

Deutsche Telekom Laboratories
An-Institut der Technischen Universität Berlin

Universal Speech Sample

For Quality Measurements in Fixed and Mobile Environments

Ulf Wüstenhagen (Deutsche Telekom Laboratories)
Jens Berger (SwissQual AG)

White Paper No. 6
March 2010


SwissQual



Table of contents

1	Introduction	4
1.1	Speech samples* to be selected and provided	4
1.2	Available speech recordings in this project.....	4
2	Phonological analysis.....	5
3	Objective analysis.....	5
4	Speaker dependency	6
4.1	Traditional narrow-band measures	6
4.2	Traditional wideband measures.....	7
4.3	Super-wideband measures.....	7
5	Selection of speech sample for further analysis	8
6	Selection of ‘Mixed’ samples.....	8
7	Subjective listening experiment	9
7.1	Test design.....	9
7.2	Test Results.....	10
8	Selection of speech samples	11
9	Comparison to objective scores	12
10	Limitations due to experimental design.....	13
11	Objective example scores for the selected speech samples	13
12	Post-processing of the selected file(s)	14
13	File naming convention	16
14	Appendix 1 Batch procedure for file processing	17

1.1.1 Universal Speech Sample

15	Appendix 2 Recording Conditions at Telekom Laboratories	19
16	List of Abbreviations	20
17	Index of figures	21
18	Index of tables	22
19	References	23

1 Introduction

The new universal speech sample became necessary for use in several objective measurement systems which are used within Deutsche Telekom. The samples which are used up to now were not fulfilling the current ITU-T Recommendations anymore. Focus of the new universal speech sample was to achieve

- Optimal lingual balance
- Recommended temporal structure level and signal to noise ratio
- Availability of the sample for future full-band audio super-wideband measurement applications.

The new universal speech samples were extensively tested by means of objective measurements and subjective evaluations in order to minimize speaker and sample dependencies as well as to guarantee a good compromise for a “German average”

This investigation can be seen as an example for selection process and can be used as a guideline for selection of universal samples in other languages.

1.1 Speech samples* to be selected and provided

1. A speech sample composed of a male and female talker. This sample should have a good approximation to the requirements given above. The target application is objective speech quality measures.
2. A speech sample spoken by a male and a speech sample (different content) spoken by a female speaker having good approximation to the requirements given above.
3. 14 further speech samples (14 contents spoken by two male and two female speakers).¹

*** In this context, the term ‘speech sample’ always refers to a sentence pair separated by a pause.**

The selection criteria of a universal speech sample should consider different characteristics:

1. Phonological balance: The sample must not show an abnormal distribution of phonemes or word structures compared to average values in German
2. Inconspicuousness in voice production: The selected speaker(s) must neither show abnormal articulation nor unnatural pronunciation. By presenting speech samples to naïve listeners, no under- or over-estimation of voice quality should be observable.

¹ Along with the two samples defined in (2), a set of 16 speech samples will be provided. This variance in content and speakers fulfills the minimum number for set-up of subjective tests according to the P.OLQA specification.

3. Transparency to objective voice quality prediction: The selected speech samples should not be subject to systematic over- or under-prediction of quality by common psycho-acoustic motivated voice quality predictors (i.e. ITU-T P.862.1)

In addition, a series of technical requirements should be met as well

1. The speech recording should follow the constraints for reference speech material given in ITU-T P.800 / P.830.
2. The temporal structure of the test speech sample must follow the requirements given in ITU-T P.800, P.862.3 and the Requirement Specification of P.OLQA. This is mainly given by the use of two sentences separated by a pause of a minimum duration.
For getting a minimum variance in the speaker’s characteristics, a composed sample of a male and a female voice is preferred for objective testing.
3. The speech sample should be made available without a post-applied restriction on bandwidth, with 48 kHz sampling frequency and a minimum resolution of 16bit linear.

