WALAILAK JOURNAL
http://wjst.wu.ac.th Engineering and Physical Sciences

IP Telephony: Comparison of Subjective Assessment Methods for Voice Quality Evaluation

Therdpong DAENGSI[1,*], Chai WUTIWIWATCHAI[2], Apiruck PREECHAYASOMBOON[3] and Saowanit SUKPARUNGSEE[4]

[1]Department of Information Technology, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
[2]Speech and Audio Laboratory, National Electronics and Computer Technology Center, Pathumthani 12120, Thailand
[3]Corporate Strategy, TOT Public Company Limited, Bangkok, Bangkok 10210, Thailand
[4]Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand

([*]Corresponding author; e-mail: therdpong1@yahoo.com)

Received: 2 December 2012, Revised: 11 June 2013, Accepted: 9 October 2013

Abstract

One problem that often occurs after installing/implementing an IP telephony system is voice quality. Although there are objective measurement tools for voice quality evaluation, the prices of these are very expensive. Therefore, back to basics, this paper focuses on subjective tests. This study used the data from three tests, consisting of listening-opinion tests, conversational tests and interview tests. All were conducted using the same IP telephony system with G.711 codec. All tests, following the ITU-T P.800, were in the best condition. The subjects who participated in the tests were 163 students and 1 worker in King Mongkut's University of Technology North Bangkok (KMUTNB). This study compared and analyzed the data from 3 kinds of subjective tests using ANOVA, to see if the data from the interview tests were consistent significantly with the data from the listening and conversational-opinion tests. The results, called the Mean Opinion Score (MOS) values, from the interview, conversational, and listening-opinion tests were 4.14, 4.16 and 4.23 respectively. Also, the analyzed result shows a p-value of 0.511. This means the MOS values from these three methods are not significantly different. Therefore interview tests can be used to evaluate voice quality, and is as good as other subjective methods, without high cost of expensive tools, making it very applicable in developing countries.

Keywords: IP Telephony, VoIP, voice quality, subjective tests, MOS

Introduction

The emergence of advanced technologies provides many advantages to human life. However, they may have some disadvantages or limitations for use. In the telecommunication industry, one problem that often occurs after installing or implementing an Internet Protocol (IP) telephony system is voice quality, because voice quality of Voice over Internet Protocol (VoIP) applications are impacted from network factors, such as packet loss, packet delay, jitter and echo [1]. This could be a cause of an argument between a system owner and the service provider, leading to delayed payment of dues. Although there are some tools available such as Perceptual Evaluation of Speech Quality (PESQ) and E-model which are objective measurement/assessment /evaluation tools used to evaluate voice quality [1], if the user is unfamiliar with these methods, trust and reliability are hard to gain. Moreover, the prices of these tools are very expensive and the charge by a third party to lease them can also be expensive. Of course, the cost

of these may be an issue if the system owner does not want to pay more which means the service provider must absorb the cost.

Therefore, this paper is to simply verify and confirm if interview tests can work well, compared with other subjective tests, and can be an alternative assessment method, instead of using an expensive objective assessment tool.

Background

This section presents the background information about IP telephony and the related subjective assessment methods. However, information on objective assessment methods can be found in [1,2].

IP telephony overview

IP telephony can be defined as the modern private branch exchange (PBX) that uses VoIP technology. However, the IP-PBX can support both IP-based packet switched technology and circuit switched technology [3], as presented in **Figure 1** [1], and can be classified into IP telephony as well. IP telephony has many features and functions similar to traditional PBX, for example, conferencing, call transferring and call picking up. It can also support adjuncts, such as, automatic call distribution system (ACD), voice messaging system (VMS), interactive voice response system (IVR), call accounting system (CAS) and predictive dialing system (PDS). IP telephony is the 3rd generation of PBX that uses the Internet Protocol (IP). Instead of using Time-Division Multiplexing (TDM), IP has been used to carry both voice signals and control signals.

Voice quality and its metric

Voice quality is a subjective topic and also an ambiguous term representing superiority of voice service. However, this term can mean different things to different people [4]. Defining 'good' voice quality may vary with business needs, cultural differences, user expectations and IP telephony systems [4].

Nevertheless, the metric of voice quality has been defined officially by the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) using a 5-point scale, where 1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent [5]. Basically, voice quality is evaluated by a group of subjects (e.g. 24 - 32 subjects) before averaging the scores from the subjects. The average score is called the Mean Opinion Score (MOS) or sometimes called the subjective MOS.

Voice quality assessment methods

There are many methods for telecom - voice quality evaluation. Those methods are classified into objective and subjective assessment methods. Objective measurement methods are classified into non-intrusive and intrusive assessment methods, as shown in **Figure 2(a)** [1]. However, it has been mentioned in [6] that objective measurements cannot predict subjective response well enough to entirely replace subjective measurements and they may not be accurate and reliable if the test conditions are changed. Therefore, this paper focuses on subjective measurement methods only. As shown in **Figure 2(b)**, there are 3 major subjective assessment methods, consisting of listening-opinion tests, conversation-opinion tests and interview tests that all are involved in this study.

