EXPERIMENTS AND MODELING OF PERCEIVED SPEECH QUALITY OF LONG SAMPLES

Xiang-Chun Tan, Stefan Wänstedt, Gunnar Heikkilä

AWARE @ Ericsson Research,
Ericsson Erisoft AB, S-971 28, Luleå, Sweden

ABSTRACT

Speech quality in cellular networks may vary significantly over time. Accessing the perceived speech quality aggregated over time, such as for an entire conversation, is desired to ensure customer satisfaction. Calculating average quality using common objective methods, which normally determine quality for short speech samples, has drawbacks. Subjective listening tests with long speech segments show that the perceived quality differs from the average quality calculated from a series of objective measurements. The overall perceived quality is affected by the brain's "ability" to forget and, hence, the last 30 to 40 s of speech form the basis for the subjective quality.

1. INTRODUCTION

Speech quality is one of the most important QoS measures in present cellular mobile systems. When several operators offer similar coverage, reliability and technical applications, speech quality becomes an important factor to compete with since a major portion of all communication is speech. Operators often try to measure speech quality to ensure that their service provides adequate quality.

Several different measures of speech quality have been devised during the time from early PSTN history until present cellular systems [1]. The quality is, however, measured/defined for short (a few seconds) speech segments. The use of such short stimuli has worked well for PSTN conversations mainly due to that the quality is quite constant [2] and, hence, an average of the series of quality values for an entire conversation resembles the aggregate perceived quality.

In a cellular network, where quality may vary significantly over time, an average may not be the best description of the aggregated speech quality. In fact, the average is believed to be a poor measure, especially for longer speech segments with large variation [1] [3] [4]. An important reason for this is the behavior of the human brain. Psychological research has shown recency and forgiveness effects to be typical of human behavior in perception of speech quality [5] [6]. In other words, the human brain forgives poor quality during a conversation as long as the quality of the ending part of the conversation is adequate. So far there appears to be no support for such effects in the common objective speech quality models.

The work presented in this article is an attempt to characterize and to model aggregate quality of long speech segments. The work is based on subjective listening tests in conjunction with an objective measure of speech quality.

2. EXPERIMENTAL SETUP

2.1. Speech Material

The original speech material is a 40-min long autobiographical presentation by a Swedish female best-seller author. From this material, segments of different lengths were extracted as test samples. The resulting segments were 30, 60, 120 and 180 s long with some variation to ensure that the segments consisted of full sentences and made sense on their own.

2.2. Channel profiles

The channel profiles are based on a typical urban (TU) environment with ideal frequency hopping. Two groups of channels (disturbances) were used. The first type of channel comprises a single disturbance on an otherwise good channel, Fig. 1(a). The position, severity and length of the disturbance varies, Table 1. The second group of channel files comprises alternating parts with good and bad quality, respectively, over the entire length, Fig. 1(b). The intervals between highs and lows are randomly varied around 7.5 s while the average C/I of the sample is kept constant, on one of three levels. For each C/I level there are four different envelope forms. The trend of the upper (A in Fig. 1(b)) and lower (B) envelopes may be constant, increase or decrease in a channel file, so that the quality trend may be positive or negative towards the end of the sample.

A total of 72 samples (36 samples of type 1 and type 2, respectively) were created with the parameter settings given in Table 1 using multi-factor design.

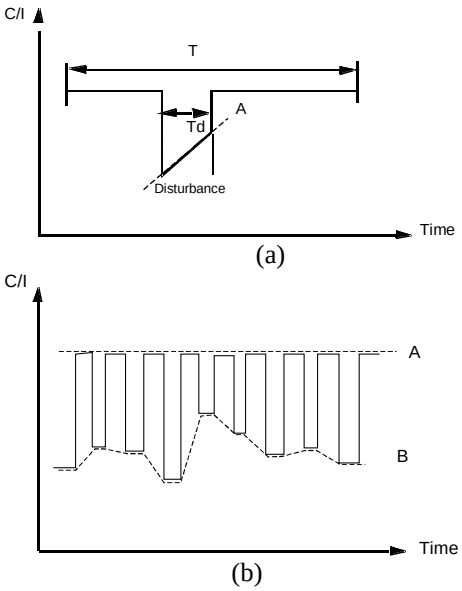

Fig. 1. Test profiles. (a) Single disturbance. (b) Multiple disturbances (A constant and B random).

