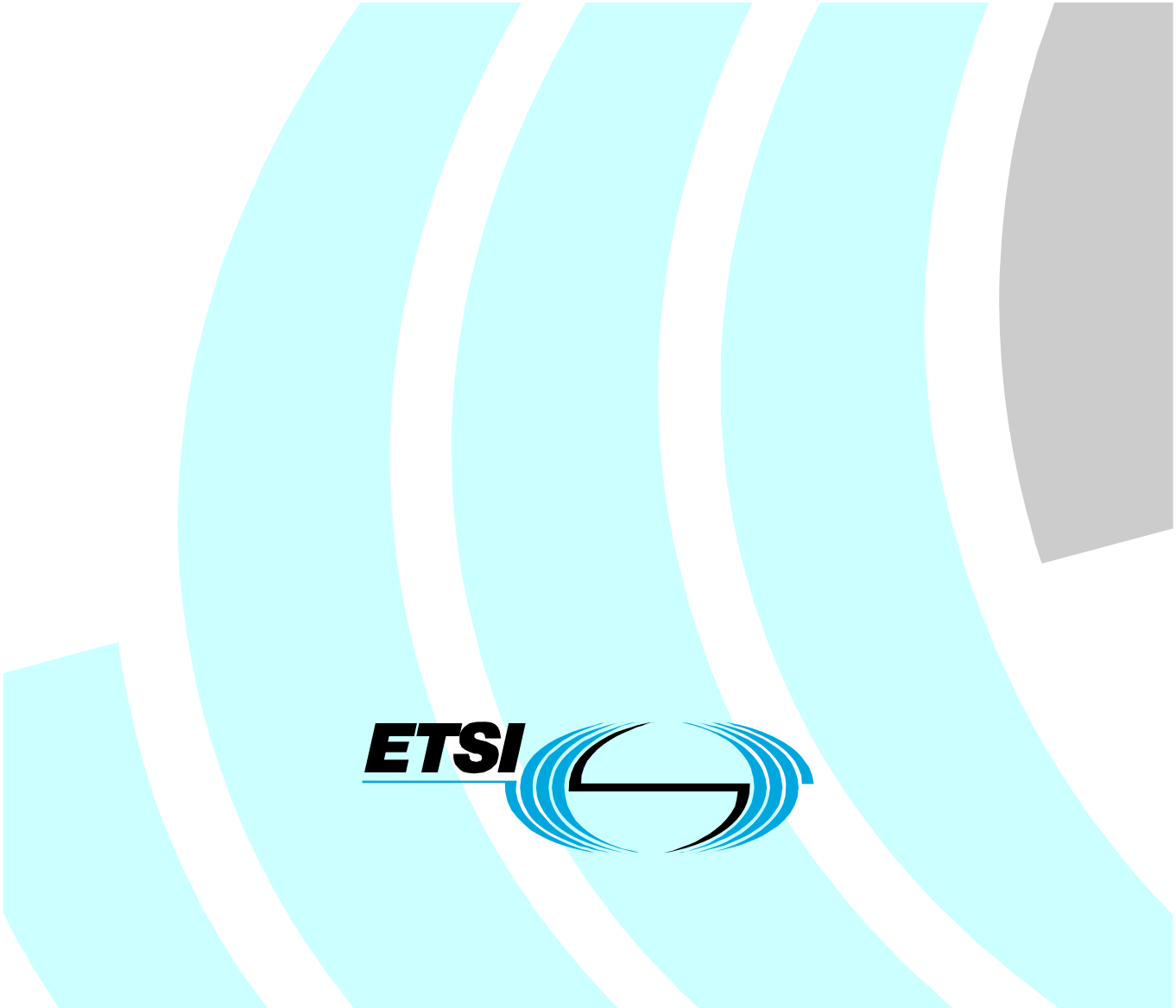
ETSI TR 102 506 V1.2.1 (2007-11)

Technical Report

Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call



Reference
RTR/STQ-000116m

Keywords
speech, quality

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://webapp.etsi.org/IPR/home.asp).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Speech Processing, Transmission and Quality Aspects (STQ).

1 Scope

The present document proposes a way to model measurement results on a per sample basis that allow to estimate the perceived end-to-end speech quality per call for narrowband circuit switched voice services in mobile networks.

The scenario is focussing on test signals between 60 seconds and 120 seconds in duration with alternating speech/silence periods as described in clause 5. The presented model is based on three studies but may not generalize to other call scenarios than those used in the underlying studies.

Throughout the present document where ITU-T Recommendation P.862.1 [2] (or ITU-T Recommendation P.862 [1]) is quoted the same applies to all measurements of listening quality. This can be listening quality scores gained by auditory tests (MOS-LQS) or objective measurements predicting MOS-LQO according to ITU-T Recommendation P.800.1 [3] covering the relevant network distortions and speech processing components in their scope.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific.

- For a specific reference, subsequent revisions do not apply.
- Non-specific reference may be made only to a complete document or a part thereof and only in the following cases:
  - if it is accepted that it will be possible to use all future changes of the referenced document for the purposes of the referring document;
  - for informative references.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

For online referenced documents, information sufficient to identify and locate the source shall be provided. Preferably, the primary source of the referenced document should be cited, in order to ensure traceability. Furthermore, the reference should, as far as possible, remain valid for the expected life of the document. The reference shall include the method of access to the referenced document and the full network address, with the same punctuation and use of upper case and lower case letters.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

2.1 Informative references

[1] 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".

[2] ITU-T Recommendation P.862.1: "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[3] ITU-T Recommendation P.800.1: "Mean Opinion Score (MOS) terminology".

[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 Recommendation P.862.3: "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2".

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

[7] CENELEC EN 60645-2:1997: "Audiometers - Part 2: Equipment for speech audiometry".

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

listening quality: quality as perceived by user in a listening situation

perceived quality: quality as perceived by a human user

speech quality per call: listening quality as perceived by a user (at the end) of a conversational call

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

EFR	Enhanced Full Rate
FR	Full Rate
HR	Half Rate
MOS	Mean Opinion Score

NOTE: Commonly used term for quality assessment.

MOS-LQO	MOS-Listening Quality from Objective testing
MOS-LQS	MOS-Listening Quality from auditory tests (Subjective)
SpQ-C	Speech (listening) Quality on Call basis
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over IP

4 General

The established way of measuring the speech quality is the measurement on a per sample basis. Much standardization work has been done by the ITU-T with the P.862 series of documents. Using that established way and taking advantage of the data acquired in that fashion one can seek to estimate the perceived speech quality of a call.

Current models of averaging over a large amount of single speech samples do not necessarily paint an accurate picture of the customer satisfaction. Since a bad sample can be outweighed by a couple of good samples. Averaging over the calls mitigates the problem but still suffers from the shortcoming that a number of good samples may outweigh a very bad sample. On the other hand threshold models that regard a call fair or poor on the basis of one or two degraded samples do not take the number of good or excellent samples into account. Models where a certain percentage of the samples need to be degraded to rate the call as bad disregards the temporal structure of the call and the relative timing of the degradation towards the end.

It is worthwhile to model the measurement results to obtain a call quality value that allows understanding the impact of varying speech quality during a conversation.

NOTE: The present document focuses on speech (listening) quality of a voice call. Conversational properties such as talker quality, round trip and other related metrics are not considered. Speech Quality of video telephony is not considered either.

5 Call properties

For the determination of the call properties like call length and the samples specifics it can be drawn on existing specification like ITU-T Recommendation P.862 [1] and TS 102 250 [4]. On that basis a reference speech quality sensitive voice call can be characterized. The standard call length for instrumental voice quality testing is defined in TS 102 250-5 [4] and the sample characteristics and evaluation is defined in ITU-T Recommendation P.862 [1] and ITU-T Recommendation P.862.3 [5]. For the structure of the call the definition needs to be done.

5.1 Call structure

Calls, be they mobile originated, mobile terminated or mobile to mobile can be divided up into different groups. Short calls of a couple of seconds where there is an announcement like pre paid account statements or voice boxes or wrong destination and conversations where the parties exchange a couple of utterances. Assumed the listening quality sensitive calls are the group where meaningful utterances are exchanged over a stretch of time, voicemail and speed dials can be excluded from the consideration. The "typical" call is a dialog like conversation, which is in line with the empirical findings.

In an idealized dialog the utterances are exchanged and distributed evenly in length and frequency. On each side a certain period of speech activity is followed by silence for the same length of time. Since the call quality on sample basis is rated for each side independently it is sufficient in an instrumental or subjective realization to feed one side with the required sample pattern.

5.2 Call length

The length of the call must give room for a couple of utterances (samples). The call length recommended in TS 102 250-5 [4] is 120 seconds which is sufficient for this requirement. In fact the average call length is well below this time. However if calls like those to the mailbox, to pre-paid account, far end voice boxes or wrong numbers are excluded from that calculation the average time of calls goes up considerably. However for practical purposes it is desirable to use call lengths that are considerably shorter that 120 seconds. The studies in annexes B and C provide results for calls with a length of 60 seconds.

5.2.1 Length of utterance (sample)

The application guideline for objective speech measurement is ITU-T Recommendation P.862.3 [5]. The typical sample of measurement systems has a length from 5 seconds to 12 seconds with a speech activity of maximum 80 %. These individual samples and their ratings are the basis of the call quality assessment. Therefore the speech activity part of the call consists of these samples.

5.2.2 Number of utterances (samples)

Depending on the length of the call in connection with length of the individual utterance it takes from five to 12 utterances and silence pairs to fill the different call lengths. From empirical evidence we know that a typical conversational call contains around 4 utterances from each side so that 5 recurrences of the speech and silent pair can be recommended. Considering that these values are applicable for short calls, longer calls can accommodate up to 12 speech and silence pairs with an individual sample length of 5 seconds.

5.3 Call design

The conversational call that is to be rated to estimate the call quality should consist of alternating phases of speech activity and silence, the length of the phases should be 5 seconds to 12 seconds and that pair recurs 5 to 12 times during the call.

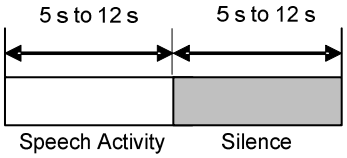


Figure 1: Structure of the speech activity silence pair

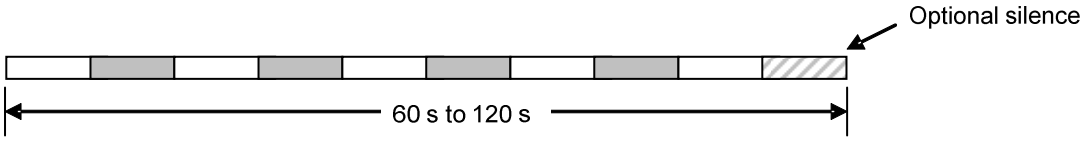


Figure 2: Structure of the call with 5 recurrences for one side

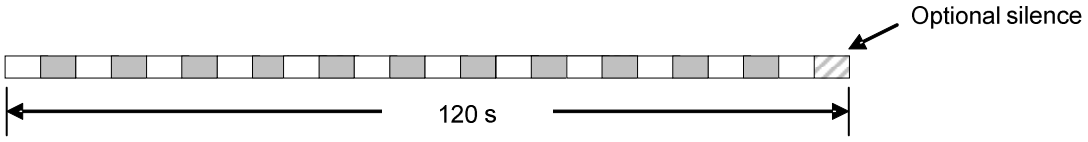


Figure 3: Structure of the call 12 recurrences

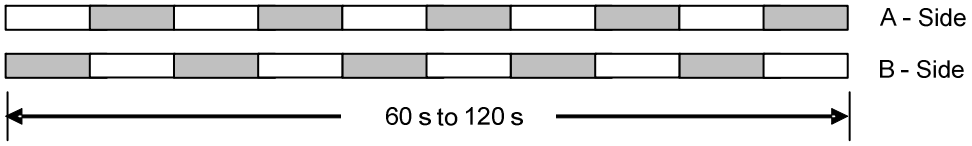


Figure 4: Structure of the call with 5 recurrences and alternating speech activity

6 Call quality on a per sample basis

In this clause a mathematical model is proposed with which the call quality of voice call can be estimated.

6.1 Evaluation of the samples

The evaluation of the individual samples is made by end-to-end speech quality measurements. They can be either evaluated by Listening Only tests acc. ITU-T Recommendation P.800 [6] or by objective prediction of those scores, e.g. by the current ITU-T Recommendation P.862.1 [2]. The use of objective prediction allows the application of the proposed model in automated network evaluation tools.

6.2 Mathematical modelling of the call quality

The desired result of the calculation is a MOS value considering the entire call in its structure. A mathematical model is necessary to aggregate the individual MOS values to one value. Two important effects are taken into account: the "recency effect" and the effect of a very bad sample in a call.

6.2.1 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 certain degradation 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 made with an individual parameter a_t at that is the weighing factor for each sample. A mathematical model here is:

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

If the time between the end of the last sample and the middle of sample i is t_i then we have for samples for t_i < 19 the following weighing factor (i is positive and needs to be between 5 and 12)

$$a_i = 1/2(19 - t_i)/19 + 1/2$$

For $t_i \geq 19$ the weighing factor is constant with a = 1/2 . This formula represents the increasing importance of a sample for the general impression the closer it is located towards the end.

6.2.2 Impact of the a single very bad sample

The correlation can be significantly improved by taking additionally into account the worst sample of the call. Empirical evidence shows that one very bad sample deteriorates the impression strongly in addition to its temporal occurrence; therefore it needs also to be taken into account. The model is extended to include the worst sample in the call.

$$MOS_{SpQ-C} = MOS_{RE} - 0{,}3\,(\overline{MOS} - \min(MOS_t))$$

6.2.3 Applicability of the mathematical model

The formula is developed for conversations with a length between 60 seconds and 120 seconds containing 5 to 12 utterances per analysed direction and with sample and pause lengths of 5 seconds to 12 seconds each.

6.2.4 Validation of the formula

The formula has 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 Call Quality in comparison with the linear averaging for all tested scenarios.

NOTE: The studies differ in the tests groups (e.g. few expert listeners in annex A and test material (different distortion patterns), therefore the range of correlations. See annexes for details).