Based on this sample, a set of post-processed samples will be provided:

- a. Band-limited: 50 ... 14000 Hz (super-wideband): This sample is for use with the upcoming Recommendation P.OLQA and for subjective tests in super-wideband mode. The bandwidth limitation will not be recognized in practical speech perception, as there are almost no spectral parts outside that band. This sample will be made available in 48 kHz and 32 kHz sampling frequency.
- b. Band-limited: 50 ... 7800 Hz (‘common’ wideband): This sample is for use in common wideband testing cases, corresponding subjective tests and the application of P.862.2 (‘PESQ-WB’). Please note that this sample should not be used as a reference signal for P.OLQA in super-wideband mode. This sample will be made available in 16 kHz sampling frequency.
- c. Band-limited: 50 ... 3700 Hz (‘common’ telephony): This sample is for use in traditional narrowband telephony testing cases, where flat input signals are required. This sample can be used as input signal for P.862.1 as well as for P.OLQA in narrowband mode.
- d. Band-limited acc. to IRS send specification (approx. 250 ... 3500 Hz with pre-emphasis): This sample is for use in common traditional narrow-band telephony testing cases, where IRSsend pre-filtered signals are required. This is the typical use case for narrow-band telephony. This sample can be used as input signal for P.862.1. Some P.OLQA candidate models may also accept this signal for the narrowband operational mode. This sample will be made available in 16 kHz and 8 kHz sampling frequency.

1.2 Available speech recordings in this project

The selection of the speech samples should be based on existing speech recordings in Deutsche Telekom Laboratories and potentially SwissQual.

Telekom Laboratories made recordings for a sub-set of 16 of the so-called ‘Free Berlin Sentences’. This set of sentences is used already

1.1.1 Universal Speech Sample

for a long time in Deutsche Telekom's formal subjective testing in the area of ITU and ETSI. These sentences were recorded in former times in narrowband, now new recordings (with different speakers) were made in full-band audio. The recording conditions can be found in

SwissQual recorded speech material for the ongoing P.OLQA activities by using native German speakers. The contents of the sentences were newly created and correspond to typical telephone conversations. These recordings were also made in full-band audio.

Based on the available recordings, the best fitting samples should be selected. The selection process is sub-divided into three steps:

1. Phonological analysis
2. Application of objective measures
3. Listening test with naïve listeners

2 Phonological analysis

It was agreed to pre-select a set of speech samples out of the available material according to a good match to the phonological constraints.

Four different sentence pairs from the Berlin recordings were selected as sufficient regarding the desired phoneme distribution as well as four sentences pairs from the SwissQual selection.

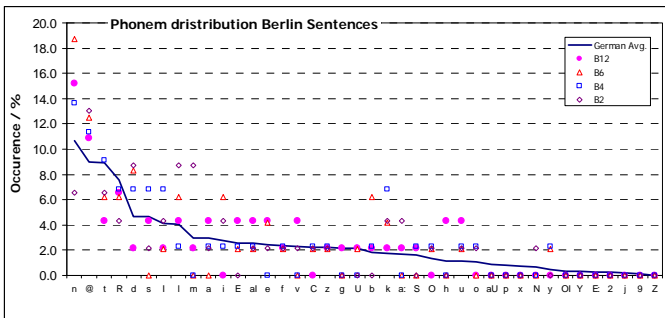


Figure 1: Phoneme distribution Berlin sequences

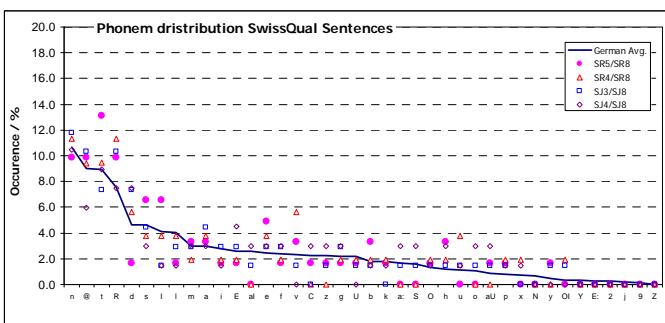


Figure 2: Phoneme distribution SwissQual sequences

3 Objective analysis

In a second step the characteristics of the selected samples were analyzed by common objective tools for speech quality prediction.

