Modeling the QoE of Rate Changes in Skype/SILK VoIP Calls
Shannon Chen, Member, IEEE, Cing-Yu Chu, Su-Ling Yeh, Hao-Hua Chu, and Polly Huang, Member, IEEE, ACM

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; 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.

Index Terms—Performance evaluation, psychophysics, quality of experience, rate adaptation, voice over IP.

I. 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 [1], [2]). 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 [3]. 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. Although people do notice rate changes would influence user’s experience [34], there still lacks a systematic study to quantitatively and qualitatively examine this issue. 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 [4] that the human “notice-ability” of a change is relative to the current experience (i.e., the degree of noticeable depends on the log of the stimulus’s intensity). To place the subject of our interest, voice-over-IP (VoIP) calls, in context, this paper investigates through a user study: 1) whether user experience is logarithmic to the sending rate [17]; 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 study uses Skype, the most popular VoIP service, as the first 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 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.
For Skype/SILK calls, 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 II presents related work on multimedia networking and psychophysics. Section III presents findings from the preliminary experiments, forming the basis for modeling the relationship of user experience and rate changes in Section IV. Using large-scale experiments, Section V derives the specifics for the models. Section VI presents an evaluation of models. Finally, this paper concludes with a description of future work.

II. BACKGROUND AND FUNDAMENTALS

A. 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.

B. 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.

C. 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. References [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-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.

D. 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 3-D 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.

E. Qoe Modeling of Speech Quality

Models assessing speech quality have long been studied [41], [42]. These models are typically classified into three categories: 1) full-reference; 2) reference-free; and 3) parametric models.

Both the original speech source and the degraded speech file are required for the full-reference models. The score of degraded speech is computed based on the comparison of the degraded signal to the original. PESQ (ITU P.862) [43] and POLQA (ITU P.863) [44] are two models that fall in this category. PESQ was first designed for narrowband speech signals and later extended to wideband signals. POLQA (ITU P.863) was finalized more recently to model the super-wideband speech signals that are used more in modern VoIP services.

The need for both the original and degraded speech signals makes full-reference models impractical assessing the quality of VoIP services in real time. Reference-free models do not require the original signal that makes online analysis of the speech quality possible. Unfortunately, existing reference-free models...
such as ITU P.563 [45] were designed to assess narrowband speech quality, not for super-wideband signals.

A widely used parametric model is E-model (ITU G.107) [46], which distributes the influence of factors such as source rate, delay, and packet loss to multiple components, each with a specific parameter. The final speech quality is computed by summing all the components. The problem of E-model is that the parameter sets are codec and transmission-mode-dependent. Until now, the number of E-models derived is limited and insufficient for super-wideband VoIP services today.

This study aims at providing parametric models for VoIP quality estimation, which is close in spirit to the existing parametric models, but differs in that the effect of rate changes, a phenomenon very common on the Internet, is systematically examined. As opposed to the existing full-reference and reference-free models, the models proposed here allow QoE assessment in real time and target modern VoIP services, which better address the motivating problem—rate adaptation for popular VoIP services such as Skype.

F. Subjective Assessment

Although not directly related to this study, an important issue that normally arises in QoE studies is subjective assessment. Among all the alternatives, the five-level MOS score recommended by ITU might be the most widely accepted. However, some researchers have argued that this ITU recommended approach is not adequate for subjective quality assessment of Internet multimedia services. Authors of [20] regarded MOS as an ordinal scale instead of an interval scale, and therefore, its value does not necessarily represent the exact difference between users’ sensory magnitude. As a result, they suggest that values of MOS should not be used for calculation directly. To solve this problem, the authors have proposed the psychometric model based on the law of categorical judgment to convert MOS values into interval scale. It is then adopted in their subsequent works [21], [22]. Since this study aims at deriving a model to predict the QoE represented in MOS instead of having mathematical operation on MOS values, it is not necessary to convert MOS into interval scale.

