Availability of Artificial Voice for Measuring Objective QoS of CELP CODECs and Acoustic Echo Cancellers

Nobuhiko Kitawaki*, Feng Wei*, Takeshi Yamada*, and Futoshi Asano**

University of Tsukuba* and AIST**, Japan
1-1-1, Tennoudai, Tsukuba-shi, 305-8573 Japan.
Tel. & Fax. +81 298 53 5526, E-mail: kitawaki@is.tsukuba.ac.jp

Topic: 2.1 Current standards comparison and evaluation
Key Words: artificial voice, Rec. P.50, CELP CODEC, echo canceller, PSQM, PESQ, Rec. P.861, P.862

Abstract

We proposed an artificial voice reflecting average characteristics of comprehensive human voice as the test signal instead of a real voice for objective quality measurement of coded speech. The artificial voice was standardized as ITU-T Recommendation P.50 in 1988. However, the performance of artificial voice was verified to the waveform codecs using an LPC cepstrum distance measure as a criterion.

Recently, there has been much progress concerning speech coding techniques based on CELP (Code Excited Linear Prediction). On the other hand, there has also been much progress concerning objective quality measures for speech coding. Those are PSQM (Perceptual Speech Quality Measure) and PESQ (Perceptual Evaluation of Speech Quality) standardized as P.861 in 1996 and P.862 in 2000, respectively.

This paper describes verification test results of the artificial voice P.50 by using PSQM method P.861 for objective QoS measurement of CELP codecs. To expand the application area for P.50, this paper also describes objective QoS measurements by the artificial voice P.50 for acoustic echo cancellers in hands-free communications .

1. Introduction

We proposed an objective QoS measurement scheme composed of objective measure and test speech signal for voiceband codecs in 1982 [1]. However, because of less information the coded speech quality at a low-bit-rate depends on the talker. The real speech signal should be selected taking account of talker dependency. In practice, only an insufficient number of speech signals were used for the objective quality measurement. Therefore, we proposed another approach which uses an artificial voice reflecting average characteristics of comprehensive human voice as the test signal instead of a real voice shown in Fig. 1 [2]. The artificial voice was standardized as ITU-T Recommendation P.50 in 1988 [3]. However, at that time, the performance of artificial voice was only verified to the waveform codecs using an LPC cepstrum distance measure as a criterion.

Recently, there has been much progress concerning speech coding techniques [4]. The CELP (Code Excited Linear Prediction) codecs have already been used in wireless personal communications and internet communications shown in Table 1, and are expected to be widely used in the third generation wireless personal multimedia communications and the broadband networks. On the other hand, there has also been much progress concerning objective quality measures for speech coding. The PSQM (Perceptual Speech Quality Measure) and PESQ (Perceptual Evaluation of Speech Quality) based on internal representations of the speech signals which make use of the psychophysical equivalents of frequency and intensity domains were standardized as Recommendations P.861 in 1996 and P.862 in 2000, respectively [5][6].

This paper describes verification test results of the artificial voice P.50 by using PSQM method P.861 for objective QoS measurement of CELP codecs. To expand the application area for P.50, this paper also describes objective QoS measurements by the artificial voice P.50 for acoustic echo cancellers in hands-free communications [7].

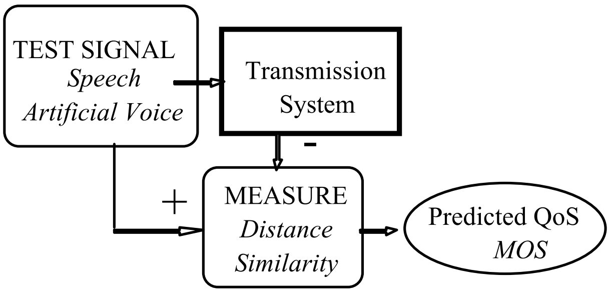


Fig.1 Objective QoS measuring scheme composed of an artificial voice and a criterion.

Table 1 G-series speech coding standardized in ITU.