Purpose of the evaluation:

It is assumed that the quality for a given processing condition should be in an acceptable range. The obtained quality will not always be the same, since the codec's or other processing components can react differently depending on the speech samples used. Additionally, the individual samples may be more or less affected by bandwidth limitations.

The following evaluation is purely done by objective measures. Differences between the speakers are to be expected. However, whether these differences are actually true or whether the measure over-reacts to some speaker characteristics cannot be proven by that evaluation.

However, an objective measure that has a narrow distribution of the individual speakers can be seen as a good predictor of the average quality, more independent from the actual sample used.

All four of the Berlin sentence pairs were spoken by four male and four female talkers. Two of the SwissQual sentence pairs (SR5/SR8 and SR4/SR8) were spoken by a male speaker, the other two pairs (SR3/SR8 and SR4/SR8) were spoken by a female talker. These samples were processed with the processing conditions listed below.

In total, $4 \times 8 + 2 \times 2 = 36$ speech samples are used for the objective analysis. To examine the dependency of individual samples on common speech processing components, all samples were transmitted over a series of processing conditions:

- **Transparent 50 ... 14'000 Hz**
- **Flat 50 ... 7000 Hz**
- Flat 100 ... 5000 Hz
- IRSSend+IRSrcv (corresponding to narrow band telephony using handsets)
- Flat 50 ... 7000 + 2% Random Packet Loss
- **Flat 50 ... 7000 + 10% Random Packet Loss**
- Flat 50 ... 7000 + AMR-WB at 23.85 kbps
- Flat 50 ... 7000 + AMR-WB at 15.85 kbps
- **Flat 50 ... 7000 + AMR-WB at 12.65 kbps**
- Flat 50 ... 7000 + AMR-WB at 8.65 kbps
- Flat 50 ... 7000 + AMR-WB at 6.65 kbps
- **IRSSend + AMR at 12.2 kbps**
- IRSSend + AMR at 10.2 kbps
- IRSSend + AMR at 7.95 kbps
- IRSSend + AMR at 7.4 kbps
- IRSSend + AMR at 6.7 kbps
- IRSSend + AMR at 5.9 kbps
- IRSSend + AMR at 5.15 kbps
- IRSSend + AMR at 4.75 kbps
- **IRSSend + 3 x AMR at 4.75kbps (as low quality anchor with coding distortions)**

1.1.1 Universal Speech Sample

- 50 ... 14000 Hz MNRU P50 6dB S/N (as low quality anchor with modulated noise)

The objective analysis will be applied to all of these 21 conditions. The six 'best fitting' speakers will be selected and finally checked by a subjective listening test. In the subjective test only the most important subset of conditions can be evaluated. These conditions are marked in **bold** in the list above.

The processed samples were evaluated by the following measures:

- P.862.1 'PESQ' (all super-wideband and wideband samples were low-pass filtered and transformed to 8kHz sampling frequency)
- P.862.2 'PESQ-WB' (all super-wideband samples were low-pass filtered and transformed to 16kHz sampling frequency)
- SQuad08 NB (all super-wideband and wideband samples were low-pass filtered and transformed to 8 kHz sampling frequency)
- SQuad08 SWB (the only available measurement algorithm for super-wideband in this project)

At first the average MOS-LQO over the speech samples per condition as well as the median were calculated. As an example, the graph below shows the average and the Median for P.862.1 'PESQ'.

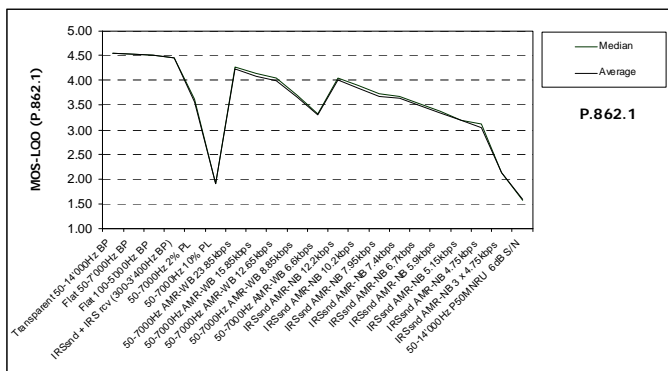


Figure 3: MOS-LQO (P.862.1)

Since there is only a minor difference between both lines, in all further diagrams only the average will be shown for comparison.