For listening-opinion tests, they are classified as, absolute category rating (ACR), degradation category rating (DCR) and comparison category rating (CCR).

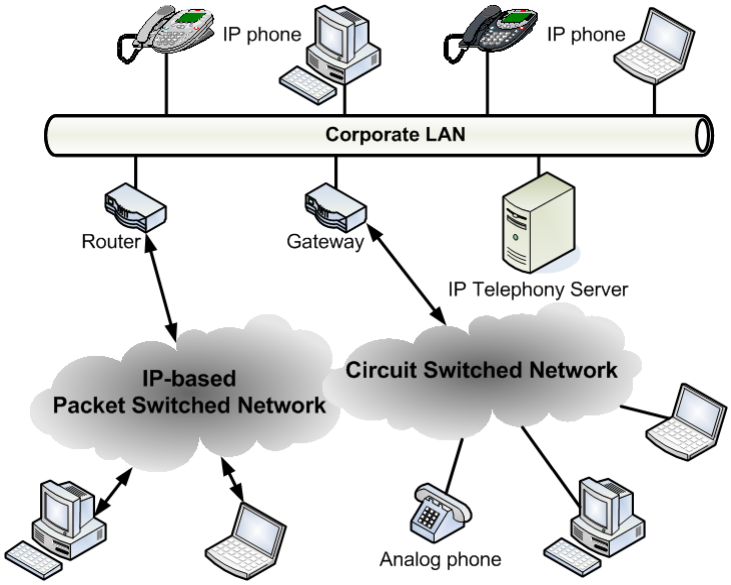

Figure 1 IP telephony system overview [1].

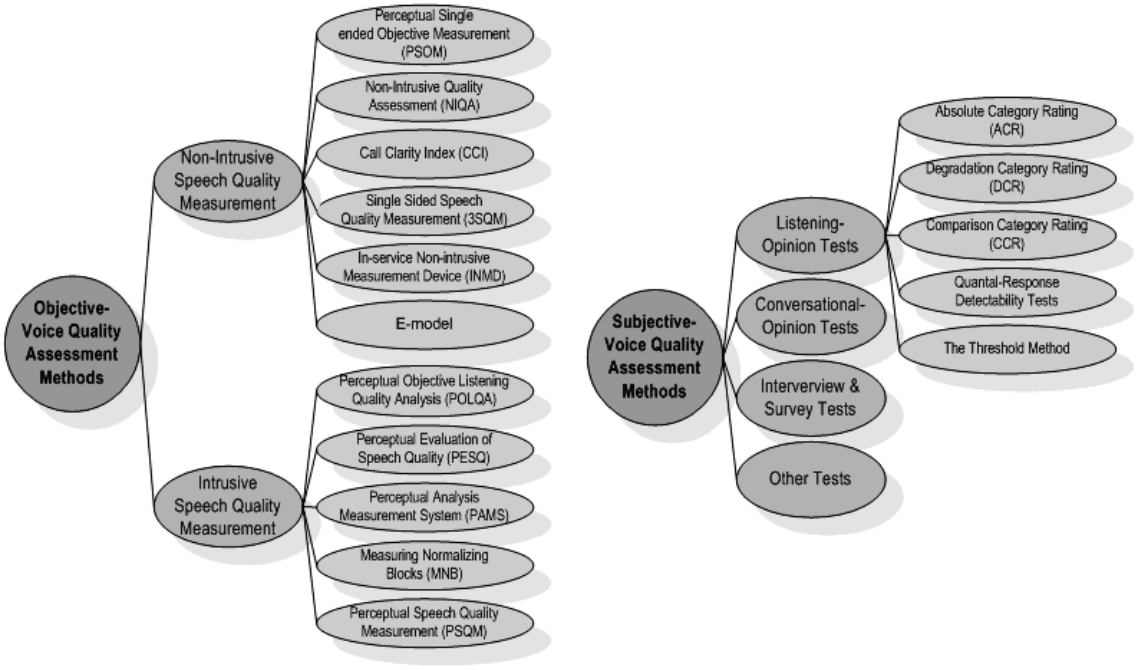

Figure 2 (a) Objective voice quality assessment methods (b) Subjective voice quality assessment methods.

ACR is preferred over DCR and CCR. The limitations of these tests are the requirements for a soundproof room and a complied standard speech set. Besides, each condition of listening-opinion tests generally requires 24 - 32 participants but each one must sit down in a soundproof room one by one per round. This means a lot of time and effort is required. Moreover, in practice, these methods are almost impossible to conduct at a customer site.

