Measuring the Perceptual Quality of Skype Sources

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, suling, haochu}@ntu.edu.tw, [†] {heidernchu, Huang.Polly}@gmail.com

ABSTRACT

The success of Skype has inspired a generation of peer-to-peer-based solutions for real-time multimedia services over the Internet. However, there lacks still a robust metric quantifying the perceptual quality of a Skype call. The widely-used PESQ (Perceptual Evaluation of Speech Quality) falls short of modeling super-wideband calls, which are characteristics of SILK – Skype’s codec made public in 2011. Towards a robust QoE (Quality of Experience) metric for VoIP call analysis, we propose a model, referred to as WF-Regression model, to capture the call rate and perceptual quality relationship. The model is shown through a user study that it is robust, R-square = 0.9990 and outperform PESQ modeling the quality of Skype calls, error ratio = 3.68% vs. 14.59%.

Categories and Subject Descriptors

C.4 [Performance of Systems]: Modeling techniques.

General Terms

Performance, Experimentation, Human Factors

Keywords

VoIP, QoE, Skype, PESQ

1. INTRODUCTION

Released by its creators of KaZaa in 2002, Skype has already achieved 2 billion downloads and 37.5 million concurrent users count in less than a decade.¹ With its recent partnering with Facebook, which brings the 750 million social network population closer to its service, Skype is estimated to be adding more than 400,000 users per day and owning one quarter of all international calling minutes happening anytime in the world.¹² The phenomenal growth of Skype has not only inspired a generation of application-level solutions for real-time multimedia services over the

Internet, but also stunned the market observers worldwide with its US\$ 8.5 billion deal with Microsoft.³

Skype draws attentions from both engineering and commercial regime due to one reason: Skype, relative to other Internet phone services, satisfies users. This echoes the rising trend to measure network services by perceptual satisfaction of users, i.e., 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” which “includes the complete end-to-end system effects” and “those that may be influenced by user expectations and context”[1].

QoE measurements are, however, difficult to acquire without application-level support and time-consuming subjective tests with real users. To facilitate assessment of voice quality without involving users, substitutive models were proposed to predict the perceived quality. Among them, the PESQ (Perceptual Evaluation of Speech Quality) model [2] proposed by ITU is the long-standing standard generally accepted for research and commercial uses.

PESQ, as a QoE metric for Skype calls, is insufficient. Despite several previous works showing that Skype provides a fair voice service [3-6], using PESQ as the performance indicator, the calls examined are limited within a small bitrate range – the range PESQ aims at predicting. SILK [7], the codec developed, submitted, and released into the public domain by Skype in 2011, is the key to the service quality of the latest versions of Skype. Compared to the PESQ model, which is designed based on psychological studies dating back to 1994 [2], SILK is designed to provide higher bandwidth and meet the demand of users in current Internet culture. More specifically, instead of accommodating only narrowband and wideband coding such in PESQ, SILK accepts two additional bandwidth modes by varying its sampling rate, which produces variable, fine-grained coding bitrates from 6 to 40 kbps. SILK’s wider encoding spectrum is designed to satisfy a keener genera-

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.

W-MUST’12, August 17, 2012, Helsinki, Finland.

Copyright 2012 ACM 978-1-4503-1476-3/12/08... \$15.00.

¹ <http://skypejournal.com/blog/category/statistics/>

² <http://www.skype.com/intl/en-us/features/allfeatures/facebook/>

³ <http://www.economist.com/node/18680040>

tion of users, while PESQ models codecs of a narrower spectrum. Evaluations of Skype calls biasing a specific bitrate range will never tell us the whole story of the prominent Internet application.

PESQ model bears another problem. It requires both the original and the degraded audio content for the computation. PESQ measures the distortion between the original and degraded speech signals and map the distortion values to mean opinion scores (MOS) [2][8]. This disqualifies PESQ as a viable feedback to the adaptation module of a network service, which tunes up or down the bitrate depending on the network capacity. To enable rate adaptation online, an ideal model shall provide predictions without referencing the original signal.

Motivated by the aforementioned deficiencies, we conducted a study on Skype and construct a new model to facilitate rate adaptation for Skype calls. Inspired by studies in Psychophysics, we propose the WF-Regression model based on the Weber-Fechner’s Law [9]. Constructed and examined by data collected from multiple subjective experiments with real human participants, we find a strong logarithmic relationship between the sending rate and the perceptual satisfaction of users, and the accuracy of our prediction model reaches a 0.9990 R-square value with a maximum 4.94% error ratio for SILK-encoded audio content.

