Modeling the QoE of Rate Changes in SKYPE/SILK VoIP Calls

Chien-nan Chen*, Cing-yu Chu[†], Su-ling Yeh[‡], Hao-hua Chu*, Polly Huang*[†]

* Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

[†] Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[‡] Department of Psychology, National Taiwan University, Taipei, Taiwan

*[†][‡] {r99944043, r99921035, suling, haochu, pollyhuang}@ntu.edu.tw

ABSTRACT

The effective end-to-end transport of delay-sensitive voice data has long been a problem in multimedia networking. One of the major issues is determining the sending rate of real-time VoIP streams such that the user experience is maximized per unit network resource consumed. A particularly interesting complication that remains to be addressed is that the available bandwidth is often dynamic. Thus, it is unclear whether a marginal increase warrants better user experience. If a user naively tunes the sending rate to the optimum at any given opportunity, the user experience could fluctuate.

To investigate the effects of magnitude and frequency of rate changes on user experience, we recruited 127 human participants to systematically score emulated Skype calls with different combinations of rate changes, including varying magnitude and frequency of rate changes. Results show that 1) the rate change frequency affects the user experience on a logarithmic scale, echoing Weber-Fechner's Law [1], 2) the effect of rate change magnitude depends on how users perceive the quality difference, and 3) this study derives a closed-form model of user perception for rate changes for Skype calls.

Categories and Subject Descriptors

H.1.2 [Models and Principles]: User/Machine Systems – *human factors*. H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – *evaluation/methodology*.

General Terms

Performance, Design, Experimentation, Human Factors.

Keywords

VoIP, Rate Adaptation, User Perception, QoE, Psychophysics.

1. INTRODUCTION

The effective end-to-end transport of delay-sensitive voice data has long been a subject of study in multimedia networking. In recent years, researchers have proposed a number of methods approaching this issue from a user-centric view (i.e., adapting the sending rate of voice calls based on user satisfaction [2][3]). Rate adaptation mechanisms ramp up the sending rate quickly when the

available bandwidth is sufficient, and carefully tune up or down the sending rate when the network becomes congested. Although users prefer calls with a higher bit rate, sending voice data with an unnecessarily high bit rate could waste network resources or result in congestion. This in turn could compromise the quality of the user's experience.

The user-centric approach is a promising approach [4]. However, most studies on this topic focus on identifying a sending rate to optimize the user experience. An issue that has largely been overlooked is how users perceive rate changes. The bandwidth available during an end-to-end connection is often dynamic. Thus, a (naive) increase in the sending rate might not always result in better user experience because 1) the change might not be detectable by the user, and 2) the change might be disturbing if the available bandwidth fluctuates. In addition, increasing the sending rate costs the system more network resources.

What we are advocating is that designing the rate adaptation mechanism is a significantly more subtle task. It is not just to determine the optimal sending rate, but to determine, for what it is worth, the optimal magnitude and frequency of changes in the sending rate.

How humans perceive the quality of voice, images, or motion pictures has long been a subject of study in psychophysics. Weber and Fechner proposed in 1834 [1] that the human "notice-ability" of a change is relative to the current experience (i.e., the degree of noticeability depends on the log of the stimulus's intensity). To place the subject of our interest, VoIP calls, in context, this study investigates 1) whether user experience is logarithmic to the sending rate, and 2) whether user experience is logarithmic to the time interval and/or the rate change magnitude. In short, does Weber-Fechner's Law apply to streaming VoIP calls?

The objective of the study is twofold. First, we address the above-raised questions. Second, we address a fundamental problem in user-centric rate adaptation mechanisms: modeling the relationship of user perception and the magnitude/frequency of rate changes in VoIP calls. This relationship is codec- and environment-sensitive. This study uses Skype, the most popular VoIP service, as the first research target. The experiments in this study are geared to emulate the specifics of Skype conversations.

The methodology is based on ITU-T P.830 [5]. First, we recorded various human speech samples. These raw audio tracks were encoded at different rates using the SILK codec [6], an open source toolkit made available by Skype. A wide range of test tracks, with varying degrees and frequency of rate changes, were synthesized. In total, 127 human participants were invited to score the synthesized tracks. Three data sets, each containing different speech contents and different sets of human participants, were

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'12, October 29–November 2, 2012, Nara, Japan.

Copyright 2010 ACM 978-1-4503-1089-5/12/10...\$15.00.

compiled independently. These data sets were first examined by ANOVA [7] tests and then used to model and verify the relationship of user perception and rate changes.

This study 1) confirms that the user experience-sending rate exhibits log-like behavior, echoing Weber's theory; 2) shows that the experience-frequency of rate change relationship also exhibits log-like behavior; 3) shows that the experience-magnitude of rate change relationship is determined by how users perceive the quality difference; and 4) derives a closed form model of user experience to rate changes has an average error ratio and R-square of 9.8% and 0.85, respectively. These findings provide a foundation for voice quality assessment and voice data delivery. For example, Skype's rate adaptation mechanism can be redesigned to optimize the user experience under dynamic network conditions.

The rest of the paper is organized as follows: Section 2 presents related work on multimedia networking and psychophysics; Section 3 presents findings from the preliminary experiments, forming the basis for modeling the relationship of user experience and rate changes in Section 4; using large-scale experiments, Section 5 derives the specifics for the models; Section 6 presents an evaluation of models; and finally, this paper concludes with a description of future work.