Table 1. Some characteristics of the two types of profiles

Parameter	Type 1	Type 2
Total time, T	30, 60, 120, 180 s	60, 120, 180 s
Disturbance, Td	10, 20, 30 s	Around 7.5 s
Location of disturbance	Beginning, middle, end	Spread over whole samples
Average C/I		7, 9, 10 dB
Disturbance C/I	4, 8 dB	
Trend for distur-bance/envelop A, B	Constant, linear increase or decrease	Constant, linear increase, dec-rease or random

2.3. Simulation of tests samples

The test samples were simulated using the GSM-EFR codec and the channel with disturbance. The original sound and channel files were input to the simulator and the processed speech as well as quality information from the air interface, e.g. BER and FER, were produced.

The Speech Quality Index (SQI) [7], an objective measure of speech quality, was calculated using the information from the radio interface. SQI is based on air-interface QoS parameters [7], namely the bit error rate and frame erasure distributions for each 2.5 increment of speech. The accuracy of SQI is comparable to e.g. PSQM [8]. The SQI for an entire speech segment, say 30 s long, is the average of 12 SQI values spaced in 2.5 s intervals.

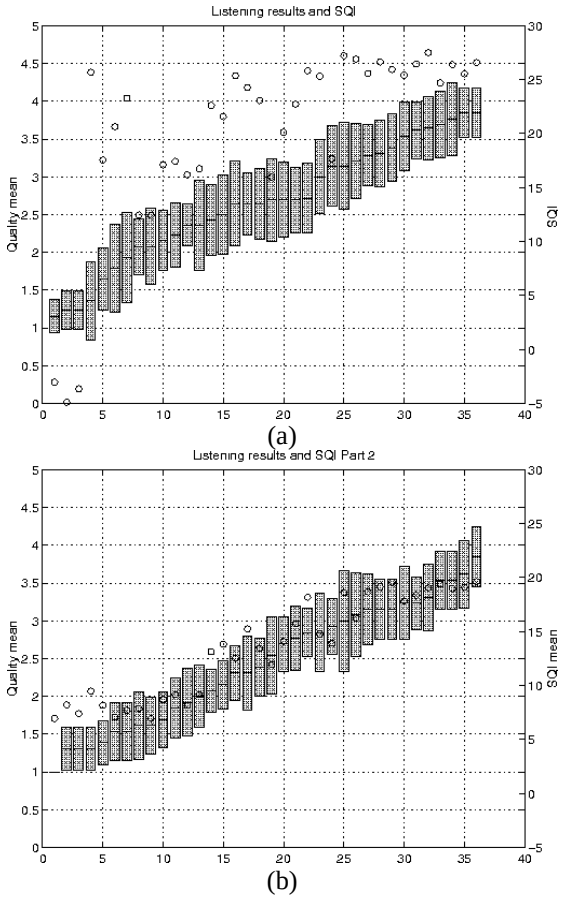

Fig. 2. Subjective and objective quality of the speech samples from the two groups of disturbances. Results of listening tests for single disturbance (a) and multiple disturbances (b), with 95% confidence intervals. The circles represent SQI averages of the samples.

2.4. Listening procedure

The subjects listened with both ears using headphones. After listening to a sample, they were prompted to grade the sample according to the recommended 5-point absolute category scale. Once all the listeners had graded a sample an average was calculated, that is, the mean opinion score (MOS).

A total of 29 people participated in the listening test. They were divided in two groups. Each listened to about 60 min of speech arranged in two 30 min sessions, corresponding to the two different types of disturbance profiles.

3. RESULTS

3.1. Subjective vs. objective quality

The listening scores and the average SQI of the two types of profiles, sorted in order of magnitude, are depicted in Fig. 2.

The correlation between the perceived quality and average SQI is high for the multiple disturbance channels, Fig. 2 (b). For the profiles with a single disturbance, Fig. 2 (a), the difference between the subjective quality and SQI is larger. This large variation indicates that human perception of aggregate quality differs from an arithmetic mean of several comparatively short quality values. The three SQI samples in the lower left corner correspond to the samples whose disturbances extend over the entire length of the sample (30 s).

Fig. 3 shows how the location of the disturbance in a sample affects the quality, when other conditions are identical. The closer the disturbance is to the end, the lower the perceived quality. The fading of memory, or the recency effect influences the perception. The average SQI is naturally not affected by the location of the disturbance.

