PERFORMANCE EVALUATION OF OBJECTIVE QUALITY MEASURES FOR CODED SPEECH

Akira TAKAHASHI, and Nobuhiko KITAWAKI (NTT) TOKYO, JAPAN*

Paolino USAI (CSELT) TORINO, ITALY

David ATKINSON (NTIA) BOULDER, UNITED STATES

* NTT Multimedia Networks Laboratories, NTT R&D Center RM2-212A

3-9-11 Midori-cho, Musashino, Tokyo 180 JAPAN

Tel.: +81-422-59-4447, FAX: +81-422-59-2330, E-mail: aki-tak@ntttqn.tnl.ntt.co.jp

ABSTRACT

A performance evaluation method for objective measures estimating subjective quality of coded speech is proposed and applied the comparison of existing objective quality measures. The measure based on Bark spectrum distortion performs the best. Comparing its estimation error with the statistical reliability of subjective quality assessment shows that objective quality measurement can be as reliable as subjective measurements for some testing conditions.

1. INTRODUCTION

Low bit-rate speech coding is a key technology for multimedia telecommunications. A number of coding algorithms have been developed for various applications. When optimizing or characterizing a codec, for example, one needs to evaluate its performance based on a subjective quality assessment, which is time-consuming and expensive. Therefore, objective quality measures that correlate well with subjective quality have been developed to estimate subjective quality.

Several objective quality measures have been proposed to the ITU-T (International Telecommunication Union - Telecommunication Standardization Sector), and a methodology to compare their performance needs to be developed.

In this paper we propose a method for demonstrating the accuracy of an objective estimation of subjective quality. We then applied the proposed method to compare the performance of objective quality measures that have been candidates for Recommendation in ITU-T.

2. METHOD FOR EVALUATING OBJECTIVE QUALITY MEASURES

Since the goal of objective quality measurement is to estimate subjective quality, its performance should be investigated in terms of the consistency between subjective and objective experimental results. This section first describes the experimental conditions for subjective and objective experiments, then proposes a performance index that will indicate the accuracy of objective estimation.

2.1. Experimental Conditions

To thoroughly investigate the performance of objective quality measures, the subjective and objective experimental results for various codecs should be compared taking into account the effects of various quality factors as shown in Table 1.

In our investigation, we used six different waveform/ CELP codecs with bitrates from 4 to 64 kbit/s (Table 2). From the viewpoint of investigating the basic performance of the objective quality measures, we took into account the effects of languages, talkers, and tandeming of the codecs (Table 3). Other factors were fixed.

Input speech was “clean” (i.e., without ambient noise) and spoken by four different talkers. The source speech samples were preprocessed (Fig. 1), then fed into the coding process (Fig. 2).

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. In the subjective experiments, there were 24 listeners and we used telephone handsets with modified IRS receiving characteristics defined by ITU-T Rec. P.830. The listening level was -15 dBPa. The experiments were carried out for the Italian and Japanese languages.

2.2. Performance Index

Conventionally, the performance of an objective quality measure is evaluated in terms of the consistency between the subjective MOS (mean opinion score) and its estimation by an objective quality measure [2]. The evaluation is done with performance indexes such as correlation coefficients and root mean square error (RMSE). The MOS is estimated by applying a predetermined relationship between a subjective MOS and an objective quality measurement value.

The subjective MOS for the same testing conditions may, however, differ from experiment to experiment, depending on experimental settings such as the nationality of the listener panel, the instructions given to the panel, and the overall quality balance in the experiment. This

Table 1 Quality factors in codec tests.

Terminal	sending/receiving acoustic characteristics, input level to a codec
Environment	ambient noise (sending/receiving)
Source signal	speech (language, talker, sentence), music
Network	tandeming, cell/packet loss, bit error

Table 2 CODEC used in experiments.

Algorithm	Bit rate [kbit/s]	Notation
ITU-T Rec. G.711 PCM	64	G.711
ITU-T Rec. G.726 ADPCM	32	G.726
ITU-T Rec. G.728 LD-CELP	16	G.728
ITU-T Rec. G.729 CS-ACELP	8	G.729
CELP*	4	CELP(4k)
US Federal Standard 1016 CELP	4.8	FS-CELP

* Pitch Synchronous Innovation CELP [1] under development.

