

Report on speech quality investigation of VoIP terminals

1. Preface

In the year 2002, Japanese telephone services face a new era after the Ministry of Public Management, Home Affairs, Posts, and Telecommunications (MPHPT) started service approval and numbering allocation for connecting providers of public switched networks (PSTN) to internet protocol (IP) networks. Voice over IP (VoIP) services may quickly replace conventional telephone services.

Our association, the Communications and Information network Association of Japan (CIAJ), which represents manufacturers supplying network devices and terminals, is greatly interested in such trends.

One of the key issues of the VoIP is speech quality. Thus, our communication quality committee of CIAJ plans to investigate the speech quality of currently marketed VoIP terminals. Based on the results, the committee intends to produce some guidelines or specifications for equipment such as the existing CIAJ standard for analog and mobile phones. The main theme of the investigation is to assess quality based on a subjective test.

The "Study Group on IP Network Technology" of the MPHPT proposed four classes of VoIP service quality (A, B, C, and D), using an R-value defined by ITU-T Recommendation G.107, and proposed an absolute transmission delay in their report in January 2002,.

We also investigate the validity of the classification with technology condition as we are now.

2. Test samples

Only samples that have been marketed were submitted to testing. Products under development or that were at the prototype stage were excluded. Because this assessment was not intended to be for commercial products, the names of the manufacturers and the product codes were kept secret. However, to motivate manufacturer participation, information for which manufacturers could identify their products was provided privately to them. If a different set up was available for the submitted sample, they requested fixing the typical values to those that manufacturers recommend to users.

Two types of typical VoIP terminals were permitted:

1) Gateway

This device is also called a terminal adapter. The subscriber's conventional analog telephone sets are connected through an FXS interface. The other side of the LAN interface can be connected to an IP network.

2) IP-Phone

This terminal is a telephone set. It has a LAN interface and can connect directly to an IP network.

Both types have jitter absorption, packet construction/deconstruction, and speech coding, and they usually have echo canceling.

Seven manufacturers participated by submitting one or two terminals each. Consequently, ten terminals were tested. Among the ten samples, five were GW and the rest were IP-Phone.

Among many standardized algorithms, three kinds of speech coding algorithms, ITU-T G.711, G.729/G.729A, and G.723.1 are mainly used for the VoIP application. In terms of G.711, an algorithm with packet loss concealment has been recommended, but no terminals implemented such functions in our sample. The G.711 and G.729/G.729A are becoming the standard in VoIP for public telecommunication services because of their speech quality. Therefore, in this assessment, G.723.1 coding algorithms were excluded. Manufacturers decided internal parameters such as the packet and jitter buffer size, including which coding algorithm was used.

3. Subjective test procedure

3.1 Confirmation of settings

Before testing the terminals in the community, subjects confirmed the settings for their samples and experienced the conversational speech quality.

3.2 Host Laboratory

The host laboratory was the administration, which managed all the test procedures and provided the facilities. The CIAJ designated the NTT Advanced Technology Corporation (NTT-AT) as the host laboratory because of its neutrality among the manufacturers and

ability to ensure confidentiality.

3.3 Network simulation

Network configuration in this test was restricted to a symmetrical network in which only a connection between the same kinds of terminal at both ends was used to avoid interoperability problems.

Three factors in IP networks affect VoIP speech quality: transmission delay (also known as absolute delay), delay variation, and packet loss. These factors have to be considered. An IP path with an increasing absolute delay is correlated with a larger delay variation because the packets pass through many routers. Consequently, a larger packet loss tends to occur. A lot of arguments were needed to determine values for these factors. We chose the following four combinations of the three factors by reviewing experimental reports presented in the ITU-T and other standard bodies to form our conditions:

Condition	Absolute delay	Delay variation	Packet loss
1 st	none	none	none
2 nd	50 ms	25 ms	1%
3 rd	100 ms	50 ms	3%
4 th	350 ms	100 ms	5%

An absolute delay here does not contain an internal delay of terminals at both sides. In terms of statistical features in delay variation and packet delay, we chose Poisson and uniform random probability, respectively, by accounting for the capability of the network simulator. We adopted “Netdisturb”, developed by France Telecom, as a network simulator because of its ease of installation and operation in Windows.

The test system configuration is shown in **Fig. 1**. All ten samples on both sides were always connected during the test. Subjects on both sides started a conversation by picking up one of their instructed handset. Speech paths in terminals that the subjects used were switched to the network simulator. The other paths were switched to a bypass route. We did this to prevent influencing other IP traffic to the given network disturbance.

3.4 Reference speech path

Within two quality classes proposed by the MPHPT study group, their expression for

class “A” is the quality experienced during a public switched telephone network (PSTN) call. The other one is class “B”, which is the quality experienced during a cellular phone call. To make quality references, we added a two-speech path to our test. For class A, a reference speech path that is configured by PCM on a one-link trunk line and a standard analog telephone were connected. For the class B reference, the speech path simulates a typical Japanese cellular telephone service configured by a low bit rate codec (4 kbit PSI-CELP adopted in a PDC half rate system) and telephone handsets. Both transmission paths were error free.