Rec.	Coding	Notes	Year
G.711	PCM	64 kbit/s	1972
G.722	SB-ADPCM	7 kHz, 64 kb/s	1988
G.723.1	ACELP MP-MLQ	dual rate 5.3 kbit/s, 6.3 kbit/s	1996
G.726	ADPCM	40, 32, 24, 16 kbit/s	1984
G.727	Em-ADPCM	5, 4, 3, 2 bit/sample	1990
G.728	LD-CELP	16 kbit/s	1992
G.729	CS-ACELP	8 kbit/s	1996

2. Artificial Voice

Figure 2 shows a generation diagram of the artificial voice recommended by P.50. The artificial voice generation method employs vector quantization to analyze a multilingual data-base composed of twenty languages and PARCOR speech synthesis. The artificial voice reflects average characteristics of comprehensive human voice: (a) long-term averaged spectra, (b) instantaneous amplitude distribution, (c) level distribution of segmental power, (d) spectral distribution of segmental power, and (e) voiced/unvoiced structure of speech waveform. Consequently, the artificial voice at most 10 seconds long can be expected to cause the same effects on objective speech quality measurement as the use of a large amount of real speech samples, and to generate the test signal by using the defined algorithm at any location.

3. Objective QoS Measurement of CELP CODECs

3.1 Experimental Conditions

To thoroughly investigate the performance of artificial voice for CELP codecs by using PSQM measure, the PSQM values derived using the artificial voice should be compared with those of real speech for various CELP codecs. The codecs were G.729, G.726, G.728, G.711, GSM-FR, IS-54, JDC-HR, and their tandeming connections shown in Table 2. In our investigations, we used seven different waveform/CELP codecs with bitrates from 5.6 to 64 kbit/s. From the viewpoints of investigating the basic performance of the artificial voice and objective quality measure, we took into account the effects of languages, talkers, and tandeming of the codecs. Other factors were fixed. That is, the number of all conditions are 29.

Real speech samples are same Japanese, English, and Italian as those used in the verification tests for standardizaion of P.861. Each spoken language is composed of two females and two males. Artificial voices are a female and a male. The input level to a codec was set to –26 dBov relative value to the overload level of linear PCM. We assumed there was no channel degradation between a coder and a decoder.

The objective quality measure is PSQM based on internal representations of the speech signal which make use of the psychophysical equivalents of frequency and intensity domains standardized as Recommendations P.861.

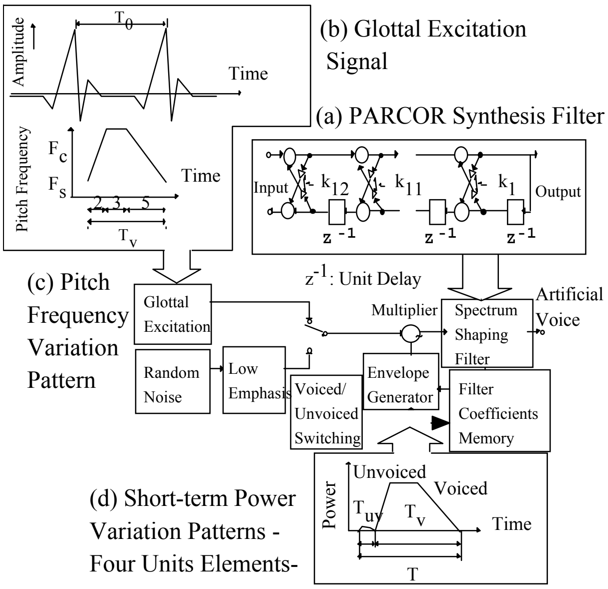


Fig. 2 Generation of the artificial voice (AV).

Table 2 CODEC used in experiment.

Rec.	Speech Coding
G.729	8 kb/s CS-ACELP
G.726	32 kb/s ADPCM
G.728	16 kb/s LD-CELP
G.711	64 kb/s PCM
GSM-FR	Full Rate GSM
IS-54	North America VSELP
JDC-HR	5.6 kb/s PSI-CELP

3.2 Performance Index

Conventionally, the performance of an artificial voice is evaluated in terms of the consistency between the objectively estimated value (PSQM) calculated by the artificial voice and that of real speech. The evaluation is done with performance indexes such as correlation coefficients and root mean square error (RMSE).

3.3 Performance Evaluation of Artificial Voice

To verify the performance of artificial voice, this section compares the objective values derived from using the artificial voice with those of real speech for various CELP and waveform codecs. Table 3 shows the RMSE and correlation coefficients of PSQM by the consistency of artificial voice and real speech. Figure 3 shows an example of male voice.

It was shown that the objective measurement PSQM for CELP and waveform codecs by artificial voice P.50 closely approximated that of real speech.

Table 3 RMSE and correlation coefficient.