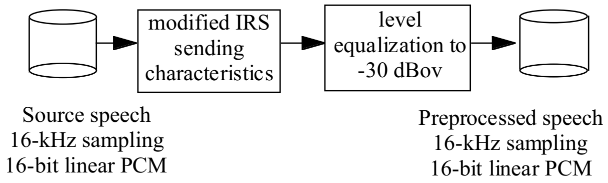


Fig. 1 Preprocessing of speech materials.

Table 3 Coding conditions.

Cond.	Codec	T	Cond.	Codec
1	G.711	1	19	G.726 & G.728
2	G.711	4	20	G.726 & G.729
3	G.711	8	21	G.726 & CELP(4k)
4	G.726	1	22	G.726 & FS-CELP
5	G.726	2	23	G.728 & G.729
6	G.726	4	24	G.728 & CELP(4k)
7	G.728	1	25	G.728 & FS-CELP
8	G.728	2	26	G.729 & CELP(4k)
9	G.728	3	27	G.729 & FS-CELP
10	G.729	1	28	CELP(4k) & FS-CELP
11	G.729	2	29	MNRU(Q = 35 dB)
12	G.729	3	30	MNRU(Q = 30 dB)
13	CELP(4k)	1	31	MNRU(Q = 25 dB)
14	CELP(4k)	2	32	MNRU(Q = 20 dB)
15	CELP(4k)	3	33	MNRU(Q = 15 dB)
16	FS-CELP	1	34	MNRU(Q = 10 dB)
17	FS-CELP	2	35	MNRU(Q = 5 dB)
18	FS-CELP	3	36	direct

T: number of tandemings
Conds. 19 - 28: asynchronous tandeming of different codecs (denoted by “Mix” hereafter)
Conds. 29 - 35: reference MNRU conditions

Table 4 RMSE in estimated equivalent-Q [dB].

Language	EPR	CD	II	CHF	PSQM
Italian	6.546	4.832	8.218	7.993	1.466
Japanese	6.338	5.160	8.471	8.204	1.810

