ESTIMATION OF 'QUALITY PER CALL' IN MODELLED TELEPHONE CONVERSATIONS

J. Berger, A. Hellenbart, R. Ullmann
SwissQual AG
Switzerland

B. Weiss, S. Möller
Deutsche Telekom Labs, Berlin
University of Technology
Germany

J. Gustafsson, G. Heikkilä
Ericsson AB
Sweden

ABSTRACT

We present a method to estimate the perceived listening quality by a subscriber at the end of a common voice telephony conversation. This method was recently introduced in ETSI STQ mobile and was approved as TR 102 506 'Speech Quality per Call' [1].

The idea is to calculate this "speech-quality-per-call" value based on short-term listening quality scores (so-called Mean Opinion Scores, MOS), as they are usually derived by subjective listening-only tests, or based on predictions of short-term scores by means of objective measures. It is shown that a pure linear averaging of short-term scores will not predict the perceived quality of the entire call sufficiently well in case of a non-stationary quality over the call. Mainly the "recency effect" and the out-weighting of very bad parts in a call have to be considered in an adequate way.

An algorithm was developed that allows the obtained "speech-quality-per-call" score to be predicted on the basis of the MOS of the individual utterances. The algorithm can be applied for various lengths of call and numbers of individual utterances. Since speech quality is usually objectively predicted in real networks the approach was also proven and confirmed for objectively obtained quality scores.

This paper follows widely [1] the work and the decisions taken within ETSI STQ mobile.

Index terms **-** Listening Quality, Speech Quality, Telephone Conversation.

1 INTRODUCTION

The established way of characterizing speech quality delivered by networks or telephony applications is the evaluation of listening quality (often simplified: speech quality) on the basis of short voice excerpts. This evaluation can be performed by means of subjective listening-only tests or – mainly in automated test systems – by an objective prediction of the listening quality, e.g. using the model described in ITU-T Recommendation P.862.1 [2].

Using that established way and taking advantage of the data collected in that fashion one can seek to estimate the perceived speech quality of an entire call. In fact, short excerpts of 4-8s length do not entirely describe a typical telephone call of 60-120s length, but may be considered as parts of such a call.

Simple approaches try to average over a large amount of short voice excerpts, but they do not necessarily paint an accurate picture of the customer satisfaction after a longer conversation. In fact, a very bad excerpt can be outweighed by a couple of good excerpts, but there is experimental evidence that bad excerpts have a stronger impact on the call quality than simple averaging suggests. On the other hand, threshold models regard a call as fair or poor on the basis of one or two degraded excerpts. Such threshold models, however, do not take the number of good or excellent excerpts into account. Finally, in a third type of models a certain percentage of the samples needs to be degraded in order to rate the call as bad; these models disregard the temporal structure of the call and the relative timing of the degradation towards the end.

Thus, it is worthwhile to develop a new model which is able to accurately predict the subjective quality rating obtained at the end of a typical call of 60-120s length on the basis of the individual conversational contributions which are typically 4-8 s long. Apart from its practical value for call quality monitoring, such a model will improve our understanding of the impact of varying speech quality during a conversation.

Note that the present document focuses on the speech (listening) quality of a voice call. Conversational properties such as talker quality, round-trip delay and its impact on the interactivity of a conversation are not considered. Speech quality of video telephony applications is not considered either.

2 CALL PROPERTIES

The determination of typical call properties like the call length and the length of conversational contributions (samples) can be based on existing specification like ITU-T Recommendation P.862.3 [3], ETSI TS 102 250 [4], and especially TS 102 250-5 [4]. On this basis, we define a new structure of a call which is sensitive to speech quality degradations.

Calls, that can be mobile-originated, mobile-terminated or mobile-to-mobile, can be divided into different groups. Short calls of a couple of seconds where there is an announcement or a voice-box message are usually too short to provide meaningful listening-quality measurements. A "typical" call contains a dialog where utterances are exchanged. Ideally, these utterances are distributed evenly in length and frequency between the two conversation partners. On each side, a certain period of speech activity is followed by a silence period of the same length. Since the call quality is rated independently on each side, it is sufficient in an instrumental or subjective model of a conversation to feed one side only with the required sample pattern.

