VoIP Quality Measurement: Insignificant Voice Quality of G.711 and G.729 Codecs in Listening-Opinion Tests by Thai Users

Therdpong Daengsi1*, Chai Wutiwiwatchai**, Apiruck Preechayasomboon***, and Saowanit Sukparungsee****

Abstract

This paper presents an introduction about the Voice over Internet Protocol (VoIP) technology and some background infor-mation about fundamental frequency of Thai language and some different characteristics of fundamental frequency of child voice, male voices and female voices. Also, voice quality indicator and an overview of voice quality measurement methods have been presented. However, only subjective measurement methods were considered in detail. Moreover, brief information about voice codecs has been described. For the approach of this study, the data set from listening-opinion tests with G.711 and G.729 has been used and analyzed. Analyzed results from T-test and ANOVA show that G.711 and G.729 provide voice quality without difference significantly, although in cases of voices from children, female and male speakers. Therefore, this paper presents the evidence which is inconsistent with the general understanding of most engineers, who implement VoIP systems.

Keywords: VoIP Quality Measurement, Listening-Opinion Tests, G.711, G.729.

1. Introduction

At present, information and communication technology (ICT) are developing rapidly. Therefore, in the world of competition, organizations that can efficiently apply ICT usually have more of a competitive advantage. VoIP technology is a kind of ICT, which is a combination of speech communication technology, called voice communication, and computer technology, called data communication. Previously, voices flow in its own networks called public switched telephone networks (PSTNs), whereas data flow in their own networks is called "data communication networks". After emerging of Internet, telecom engineers, computer scientists and researchers applied the data communication network to carry voice over IP networks to avoid call costs from PSTN operators and also to utilize the data network that might be available.

Like other countries, VoIP has been used in Thailand broadly recently. For personal use, some people know and use VoIP applications such as, Skype and Google Talk, to chat with friends as overseas calls or long distance calls. For private sector, some companies use IP telephony systems in some parts of their organizations already [1], [2]. Whereas some organizations of government such as Thai Customs Department, Ministry of Finance, has implemented a VoIP system with over 2,000 IP phones that covers over 70 customs houses nationwide, with the expectation to reduce telephone cost at least 30% [3]. For the side of operators, including TOT, True and CAT, they issued their VoIP services already, particularly, for overseas call services such as TOTnetcall and 008 by TOT, NetTalk by True and 009 by CAT [4], [5], [6], [7]. However, the direction of VoIP services in Thailand is controlled appropriately. Of course, this role is acted by the regulator, called National Broadcasting and Telecommunications Commission (NBTC), the National Telecommunications Commission (NTC) former. NBTC holds a numbering plan that supports the numbering of VoIP technology especially, consisting of 061-xxx-xxxx to 065x-xxx-xxxx and 067-xxx-xxxx to 068-xxx-xxxx [8].

However, there are limitations when using VoIP. One of those is voice quality which can be degraded by packet loss and packet delay in IP networks. It is mentioned that people can perceive the voice delay if packet delay is more than 150 ms and may want to stop conversation if packet delay is more than 400 ms, whereas, it can be noticed if packet loss is more than 1% and it may not be unacceptable if packet loss is more than 3% [9]. To compensate effects from issue in IP networks,

* *Faculty of Information Technology, King Mongkut's University of Technology North Bangkok.*

** *Human Language Technology Laboratory (HLT) National Electronics and Computer Technology Center.*

*** *TOT Innovation Institute, TOT Public Company Limited.*

**** *Applied Statistics Dept., Faculty of Applied Science, King Mongkut's University of Technology North Bangkok.*

selection of codec that can provide high voice quality can be applied. Nevertheless, G.711 codec is generally used for LAN, while G.729 codec is generally used for WAN. According to the information presented in [9], it is mentioned that the mean opinion score, which is the scale of voice quality, of G.711 is 4.1, while, the score of G.729 is 3.92. However, it can be implied that the results were based on English, which is a non-tonal language, unlike Thai. Therefore, this paper presents the background information about tonal feature of Thai language, characteristics of voice of different types of speakers, official scale of voice quality and voice quality measurement methods, before presenting the comparison between voice quality of G.711 and G.729. The data set from listening-opinion tests, based on Thai, was analyzed to investigate if perceptual voice quality of G.711 is better than G.729 as mention in [9] or not, when carrying different types of voices that recorded by different types of Thai native speakers, consisting of children, adult female and male speakers, to Thai native speakers.