	Study "Annex B" (English) 5 seconds samples		**Study "Annex C" (German) 5 seconds to 6 seconds samples**		**Study "Annex A" (German) 12 seconds samples**
Call length	120 seconds	60 seconds	120 seconds	60 seconds	120 seconds
Lin. Average with MOS-LQS (RMSE)	92 % (0,66)	88 % (0,63)	83 % (0,51)	85 % (0,49)	57 % (0,84)
CallQuality model with MOS-LQS (RMSE)	98 % (0,21)	97 % (0,22)	93 % (0,31)	94 % (0,26)	84 % (0,37)
CallQuality model with MOS-LQO P.862.1 (RMSE)	97 % (0,32)	96 % (0,33)	84 % (0,42)	89 % (0,35)	80 % (0,43)

7 Conclusion

The perceived speech quality is not a simple aggregation (average) of the rated samples in a call. The experimental evidence shows that the impact of a degraded speech is not simply outweigh by a longer stretch of good or acceptable listening quality. For single calls the temporal structure of the call must be considered. Lower listening quality towards the end of a call has a stronger impact on the overall rating of a call than degraded parts in the beginning.

With the presented formula in clause 6.2.2 it is possible to estimate the perceived (subjective) speech quality of a call for each side on the basis of (objectively or subjectively) rated samples.

Annex A:
Empirical Study on the perceived call quality: PESQ_mobil

In this annex an excerpt of the study "Ergebnisbericht (Study) Berkom, PESQ-mobil" (in German), J. Berger, T-Systems is presented. This study addresses a wider range than the evaluating a model for prediction of Speech Quality per Call. This annex is focused strongly on topics related to the present document.

A.1 Test concept and speech recordings

A.1.1 Test description of the overall project

In automatic measurement systems for speech quality evaluation in practical use, short speech samples (4 seconds to 8 seconds) are transferred over a telephone connection and evaluated with an algorithm. At the end of every call, several measured speech quality samples are available, which will be averaged usually. With these measured quality results, the assessment of a listening person being in a dialogue situation is emulated and thereby the overall quality of a telephone call is described. This overall quality of the complete call should be called Speech Quality per Call (SpQ-C).

Existing problems by this usage:

- The measurement cycle is shorter than an average real phone call.
- The measurement result is based on "speech samples", which are restricted because of their short length in variability and its phonetics.
- Due to different and time varying quality conditions of the connection during a measurement cycle, an average of these single speech samples is only for limited use for the prediction of the Speech Quality per Call.

The speech quality assessments a person gives after a phone call is highly stamped by the time of appearance of a possible distortion. This influence of the different quality states during a call on the overall result respects both the time difference of a quality state at the time of the assessment and the loss in means of semantics. It can be assumed, that a distortion at the beginning of the call is already forgotten at its end.

To evaluate this effect, a listening situation as natural as possible had been designed and test persons assessed the experienced listening quality. The task of this investigation lies in the modelling of the assessment of a longer conversation with varying listening quality. Therefore a conversation was modelled by a series of single "speech samples". The assessment of the complete modelled conversation at its end by human listeners forms the reference of the model. These "target values" for the Speech Quality per Call are to be emulated by an weighted average of short term scores as they could be derived by an instrumental measurement method as well. Here it is assumed that this instrumental method is in the position to assess a static quality like a human listener.

The intention had been to find a mathematical description for the consecutive speech parts to be able to calculate an overall quality score, which emulates the assessments of the test persons. The method to develop a model for prediction of Speech Quality per Call by means of instrumental measures can be divided into three steps:

1) Modelling and assessing of simulated conversations in a subjective test (gaining the "target values").
2) Assessing short parts of the conversation (single samples, "per sample scores") subjectively and developing a model to predict the "target values" by processing that single scores.
3) Replacing the subjective per-sample scores by instrumental gained scores in model obtained in 2). Here ITU-T Recommendation P.862 [1] was used.

The listening test used samples, which had been designed in a way, that they should partly cover the awaited distortions in UMTS or VoIP. Particularly the distortion with longer duration and the accumulate appearance of short distortions are of central interest.

A.2 Design of an auditory test methodology to assess the speech material

A.2.1 Structure of the quality assessment

The quality assessment with auditory tests with test persons is separated into two parts:

1) Simulation of a conversation.

2) Assessment of shorter conversation parts without personal activity.

For this study eight employees from T-Systems Nova, Berkom had been invited. Before the test started, all persons had been tested for normal hearing. The 6 men and 2 women were of the age between 21 years and 45 years and German native speakers. All invited employees had been working in the quality and acceptance department. This means that the test environment had been well known and they had no problems with their tasks and the way they had to give their assessments. None of them had taken part in the development of this test.

A.2.2 Simulation of a conversation

A typical speech situation is a dialogue between two persons, thus the situation is divided in parts with hearing activity and speech activity. The interest for the content of both persons is supposed. For that reason typical contents of telephone conversations were chosen (e.g. request for a rental car).

The realization of such a modelled conversation consists of a series of shorter "utterances", which have a pause between them for interaction but are connected logically with regard to the content of the presentation. Instead of the own speech activity as an interaction a content orientated task is to be done (e.g. keyword spotting). The speech material is constructed in such a way that 4 breaks are possible. After each replay of the whole simulated conversation the test person is asked for an assessment for the complete simulated call.

Experiment 1 equals the automatically test methodology half duplex. The used speech material consists of 5 speech parts (samples), which correspond to the utterances of one party. The design of the speech material is shown in figure A.1. After a 12 seconds speech sample there is a 12 seconds pause during which the test person had to perform a content regarding task. At the end of this experiment a score for the Speech Quality per Call is obtained.

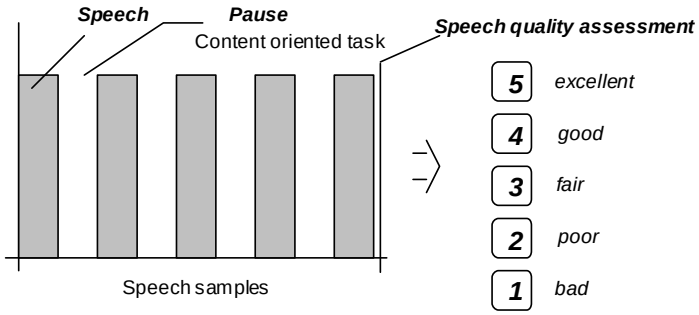


Figure A.1: Schematic presentation of the speech situation assessment

A.2.3 Assessment on an individual per-sample basis

In the second experiment the test persons listen to the small conversational parts (samples, 12 seconds in length) which were replayed in a casual sequence. This means that the different parts will be individually presented and assessed. This scenario corresponds to an automatic test situation with only uplink or downlink speech samples of a short length. This matches a simplified test according to ITU-T Recommendation P.800 [6] with short speech samples. (At the end of this test "per-sample" scores for each individual part of the simulated conversations, as average for each sample, were available).

A.2.4 Distortion types for the voice transmission

The focus of this research is on the influence of the time variable transmission faults on the perceived speech quality at the end of the call. It is assumed that difference of the time of the distortion to the time of the assessment and its intensity and length have the strongest influence. Based on this, distortion patterns are designed which will be shown in the following figures. Each pattern consists of five speech samples and reflects the temporal structure of the simulated conversation. A difference was made between distortions perceptible over the complete sample (such as vocoders) and "bursty" distortions such as interruptions.

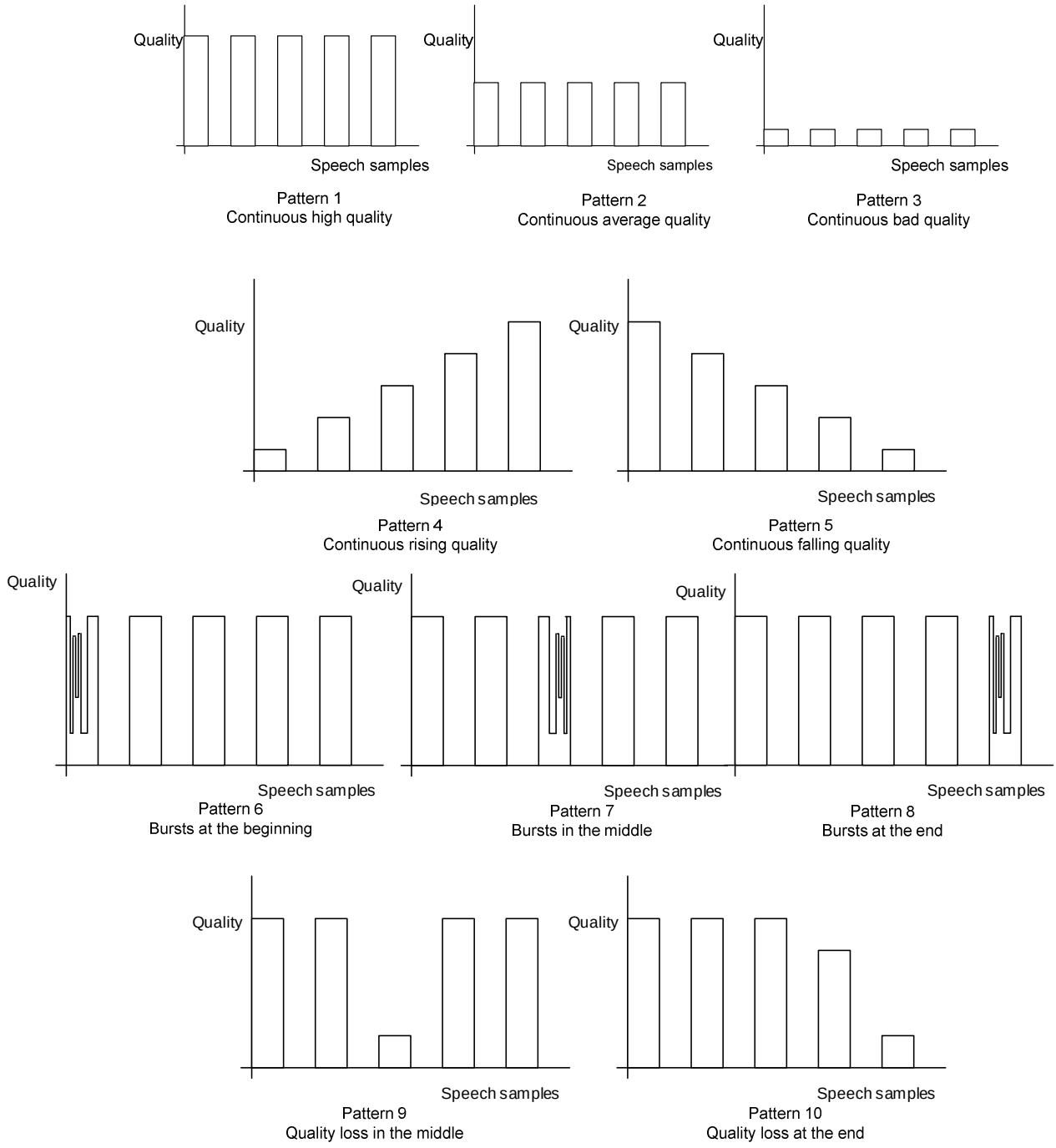


Figure A.2: The temporal structure of the ten quality pattern

All these examples are spoken by two different speakers and have different content.

A.2.5 Structure of the speech material

Four modelled conversation-examples with a longer period consisting of a series of five individual parts (speech samples) that model a real telephone situation two of them were actually used in the investigation.

Of interest in this evaluation is the influence of the time distance between the occurrence of the distortion and the end of the transmission ("recency-effect") on the overall quality scored at the end. Experiments have shown that the gradient of the influence is decreasing with the time distance of the assessment. Later the influence nears zero (the influence of the distortion is constant). The band of 50 seconds to 90 seconds before transmission end is seen for interest in this evaluation. This means that the simulated dialogues should have at least this length.

The speech samples used for this auditory evaluation should be like a natural telephone situation, e.g. renting a car. They are structured in the way that they are constructed out of 5 samples of 12 seconds to 13 seconds with active speech. After each 12 seconds part, a pause of 12 seconds length is implemented. This results in an overall transmission length of 110 seconds.

Speech activity

The speech material used in this evaluation is small parts of a conversation called speech samples. This means one person is speaking, the other one is listening. The term *Text* describes the content (e.g. car rental), the subparts 1.1, 1.2, etc. describe the individual phrases in this context. A phrase, spoken by a speaker, forms a 12 seconds speech sample. The speech activity of the simulated dialogues is shown in table A.1.

Table A.1: Activity of speech samples

Speech part	Speech activity	Speech part	Speech activity
Text 1.1: female 2, Sample 1	88 %	Text 2.1: male 2, Sample 1	94 %
Text 1.2: female 2, Sample 2	96 %	Text 2.2: male 2, Sample 2	90 %
Text 1.3: female 2, Sample 3	93 %	Text 2.3: male 2, Sample 3	92 %
Text 1.4: female 2, Sample 4	85 %	Text 2.4: male 2, Sample 4	94 %
Text 1.5: female 2, Sample 5	92 %	Text 2.5: male 2, Sample 5	93 %

Together with the implemented pauses an overall speech activity of about 50 % is reached.

A.2.6 Quality of the speech material

The speech material is transmitted over test calls in a live network. One time the material is transmitted over a transmission with the best possible available call quality to achieve the best speech quality in a real network. Then the connection is influenced to reduce the speech quality. The material is degraded in a way that it covers all necessary quality states for this test. This test requires the whole range from excellent/good to bad.

A.2.7 Results

In the first part of the test, the test persons listened to the simulated dialogues (all 10 fault patterns of every speaker (see clause A.2.6)) one time. An average over 8 individual assessments is the result. The scores for the overall (per call) quality obtained in the auditory experiment are shown in figure A.3.