2. BACKGROUND AND FUNDAMENTALS

2.1 Quality of Service

Whether a network application is providing a "good enough" service is traditionally measured by throughput, loss rate, delay, and delay jitter [8]. These measurements are referred to as Quality of Service (QoS) measurements. For years, measured QoS represented the performance of a network application. The quality of the best-effort data, transported through TCP [9], is typically measured primarily by throughput, whereas the quality of real-time streaming media, transported through UDP [10], is typically measured by additional metrics such as loss rate, delay, and delay jitter.

Early rate control mechanisms attempted to improve QoS for the data delivered [11]. For example, TCP employs an additive-increase-multiplicative-decrease (AIMD) policy to control the send-window size. This allows the exploration of maximum available bandwidth and avoids data loss during network congestion. AVoIP [12], designed to transport voice data, employs a similar AIMD policy to control the sending rates of VoIP calls.

2.2 Quality of Experience

An increasing trend in this area is to measure network services using the Quality of Experience (QoE). As defined by the International Telecommunication Union (ITU), QoE is "the overall acceptability of an application or service, as perceived subjectively by the end-user," and "includes the complete end-to-end system effect" and "may be influenced by user expectations and context" [13]. The metrics of QoE, such as responsiveness [14] and the mean opinion score (MOS) [15], directly reflect the user's perception of the network services.

However, QoE measurements are difficult to acquire without application-level support [14][15]. Responsiveness might require data content analysis, and MOS requires user feedback. Rate adaptation mechanisms based on these measurements are inherently difficult to implement. Furthermore, the delay of acquiring these measurements might exceed the granularity of network dynamics, rendering the approach impractical.

2.3 QoS-QoE Synergy

Despite their differences, QoS and QoE are not in competition to each other. Rather, they are complementary pieces in the long-standing puzzle of assessing the quality of multimedia services in real time. Recent research dedicated to map the objective, network-centric QoS to the direct, user-centric QoE has enabled the practical implementation of service quality assessment in real time. The authors of [14] proposed a formula mapping bit rate, loss, delay, and jitter to user satisfaction in Skype calls. [16] and [17] investigated the QoS and QoE of a wide range of network services, including Web page browsing, photo sharing, file downloading, and VoIP. These QoS-to-QoE mappings exhibit logarithmic relationships.

Previous QoS and QoE mapping techniques have provided insights on how users perceive multimedia streams in steady states. However, they are insufficient to derive adaptation strategies that deliver real-time multimedia streams, which are under the influence of frequent network dynamics. Thus, this study derives the relationship between rate changes and human perception. This is an essential issue that has not yet been thoroughly investigated.

2.4 QoE-Based Adaptation

Similar to the ultimate goal of this study, previous researchers have proposed a number of adaptation schemes using QoE metrics as the criteria. For a 3D tele-immersive video service, the authors of [18] identified two metrics, "Just Noticeable Degradation" and "Just Unacceptable Degradation," and proposed a quality adaptor based on these two metrics to reduce resource usage while enhancing perceived visual quality. For a mobile video service, the authors of [19] defined a QoE metric, "Acceptability," to indicate whether users are pleased with the service under different content types and codec settings. These QoE metrics are binary in that their assessment results are acceptable or unacceptable. To facilitate fine-grained quality assessment, and therefore, to enable more sophisticated adaptation, this study investigates the effect of rate changes on user experience and derives a user experience model in the continuous space.

2.5 Psychophysics

Weber-Fechner's Law [1] provides a plausible explanation of the logarithmic relationships between various QoS and QoE metrics. Studies on the relationships between stimulus and human perception date back to 1834, when Ernst Heinrich Weber published his insights into the human sensory system. Quantitatively, Weber showed that the ratio of the noticeable threshold of stimulus intensity change to the intensity of original stimulus is a constant:

$$\frac{\Delta I}{I} = K$$

Figure 1. The quality fluctuation.

where ΔI is the amount of intensity difference being just noticeable, I is the original intensity, and the constant K is called the Weber fraction. Take weight lifting as an example. Assume that one starts with a 25 kg object, and the carrier does not notice the difference until it increases by 5 kg. Here, K for weight lifting is $1/5$, and the just-noticeable increment is 1 kg if the carrier starts from 5 kg.

Shortly after the publication of Weber's Law, Gustav Theodor Fechner presented a mathematic model known as Weber-Fechner's Law. Given a stimulus S and its responding quantitative perception P , the relationship between S and P is as follows:

$$dP = k \times \frac{dS}{S}$$

where dS and dP are the differences of stimulus and perception, respectively, and k is a constant scale factor. Integrating both sides of the equation produces the following:

$$P = k \times \ln S + c$$

where c is the constant introduced by integration. Weber-Fechner's Law is applicable to a wide range of human perceptions [20][21][22][23], including hearing, vision, taste, sense of touch and heat, and even in temporal, spatial, and numerical cognitions.

3. PRELIMINARY EXPERIMENT

The purpose of the preliminary experiment in this study was to examine, through user tests, whether changes in sending rates decreases the user experience, and at what scale the mean opinion score (MOS) is influenced by the magnitude and frequency of changes. For clarity, we define magnitude and frequency, the two key aspects of a rate change, as follows. As shown in Fig. 1, magnitude is determined by a pair of bitrates, namely the high rate (hr) and the low rate (lr), whereas frequency is defined as the time interval (ΔT) between two adjacent rate changes.