In addition, Bouch et al. [23] have proposed a three-dimensional model including user satisfaction, task performance, and user cost to assess the quality of service perceived by end-users. They also argued that MOS is not suitable for measuring user satisfaction since it is more likely that the responses will be concentrated at the lower scale [24]. Moreover, due to the cultural and linguistic difference, the scale of MOS is not always internationally interval or internationally ordinal. Therefore, they have designed a slider-based approach named QUASS [25] to avoid the above shortcomings, and it is able to capture the dynamics of user satisfaction. However, how to interpret a series of continuous scores introduced by the slider and convert the result into a unique QoE metric for model fitting remains to be solved. Moreover, the slider-based tool is not publicly available for use. As a result, we chose to adopt the five-level MOS suggested by ITU as the QoE metric for assessing quality perceived by user. Hobfeld [35] suggests that MOS should be used in conjunction with another QoE measure called “SOS” to compensate the problem that the average MOS cannot assess user rating diversity. Since the focus of this study is to capture the average behavior and establish models accordingly, we use the average MOS alone for QoE measurement.

G. Psychophysics

Weber–Fechner’s Law [4] 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

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

where $\Delta I$ 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 mathematical 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 [26]–[29], including hearing, vision, taste, sense of touch and heat, and even in temporal, spatial, and numerical cognitions.

III. 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 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 bit rates, namely the high rate ($hr$) and the low rate ($lr$), whereas frequency is defined as the inverse of time interval ($1/\Delta T$) between two adjacent rate changes.

Note that the time granularity of a rate adaptation decision is often finer than a call. That is, the sending bit rate may be adjusted multiple times in the call duration. Given a change in available bandwidth, the system will need to know how the user perceives the change in order to adapt, as opposed to wait until the call ends. The insight we seek to unveil is thus how users perceive a change, as opposed to a call. To this end, we take a systematic approach to investigate the perceived quality when
the sending bit rate fluctuates between two rates with a certain frequency. These synthesized audio tests do not resemble actual voice calls. However, each test represents a specific rate change. By varying the two rates and the frequency, we exhaust a wide range of rate changes that enables us to analyze the relationship of user experience to the key parameters of a rate change.

A. 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 connotations. 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.

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 bit rates, uniformly from 40.6 to 5.6 kb/s, as shown in Fig. 2.

Variable-Rate Tracks: We synthesized test tracks with varying qualities by combining two audio tracks with different bit rates into one with different time intervals. The high and low bit rates were chosen from 40.6, 28.9, 17.2, or 5.6 kb/s. This pairing produced six 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 bit rates 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 bit rate 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 [31] using a 5-point MOS, where 5 represents the best quality, and 1 represents 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-kb/s bit rate 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. The reference track is inserted twice in a session. Each participant rated 42 audio tracks, taking slightly longer than 20 min to complete. The scores were then calibrated as detailed in (6) and (8) in Section V.

B. 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 that echoes the finding in [36]. The x-axis indicates the sending rate, and the 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-bit rate relationship is sublinear). 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 kb/s apart. This is a shortcoming in support of Weber–Fechner’s Law in the MOS–bit-rate relationship. The region between 7 and 10 kb/s is where the regression suggests a rapid increase in the MOS. However, data points are insufficient to confirm that the relationship is logarithmic in the 7–10-kb/s region. The large-scale experiments described in Section V consider this shortcoming and address this issue.

C. 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 \(\{hr, lr\} = (40.6, 17.2)\) kb/s. The MOSs of the fixed-rate 17.2-, 28.9-, and 40.6-kb/s 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 kb/s) 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) kb/s set). This phenomenon is particularly distinct when the change frequency is high. 2) The MOS of the fixed average-rate track, the 28.9-kb/s test, exceeds that of the variable-rate track (40.6, 17.2) kb/s. 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.

D. 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 $[hr, lr] = (28.9, 17.2)$ kb/s and $(40.6, 5.6)$ kb/s. The two tracks share the same 20.6-kb/s average rate. However, they have relatively different MOSs. Regardless of $\Delta T$, tracks that are smaller in change magnitude, $(28.9, 17.2)$ kb/s, are consistently better than those that are large in change of magnitude, $(40.6, 5.6)$ kb/s. 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 kb/s are the variable-rate tracks with $hr = 40.6$ kb/s, and those labeled 28.9 kb/s are $hr = 28.9$ kb/s. The MOS generally declines as the magnitude increases, echoing the results above. Particularly, the MOS of the tracks with $hr = 40.6$ kb/s declines slower than that of the $hr = 28.9$ kb/s 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 IV.

