ETSI TR 102 506 V1.1.1 (2006-10)

Technical Report

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

Reference

DTR/STQ-00088m

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

Important notice

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, please send your comment to one of the following services:
http://portal.etsi.org/chaircor/ETSI_support.asp

Copyright Notification

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 restricted to test signals between 100 seconds and 120 seconds in duration with alternating speech/silence periods as described in clause 5. The presented model is based on a single study and may not generalize to other call scenarios than those used in the underlying study.

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 P.800.1 [5] covering the relevant network distortions and speech processing components in their scope.

2 References

For the purposes of this Technical Report (TR) the following references apply:

[1] ITU-T Recommendation P.862: "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband 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] Ergebnisbericht (Study) Berkom PESQ-mobil.

[4] Presentation on Call Quality document 08TD17 from STQMobile #8 September 2005.

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

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

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

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

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 [6]. 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 [6] and the sample characteristics and evaluation is defined in ITU-T Recommendation P.862 [1] and ITU-T Recommendation P.862.3 [7]. 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 [4].

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 [6] is 120 s 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 [4]. Hence the recommended call length and the empirical evidence can be reconciled.

5.2.1 Length of utterance (sample)

The application guideline for objective speech measurement is ITU-T Recommendation P.862.3 [7]. The typical sample of measurement systems has a length from 5 s to 12 s 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)

Given the recommended length of the call in connection with length of the individual utterance it takes around five to six utterances and silence pairs to fill the 120 s. From empirical evidence we know that a typical conversational call contains around 4 utterances from each side [4] so that 5 recurrences of the speech and silent pair are recommended.

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 10 s to 12 s and that pair recurs 5 times during the call.

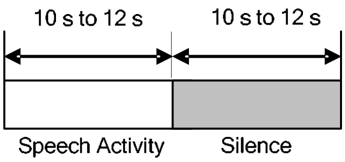


Figure 1: Structure of the speech activity silence pair

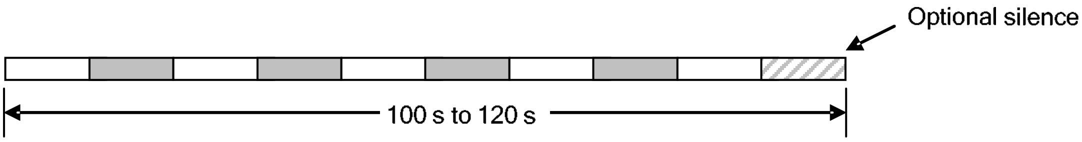


Figure 2: Structure of the call for one side

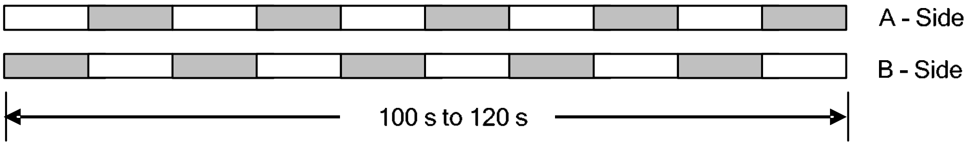


Figure 3: Structure of the call with alternating speech activity

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. The perceived speech quality of the samples is rated according to ITU-T Recommendation P.862.1 [2].

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 where n = 5 for the presented case. The weighing is made with an individual parameter a_t at that is the weighing factor for each sample. A mathematical model here is:

$$MOS_{RE} = \frac{\sum_{t=1}^{n} a_t MOS_t}{\sum_{t=1}^{n} a_t}$$

$$n = 5; a_1 = 0{,}4; a_2 = 0{,}5; a_3 = 0{,}6; a_4 = 0{,}8; a_5 = 1{,}0$$

The parameter become greater towards the end starting with a_1 = 0,4 to 1,0 for a_5. The increase of the parameter

represents the increasing importance of a sample for the general impression the closer it is located towards the end. Comparison with subjective data shows that this model achieves a correlation with empirical evidence of around 65 % (up from 57 % for the simple arithmetical average) ([3] see annex).

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.

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

The formula above is showing a correlation of 85 % with empirical data ([3] see annex).

6.2.3 Requirements for the applicability of the mathematical model

The application of the formula requires the conversation length to be 90 s to 120 s. There must be 5 utterances per analysed direction and the sample and the pause have to be 10 s to 12 s each. This results into a temporal separation of 20 s to 24 s between each quality assessment.

