A Dependence between Average Call Duration and Voice Transmission Quality: Measurement and Applications

J. Holub¹, J. G. Beerends², R. Šmíd¹
¹FEE CTU, Dept. of Measurement, Czech Republic
²TNO Telecom, The Netherlands
holubjan@feld.cvut.cz, J.G.Beerends@telecom.tno.nl

Abstract

This contribution deals with the estimation of the relation between speech transmission quality and average call duration for a given network and portfolio of customers. It uses non-intrusive speech quality measurements on live speech calls. The basic idea behind this analysis is an expectation that the average call duration is higher for calls of a good quality since users are less disturbed during their conversation by transmission impairments. Both traditional and state of the art methods for non-intrusive speech transmission quality measurements and also for call duration monitoring are briefly described and two basic ways of the dependency analysis are shown. The relations for result uncertainty estimation are derived. A numerical example is given based on a limited training database that is constructed on the basis of real network data. Finally, potential applications are presented and discussed. The first results indicate that in the low to medium quality range (Mean Opinion Scores between 1 and 3) the call duration increases with increasing speech quality.

1. Introduction

Speech transmission during any call in the telecommunication network is affected by many impairments like delay, echo, various kinds of noise, speech (de)coding distortions and artifacts, temporal and amplitude clipping etc. Each transmission impairment has a certain perceptual impact on the speech transmission quality. The overall quality can be evaluated and expressed in terms of a Mean Opinion Score (MOS) covering the range from 1 (bad) to 5 (excellent).

Each call has certain Call Duration (CD) that is primarily influenced by purpose of the call, by habits and number of topics discussed by call participants etc. and appears to be a virtually random variable covering a range between a few seconds and several tens of minutes (or rarely up to few hours).

Apparently, there is no direct relation between MOS and CD of the calls. In many calls this is true but there are conversational cases when increased speech transmission quality can positively influence call duration. With high conversational speech quality it is more likely that the participants may find additional topics to discuss (that were not a primary reason for that call), being less disturbed by transmission impairments. Those calls contribute to bias in the distribution of CDs over various MOS intervals. As a result, after processing a significant number of calls in the given network, a slight increase of average CD for increasing MOS can be identified. Such a result can be conveniently used by network operators to predict potential increase of call duration after network technology upgrade or reconfiguration that may affect the speech transmission quality. Other potential application fields are optimization of technology investments or suitable operator’s selection of recommended user terminals.

2. Non-intrusive Speech Transmission Quality Measurements

The following three groups of speech transmission quality measurement can be distinguished: Listening and conversational tests, intrusive objective measurements and non-intrusive objective measurements.
2.1. Listening and Conversational Tests

A trivial method of measuring quality would be to ask callers for their opinion after a call has been made. Due to obvious practical problems related to this approach, listening and conversational tests have been standardized instead as the methods for subjective determination of transmission quality. These tests relate real world distortions created in a laboratory environment to the subjectively perceived quality. E.g. recommendation [3] describes approved methods which are considered to be suitable for determining how satisfactory given telephone connections may be expected to perform. They contain recommended subjective evaluation procedures for conversational and listening-only tests. For establishing a relation between CD and MOS subjective tests are not practical.

2.2. Intrusive Objective Measurements

Intrusive measurements of speech transmission quality usually require special test calls generated by the measurement system and require that the original (non-distorted) speech sample is available to the measurement algorithm. The algorithm itself then compares original and transmitted speech samples and identifies and integrates the perceptual differences between them. Known psycho-acoustical aspects of human hearing (human ear loudness and frequency resolution and sensitivity, temporal and frequency masking, etc.) are/should be modeled by the algorithm to estimate the subjectively perceived quality in terms of the MOS value as would have been obtained in a listening tests. A typical example of an intrusive algorithm is PESQ [4],[10],[11],[12]. The correlation coefficient between the PESQ MOS estimate and the related MOS from a formal listening tests is in most cases above 0.9. PESQ was validated for various transmission and coding technologies including mobile networks and Voice over Internet Protocol (VoIP) transmissions. The typical length of the analyzed speech samples is 8-12 s. For establishing a relation between CD and MOS intrusive measurements are also not practical.

2.3. Non-Intrusive Objective Measurements

Passive monitoring of on-going calls in the network appears as the only suitable approach for our purpose. INMD’s (In-service, non-intrusive monitoring devices), as specified in ITU-T rec. P.561 [5], typically use a non-intrusive approach, allowing to monitor large amounts of live traffic. Impairment-related parameters like echo, signal to psophometrically weighted noise ratio etc. are collected from the voice channel (the recommended length of the speech sample in this case is 20s of active speech per direction) and combined by means of a suitable algorithm (P.562 [6]), to a final estimation of speech quality (CCI – Call Clarity Index). The results that are presented in this paper are all based on this type of non-intrusive objective measurements.

Another ITU-T non-intrusive speech quality analysis approach is known under the working title P.SEAM [8]. It has recently been approved by ITU-T after a selection procedure where two algorithms were compared [7], [8]. The P.SEAM combines three non-intrusive algorithms and achieves a correlation coefficient with listening tests of around 0.8.