3.1 Methodology

Audio Source: Following the recommendations of ITU-T P.830 [5], the source material consisted of a number of simple, short, meaningful sentences with no obvious contextual connections. Two female and two male speakers were recruited to produce the audio samples to avoid bias caused by contextual and speaker characteristics. Each audio sample lasted 30 s, and the sampling rate of each recording was a standard 44.1 kHz.

Figure 2. MOS-bitrate plot of fixed rate tests.

Fixed-Rate Tracks: This study used Skype, a state-of-the-art commercial VoIP application, as the experimental environment. The latest version of Skype [24] uses the SILK [6] audio codec for its PC-to-PC service. Thus, the audio source was encoded by SILK for all the experiments. Based on the limitations of the SILK freeware's functionality, the audio sources were encoded in 10 different bitrates, uniformly from 40.6 to 5.6 kbps, as shown in Fig. 2.

Variable-Rate Tracks: We synthesized test tracks with varying qualities by combining two audio tracks with different bitrates into one with different time intervals. The high and low bitrates were chosen from 40.6, 28.9, 17.2, or 5.6 kbps. This pairing produced 6 pairs of high-low rate changes, indicated as (hr, lr). The time interval between rate changes was chosen from 1, 2, 3, 5, and 10 s. Thirty test tracks were generated to form the variable-rate test group for the experiment. For most video codecs capable of generating fixed-rate and variable-rate video streams, a video stream of variable bitrates is very likely different from composition of fixed-rate streams. However, SILK works in such a manner that, given a bandwidth limit, it generates an audio stream at the given limit. Skype sends a variable bitrate stream by feeding SILK a sequence of available bandwidths measured from the network. The manner in which the variable-rate tracks are synthesized in the study utilizes this Skype characteristic.

Number of Participants: Each test track was rated by 14 non-expert participants [25] using a 5-point MOS, where 5 representing the best quality, and 1 representing the worst [15]. Among all the participants, there were 10 males and 4 females whose ages ranged from 23 to 31 years, with a mean age of 25.6 years. All of them are graduate students. The original audio source of the 44.1 kbps bitrate was presented to participants at the beginning of the experiment for two reasons: to provide a reference of the most desirable case, and to allow the participants to focus on experiencing the quality of the audio for subsequent samples, instead of focusing on the content of the conversation. Other than the reference track, the remaining 40 test tracks were randomly ordered to avoid time-dependent bias. Each participant rated 41 audio tracks, taking slightly longer than 20 min to complete.

3.2 Fixed-Rate Results

Increasing the sending rate does not produce a proportional improvement in user experience. Fig. 2 shows the MOS of the fixed-rate tracks. The x-axis indicates the sending rate, and the

Figure 3. MOS- ΔT plot of (40.6, 17.2) kbps set.

Figure 4. MOS- ΔT plot of (28.9, 17.2) kbps and (40.6, 5.6) kbps tests.

y-axis indicates the corresponding MOS. The opinion score increases as the sending rate increases. In particular, the degree of MOS increase is not proportional to that of the rate increase (i.e., the MOS-bitrate relationship is sub-linear). Applying a regression test to the data set shows a logarithmic fit to the convexity with a substantially high R-square value: 0.9607. This is a strong indication that the MOS (P) and sending rate (S) relationship exhibit Weber-Fechner's Law.

In the preliminary experiments, the sending rates tested were evenly distributed, approximately 11 kbps apart. This is a shortcoming in support of Weber-Fechner's Law in the MOS-bitrate relationship. The region between 7 and 10 kbps is where the regression suggests a rapid increase in the MOS. However, data points are insufficient to confirm that the relationship is logarithmic. The large-scale experiments described in Section 5 consider this shortcoming and address this issue.

3.3 Variable-Rate Results

Fluctuation in sending rates significantly affects the user experience. Thus, keeping the rate low and steady might be significantly better than maximizing it. Fig. 3 plots the MOS of variable-rate tracks, where $(hr, lr) = (40.6, 17.2)$ kbps. The MOSs of the fixed-rate 17.2, 28.9, and 40.6 kbps tracks are 4.0, 4.7, and 4.8, respectively, and are indicated in the figure to highlight the following findings: 1) User experience at the fixed low rate (17.2 kbps) can be better than that of a dynamic one that runs no lower than the low rate throughout the track (i.e., those in the (40.6, 17.2) kbps set). This phenomenon is particularly distinct when the

Figure 5. MOS- $(hr - lr)$ plot of $hr = 40.6$ and $hr = 28.9$ kbps tests.

Figure 6. MOS- ΔT plot of (40.6, 28.9) kbps, (40.6, 17.2) kbps, (40.6, 5.6) kbps tests.

change frequency is high. 2) The MOS of the fixed average-rate track, the 28.9 kbps test, exceeds that of the variable-rate track (40.6, 17.2) kbps. This suggests that the average scores of the two steady tracks cannot quantitatively represent the user experience of a dynamic audio track. Thus, modeling the user experience of the fixed-rate audio streams is insufficient to derive values as measured under variable rates.