2. BACKGROUND

2.1 Thai Language and Fundamental Frequency

Like Mandarin-Chinese, Vietnamese and Laos, Thai is one of the tonal language in the world, used as the official language in Thailand. There are five tones, consisting of middle, low, falling, high and rising [10]. Tone is about changing of fundamental frequency (F0) and each tone has different F0 pattern, as shown in Figure 1 [11]. Therefore, instead of using the term, 'a tonal language', it could be said that Thai is an F0 sensitive language.

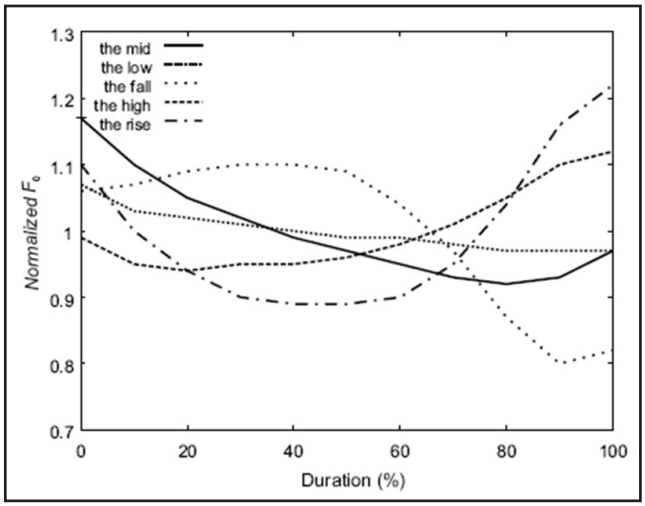
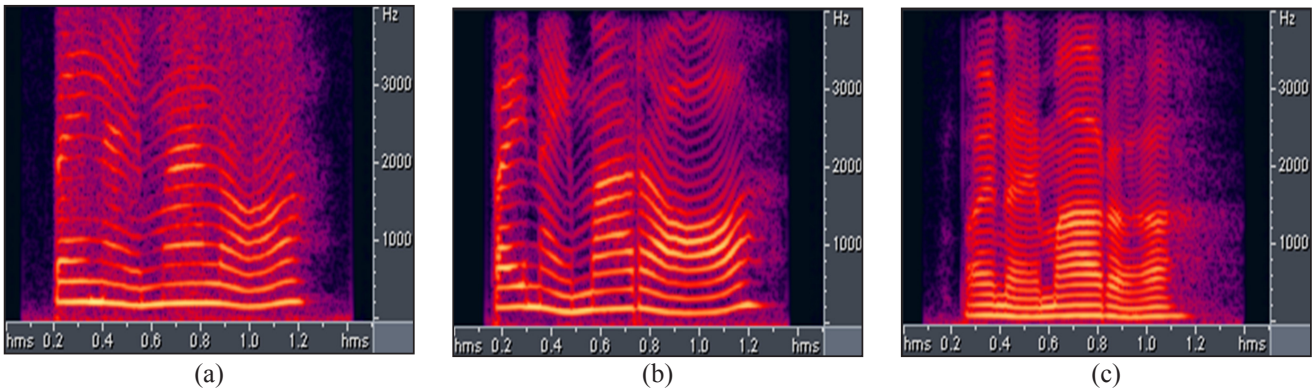


Figure 1 *An example of five tone patterns in Thai language using normalized F0.*

There was a comparison between differences in spoken frequency of a female group and a male group who use Mandarin [12]. It has been found that the highest and lowest ranges of F0 for the female group were greater than the male group. Also, there is a study on F0 contours of Thai expressive speech using Fujisaki's model [13]. It presented the results which could be implied that the F0 of female voices are higher than male voices significantly. Not only F0 but also other characteristics of tone and phase of those are different. Moreover, when considering the spectrum view of the same spoken sentence from a child, an adult female and male speaker, using audio files which are parts of Thai Speech Set for Telephonometry (TSST) [14], the result is consistent with the previous studies because it can be seen that there are differences of frequency characteristic of a spoken sentence by those, as in Figure 2.