At first, the average values (i.e. the averaged MOS predictions over all samples of one condition) are shown per prediction method. This gives an idea about systematic differences between the methods caused by the processing conditions (but still not by speakers or samples).

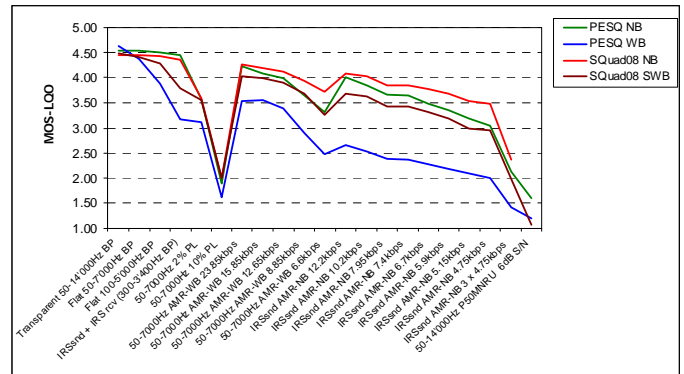


Figure 4: MOS-LQO

It can be seen that the basic 'shape' of the ratings is similar. However, there are biases to observe. It is mainly caused by the different scales of the two narrowband measures ('red' SQuad08 NB and 'green' P.862.1 'PESQ') compared to the super-wideband measure SQuad08 SWB (brown). The P.862.2 'PESQ-WB' (blue) measures show abnormal behavior, the rated scores are far too low.

4 Speaker dependency

4.1 Traditional narrow-band measures

In a next step, the dependencies of the predicted MOS scores on the speaker should be analyzed. For this purpose, the sentences spoken by one speaker are averaged. Thus, we get one line per speaker in the diagram. The average over all speakers is given as reference too.

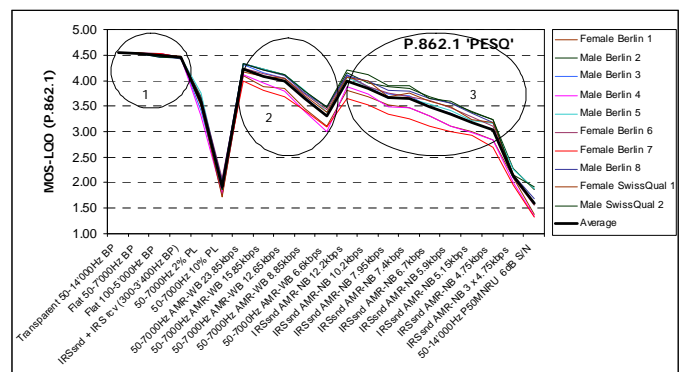


Figure 5: MOS-LQO (P.862.1)

At first, we have to consider that P.862.1 is a narrowband measure. Thus, all samples just limited in audio bandwidth will not be seen as degraded by that measure (area 1). The bandwidth limitation happens outside of the analyzed scope of P.862.1.

In area 2 we see the AMR-WB conditions for decreasing bit-rates. The tendency is clearly visible; however, there are talker dependencies covering a range of up to 0.5 MOS. A similar picture can be seen for the AMR-NB conditions. The bit-rates are well scored; however there is an even higher talker dependency.

From the point of view of use for P.862.1, the talker 'Female Berlin 7' should not be considered in the further selection, due to systematic low scores.

1.1.1 Universal Speech Sample

In a next step, the SQud08 algorithm in narrowband mode should be used for evaluation. At first we have to state that the pure band-width limitations will also not be taken into account due to the narrow-band only evaluation (see also area 1 in Figure 5). Furthermore, it can be seen that SQud08 NB is much less speaker dependent for the AMR-WB as well as for the AMR-NB codec's.