E. 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–$\Delta T$ plot for three of the variable-rate cases, where $[hr, lr] = (40.6, 28.9)$ kb/s, $(40.6, 17.2)$ kb/s, and $(40.6, 5.6)$ kb/s. As the frequency of change decreases, the MOS increases. Furthermore, the logarithmic trend indicates that the MOS–$\Delta T$ 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) kb/s case. This may be attributed to the similarity of the two bit rates. According to the post-experiment feedbacks, participants were indifferent to, or did not notice, the quality changes between two similar rates. Although the MOS–$\Delta T$ relationship generally exhibits logarithmic behavior, it is subtler and depends on the specific and of a rate change.

IV. PROPOSED MODEL

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

A. Fixed-Rate Model

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

$$f_{Y|X}(hr) = \gamma \times \ln(hr - \alpha) + \beta$$

(4)
where \( br \) is the bit rate, and \( \alpha, \beta, \gamma \) are coefficients to be determined by SILK characteristics. The bit-rate shift \( (\alpha) \) is caused by the limit of human perception. According to the formula, the \( br - \alpha \) term, being inside the operator, can only be positive. When the quality drops below, users are unable to notice any difference. Based on a large-scale experiment, Section V presents the specifics of the coefficients.

B. Variable-Rate Model

The variable-rate model is based on the finding outlined in Section III-D that the MOS of variable-rate tracks depends on \( hr \) and \( lr \), and the finding of Section III-E that the MOS is logarithmic to \( \Delta T \). To capture these effects, this study proposes the following closed-form formula for rate-changing SILK streams:

\[
f_{\text{FULL}}(hr, lr, \Delta T) = \text{SCALE}(hr, lr) \times \ln(\Delta T) + \text{SHIFT}(hr, lr)
\]

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

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

\( \text{SCALE}() \) represents the influence of \( hr \) and \( lr \) of the rate change on \( \ln(\Delta T) \). This helps rescale the logarithmic effect introduced by \( \Delta T \) and control the degree of quality change when \( \Delta T \) varies. Hence, we can interpret \( \text{SCALE}() \) 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 \( \text{SCALE}() \). The finding in Section III-D suggests that a larger difference between \( hr \) and \( lr \) indicates higher sensitivity. Thus, \( \text{SCALE}() \) is a function that increases with the rate difference. Another caution to take with \( \text{SCALE}() \) is that, according to Section III-D, 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 \( hr = 40.6 \text{ kb/s} \) case, but faster in the \( hr = 28.9 \text{ kb/s} \) case. Thus, the \( \text{SCALE}() \) term captures this dimension of interaction between the rate change frequency and the level of sending rates as well.

\( \text{SHIFT}() \) is the remaining portion of MOS that is not influenced by \( \Delta T \). This value can be derived by projecting the value of MOS when \( \Delta T \) approaches the duration of the audio track. As \( \Delta T \) 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 III-C, 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 \( \Delta T \), its value varies by case in the preliminary results. Therefore, the \( \text{SHIFT}() \) subroutine is formulated with an association to \( hr \) and \( lr \).

V. LARGE-SCALE EXPERIMENT

This section reports a large-scale experiment conducted for two purposes: to reexamine the MOS–bit-rate and MOS–\( \Delta T \) relationships using ANOVA tests to verify the proposed models, and to derive the unknown coefficients and the exact forms of \( \text{SCALE}() \) and \( \text{SHIFT}() \) in the proposed formulas.

A. 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 bit rates. The expected MOSs were estimated by the logarithmic fit obtained from the preliminary experiments to address the shortcoming mentioned in Section III-B. The sending rates were \( \{r1, r2, r3, r4, r5, r6, r7, r8, r9\} = \{40.6, 27.7, 19.4, 14.1, 10.7, 8.5, 7.1, 6.1, 5.6\} \text{ kb/s} \).

