Subjective quality index for compatibly evaluating narrowband and wideband speech

Akira Takahashi, Atsuko Kurashima, and Hideaki Yoshino

NTT Service Integration Laboratories, Japan

{takahashi.akira}{kurashima.atsuko}{yoshino.hideaki}@lab.ntt.co.jp

Abstract

Telecommunications services using wideband speech media have been developed through the evolution of conventional IP telephony. These include not only simple wideband IP telephony, but also video telephony and video conferencing applications exploiting IP technologies. Since such services coexist with conventional telephony services that use narrowband/telephone-band speech media, it is important to develop a stable subjective quality index that can compare narrowband and wideband services taking into account the quality enhancement achieved by widening the speech bandwidth. We proposed to use an Absolute Category Rating (ACR) test as a subjective assessment method in which speech with various bandwidths was tested in the same context. We investigated the validity of the proposed method from the viewpoints of compatibility with the conventional Mean Opinion Score (MOS), repeatability, and sensitivity in evaluating coding and packet-loss distortion.

Keywords

wideband, subjective quality, MOS

1 Introduction

Wideband speech services are expected in a number of telecommunications scenarios, such as wideband IP telephony and videoconferencing over IP networks. It appears that telephone-band and wideband speech media will coexist for the time being. Therefore, we think it is highly important to evaluate the quality of speech taking into account its bandwidth, as well as other quality degradations such as coding distortion and packet loss.

So far, the quality of wideband speech, which is defined as a bandwidth between 3400 and 7000 Hz in this paper, has usually been assessed individually. Therefore, it is difficult to evaluate the trade-off in characteristics between widening the speech bandwidth and speech coding/packet-loss distortion because of limited network/terminal resources. However, in the near future, users will need to choose a service from a menu with different speech bandwidths. Therefore, it is essential to develop reliable and effective quality assessment methodologies for use in such scenarios.

To evaluate the quality of speech with an individual bandwidth such as the telephone band or the 7-kHz band, we have standardized methodologies, that is, Recommendations P.800 [1] and P.830 [2], which ensure stable and reliable subjective evaluation. We call such an individual evaluation a “subset evaluation” in this paper. However, we do not know much about users’ perception of quality when speech with various bandwidths is presented in the same context. We call the speech quality evaluation that takes into account the bandwidths effect “global evaluation” in this paper.

Conventionally, the subset MOS for telephone band (300-3400 Hz) or wideband (100-7000 Hz) has been used for assessing subjective speech quality [2]. However, because we do not know much about the relationship between these indices, we cannot compare them. This means we cannot compare the quality of narrowband speech to wideband speech although there is trade-off between them when determining the bandwidth within a given bitrate. Therefore, we definitely need a common index that compares the quality of speech samples that have various bandwidths.

France Telecom R&D investigated this issue and made some important remarks [3]. One of those is that the MOS obtained in the global evaluation has a systematic difference from that obtained in the subset evaluation of telephone-band speech, although the MOSs for wideband speech in both global and subset contexts are consistent. This remark is supported based on more extensive subjective experiments reported in this paper.

This paper proposes to adopt the global Mean Opinion Score (MOS) obtained in the global evaluation as a quality index for narrowband and wideband speech. We first investigate the relationship between the global MOS and subset MOS of speech with various bandwidths. Next, we confirm the repeatability of the global MOS. Finally, we also evaluate the sensitivity of such evaluations to quality degradation other than bandwidth limitation, such as coding and packet-loss distortions.

2 Relationship between global and subset MOS

2.1 Experimental conditions

2.1.1 Signal processing

We used four females and four males as talkers in the experiments. For each talker, we used a sentence pair lasting eight seconds including silent time. We passed these speech samples through band-/low-pass filters and the modulated noise reference unit (MNRU) defined by ITU-T Recommendation P.810 [4]. The testing conditions are summarized in Table 2.1. The block labelled “BPF/LPF” indicates band-pass filtering and low-pass filtering, respectively. The Q value is the signal-to-noise ratio of the MNRU system.