In case of mobile networks where the radio link can be considered as the weakest part, algorithms that estimate MOS from radio link parameters like Receiving signal level (RxLev) or Receiving Signal Bit Error Rate (RxQual) may be conveniently used [2]. Recent research [1],[13] shows that a correlation with PESQ results above 0.90 can be achieved in this case. However, the quality estimating algorithm has to be trained for the used network technology and requires retraining after any significant network change or upgrade.

3. Call Duration Measurements

Call Duration (CD) measurement is much more simple than MOS estimation. There are more definitions of CD available, differing in how they define start and end points of the call. For our purpose, the ideal CD should contain only the time duration between the end of a successful call setup and the call termination by one of the callers. In this way time periods like call setup and call release are excluded from the measurement. As a source of CD data in given network it may be possible to use either INMD records (call duration is a non-mandatory parameter of INMD) or data from Call Detail Records as acquired directly by the switches primarily for billing purposes. In the second case, it is necessary to correlate properly data from MOS measurements and CDR databases using a proper common parameter(s).

When MOS estimation based on radio link parameters is used in the GSM network, CDRs acquired on A-bis and A-interfaces are a convenient source of data since A-bis based CDRs contain thresholded RxQual and RxLev parameters and A interface based CDRs contain CD parameter.
4. CD versus MOS Dependence Analysis

We assume to have as an input data QCDR (Quality Call Detail Record) database, containing at least MOS and CD for each analyzed call. The analyzed calls should be acquired on average link selection that, as for its structure, corresponds to link structure of the entire network. In the following example, a limited training database that has been constructed on the basis of real network data will be used. An example of QCDR records (first 4 out of 1 000 000 data pairs) is given in the Tab. 1 and the first 1000 pairs CD-MOS are depicted on the Fig. 1. Distributions of MOS and CD values are shown in Fig. 2 and 3, respectively.

<table>
<thead>
<tr>
<th>MOS</th>
<th>2.84</th>
<th>3.51</th>
<th>4.06</th>
<th>3.08</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD (s)</td>
<td>5.0</td>
<td>55.1</td>
<td>113.6</td>
<td>138.3</td>
<td>...</td>
</tr>
</tbody>
</table>

For the given training QCDR database, the average MOS is \( \text{MOS}_{\text{aver}} = 3.62 \) and total average CD is \( \text{ACD} = 107.1 \text{s} \). The dependence of average CD on MOS can be evaluated either by direct polynomial fit, or the aggregation of CD values according similar MOS estimates. There are two elementary ways of this aggregation, fixed MOS interval or fixed variance approach.

4.1. Raw Data Fitting

The most straightforward calculation way is the direct polynomial fit of the raw QCDR data. The examples are shown in Fig. 4 to 6.

The output of the fitting is a set of polynomial coefficients. The fit should be such that the resulting curve is monotone over the whole MOS range. No information about validity MOS range and about dependency uncertainty for different MOS values is available. The correlation coefficient can be calculated...
to estimate how much variance of CD the MOS model can explain. Based on the raw data as given in Fig. 1 one can expect that high correlations can be found in the MOS range 2 to 3.5. It is necessary to remove masking points of high MOS where the correlation is not so significant. The results for the correlation calculation for the different MOS thresholds are shown in Table 2.

Table 2 Correlation coefficient for different number of QCDR samples used

<table>
<thead>
<tr>
<th>MOS upper threshold</th>
<th>Number of calls</th>
<th>Correlation coefficient $R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>26</td>
<td>0.3951</td>
</tr>
<tr>
<td>2.0</td>
<td>778</td>
<td>0.5317</td>
</tr>
<tr>
<td>2.5</td>
<td>7913</td>
<td>0.6055</td>
</tr>
<tr>
<td>3.0</td>
<td>56799</td>
<td>0.6575</td>
</tr>
<tr>
<td>3.5</td>
<td>295850</td>
<td>0.6352</td>
</tr>
<tr>
<td>4.0</td>
<td>924698</td>
<td>0.2910</td>
</tr>
<tr>
<td>4.5</td>
<td>1000000</td>
<td>0.1232</td>
</tr>
</tbody>
</table>

4.2. Fixed MOS Interval Aggregation

Average CD is calculated on calls which MOS estimates fall into certain interval (e.g. 2.0-2.1). Consequently, the MOS interval is shifted and the procedure is repeated (2.1-2.2, 2.2-2.3, ...). In addition, a variance analysis is performed for each interval to obtain confidence intervals (measurements of the uncertainty). The main advantage of this approach is that sorted QCDR database is not required, however, there are two obvious disadvantages: The resulting uncertainty changes from interval to interval and database processing is relatively slow since multiple search operations must be performed.

Optimal number of MOS intervals into which the range 1 to 5 is to be divided depends on total number of analyzed call available. If the number of intervals is low, the dependency approximation is too rough (containing few linear parts only). On the other hand, if the number of MOS interval is too high, there are only few calls even in the most frequently occurring MOS intervals and resulting uncertainty is high. For the previously described QCDR database example, the solutions for number of MOS intervals of 8, 100 and 1000 is shown in the Figures 7, 8 and 9.