Variable-Rate Tracks: Each variable-rate track contains a high rate and a low rate selected from the nine bit rates 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 30-s 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 6–40 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 ages 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. Furthermore, none of them had participated in the preliminary experiment. The experiment duration, including the time spent listening to the reference track 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 [32]. 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 kb/s) reference track for every nine test tracks.

For data processing, a differential quality score (DMOS) [32] was computed between each track and its corresponding reference using the following formula:

\[
DMOS = \text{MOS}_{\text{TEST}} - \text{MOS}_{\text{REF}} + 5.
\]

B. ANOVA Tests

To investigate whether the influence of bit rate, rate change magnitude, and rate change frequency is significant, we conducted a series of ANOVA tests [33]. To alleviate the effect introduced by individual participant’s preference, all the scores are normalized using the z-score

\[
z\text{-score} = \frac{DMOS - \mu}{\sigma}.
\]
In that, μ and σ are the mean and standard deviation of all scores obtained from tracks of the same \((hr, lr)\) pair. The z-score was then used for ANOVA tests.

**MOS–Bit Rate:** The one-way within-subject ANOVA test was performed to determine the significance of the bit rate’s influence to MOS. The \(p\) value for the MOS–bit-rate test is \(1.68e-42\) and the \(f\) value is 73.8, indicating a significant influence. The R-square value of a logarithmic fit to MOS–bit-rate relationship is 0.9645, confirming Weber–Fechner’s Law observed in the preliminary experiment.

**Interaction Between \(\Delta T\) and \((hr, lr)\):** To confirm that an interaction exists between these two factors, a two-way mixed-design ANOVA was conducted on \(\Delta T\) and the difference of \(hr\) and \(lr\). The \(p\) value of interaction is \(8.23e-20\), and \(f\) value is 2.55, which strongly supports the significance of an interactive term. This confirms the multiplication of \(\text{SCALE}(hr, lr)\) and \(\ln(\Delta T)\) in the proposed variable-rate model. Fig. 7 depicts the MOS–\(\Delta T\) relationships of a number of variable-rate tracks. The regression fits are not parallel to each other, suggesting an interaction between \(\Delta T\) and \((hr, lr)\).

**MOS–\(\Delta T\):** For each given pair of \(hr\) and \(lr\), one-way within-subject ANOVA was conducted with respect to \(\Delta T\) of the rate changes. A \(p\) value less than 0.05 indicates that \(\Delta T\) has a significant influence on MOS. Otherwise, it suggests that participants are unable to notice the change of bit rates with varying \(\Delta T\).

Table I shows the variable-rate data sets with \(p\) values exceeding 0.05. The \(hr\) and \(lr\) pairs are all similar to each other in these data sets. This echoes the odd case identified in Section III-E, in which \(hr\) would be irrelevant if the rate change is indiscernible. Fig. 8 shows the MOS–\(\Delta T\) relationships for some of the variable-rate tracks in Table I. The MOS does not vary significantly across \(\Delta T\) in these data sets. This effect is captured by the \(\text{SCALE}(\cdot)\) subroutine, which tends to give a small value when the difference between the two bit rates is small.

### C. Model Specifics

After validating the influence of each factor with ANOVA tests, we further derive the model specifics in this section. The goal here is to predict the MOSs. The z-scores used for ANOVA tests are not quite appropriate. In the meantime, DMOS might exceed 5, which is out of the range of MOS. We applied a 2-point crushing function [32] to prevent DMOS from unduly influencing the overall MOS

\[
\text{DMOS}_{\text{crushed}} = \frac{7 \times \text{DMOS}}{2 + \text{DMOS}}, \quad \text{when DMOS > 5}. \quad (8)
\]

The \(\text{DMOS}_{\text{crushed}}\) was then used for model construction.