3.4 Effects of Rate Change Magnitude

Variable-rate tracks with the same average rate do not produce the same level of user experience. For variable-rate tracks with the same average sending rate, participants prefer the one with a smaller rate change magnitude. Fig. 4 shows the MOSs of two variable-rate tracks, where $(hr, lr) = (28.9, 17.2)$ kbps and $(40.6, 5.6)$ kbps. The two tracks share the same 20.6 kbps average rate. However, they have relatively different MOSs. Regardless of ΔT , tracks that are smaller in change magnitude, (28.9, 17.2) kbps, are consistently better than those that are large in change of magnitude, (40.6, 5.6) kbps. This observation suggests that the MOS depends on the rate change magnitude.

The rate change magnitude does not directly determine the user experience, which instead depends on the specifics of hr and lr . Fig. 5 shows the MOS (y-axis) across rate change magnitudes (x-axis). The data points labeled 40.6 kbps are the variable-rate tracks with $hr = 40.6$ kbps, and those labeled 28.9 kbps are $hr = 28.9$ kbps. The MOS generally declines as the magnitude increases, echoing the results above. Particularly, the MOS of the tracks with $hr = 40.6$ kbps declines slower than that of the

$hr = 28.9$ kbps tracks. This indicates that the MOS is a function of hr and lr , and not only the change magnitude ($hr - lr$). This finding sets the stage for the modeling task in Section 4.

3.5 Effects of Rate Change Frequency

Frequent changes in sending rate frustrate users. The results of the variable-rate track tests indicate a negative correlation between rate change frequency and MOS. Fig. 6 shows the resulting MOS- ΔT plot for three of the variable-rate cases, where $(hr, lr) = (40.6, 28.9)$ kbps, $(40.6, 17.2)$ kbps, and $(40.6, 5.6)$ kbps. As the frequency of change decreases, the MOS increases. Furthermore, the logarithmic trend indicates that the MOS- ΔT relationship likely obeys Weber-Fechner's Law.

Regression tests were conducted on each of the variable-rate track results. The R-square values are all higher than 0.9, except for the $(28.9, 17.2)$ kbps case. This may be attributed to the similarity of the two bitrates. According to the post-experiment feedbacks, participants were indifferent to, or did not notice, the quality changes between two similar rates. Although the MOS- ΔT relationship generally exhibits logarithmic behavior, it is subtler and depends on the specific hr and lr of a rate change.

4. PROPOSED MODEL

Based on the findings of Section 3, this section proceeds one step further and proposes models that quantify the user experience for fixed-rate and variable-rate Skype calls.

4.1 Fixed-Rate Model

The fixed-rate model is straightforward. Based on the logarithmic relationship observed in MOS-bitrate, as described in Section 3.2, this study proposes a closed-form formula to predict the MOS from the sending rate as follows:

$$f_{FIX}(br) = \gamma \times \ln(br - \alpha) + \beta$$

where br is the bitrate, and α , β , and γ are coefficients to be determined by SILK characteristics. The bitrate shift (α) is caused by the limit of human perception. According to the formula, the $br - \alpha$ term, being inside the $\ln$ operator, can only be positive. When the quality drops below α , users are unable to notice any difference. Based on a large-scale experiment, Section 5 presents the specifics of the coefficients.

4.2 Variable-Rate Model

The variable-rate model is based on the finding outlined in Section 3.4 that the MOS of variable-rate tracks depends on hr and lr , and the finding of Section 3.5 that the MOS is logarithmic to ΔT . To capture these effects, this study proposes the following closed-form formula for rate-changing SILK streams:

$$f_{FLUC}(hr, lr, \Delta T) = SCALE(hr, lr) \times \ln(\Delta T) + SHIFT(hr, lr)$$

where the effect of change frequency is distributed to the logarithm term, and the effect of hr and lr is distributed to the two subroutines, $SCALE()$ and $SHIFT()$.

- $\ln(\Delta T)$ represents the MOS- ΔT relationship, in which the MOS increases logarithmically to ΔT .

- $SCALE()$ represents the influence of hr and lr of the rate change on $\ln(\Delta T)$. This helps rescale the logarithmic effect introduced by ΔT and control the degree of quality change when ΔT varies. Hence, we can interpret $SCALE()$ as the sensitivity to the rate change frequency. A (hr, lr) pair is more sensitive to the rate change frequency if it has a larger value of $SCALE()$. The finding in Section 3.4 suggests that a larger difference between hr and lr

indicates higher sensitivity. Thus, $SCALE()$ is a function that increases with the rate difference. Another caution to take with $SCALE()$ is that, according to Section 3.4, the sensitivity to rate change frequency is not solely determined by the rate change magnitude. Instead, it is lower when the corresponding hr and lr are high. Fig. 5 shows that the decline of MOS to $hr - lr$ is slower in the $hr = 40.6$ kbps case, but faster in the $hr = 28.9$ kbps case. Thus, the $SCALE()$ term captures this dimension of interaction between the rate change frequency and the level of sending rates as well.

- $SHIFT()$ is the remaining portion of MOS that is not influenced by ΔT . This value can be derived by projecting the value of MOS when ΔT approaches the duration of the audio track. As ΔT grows, the effect of fluctuation decreases and the variable-rate case becomes indistinguishable from a fixed-rate version. The quality of this fixed rate equivalent is called the *dominant quality of the fluctuation*. As mentioned in Section 3.3, this dominant quality is not the average quality of the high and low rates. The dominant quality is the quality a user expects when the effect of fluctuation decreases. For the modeling task, the resulting MOS serves as an anchor point of the formula. Although the term does not depend on ΔT , its value varies by case in the preliminary results. Therefore, the $SHIFT()$ subroutine is formulated with an association to hr and lr .