(a) (b) (c)

Figure 2 *Examples of spectrum view of narrow band voices by (a) a girl (b) an adult female and (c) an adult male speaker respectively, when saying the Thai sentence "ไปไหนมาเหรอ" (Paj Nai Ma Roi) which means "Where have you been?". It can be seen that the most contours of voice frequencies of (a) shows the highest, whereas (c) shows the lowest.*

2.2 Voice Quality Indicator [15], [16]

The term "quality" of one person may differ to another person. Besides, its meaning in one culture might be different too. That means it could be very, depending on, for example, individuality and culture. Moreover, balance of conditions for VoIP measurements in different laboratories may reflex different perception of quality. For quality of service of a network, the indicators to tell an administrator are, for example, packet delay, packet loss and jitter. However, ITU-T recommended using the voice quality indicator, called mean opinion score or MOS, in short, to evaluate voice quality. Basically, MOS is obtained by evaluation subjectively. A group of people, generally 24-32 subjects, are required to vote the quality of voice from both female and male speakers, via a listening system (e.g. VoIP system). The score for voting could be presented as in Table 1,. After voting by enough number of subjects, the mean of score would be calculated. That is the source of its name.

Table 1 *Scale of opinion scores and meaning [16].*

Opinion Score	Meaning
5	Excellent; no perceptible impairments
4	Good; barely perceptible but not annoying impairments
3	Fair; perceptible and slightly annoying impairments
2	Poor; annoying but not objectionable impairments
1	Bad; very annoying and objectionable impairments

2.3 Overview on VoIP Quality Measurement Methods

There are two main methods for VoIP quality measurement. The traditional methods are subjective methods that require a group of subject to participate in a laboratory to evaluate the voice quality, provided by the system with a condition that is interested in. Whereas, the objective methods are the modern methods, that are more popular than the traditional methods at present. However, each main method has sub-methods, as presented in figure 3 [17], [18]. While the comparison, between these two main methods, is presented in Table 2.

Although the objective methods are very popular, they cannot work accurately and reliably without calibrating with the results from the subjective tests. Therefore, subjective methods are very importance. According to figure 3, there are three main subjective methods that they can be described as follows:

A. Conversation-opinion tests

ITU-T recommended in [17] about the test facilities, experiment design, conversation task and test procedure for

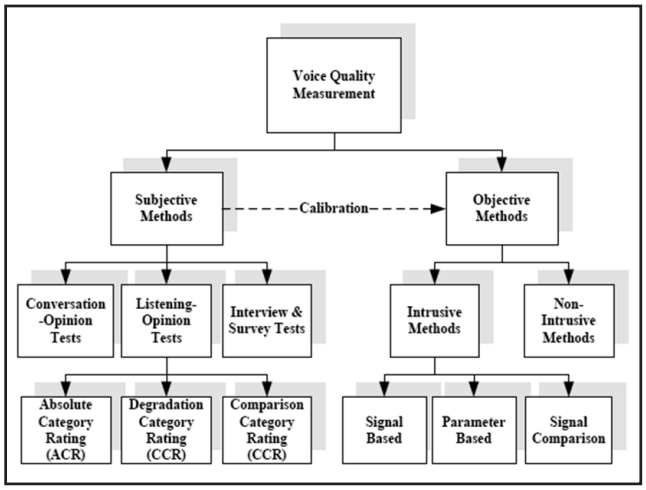


Figure 3 *Voice quality measurement overview.*

Table 2 *Subjective measurement methods VS objective measurement methods.*

	Voice Quality Measurement	
	Subjective	**Objective**
Accuracy and Reliability	High	Medium - high
Management skill requirement	High	Low
Endeavor requirement	High	Low
Automatic measurement	No	Yes
Special test facilities requirement	Soundproof room(s)	objective measurement tool (e.g. E-model measurement tool)
Time consumption	Very long (e.g. 5 minutes per participant)	Short
Collaboration requirement	High (e.g. 24-32 participants per condition)	Low
Cost	High (for conducting subject to participate, to employ a research assistant and to prepare standard test facilities, e.g. a sound proof room)	High (for a standard tool, e.g. E-model measurement tool)

conversation-opinion tests. These kinds of subjective method are the most realistic. These can be used to test voice quality under the conditions of delay and echo that might degrade voice quality, while two telephone users are talking together. In the test, two inexperienced subjects are recommended to conduct to sit in separate soundproof room, before conversation via IP phone. That means, these methods require two good soundproof rooms. Of course, high cost is required if it is about building two soundproof rooms with very low reverberation time and very low room noise. However, for

pros of conversation-opinion tests, these methods are not necessary to prepare a speech database.