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(bit rate) and QoE(MOS) for a fixed-rate Skype VoIP service can be described using the following formula:

\[
f_{\text{FIX}}(br) = \ln(br - 4.091) + 1.515 \quad (9)
\]

where \(br\) is the bit rate of the service. The lower bound of user perception can be further inferred by

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

\[
\rightarrow br' = 4.091 + e^{0.515 \approx 5.0} \text{ (kb/s)} \quad (10)
\]

where \(br'\) is the bit rate that provides the baseline user satisfaction (MOS = 1). In other words, any bit rate lower than \(br'\) does not affect the QoE. The resulting \(br'\) is 5.0 kb/s is close to the lower bound of SILK’s encoding capability (5.6 kb/s), suggesting that SILK is likely designed with insights to the limits of human perception, e.g., the one indicated by \(\alpha\) in Section IV-A.

2) **\(\text{SCALE}(\cdot)\) Subroutine of Variable-Rate Formula:** The \(\text{SCALE}(\cdot)\) term is named as such because it is the scaling
TABLE II
COEFFICIENT OF POLYNOMIAL FIT TO SCALE(·)

<table>
<thead>
<tr>
<th>$P_{00}$</th>
<th>$P_{10}$</th>
<th>$P_{01}$</th>
<th>$P_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02122</td>
<td>0.06465</td>
<td>-0.001637</td>
<td>-0.004956</td>
</tr>
<tr>
<td>$P_{31}$</td>
<td>$P_{02}$</td>
<td>$P_{30}$</td>
<td>$P_{21}$</td>
</tr>
<tr>
<td>0.000538</td>
<td>3.488e-005</td>
<td>-9.562e-007</td>
<td>8.639e-005</td>
</tr>
<tr>
<td>$P_{12}$</td>
<td>$P_{03}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3.903e-005</td>
<td>-2.21e-007</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 IV-B, 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–$\Delta T$ 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

$$\text{SCALE} (hr, lr) = f(x, y) = p_{00} + p_{10}x + p_{01}y + p_{20}x^2 + p_{11}xy + p_{02}y^2 + \cdots$$  \hspace{1cm} (11)

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

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 II presents the corresponding coefficients of the polynomial.

To gain a complete picture of how $hr$ and $lr$ influence SCALE(·), this study plots the derived polynomial and depicts the resulting 3-D 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 $\Delta T$: 3) After $hr - lr$ exceeds a certain level, such as 15 kb/s, 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 $\Delta T$ 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.

3) SHIFT() Subroutine of Variable-Rate Formula: This term is derived by projecting MOS when $\Delta T$ 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 $D(\cdot)$. For each ($hr, lr$) pair, this study identifies the regression fit of the MOS–$\Delta T$ 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 $hr$ and $lr$ based on the $f_{HRX}(hr)$ formula. The MOS of $hr$ and $lr$ are denoted as MOS$_h$ and MOS$_l$, respectively.

The findings are the following: 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 kb/s (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 $D(\cdot)$ equals MOS$_h$ for cases where $hr \leq 14.1$ kb/s. 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 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 subroutine can be formulated as follows:

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

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

(12)

The term “$\max(0.55,*$)” 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 $lr$.

VI. EVALUATION AND DISCUSSION

In this section, we first examine the accuracy of the derived model and then compare the result to widely accepted Perceptual Evaluation of Speech Quality (PESQ).

A. Prediction Accuracy

The evaluation of the proposed model is twofold. First, we evaluate the goodness of fit[7] of the mathematical 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 $\text{SCALE}()$ and $\text{SHIFT}()$. 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 and $function and the logarithmic approximation do capture the characteristics of the data. The scatter plot in Fig. 15 shows the fitted score (as the $y$-axis) and the user score (as the $x$-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.

This study uses two additional data sets to evaluate prediction accuracy. The first data set is the result of the preliminary experiment in Section III, 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. Four of them were undergraduate students, and the rest were 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 $\text{SCALF}(\cdot)$ in the preliminary experiments is unclear. The term $D(\cdot)$ 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.