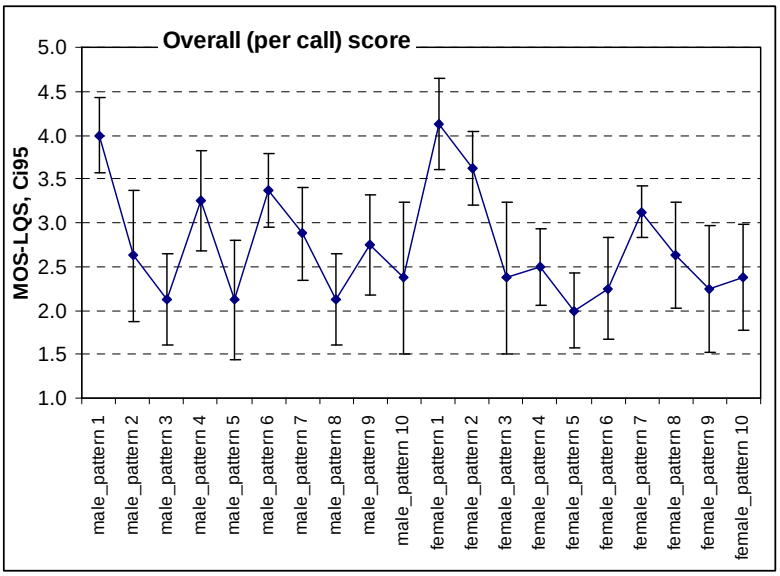


Figure A.3: Auditory MOS "per call" per pattern

In a second part the test persons listened to the separated speech samples twice during the test. This means that the MOS value represent the average of 16 individual assessments. Here the Ci95 is smaller due to the higher number of individual results (16) and a smaller inter-individual deviation in the scores.

Because of the two separated tests an integrative overall quality assessment and also an individual speech part assessment exist.

A.3 Modelling the overall quality mathematically on basis of the MOS-values

A.3.1 Modelling of Speech Quality by averaging per-sample scores

Figure A.4 shows the simple arithmetical average of the auditory MOS assessment of the individual speech samples to the overall quality assessment (Speech Quality per Call). It can easily be seen that a pure average will not be applicable for predicting the Speech Quality per Call.

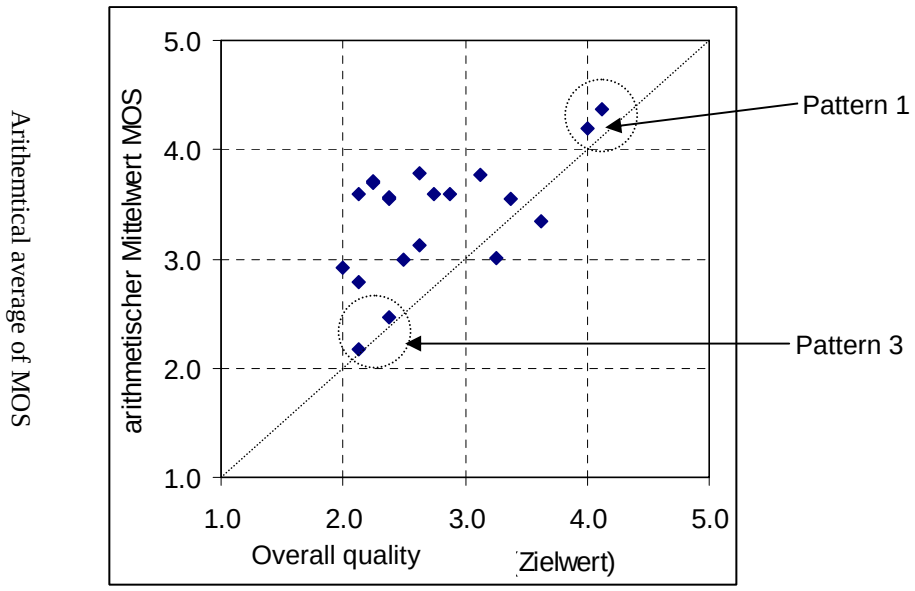


Figure A.4: Arithmetical average of the MOS assessment of the individual speech parts to the overall quality assessment

Only in the case of static quality over the complete "call" modelled by patterns 1, 2 and 3 in clause A.2.4 the simple averaging gives reliable results. For varying quality, the arithmetical average seems to be too optimistic for the prediction of Speech Quality per Call.

The linear correlation coefficient is about 57 %. This leads to the result, that the arithmetical average should not be used for describing the Speech Quality per Call.

In the scenarios in which a quality drop within one speech part occurs, the overall quality is below the average. *A possible reason could be that the overall assessment is disproportionately influenced by a strong quality drop in a longer speech presentation. This degradation is the stronger the higher the presented quality is.* This influence occurs at first independent of the time within the 1 minute 40 seconds dialogue. In tendency it can be said, that the influence is stronger the later the distortion occurs. This conclusion corresponds with the statements of the test persons.

A.3.2 Modelling of Speech Quality by consideration of the "recency effect"

First the recency effect (model 1) will be modelled. With the weighting of the individual speech samples in the modelled conversation (figure A.5) good results can be achieved:

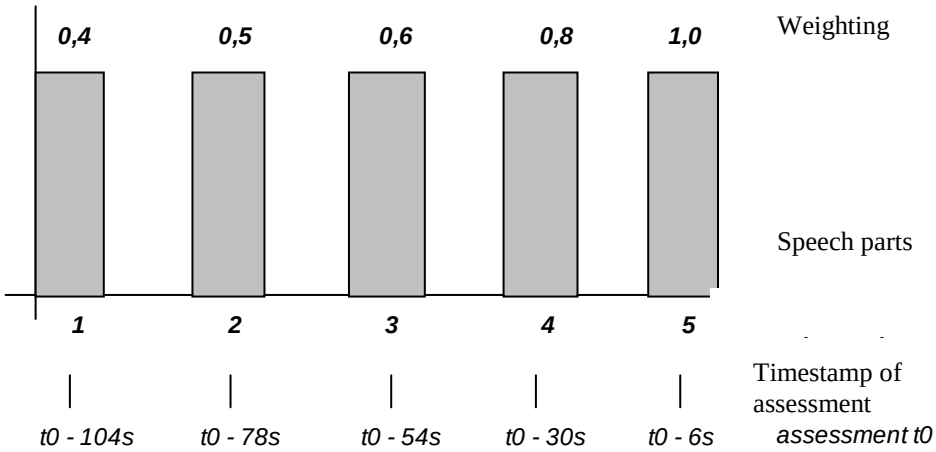


Figure A.5: Weighting of the individual speech parts

The weighting coefficients have been designed in a simple way, because with a more precise model there is the risk of "over adaptation" to this single experiment due to the limited amount of data.

With these coefficients the weighted average results in a correlation of about 65 %.

The weighted average MOS_{Mod1} can be calculated as follows:

$$\overline{MOS}_{Mod1} = \sum_{t=1}^{5}(a_t MOS_t) \Big/ \sum_{t=1}^{5} a_t$$ t: speech parts, a: weighting coefficient

A.3.3 Modelling of Speech Quality with consideration of a bad sample

In a further step (model 2) the over proportional high degraded speech samples will be considered too. Therefore the difference of the average of all five individual speech parts to the lowest of one speech part result is used. This difference will then be weighted and directly subtracted from the result of the first model.

$$MOS_{SpQ-C} = MOS_{RE} - \frac{2}{5}(\overline{MOS} - \min(MOS_t))$$

This step is the more important one compared to the already modelled "recency effect". It reflects the non-linear averaging of perceived quality by a user. It corresponds with the hypothesis that a degradation (burst) is a topic of interest rather than good quality (which is assumed as normal). Thus, this topic of interest will dominate the quality assessment at the end.

Applying this model, shown as filled squares in figure A.6, the correlation is about 85 %.

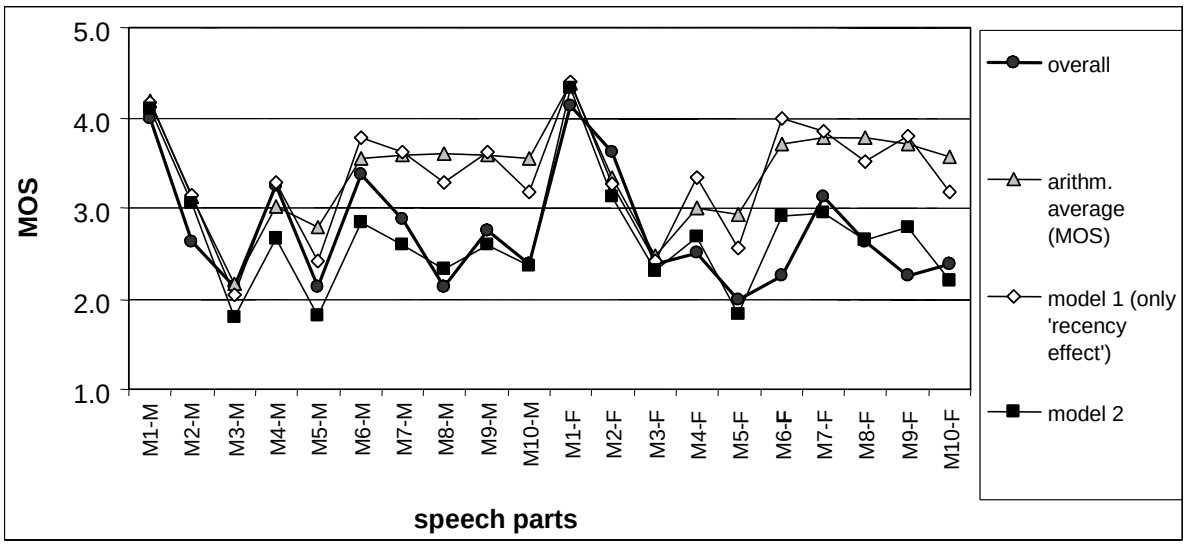


Figure A.6: Individual results by using model 2

It can be concluded that the Speech Quality per Call can be predicted reliably by a two step model as shown in clauses A.3.2 and A.3.3.

A.4 Assessment of the speech material by ITU-T Recommendation P.862

In this clause it is shown how accurate the per-sample subjective scores are when predicted by instrumental measures and whether they can be used to predict Speech Quality per Call by applying the model developed in clause A.3.

To assess the signal of interest by ITU-T Recommendation P.862 [1], it will be compared with a non-influenced reference signal. A non-filtered and non band limited (130 Hz to 3 500 Hz) signal had been used as reference signal. This equals a "flat source" in Literature as source signal.

A.4.1 Assessment of the separated speech parts

The individual speech samples of the modelled conversations were assessed by means of ITU-T Recommendation P.862 [1]. The speech signals used are identical to those one scored subjectively in clause A.2.3.

At first, all separated speech parts had been assessed by ITU-T Recommendation P.862 [1], 27 for male and 26 for female speakers. All of these speech parts have a length of 12 seconds.

In figure A.7, the ITU-T Recommendation P.862 [1] results are compared to the MOS values of the listening test. On the x-axis, the MOS values, on the y-axis the ITU-T Recommendation P.862 [1] results are displayed. The achieved correlation between the ITU-T Recommendation P.862 [1] results and the MOS values is 97,5 %. However because of the small amount of MOS-values from the listening test, this comparison should be treated carefully.

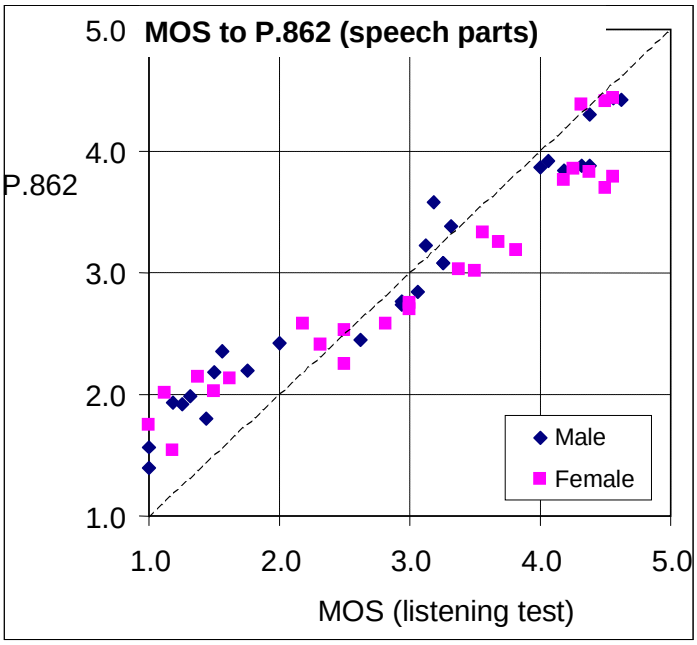


Figure A.7: ITU-T Recommendation P.862 [1] results in comparison to the separated speech parts

The more optimistic assessment of ITU-T Recommendation P.862 [1] for lower values is remarkable. This effect has already been seen in previous research. The reproduction of the ranking worked well.

To be able to compare the ITU-T Recommendation P.862 [1] values with the values from the auditory test, an easy linear transformation is designed:

$$P.862 = 1{,}04 + 1{,}34 \times P.862$$

This function stretches the width of the ITU-T Recommendation P.862 [1] results. In figure A.8 the results are shown.

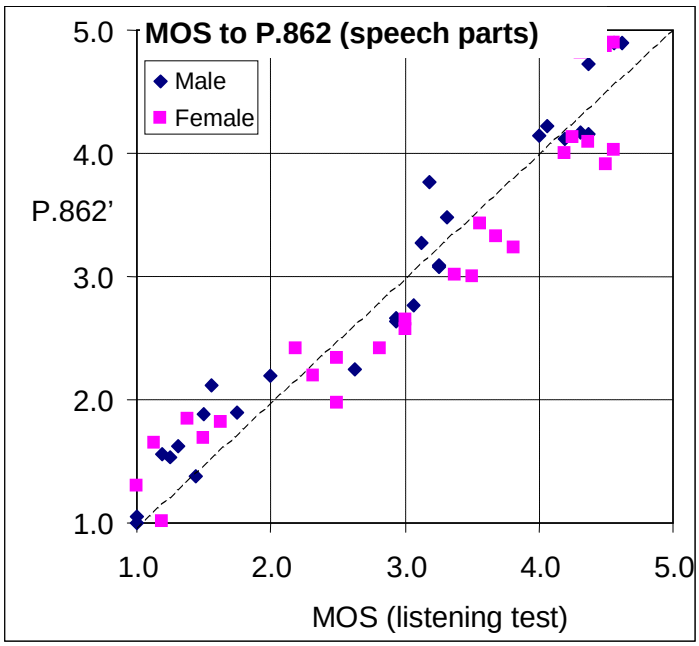


Figure A.8: ITU-T Recommendation P.862 [1] results in comparison to the separated speech parts with the scaling function

The transformed ITU-T Recommendation P.862 [1] values will be used for result evaluation.

A.4.2 Result presentation

Generally it can be said, that ITU-T Recommendation P.862 [1] has no problems in assessing the used individual speech samples in the right way. There are no outliers. The correlation of the ITU-T Recommendation P.862 [1] results with the results of the listening test is very high.