The length of the call must give room for a couple of utterances (samples). The call length recommended in TS 102 250 [4] is 120s which sufficiently fulfils this requirement. In fact, the average call length may reach this 120s, however, the median of the call durations is well below. For practical purposes it is desirable to use call lengths that are considerably shorter that 120 seconds too. Thus, the subjective evaluations made in this paper used conversation lengths of 60s and 120s.

3 CALL STRUCTURE AND SUBJECTIVE EVALUATION

In [3], [4] and [5], a typical utterance in telephony calls is assumed to be 5 to 12s long. Depending on the length of the call in connection with the length of the individual utterances, it takes from 5 to 12 utterance-and-silence pairs to fill a call length of 60s or 120s. From empirical evidence we know that a typical conversational call contains around 4 utterances from each side, so that 5 recurrences of the utterance-and-silence pairs can be recommended. Considering that these values are applicable to short calls, longer calls can accommodate up to 12 utterance-and-silence pairs with an individual sample length of 5s.

3.1 CALL DESIGN

The conversational call that is to be rated to estimate the call quality is assumed to consist of alternating phases of speech activity and silence, the length of the phases should be 5 s to 12 s, and the phases should recur 5 to 12 times during a call. It is further assumed that the call structure is symmetric. This means that each partner has active speaking parts of a given length and the same duration for listening to the other partner. Thus, a conversational exchange can be defined by an active part and a corresponding listening part (Fig. 1).

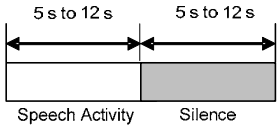

Fig. 1: Structure of the utterance-and-silence pair

An entire call can then be described by a series of such utterance-and-silence pairs, as it is shown in Fig.2.

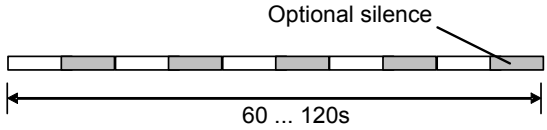

Fig. 2: Structure of a modelled conversation (one side)

Under the assumption that later in automated test systems both sides (A and B in Fig. 3) can be equipped with measurement devices, an inverse structure can be used at the other end of the connection, see Fig. 3.

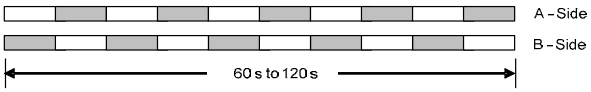

Fig. 3: Structure of the call with 5 recurrences and alternating speech activity

3.2 MODELLING OF THE CONVERSATION

The conversation should be modelled in such a way that it can be scored in a listening-only context. However, in a real conversation the talkers own speech activity is alternating with listening parts. The pure listening to a 60s or 120 s recording of speech interrupted by pauses will surely not model the actual behaviour in a conversation. For that reason, the missing speech activity was replaced by another activity to be performed by the listening person. We chose a keyword spotting, where the listener has to follow the content of the voice signal, and to answer a question related to the current utterance speaking aloud (tests in German) or tagging the answer on paper (tests in English).

The voice excerpts to be presented in the test were recorded in a studio environment. The content can be considered as typical for a telephone conversation (e.g. calling a rental car company, making a medical appointment). The individual excerpts were band-limited to telephone band and some of them were transmitted over various telephone channel to introduce artefacts and severe distortions. A goal of the simulation of the conversation was to consider different profiles of quality over time in our model. Thus, the presentation could start with a low quality excerpt but already in the second excerpt the quality is high. All in all, 10 different 'quality profiles' were used in each of the individual investigations. An example is shown in Fig. 4.