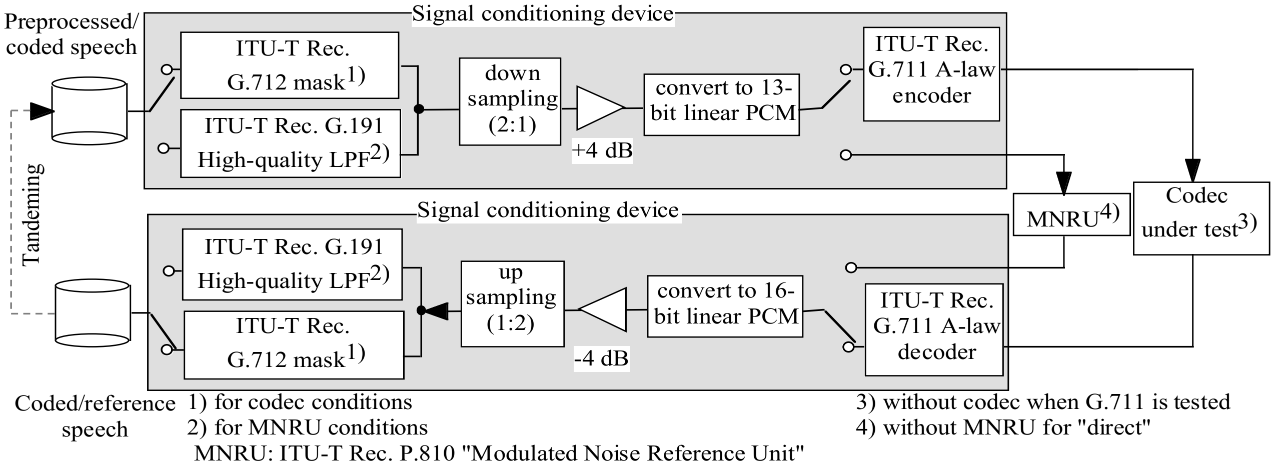

Preprocessed/
coded speech
Signal conditioning device
ITU-T Rec. G.712 mask1)
ITU-T Rec. G.191 High-quality LPF2)
down sampling (2:1)
+4 dB
convert to 13-bit linear PCM
ITU-T Rec. G.711 A-law encoder
Tandeming
MNRU4)
Codec under test3)
Signal conditioning device
ITU-T Rec. G.191 High-quality LPF2)
ITU-T Rec. G.712 mask1)
up sampling (1:2)
-4 dB
convert to 16-bit linear PCM
ITU-T Rec. G.711 A-law decoder
Coded/reference speech
1) for codec conditions
2) for MNRU conditions
MNRU: ITU-T Rec. P.810 "Modulated Noise Reference Unit"
3) without codec when G.711 is tested
4) without MNRU for "direct"

Fig. 2 Processing of coding/reference conditions.

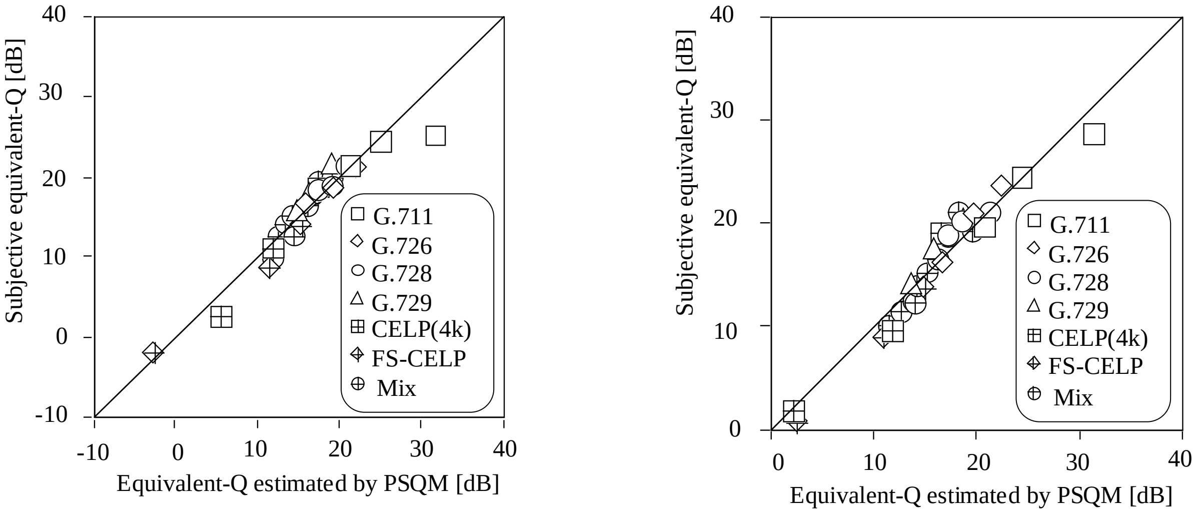
means that the estimated MOS may diverge from the subjective MOS not only because of poor performance of an objective quality measure, but also because the MOS is so experimentally dependent.

The equivalent-Q conversion method is often used in subjective quality assessments to avoid the experiment-dependency of the MOS. Equivalent-Q is defined as the SNR of MNRU (ITU-T Rec. P.810: Modulated Noise Reference Unit) speech whose quality is equivalent to that of a codec. (The SNR of the MNRU speech is referred to as “Q.”)

Since the relative quality between coded and reference speech is expected to be preserved over experiments, we can appropriately compare the subjective quality and its objective estimation in the equivalent-Q domain. Therefore, we propose the RMSE of the estimated equivalent-Q as an index of the performance of an objective quality measure.

3. PERFORMANCE EVALUATION OF OBJECTIVE QUALITY MEASURES

Applying the evaluation method described in Section 2, we evaluated the performance of objective quality measures that were candidates to become the ITU-T standard measure: the Coherence Function (CHF) [3], LPC Cepstrum Distance (CD) [4], Information Index (II) [5], Perceptual Speech Quality Measure (PSQM) [6], and Ex-



(a) Japanese

(b) Italian

Fig. 3 Relationship between subjective and objective evaluation for PSQM.