Consequently, the usage of the ITU-T Recommendation P.862 [1] results gained by evaluation of the individual samples will lead to the same results as the use of auditory score samples.

The deviation is shown as a scatter plot below. The resulting pattern is similar to the one in figure A.4 using the auditory MOS. For non-varying quality patterns the prediction shows fewer differences to the target values than for the conditions with varying quality during the call.

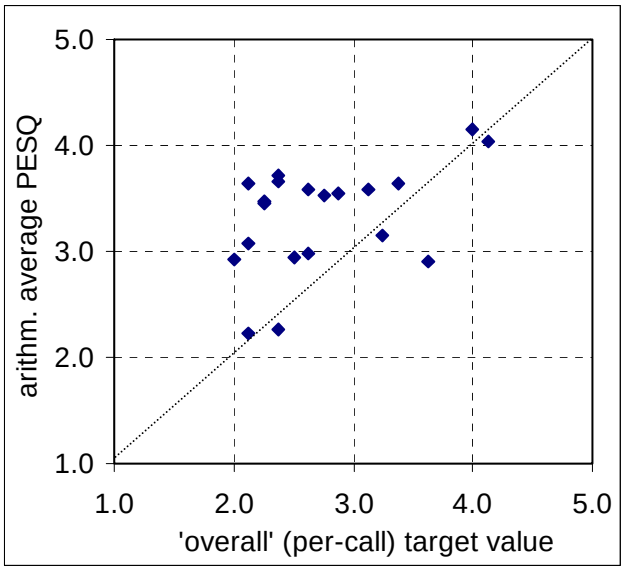


Figure A.9: Arithmetical average of the ITU-T Recommendation P.862 [1] assessment of the individual speech samples to the overall quality assessment

The comparison shows that ITU-T Recommendation P.862 [1] is like the auditory MOS not able to predict the per-call quality by simply averaging the per-sample results.

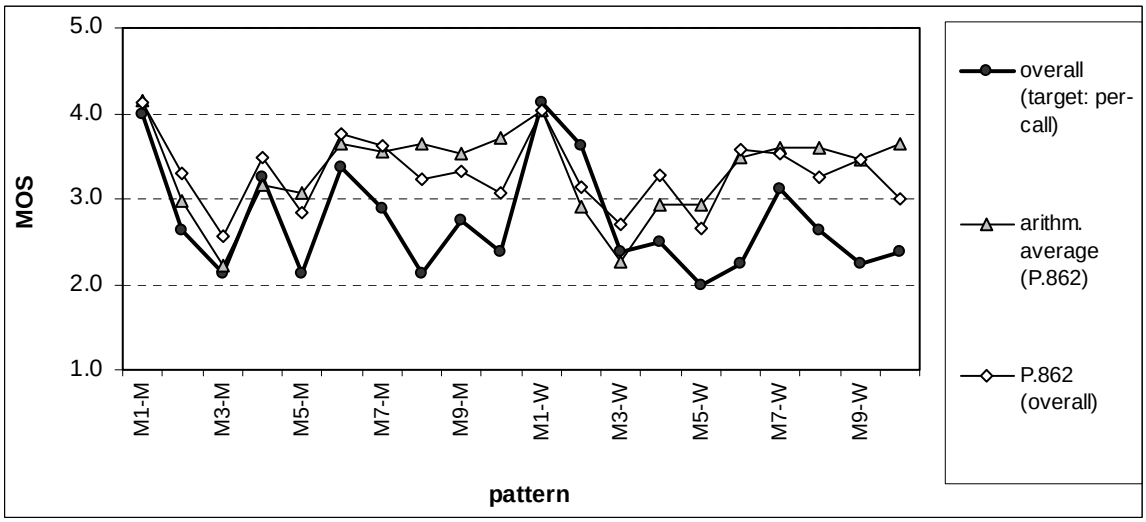


Figure A.10: Individual results by using ITU-T Recommendation P.862 [1]

A.4.3 Usage of the model with the ITU-T Recommendation P.862 results

Since ITU-T Recommendation P.862 [1] is trained to predict listening quality for speech samples between 5 seconds and 16 seconds in length, it can be expected that the results from the auditory tests and the results from the algorithm show the same distribution (see figure A.7). Consequently, it can be assumed that the individual results can also be used as an input for the per-call quality model developed in clause A.3.

In this last step the model will be applied on the transformed ITU-T Recommendation P.862 [1] per-sample results. The procedure is the same as in the previous clause; only transformed ITU-T Recommendation P.862 [1] results instead of the MOS values will be used. In figure A.11 the results are shown.

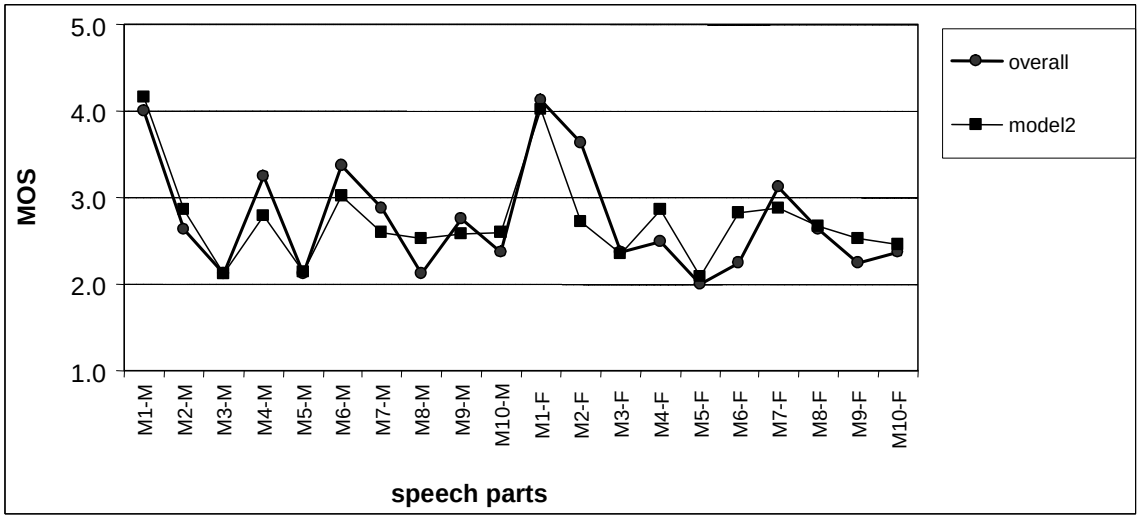


Figure A.11: Individual results by using model 2 with the ITU-T Recommendation P.862 [1] results

The correlation between using model 2 with the ITU-T Recommendation P.862 [1] results and the overall MOS results is about 85 %. It can be concluded that ITU-T Recommendation P.862 [1] (as a predictor of MOS-LQS) can be used in the introduced model for the prediction of Speech Quality per Call without considerable impact compared to the usage of subjectively scored speech samples. It has to be noted that the mentioned accuracy of ITU-T Recommendation P.862 [1] depends on the distortion types in the network. It has to be guaranteed that they are covered by the scope of ITU-T Recommendation P.862 [1]. Otherwise ITU-T Recommendation P.862 [1] must not be applied in such a context. Within this empirical study GSM-FR, GSM-EFR and GSM-HR speech codecs were used, which are covered by the scope of ITU-T Recommendation P.862 [1].

A.5 The rating of the samples

A.5.1 Rating of the calls

Table A.2

Pattern	Votes	Structure					MOS (Call)	StDev (Call)
		Utterance A	Utterance B	Utterance C	Utterance D	Utterance E		
Pattern 1 - male	8	M_A_hm	M_B_hm	M_C_hm	M_D_hm	M_E_hm	**4,00**	*0,53*
Pattern 2 - male	8	M_A_mm	M_B_mm	M_C_mm	M_D_mm	M_E_mm	**2,63**	*0,92*
Pattern 3 - male	8	M_A_ml	M_B_ml	M_C_ml	M_D_ml	M_E_ml	**2,13**	*0,64*
Pattern 4 - male	8	M_A_ll	M_B_ml	M_C_mm	M_D_fr	M_E_hh	**3,25**	*0,71*
Pattern 5 - male	8	M_A_hh	M_B_fr	M_C_mm	M_D_ml	M_E_ll	**2,13**	*0,83*
Pattern 6 - male	8	M_A_b2	M_B_hm	M_C_hm	M_D_hm	M_E_hm	**3,38**	*0,52*
Pattern 7 - male	8	M_A_hm	M_B_hm	M_C_b3	M_D_hm	M_E_hm	**2,88**	*0,64*
Pattern 8 - male	8	M_A_hm	M_B_hm	M_C_hm	M_D_hm	M_E_b2	**2,13**	*0,64*
Pattern 9 - male	8	M_A_hm	M_B_hm	M_C_ll	M_D_hm	M_E_hm	**2,75**	*0,71*
Pattern 10 - male	8	M_A_hh	M_B_hh	M_C_hm	M_D_mm	M_E_ll	**2,38**	*1,06*
Pattern 1 - female	8	F_A_hm	F_B_hm	F_C_hm	F_D_hm	F_E_hm	**4,13**	*0,64*
Pattern 2 - female	8	F_A_mm	F_B_mm	F_C_mm	F_D_mm	F_E_mm	**3,63**	*0,52*
Pattern 3 - female	8	F_A_ml	F_B_ml	F_C_ml	F_D_ml	F_E_ml	**2,38**	*1,06*
Pattern 4 - female	8	F_A_ll	F_B_ml	F_C_mm	F_D_fr	F_E_hh	**2,50**	*0,53*
Pattern 5 - female	8	F_A_hh	F_B_fr	F_C_mm	F_D_ml	F_E_ll	**2,00**	*0,53*
Pattern 6 - female	8	F_A_b1	F_B_hm	F_C_hm	F_D_hm	F_E_hm	**2,25**	*0,71*
Pattern 7 - female	8	F_A_hm	F_B_hm	F_C_b1	F_D_hm	F_E_hm	**3,13**	*0,35*
Pattern 8 - female	8	F_A_hm	F_B_hm	F_C_hm	F_D_hm	F_E_b1	**2,63**	*0,74*
Pattern 9 - female	8	F_A_hm	F_B_hm	F_C_ll	F_D_hm	F_E_hm	**2,25**	*0,89*
Pattern 10 - female	8	F_A_hh	F_B_hh	F_C_hm	F_D_mm	F_E_ll	**2,38**	*0,74*

A.5.2 Rating of the utterances

Table A.3

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862 (Utt)
M_A_hm	16	**4,38**	*0,72*	3,88
M_B_hm	16	**4,31**	*0,70*	3,88
M_C_hm	16	**4,00**	*0,63*	3,87
M_D_hm	16	**4,06**	*0,57*	3,93
M_E_hm	16	**4,19**	*0,66*	3,85
M_A_mm	16	**3,25**	*0,58*	3,07
M_B_mm	16	**2,94**	*0,57*	2,76
M_E_mm	16	**3,25**	*0,77*	3,08
M_A_ml	16	**2,63**	*0,72*	2,45
M_C_ml	16	**2,00**	*1,03*	2,42
M_E_ml	16	**1,75**	*0,58*	2,20
M_A_ll	16	**1,44**	*0,51*	1,80
M_B_ml	16	**2,94**	*0,44*	2,74
M_C_mm	16	**3,13**	*0,72*	3,22
M_D_fr	16	**3,19**	*0,75*	3,58
M_E_hh	16	**4,38**	*0,72*	4,30
M_A_hh	16	**4,63**	*0,50*	4,43
M_B_fr	16	**3,31**	*0,60*	3,38
M_D_ml	16	**1,56**	*0,63*	2,36
M_E_ll	16	**1,31**	*0,48*	1,99
M_A_b2	16	**1,19**	*0,40*	1,94
M_C_b3	16	**1,00**	*0,00*	1,56
M_E_b2	16	**1,25**	*0,45*	1,92
M_C_ll	16	**1,00**	*0,00*	1,39
M_B_hh	16	**4,56**	*0,63*	4,43
M_D_mm	16	**3,06**	*0,77*	2,84
M_E_ll	16	**1,50**	*0,52*	2,18

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862 (Utt)
F_A_hm	16	**4,38**	*0,62*	3,83
F_B_hm	16	**4,19**	*0,75*	3,76
F_C_hm	16	**4,50**	*0,63*	3,70
F_D_hm	16	**4,25**	*0,77*	3,86
F_E_hm	16	**4,56**	*0,51*	3,78
F_A_mm	16	**3,81**	*0,66*	3,19
F_B_mm	16	**3,50**	*0,63*	3,02
F_E_mm	16	**3,00**	*0,73*	2,76
F_A_ml	16	**2,50**	*0,63*	2,52
F_C_ml	16	**2,81**	*0,66*	2,58
F_E_ml	16	**2,19**	*0,66*	2,58
F_A_ll	16	**1,38**	*0,62*	2,15
F_B_ml	16	**2,50**	*0,52*	2,25
F_C_mm	16	**3,00**	*0,89*	2,69
F_D_fr	16	**3,56**	*0,73*	3,33
F_E_hh	16	**4,56**	*0,51*	4,43
F_A_hh	16	**4,50**	*0,63*	4,41
F_B_fr	16	**3,69**	*0,79*	3,26
F_D_ml	16	**2,31**	*0,70*	2,41
F_E_ll	16	**1,13**	*0,34*	2,01
F_A_b1	16	**1,00**	*0,00*	1,75
F_C_b1	16	**1,50**	*0,63*	2,03
F_E_b1	16	**1,63**	*0,50*	2,14
F_C_ll	16	**1,19**	*0,40*	1,53
F_B_hh	16	**4,31**	*0,70*	4,38
F_D_mm	16	**3,38**	*0,81*	3,03

Legend of the naming convention

*_hh:	high quality	e.g. transparent transmission
*_hm:	high-med quality	e.g. EFR without channel distortions
*_mm:	med quality	e.g. EFR in non-optimal conditions
*_ml:	med-low quality	e.g. EFR / FR in bad channel conditions but no muting
*_ll:	low quality	e.g. EFR / FR in very bad channel conditions but no muting
*_fr:	FullRate	e.g. FR without channel distortions
*_b?:	Bursts	e.g. EFR / FR with bursty mutings

Annex B:
Empirical Study on the perceived call quality with English samples (Ericsson AB, 2007)

B.1 Introduction

ETSI STQ Mobile develops an objective model for measuring speech quality per call. The model is developed in a joint project, where Ericsson participates. Subjective test results are used to build the model. Ericsson has done subjective tests with English speech material and the results are presented in the present document.

