A Framework to Analyze Quality of Service (QoS) for Text-To-Speech (TTS) Services

Mohd Farhan Md Fudzee[1], Mohamud Hassan[1],
Hairulnizam Mahdin[1](✉), Shahreen Kasim[1], and Jemal Abawajy[2]

[1] Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia
{farhan, hairuln, shahreen}@uthm.edu.my

[2] School of Information Technology, Deakin University, Burwood, Australia
jemal@deakin.edu.au

Abstract. Quality of service (QoS) evaluation is vital for text-to-speech (TTS) web service applications. Most of the current solutions focus on either evaluating functional or nonfunctional attributes of the TTS. In this paper, we propose a QoS framework to evaluate and analyze the perceived QoS that combines general and specific mechanisms for measuring both functional and nonfunctional requirements of speech quality. General mechanism measures the response time of TTS services while specific mechanism measures intelligibility and naturalness through subjective quality measurements, which are mapped onto mean opinion score (MOS). The result shows the workability of the framework, tested by predetermined users to three services: service1 (Fromtexttospeech) resulting 47.84%; service2 and service3 (NaturalReader and Yakitome) are 31.62 and 21.53% respectively. The TTS services evaluation can be to enhance the user experience.

Keywords: Quality of Services (QoS) · Text To Speech (TTS) · Mean Opinion Score (MOS) · Intelligibility · Naturalness · Response time · Quality attributes

1 Introduction

Text to speech systems are vital in our everyday activities. The use of speech synthesis and building voices became common due to the rapid advancement in information technology and communications. The Voice User Interface (VUI) plays huge role in technology such as computer systems, mobile multimedia and voice-enabled equipment. Speech Understanding and Synthesis Technology are among the frequently used technology to support users. Ultimately, high quality synthesized outputs are preferred. Thus, evaluating the quality of web text to speech (TTS) services are important. TTS is useful in the areas like disabled, education, consumer, computer interface and telecommunications. The Quality of Service (QoS) in multimedia conversion of TTS examines the performance in terms of accessibility, media conversion availability, conversion accuracy, and user satisfactions. QoS determines how well a service performs while functionality determines what a service does [1, 2].

T. Herawan et al. (eds.), *Recent Advances on Soft Computing and Data Mining*,
Advances in Intelligent Systems and Computing 549, DOI 10.1007/978-3-319-51281-5_59

Quality evaluations of TTS web services can be classified into functional and nonfunctional requirements. Functional requirement focuses on what TTS service does, while nonfunctional requirements also known as quality attributes is used to determine the quality of services requirements. There are general and specific QoS mechanisms to evaluate functional and nonfunctional requirements of TTS services. Currently, there is less effort to integrate both mechanisms into single solution to analyze QoS from the end-user perspective to provide users with capacity to enhance their experience.

In this paper, a QoS analysis framework for text to speech services is proposed. This framework is aimed to analyze and examine the quality of services of the multimedia conversion text to speech on the web and measure TTS performance in term of content accessibility, response time, and voice intelligibility and naturalness by comparing three web TTS services to enhance the quality of experience of the online users. The remainder of this paper is structured as follows. A brief review of previous works is discussed in Sect. 2. In Sect. 3, we describe the proposed framework. The results of the web QoS for TTS are analyzed in Sect. 4. Finally, some concluding remarks are given in Sect. 5.

2 Related Work

Recently, a number of works have focused on developing subjective QoS evaluation frameworks. For instance, a probabilistic model was introduced by Wang et al. [3]. Based on the received speeches, the system will calculate the confidence score of multiple different levels using a constrained generalized posterior probability (CGPP) algorithm. This method calculates the received speech input in multiple different levels (e.g. phoneme, syllable, and word and/or utterance level) using CGPP algorithm. This QoS service evaluation needs complex data processing, which consumes a lot of time. Also, it is very expressive for end users.