5. LARGE-SCALE EXPERIMENT

This section reports a large-scale experiment conducted for two purposes: to re-examine the MOS-bitrate and MOS- ΔT relationships using ANOVA tests to verify the proposed models, and to derive the unknown coefficients and the exact forms of $SCALE()$ and $SHIFT()$ in the proposed formulas.

5.1 Methodology

Audio Source: The audio source is the same as that used in the preliminary experiments.

Fixed-Rate Tracks: Nine rates were selected in this set of experiments. Unlike the preliminary experiments, the chosen rates were separated evenly by their expected MOS, and not by the bitrates. The expected MOSs were estimated by the logarithmic fit obtained from the preliminary experiments to address the shortcoming mentioned in Section 3.2. The sending rates were $(r1, r2, r3, r4, r5, r6, r7, r8, r9) = (40.6, 27.7, 19.4, 14.1, 10.7, 8.5, 7.1, 6.1, 5.6)$ kbps.

Variable-Rate Tracks: Each variable-rate track contains a high rate and a low rate selected from the nine bitrates used in the fixed-rate group. This pairing produced 36 (hr, lr) combinations. The frequencies used here are the same as those in the preliminary experiments: 1, 2, 3, 5, and 10 s.

Number of Participants: The participants in this study rated 189 thirty-second tracks, of which 180 were variable-rate tracks and 9 were fixed-rate tracks. Based on the recommendation of ITU-T P.911 [25], which states that each track should be rated by 6 to 40 participants and each experiment should not exceed 30 min, we recruited 127 human participants, and each of them rated 45 randomly chosen tracks (5760 scores acquired). Among all the participants, there were 96 males and 31 females. Their age ranged from 18 to 29 years, with an average of 22.8 years. All of them had received undergraduate level education, and 61 of them had also received graduate level education. Besides, none of them had participated in the preliminary experiment. The experiment duration, including the time spent listening to the reference track

Figure 7. Non-parallel MOS- ΔT relationships indicate interaction.

Figure 8. MOS- ΔT plots of indiscernible tests.

presented at the beginning and inserted in the experiment for score calibration, was approximately 25 min. Each track was rated by at least 30 participants.

Score Calibration: Forty-five tracks are quite a few to rate. This study uses a modified Absolute Category Rating with the Hidden Reference (ACR-HR) approach to calibrate scores that might be biased because of fatigue [26]. The main idea of ACR-HR is to play a reference track before each test track, so that the participants can adjust the MOS of each track based on the corresponding reference. ACR-HR provides a good approach to compare the scores of different tracks, but also prolongs the test duration. As a compromise to the original ACR-HR, which essentially doubles the experimentation time, we modified ACR-HR to insert one high-quality (44.1 kbps) reference track for every nine test tracks.

For data processing, a differential quality score (DMOS) [26] was computed between each track and its corresponding reference using the following formula:

$$DMOS = MOS_{TEST} - MOS_{REF} + 5$$

Because DMOS might exceed 5, we applied a 2-point crushing function [26] to prevent DMOS from unduly influencing the overall MOS:

$$DMOS_{crushed} = \frac{7 \times DMOS}{2 + DMOS}, \text{ when } DMOS > 5$$

Figure 9. MOS-bitrate plot of fixed-rate tests in large-scale experiments.

5.2 ANOVA Tests

MOS-bitrate: The one-way ANOVA test was performed to determine the significance of the bitrate's influence to MOS. The p value for the MOS-bitrate test is $1.38e-58$, indicating a significant influence. The R-square value of a logarithmic fit to MOS-bitrate relationship is 0.9645, confirming Weber-Fechner's Law observed in the preliminary experiment.

Interaction between ΔT and (hr, lr) : To confirm that an interaction exists between these two factors, a two-way ANOVA was conducted on ΔT and the difference of hr and lr . The p value of interaction is $7.94e-14$, which strongly supports the significance of an interactive term. This confirms the multiplication of $SCALE(hr, lr)$ and $\ln(\Delta T)$ in the proposed variable-rate model. Fig. 7 depicts the MOS- ΔT relationships of a number of variable-rate tracks. The regression fits are not parallel to each other, suggesting an interaction between ΔT and (hr, lr) .

MOS- ΔT : For each given pair of hr and lr , one-way ANOVA was conducted with respect to ΔT of the rate changes. A p value less than .05 indicates that ΔT has a significant influence on MOS. Otherwise, it suggests that participants are unable to notice the change of bitrates with varying ΔT .

Table 1. Indiscernible variable-rate data sets

Test	p value	Test	p value	Test	p value
r1r2	.31	r6r7	.31	r7r8	.26
r3r4	.42	r6r8	.11	r7r9	.34
r4r5	.31	r6r9	.09	r8r9	.32

Table 1 shows the variable-rate data sets with p values exceeding .05. The hr and lr pairs are all similar to each other in these data sets. This echoes the odd case identified in Section 3.5, in which ΔT would be irrelevant if the rate change is indiscernible. Fig. 8 shows the MOS- ΔT relationships for some of the variable-rate tracks in Table I. The MOS does not vary significantly across ΔT in these data sets. This effect is captured by the $SCALE()$ subroutine, which tends to give a small value when the difference between the two bitrates is small.

Figure 10. 3D plot of SCALE().