B.2 Test design

We recorded the speech material with native English speakers - two male and two female speakers. We recorded one side of a conversation and we used four different scenarios. The recorded speech files were then coded with AMR speech codec and degraded with a GSM AMR simulator.

The speech files were tested in a subjective test. All test persons were native English speakers. The test was divided into three parts:

1) 120 seconds calls.

2) 60 seconds calls.

3) 5 seconds utterances.

About half of the test persons listened on the 60 seconds conversations first and the other half listened on the 120 seconds conversions first. All test persons listened on the 5 seconds utterances last.

The test persons listened on one side of a conversation for the 60 seconds and 120 seconds calls. The longer call consisted of 12 utterances and the short calls consisted of 6 utterances. The utterances were of quality level ranging from 1 to 5. The quality 5 was the best and the quality 1 worst. The profiles for the calls are described in Appendix I below. The test persons answered a question after each utterance and marked the answer on a paper. They did this to keep them concentrated on what they were listening to. After each call the test person scored the call using a standard ACR scale:

Quality of the speech:

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

The test person entered the score on a terminal with a number of buttons on.

Finally the test persons listened to and scored the individual utterances from the calls. They used the same standard ACR scale as for the 120 seconds and 60 seconds call test.

B.3 Test results

These are the results from the subjective tests. The speech files for the calls are named as <speaker>_<profile> and the speech files for the utterances are named <speaker>_<utterance_number>_<quality_level>.

For example: The 60 seconds file "f1_18" is recorded with female speaker 1 and with profile number 18. The utterance speech file "m2_u3_q4" is recorded with male speaker 2, is utterance 3 in the call and has quality level 4.

The other columns in the tables are:

- Votes: The number of test persons scoring the speech file.
- MOS: The subjective MOS from our tests.
- Std: Standard deviation.
- CI: 95 % confidence interval.
- CIL: Lower 95 % confidence interval level.
- CIU: Upper 95 % confidence interval level.
- P.862: Raw "PESQ score".
- P.862.1: Mapped "PESQ score".

B.3.1 Results for 60 seconds calls

Table B.1

File	Votes	MOS	Std	CI	CIL	CIU
f1_11	25	2,56	0,77	0,30	2,26	2,86
f1_12	26	1,96	0,92	0,35	1,61	2,31
f1_13	25	4,40	0,71	0,28	4,12	4,68
f1_14	26	3,12	0,86	0,33	2,78	3,45
f1_15	26	1,85	0,73	0,28	1,56	2,13
f1_16	26	3,38	0,85	0,33	3,06	3,71
f1_17	26	3,15	0,92	0,36	2,80	3,51
f1_18	26	2,81	0,80	0,31	2,50	3,12
f1_19	26	4,15	0,67	0,26	3,89	4,41
f1_20	26	4,15	0,73	0,28	3,87	4,44
m2_11	24	2,50	0,88	0,35	2,15	2,85
m2_12	26	2,19	0,69	0,27	1,93	2,46
m2_13	26	4,38	0,80	0,31	4,08	4,69
m2_14	25	3,04	0,84	0,33	2,71	3,37
m2_15	25	2,12	0,60	0,24	1,88	2,36
m2_16	26	3,81	0,75	0,29	3,52	4,10
m2_17	24	3,33	0,70	0,28	3,05	3,61
m2_18	26	2,88	0,91	0,35	2,54	3,23
m2_19	26	3,73	0,67	0,26	3,47	3,99
m2_20	26	3,96	0,72	0,28	3,68	4,24

B.3.2 Results for 120 seconds calls

Table B.2

File	Votes	MOS	Std	CI	CIL	CIU
f2_01	26	2,23	0,65	0,25	1,98	2,48
f2_03	26	4,31	0,74	0,28	4,02	4,59
f2_05	25	2,32	0,75	0,29	2,03	2,61
f2_07	25	3,28	0,74	0,29	2,99	3,57
f2_10	26	3,81	0,69	0,27	3,54	4,07
m1_02	26	1,88	0,65	0,25	1,63	2,14
m1_04	26	2,96	0,87	0,33	2,63	3,30
m1_06	26	3,92	0,84	0,32	3,60	4,25
m1_08	26	3,23	0,71	0,27	2,96	3,50
m1_09	25	3,92	0,64	0,25	3,67	4,17

B.3.3 Results for the utterances

Table B.3

File	Votes	MOS	Std	CI	CIL	CIU	P.862	P.862.1
f1_u1_q1	26	1,27	0,45	0,17	1,10	1,44	1,17	1,20
f1_u1_q3	26	3,62	0,90	0,35	3,27	3,96	3,18	3,09
f1_u1_q4	26	3,62	0,90	0,35	3,27	3,96	3,13	3,02
f1_u1_q5	26	4,58	0,58	0,22	4,35	4,80	3,96	4,11
f1_u2_q1	26	1,19	0,40	0,15	1,04	1,35	1,14	1,20
f1_u2_q2	26	2,50	0,71	0,27	2,23	2,77	2,43	2,05
f1_u2_q3	26	2,85	0,73	0,28	2,56	3,13	2,87	2,63
f1_u2_q4	26	3,46	0,86	0,33	3,13	3,79	3,07	2,93
f1_u2_q5	26	4,62	0,53	0,20	4,41	4,82	3,96	4,12
f1_u3_q3	26	2,77	0,82	0,31	2,46	3,08	2,49	2,12
f1_u3_q4	26	3,96	0,87	0,33	3,63	4,30	3,33	3,31
f1_u3_q5	26	4,46	0,65	0,25	4,21	4,71	4,01	4,16
f1_u4_q3	26	3,12	0,82	0,31	2,80	3,43	3,00	2,83
f1_u4_q4	26	3,65	0,80	0,31	3,35	3,96	3,07	2,92
f1_u4_q5	26	4,77	0,43	0,17	4,60	4,93	4,06	4,21
f1_u5_q1	26	1,15	0,37	0,14	1,01	1,30	1,28	1,24
f1_u5_q2	26	2,04	0,66	0,25	1,78	2,29	1,86	1,53
f1_u5_q3	26	2,96	0,77	0,30	2,66	3,26	3,02	2,84
f1_u5_q4	26	2,65	0,75	0,29	2,37	2,94	2,92	2,71
f1_u5_q5	26	4,42	0,64	0,25	4,18	4,67	4,01	4,16
f1_u6_q1	26	1,31	0,47	0,18	1,13	1,49	1,40	1,28
f1_u6_q3	26	3,42	0,90	0,35	3,08	3,77	3,10	2,97
f1_u6_q4	26	3,88	0,86	0,33	3,55	4,22	3,22	3,15
f1_u6_q5	26	4,42	0,86	0,33	4,09	4,75	4,06	4,21
f2_u1_q1	26	1,23	0,43	0,17	1,07	1,40	0,81	1,12
f2_u1_q4	26	3,50	0,71	0,27	3,23	3,77	2,65	2,32
f2_u1_q5	26	4,54	0,51	0,20	4,34	4,73	3,58	3,66
f2_u2_q1	26	1,27	0,45	0,17	1,10	1,44	1,05	1,17
f2_u2_q4	25	3,80	0,82	0,32	3,48	4,12	2,68	2,37
f2_u2_q5	26	4,38	0,64	0,24	4,14	4,63	3,68	3,79
f2_u3_q2	26	2,15	0,73	0,28	1,87	2,44	2,01	1,64
f2_u3_q4	26	3,23	0,65	0,25	2,98	3,48	3,14	3,03
f2_u3_q5	26	4,31	0,62	0,24	4,07	4,55	3,84	3,98
f2_u4_q1	26	1,15	0,37	0,14	1,01	1,30	1,17	1,21
f2_u4_q2	26	1,92	0,56	0,22	1,71	2,14	2,05	1,67
f2_u4_q5	26	4,38	0,70	0,27	4,12	4,65	3,75	3,87
f2_u5_q3	26	2,81	0,85	0,33	2,48	3,13	2,58	2,24
f2_u5_q4	26	3,12	0,71	0,27	2,84	3,39	3,01	2,83
f2_u5_q5	26	4,27	0,72	0,28	3,99	4,55	3,84	3,98
f2_u6_q3	26	3,38	0,64	0,24	3,14	3,63	2,96	2,77
f2_u6_q4	26	3,31	0,68	0,26	3,05	3,57	3,01	2,83
f2_u6_q5	26	4,50	0,76	0,29	4,21	4,79	3,87	4,02
f2_u7_q3	26	3,50	0,81	0,31	3,19	3,81	3,01	2,84
f2_u7_q4	26	3,65	0,94	0,36	3,29	4,01	3,09	2,95
f2_u7_q5	26	4,46	0,71	0,27	4,19	4,73	3,88	4,03
f2_u8_q3	26	2,58	0,58	0,22	2,35	2,80	2,84	2,59
f2_u8_q4	26	3,50	0,71	0,27	3,23	3,77	3,06	2,91
f2_u8_q5	26	4,46	0,51	0,20	4,27	4,66	3,93	4,08
f2_u9_q1	26	1,31	0,47	0,18	1,13	1,49	1,40	1,28
f2_u9_q4	26	2,96	0,77	0,30	2,66	3,26	2,72	2,42
f2_u9_q5	26	4,38	0,64	0,24	4,14	4,63	3,83	3,98
f2_u10_q4	26	3,62	0,85	0,33	3,29	3,94	3,03	2,86
f2_u10_q5	26	4,69	0,47	0,18	4,51	4,87	3,80	3,93
f2_u11_q4	26	3,27	0,96	0,37	2,90	3,64	2,97	2,78
f2_u11_q5	26	4,31	0,68	0,26	4,05	4,57	3,80	3,94
f2_u12_q3	26	3,00	0,85	0,33	2,67	3,33	2,65	2,32
f2_u12_q4	26	3,31	0,84	0,32	2,99	3,63	2,92	2,71
f2_u12_q5	26	4,62	0,57	0,22	4,40	4,83	3,84	3,98
m1_u1_q3	26	2,69	0,55	0,21	2,48	2,90	2,80	2,52
m1_u1_q5	26	4,31	0,62	0,24	4,07	4,55	4,05	4,21

File	Votes	MOS	Std	CI	CIL	CIU	P.862	P.862.1
m1_u2_q3	26	3,04	0,87	0,33	2,70	3,37	3,27	3,23
m1_u2_q5	26	4,19	0,80	0,31	3,88	4,50	4,09	4,24
m1_u3_q3	26	3,50	1,07	0,41	3,09	3,91	3,07	2,93
m1_u3_q4	26	3,19	0,80	0,31	2,88	3,50	3,00	2,82
m1_u3_q5	26	4,42	0,50	0,19	4,23	4,62	4,13	4,28
m1_u4_q3	26	3,27	0,72	0,28	2,99	3,55	2,88	2,65
m1_u4_q4	26	3,58	0,70	0,27	3,31	3,85	3,22	3,15
m1_u4_q5	26	4,46	0,65	0,25	4,21	4,71	4,10	4,25
m1_u5_q3	26	3,69	0,79	0,30	3,39	4,00	3,44	3,47
m1_u5_q5	26	4,54	0,58	0,22	4,31	4,76	4,17	4,31
m1_u6_q3	26	3,19	0,85	0,33	2,87	3,52	3,20	3,12
m1_u6_q5	26	4,50	0,65	0,25	4,25	4,75	4,10	4,25
m1_u7_q3	26	3,46	0,90	0,35	3,11	3,81	3,47	3,51
m1_u7_q5	26	4,50	0,58	0,22	4,28	4,72	4,14	4,28
m1_u8_q3	26	3,27	0,83	0,32	2,95	3,59	3,40	3,41
m1_u8_q5	26	4,38	0,71	0,27	4,11	4,66	4,05	4,20
m1_u9_q2	26	2,27	0,60	0,23	2,04	2,50	2,34	1,96
m1_u9_q3	26	3,08	0,56	0,22	2,86	3,29	3,29	3,25
m1_u9_q5	26	4,31	0,79	0,30	4,00	4,61	4,13	4,28
m1_u10_q2	26	2,00	0,57	0,22	1,78	2,22	2,37	1,99
m1_u10_q3	26	3,31	0,79	0,30	3,00	3,61	3,09	2,96
m1_u10_q5	26	4,19	0,74	0,29	3,91	4,48	4,11	4,25
m1_u11_q1	26	1,92	0,63	0,24	1,68	2,16	1,57	1,36
m1_u11_q3	26	3,42	0,81	0,31	3,11	3,73	3,14	3,03
m1_u11_q5	26	4,50	0,51	0,20	4,30	4,70	4,06	4,21
m1_u12_q1	26	1,42	0,58	0,22	1,20	1,65	1,61	1,38
m1_u12_q3	26	2,81	0,63	0,24	2,56	3,05	2,38	2,00
m1_u12_q5	26	4,46	0,65	0,25	4,21	4,71	4,15	4,30
m2_u1_q1	25	1,44	0,51	0,20	1,24	1,64	1,54	1,34
m2_u1_q3	26	2,88	0,71	0,27	2,61	3,16	3,12	3,00
m2_u1_q4	26	3,00	0,63	0,24	2,76	3,24	3,10	2,97
m2_u1_q5	26	4,38	0,64	0,24	4,14	4,63	4,13	4,28
m2_u2_q1	26	1,88	0,86	0,33	1,55	2,22	1,95	1,59
m2_u2_q2	26	1,96	0,66	0,25	1,71	2,22	2,20	1,80
m2_u2_q3	26	2,81	0,69	0,27	2,54	3,07	2,90	2,68
m2_u2_q4	26	3,42	0,76	0,29	3,13	3,71	3,37	3,37
m2_u2_q5	26	4,73	0,45	0,17	4,56	4,90	4,03	4,18
m2_u3_q3	26	3,65	0,69	0,27	3,39	3,92	3,21	3,13
m2_u3_q4	26	3,58	0,64	0,25	3,33	3,82	3,27	3,23
m2_u3_q5	26	4,62	0,57	0,22	4,40	4,83	4,10	4,25
m2_u4_q3	26	2,81	0,69	0,27	2,54	3,07	2,49	2,13
m2_u4_q4	26	3,38	0,94	0,36	3,02	3,75	3,22	3,15
m2_u4_q5	26	4,38	0,64	0,24	4,14	4,63	3,98	4,14
m2_u5_q1	26	1,42	0,50	0,19	1,23	1,62	1,53	1,34
m2_u5_q2	26	2,19	0,69	0,27	1,93	2,46	1,96	1,60
m2_u5_q3	26	2,27	0,72	0,28	1,99	2,55	2,73	2,44
m2_u5_q4	26	3,31	0,74	0,28	3,02	3,59	3,29	3,25
m2_u5_q5	26	4,04	1,08	0,41	3,62	4,45	4,14	4,29
m2_u6_q1	26	1,46	0,58	0,22	1,24	1,69	1,57	1,36
m2_u6_q3	26	2,96	0,82	0,32	2,64	3,28	2,78	2,50
m2_u6_q4	26	3,69	0,84	0,32	3,37	4,01	3,16	3,06
m2_u6_q5	26	4,62	0,50	0,19	4,42	4,81	3,98	4,13