On the other hand, Remes et al. proposed frequency-weighted segmental Signal-to-noise ratio (SNR) quality measurement that exhibited a performance using standardized perceptual evaluation of speech quality (PESQ) objective evaluation measure [4]. It will allow capturing the automated quality evaluation. This method has a drawback for its relaying automation but it performs when it comes to intelligibility and naturalness, which reflect the user experience. Other author developed TTS quality evaluation using E-model. E-model is a computation model which takes into account all links between transmission parameters [5]. This model requires the individual transmission path parameters not being assessed separately but rather all their possible combinations and corresponding interaction are considered. This can be achieved using quality estimation based on system approach of computation model. The computation model approach has downward because it depends on transmission path instead of the experience of the quality services.

As mentioned earlier, the significant issue of TTS services is QoS. Even though there are some researches on how to determine the qualities of TTS media conversions however, previous researches on quality have concentrated only on server side and media conversion stage which involves implementation part of text to speech applications. Alternatively, we highlight the crucial needs for a model to be used as guidelines for service users to get the best TTS application services and enhance the user experience. It is not easy for Internet users to captivate the TTS online services. Thus, it's very significant to write a generic tool that provides client-side measurements for the performance of the TTS services available online [6].

3 TTS Evaluation Framework

In this section, the framework for QoS analysis for TTS media conversion has been proposed. The two elements in this framework are user's requirement and perception to deliver quality of services of text to speech services. To evaluate non-functional quality attributes for TTS web services, we analyze and measure general and specific QoSs. The proposed framework gives great flexibility, which provides analytical solutions for the end users as illustrated by the Fig. 1.

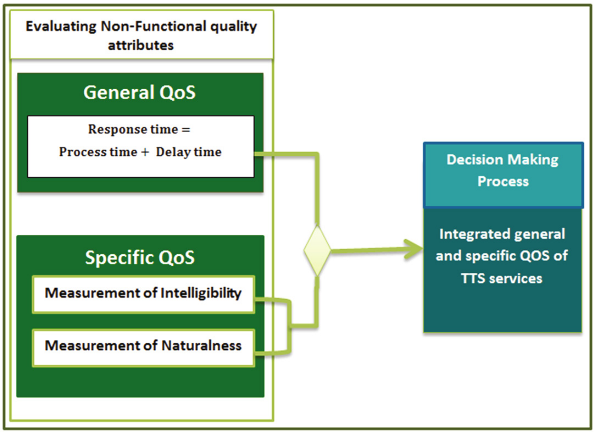
Evaluating Non-Functional quality attributes
General QoS
Response time =
Process time + Delay time
Specific QoS
Measurement of Intelligibility
Measurement of Naturalness
Decision Making Process
Integrated general and specific QOS of TTS services

Fig. 1. Proposed evaluation framework for QoS of TTS.

3.1 General QoS Attributes

General QoS attributes estimates the response time of the services. Measuring response time performance is one of the most significant factors in the servers since service providers need to manage and enhance the delay.

$$\mathrm{R(t)} = P(t) + D(t) \tag{1}$$

where R(t) = response time, P(t) is process time and D(t) delay time.

3.2 Specific QoS Attributes

The second part of the framework is the specific QoS attributes, that consists of evaluations for intelligibility and naturalness of synthetic speech using subjective measures. In this study we use Mean Opinion Scores (MOS) in calculating and measuring the performance of the intelligibility and naturalness of TTS service.

The framework combines the attributes of general and specific QoS. Specifically, the response time is captured for general QoS while a five point mean opinion score is used to estimate the scale for user experience of the intelligibility and naturalness of text to speech (TTS) services.

3.3 Experimental Setup

To implement the QoS analysis for TTS online services, we consider three TTS online services to examine the quality of the speech that online system generates. Table 1 shows the list of TTS services that we use to examine the QoS analysis. We compare these services in term of file size, duration, delay time and then the response time is calculated.

Table 1. List of used TTS services.