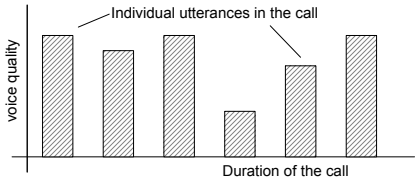

Fig. 4: Example of varying speech quality over the call duration

3.3 SUBJECTIVE TESTING

The idea of the test methodology and a first mathematical model were already introduced in an internal study of Deutsche Telekom in 2002 [6]. Since this study was carried out with only a few experienced test subjects, a validation of the approach was required. For this aim, two studies were conducted by ETSI STQ mobile, one with German and one with English speech samples. The studies investigate call lengths of 60s and 120s to cover the range the model should be designed for (see 3.1 Call Design). In each of the studies, subjective experiments were carried out in two steps:

First, the "speech-quality-per-call" was evaluated at the end of the presentation of the simulated conversations. Each experiment contains 20 (for the 120s series) to 40 (for the 60s series) different presentations of simulated conversations. Conversations varied with respect to the used quality profile (10 in each series), the speaker (two male/two female), and the text contents ('car rental', 'dentist' and similar).

The individual modelled conversations were presented in a random order in a listening-only context similar to the one described in [6]. The subjects (between 24 and 28 in the individual experiments) were requested to listen to the excerpts and to do a keyword-related activity during the silence period. At the end of the presentation of a modelled conversation, the subjects voted the perceived listening quality for the entire presentation. The common five-point category MOS scale accordingly to ITU-T P.800 was used for scoring [5]. At the end of this experiment, the "speech-quality-per-call" scores were derived by averaging the individual scores of the subjects for each presentation. These values form the target values for the algorithm to be developed.

In a second step, all the individual utterances (the 5 to 12s excerpts the modelled conversation consist of) of each study were scored by all subjects in a separate session, using the common listening-only test procedure according to [5]. At the end of this experiment, MOS values for all short-term excerpts were available, in addition to the target values describing the quality for the complete modelled conversation consisting of a series of excerpts.

4 MODELLING CALL QUALITY

The model to-be-developed should aggregate the individual MOS values to one value, considering the temporal structure of good and bad excerpts within a simulated call. Two important effects are taken into account: the "recency effect" and the effect of a very bad sample in a call. Already in [6] it was shown that a simple averaging of all of the short term evaluations will not form a confident predictor of the "speech-quality-per-call". Only in case of constant quality over the entire call averaging may be sufficient, but in realistic cases, where the quality may drop anywhere during a conversation, the linear average fails in predicting the targeted scores.

4.1 IMPACT OF BAD SAMPLES TOWARDS THE END OF A CALL

The impact of degradations that occur towards the end of a call are considered in the so called "recency effect". The closer a "bad event" is towards the end of a conversation, the stronger is its impact on the overall rating of the entire call. In the chosen call structure, the speech samples are numbered from 1 to n. The weighing is performed by a weighting factor a_i in the following way:

$$MOS_{RE} = \frac{\sum_{i=1}^{n} a_i MOS_i}{\sum_{i=1}^{n} a_i}$$

If the time between the end of the last sample and the middle of sample i is t_i, then we apply the following weighting factors:

$$a_i = \begin{cases} 0.5\dfrac{(19 - t_i)}{19} + 0.5 & for \quad t_i < 19; i \in [5;12] \\ 0.5 & otherwise \end{cases}$$

This formula represents the increasing importance of a sample for the overall quality the closer it is located towards the end of a call. The coefficients were gained by a multiple linear regression at first. To avoid over-training on the restricted amount of test data, we simplified the linear function manually in a second step.

4.2 IMPACT OF A SINGLE BAD SAMPLE

The model can be significantly improved by taking additionally into account the worst sample of the call. Empirical evidence shows that one very bad sample strongly deteriorates the overall quality, in addition to its temporal occurrence. Thus, the model is extended to include the worst sample in the call in the following way:

$$MOS_{SpQ-C} = MOS_{RE} - 0.3\,(\overline{MOS} - \min(MOS_i))$$

The formulae were developed for conversations with a length between 60s and 120s containing 5 to 12 utterances per analyzed direction and with sample and pause lengths of 5s to 12s each. Thus, the same formula can be applied to the entire range of call and utterance length' as defined in '3.1 Call Design'.