### B. Comparison to PESQ

The PESQ model proposed by ITU is a long-standing standard generally accepted for research and commercial uses. Despite several previous works using PESQ as the performance indicator to show that Skype provides a fair voice service [22]–[24], PESQ is not a proper QoE metric for the SILK codec and cannot be used in real-time services. Its limitations come from: 1) it requires both the original and degraded audio tracks for quality computation which is not feasible in a real-time system; 2) PESQ only supports sampling rates in 8k and 16k Hz, which only cover the narrow and wide bands of SILK but not the medium and super-wide bands; and 3) whether the performance indicator designed for old audio codecs is still applicable to a modern one is doubtful. Due to the limitations stated above, authors of [37] could only build up a model for narrow and wide bands of SILK. In order to validate if our models are more capable of capturing the user perception, a comparison to PESQ is presented.

1) Fixed Rate: For the fixed-rate comparison, we included tracks in both the preliminary and large-scale experiment. Due to the limitation of PESQ, only audio tracks that were encoded under 8k or 16k sampling rates were used for the comparison. Therefore, only nine bit rates were chosen, namely 28.9, 25.0, 14.1, 13.3, 9.5, 8.2, 7.1, 6.6, and 5.6 kb/s.

The accuracy of prediction of both models on SILK codec can be seen from Fig. 17. In Fig. 17(a), we plot the user-rated MOS on the x-axis, the model predicted MOS on the y-axis, and a diagonal auxiliary line that represents a perfect prediction. The results of our model and PESQ are both plotted. The quality of SILK calls tends to be underestimated or overestimated by PESQ, causing the plot of PESQ prediction to form a line with a milder slope. Except the mid-quality track with MOS = 3.3, our model outperforms PESQ in all cases, providing predictions with a 0.9601 R-square value. In Fig. 17(b), the cumulative distribution function (CDF) of MOS errors made by the two models are plotted. The error of our model is bounded by 0.22 while more than 50% of the predictions of PESQ deviate from the user-rated score by more than 0.49. From the statistics of accuracy of the two models, listed in Table III, we can see PESQ being a poor predictor on newly proposed codec such as SILK. With its average error ratio being 14.59%, the prediction of PESQ can be deviated from the true value by 0.41 as indicated by the root-mean-square error (RMSE).

Comparing the predictions of PESQ and our model on SILK-encoded contents, we can see PESQ being inferior on processing audio content with fairly high and fairly low qualities. The result shows that PESQ is too conservative on its prediction and
hence gives underestimated scores when the quality is high and overestimated ones when the quality is low.

2) Variable Rate: A similar comparison was also conducted for variable tracks. However, it is more difficult for variable rate tracks to satisfy the requirement of PESQ since both the $hr$ and $lr$ need to be encoded using the same sampling rate, 8k or 16k Hz. As a result, only combinations of four bit rates from the large-scale experiment—8.2, 7.1, 6.6, and 5.6 kb/s—can be used, and hence six $(hr, lr)$ pairs were produced. With five different $\Delta T'$ for each $(hr, lr)$ pair, we have totally 30 variable tracks for the comparison.

The accuracy of prediction of both the proposed model and PESQ are indicated in Fig. 18 and Table IV. The findings here are as follows: 1) Table IV shows that the goodness of fit of both models is inferior to the fixed-rate case, which can also be observed in Fig. 18(a). However, the proposed model still slightly outperforms PESQ. 2) As shown in Fig. 18(a), it can be observed that the range of the predicted scores given by PESQ is small, thus resulting in a mild slope. This implies that PESQ might not take the frequency of rate change into consideration. The proposed model, on the other hand, is capable of capturing this characteristic and gives a wider range of predicted scores.

3) From the CDF of MOS error shown in Fig. 18(b), we can see that the 50th percentile of the proposed model is 0.18, while it is 0.29 for PESQ. This is also supported by the RMSE and average error ratio in Table IV. 4) Although both the RMSE and average error ratio of the proposed model are better than PESQ, it can be noticed that both metrics of the proposed model are slightly worse than the fixed-rate case. For PESQ, these two metrics are slightly better than the fixed-rate case.