4.3 Fixed Variance Aggregation

In the third approach, the database is separated into groups with fixed number of calls per group that assures stable uncertainty. The main disadvantage of this approach is the requirement for MOS-sorted database, however, the constant uncertainty and fast processing (after the database is sorted) are to be considered as significant pros of this method.

For our QCDR database example, the cases for 100, 1000 and 100 000 calls per group are depicted in the Figures 10, 11 and 12.

It is obvious that optimal aggregation depends on the consequent application since few intervals with many calls in each of them smooth results and reduce confidence interval but do not provide results on both ends of the MOS scale where the call occurrence is not so dense and vice versa.
Fig. 7 Data aggregation into 8 equidistant MOS intervals (95% confidence borders for the regression lines are shown).

Fig. 8 Data aggregation into 100 equidistant MOS intervals (95% confidence borders for the regression lines are shown).

Fig. 9 Data aggregation into 1000 equidistant MOS intervals (95% confidence borders for the regression lines are shown).

Fig. 10 Data aggregation into MOS intervals containing 1000 calls each (95% confidence borders for the regression lines are shown).

Fig. 11 Data aggregation into MOS intervals containing 10 000 calls each (95% confidence borders for the regression lines are shown).

Fig. 12 Data aggregation into MOS intervals containing 100 000 calls each (95% confidence borders for the regression lines are shown).
4.4 Uncertainty Analysis

In most applications, including the one shown in the next chapter 5, it is useful to evaluate the sensitivity coefficient \( a \) that is defined as

\[
y = ax + b \quad (1)
\]

Where \( x \) is an aggregated MOS and \( y \) is the related averaged CD (symbols \( x \) and \( y \) are used to keep the following equations short). Linear increase in the neighborhood of MOS interval of interest is expected. To estimate the sensitivity coefficient \( a \) from \( n \) neighboring pairs of \((x_i,y_i)\), the least square method can be used:

\[
\sum_{i=1}^{n} (ax_i + b - y_i)^2 = \min \quad (2)
\]

It can be shown that in this case

\[
a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \quad (3)
\]

and

\[
b = \frac{\sum y_i - ax}{n} \quad (4)
\]

The uncertainty of the sensitivity coefficient \( a \) is then expressed as

\[
u_a = \sqrt{\frac{1}{k_i} \left[ \left( \frac{\partial a}{\partial x_i} \sigma_{x_i} \right)^2 + \left( \frac{\partial a}{\partial y_i} \sigma_{y_i} \right)^2 \right]} \quad (5)
\]

where \( \sigma \) is standard deviation of \( x \) respectively \( y \) for relevant set of calls (=in the \( i \)-th interval), \( k_i \) is number of calls in the \( i \)-th interval and and

\[
\frac{\partial a}{\partial x_i} = \frac{(y_i - \bar{y})(\sum x_i^2 - n\bar{x}) - 2(\sum x_i y_i - n\bar{x}\bar{y})(x_i - \bar{x})}{(\sum x_i^2 - n\bar{x}^2)^{3/2}} \quad (6)
\]

5. Applications

The above mentioned approach can be conveniently used for predictions of network technology upgrade impacts. The average MOS increase \( MOS_{inc} \) caused by potential upgrade is estimated by means of intrusive speech transmission quality measurements in the test laboratory where demonstration or trial system installation of the considered new technology is available.

In simplified case when the dependence between ACR and MOS is approximately linear with sensitivity coefficient \( a \), the average increase of call durations will be

\[
CD_{inc} = aMOS_{inc} \quad (8)
\]

The resulting uncertainty of \( CD_{inc} \) is then given by sum of the uncertainty of \( a \) and that of \( MOS_{inc} \), evaluated by statistical analysis of intrusive measurement results. In case of strongly non-linear dependency, the results can be obviously evaluated more accurately e.g. by approximating the curve by several linear parts and by repeating the calculation (3) to (8) for each such part separately. The final result is obtained then as an average of calculated estimations of average call duration increments, weighted by the corresponding number of calls in the relevant MOS interval.

Special attention should be devoted to representative structure of the used QCDR database since it was observed that e.g. prepaid customers are generally less sensitive to speech quality than customers using regular tariff. Also the analyzed dependency differs for working days and weekend traffic or for busy hours and night traffic of any working day. The QCDR database should either contain all before mentioned cases in proper balance that corresponds to the average structure of calls in given network or each such category should be analyzed separately.
6. Conclusions

The problem of estimation of dependency between average call duration and speech transmission quality has been discussed. The straightforward polynomial fitting is useful in case when no detailed information about result uncertainty in required. The analytical expressions for sensitivity coefficient and its uncertainty for simplified linear case were shown after introducing two aggregation methods. An example of method application was presented, too. A practical example on limited training database is given. The first results indicate that in the low to medium quality range (Mean Opinion Scores between 1 and 3) the call duration increases with increasing speech quality. The method needs further verifications on statistically valid QCDR databases.

7. References