Service name	Link
FromTextToSpeech	http://www.fromtexttospeech.com/
Natural Reader	http://www.naturalreaders.com/
Yakitome	https://www.yakitome.com/upload/from_text

3.4 Calculating Response Time

The length of the TTS speech is the duration of the audio where it is achieved by capturing the length or the duration of the audio that have been converted. An array of arbitrary parameter values from starting time of the speech and end of the speech audio time for the duration is used. The duration parameter is the amount of time speech in seconds values that will be calculated according to the parameter values. To get the response time of the speech of the TTS service, we calculate the processed time and deduct the delay time of the audio speech read.

$$\mathrm{R(t)} = P(t) + D(t) \tag{2}$$

R(t) represents the response time of the converted speech in seconds, P(t) is the process time or the duration time of the TTS in seconds and the D(t) stands for the amount of seconds delayed from the started period.

3.5 MOS Design and Structure

Mean opinion scores (MOS) of quality of services (QoS) analysis measure Listening Quality of mean opinion score (MOS-LQ) of the speech audio that is being listened by the users [7]. This value takes into consideration the speech intelligibility and naturalness and from this data calculates how users would rate the audio quality they hear. MOS-LQ scale is divided into two major areas, intelligibility and naturalness of the text to speech application. To minimize the complex nature of the MOS, we use various question logic and rating scales ranging from "strongly agree" to "disagree" and "not at all" to "very often" as part of ITU recommendation.

3.5.1 MOS Intelligibility

Mean opinion scores (MOS) consist of five parts that focus on intelligibility of the web speech for TTS services. Intelligibility test will measure how the speech is comprehensible to the users and in what degree that can the TTS speech be understood. The users will scale questions on quality of services on spoken clarity, explicitness, comprehensibility, perspicuity, and precision. Intelligibility consists of five parts which are:

- Overall impression: rates the quality of the TTS audio
- Listening effort: rates the effort that are required to make in order to understand the message
- Pronunciation: irregularities in pronunciation of the TTS speech
- Speaking rate: rates the average delivery of the speed of TTS speech
- Pleasantness: measures the pleasantness of the voice.

3.5.2 MOS Naturalness

Questions on naturalness will examine the TTS web service's ability on how it is perceived as natural degrade human perception of quality. Naturalness essentially considers if the text is read in a natural human sounding manner. Naturalness questions consist of:

- Naturalness: rates the overall quality of the naturalness of the audio.
- Audio flow: rates the continuity or flow of the audio.
- Ease of listening: how easy or difficult to listen to the voice of TTS for long period of time.
- Comprehension problems: considers if it finds certain word hard to understand.
- Articulation: if the sounds in the audio distinguishable
- Acceptance: do you think that this voice can be used for TTS service users.

Client rates the services using MOS scale and the client scaled the preferred services by comparing the response time of the services. Table 2 shows some examples of categorization of quality of services (QoS).

Table 2. Quality of Services (QoS) categorization.

QoS attributes	Service's QoS
Response time	Negative relations
MOS	Rating

MOS data is computed as normalized score between 1 to 5 to calculate the aggregate score and represented as following [8]:

$$n_l = \sum\nolimits_{m=1}^{j} q_m \times w_m \tag{3}$$

n_l is the node score for response time of the services, where q_m is the quality of services (QoS) relation associated from the rated data and w_m denoted as weight of the QoS relation.

$$q_m = \frac{v_i min}{v_i} \quad if\, v_i \min > 0 \tag{4}$$

where v_i represents the quality of services (QoS) value for specific services *(i).* In the case for $if\ \ v_i \min > 0$ response time must be greater than zero, because it is impossible to have zero or less than zero for the response time. q_m is quality measurement of specific services.

Twenty users were initially asked to evaluate three online TTS services, from which all of the participants completed the quality evaluations of TTS services. The participants consisted undergraduate as well as post graduate students and independent parties, and were asked to take MOS questions to scale the quality of services for three different TTS web services. The experiment was carried out in the university in non-soundproof space with basic laptop, headphone and Wi-Fi internet to do the quality analysis testing for TTS services [9].