Figure 11. Contour of SCALE().

5.3 Model Specifics

1) *Coefficients of Fixed-rate Formula*: As Fig. 9 shows, the data are fitted by the proposed fixed-rate formula with $\alpha = 4.091$, $\beta = 1.515$, and $\gamma = 1.000$. Thus, the relationship between QoS(bitrate) and QoE(MOS) for a fixed-rate Skype VoIP service can be described using the following formula:

$$f_{FIX}(br) = \gamma \times \ln(br - 4.091) + 1.515$$

where br is the bitrate of the service. The lower bound of user perception can be further inferred by:

$$1(MOS) = f_{FIX}(br') = \ln(br' - 4.091) + 1.515$$

$$\rightarrow br' = 4.091 + e^{-0.515} \approx 5.0(kbps)$$

where br' is the bitrate that provides the baseline user satisfaction (MOS=1). In other words, any bitrate lower than br' does not affect the QoE. The resulting $br' = 5.0$ kbps is close to the lower bound of SILK's encoding capability (5.6 kbps), suggesting that SILK is likely designed with insights to the limits of human perception, e.g., the one indicated by α in Section 4.1.

2) *SCALE() Subroutine of Variable-rate Formula*: The SCALE() term is named as such because it is the scaling coefficient of the $\ln(\Delta T)$ fit and it represents the sensitivity of users to the rate change frequencies. The preliminary experiments in this study show that the MOS and the magnitude of rate changes are positively correlated, meaning that SCALE() in the formula is an increasing function of $hr - lr$. However, as discussed in Section 4.2, SCALE() returns a small value when the levels of sending rates are high. To identify the exact relationship of SCALE() to (hr, lr) , we logarithmically fit the MOS- ΔT relationship to each (hr, lr) pair. The coefficient of the $\ln(\Delta T)$ component serves as the value of SCALE() for the rate pair.

Given that the SCALE() term is not simply decided by $hr - lr$, but also the level of hr and lr , we take the analysis to another space, where x is $hr - lr$ and y is $hr + lr$. The two variables proposed here are designed to capture the complexity of perceptual sensitivity to rate change frequency. The term $hr - lr$ allows a large SCALE() when the rate difference is large, whereas the term $hr + lr$ allows a small SCALE() when both hr and lr are large. After applying polynomial regression to the new variables, the resulting polynomial form of SCALE() can be expressed as:

$$SCALE(hr, lr) = f(x, y) = p_{00} + p_{10}x + p_{01}y + p_{20}x^2 + p_{11}xy + p_{02}y^2 + \dots$$

where $x = hr - lr$ and $y = hr + lr$.

The wider variety of fluctuations tested in the large-scale experiment enables polynomial regression to a higher degree. After exploring different degrees, we find that the cubic polynomial function fits sufficiently well (i.e., the fit is not significantly better taken to a degree higher than 3). Table 2 presents the corresponding coefficients of the polynomial.

Table 2. Coefficient of Polynomial Fit to SCALE()

p_{00}	p_{10}	p_{01}	p_{20}
0.02122	0.06465	-0.001637	-0.004956
p_{11}	p_{02}	p_{30}	p_{21}
0.001538	3.488e-005	-9.562e-007	8.639e-005
p_{12}	p_{03}		
-3.903e005	-2.21e-007		

To gain a complete picture of how hr and lr influence SCALE(), this study plots the derived polynomial and depicts the resulting 3D surface in Fig. 10. To facilitate the discussion, Fig. 11 shows the contour of SCALE().

As shown in these figures, the relationship between SCALE() and (hr, lr) is as follows: 1) With the same $hr + lr$, SCALE() generally increases as $hr - lr$ increases. This confirms the findings in the previous sections that the rate change becomes more disturbing when users are more aware of the quality difference. 2) Along each contour line, SCALE() tends to be smaller when $hr + lr$ becomes large. A larger $hr + lr$ means that both hr and lr are high. Based on the fixed-rate model, the quality differences between two high rates are relatively smaller. Consequently, users are less aware of the difference. Thus, the case leads to a smaller SCALE(), or a lower sensitivity to ΔT . 3) After $hr - lr$ exceeds a certain level, such as 15 kbps, SCALE() declines when $hr + lr$ decreases. This means that if the quality difference is large while both hr and lr are low, the sensitivity to ΔT decreases with the sending rates. This indicates that, despite users being able to perceive a rate change, reducing the frequency of rate change does not improve the quality of experience because the effect of rate change frequency is bounded by users' perceptual limit to low-rate audio tracks.

Figure 12. Dominant quality.

Figure 13. Normalized dominant quality.

3) *SHIFT()* Subroutine of Variable-rate Formula: This term is derived by projecting MOS when ΔT approaches the length of the audio track (i.e., 30 s). The MOS at $\Delta T = 30$ is called the dominant quality of the variable-rate track, and denoted as $D()$. For each (hr, lr) pair, this study identifies the regression fit of the MOS- ΔT relationship and then projects the MOS value for $\Delta T = 30$. These values are plotted in Fig. 12 and indicated as the Dominant Quality data set. This figure also plots the MOS of the hr and lr based on the $f_{FIX}(br)$ formula. The MOS of hr and lr are denoted as MOS_h and MOS_l , respectively.