2.1.2 Presentation of speech samples

In our investigation, we conducted two kinds of subjective quality experiments; one is called a “subset test,” in which only the speech signals with the same bandwidth were tested. For instance, a subset test consisted only of speech with a bandwidth of 0 – 3400 Hz, labelled as “A” in Table 2.1. The other test is called a “global test,” in which all the conditions shown in Table 2.1 were presented randomly to the subjects. In this global test, subjects listened to speech with arbitrary bandwidths and arbitrary Q-values.

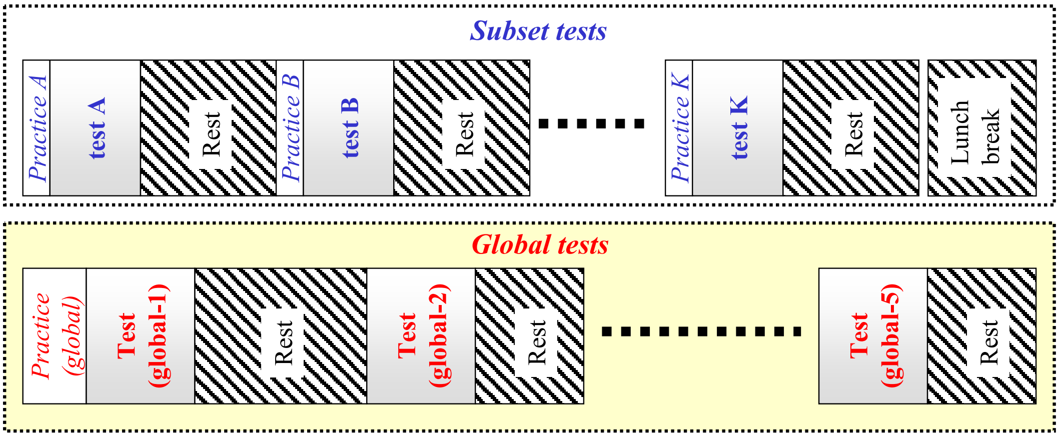
The daily schedule of subjects is illustrated in Figure 2.1. We used eight subjects per day and divided them into two groups. The presentation order, that is, the randomization pattern was different for different groups.

The subjects first experienced subset tests. Each test had a training period consisting of three speech samples and 32 speech samples for evaluation. The training period has the effect of resetting the opinion criteria in the

Table 2.1 Signal processing conditions.

Subset	Q [dB]	Bandwidth (LPF)		Subset	Q [dB]	Bandwidth (BPF)	
		Fl [Hz]	Fh [Hz]			Fl [Hz]	Fh [Hz]
A	15	0	3400	F	15	100	3400
	25	0	3400		25	100	3400
	35	0	3400		35	100	3400
	100	0	3400		100	100	3400
B	15	0	4000	G	15	100	7000
	25	0	4000		25	100	7000
	35	0	4000		35	100	7000
	100	0	4000		100	100	7000
C	15	0	5000	H	15	200	3400
	25	0	5000		25	200	3400
	35	0	5000		35	200	3400
	100	0	5000		100	200	3400
D	15	0	6000	I	15	200	7000
	25	0	6000		25	200	7000
	35	0	6000		35	200	7000
	100	0	6000		100	200	7000
E	15	0	7000	J	15	300	3400
	25	0	7000		25	300	3400
	35	0	7000		35	300	3400
	100	0	7000		100	300	3400
				K	15	300	7000
					25	300	7000
					35	300	7000
					100	300	7000

Figure 2.1 Schedule of subjective tests.



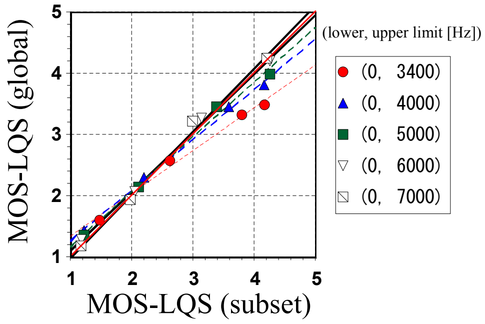


Figure 2.2 Relationship between subset and global MOS (LPF).

previous session and reestablishing them for the coming session.