B. Listening-opinion tests [14], [17], [19]

These kinds of tests have been classified into absolute category rating (ACR), degradation category rating (DCR), and comparison (CCR), as in Figure 3. Although listening-opinion tests cannot reach the realism as conversation-opinion tests, ITU-T recommended using ACR for listening-opinion tests. However, a difficulty of listening-opinion tests is about source recordings of speech set for testing that requires, for example, very good recording environment (e.g. reverberation time is less than 300 ms and the room noise is lower than 30 dBA), good recording system, and appropriate speech material. However, instead of recording source of speech by self, there are available speech database. For example, Multi-Lingual Speech Da-tabase for Telephonometry 1994, that was developed by NTT-Advanced Technology Corporation in Japan. This speech database contains 21 language of speech, including Thai. Whereas, Thai Speech set for Telephonometry (TSST) that was developed under funding support of Human Language Technology Laboratory (HLT) of NECTEC, is an option without charge.

C. Interview and Survey Tests [17]

These kinds of tests are options to evaluate voice quality with subjects. These can be used to interview and/or survey from the real users. For example, after finishing the new VoIP system installation for a customer, the implementer might use the survey tests to obtain the evidence from users, about voice quality provided by the new VoIP system, in order to hand-over the VoIP system to a customer. However, a lot of endeavor is required to gather opinions about voice quality from at least 100 interviewees that is the trade-off with conducting outside a soundproof room.

2.4 Voice Codecs

Voice Codec is an important part of VoIP applications because it is used to change voice signals into voice packets at the source before being carried via IP networks, and change them as voice signals at the destination. In telecom industries, G.711 is usually the first choice in local area networks (LANs), such as, within the same office and/or the same building. While G.729 is usually the first option for WAN, such as, between two branches in different cities, because its payload requirement is only 8 kbps, while the requirement of G.711 is 64 kbps [20]. Generally, engineers, who implement VoIP systems, under-stand that G.711 provides better voice quality than G.729. Therefore, in this study, these two codecs were focused. Details about G.711 and G.729 can be presented as in Table 3. Normally, each VoIP application has a default codec.

3.1 Materials

In this study, some part of the data set from the listening-opinion tests, that was conducted at the good standard environment (reverberation time < 300 ms and room noise < 35 dBA) with the best possible condition of VoIP system by applying some audio files from TSST, were used to investigate the perception of voice quality of Thai subjects that provides by G.711 and G.729. For the procedure of listening–opinion tests to create the data set, each subject had to listen to a speech list, which consists of 10 different speech groups from 10 speakers (2 children, 4 female and 4 male speakers), that was given randomly once, and evaluated it using a paper-based form. Therefore, each subject gave 10 values of opinion scores, as presented in the Appendix. For the subjects who joined the test, the majority were undergraduate students from the faculty of science, KMUTNB. The majority of them were about 20-24 years old. However, the outliers have been defined and discarded (e.g. the highest and the lowest average values of speakers for each codec) before analyzing the data set.

3.2 Methodology

Both G.711 and G.729, the data from different types of voices, consisting of voices of children, voice of adult female speakers and voices of adult male speakers have been

Table 3 *Characteristics of G.711 and G.729 [9].*

Codec	Bitrate (kbps)	Frame (ms)	Bits per frame	Algorithmic delay (ms)	Codec delay (ms)	Compression algorithm	Complexity (MIPS)	MOS
G.711	64	0.125	8	0.125	0.25	PCM	< 1	4.1
G.729	8	10	80	15	25	CS-ACELP	< 20	3.92

Table 4 *Analyzed results about G.711 vs G.729*

Hypotheses	p-value
H1: MOS of G.711(all) vs MOS of G.729(all)	0.443
H2: MOS of G.711(Children) vs MOS of G.711(female) vs MOS of G.711(male) vs MOS of G.729(Children) vs MOS of G.729(female) vs MOS of G.729(male)	0.066