Condition	Talker	RMSE	Correlation
Single CODEC	male	0.1031	0.9976
	female	0.5366	0.8965
Tandeming	male	0.3174	0.9548
	female	0.3681	0.8524
Total	male	0.3141	0.9696
	female	0.4317	0.8695

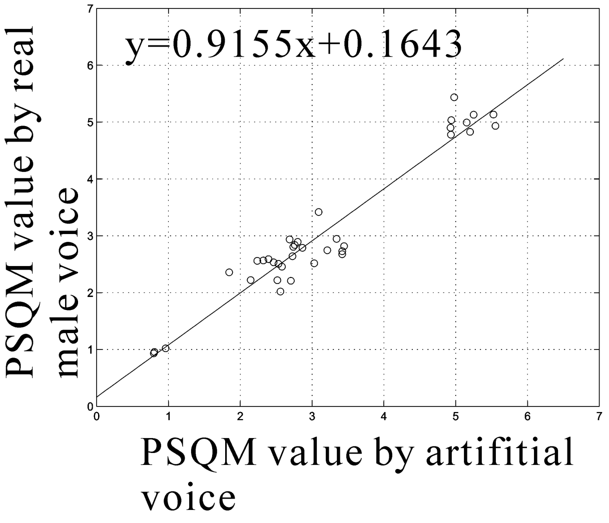


Fig. 3 An example of male voice.

4. Objective QoS Measurement of Acoustic Echo Cancellers

4.1 Application Expanding of Artificial Voice

Performance of the echo canceller is evaluated by residual echo characteristics expressed in echo return loss enhancement (ERLE). The ERLE can be conventionally measured by putting white Gaussian noise into the echo canceller system. However, white Gaussian noise is not adequate as the test signal for measuring the performance of the echo canceller, since the performance may depend on the characteristics of input test signal, and the characteristics of the white Gaussian noise differ from those of real spoken language. Therefore, this section discusses appropriate characteristics of spoken language required for objective quality evaluation of echo cancellers to expand the application area of the artificial voice P.50.

4.2 Echo Cancellers used in Verification Test

Figure 4 shows schematic diagram for measuring residual echo characteristics. Test signal is input to the telephone network system equipped with echo canceller, and emitted to the room from loudspeaker at the receiving end. Acoustic echo and ambient noise are input to the microphone at the receiving room. Adaptive filter of the echo canceller generates echo replica by estimating impulse response in the acoustic echo path and cancels a room acoustic echo. Residual echo characteristics are expressed in ERLE.

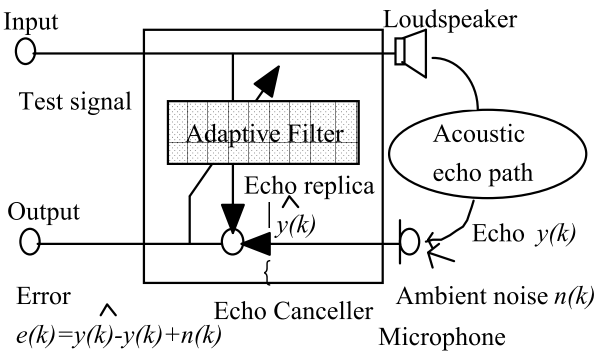


Fig. 4 Measurement of residual echo cancellers.

Table 4 shows echo cancellers used in this verification test. They are EC-A ("Natural Talk PRO") and EC-B ("Smooth Talk 200") produced by NTT (Nippon Telegraph and Telephone Corporation). The algorithm of echo canceller is based on exponentially weighted step-size projection (ES-Projection) method [8][9]. This algorithm uses a different step size for each coefficient of an adaptive transversal filter, and these step sizes are time-invariant and weighted in proportion to the expected variation of a room impulse response. For speech input, ES-Projection method has a convergence speed four times that of the normalized LMS (NLMS) but its computational load is almost the same.

Table 4 Acoustic echo cancellers used in tests.

	EC-A	EC-B
Pass Band	0.1-7.3 kHz	0.1-3.6 kHz
Algorithm	ES-Projection	ES-Projection
Eli. Time	250 ms	100 ms
Eli. Mag.	35 dB	Unkwon

4.3 Test Signals

Table 5 shows test signals studied in this experiment. To study of the characteristics of spoken languages required for objective quality measurement of echo cancellers, following five test signals are examined: real voice (RV), white Gaussian noise (WN), frequency weighted Gaussian noise (FWN), artificial voice (AV), and composite source signal (CSS).

WN having flat frequency spectrum is conventional test signal for echo canceller performance. FWN reflects the long-term average spectra of the spoken language to the white Gaussian noise. AV reflects the average characteristics of the spoken language such as long-term average spectra, instantaneous amplitude distribution, level distribution of segmental power, spectral distribution of segmental power, voiced/unvoiced structure of speech waveform, and short-term spectral characteristics. CSS is composed of above artificial voice (AV) sequence in the voiced intervals, white Gaussian noise (WN) sequence, and silent sequences, mixed in random. RV is Japanese short sentences, and is thought as reference criterion for performance evaluation of each test signal.