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 [3] 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 s to 8 s) 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 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 construced 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 s speech sample there is a 12 s 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.

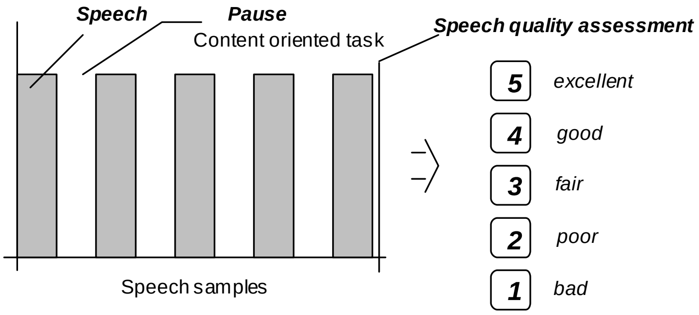


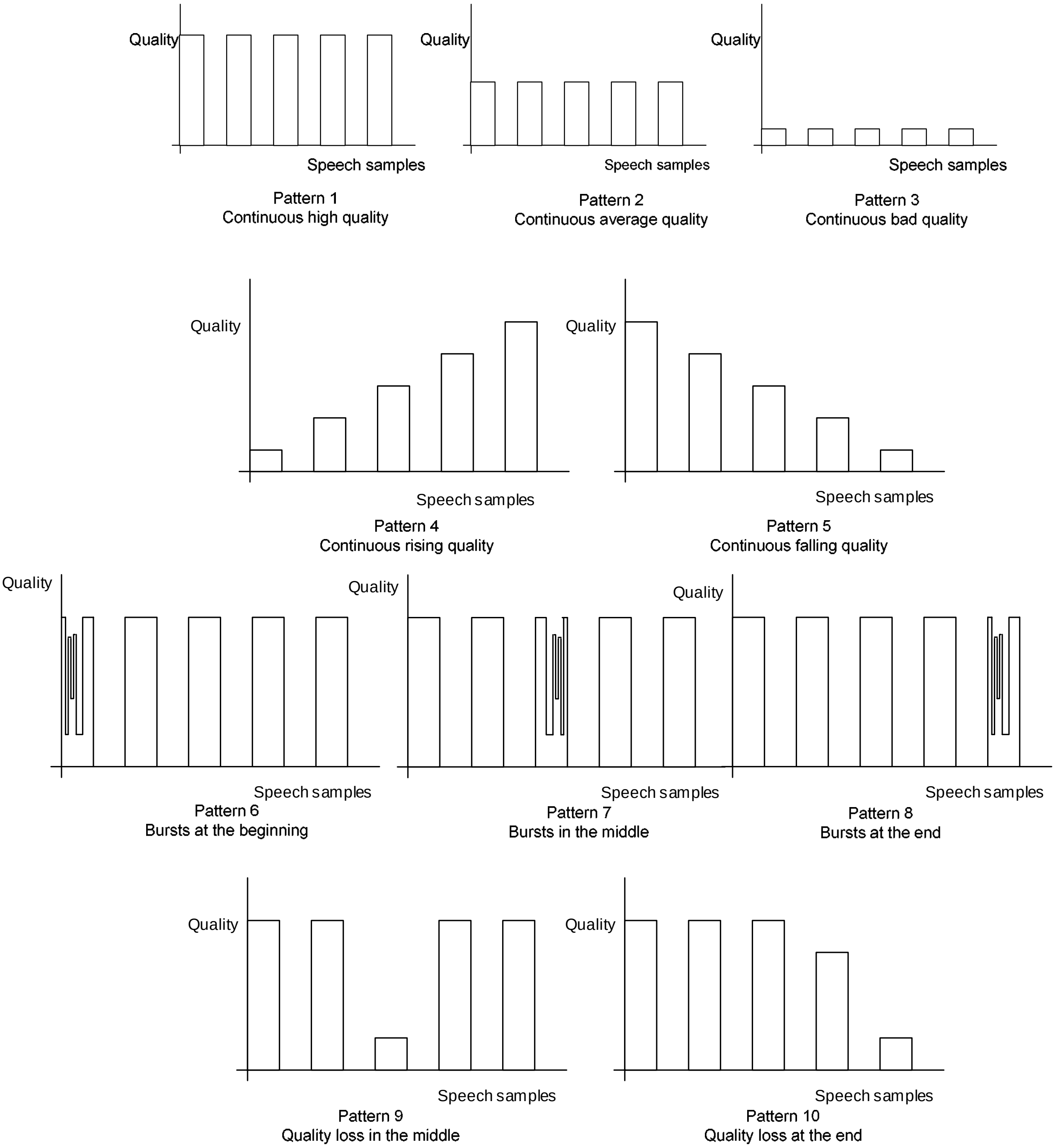
Figure A.1: Schematical presentation of the speech situation assessment