Note: Difference is significant if p-value < 0.05

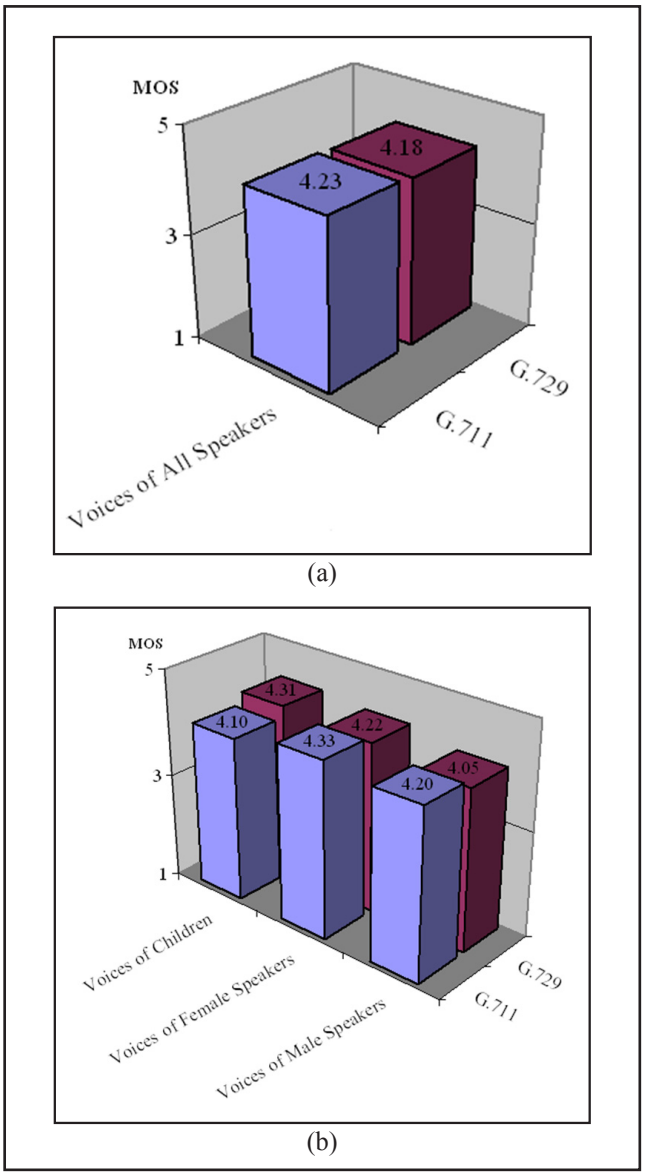


Figure 4 *MOS of G.711 VS G.729 from (a) all peakers and (b) separated types of voices.*

analyzed. Of course, it was analyzed if each codec provides different types of voices with different voice quality significantly or not. T-test and ANAOVA, statistic tools were selected to analyze the hypotheses as follows:

$H1_0$: The mean opinion scores of Thai subjects to voices of all speakers provides by G.711 and G.729 are the same.

$H1_1$: The mean opinion scores of Thai subjects to voices of all speakers provides by G.711 and G.729 are different.

$H2_0$: The mean opinion scores of Thai subjects to voices of children provides by G.711, voice of children provides by G.729, voices of female speakers provides by G.711, voices of female speakers provides by G.729, voices of male speakers provides by G.711, and voices of male speakers provides by G.729 are the same.

$H2_1$: The mean opinion scores of Thai subjects to voices of children provides by G.711, voice of children provides by G.729, voices of female speakers provides by G.711, voices of female speakers provides by G.729, voices of male speakers provides by G.711, and voices of male speakers provides by G.729 are different.

4. ANALYZED RESULT

From the data, the average scores of G.711 and G.729 could be presented as in figure 4 (a), while figure 4 (b) presented average scores for each type of voices.

In figure 4 (b), it can be seen that even MOS from all speakers or each type of speakers are different but not much. Therefore, the hypothesis H0 must be analyzed, the analyzed results presented in Table 4.

5. DISCUSSION AND CONCLUSION

Normally, each hypothesis was tested, if it reaches 95% confidence or not. If p-value is more than 0.05, H0 would be re-jected and H1 would be accepted, before describing that they are not different significantly. Therefore after considering the p-values in Table 4., both H1 and H2, they confirmed that there is no significant between MOS of G.711 and MOS of G.729, even though they have been used to carry different types of voice that p-value = 0.066.