After these 11 subset tests, the subjects were asked to take part in the global test, which was preceded by a training session intended to let them form their opinion criteria. This training session consisted of ten speech samples with a wide range of bandwidths and Q-values. Each session consisted of 80 speech samples to be evaluated; there was a short break after the evaluation of 40 speech samples, except in the first session, in which there were 32 speech samples plus ten training samples.

We repeated these tests ten times for different subjects to obtain opinion data from 40 subjects (20 females and 20 males).

2.1.3 Listening conditions

We used binaural headphones (SENNHEISER HD250) whose frequency responses were equalized to be as flat as possible. The listening level was -18 dBPa at each ear reference point. We did not add any environmental noise in the receiving room.

2.2 Results

Experimental results for the low-pass filter conditions (A – E) and the band-pass filter conditions (F – K) are shown in Figures 2.2 and 2.3, respectively[1].

The global test gives consistent results with the subset tests for speech that has an upper band limit of 6000 Hz or more as indicated in Figure 2.2. However, when testing speech that had an upper limit of 5000 Hz or less, there were significant gaps between the results from subset tests and those from the global test. These results support the conclusion derived in [3]. It may be possible to map the MOS-LQS of subset tests to those of the global tests, as suggested also in [3].

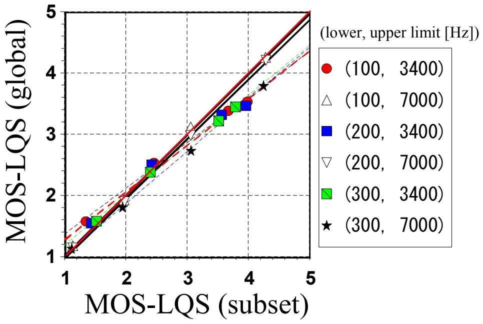


Figure 2.3 Relationship between subset and global MOS (BPF).

One can see from Figure 2.3 that the finding presented in Figure 2.2, that is, the global test gave results consistent with the subset tests for wideband speech, also holds if the lower limit of the band-pass filter is up to 200 Hz. For narrowband speech, as expected, the subset tests always gave different results from the global test, regardless of the lower limit of the bandpass filter.

3 Repeatability of global MOS

Because subjects are required to rate the quality of speech with various bandwidths in the global evaluation, results may become unstable. That is, different subjects may give a quite different score for the same testing condition due to their preference for speech bandwidth.

To investigate the stability of the global MOS, we repeated the global evaluation described in Section 2 using different subjects. The results are compared with those of the previous test in Figure 3.1. This ensures that the global MOS is quite stable even for different subjects.

4 Sensitivity of global MOS

It is often difficult for subjects to evaluate multiple kinds of degradations at the same time. For example, when one presents speech samples with various speech bandwidths, subjects may focus only on the speech bandwidths, being less sensitive to other test factors, such as coding distortion and packet loss.

In our study, we investigated the above-mentioned issue by carrying out two subjective experiments in which we evaluated the effects of coding distortion and packet loss. One was a subset evaluation that employed only wideband speech, and the other was a global evaluation including narrowband and wideband speech. The wideband speech was completely the same in both tests.

The testing conditions are summarized in Table 4.1. When generating random/bursty packet loss, we used the discrete Gilbert-Elliot channel model which is

[1] The 95% confidence intervals in subjective testing were between 0.04 and 0.11 on the MOS-LQS scale.

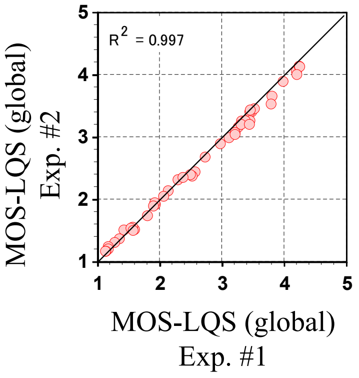


Figure 3.1 Reproducibility of global MOS.

adopted in ITU-T Recommendation G.191 [5]. There are a total of 36 conditions[2] for G.711 and 45 for other codecs. The numbers of subjects were 48 and 40 for the global and subset evaluations, respectively. Listening conditions were the same as the ones described in Section 2.1.3.