Having argued the need to extend QoE metric design for a better understanding of Skype, we would like to stress, however, that our purpose is not to replace PESQ with the WF-Regression model. The two are in fact analyzing different types of contents. PESQ predicts the user-rated outcomes via analyzing the difference between the two signals on the sender and the receiver sides which accounts network impairments. On the other hand, our model analyzes the relationship between encoding bitrate of modern SILK codec and the perceptual quality in result. The purpose is to aid the decision of internal rate adaptation module of Skype in a timely manner. In other words, the proposed model focuses on perceptual quality evaluation of Skype sources on the sender side.

Finally, in order to back up our statement that PESQ being incapable of predicting performance of contemporary codecs, we compare its prediction on SILK and on AMR-WB [10], the wideband version of a coeval codec standard with PESQ which is broadly used by GSM- and 3G-capable devices. The result supports our claim and shows that the predictions of PESQ are too conservative on newer codecs that produce content in multiple bitrates forming a wider quality spectrum.

Our contribution is three-fold. 1) We confirm the insufficiency of the broadly accepted PESQ model, and proof by actual user studies that it is outmoded with contemporary codec and sensory of users; 2) we verify the user experience versus bitrate relationship exhibits a log-like behavior,

echoing the Weber-Fechner’s theory; and 3) we devise an objective and perceptual user satisfaction model in which the parameters are all easily measurable and computable online.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction of the SILK codec and describes the subjective experiment that served as the foundation of our model construction. In Section 3, the data collected from the experiment is examined first by the ANOVA test before we derive our WF-Regression model base on the insight of Weber-Fechner’s Law. The model is then evaluated in Section 4, in which the accuracies of predictions on SILK and on AMR-WB encoded contents are shown. Finally, we conclude the work in Section 5.

2. EXPERIMENT

In this section, we detail the methodology and findings modeling the relationship of user experience to voice data rate of SILK, the most up-to-date codec currently used by Skype. The objective of this user study is to provide as a foundation for the succeeding MOS-bitrate model construction of the codec.

2.1 The SILK Codec

SILK is the speech codec of Skype in its latest versions and has been made publically available as an IETF draft [7] since 2010. It is designed for real-time, packet-based voice communication. The goal of SILK is to achieve high coding efficiency as well as being able to adapt its operating point to the characteristics and limitations of network, hardware, and audio signal. In order to fulfill this goal, SILK is a variable bitrate codec that allows even finer-grained output bitrate. This is enabled by adjusting various input parameters to the encoder. These parameters are:

Sampling Rate: To support not only PC-to-PC communications but also communications for PSTN and mobile phones, SILK is designed to accept bandwidth from narrowband to super wideband by varying the sampling rate. The possible sampling rates are listed in Table 1.

Table 1. SILK-sampling rate in each bandwidth mode

Narrowband	8 kHz
Mediumband	8 or 12 kHz
Wideband	8, 12 or 16 kHz
Super Wideband	8, 12, 16 or 24 kHz

Packet Rate: SILK encodes frames of 20 milliseconds and can combine one to five frames into one payload, hence creating inter-packet gaps from 20, 40, 60, 80 to 100 milliseconds. Lower packet rate reduces the output bitrate for that it conserves more bits sending fewer packet headers. It,

however, increases the latency and the resulting call is more sensitive to packet losses.

Packet Loss and FEC: SILK exploits inter-frame dependencies and is able to adjust on the fly between the trade-off of bitrate conservation vs. error propagation. In addition, SILK also adopts in-band Forward Error Correction (FEC) to provide robustness against packet losses. It re-encodes packets with perceptually important speech information using a lower bitrate and attaches it to the subsequent packet.

Complexity: To accommodate hardware platforms of different computation capacity, SILK provides several options that can be enabled or disabled during the encoding process. This is achieved by changing the order of each filter at the cost of increasing the bitrate by a few percent.

Bitrate: The bitrate of SILK can be set between 6 and 40 kbps in an arbitrary, fine-grained manner. When the bitrate is given to the SILK SDK, unless the above 4 parameters are specified, the default values and rules are used and the audio stream of the exact bitrate is generated.

2.2 Methodology

Audio Source: Following recommendations of ITU-T P.830 [11], our source material consists of a number of simple, short, meaningful sentences which do not have obvious contextual connections. Two female and two male speakers are recruited to produce the voice of the audio source. Length of the source material is 30 seconds and the sampling rate of recording is the standard 44.1 kHz.