For conversation-opinion tests, these tests are also recommended in ITU-T Recommendation P.800 [5]. The most obvious advantages of these methods are being able to gather opinion scores from two subjects at the same round of testing without using a speech set that should comply with the ITU-T standard. Therefore, these methods take a shorter time than listening-opinion tests. However, the disadvantages are the requirement of two separate soundproof rooms. Some examples of scenarios and tasks for these tests can be found in ITU-T Recommendation P.805 [7].

ITU-T Recommendation P.800 also briefly describes interview and survey tests [5]. Some examples of the survey forms are presented in ITU-T Recommendation P.82 [8]. The advantage of these tests is that tests can be conducted outside soundproof rooms. However, it does not mean the tests can be conducted within a very noisy area because noises affect the results of the scores from users. As for disadvantages, each condition of the test requires at least 100 subjects. Therefore a lot of time and effort is required. In particular, collaboration from a lot of subjects can be a problem.

G.711 codec [2,9,10]

G.711 is a narrow band codec developed by ITU-T. It has been classified into u-law and A-law. The u-law is mainly used in the USA, Canada and Japan, while the A-law is used in the rest of the world, including Southeast Asia. Without header and trailer consideration, it consumes about 64 Kbps of bandwidth per channel, which is not appropriate for use over a WAN link. Thus, it is recommended to use in LAN.

Materials and methods

This study used the data from three kinds of subjective assessment methods that consisted of ACR listening-opinion, conversational-opinion and interview tests that have been presented in [11-13]. All tests were conducted using the same testbed system that works as an IP Telephony system, using G.711 codec. It was implemented using Asterisk (open-source software) and IP phones using Session Initiation Protocol (SIP).

However, for the ACR listening-opinion tests, some parts of TSST [14] were used as the speech materials. All tests followed the ITU-T Recommendation P.800, under so called 'direct conditions' as these are the best conditions for the system to provide an accurate experiment and is equivalent to Q_N of infinity as mentioned in ITU-T P.810 [5,15]. The laboratory was at the Central Library at KMUTNB which has a soundproof room (e.g. background noise of < 32 dBA) [16].

The participants were native Thai speakers. Firstly, the listening-opinion test data of 247 valid records were obtained from 32 subjects (17 female and 15 male subjects). Secondly, the conversational-opinion test data of 32 records were obtained from other 32 subjects (14 female and 18 male subjects). Lastly, the interview test data totaling 100 records (40 female and 60 male subjects) were obtained. All were students at KMUTNB, except one person who was a worker.

This study compared the data from three kinds of subjective tests and analyzed the results using ANOVA, a statistical tool, to see if the data from the interview tests were consistent significantly with the data from listening-opinion and conversation-opinion tests, using the hypothesis as follows;

H_0: The subjective MOS values from three assessment methods are the same.

H_1: The subjective MOS values from three assessment methods are different.

Results, analysis and discussion

The data from three tests are presented graphically as in **Figure 3**, whereas, the analyzed results of the hypothesis test using the statistical tool ANOVA are given in **Table 1**.

From **Figure 3**, it is surprising that the subjective MOS values from the conversational-opinion tests and the interview tests are almost the same, being 4.16 and 4.14 respectively. Whereas, the subjective MOS values of listening-opinion test is 4.23, which is the highest. The highest subjective MOS may come from the fundamental of the listening-opinion tests that makes subjects concentrate when alone in a soundproof room, while other tests are more relaxed. Nevertheless, the differences in the results from the different assessment methods may come from the individual variation as mentioned in [17].

However, after the hypothesis testing, the analyzed result shows a p-value of 0.511 that is greater than 0.05 significantly, considering a 95 % confidence interval, as in **Table 1**. Therefore, H_0 is accepted. It means the subjective MOS values from 3 assessment methods are the same, which is consistent with the expectation after gathering the data from all 3 tests for the first time.

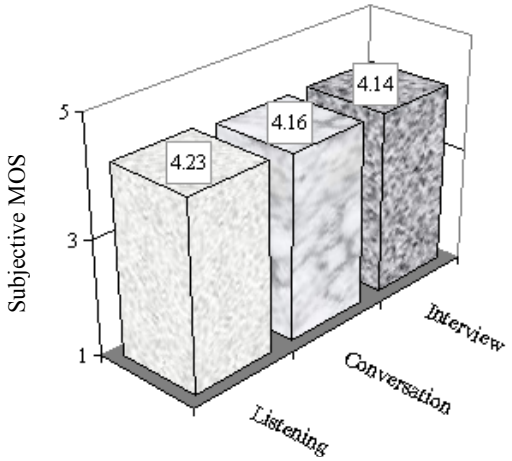

Figure 3 Comparison of MOS values from ACR listening-opinion, conversational-opinion and interview tests with a standard deviation of 0.70, 0.51 and 0.60 respectively.