B.3.4 Correlation Between MOS and P.862.1 for the individual utterances

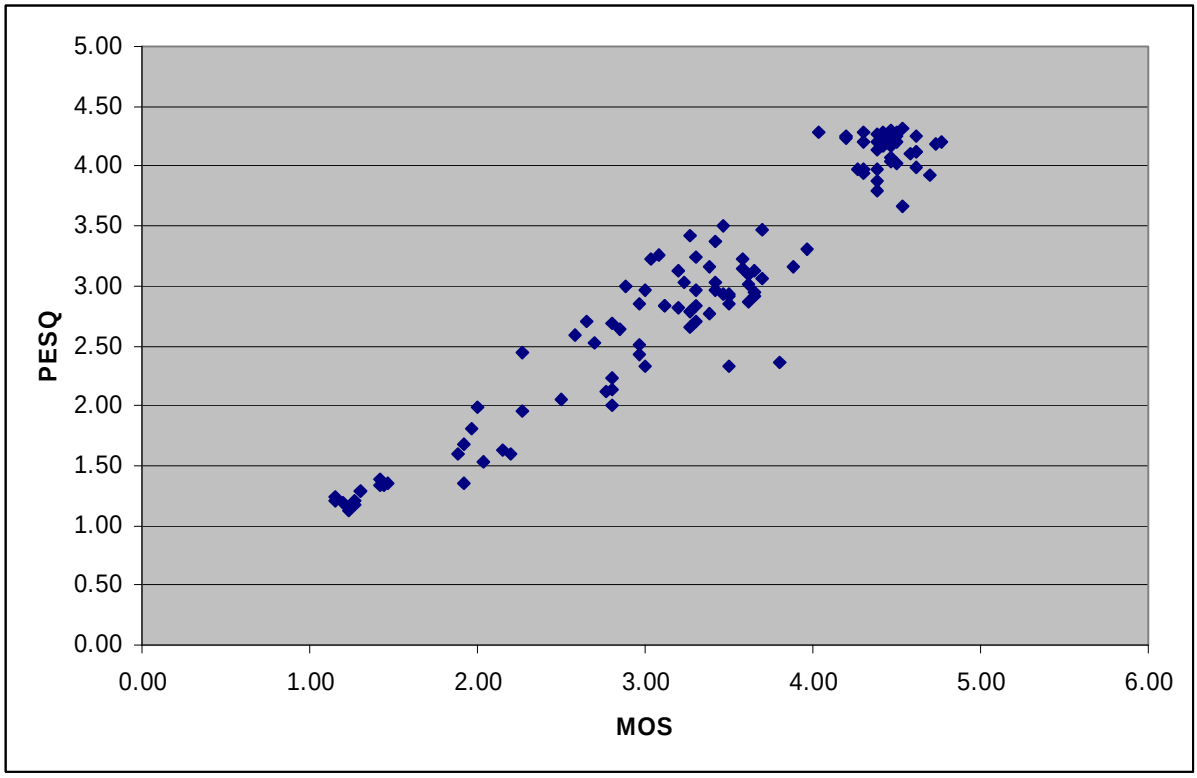


Figure B.1

B.4 Call profiles

The test profiles for the call quality test are described in the tables below. Profile 1 to profile 10 describe the 120 seconds calls and profile 11 to 20 the 60 seconds calls.

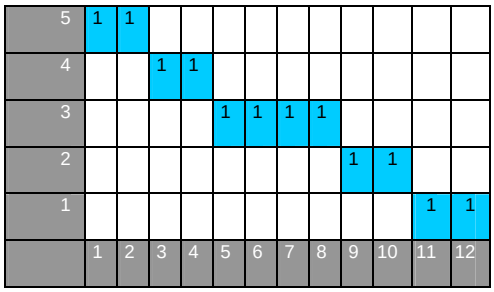
B.4.1 Quality profiles for 120 seconds calls

Table B.4

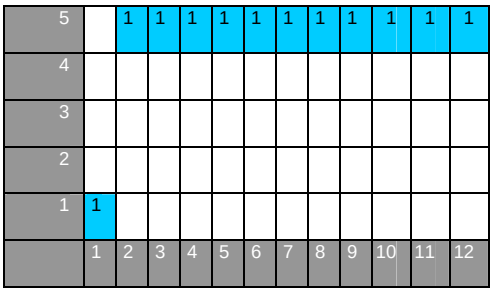
Qlevel	Profile 1											
5											1	1
4									1	1		
3					1	1	1	1				
2			1	1								
1	1	1										
	1	2	3	4	5	6	7	8	9	10	11	12

Qlevel	Profile 6											
5	1	1	1	1					1	1	1	1
4												
3					1	1	1	1				
2												
1												
	1	2	3	4	5	6	7	8	9	10	11	12

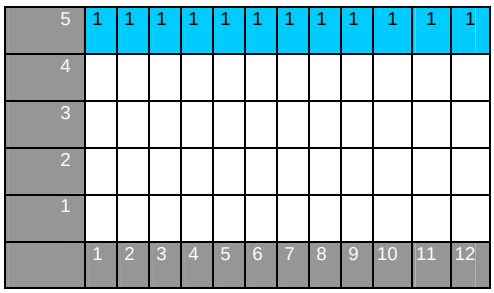
Qlevel Profile 2



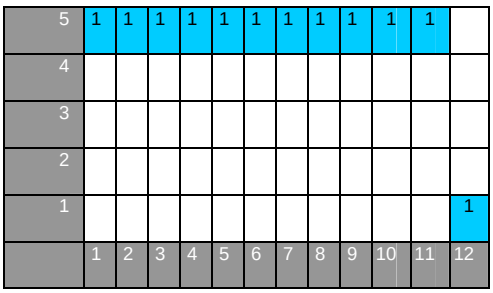
Qlevel Profile 7



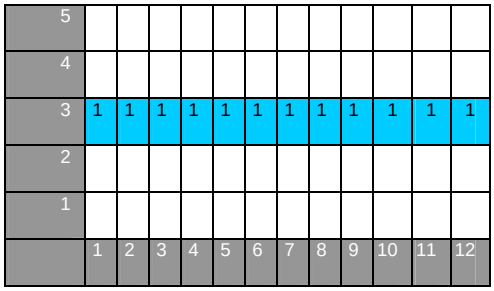
Qlevel Profile 3



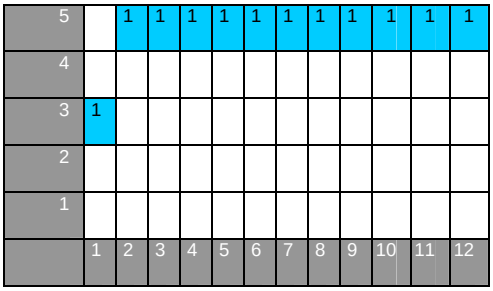
Qlevel Profile 8



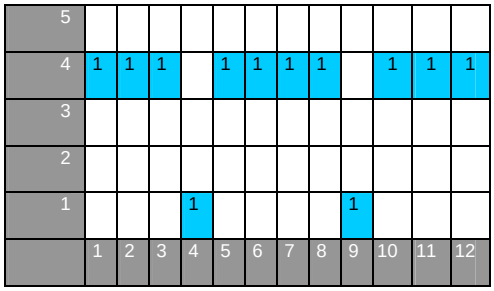
Qlevel Profile 4



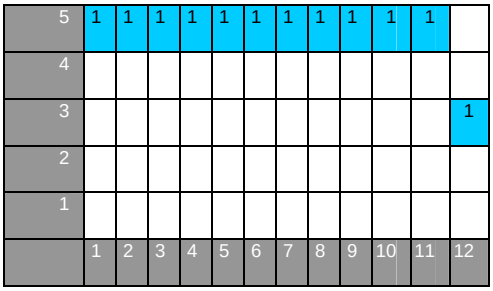
Qlevel Profile 9



Qlevel Profile 5



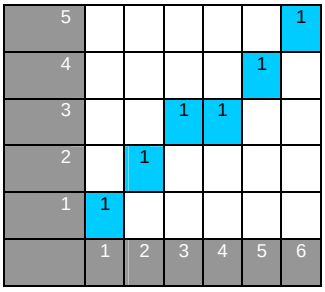
Qlevel Profile 10



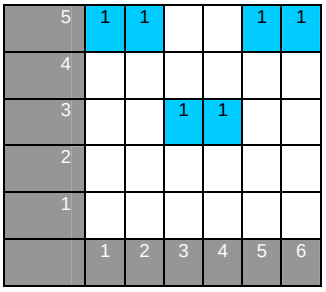
B.4.2 Quality profiles for 60 seconds calls

Table B.5

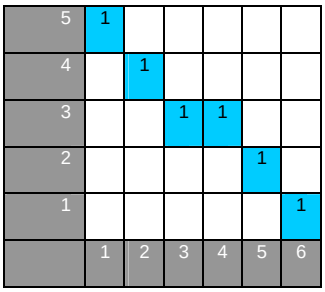
Qlevel Profile 11



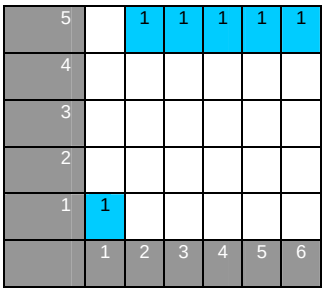
Qlevel Profile 16



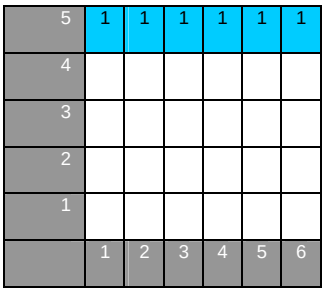
Qlevel Profile 12



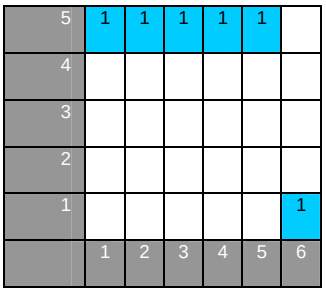
Qlevel Profile 17



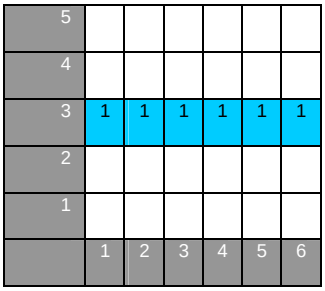
Qlevel Profile 13



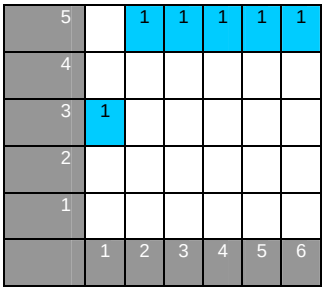
Qlevel Profile 18

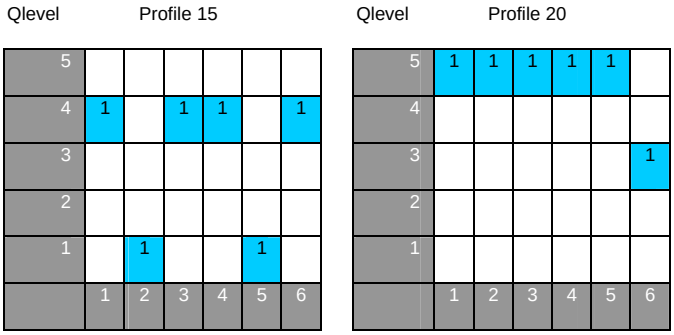


Qlevel Profile 14



Qlevel Profile 19




Qlevel
Profile 15
5
4
3
2
1
1
2
3
4
5
6
Qlevel
Profile 20
5
4
3
2
1
1
2
3
4
5
6

Annex C:
Study on the perceived call quality with German samples (T-Labs, 2007)

C.1 Introduction

Two two-part experiments were conducted to assess perceived quality associated with short speech samples as well as simulated telephone conversations (1 minute, 2 minutes), consisting of these samples. In the first part of the experiments, subjects had to listen to short stimuli and verbally answer questions regarding the content of the stimulus just heard. After five of these stimuli, they had to rate the quality of the whole simulated conversation on an ACR-scale. Subjects were instructed to try to put themselves in the position of an interlocutor. The five stimuli were degraded and formed different profiles. In the second part, which was identical for both experiments, all samples used were rated separately without interaction.

C.2 Test Design

C.2.1 Material

High quality recordings (48 kHz, 16-bit) from four speakers (two male, two female) were used to produce the material for both experiments. Every speaker simulated a dialog partner of a telephone conversation regarding a unique topic (zoo visit, car rent, kitchen purchase order, making an appointment with a dentist) (see V1.1.1 of the present document). 10 samples (5 seconds to 6 seconds) of each recording were used. These samples were down-sampled to 8 kHz, filtered by IRS (Intermediate Reference System) and processed several times (1 to 9) over real mobile channels in order to obtain natural material with different degrees of degradation.

Out of these samples one "dialog" was created, with five stimuli and four breaks for subject interaction. The stimuli and pauses had the same length, with a fixed break duration. So for the experiment with longer dialogs, two succeeding samples build one stimulus (11 seconds to 12 seconds). All 10 samples were used for one dialog with 12 s breaks, resulting in 105 seconds to 108 seconds before rating (2 minutes test). For the test with shorter dialogs, each sample formed a stimulus with breaks 6,5 seconds long, resulting in 53 seconds to 56 seconds before rating (1 minute test). For one of the 1 minute test dialogs, either the first five or the last five samples were used to ensure coherent contents.