Test Tracks: 9 different rates are selected. One particular caution we take is that, hinted by various previous works on QoS-QoE relationship studies [12-13], the chosen rates are separated exponentially to ensure an even sampling of resulting perceptual qualities. The density of rate choosing is higher at low quality region and lower at high quality region to ensure we get enough samples at the low rates where perceptual quality changes more rapidly. The sending rates are 40.6, 27.7, 19.4, 14.1, 10.7, 8.2, 7.1, 6.6, and 5.6 kbps.

Participants: Approximately 30 participants are recruited for the study. They are chosen with diversity in age and gender. Each test track is rated by more than 12 non-expert participants using 5-point mean opinion score (MOS), 5 for the most desirable quality and 1 for the least. The original audio source is presented to the participants at the beginning of the experiment for reference. Other than the reference track, the rest of the test tracks are randomly ordered to avoid time-dependent bias. In each term, a participant rates no more than 11 audio tracks, which takes less than 10 minutes to complete.

Score Calibration: To calibrate scores that might be biased due to fatigue, we take a modified Absolute Category Rating with Hidden Reference (ACR-HR) approach [14].

In that, the differential quality scores (DMOS) are computed between each track and its corresponding reference track.

2.3 Results

Figure 1 depicts the MOS-bitrate relationship for tracks encoded by SILK. The x-axis indicates the sending rate and the y-axis the corresponding MOS. The error bar indicates the range of the resulting MOS extended one standard deviation above and below the average.

3. MODELING

Our purpose in this section is to construct model that maps between bitrate and MOS for SILK codec. From Figure 1, we find a logarithmic relationship in between the user experience and sending rate which is supported by Psycho-physical studies that models human perception under different quantified physical stimuli.

3.1 Data Analysis

As a precaution before we derive the models for MOS-bitrate relationships, one-way ANOVA [15] test is conducted to examine the significance of bitrate's influence to MOS. For SILK, the resulting p-value is 1.38e-58, indicating the influence is significant. This significance is due to the fact that SILK has an independent error correction module and hence every additional bit in SILK tracks contributes to the enhancement of quality.

3.2 Weber-Fechner's Law

The Weber-Fechner's Law [9] in psychophysics provides a plausible explanation to the logarithmic relationships observed between various QoS and QoE metrics. Studies of relationships between stimulus and human perception date back to 1834, when Ernst Weber published his insight upon the human sensory system. Quantitatively speaking, Weber found that the ratio of the noticeable threshold of stimulus intensity change to the intensity of original stimulus is a constant, i.e.,

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

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 from carrying a 25 kg object and the carrier does not notice the difference until we increase the weight by 5 kg. The just-noticeable increment will then be 1 kg, if the carrier starts from 5 kg. Here, the K for weight lifting is 1/5.

Shortly after the publishing of Weber's Laws, Gustav Fechner stated a mathematic model as an extension known as the Weber-Fechner's Law. Given a stimulus S and its responding quantitated perception P , the relationship between S and P is as follows.

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

Figure 1. MOS-bitrate plot of SILK's WF model.

Figure 2. MOS-bitrate plot of SILK's user-rated scores.

where dP and dS are the differences of stimulus and perception, respectively, and k is a constant scale factor. By integrating both sides, one obtains the following.

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

where c is the constant introduced by integration.

To put the subject of our interest, VoIP calls, in context, the strong showing that MOS (perception) and bitrate (stimulation) are logarithmically related exhibits the very essence of Weber-Fechner's Law.

3.3 WF-Regression Model

Hinted by the Weber-Fechner's Law, we propose our WF-Regression Model based on the logarithmic relationship observed. The closed form formula to predict MOS from the sending rate is as follows.

$$MOS = \gamma \times \ln(br - \alpha) + \beta$$

where br is the bitrate, and α , β , γ are codec dependent coefficients. The coefficients derived from logarithmic fits for SILK are $(\alpha, \beta, \gamma) = (4.02, 1.52, 1.00)$ with a 0.9645 R-square value.

The shifted bitrate (α) is due to the limit of human perception. When the quality drops below a certain threshold, users are unable to notice the difference. The other coefficients are more related to the target use of individual codecs. For codecs that target voice services under a wide range of available bandwidth (e.g., over the Internet), the extra bits are used to elaborate the details of audio content when the bandwidth is abundant. This results in a wider quality range and hence incurs a scaling factor (γ) no less than 1. On the other hand, codecs that aim to provide loss-resilient voice services given limited bandwidth (e.g., over a 3G wireless network), the extra bits are often redundant data to facilitate error correction. Thus the audio quality of the decoded content bears little differences under different sending rates and hence incurs a small scaling factor that restricts the range of output.