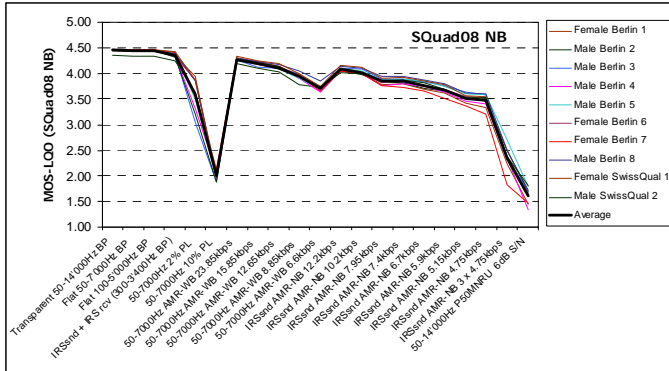


Figure 6: MOS-LQO (SQud 08 NB)

If one speaker should be flagged as a bit problematic, it would be the talker 'Female Berlin 7' as well.

4.2 Traditional wideband measures

When looking at wideband measures, we have at first the wide-band extension of PESQ (P.862.2). One needs to keep in mind that this version is known for inaccurate predictions especially in case of narrowband or intermediate bandwidth conditions. P.862.2 was accepted as a temporary Recommendation to be replaced by a more appropriate successor in a short time. Probably, P.OLQA in super wideband mode will supersede P.862.2 soon.

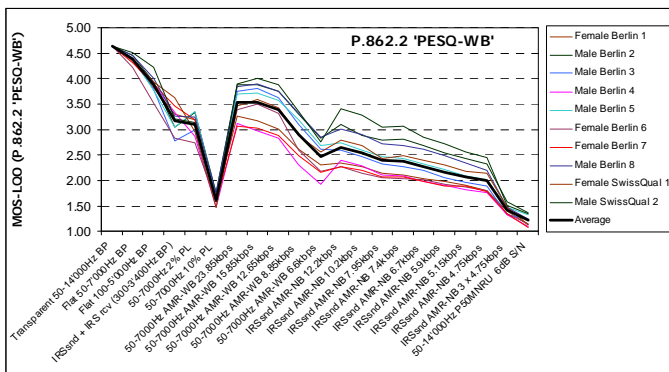


Figure 7: MOS-LQO (P.862.2 'PESQ-WB')

Firstly, we do have a measure that takes into account bandwidth limitations (at least below 8 kHz). This is what we would expect from a wideband measure.

Secondly, the talker dependency can be seen clearly in the predicted MOS scores. This variability already appears in case of plain bandwidth limitations, but even much more for both codec's. By having a closer look at the talker averages, it could be derived that the male speakers (blue/green colors) get better

scores, while female speakers (red/brown) receive lower ones. In principle, this could be explained by the different spectral distribution and the higher amount of higher frequencies in female voice. However, the spread of MOS values appears quite large.

This range of predicted scores is 1.0 MOS. What's more, we have to consider that we have already four sentence pairs (samples) averaged (two for SwissQual talkers) before plotting the results. The 'per-sample' deviation might be even larger.

4.3 Super-wideband measures

The only available measure for super-wideband is SwissQual's P.OLQA candidate SQud08. Compared to the previous measure P.862.2 PESQ WB it considers the entire audio bandwidth up to 14'000Hz as targeted in this project.

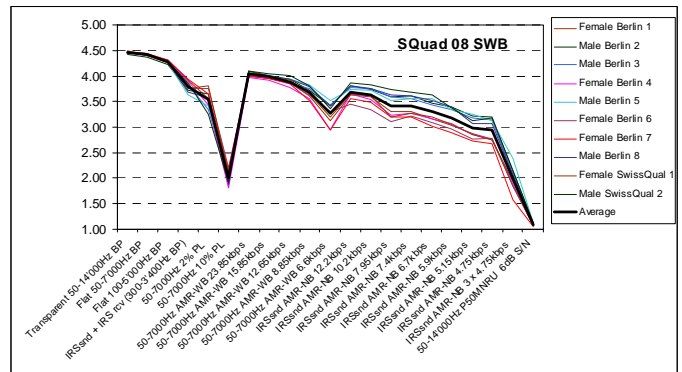


Figure 8: MOS-LQO (SQud08 SWB)

The pure bandwidth reduction shows the expected degradation.