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 s 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 [8] 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.



Pattern 1
Continuous high quality

Pattern 2
Continuous average quality

Pattern 3
Continuous bad quality

Pattern 4
Continuous rising quality

Pattern 5
Continuous falling quality

Pattern 6
Bursts at the beginning

Pattern 7
Bursts in the middle

Pattern 8
Bursts at the end

Pattern 9
Quality loss in the middle

Pattern 10
Quality loss at the end

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 s to 90 s 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 s to 13 s with active speech. After each 12 s part, a pause of 12 s length is implemented. This results in an overall transmission length of 110 s.

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 s speech sample. The speech activity of the simulated dialogues is shown in the next table:

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.

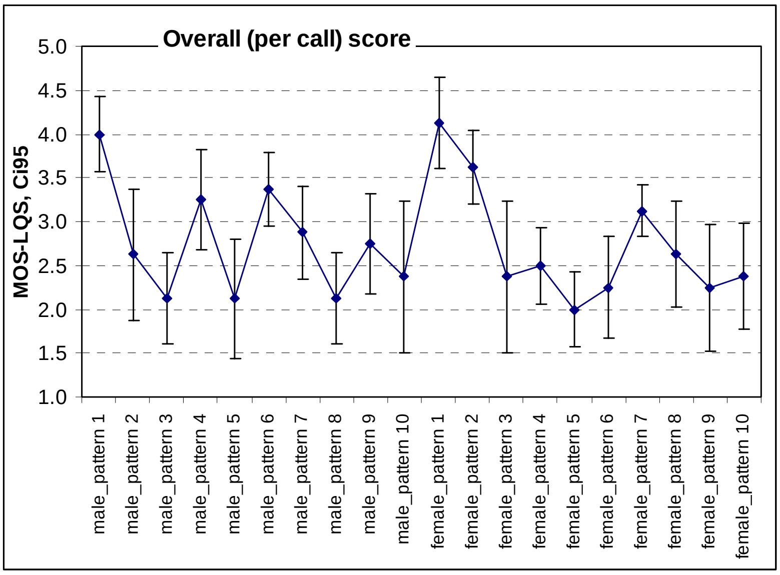


Figure A.3: Auditory MOS "per call" per pattern

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.

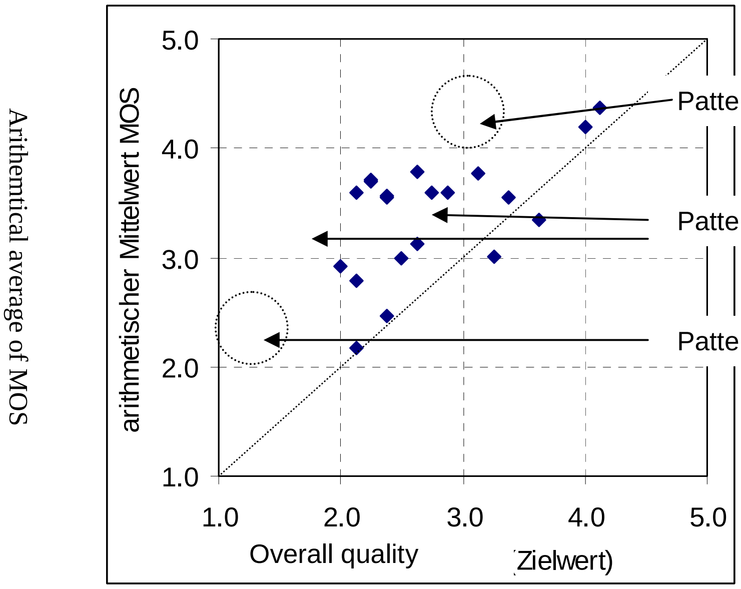


Figure A.4: Arithmetical average of the MOS assessment of the individual speech parts to the overall quality assessment