4 Result

To implement the QoS analysis, we consider 3 TTS online services to examine the quality of the speech that online system generates. Table 1 shows the list of TTS services that we use to examine the QoS analysis. We compare these services in term of file size, duration, delay time and then the response time is calculated.

4.1 TTS Speech Test

Application Response records TTS speech response times for actual end-user activity to provide insight to the user's experience. It collects the speech data file name, size and processing time. After receiving collected information about converted speech file it calculates the delay time and response times to compare the performance of the text to speech (TTS) web services. Table 3 shows the example of the mean for response time. In this experiment, the file size used are (408, 591 and 151) KB.

4.2 MOS Analysis Result

The outcome of the mean opinion scores (MOS) is analyzed by comparing the three different services. The main aim is to give the users an overall outlook of the quality of

Table 3. Quality of Services (QoS) categorization.

Service	Process time	Delay time	Response time
fromtexttospeech.com	00:1:12	00:00:6	00:1:18
Natural Reader	00:1:15	00:00:17	00:1:32
Yakitome	00:1:35	00:00:26	00:2:01

the services (QoS) of the TTS web applications. The collected user observation on TTS services aimed to measure the intelligibility, naturalness and reading comprehension of the web text to speech TTS service.

4.2.1 MOS Intelligibility Result Analysis

Mean opinion score for the text to speech performance summarized in the Table 4 below. The overall average score is shown from 1 to 5 MOS scale.

Table 4. MOS intelligibility performance of compared online TTS services.

MOS intelligibility	fromtexttospeech.com	Natural Reader	Yakitome
Overall impression	4.2	2.2	1.4
Listening effort	3.4	2.3	2
Pronunciation	3.5	1.2	3
Speaking rate	3	1.5	2
Pleasantness	3.5	2	2.3

Table 4 illustrates that the overall impression of the quality of the TTS speech, fromtexttospeech.com is high compared to other services by obtaining 4.2 while Natural Reader and Yakitome obtain 2.2 and 1.4 rate respectively. Intelligibility performance for fromtexttospeech.com service surpasses both Natural Reader and Yakitome on listening effort, pronunciation, speaking rate, as well as pleasantness. MOS Naturalness result analysis.

Table 5 shows the metric performance mean opinion score (MOS) for naturalness. It compares all three text to speech web services (fromtexttospeech.com, Natural Reader and Yakitome).

Table 5 shows that text to speech (TTS) services that we used for this testing. Although the two services have the same audio flow Fromtexttospeech: 3.5, NaturalReader: 3.5 while Yakitome: 3.5. The TTS web service Fromtexttospeech has more human understanding and it's more preferred by the participants compared to Natural Reader and Yakitome. NaturalReader has the least comprehension rate, where it contains certain words which are very difficult to understand. Although, Yakitome, is less articulate than the rest of web TTS services (Articulation: Fromtexttospeech: 4, NaturalReader: 2.4 and Yakitome: 2.1). It is more efficient that Fromtexttospeech in overall quality of services in MOS naturalness. Participants can easily read and understand the outcome of the quality of service analysis for TTS web services and suggested that Fromtexttospeech voice can be useful for TTS service users.

Table 5. MOS naturalness performance of compared online TTS services.

MOS intelligibility	fromtexttospeech.com	Natural Reader	Yakitome
Naturalness	3.6	1.5	2.2
Audio flow	3.5	3.5	3.2
Easy listening	4	3	2.3
Comprehension	2.5	2	1
Articulation	3	1.2	1
Acceptance	4	2.4	2.1

4.3 Overall Quality Analysis

The result of the TTS services comparison shows that the overall quality of services (QoS) performance where Q = {intelligibility, naturalness, responseTime} such that intelligibility = (service1 = 1, service2 = 0.75, and service3 = 0.25), naturalness = (service1 = 1, service2 = 0.7, and service3 = 1), responseTime = (service1 = 1, service2 = 0.85, and service3 = 0.64), response time has negative relationship quality attributes while intelligibility and naturalness both QoS attributes are based on rating [10]. All QoS attributes are shown in Table 6 below.

$$AgS(q) = \sum\nolimits_{m=1}^{k} n_m \tag{5}$$

where *AgS(q)* is the aggregate score for given quality attribute, *k* is maximum number of *q*.