The quality degradation characteristics in terms of packet-loss rate for G.722 codec with a packet length of 20 ms are demonstrated in Figure 4.1. The experimental results indicate that the quality evaluation sensitivity of global MOS is comparable to that of subset MOS. By evaluating the statistical significance of difference between global and subset MOS with a significance level of 5%, we found that the global evaluation does not degrade the sensitivity of MOS in any case from the statistical viewpoint.

5 Conclusion

The use of global MOS as an index for evaluating narrowband and wideband speech on the same scale, taking into account the difference in speech bandwidth was proposed in this paper. We investigated the relationship between the proposed index and the conventional MOS scales. In addition, we verified the rationale of the proposed index from the viewpoints of its reproducibility and sensitivity. We concluded that the global MOS preserves the reproducibility and sensitivity of conventional MOS scores, enabling the comparison of subjective quality among various speech bandwidths.

Table 4.1 Coding and packet-loss conditions.

			subset evaluation	global evaluation
packet length [ms]	telephone -band	G.711 with PLC	▨	10, 20, 40, 100
		G.729	▨	10, 20, 50
		MNRU	▨	(Q = 5, 15, 25, 35, 40 dB)
	wideband	G.711 with PLC*1	10, 20, 40	
		G.722@64kb/s	10, 20, 40	
		G.722.1@32kb/s	20, 40, 80	
		MNRU	(Q = 15, 25, 30, 35, 40 dB)	
packet-loss rate [%]			0, 1, 3, 5, 10	
packet-loss characteristics			random/bursty	

*1 Simple extension of G.711 with PLC to wideband speech. The signal processing frame was doubled so that the frame size in the time domain was unchanged.

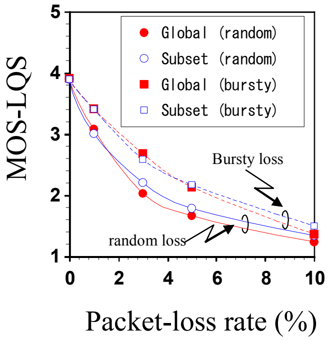


Figure 4.1 Sensitivity of MOS vs. packet-loss degradation.

6 References

[1] ITU-T Recommendation P.800, “Methods for subjective determination of transmission quality,” Aug. 1996.

[2] ITU-T Recommendation P.830, “Subjective performance assessment of telephone-band and wideband digital codecs,” Aug. 1996.

[3] Vincent Barriac, et al., “Discussion on unified objective methodologies for the comparison of voice quality of narrowband and wideband scenarios,” Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction, June 2004.

[4] ITU-T Recommendation P.810, “Modulated noise reference unit (MNRU),” Feb. 1996.

[5] ITU-T Recommendation G.191, “Software tools for speech and audio coding standardization,” Nov. 2000.

Appendix

In this appendix, we further confirm the relationship between subset and global MOS, which is investigated

[2] 4 packet-length conditions × (1 error-free condition + 2 packet-loss characteristics conditions × 4 packet-loss rate conditions)

in Section 2, using the subjective data obtained in the experiments described in Section 4.

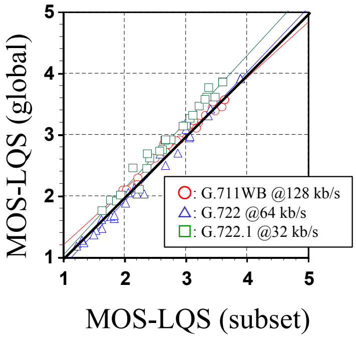


Figure A.1 Relationship between subset and global MOS.

The subset and global MOS are compared in Figure A.1. The claim in Section 2 that global MOS corresponds to subset MOS when lower and upper limits in speech bandwidth are below 200 Hz and above 6000 Hz[3], respectively, still holds in this case. However, there is a slight difference in the evaluation of the G.722.1 codec in comparison with other codecs. That is, the G.722.1 codec was evaluated more highly when presented with telephone-band codecs than with wideband codecs only. We need to further investigate the codec dependence in the relationship between subset and global MOS.

[3] The effective bandwidth was 100 - 7000 Hz in these experiments.