Not all degradation pattern could be built out of the processed samples. Where necessary, those qualities needed were simulated due to AMR codec (Adaptive Multi-Rate) with frame losses or GSM-HR codec (Global System for Mobile communication, Half Rate) (occurence:16 % in the 1 minute test, 19 % in the 2 minutes test). The use of processed samples was preferred over simulated ones to restrict the number of those. Because of that, not all dialogs turned out as good realizations of the profiles. Especially, dialogs of profiles with constant poor stimuli were not that constant.

For the 1 minute test, 40 dialogs were created, with all ten profiles for every speaker, using 5 times the first, 5 times the second part of the 10 samples. For the 2 minutes test, 10 profiles were produced for both gender.

C.2.2 Subjects

24 naive subjects participated in every experiment; 9 female, 15 male each (aged between 17 and 48). All of them were paid and recruited outside the laboratory. All reported normal hearing, which was controlled with an audiometer (DIN EN 60645-2 [7]) prior to the experiments.

C.2.3 Procedure

Every experiment consisted of two parts. After the audiometer test, the subjects individually sat in a silent room in front of a computer screen. The stimuli were presented via a handset (Post FeTAp 752). The subjects were told to concentrate on the content of the stimuli. After presentation, a short question popped up at the screen with three possible answers (one correct) and the option "I could not understand / cannot remember". After five stimuli, no question appeared. Instead, the subjects had to rate the overall quality of transmission on an ACR chart on the screen. This rating was done with a computer mouse to end the simulated conversation explicitly with this change in modality. After one dialog, the next one started immediately. As training, one additional profile was presented at the beginning of the test. This one had a falling quality contour, so the subjects could get an impression of the quality range. After half of the dialogs, there was a pause of 5 minutes to 10 minutes. The dialogs were presented in five different pseudo-randomized orders preventing directly succeeding stimuli of one speaker and series of profiles. Every content of the dialogs was presented 5 times with different profiles to one subject. For each of theses contents, 2 or 3 different questions were formulated. To make every screen picture unique, the incorrect answers were always different as well as the position of all three answers. Regardless of the experiment, one performance took about 55 minutes to 65 minutes, including audiometer test and briefing.

In the second part of the experiments, every sample (not stimulus) used in both experiments has been rated in random order on ACR. This part included also one break of 5 minutes to 10 minutes (35 minutes to 45 minutes altogether). As the first three of all 48 subjects did only rate such samples they had listened to, there are 45 to 48 ratings for each sample.

C.2.4 Results

The correlation between overall judgments (MOS) and those predicted by P.862.1 is lesser than one could expect. This is caused by non-optimal prediction of the short term MOS that combines with the special distribution of simulated samples in profile positions with strong degradation. However, the increase due to the model compared to the plain mean (P.862.1) is comparable to the one with MOS values.

C.3 Detailed test results 60 seconds calls

C.3.1 Rating of the calls

Table C.1

Pattern	Votes	Structure					MOS (Call)	StDev (Call)
		Utterance A	Utterance B	Utterance C	Utterance D	Utterance E		
P 1 - male1	24	M1_A_hm	M1_B_hm	M1_C_hm	M1_D_hm	M1_E_hm	3,96	0,42
P 2 - male1	24	M1_A_mm	M1_B_mm	M1_C_mm	M1_D_mm2	M1_E_mm	3,67	0,32
P 3 - male1	24	M1_A_ml	M1_B_ml2	M1_C_ml	M1_D_ml	M1_E_ll	1,08	0,12
P 4 - male1	24	M1_A_ll	M1_B_ml	M1_C_hm	M1_D_hm	M1_E_hh	2,96	0,47
P 5 - male1	24	M1_A_hh	M1_B_hm5	M1_C_mm	M1_D_ml2	M1_E_ll	1,96	0,29
P 6 - male1	24	M1_A_b	M1_B_hm3	M1_C_hm3	M1_D_hm3	M1_E_hm3	3,33	0,37
P 7 - male1	24	M1_A_hm2	M1_B_hm2	M1_C_b	M1_D_hm2	M1_E_hm2	3,04	0,34
P 8 - male1	24	M1_A_hm4	M1_B_hm4	M1_C_hm4	M1_D_hm4	M1_E_ll	2,13	0,31
P 9 - male1	24	M1_A_hm	M1_B_hm	M1_C_ll	M1_D_hm	M1_E_hm	3,54	0,39
P 10 - male1	24	M1_A_hm	M1_B_hm	M1_C_hm	M1_D_mm	M1_E_ll2	2,92	0,33
P 1 - male2	24	M2_A_hm3	M2_B_hm3	M2_C_hm3	M2_D_hm3	M2_E_hm3	3,96	0,38
P 2 - male2	24	M2_A_mm	M2_B_mm	M2_C_mm	M2_D_mm	M2_E_mm	3,00	0,50
P 3 - male2	24	M2_A_ml	M2_B_ml2	M2_C_ml2	M2_D_ml2	M2_E_ml2	2,00	0,35
P 4 - male2	24	M2_A_ll	M2_B_ml2	M2_C_mm2	M2_D_mm2	M2_E_hh	3,33	0,44
P 5 - male2	24	M2_A_hh	M2_B_hm4	M2_C_mm2	M2_D_ml	M2_E_ll2	1,63	0,30
P 6 - male2	24	M2_A_b	M2_B_hm	M2_C_hm	M2_D_hm	M2_E_hm	2,63	0,39
P 7 - male2	24	M2_A_hm2	M2_B_hm	M2_C_b	M2_D_hm	M2_E_hm	3,58	0,39
P 8 - male2	24	M2_A_hm2	M2_B_hm	M2_C_hm	M2_D_hm	M2_E_b	2,67	0,41
P 9 - male2	24	M2_A_hm2	M2_B_hm	M2_C_ll	M2_D_hm	M2_E_hm	3,79	0,43
P 10 - male2	24	M2_A_hm4	M2_B_hm4	M2_C_hm4	M2_D_mm2	M2_E_ll	3,33	0,34
P 1 - female1	24	F1_A_hm3	F1_B_hm3	F1_C_hm3	F1_D_hm3	F1_E_hm3	4,04	0,36
P 2 - female1	24	F1_A_ml	F1_B_ml	F1_C_mm	F1_D_mm	F1_E_mm	3,83	0,27
P 3 - female1	24	F1_A_ml	F1_B_ml	F1_C_ml	F1_D_ml	F1_E_ll	2,29	0,32
P 4 - female1	24	F1_A_ll	F1_B_ml	F1_C_mm	F1_D_hm	F1_E_hh	3,21	0,37
P 5 - female1	24	F1_A_hh	F1_B_hm3	F1_C_mm2	F1_D_ml2	F1_E_ll2	2,58	0,35
P 6 - female1	24	F1_A_b	F1_B_hm	F1_C_hm	F1_D_hm	F1_E_hm	3,50	0,37
P 7 - female1	24	F1_A_hm3	F1_B_hm3	F1_C_b	F1_D_hm3	F1_E_hm3	3,25	0,44
P 8 - female1	24	F1_A_hm3	F1_B_hm3	F1_C_hm3	F1_D_hm3	F1_E_b	3,21	0,33
P 9 - female1	24	F1_A_hm3	F1_B_hm3	F1_C_ll	F1_D_hm3	F1_E_hm3	3,83	0,39
P 10 - female1	24	F1_A_hm2	F1_B_hm	F1_C_hm	F1_D_mm	F1_E_ll	3,04	0,36
P 1 - female2	24	F2_A_hm	F2_B_hm	F2_C_hm	F2_D_hm	F2_E_hm	3,83	0,32
P 2 - female2	24	F2_A_mm3	F2_B_mm3	F2_C_mm3	F2_D_mm3	F2_E_mm3	4,00	0,33
P 3 - female2	24	F2_A_ml	F2_B_ml2	F2_C_ml	F2_D_ml2	F2_E_ml	2,29	0,44
P 4 - female2	24	F2_A_ll	F2_B_ml	F2_C_mm	F2_D_hm2	F2_E_hh	3,25	0,38
P 5 - female2	24	F2_A_hh	F2_B_hm4	F2_C_mm2	F2_D_ml2	F2_E_ll	2,75	0,31
P 6 - female2	24	F2_A_b	F2_B_hm	F2_C_hm	F2_D_hm	F2_E_hm	3,08	0,45
P 7 - female2	24	F2_A_hm	F2_B_hm	F2_C_b	F2_D_hm2	F2_E_hm	3,00	0,35
P 8 - female2	24	F2_A_hm3	F2_B_hm3	F2_C_hm3	F2_D_hm3	F2_E_b	2,96	0,38
P 9 - female2	24	F2_A_hm	F2_B_hm	F2_C_ll	F2_D_hm2	F2_E_hm	3,58	0,30
P 10 - female2	24	F2_A_hm3	F2_B_hm3	F2_C_hm3	F2_D_mm2	F2_E_ml	3,67	0,20

C.3.2 Rating of the utterances

Table C.2

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
M1_A_b	48	1,35	0,31	1,74
M1_A_hh	45	3,44	0,54	4,52
M1_A_hm	45	4,31	0,36	3,95
M1_A_hm2	45	3,98	0,47	3,94
M1_A_hm4	45	3,69	0,46	3,94
M1_A_ll	45	2,07	0,45	1,74
M1_A_ml	45	2,73	0,45	2,58
M1_A_mm	45	3,67	0,44	3,37
M1_B_hm	45	4,02	0,41	4,08
M1_B_hm2	48	3,69	0,46	4,01
M1_B_hm3	48	3,71	0,46	4,09
M1_B_hm4	45	3,58	0,39	4,09
M1_B_hm5	45	3,89	0,48	4,11
M1_B_ml	45	3,2	0,44	3,06
M1_B_ml2	45	2,25	0,39	2,49
M1_B_mm	45	3,69	0,33	3,82
M1_C_b	48	1,42	0,29	1,95
M1_C_hm	45	3,69	0,57	4,1
M1_C_hm3	48	3,92	0,48	4,09
M1_C_hm4	45	3,76	0,45	4,11
M1_C_ll	45	2,31	0,49	2,79
M1_C_ml	48	1,17	0,25	1,42
M1_C_mm	45	3,64	0,43	3,16
M1_D_hm	48	3,92	0,45	4,06
M1_D_hm2	48	3,65	0,5	3,91
M1_D_hm3	48	3,79	0,45	4,03
M1_D_hm4	45	3,47	0,51	3,95
M1_D_ml	48	1,06	0,25	1,31
M1_D_ml2	45	2,53	0,44	2,76
M1_D_mm	45	3,67	0,5	3,24
M1_D_mm2	45	3,64	0,45	3,58
M1_E_hh	48	3,71	0,46	4,5
M1_E_hm	45	3,92	0,49	3,95
M1_E_hm2	45	3,76	0,48	4
M1_E_hm3	48	3,94	0,49	3,97
M1_E_ll	48	1,1	0,34	1,3
M1_E_ll2	45	1,98	0,37	2,33
M1_E_mm	45	3,47	0,49	3,14

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
M2_A_b	45	1,75	0,47	1,73
M2_A_hh	45	4,18	0,45	4,53
M2_A_hm2	48	3,9	0,45	4,11
M2_A_hm3	48	4,08	0,41	4,1
M2_A_hm4	48	4,17	0,45	4,11
M2_A_ll	48	2,67	0,38	2,26
M2_A_ml	48	2,52	0,72	3,18
M2_A_mm	48	3,08	0,52	3,36
M2_B_hm	48	3,31	0,61	4,11
M2_B_hm3	48	3,79	0,43	4,12
M2_B_hm4	48	4,17	0,42	4,1
M2_B_ml2	48	2,24	0,37	1,86
M2_B_mm	48	2,02	0,41	2,64
M2_C_b	45	2,38	0,45	1,89
M2_C_hm	48	4,15	0,43	4,11
M2_C_hm3	48	4,15	0,41	4,13
M2_C_hm4	48	4,13	0,43	4,03
M2_C_ll	48	2,78	0,38	2,1

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
M2_C_ml2	48	1,91	0,44	1,78
M2_C_mm	48	2,88	0,46	2,82
M2_C_mm2	48	2,98	0,47	2,99
M2_D_hm	48	3,96	0,43	4,06
M2_D_hm3	48	4,15	0,41	4,01
M2_D_ml	48	1,92	0,36	2,06
M2_D_ml2	48	3,67	0,54	2,61
M2_D_mm	48	3,9	0,45	3,36
M2_D_mm2	45	4,09	0,42	3,82
M2_E_b	48	1,53	0,33	1,8
M2_E_hh	48	4,08	0,48	4,53
M2_E_hm	48	3,71	0,43	3,93
M2_E_hm3	48	4,17	0,44	4,07
M2_E_ll	48	2,52	0,45	1,98
M2_E_ll2	45	1	0	1,1
M2_E_ml2	45	2,07	0,55	2,66
M2_E_mm	48	3,44	0,48	3,3

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
F1_A_b	45	1,96	0,4	1,96
F1_A_hh	48	4,36	0,43	4,47
F1_A_hm2	48	4,27	0,43	3,89
F1_A_hm3	48	4,06	0,45	3,89
F1_A_ll	45	2,09	0,44	1,83
F1_A_ml	48	2,77	0,54	2,97
F1_B_hm	48	4,23	0,44	4,07
F1_B_hm3	48	4,29	0,4	4,04
F1_B_ml	48	3,08	0,46	2,81
F1_C_b	45	1,31	0,28	2,09
F1_C_hm	48	4,1	0,47	3,96
F1_C_hm3	48	3,98	0,5	3,98
F1_C_ll	48	2,06	0,44	1,87
F1_C_ml	48	2,29	0,46	1,86
F1_C_mm	48	3,6	0,48	3,28
F1_C_mm2	48	3,73	0,53	3,63
F1_D_hm	48	4,19	0,54	3,99
F1_D_hm3	48	4,25	0,46	3,99
F1_D_ml	48	2,06	0,4	2,15
F1_D_ml2	48	2,08	0,51	1,7
F1_D_mm	48	4,23	0,44	3,57
F1_E_b	45	2,58	0,43	1,84
F1_E_hh	45	4,22	0,44	4,49
F1_E_hm	48	4,08	0,45	3,95
F1_E_hm3	48	4,29	0,43	4,02
F1_E_ll	48	1,96	0,34	1,87
F1_E_ll2	45	2,25	0,59	1,79
F1_E_mm	48	3,77	0,5	3,42