Table 5 RMSE in estimated equivalent-Q [dB] for each talker.

	EPR	CD	II	CHF	PSQM
Japanese					
Talker: F1	6.275	4.937	10.264	8.526	2.918
Talker: F2	5.480	5.099	7.837	8.347	4.226
Talker: M1	7.151	5.942	9.737	9.248	2.817
Talker: M2	7.123	6.035	8.362	7.980	2.568
Average	6.507	5.503	9.050	8.525	3.132
Italian					
Talker: F1	8.081	3.937	12.209	10.869	4.441
Talker: F2	5.270	5.431	6.664	6.943	5.061
Talker: M1	10.130	8.930	11.024	11.137	4.520
Talker: M2	6.319	6.021	5.927	6.067	2.094
Average	7.450	6.080	8.956	8.754	4.029

Table 6 RMSE in estimated equivalent-Q [dB] for each category of coding conditions.

	EPR	CD	II	CHF	PSQM
Japanese					
Waveform	3.833	6.997	6.696	5.249	2.814
CELP	6.747	5.109	8.631	8.353	1.616
Mix	7.003	3.722	9.199	9.390	1.136
Italian					
Waveform	3.343	5.053	7.440	4.762	1.355
CELP	7.213	5.164	8.328	8.365	1.587
Mix	7.130	4.250	8.523	9.017	1.376

pert Pattern Recognition (EPR) using the CHF, CD, and II as distortion measures [7].

3.1. Performance Comparison

The RMSE of objective measurement was calculated in the following manner:

Step 1) averaging subjective/objective measurement values for four talkers,
Step 2) transforming them into the Q domain,
Step 3) taking RMS of the differences between subjective and objective equivalent-Q.

Table 4 shows the RMSE of equivalent-Q estimated by each of the five objective quality measures. The PSQM best estimated the subjective quality for both languages. Figure 3 demonstrates the relationship between subjective equivalent-Q and its estimation by the PSQM.

We further investigated this superiority of the PSQM from different points of view. Table 5 shows the RMSEs for individual talkers. To obtain these, we calculated subjective and objective equivalent-Q for individual talkers, then applied Step 3 in the above procedure. Although there is one case where the CD performs better than the PSQM (Italian talker F1), the difference is relatively small. Table 6 analyzes the results for three types of coding conditions: Waveform (Conditions 1 - 6), CELP (Conditions 7 - 18), and Mix (Conditions 19 - 28). It shows that the superiority of the PSQM does not depend on the codec.

3.2. Comparison with Confidence Intervals of Subjective Testing Results

In Section 3.1, we compared five objective quality measures, and found that the PSQM performed the best. This does not necessarily mean, however, that objective measurement by the PSQM can substitute for subjective quality assessment. In this section, we evaluate the accuracy of objective estimation by the PSQM in comparison with the statistical reliability of the subjective experimental results.

To do this, we first derived the 95% confidence interval for each testing condition in the MOS domain (the degree of freedom is 95), then transformed this into the Q domain. Finally, we calculated the RMS of these confidence intervals in the Q domain, and compared it with the

Table 7 Comparison with confidence intervals of subjective testing results for each talker.

	F1	F2	M1	M2
Japanese				
RMS of one-sided 95% confidence interval	1.540	1.607	1.683	1.272
RMSE of estimation by PSQM	2.918	4.226	2.817	2.568
Italian				
RMS of one-sided 95% confidence interval	5.410	2.335	3.090	2.268
RMSE of estimation by PSQM	4.441	5.061	4.520	2.094

Table 8 Comparison with confidence intervals of subjective testing results for each category of coding conditions.

	Waveform	CELP	Mix
Japanese			
RMS of one-sided 95% confidence interval	1.268	3.120	0.833
RMSE of estimation by PSQM	2.814	1.616	1.136
Italian			
RMS of one-sided 95% confidence interval	4.607	1.538	1.227
RMSE of estimation by PSQM	1.355	1.587	1.376

RMSE by the PSQM.