The findings are 1) the dominant quality is generally bounded by MOS_h and MOS_l . It can be slightly better than MOS_h when the hr is high and the rate change magnitude is relatively small (e.g., r1r3 and r2r4). 2) For cases before r3r9, the dominant quality decreases as the rate change magnitude increases. This suggests that when the variation increases, the changes become disturbing. 3) The behavior of the dominant quality diverges as both hr and lr drop below 14.1 kbps (i.e., the cases from r4r5 and on). According to the analysis mentioned earlier, users cannot easily perceive the changes when both rates of the fluctuation are low. For these cases, the curves of the dominant quality and MOS_h are intertwined.

The results presented here show that $D()$ equals MOS_h for cases where $hr \leq 14.1$ kbps. For the remaining cases, this study examines if a closer relationship exists between the dominant quality and the rate change magnitude. Fig. 13 shows the MOS of dominant quality normalized by MOS_h and MOS_l , where they are 100% and 0%, respectively. Fig. 14 shows the MOS of dominant

Figure 14. Normalized dominant quality- MOS_{diff} plot.

quality with the difference between MOS_h and MOS_l , denoted as MOS_{diff} , as its x-axis. A significant linear relationship exists between the normalized dominant quality and MOS_{diff} . A linear regression fit to the relationship in Fig. 14 produces $D()$ for cases in which $hr > 14.1$ kbps. Thus, the *SHIFT()* subroutine can be formulated as follows:

$$SHIFT(hr, lr) = D(hr, lr) - SCALE(hr, lr) \times \ln(30)$$

$$D(hr, lr) = \begin{cases} MOS_h, & \text{if } hr \leq 14.1, \\ \max(0.55, 1.5332 - 0.371 \times MOS_{diff}) \times MOS_{diff} + MOS_l, & \text{else} \end{cases}$$

The term “ $\max(0.55, \bullet)$ ” in the formula indicates that the relationship between MOS_{diff} and $D()$ for $hr > 14.1$ kbps is more complex than simply linear. This term is based on the bounded decrease of the plot in Fig. 14. This figure shows that the degradation of the normalized dominant MOS is limited. A bound near 50% is apparent, indicating that the lower bound of dominant quality is approximately the average MOS value of hr and lr .

6. EVALUATION AND DISCUSSION

The evaluation of the proposed model is twofold. First, we evaluate the goodness of fit [7] of the mathematic form of the model with respect to the training data by examining its R-square value. A high R-square value confirms the robustness of model derivation, including the establishment of subroutines *SCALE()* and *SHIFT()*. Second, the model is tested using two sets of data that are independent of the model construction. The purpose of this phase is to evaluate the prediction accuracy of the model by showing its average error ratio for the two data sets. In summary, the purpose of presenting the R-square and average error ratio is to examine whether the proposed model is capable of capturing the average human perception. To this end, both metrics are computed based on the average score of each track from the three data sets.

The resulting R-square of the proposed model on the training data is 0.8504, which indicates that the data do not deviate significantly from the statistical fit. Thus, the proposed *SCALE()* and *SHIFT()* functions and the logarithmic approximation do capture the characteristics of the data. The scatter plot in Fig. 15 shows the fitted score (as the x-axis) and the user score (as the y-axis) for each rate pair and time interval test set. The proposed model closely follows the measured data as the points fall densely around the 45° line, indicating the goodness of fit and supporting the robustness of model derivation.

Figure 15. Goodness of fit of training data.

This study uses two additional data sets to evaluate prediction accuracy. The first data set is the result of the preliminary experiment in Section 3, where data is collected independently of those used for model construction. The other data set is taken from an additional set of a subjective experiment, in which the audio content, speakers, and participants are all different from the prior experiments. In this additional experiment, 11 male and 3 female participants were recruited. Their ages ranged from 21 to 28 with an average of 24.8 years. 4 of them were undergraduate students and the rest are graduate students. For simplicity, the results of the preliminary and this new experiment are labeled as data sets I and II, respectively.

The resulting average error ratios of the two data sets are 9.8% and 9.7%, respectively, indicating that the proposed model is able to capture the average user experience. In addition to the numerical result, Fig. 16 shows the accuracy of the proposed model. In this figure, the predictions of both data sets closely track the measured data, regardless of the difference in audio content and speaker. The average error ratios of the two data sets are similar, confirming that ITU guidelines for subjective study can operatively minimize the contextual and speaker bias.

Three major issues to address next include 1) in-depth exploration of the terms shown after numerical fit, 2) whether Weber-Fechner's Law exists in other VoIP services, and 3) a model that produces distribution as its output instead of a single average value. First, the relationship between the rate change magnitude and $SCALE()$ in the preliminary experiments is unclear. The term $D()$ has a relatively complex structure. Furthermore, there could be a certain connection between the fixed-rate and variable-rate models, which might lead to one concise model. Second, although the good average error ratio and R-square over three sets of experiments support the proposed models, it is unclear whether the relationships of MOS to magnitude/frequency of rate changes will remain the same for a different codec or experimental setting (e.g., mobile or lossy VoIP calls). Third, the current model has the ability to handle average user perception and provide suggestions for designing rate adaptation mechanisms. However, the variance among each user has not been fully explored. Adding probabilistic components might produce a model that can better describe user variance and provide more realistic predictions. These are subjects for future research.