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 disproportiontely 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:

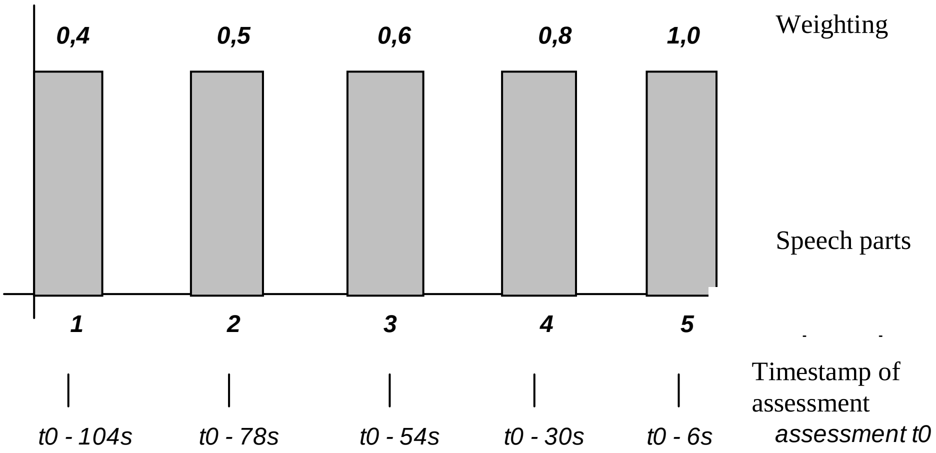


Figure A.5: Weighting of the individual speech parts

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 $\mathrm{MOS}_{\mathrm{Mod1}}$ can be calculated as follows:

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

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 %.

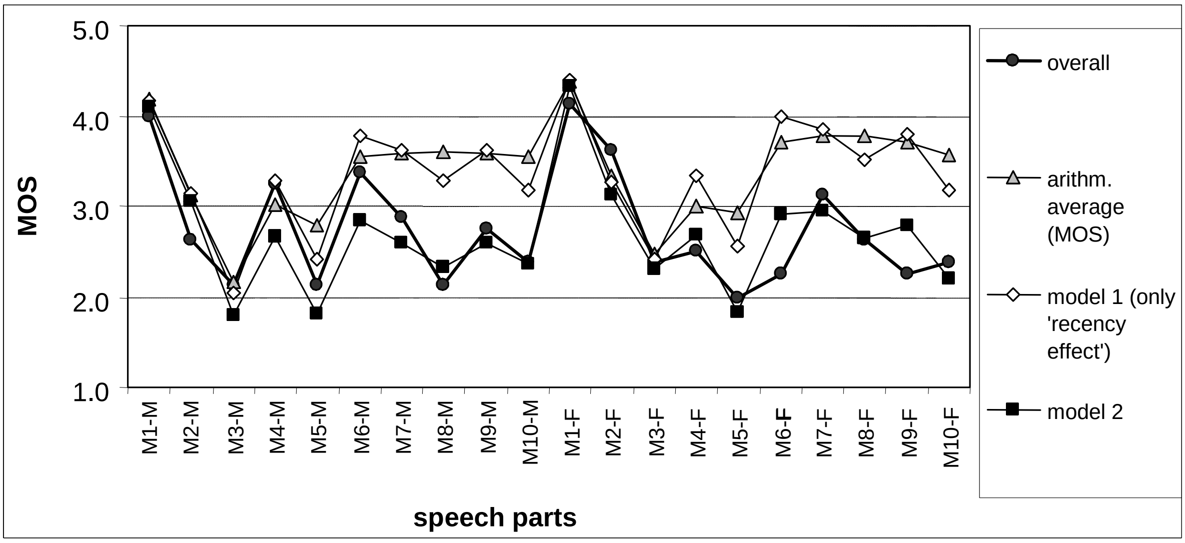


Figure A.6: Individual results by using model 2

It can be concluded that the Speech Quality per Call can be predicted reliably by a two step model as shown in clauses A3.2 and A3.3.

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 s.

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.