After analyzing the data set without the outliers, gathered from the testing of perceptual voice quality provided by G.711 and G.729, with Thai native listeners by listening to three types of voices from 8 speakers (2 children, 3 female speakers and 3 male speakers), it could be concluded in this paper that there is no significant difference of voice quality. Therefore, based on Thai users, it is strongly recommended that G.729 can be used instead of G.711, in order to reduce voice traffic in IP net-works,

6. ACKNOWLEDGMENT

Gratitude is due to Dr. Gareth Clayton, the advisor of the first author who sadly passed away last year. Thank you very much all participants who joined the subjective-listening tests, without you all, the data set would not have been created and this paper could not have been written. Lastly, thank you very much Mr. Gary Sherriff, the international coordinator in the faculty of information technology, KMUTNB, who is always kind to the first author, for editing.

7. References

[1] T. Daengsi and A. Preechayasomboon, "Case Study: AIA Insurance – Migration Project Experience," *NCCIT 2008*, Mahasarakam, Thailand, May, 2008.

[2] T. Daengsi and A. Preechayasomboon, "Case Study: AMEX Thailand - PABX Migration Experience," *NCCIT 2009*, Bangkok, Thailand, May, 2009.

[3] Thai Customs Department, "Biddin, "http://www.customs.go.th/UploadFile/Bidding/B0510010.pdf, Jul, 2011.

[4] TOTnetcall, "About TOTnetcall". http://www.totnetcall.com/VoIP/tdetinfo.aspx, Jul, 2011.

[5] TOT, "008". http://www.tot.co.th/index.php?option=com_linkcontent&categoryid=90&Itemid=128&lang=en, Jul, 2011.

[6] Truenettalk, "About NetTalk". http://www.truenettalk.com/th/about_truenettalk/what_is_truenettalk.html, Jul, 2011.

[7] CAT Contact Center, "CAT009 Service". http://www.contactcenter.cattelecom.com/thai/oversea/cat009_info.asp, Jul, 2011.

[8] NTC, "NTC Announcement: Numbering Plan". http://www.ratchakitcha.soc.go.th/DATA/PDF/2549/00189184.PDF, Jul, 2011.

[9] S. Karapantazis and F.-N. Pavlidou, "Voip: A comprehensive survey on a promising technology," *Computer Networks*, vol. 53, no. 12, pp. 2050-2090, Aug, 2009.

[10] C. Wutiwiwatchai and S. Furui, "Thai speech processing technology: A review," Speech communication, Vol. 49, pp. 8-27, Jan, 2007.

[11] N. Thubthong, "A study of various linguistic effects on tone recognition in Thai continuous speech," Ph.D. Dissertation, Chulalongkorn Univ. Bangkok, Thailand, 2001

[12] S. H. Chen, "Sex differences in frequency and intensity in reading and voice rang profiles for Taiwanese adult speakers," *Folia Phoniatr Logop*, Vol. 59, pp. 1-9, 2007.

[13] S. Chomphun, "Analytical study on fundamental frequency contours of Thai expressive speech using Fujisaki's model," *Journal of Computer Science*, Vol. 6 (1), pp. 36-42, 2010.

[14] T. Daengsi, A. Preechayasomboon, S. Sukparungsee and C. Wutiwiwatchai, "The development of a Thai speech set for telephonometry," *O-COCOSADA2010*, Kathmandu, Nepal, Nov, 2010.

[15] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," Psytechnics, Sep, 2003.

[16] Detech Networks, "Voice quality beyond IP QoS," White paper, Jan, 2007

[17] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality", Aug, 1996.

[18] Z. Ren and H. Wang, "Chapter 8 Measurement and analysis on the quality of Skype VoIP," *VoIP Handbook: Applications, Technologies, Reliability*, and Security, CRC Press, 2009.

[19] NTT-AT, "Multi-lingual speech database for telephonometry 1994". http://www.ntt-at.com/products_e/speech/index.html, Jul, 2011.

[20] Avaya Labs., "Avaya IP Voice Quality Network Requirements," Avaya Inc, CO, Apr, 2006.