4. EVALUATION AND DISCUSSION

The evaluation of our model is two-folds. First, in order to evaluate the accuracy of our model with enough generality, we conduct an additional subjective experiment to collect a set of independent data. In the settings of this additional

experiment, the chosen coding rates as well as the human participants are different from the previous experiment which contributes to the construction of our model. Second, our model is compared with the well-accepted PESQ model. In order to cross-compare our model to PESQ, we conduct the same subjective experiment for AMR-WB, a contemporary codec standard in the public domain which is widely used in most high-end digital phones as well as GSM- and 3G-capable devices.

4.1 Accuracy of Prediction

4.1.1 Evaluation Experiment

In order to evaluate the performance and to verify the generality of our model, an additional subjective experiment is conducted. The settings of this experiment are the same as we listed in Section 2.2 except that the chosen bitrates and the human participants are different from the previous one. 10 bitrates are chosen uniformly within the coding capability of SILK in this experiment, namely: 40.6, 36.6, 32.8, 28.9, 25.0, 21.1, 17.2, 13.3, 9.5, and 5.6 kbps.

4.1.2 Result

The result of this evaluation dataset is plotted in Figure 2. The dots in the figure represent the average MOS of participants for each audio test case. As we can see from the figure, the prediction of our model follows closely the user-rated scores. The robustness of our model is also supported by multiple statistical metrics in Table 2. The goodness of fit (R-square) is 0.9990 and the correlation (Pearson's r) is 0.9932. These values indicate the bitrate-MOS relationship of SILK codec does comply with the Weber-Fechner's Law of logarithmic relationship between physical and perceptual quantities. As for the numerical deviation of our model, both root-mean-square error (RMSE) and error ratio indicate it being negligible.

4.2 Comparison with PESQ

In this section we compare the performance of WF-Regression model and the PESQ model. The comparison contains two stages. First, we use the two models to predict the MOS outcomes of SILK encoded audio tracks and compare their accuracy. Second, we apply the same modeling methodology as Section 2.2 on AMR-WB codec to get

Table 2. Accuracy of WF-Regression Model.

R-square	0.9990
Pearson's r	0.9932
RMSE	0.03
Max. Err. Ratio	4.94%

Figure 3. Prediction error: SILK. (o: WF-Reg.; x: PESQ)

Figure 4. Prediction error: AMR-WB (o: WF-Reg.; x: PESQ)

an AMR-WB version of the WF-Regression model. We then compare the predictions of this model with PESQ.

4.2.1 SILK

Due to the limitation of PESQ, only audio tracks which were encoded under 8k or 16k sampling rates can be fed into the model. Among the 17 test cases (9 in model construction + 10 in evaluation experiment – 2 replications: 40.6 and 5.6 kbps), we have bitrates span across SILK’s narrowband, mediumband, wideband, and super wideband coding modes. PESQ can only accommodates 9 (28.9, 25.0, 14.1, 13.3, 9.5, 8.2, 7.1, 6.6, and 5.6 kbps) of our test cases.

The accuracy of prediction of both models on SILK codec can be seen from Figure 3. As shown in Figure 3a, where 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 WF-Regression model and PESQ are both plotted. The quality of SILK calls tends to be underestimated 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 Figure 3b, the CDF of MOS errors made by the two models are plotted. The error of our model is bounded by 0.22 while more than

Table 3.
Accuracy of WF-Reg. vs. PESQ model on SILK.

Codec	SILK	
Model	WF-Reg.	PESQ
R-square	0.9601	0.7841
Pearson’s r	0.9940	0.9980
RMSE	0.16	0.41
Max. Err. Ratio	8.69%	39.12%
Avg. Err. Ratio	3.68%	14.59%

Table 4.
Accuracy of WF-Reg. vs. PESQ model on AMR-WB.

Codec	AMR-WB	
Model	WF-Reg.	PESQ
R-square	0.7878	0.6289
Pearson’s r	0.8886	0.8925
RMSE	0.10	0.13
Max. Err. Ratio	5.60%	8.20%
Avg. Err. Ratio	2.18%	2.86%

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 3, 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.

Comparing the predictions of PESQ and WF-Regression model on SILK-encoded contents, we can see that 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. Considering the design era of PESQ, our result also calls for subjective studies of wide spectrum codecs which should give rise to an extended PESQ.