For this subjective experiment, with 24 listeners, the RMS of the 95% confidence intervals were 2.183 dB for Japanese and 2.470 dB for Italian. Comparing these values to the RMSE values of the PSQM (from Table 4), we can conclude that the objective quality measurement by the PSQM is as reliable as these subjective results for estimating the quality of a waveform or CELP codec between 4 and 64 kbit/s.*

We compared the performance of the PSQM quality predictions with subjective scores based on talker (Table 7) and coding category (Table 8).

For the talker-base analysis, the 95% confidence intervals were derived in the MOS domain for individual talkers per condition (degree of freedom is 23), then transformed into the Q domain. Table 7 shows that the RMSE by PSQM is larger than the RMS of 95% confidence intervals with a few exceptions. This implies that the performance of the PSQM degrades if only a few talkers (or speech samples) are used in the measurement.

In Table 8, the RMSE by PSQM is less or comparable to RMS of 95% confidence intervals with one exception: “Japanese Waveform.” In Fig. 3 (a), the objective estimation error is very large (approx. 5 dB) for one of the “G.711” conditions (Condition #1), making the RMSE for “Japanese Waveform” much larger than that for other categories. This is simply because of instability of the equivalent-Q conversion and does not mean the PSQM is inapplicable to waveform codings.** In fact, for other waveform coding conditions, the PSQM works quite satisfactorily.

4. CONCLUSION

We proposed a method for evaluating the performance of objective quality measures. Applying this method to five objective quality measures that were candidates to become the ITU-T standard measure, we found that the PSQM, which is based on the loudness of the coding distortion in the Bark-spectral domain, gave the best performance, regardless of language, talker, or codec.

Comparing the accuracy of the estimation by the PSQM with confidence intervals of the subjective equivalent-Q, we concluded that, under some testing conditions, the objective quality measurement by the PSQM can be as reliable as a subjective assessment in terms of estimating the mean quality for several talkers.

While the validity does not depend on the coding schemes, the performance of the PSQM degrades when estimating the quality of a codec for individual talkers.

The methodology and evaluation results discussed in this paper were reflected in the study of objective speech quality measures in ITU-T SG12. Based on this investigation, the new ITU-T Recommendation P.861 “Objective quality measurement of telephone-band (300-3400 Hz) coded speech” using the PSQM as an objective quality measure was standardized.

The validity of the Rec. P.861 was verified only for the evaluation of the effects of tandemings. Its applicability to evaluating the effects of other quality factors such as cell/ packet loss and ambient noise is still under study.

5. REFERENCES

[1] K. Mano and T. Moriya, "Improved 4-kbit/s PSI-CELP coding using pitch and phase adaptation," Proc. 1995 IEEE workshop on speech coding for telecommunications, pp. 41-42, Sept. 1995.

[2] e.g., H. Irii, “Comparison of objective speech quality assessment methods based on international subjective evaluations of universal codecs,” ICC’91, pp.1726-1730, 1991.

[3] CCITT Contribution COM XII-60, "Evaluation of non-linear distortion via the coherence function (BNR)," April 1982.

[4] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems, " IEEE J. on Selected Areas in Communication, vol. 6, no. 2, pp. 242-248, Feb. 1988.

[5] J. Lalou, "The information index: an objective measure of speech transmission performance," Annales des Telecommunications, vol. 45, no. 1-2, CNET/France, pp. 47-65, 1990.

[6] J. G. Beerends, J. A. Stemerdink, "A perceptual speech quality measure based on a psychoacoustic sound representation," J. Audio Eng. Soc., vol. 42, no. 3, pp. 115-123, March 1994.

[7] ITU-T P-series Recs. Supplement 3 Annex G (1993, not in force since April 1997 on).

* It should be noted that the values of the 95% confidence interval vary as a function of 1/sqrt(N), where N is the number of opinion votes in the subjective test. Therefore, for a large number of listeners, for example, the PSQM may not provide predictions that are of the same reliability as the subjective testing results.

** Similar instability is observed in the subjective results for "Italian Waveform" in Table 8. In general, for very high/ low Q-regions, the Q vs. subjective/objective quality curve is almost flat, and equivalent-Q conversion is sometimes too sensitive.