Table 1 Hypothesis test - result.

Hypothesis	**p-value**
Comparison of subjective MOS values: ACR Listening, Conversation and Interview Tests	0.511

Remark: Significant at p-value < 0.05.

Conclusions

After this study, it can be briefly concluded that interview tests can be used to evaluate voice quality with high accuracy and reliability; as good as the other subjective assessment methods that require soundproof rooms. It follows that voice quality of IP telephony system can be evaluated using interview tests, without the need for expensive-objective measurement tools, particularly useful in developing countries.

Acknowledgements

We thank the anonymous reviewers for very useful comments. Thanks also to Mr. Wiwat Suwanuntawong, of the Central Library Studio, KMUTNB, for support. All the participants who joined the tests are gratefully acknowledged. Lastly, the first author would like to express gratitude to Dr. Gareth Clayton, the advisor who sadly passed away.

References

[1] F De Rango, M Tropea, P Fazio and S Marano. Overview on VoIP: Subjective and objective measurement methods. *Int. J. Comput. Sci. Netw. Secur*. 2006; **6**, 140-53.
[2] S Karapantazis and FN Pavlidou. VoIP: A comprehensive survey on a promising technology. *Comput. Netw.* 2009; **53**, 2050-90.
[3] A Sulkin. *PBX Systems for IP Telephony*. McGraw-Hill, New York, 2002, p. 150-3.
[4] Avaya IP Voice Quality Network Requirements. Available at: http://downloads.avaya.com/css/P8/documents/100018203, accessed November 2012.
[5] ITU-T Recommendation P.800 (Methods for subjective determination of transmission quality). Available at: http://www.itu.int/rec/T-REC-P.800-199608-I/en, accessed November 2012.
[6] T Triyason and P Kanthamanon. Perceptual evaluation of speech quality measurement on Speex codec VoIP with Tonal language Thai. *In*: Proceeding of the 5th International Conference on Advances in Information Technology, Bangkok, Thailand, 2012, p. 181-90.
[7] ITU-T Recommendation P.805 (Subjective evaluation of conversational quality). Available at: http://www.itu.int/rec/T-REC-P.805-200704-I/en, accessed November 2012.
[8] ITU-T Recommendation P.82 (Methods for evaluation of service from the standpoint of speech transmission quality). Available at: http://www.itu.int/rec/T-REC-P.82-198811-W/en, accessed November 2012.
[9] ITU-T Recommendation G.711 (Pulse code modulation (PCM) of voice frequencies). Available at: http://www.itu.int/rec/T-REC-G.711-198811-I/en, accessed November 2012.
[10] Avaya IP Telephony Implementation Guide. Available at: https://downloads.avaya.com/elmodocs2/comm_mgr/r3_1/avaya-iptel-imp-guide3.1.pdf, accessed November 2012.
[11] T Daengsi, C Wutiwiwatchai, A Preechayasomboon and S Sukparungsee. VoIP quality measurement: Insignificant voice quality of G.711 and G.729 codecs in listening-opinion tests by Thai users. *Inform. Tech. J.* 2012; **8**, 77-82.
[12] T Daengsi, C Wutiwiwatchai, A Preechayasomboon and S Sukparungsee. A study of VoIP quality evaluation: User perception of voice quality from G.729, G.711 and G.722. *In*: Proceeding of the 9th Annual IEEE Consumer Communications and Networking Conference, Las Vegas, Nevada, USA, 2012, p. 342-5.
[13] T Daengsi, C Wutiwiwatchai, A Preechayasomboon and S Sukparungsee. Speech quality assessment of VoIP: G.711 VS G.722 based on interview tests with Thai users. *Int. J. Inform. Tech. Comput. Sci.* 2012; **4**, 19-25.
[14] T Daengsi, A Preechayasomboon, S Sukparungsee, P Chootrakool and C Wutiwiwatchai. The Development of a Thai Speech Set for Telephonometry. Available at: http://desceco.org/O-COCOSDA2010/proceedings/paper_53.pdf, accessed November 2012.
[15] ITU-T Recommendation P.810 (Modulated Noise Reference Unit (MNRU)). Available at: http://www.itu.int/rec/T-REC-P.810-199602-I/en, accessed November 2012.
[16] T Daengsi and K Tontiwattanakul. A case of improvement of building acoustics using available equipments and limited resources. *In*: Proceeding of the 6th Naresuan Research Conference, Phitsanulok, Thailand, 2010, p. 2-13.
[17] AW Rix. Comparison between subjective listening quality and P.862 PESQ score. Available at: http://82.165.4.28/com/docs/downloads/pesq/pdf/wp_sub_v_pesq.pdf, accessed November 2012.