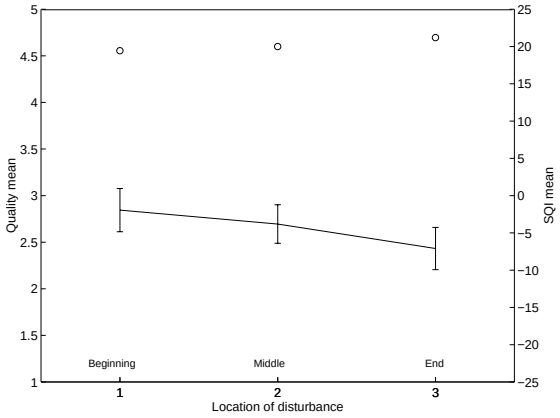

Fig. 3. Subjective and objective quality of type 1 samples with 95% confidence intervals. Circles represent average SQI.

3.2. Initial models

As has been shown in the above section, the recency effect is involved in human perception of long speech samples. This is confirmed with PLS (partial least square) models relating the quality to SQI mean. The model results indicate that for multiple disturbances the average is a good measure of the aggregated quality; the cross-validated correlation coefficient (Q^2) is 0.9. For the single disturbances and average SQI, on the other hand, Q^2 is only 0.5.

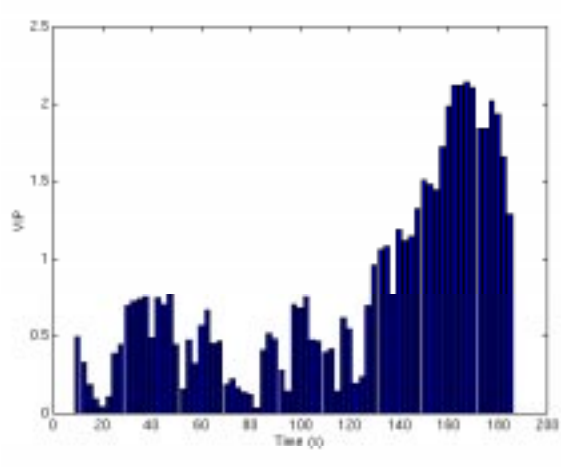

Fig. 4. Variable importance plot of the SQI series from a 180 s long speech sample. Each bar represents the importance of a single SQI.

Multivariate analysis was carried out to model the forgiveness effect. The SQI values of 2.5-s intervals were the initial variables and the overall quality the dependable variable. The variable importance (VIP) was calculated to see how the overall quality was affected by the series of individual SQI values. VIP represents the sum of variable influence over all model dimensions [9]. Here, a SQI variable with a VIP larger than one has an above average influence on overall quality. The VIP plot of the SQI model shown in Fig. 4 indicates that the SQI values in the last quarter of the series are the most important for the perceived quality.

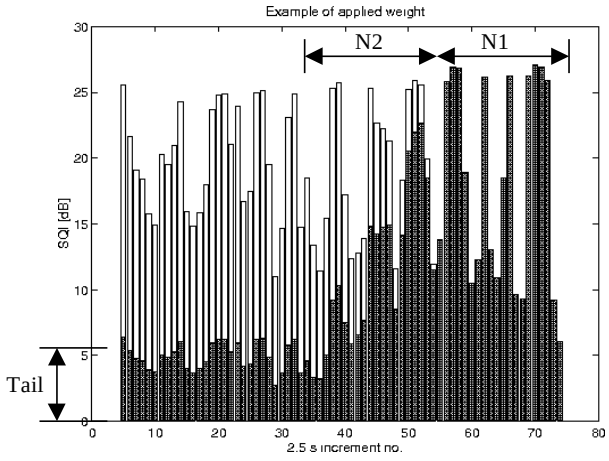

Fig. 5. Example of weights applied to SQI data for a 180-s speech sample. Empty bars are primary data, filled bars represent data with applied weights. N1=20 (50 s), N2=20, linear weights for part N2 ending with tail 0.25.

3.3. Weighted average model

Since the quality close to the end of a speech sample is more important for the perceived aggregated quality, weights are applied to the SQI series. The SQI variables near the end (N1) are kept unchanged, while those in the middle (N2) are multiplied by weights less than one depending on the selected weight function. The weight of the remaining samples (tail) may be zero in some cases. An example of weights applied to a sequence of SQI variables is shown in Fig. 5.