4.2.2 AMR-WB

AMR-WB [10] is the contemporary audio codec with PESQ adopted by both ITU-T and 3GPP as a standard for GSM, 3G wireless communications, as well as the wire-line phone service to eliminate the need for transcoding. It extends the sampling frequency range from 200-3400 Hz to 50-7000 Hz, which provides better intelligibility, natural-

ness and feeling of transparent communication. The flexibility enables the codec to generate audio streams of different bitrates ranging from 6.6 to 23.85 kbps under its 9 coding modes: 6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, and 23.85 kbps, which allows the phone service adjusting in finer granularity, relative to G.7xx series, to fluctuating network conditions in packet networks.

For AMR-WB codec, all 9 coding modes fit the sampling rate requirement of PESQ model, so all 9 bitrates are used in the succeeding analysis. We treat AMR-WB as the subjecting codec and conduct the model construction experiment the same as in Section 2.2. The same logarithmic relationship between bitrate and MOS indicated by the Weber-Fechner's Law is revealed and the resulting parameters of the AMR-WB version of our WF-Regression model are $(\alpha, \beta, \gamma) = (6.56, 3.45, 0.10)$.

Similar to the comparison procedures of SILK results, Figure 4 provides visual indication of the accuracy of prediction. From Figure 4a, we can see that for contemporary codec as AMR, the prediction of PESQ and our model have rather similar distributions. From Figure 4b, the fact that the plotting of the two models entangles with each other indicates the similar prediction accuracy of the two. Likewise, accuracy of both models for AMR-WB encoded tracks is detailed in Table 4. Although our model performs generally better than PESQ, the improvement is not significant. The PESQ model for AMR-WB calls still provides a reliable prediction with its error ratio being bounded under 8.2%.

The fact that the two models give similar predictions on AMR-encoded contents raises also the confidence of our subjective experiments. Without this similarity, the insufficiency of PESQ's predictions on SILK codec could have been incurred by experiment error of our subjective test. Given that the correlation between the predictions of the two models is as high as 0.9808 and the p-value resulted from ANOVA test is 0.38, the two sets of predictions can be seen as identical, which support the reliability of our experiment setup.

5. CONCLUSION

We measure and model the perceptual quality of Skype calls based on the Weber-Fechner's Law of psychophysics. In that, the relationship between coding bitrate of Skype sources and the perceptual quality is strongly logarithmic. The model is robust with high prediction accuracy. The property that it can be calculated in real time shall assist the decision of rate adaptation for Skype calls under dynamic network environment. In addition, we have demonstrated the insufficiency of existed perceptual evaluation model, PESQ. Our result shows that the prediction of PESQ being too conservative with current codec which provide variable coding rates in a wider range.

6. REFERENCES

- [1] ITU-T Recommendation P.10/G.100. Vocabulary for performance and quality of service. ITU, 2006.
- [2] ITU-T Recommendation P.862. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU, 2001.
- [3] M. Goudarzi, L. Sun and E. Ifeachor. Modeling Speech Quality for NB and WB SILK Codec for VoIP Applications. In *Proc. of NGMAST'11*, 2011.
- [4] M. Cardenete-Suriol, J. Mangles-Bafalluy, A. Maso, and M. Gorricho. Characterization and Comparison of Skype Behavior in Wired and Wireless Network Scenarios. In *Proc. of GLOBECOM'07*, 2007.
- [5] K. Kim and Y.-J. Choi. Performance comparison of various VoIP codecs in wireless environments. In *Proc. of ICUIMC'11*, 2011.
- [6] B. Sat and B. W. Wah. Payout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality. In *Proc. of MULTIMEDIA'07*, 2007.
- [7] IETF Internet-Draft: draft-vos-silk-02, SILK Speech Codec, 2010.
- [8] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. In *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2001.
- [9] K. R. Boff, L. Kaufman, and J. P. Thomas. Handbook of Perception and Human Performance. Wiley- Interscience, 1986.
- [10] 3GPP TS26.171: Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description, 2011.
- [11] ITU-T Recommendation P.830. Subjective performance assessment of telephone-band and wideband digital codecs. 1996.
- [12] 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.
- [13] P. Reichl, S. Egger, R. Schatz, and A. Dalconzo. The logarithmic nature of QoE and the role of the Weber-Fechner Law in QoE assessment. In *Proc. of ICC'10*, 2010.
- [14] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. 2008.
- [15] R. A. Fisher. Statistical methods for research workers. Oliver and Boyd, 1925.