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
F2_A_b	48	1,52	0,34	1,92
F2_A_hh	45	4,16	0,48	4,49
F2_A_hm	48	3,94	0,49	3,96
F2_A_hm3	48	4,42	0,43	3,88
F2_A_ll	48	1,98	0,42	1,83
F2_A_ml	45	3,22	0,38	2,26
F2_A_mm3	48	3,81	0,41	3,44
F2_B_hm	48	3,67	0,42	3,84
F2_B_hm3	48	4,1	0,45	3,85
F2_B_hm4	48	3,94	0,44	3,63
F2_B_ml	45	2,89	0,5	2,8
F2_B_ml2	48	2,63	0,52	2,07
F2_B_mm3	45	2,93	0,45	3,15

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
F2_C_b	45	1,47	0,33	1,94
F2_C_hm	48	4,13	0,5	3,91
F2_C_hm3	48	4,1	0,45	3,97
F2_C_ll	48	2,9	0,45	2,51
F2_C_ml	48	2,23	0,44	2,04
F2_C_mm	45	3,22	0,44	3,46
F2_C_mm2	45	2,44	0,51	3,03
F2_C_mm3	45	3,69	0,48	3,63
F2_D_hm	48	4,13	0,39	3,91
F2_D_hm2	48	3,71	0,45	3,82
F2_D_hm3	48	4,13	0,5	3,81
F2_D_ml2	45	3,02	0,45	2,54
F2_D_mm2	45	3,87	0,51	3,7
F2_D_mm3	45	3,24	0,34	3,27
F2_E_b	48	1,77	0,48	1,66
F2_E_hh	45	3,91	0,44	4,47
F2_E_hm	48	4,19	0,41	3,87
F2_E_ll	48	2,02	0,41	1,74
F2_E_ml	45	2,73	0,43	2,39
F2_E_mm3	45	3,76	0,5	3,29

C.4 Detailed test results 120 seconds calls

C.4.1 Rating of the calls

Table C.3

Pattern	Votes	Structure										MOS (Call)	StDev (Call)
		Utterance A	Utterance B	Utterance C	Utterance D	Utterance E	Utterance F	Utterance G	Utterance H	Utterance I	Utterance J		
P 1 - male	24	M1_A_hm	M1_B_hm	M1_C_hm2	M1_D_hm2	M1_E_hm	M1_F_hm	M1_G_hm	M1_H_hm	M1_I_hm	M1_J_hm	4.17	0.37
P 2 - male	24	M2_A_mm	M2_B_mm	M2_C_mm	M2_D_mm	M2_E_mm	M2_F_mm	M2_G_mm	M2_H_mm	M2_I_mm	M2_J_mm	3.25	0.36
P 3 - male	24	M1_A_ml	M1_B_ml	M1_C_ml	M1_D_ml	M1_E_ml	M1_F_ml	M1_G_ml	M1_H_ml	M1_I_b	M1_J_b	1.46	0.28
P 4 - male	24	M2_A_ll	M2_B_ll	M2_C_ml	M2_D_ml	M2_E_mm	M2_F_mm	M2_G_hm	M2_H_hm	M2_I_hh	M2_J_hh	2.33	0.34
P 5 - male	24	M1_A_hh	M1_B_hh	M1_C_hm	M1_D_hm	M1_E_mm	M1_F_mm	M1_G_ml2	M1_H_ml2	M1_I_b	M1_J_b	1.75	0.31
P 6 - male	24	M2_A_b	M2_B_b	M2_C_hm	M2_D_hm	M2_E_hm2	M2_F_hm2	M2_G_hm2	M2_H_hm2	M2_I_hm	M2_J_hm	3.42	0.41
P 7 - male	24	M1_A_hm	M1_B_hm	M1_C_hm2	M1_D_hm2	M1_E_b	M1_F_b	M1_G_hm	M1_H_hm	M1_I_hm	M1_J_hm	2.92	0.39
P 8 - male	24	M1_A_hm	M1_B_hm	M1_C_hm2	M1_D_hm2	M1_E_hm	M1_F_hm	M1_G_hm	M1_H_hm	M1_I_b	M1_J_b	2.42	0.46
P 9 - male	24	M2_A_hm	M2_B_hm	M2_C_hm	M2_D_hm	M2_E_ll	M2_F_ll	M2_G_hm2	M2_H_hm2	M2_I_hm	M2_J_hm	3.33	0.34
P 10 - male	24	M2_A_hm3	M2_B_hm3	M2_C_hm3	M2_D_hm3	M2_E_hm3	M2_F_hm3	M2_G_mm	M2_H_mm	M2_I_ll	M2_J_ll	3.33	0.39
P 1 - female	24	F1_A_hm	F1_B_hm	F1_C_hm	F1_D_hm	F1_E_hm	F1_F_hm	F1_G_hm	F1_H_hm	F1_I_hm	F1_J_hm	4.13	0.34
P 2 - female	24	F1_A_mm	F1_B_mm	F1_C_mm	F1_D_mm	F1_E_mm	F1_F_mm	F1_G_mm	F1_H_mm	F1_I_mm	F1_J_mm	3.75	0.4
P 3 - female	24	F1_A_mm	F1_B_mm	F1_C_ml	F1_D_ml	F1_E_ml	F1_F_ml	F1_G_ml	F1_H_ml	F1_I_ll	F1_J_ll	2.58	0.28
P 4 - female	24	F2_A_ll	F2_B_ll	F2_C_ml	F2_D_ml	F2_E_mm	F2_F_mm	F2_G_hm2	F2_H_hm2	F2_I_hh	F2_J_hh	3.17	0.37
P 5 - female	24	F2_A_hh	F2_B_hh	F2_C_hh	F2_D_hm	F2_E_mm	F2_F_mm	F2_G_ml	F2_H_ml	F2_I_ll	F2_J_ll	2.83	0.46
P 6 - female	24	F2_A_b	F2_B_b	F2_C_hm	F2_D_hm2	F2_E_hm	F2_F_hm	F2_G_hm	F2_H_hm	F2_I_hm	F2_J_hm	3.13	0.42
P 7 - female	24	F2_A_hm	F2_B_hm	F2_C_hm	F2_D_hm2	F2_E_b	F2_F_b	F2_G_hm	F2_H_hm	F2_I_hm	F2_J_hm	2.92	0.35
P 8 - female	24	F2_A_hm	F2_B_hm	F2_C_hm	F2_D_hm2	F2_E_hm	F2_F_hm	F2_G_hm	F2_H_hm	F2_I_b	F2_J_b	3	0.39
P 9 - female	24	F1_A_hm	F1_B_hm	F1_C_hm	F1_D_hm	F1_E_ml	F1_F_ml	F1_G_hm	F1_H_hm	F1_I_hm	F1_J_hm	3.83	0.34
P 10 - female	24	F1_A_hm	F1_B_hm	F1_C_hm	F1_D_hm	F1_E_hm	F1_F_hm	F1_G_mm	F1_H_mm	F1_I_ll	F1_J_ll	3.54	0.43

C.4.2 Rating of the utterances

Table C.4

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
M1_A_hh	48	4,06	0,39	4,50
M1_A_hm	48	4,15	0,47	3,90
M1_A_ml	48	3,13	0,47	3,18
M1_B_hh	48	3,77	0,46	4,51
M1_B_hm	48	4,11	0,45	4,11
M1_B_ml	48	2,62	0,43	3,12
M1_C_hm	48	4,00	0,47	4,10
M1_C_hm2	48	4,02	0,50	4,03
M1_C_ml	48	1,42	0,29	1,95
M1_D_hm	48	3,65	0,50	3,91
M1_D_hm2	48	3,92	0,45	4,06
M1_D_ml	48	3,77	0,46	3,69
M1_E_b	48	2,04	0,39	1,93
M1_E_hm	48	3,92	0,49	3,95
M1_E_ml	48	2,28	0,45	2,39
M1_E_mm	48	3,60	0,48	3,81
M1_F_b	48	1,35	0,31	1,74
M1_F_hm	48	3,89	0,49	4,08
M1_F_ml	48	2,49	0,44	2,48
M1_F_mm	48	3,45	0,47	3,67
M1_G_hm	48	3,71	0,46	4,09
M1_G_ml	48	2,25	0,39	2,49
M1_G_ml2	48	3,32	0,55	3,43
M1_H_hm	48	3,92	0,48	4,09
M1_H_ml	48	1,17	0,25	1,42
M1_H_ml2	48	1,36	0,31	1,69
M1_I_b	48	1,06	0,25	1,31
M1_I_hm	48	3,79	0,45	4,03
M1_J_b	48	1,10	0,34	1,30
M1_J_hm	48	3,94	0,49	3,97

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
M2_A_b	48	1,75	0,47	1,73
M2_A_hm	48	3,90	0,45	4,11
M2_A_hm3	48	4,11	0,45	4,09
M2_A_ll	48	1,66	0,45	1,51
M2_A_mm	48	3,08	0,52	3,36
M2_B_b	48	1,64	0,38	1,98
M2_B_hm	48	3,31	0,61	4,11
M2_B_hm3	48	3,77	0,45	4,06
M2_B_ll	48	1,66	0,43	1,62
M2_B_mm	48	2,02	0,41	2,64
M2_C_hm	48	4,15	0,43	4,11
M2_C_hm3	48	4,38	0,43	4,18
M2_C_ml	48	2,40	0,40	1,87
M2_C_mm	48	2,88	0,46	2,82
M2_D_hm	48	3,96	0,43	4,06
M2_D_hm3	48	4,26	0,43	4,06
M2_D_ml	48	2,47	0,40	1,80
M2_D_mm	48	3,90	0,45	3,36
M2_E_hm2	48	3,71	0,43	3,93
M2_E_hm3	48	4,00	0,47	4,04
M2_E_ll	48	1,77	0,41	1,57
M2_E_mm	48	3,44	0,48	3,30
M2_F_hm2	48	4,08	0,41	4,10

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
M2_F_hm3	48	4,17	0,45	4,11
M2_F_ll	48	2,68	0,43	2,29
M2_F_mm	48	2,52	0,72	3,18
M2_G_hm	48	4,17	0,42	4,10
M2_G_hm2	48	3,79	0,43	4,12
M2_G_mm	48	3,87	0,48	3,75
M2_H_hm	48	4,13	0,43	4,03
M2_H_hm2	48	4,15	0,41	4,13
M2_H_mm	48	2,98	0,47	2,99
M2_I_hh	48	4,21	0,46	4,51
M2_I_hm	48	4,15	0,41	4,01
M2_I_ll	48	1,92	0,36	2,06
M2_I_mm	48	3,67	0,54	2,61
M2_J_hh	48	4,08	0,48	4,53
M2_J_hm	48	4,17	0,44	4,07
M2_J_ll	48	2,52	0,45	1,98
M2_J_mm	48	4,04	0,47	3,48

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
F1_A_hm	48	4,27	0,43	3,89
F1_A_mm	48	2,77	0,54	2,97
F1_B_hm	48	4,23	0,44	4,07
F1_B_mm	48	3,08	0,46	2,81
F1_C_hm	48	4,10	0,47	3,96
F1_C_ml	48	2,29	0,46	1,86
F1_C_mm	48	3,60	0,48	3,28
F1_D_hm	48	4,19	0,54	3,99
F1_D_ml	48	2,06	0,40	2,15
F1_D_mm	48	4,23	0,44	3,57
F1_E_hm	48	4,08	0,45	3,95
F1_E_ml	48	1,96	0,34	1,87
F1_E_mm	48	3,77	0,50	3,42
F1_F_hm	48	4,06	0,45	3,89
F1_F_ml	48	2,34	0,48	1,81
F1_F_mm	48	3,74	0,50	3,21
F1_G_hm	48	4,29	0,40	4,04
F1_G_ml	48	2,30	0,40	2,43
F1_G_mm	48	4,21	0,46	3,70
F1_H_hm	48	3,98	0,50	3,98
F1_H_ml	48	2,06	0,44	1,87
F1_H_mm	48	3,73	0,53	3,63
F1_I_hm	48	4,25	0,46	3,99
F1_I_ll	48	2,08	0,51	1,70
F1_I_mm	48	3,96	0,47	3,71
F1_j_hm	48	4,29	0,43	4,02
F1_J_ll	48	2,25	0,59	1,79
F1_J_mm	48	3,04	0,46	2,96

Utterance	Votes	MOS (Utt)	StDev(Utt)	P.862.1 (Utt)
F2_A_b	48	1,52	0,34	1,92
F2_A_hh	48	3,91	0,46	4,49
F2_A_hm	48	3,94	0,49	3,96
F2_A_ll	48	1,98	0,42	1,83
F2_B_b	48	1,89	0,43	2,10
F2_B_hh	48	3,55	0,47	4,50
F2_B_hm	48	3,67	0,42	3,84
F2_B_ll	48	1,94	0,46	1,92
F2_C_hh	48	4,02	0,43	4,50
F2_C_hm	48	4,13	0,50	3,91
F2_C_ml	48	2,90	0,45	2,51
F2_D_hm	48	3,71	0,45	3,82
F2_D_hm2	48	4,13	0,39	3,91
F2_D_ml	48	2,53	0,40	2,23
F2_E_b	48	1,77	0,48	1,66
F2_E_hm	48	4,19	0,41	3,87
F2_E_mm	48	3,53	0,50	3,48
F2_F_b	48	1,72	0,49	1,60
F2_F_hm	48	4,42	0,43	3,88
F2_F_mm	48	3,81	0,41	3,44
F2_G_hm	48	4,10	0,45	3,85
F2_G_hm2	48	3,94	0,44	3,63
F2_G_ml	48	2,63	0,52	2,07
F2_H_hm	48	4,10	0,45	3,97
F2_H_hm2	48	3,94	0,43	3,86
F2_H_ml	48	2,23	0,44	2,04
F2_I_b	48	2,04	0,47	2,07
F2_I_hh	48	4,02	0,51	4,48
F2_I_hm	48	4,13	0,50	3,81
F2_I_ll	48	1,96	0,56	1,80
F2_J_b	48	1,77	0,30	1,66
F2_J_hh	48	3,72	0,45	4,50
F2_J_hm	48	4,09	0,44	3,78
F2_J_ll	48	2,02	0,41	1,74

History

Document history		
V1.1.1	October 2006	Publication
V1.2.1	November 2007	Publication