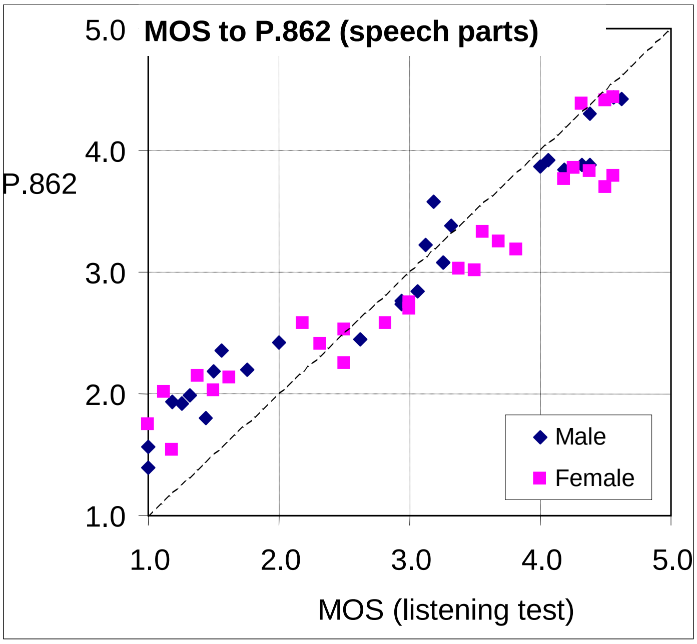


Figure A.7: ITU-T Recommendation P.862 [1] results in comparison to the separated speech parts

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:

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

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

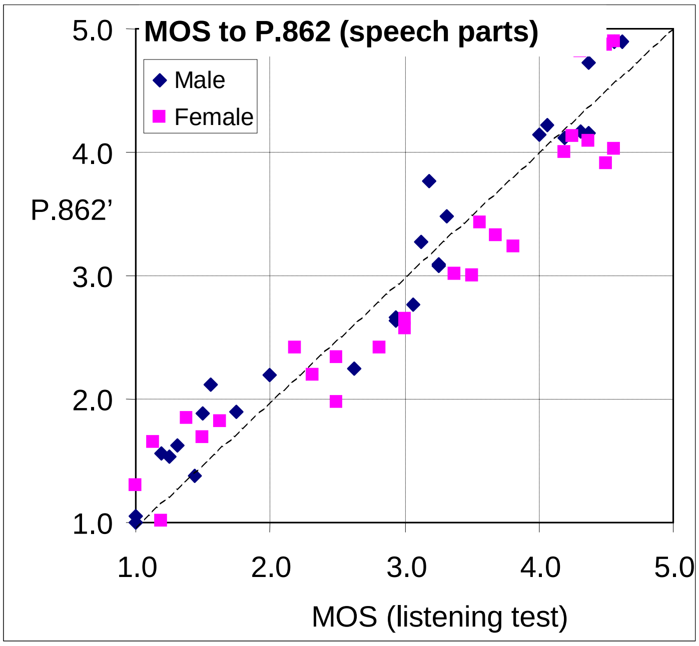


Figure A.8: ITU-T Recommendation P.862 [1] results in comparison to the separated speech parts with the scaling function

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.

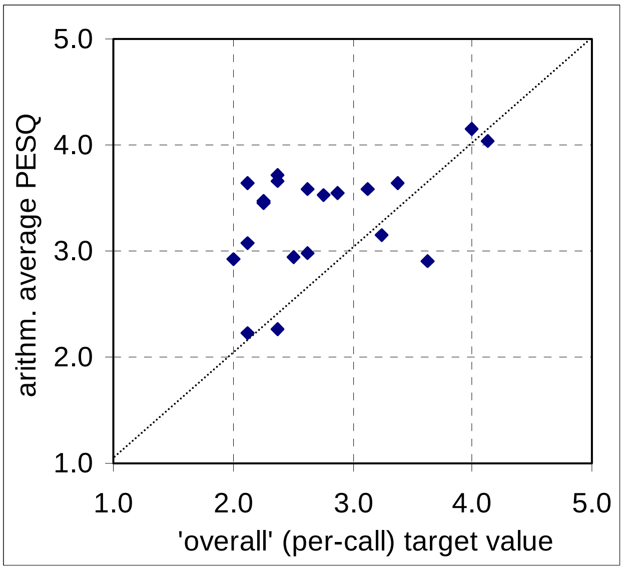


Figure A.9: Arithmetical average of the ITU-T Recommendation P.862 [1] assessment of the individual speech samples to the overall quality assessment

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.