The basic approach of this formula was taken from [6], the coefficient was optimized by an automated iteration loop. Due to the restricted amount of test data the coefficient is limited to one digit only.

Correlation coefficient (RMSE in parentheses)	Study 'English' (5s samples)		Study 'German' (5-6s samples)	
Call length	**120s**	**60s**	**120s**	**60s**
Linear Average with MOS-LQS	0.92 (0.66)	0.88 (0.63)	0.83 (0.51)	0.85 (0.49)
CallQuality model with subjective MOS	0.98 (0.21)	0.97 (0.22)	0.93(0.31)	0.94 (0.26)
CallQuality model with predicted MOS (P.862.1)	0.97 (0.32)	0.96 (0.33)	0.84 (0.42)	0.89 (0.35)

4.3 VALIDATION OF THE MODEL

The formulae have been validated with modelled conversations with various lengths and different speech sample lengths in German and English. The scores predicted by the formula show a significant gain in correlation with the subjectively obtained scores for the "speech-quality-per-call" in comparison with the linear averaging for all tested scenarios.

Table 1 shows the obtained results in a condensed manner. The four columns represent the four experiments (two simulated call lengths for each study). The individual numbers were derived by applying different models and methods for predicting the "speech-quality-per-call" scores. We have limited the presentation here to Pearson's correlation coefficient and the RMSE of the prediction.

In the first row, we have averaged the individual short term MOS as a prediction of the "speech-quality-per-call" scores. It can be seen that the correlation coefficients seems to by sufficiently high, but the residual prediction error is above 15% of the used 1 to 5 point scale.

By applying the introduced formulae the prediction confidence is significantly increased. This confirms our initial assumption that the simple averaging will not reflect the user's behaviour at the end of the call.

So far, we applied the model only to the subjectively-derived MOS values. However, the model can equally well be used with instrumentally-derived values provided by automated test tools, since subjective testing cannot be applied in real field measurements. With ITU-T P.862.1 [2] we have an objective model that is applicable for evaluation of short voice samples and predicts the MOS. Consequently, the model should be able to use those predictions too instead of subjective sores. The final row of Table 1 shows the performance of that model by using objective MOS values as an input. The performance remains high. The difference between both studies in this case can be explained by the different transmission conditions used. The "English" study uses only different AMR speech codec varieties; the "German" study involved a wider range of distortions. Here the deviation of the predicted MOS derived by [2] to the subjective scores is commonly larger, which leads to a worse approximation of the short term listening quality values by our model, i.e. lower correlations.

5 CONCLUSION AND OUTLOOK

The perceived speech quality is not a simple aggregation (average) of the quality of individual samples in a call. Experimental evidence shows that the impact of a degraded speech excerpt can not simply be undone by a longer stretch of good or acceptable listening quality. For single calls the temporal structure of the call must be considered.

With the presented model it is possible to estimate, with high accuracy, the perceived (subjective) speech quality of a call for each side on the basis of (objectively or subjectively) rated samples for a given frame of controlled call structures.

The weightings we found will reflect the range of structures of our simulated calls. In calls with different structures (e.g. continuously read texts) they may be different.

6 REFERENCES

[1] ETSI TR 102 506: Estimation of Quality per Call.

[2] ITU-T Rec. P.862: "Perceptual Evaluation of Speech Quality"; P.862.1: "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[3] ITU-T Rec. P.862.3: "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2".

[4] ETSI TS 102 250 (all parts): "Speech Processing, Transmission and Quality Aspects (STQ); QoS aspects for popular services in GSM and 3G networks".

[5] ITU-T Rec. P.800: "Methods for subjective determination of transmission quality".

[6] J. Berger: "Call Quality – PESQ mobil". Internal Report, T-Nova Berkom, 2002 (in German), Extract is published in Annex A of [4].

ACKNOWLEDGMENT

The authors would like to thank T-Mobile Deutschland GmbH and E-Plus, Germany, for their generous contribution to the German study. Furthermore, we gratefully acknowledge the support of the entire ETSI STQ mobile group for carrying out this investigation.