AV is recommended as P.50 used for measuring coded speech performances [3]. CSS specified by P.501 was proposed for measuring performance of telephonometry devices [10], and FWN specified by P.64 is used for measuring Loudness Rating (LR) in ITU-T [11]. Real voices composed of Japanese short sentences spoken by 6 males and 5 females.

4.4 Measurement of Residual Echo Characteristics

ERLE is defined as logarithmic power ratio of the output from the loudspeaker and residual echo. The loudspeaker and microphone arrangement was set up conforming to ITU-T Recommendation P.340 which specifies acoustic measurement of hands-free telephone system. Sound level from the loudspeaker was 73 dB, and ambient noise level was 45 dB.

Table 5 Test signals studied in this experiment.

Test Signal	Symbol	Rec.	Note
Real Voice	RV		reference
White Noise	WN		conventional
Freq. Weighted N.	FWN	P.64	long-term spect.
Artificial Voice	AV	P.50	time and freq.
Comp. Source S.	CSS	P.501	AV & WN

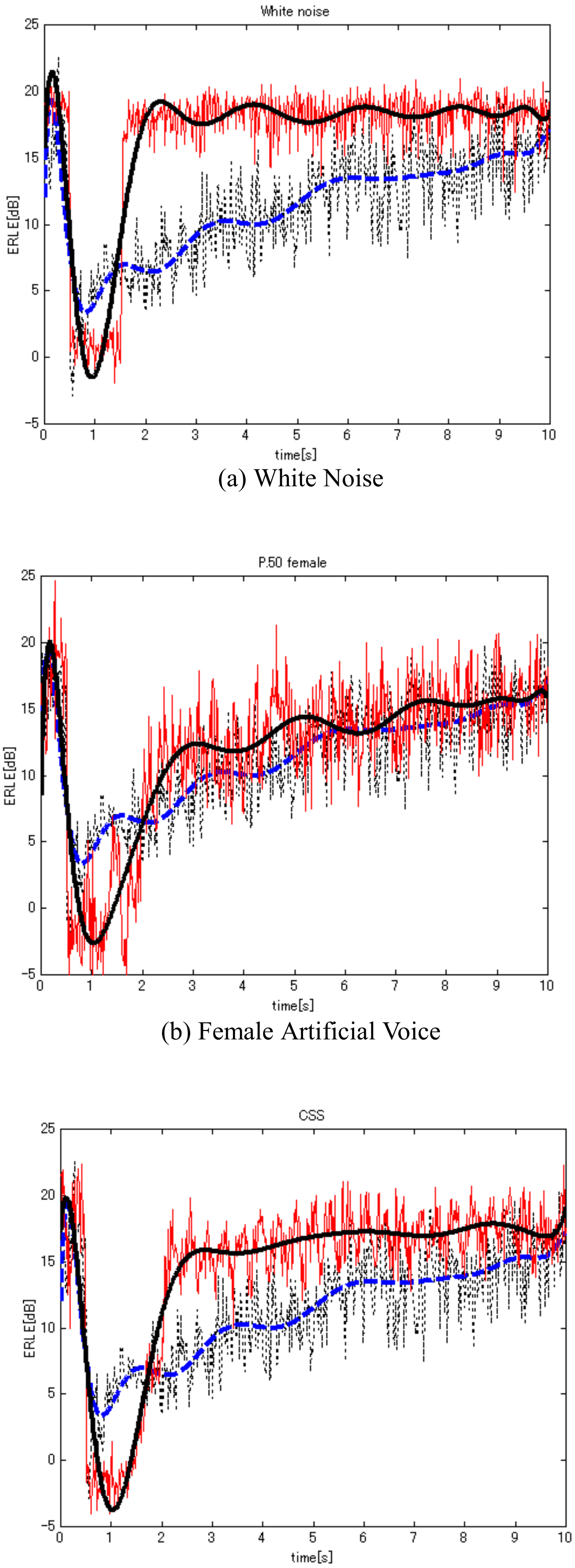
ERLE characteristics were measured by five test signals, and compared with those of RV as the reference criterion. Figure 5 (a)-(c) shows an example of ERLE characteristics of each test signal and that of RV for echo canceller EC-A. In the Figure, the dark colored curves are for real voices (RV) as the reference, and the light colored curves are for each test signal other than real voices. Smooth curves approximate each ERLE characteristics, in which dotted curve shows for RV, and solid curve shows for test signal.

