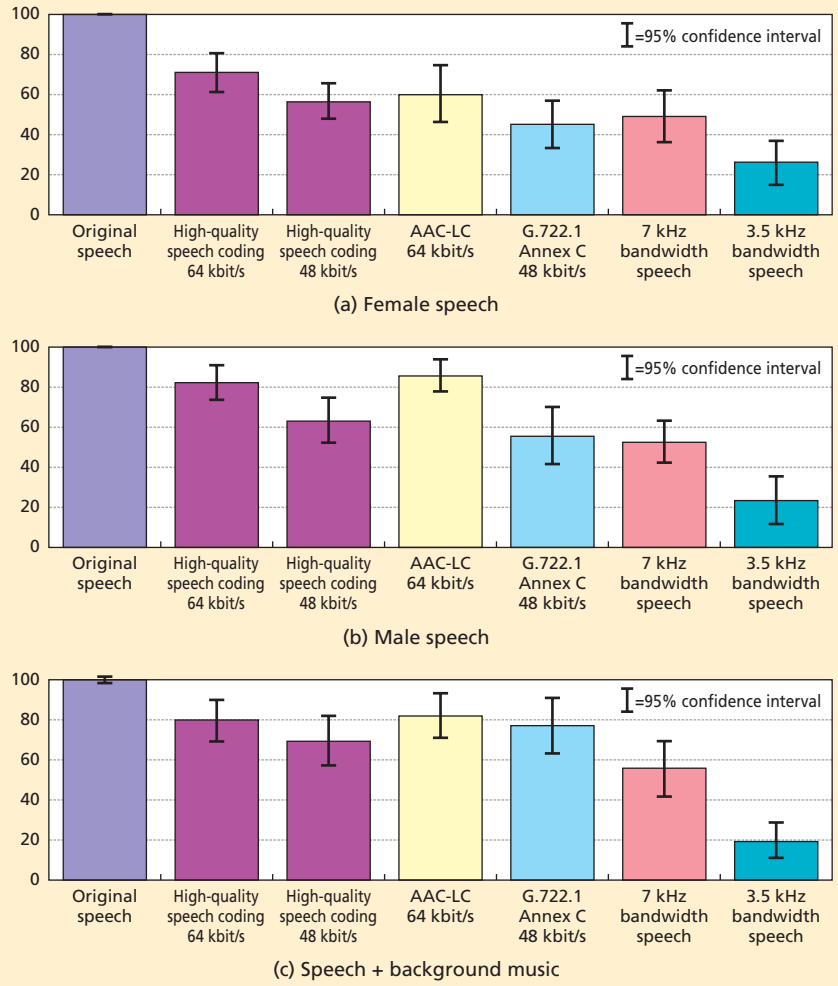

Figure 2 Subjective evaluation results

Photo 1 Display example of high-quality VoIP prototype software

and call control is done with Session Initiation Protocol (SIP) [*18][9].

4. Conclusion

We presented high-quality speech coding technology that enables super-wideband speech communication services. Subjective evaluation results confirmed that the high-quality speech coding shows better speech quality than that of wideband speech, and equivalent quality to that of existing super-wideband speech codecs. We also introduced VoIP prototype software that adopts this technology.

This technology can realize more natural conversations over the mobile terminals by extending the speech bandwidth and is expected to enhance future speech communication services.

For future work, we plan to develop additional functionality for the entire VoIP system while considering requirements and objectives of actual VoIP services and to investigate technologies for new applications that use the advantages of super-wideband speech.

References

[1] 3GPP TS26.090: "Adaptive Multi-Rate (AMR) speech codec; Transcoding functions," 1999.
[2] ITU-T Recommendation G.711: "Pulse Code Modulation (PCM) of Voice Frequencies," 1988.
[3] ITU-T Recommendation G.729 Annex A: "Coding of Speech at 8 kbit/s using conjugate structure algebraic-code-excited linear prediction (CS-ACELP) Annex A: Reduced complexity 8 kbit/s CS-ACELP speech codec," 1996.
[4] ITU-T Recommendation G.729: "Coding of Speech at 8 kbit/s using conjugate structure algebraic-code-excited linear prediction (CS-ACELP)," 1996.
[5] ITU-T Recommendation G.722.1: "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," 2005.
[6] ISO/IEC 14496-3: 2001: "Information technology—Coding of audio-visual objects—Part 3: Audio," 2001.
[7] ITU-R Recommendation BS.1534-1: "Method for the subjective assessment of intermediate quality level of coding systems," 2003.
[8] IETF RFC3261: "SIP: Session Initiation Protocol," 2002.
[9] IETF RFC1889: "RTP: A Transport Protocol for Real-Time Applications," 1996.

*14 **Pre-echo**: A phenomenon in which a frequency domain quantization error just prior to an onset of attack in audio signal is perceived as an echo-like distortion.
*15 **Statistical test of significance**: A method used to determine whether or not the difference between two values is statistically significant. If the difference between two compared values is within the range of the confidence interval calculated from the dispersion in the corresponding values, there is no statistically significant difference; otherwise, the difference is statistically significant.
*16 **Windows Mobile®**: Registered trademark of the Microsoft Corporation in the United States and other countries.
*17 **RTP**: A real-time multimedia transport protocol via IP networks defined by the Internet Engineering Task Force (IETF).