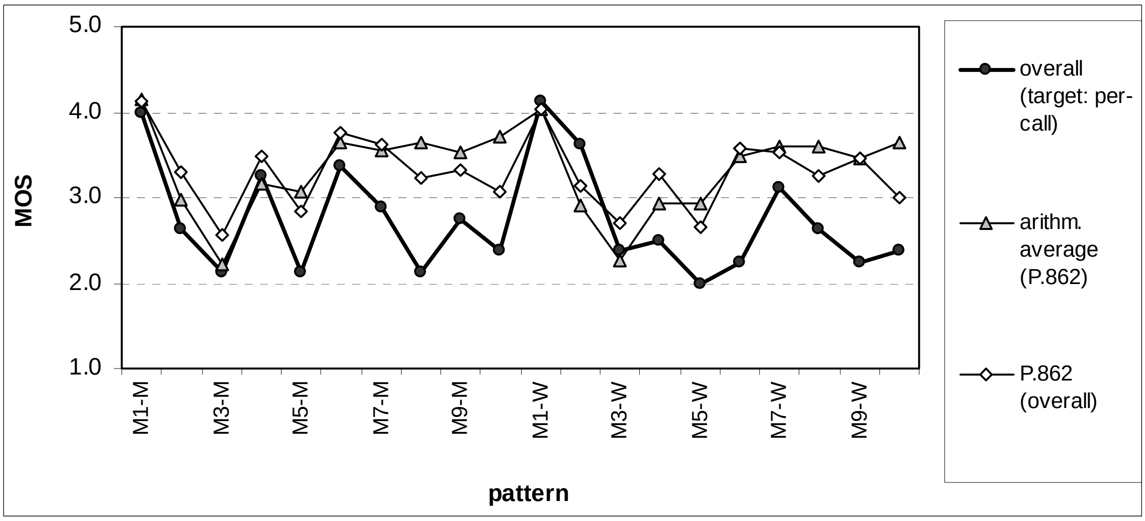


Figure A.10: Individual results by using ITU-T Recommendation P.862 [1]

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 s and 16 s 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:

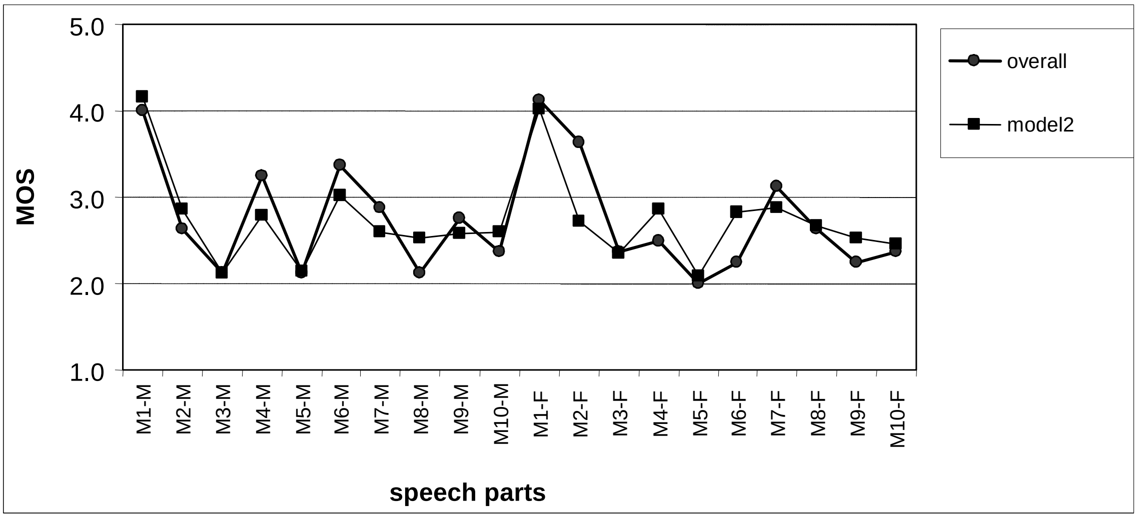


Figure A.11: Individual results by using model 2 with the ITU-T Recommendation P.862 [1] results

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

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

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

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

History

Document history		
V1.1.1	October 2006	Publication