3.5 Other test condition

Because loudness of perceived speech seriously influences speech quality, the overall loudness rating (OLR) of each speech path was adjusted to 10 dB within 2 dB by measurement. 30 dBA of room noise was generated in both rooms to reduce talker echo using acoustical coupling at the far end.

3.5 Quality rating

Two kinds of subject opinion were collected.

First: A five-point absolute category rating.

Point	Quality
5:	Excellent
4:	Good
3:	Fair
2:	Poor
1:	Bad

Second: Three grades determined whether or not speech quality was permissible:

Speech quality was permissible if it was the same charge as that for conventional telephone services (Permissible)

Speech quality was permissible if it was less expensive than that for conventional telephone services (Conditionally permissible)

Speech quality was not permissible if it was less expensive than that for conventional telephone services. (Reject)

Subjects input their opinion through a keypad so that they could not see their scoring carrier.

3.6 Conversation method

Both subjects were instructed to use a set of abstract patterns for topics of conversation named one word, as found in the traditional conversational test introduced in an old CCITT recommendation. For this method, subjects on both sides alternately conducted Q&A dialog for schematic features of the pattern. A one-minute period was allocated for each condition, and subjects voted their opinion after finishing every condition.

Only handsets were presented in front of the subject, and telephone housings were kept away from him or her to reduce physiological bias in evaluating. Forty subjects were recruited from the community.

Before the experiment, subjects were made to practice for 6 conditions. During practice, the experimenter checked whether or not subjects responded correctly after receiving instructions.

4 Delay measurement

Absolute delay was measured for each test speech path using Agilent's Telegra VQT. For the GW, the segment between the FXS interfaces of both terminals was measured. For the IP-Phone, the segment between the handset modular jack in both terminals was measured.

5 Results

5.1 Mean opinion score

Mean opinion scores (MOS) are summarized in **Fig. 2**. In the figure, each terminal was alphabetically coded to ensure confidentiality. Results show that the speech sample is lower than the MOS for a PSTN telephone, at approximately 4 points, but exceeds the MOS for a cellular phone. The fact that a few samples without disturbance are close to the quality of a PSTN telephone is quite convincing for the following reasons. In principle, samples that use G.711 speech coding are of the same quality as PSTN telephones in terms of coding distortion. They may have little difference in speech loudness and frequency response. Because most of the samples adopt G.729/G.729A, about a 0.5 lower MOS is caused because of the speech coding cannot avoid greater coding distortion than PCM.

In a moderate IP disturbance, samples provide the same speech performance as that in

a cellular phone. At the most serious level of disturbance, samples cannot reach the quality of a cellular phone, although they exceed the 2-point MOS level.

5.2 Permissible rate

A permissible percentage for each condition is shown in **Fig. 3**. The percentage of rejection increases as the size of the disturbance increases, and it reaches about 40 percent in the worst condition. However, the conditionally permissible rate is relatively stable regardless of the IP disturbance.

The correlation between the permissible rate and the MOS is shown in **Fig. 4**. From this relationship, a MOS above 3 points is necessary to satisfy a permissible rate of more than 50%. The rejection rate can be suppressed to less than 20% at this MOS value.

5.3 Delay measurement

Results for the absolute delay of each test conditions are shown in **Fig. 5**. The delay increased according to the additional delay caused by the IP disturbance simulator. This ensures an additional delay correctly controlled by the simulator. When no disturbance was added, the measured delay was the sum of the internal delay of the terminals on each side. It was determined that such an internal delay was in the range of 80 to 170 ms.

6 Conclusion and Observation

We can conclude the following facts from this investigation.

- 1) Using the current technology, a VoIP service connecting the same terminal on both ends can give better speech quality than a cellular phone until a weak network disturbance is caused, but it does not reach the PSTN level.

At the transition era when one side were to be connected to a PSTN, speech quality would be worse than these results due to echo caused by an insufficient return loss at a 2 wire-4 wire conversion in a VoIP terminal and a PSTN GW.

PSTN quality can be achieved if a sufficiently broader bandwidth and reduction in network load are guaranteed and if PCM speech coding become mainstream in the future.

2) R-values at the boundary of each quality class by the MPHPT Study Group correspond to 4.03 and 3.60 for class A and B if we convert them using ITU-T Recommendation G.107. However, as for the results of this investigation, although the quality of all terminals was better than that in cellular phones for none or weak disturbance, all of the samples are ranked below class B. This is because the boundaries were derived from subjective tests for North American or European subjective tests. Japanese are well known to be more demanding with regard to quality. We should consider such differences due to nationality when we plan guidelines or domestic standards.

3) In defining the quality class for the study group, class A was less than 100 ms, and class B was less than 150 ms. This included the terminal delay. Because the terminal itself has more than a 100 ms delay, the boundary for each class should be wider when taken into the actual terminals.

7 Future work

We are planning our next investigation of VoIP speech quality. In it, we will consider using a PSTN network for the other end.