The findings above show that the proposed model still outperforms PESQ in the variable-rate case, but not as much as in the fixed-rate case. It should be noticed that the variable-rate tracks chosen for comparison are biased to rates with low quality due to the limitation of PESQ. As a result, a complete comparison across the whole coding range of SILK cannot be conducted.

VII. CONCLUSION AND FUTURE WORK

The findings of this study provide a foundation for rate adaptation mechanism design in real-time voice data delivery. For popular Skype/SILK calls, this study: 1) confirms that user experience versus the bit-rate 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.

Admittedly, 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. Just as how analyzing measured loss rates and end-to-end delays is important to facilitate network designs optimizing for QoS, measuring and analyzing actual user experience is critical to close the loop for QoE-centric designs. As the community grows more interested in delivering multimedia content based on QoE, there lack studies that measure actual user experience. This work is one of first few that embark a sizeable user study. Despite the effort, this work by no means addresses all problems. It is only a beginning and calls for works in three directions: 1) QoE measurement; 2) network design; and 3) multimedia content.

A. QoE Measurement

In this study, we quantify the relationship between user experience and bit rate changes for Skype/SILK calls. As widely known, network impairments such as packet loss, delay, and jitter might also play significant roles. More user studies and modeling effort will be necessary to develop a comprehensive model such as that proposed in [14]. To this extent, investigating how loss rate, as well as the loss distribution, affect the user experience could result in an experiment that requires an explosive number of test subjects. The problem is similar for delay and delay jitter. As crowdsourcing platforms such as Amazon’s Mechanism Turk [38] are drawing attention from the content research community [39], we see hope as well that a large-scale user study on how loss and delay affect user experience can be realized.

B. Network Design

A solid understanding of multimedia content is fundamental to better engineering of the network applications. The findings of this work affect the designs of bandwidth allocation and rate adaptation as follows. 1) As suggested by the fixed rate model, given the same amount of increase in bit rate, user experience improves marginally when the call quality is already good. A bandwidth allocation mechanism may prioritize and favor allocating network bandwidth to calls of lower quality. This approach would maximize the overall QoE. 2) We showed also how rate change magnitude and frequency influence the perceived quality and rapid changes could worsen the quality of experience. Therefore, one might consider keeping the bit rate at a lower level instead of raising the bit rate at any possible time when the network bandwidth fluctuates. As mentioned in Section III, these findings can provide a basis for rate adaptation, and the derived variable-rate model can serve as a utility function.

![Fig. 18. Prediction error: SILK versus PESQ for variable-rate tracks.](image-url)
Based on the history of network condition, one may possibly predict the available bandwidth in the near future. If the available bandwidth increases at a given time while the prediction indicates this increase is only temporary, the derived model would tell the system if it is worthwhile raising the bit rate with the risk of introducing quality fluctuation to users.

C. Multimedia Content

Taking Skype as the first research target, this study investigates the influence of different bit rates and rate changes of audio streams produced by the SILK codec. However, SILK is not the only audio codec adopted by today’s VoIP applications. A question raised would be if the findings in this study are also applicable to other audio codecs. The audio codecs can be classified into two categories: fixed-rated and variable-rate codecs. Since fixed-rate audio codecs provide only fixed coding quality regardless of the available bandwidth, we focus on the discussion of variable-rate codecs.

Traditional variable-rate audio codecs, e.g., AMR-WB, define a set of coding bit rates that applications can choose based on different network or channel conditions. With the growth of network capacity, it is a trend that modern audio codecs such as SILK support a wider range of coding bit rate and the output bit rate can be set in fine granularity. For these audio codecs, it is important to know if the properties identified in this study also hold for different audio codecs.

Although it is needed to apply similar methodology to examine different audio codecs quantitatively, we can still provide a qualitative discussion over this issue. First, recent works [17], [40] have demonstrated that the logarithmic relationship between QoS and QoE is commonly observed in many Internet applications, which supports the applicability of Weber–Fechner Law in the network field. Moreover, these works also conducted subjective evaluation on audio codecs that are different from SILK, and the result echoes the findings in this study that the influence of coding bit rate is logarithmic. Second, the design philosophy of most audio codecs is to first extract perceptually important components to form a basis. On top of this basis, increasing the coding bit rate normally implies adding more details that are less perceivable to users. As a result, the relationship between bit rate and perceived quality tends to be sublinear. Therefore, we believe the main findings in this study are qualitatively applicable to most variable-rate codecs.