The overall quality observations and the conclusion indicates that service1 (Fromtexttospeech) has 47.84% is acceptable TTS service provider where service2 and service3 (NaturalReader and Yakitome) are close with 31.62 and 21.53% respectively and less preferred because of the focus and attention needed from general comprehension for the speech.

Table 6. QoS performance with the computed aggregate score.

Service	Intelligibility	Naturalness	Response time	Aggregate score
fromtexttospeech	1	1	1	47.48%
Natural Reader	0.57	0.5	0.85	31.62%
Yakitome	0	0.4	0.64	21.53%

5 Conclusion

This research work covers text to speech media conversion users on the internet speech quality of services measurement. This work proposed client side QoS analysis framework for TTS services on web environment. This framework is designed to give a comparative estimation for specific quality attributes such as intelligibility, reading comprehension, and naturalness, as well as general quality attributes for performance requirements including accessibility and response time of the TTS services. It gives the capability of measuring the quality speech of TTS and estimate user perception of the services to provide feedback to the end users.

The analysis for speech quality could be extended with other adjustment variables such as QoS estimation for speech to text (STT), video quality analysis and other media conversion to be integrated with the user experience to provide better online media services.

Acknowledgments. The authors would like to acknowledge the Malaysian Ministry of Higher Education for the Fundamental Research Grant Scheme vot 1238. This research also supported by GATES IT Solution Sdn. Bhd under its publication scheme.

References

1. Patil, M., Kawitkar, R.S.: "Syllable" concatenation for text to speech synthesis for Devnagari script. Int. J. Adv. Res. Eng. Comput. Sci. Softw. **2**(9), 180–184 (2012)
2. Md Fudzee, M.F., Abawajy, J.: A protocol for discovering content adaptation services. In: Xiang, Y., Cuzzocrea, A., Hobbs, M., Zhou, W. (eds.) ICA3PP 2011. LNCS, vol. 7017, pp. 235–244. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24669-2
3. Wang, L., et al.: Evaluating text-to-speech intelligibility using template constrained generalized posterior probability. U.S. Patent Application (2012)
4. Remes, U., Reima, K., Mikko, K.: Objective evaluation measures for speaker adaptive HMM-TTS systems. In: Proceedings of 8th ISCA Speech Synthesis Workshop (2013)
5. Möller, S., Wai, Y.C., Cote, N., Falk, T., Raake, A., Waltermann, A.: Speech quality estimation: models and trends. IEEE Sign. Process. Mag. **28**, 18–28 (2011)
6. Egger, S., et al.: Waiting times in quality of experience for web based services. In: 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX). IEEE (2012)
7. Streijl, C.R., Winkler, S., Hands, D.S.: Mean Opinion Score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Syst. **22**, 213–227 (2014)
8. Md Fudzee, M.F., Abawajy, J.: Request-driven cross-media content adaptation technique. In: Ragab, K., Helmy, T., Hassanien, A.E. (eds.) Developing Advanced Web Services Through P2P Computing and Autonomous Agents: Trends and Innovations, chap. 6, pp. 91–113. IGI Global (2010)
9. Eyben, F., et al.: Unsupervised clustering of emotion and voice styles for expressive TTS. In: International Conference on IEEE Acoustics, Speech and Signal Processing (ICASSP) (2012)
10. Md Fudzee, M.F., Abawajy, J.: Management of Service level agreement for service-oriented content adaptation platform. In: Network and Traffic Engineering in Emerging Distributed Computing Applications, pp. 21–42 (2012)