The optimum set of weight variables was found by optimizing the model performance; that is, for each combination of weights a Q^2 for the model was calculated until a maximum Q^2 was found. The final model parameters are N1 = 5 (12.5 s), N2 = 30 (75 s) and a linear weight function for N2 ending at zero (tail=0). The weighted average of SQI is calculated as:

$$SQI_{wt} = \frac{\sum_{i=1}^{N1} SQI_i + \sum_{i=1}^{N2} wt_i \bullet SQI_{i+N1}}{N1 + \sum_{i=1}^{N2} wt_i}$$

and used as the input variable in PLS modeling.

Fig.6 depicts the resulting model vs. the observed values. The model (Q^2 = 0.65), which covers both types of disturbances, is much improved compared to a pure average model (Q^2 = 0.57). The 3 outliers in the lower left part of the figure represent samples with disturbances extending the entire sample (30 s).

4. DISCUSSIONS AND CONCLUSIONS

The listeners are more sensitive to the quality of the later part of a speech segment. The last 30 to 40 s of speech appears to play an important role in the grade of the whole sample even though the sample is much longer. Simply put, if the message is longer than 30 s the listeners start forgetting the quality of the part of the sample prior to the last 30 s and forgive parts with deteriorated quality early in a sample, if the quality of the ending part is good. On the other hand, the listeners perceive severe deterioration in quality if the disturbance is heard recently.

The effect of “forgiveness” is not as pronounced for the channels with multiple impairments. The reason for this may be that the variations in quality along the entire sample results in less contrast of memory of quality which occur in the single disturbance. In other words, there is really no time for the listener to adjust to “good” quality or to “forget” since the quality varies all the time, that is, at least twice every 30 s. Therefore the mean SQI model predicts the quality rather well for the samples with multiple disturbances.

Adding weights to emphasize the importance of final parts of the sample results in a better model with higher correlation to the perceived quality compared to a simple arithmetic mean.

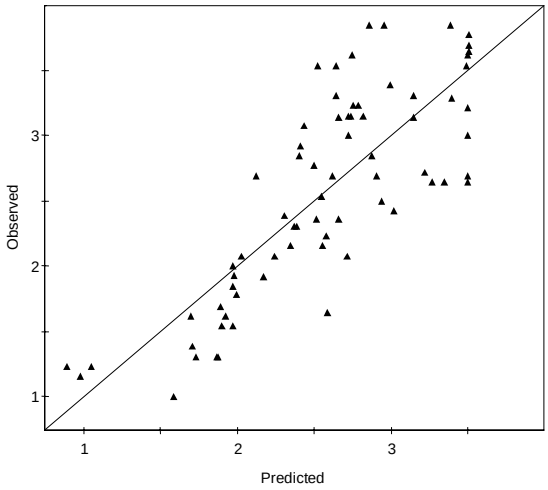

Fig. 6. Prediction vs. observed overall quality using weighted average of SQI.

REFERENCES

[1] ITU-T P.800 Methods for subjective determination of transmission quality.

[2] Rosenbluth, J.H. (1998) Testing the quality of connections having time-varying impairments. ITU Study group 12 - Contribution 64. COM 12-64-E

[3] Chateau, N. (1999) Continuous assessment of time-varying subjective vocal quality and its relationship with overall subjective quality. ITU Study group 12 - Contribution 94. COM 12-94-E

[4] Hollier, M. (1997) An experimental investigation of the accumulation of perceived error in time-varying speech distortions. ITU Study group 12 - Contribution 21. COM 12-21-E

[5] Watson, A. and Sasse, M.A. (1998) Measuring perceived quality of speech and video in multimedia conferencing applications. ACM Multimedia 98.

[6] Watson, A. and Sasse, M.A. (1998) Multimedia conferencing via multicast: Determining the QoS required by the end user.

[7] Karsson, A., Heikkilä, G., Minde, T., Nordlund, M. and Timus, B.. (1999) Radio link parameter based speech quality index – SQI, 1999 IEEE Speech Coding Workshop, Haikko Manor, Porvoo, Finland, June 20-23.

[8] ITU-T P.861 Telephone transmission quality methods for objective and subjective assessment of quality.

[9] SIMCA 7.0: A New Standard in Multivariate Data Analysis. User’s guide to SIMCA, UMETRI, 1998.