Beside VoIP services, video streaming is another relevant application that is of rising user demand. It is intuitive to ask if the same methodology can also be applied on video codecs to investigate the relationship between the coding bit rate of video codecs and user perception, thus enabling the design of rate adaptation schemes for video streaming applications. However, the design of video codecs is significantly more complex, and the video quality under the same coding bit rate can be quite different depending on specific parameters such as frame rate, video resolution, quantization step, and so on. All these factors would have significant impact on user perception. Although it might be more time-consuming, isolating each factor and identifying the corresponding impact would be the most essential task.

References

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Chen et al.: Modeling QoE of Rate Changes in Skype/Silk VoIP Calls

Shannon Chen (M’12) received the B.S. degree in computer science and the M.S. degree in networking and multimedia from National Taiwan University, Taipei, Taiwan, in 2009 and 2012, respectively, and is currently pursuing the Ph.D. degree in computer science at the University of Illinois at Urbana-Champaign, Urbana, IL, USA.

His research interests include networking, multimedia, and psychology.

Cing-Yu Chu received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2009 and 2012, respectively, and is currently pursuing the Ph.D. degree in electrical engineering at the Polytechnic Institute of New York University, Brooklyn, NY, USA.

His research interests include multimedia networking and network resilience.

Su-Ling Yeh received the Ph.D. degree in psychology from the University of California, Berkeley, CA, USA.

Since 1994, she has been with the Department of Psychology, National Taiwan University, Taipei, Taiwan. Her current research interests include multisensory integration and effects of attention on perceptual processes.

Prof. Yeh is an Associate Editor of the Chinese Journal of Psychology and serves in the Editorial Board of Frontiers in Perception Science. She is a recipient of the Distinguished Research Award of the National Science Council of Taiwan and is a Distinguished Professor of National Taiwan University.

Hao-Hua (Hao) Chu received the B.S. degree from Cornell University, Ithaca, NY, USA, in 1993, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 1999, both in computer science.

He is a Professor with the Department of Computer Science and Information Engineering and Graduate Institute of Networking and Multimedia, National Taiwan University (NTU), Taipei, Taiwan.

Prior to joining NTU, he worked with NTT DoCoMo USA Labs, San Jose, CA, USA, and Intel Corporation, Folsom, OR. His research areas are in ubiquitous computing, sensor/wireless networks, and persuasive technologies.

Polly Huang (M’99) received the B.S. degree in mathematics from National Taiwan University, Taipei, Taiwan, in 1993, and the Ph.D. degree in computer science from the University of Southern California, Los Angeles, CA, USA in 1999.

She is a Professor with the Department of Electrical Engineering, the Graduate Institute of Communication Engineering (joint appointment), and the Graduate Institute of Networking and Multimedia (courtesy), National Taiwan University. Prior to joining National Taiwan University, she spent time in the early stage of her career with AT&T Labs—Research, Florham Park, NJ, USA; the Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland; and the University of California, Los Angeles (UCLA), CA, USA.

She is also a Visiting Scientist with MIT CSAIL, Cambridge, MA, USA, from 2013 to 2014. Her research interests span wireless sensor and multimedia networking.

Dr. Huang is a member of the Association for Computing Machinery (ACM). She serves currently as an Associate Editor of the ACM Transactions on Sensor Networks and an Editor of the IEEE ComSoc Journal of Communications and Networks. She has served as a TPC member for major conferences such as ACM SenSys, ACM SIGCOMM, and ACM Multimedia. Her team is the recipient of the IEEE/ACM IPSN 2012 Best Paper Award, ACM SIGCOMM W-MUST 2012 Best Paper Award, ACM ASPLOS 2012 Best Poster Award, ACM SenSys 2010 Best Presentation Award, and IS 2000 Best Paper Award.