7. CONCLUSION

The findings of this study provide a foundation for rate adaptation mechanism design in real-time voice data delivery. This study 1)

Figure 16. Accuracy of prediction.

confirms that user experience versus the bitrate relationship exhibits a log-like behavior, echoing Weber's theory; 2) shows that the experience versus frequency of the rate change relationship also exhibits a log-like behavior; 3) shows that the experience versus magnitude of the rate change relationship is determined by how users perceive the quality difference; and 4) derives a closed form model of user experience to rate changes with an average error ratio and R-square of 9.8% and 0.85, respectively.

Although this study was motivated by a networking problem (i.e., how to better provide multimedia services over the Internet), the work itself is content-centric in that it focuses on how users perceive the quality of the multimedia content delivered. In addition to its technical findings, this study indicates the need for interdisciplinary research that considers both content and network. A solid understanding of multimedia content is fundamental to better engineering of the network applications. For example, knowing that, given the same amount of increase in bitrate, user experience improves marginally when the call quality is already good, a rate adaptation mechanism should allocate network bandwidth to calls of lower quality. This approach can maximize the overall QoE. This study also shows that frequent changes in the sending rate decrease the perceived quality. Thus, a rate adaptation mechanism should not increase the sending rate at any given opportunity, but instead, dynamically adapt adjustment sensitivity based on network stability.

8. REFERENCES

- [1] K. R. Boff, L. Kaufman, and J. P. Thomas, "Handbook of Perception and Human Performance", Wiley-Interscience, ISBN 0-47-182957-9, 1986.
- [2] J. Matta, C. Pépin, K. Lashkari, and R. Jain, "A Source and Channel Rate Adaptation Algorithm for AMR in VoIP Using the E-model," in Proceedings of ACM Network and Operating System Support for Digital Audio and Video, NOSSDAV, 2003.
- [3] Z. Qiao, L. Sun, N. Heilemann, and E. Ifeachor, "A New Method for VoIP Quality of Service Control Use Combined Adaptive Sender Rate and Priority Marking," in Proceedings of IEEE International Communications Conference, ICC, 2004.
- [4] T.-Y. Huang, P. Huang, K.-T. Chen, and P.-J. Wang, "Can Skype be More Satisfying? A QoE-Centric Study of the FEC Mechanism in the Internet-Scale VoIP System," IEEE Network, Vol. 24(2), pp.42–48, 2010.

- [5] ITU-T Recommendation P.830, "Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs" 1996.
- [6] Skype Developer, <http://developer.skype.com/silk>.
- [7] R. A. Fisher, "Statistical Methods for Research Workers", Oliver and Boyd, ISBN 0-05-002170-2, 1925.
- [8] R. Beuran, M. Ivanovici, and B. Dobinson, "Network Quality of Service Measurement System for Application Requirements Evaluation", in Proceedings of IEEE International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS, 2003.
- [9] RFC 793, "Transmission Control Protocol," 1981.
- [10] RFC 768, "User Datagram Protocol," 1980.
- [11] J. Bolot and T. Turletti, "A Rate Control Mechanism for Packet Video in the Internet", in Proceedings of IEEE International Conference on Computer Communications, INFOCOM, 1994.
- [12] A. Barberis, C. Casetti, J.C. De Martin, and M. Meo, "A Simulation Study of Adaptive Voice Communications on IP Networks." Computer Communications, Vol. 24(9), pp.757–767, 2001.
- [13] ITU-T Recommendation P.10/G.100, "Vocabulary for Performance and Quality of Service", 2008.
- [14] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei, "Quantifying Skype User Satisfaction", in Proceedings of ACM Communications and Computer Networks, SIGCOMM, 2006.
- [15] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality," 1996.
- [16] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A Generic Quantitative Relationship between Quality of Experience and Quality of Service" IEEE Network, vol. 24(2), pp.36-41, 2010.
- [17] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment," in Proceedings of IEEE International Communications Conference, ICC, 2010.
- [18] W. Wu, A. Arefin, G. Kurillo, P. Agarwal, K. Nahrstedt, and R. Bajcsy, "Color-plus-Depth Level-of-Detail in 3D Tele-immersive Video: A Psychophysical Approach", in Proceedings of ACM Multimedia, MM, 2011.
- [19] W. Song, D. W. Tjondronegoro, and M. Docherty, "Saving Bitrate vs. Pleasing Users: Where is the Break-Even Point in Mobile Video Quality", in Proceedings of ACM Multimedia, MM, 2011.
- [20] J. Shen, "On the Foundations of Vision Modeling: I. Weber's law and Weberized TV restoration," Physica D: Nonlinear Phenomena, vol. 175(3-4), pp.241–251, 2003.
- [21] E. D. Scheirer, "Tempo and Beat Analysis of Acoustic Musical Signals," Journal of the Acoustical Society of America, vol. 103(1), pp.588–601, 1998.
- [22] R. S. Moyer and T. K. Landauer, "Time Required for Judgments of Numerical Inequality," Nature, vol. 215(5109), pp.1519–20, 1967.
- [23] M. R. Longo and S. F. Lourenco, "Spatial Attention and the Mental Number Line: Evidence for Characteristic Biases and Compression," Neuropsychologia, vol.45, pp.1400-1406, 2007.
- [24] Skype for Windows - v5.10.1.115, released on July 5, 2012, <http://www.skype.com/get-skype/>.
- [25] ITU-T Recommendation P.911, "Subjective Audiovisual Quality Assessment Methods for Multimedia Applications," 1998.
- [26] ITU-T Recommendation P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," 2008.