ERLE characteristics measured by artificial voice according to P.50 are almost equivalent to those of real voices. However, ERLE characteristics measured by other test signal than P.50 differ from those of real voices, and consequently ERLE characteristics are very rapidly recovered compared with those of real voices. This may cause over-estimation for the echo canceller system if CSS, white Gaussian noise, and frequency weighted Gaussian noise are used as a test signal.

4.5 Performance Index

The performance of a test signal is evaluated in terms of the consistency between the approximate curve for each test signal and that for real voice. The evaluation is done with performance index as root mean square error (RMSE). Table 6 shows RMSE for each test signal. It is concluded that female and male artificial voices reveal best performance among test signals studied in this experiment.

Another performance index is convergence time. In this paper, convergence time is defined as the time to reach ERLE of 15 dB from the least value. Table 7 shows convergence time of the echo canceller system by each test signal. It is concluded that convergence time measured by artificial voice is the nearest to that of real voice.



(a) White Noise

(b) Female Artificial Voice

© Composite Source Signal

Fig. 5 An example of ERLE characteristics of each test signal and real voice for EC-A.

Table 6 RMSE between test signal and real voice.

Test Signal	EC-A	EC-B
WN	39.8	72.1
FWN	38.0	56.9
AV-Male	17.0	17.0
AV-Female	17.9	17.7
CSS	30.2	40.4

Table 7 Convergence time by each test signal.

•]‰¿–pM †	EC-A	EC-B
RV	8.7	impossible
WN	1.7	2.2
FWN	3.8	3.9
AV-Male	7.3	impossible
AV-Female	7.2	impossible
CSS	3.6	4.4

5. Conclusion

This paper verified the availability of the artificial voice recommended by ITU-T Rec. P.50 for both objective QoS measurements of CELP codecs by PSQM and acoustic echo cancellers. It is concluded that artificial voice having average characteristics of spoken language in time and frequency domain are required for and satisfied with objective QoS measurement of CELP codec and acoustic echo canceller. Therefore, we propose that current recommendations should be specified on this point.

Acknowledgments

We would like to thank Ms. Mari Saito, Ms. Keiko Kawaguchi, Mr. Yasuhiro Hotta, University of Tsukuba, and Ms. Junko Sasaki, NTT-AT, for their cooperation to this work..

Reference

[1] Nobuhiko Kitawaki, Kenzo Itoh, Masaaki Honda, and Kazuhiko Kakehi: COMPARISON OF OBJECTIVE SPEECH QUALITY MEASURES FOR VOICEBAND

CODECS, IEEE Int. Conf. Acoust. Speech, Signal Processing, ICASSP'82, pp. S9.5.1-9.5.4, 1982.

[2] Kenzo Itoh, Nobuhiko Kitawaki, Hiromi Nagabuchi, and Hiroshi Irii: A New Artificial Speech Signal for Objective Quality Evaluation of Speech Coding Systems, IEEE Trans. Commun., vol.42, no.2/3/4, pp.664-672, 1994.

[3] ITU-T Recommendation P.50: Artificial voices, 1993.

[4] Nobuhiko Kitawaki : Perceptual QoS Assessment for Wireless Personal Communications, IEEE The Fourth International Symposium on Wireless Personal Multimedia Communications, Proc. WPMC'01, pp. 541-546, September 2001.

[5] ITU-T Recommendation P.861: Objective quality measurement of telephone-band (300-3400 Hz) speech codecs, 1998.

[6] ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[7] Nobuhiko Kitawaki, Takeshi Yamada, and Futoshi Asano: Objective QoS Measurement for Hands-Free Telecommunications, 2001 Asia-Pacific Symposium on Information and Telecommunication Technologies, Proc. APSITT'01, November 2001.

[8] Shoji Makino, Yutaka Kaneda and Nobuo Koizumi: Exponentially weighted step-size NLMS adaptive filter based on the statistics of a room impulse response, IEEE Trans. Speech & Audio, vol. 1, no. 1, 1993.

[9] Shoji Makino and Yutaka Kaneda: Exponentially weighted step-size projection algorithm for acoustic echo cancellers, Trans. IEICE Japan, E75-A, pp. 1500-1508, Nov. 1992.

[10] ITU-T Recommendation P.501: Test signals for use in telephonometry on CD-ROM, 1996.

[11] ITU-T Recommendation P.64: Determination of sensitivity/frequency characteristics of local telephone systems, 1997.