In combination with the codec's we can state that the AMR-WB is much more realistically scored (in comparison to P.862.2 where even AMR-WB at 23.85 just reaches MOS = 3.5). SQud08-SWB goes to MOS = 4.1 here. Under clean conditions, the AMR-WB at 23.05 is even a bit better (MOS = 4.15, not in the graph) as known also from subjective testing.

At higher bitrates there is nearly no talker dependency in the results, but the dependency increases with lower bitrates. This can be explained by the individual amount of higher frequencies in the samples. They are more affected by the compression. Consequently, female voices are more disadvantaged here. By having a closer look again at the talker averages, it could be derived that the male speakers (blue/green colors) get better scores, while female speakers (red/brown) receive lower ones.

Analyzing the results for AMR-NB, we see again that male talkers are resulting in higher scores, whilst the female voices will be scored lower. The most probable explanation is that the male voices are less affected by the bandwidth limitation to narrowband, and less high-frequency content is missing compared to the female voices. Thus, the remaining higher frequencies are also less affected by the compression (AMR inserts more compression artifacts in the higher bands).

5 Selection of speech sample for further analysis

Based on the objective analysis a selection of the most suitable sentence pairs was done. The following sentence pairs were selected for consideration in the listening experiment.

- Berlin Female 1, Sample 06
- Berlin Female 1, Sample 12
- Berlin Male2, Sample 02
- Berlin Male2, Sample 04
- Berlin Male3, Sample 02
- Berlin Male3, Sample 04
- Berlin Female 4, Sample 06
- Berlin Female 4, Sample 12

- SwissQual Male 4/8
- SwissQual Male 5/8
- SwissQual Female 3/8
- SwissQual Female 4/8

6 Selection of 'Mixed' samples

For automated quality test tools in particular, the use of voice samples composed of a male and female voice is interesting. Based on the phoneme distribution and the objective analysis of the speakers and sentence pairs, a sub-selection of six of those samples was made.

Berlin Sentence Pair 12: *'Im Fernsehen wurde alles gezeigt – Alle haben nur einen Wunsch'*

Spoken by:

- Male2 – Female 1
- Male3 – Female 1
- Male2 – Female 4
- Male3 – Female 4

Out of the SwissQual sentences the pairs 8/4 and 5/8 were selected:

'Er wird bald wieder gesund. – Der Storch hat auf dem Kirchendach sein Nest gebaut.'

and

'Du wirst heute noch den Klempner anrufen. – Hast Du Deine Sommerferien schon geplant?'

Both pairs are spoken by SwissQual's male and female talker.

The following graph shows the phoneme distribution of the three selected sentence pairs.

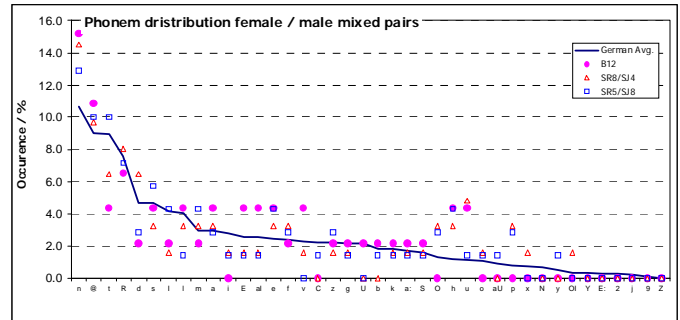


Figure 9: Phoneme distribution female / male mixed pairs

This selection was made based on the best phonological match as well as on the obtained objective results.

The following graphs show the results gained by Squad08 SWB, PESQ WB and -NB and Squad08 NB with these male/female mixed samples.

We should expect a narrower distribution closer to the average for all selected samples. Especially the samples where a male and a female voice are mixed should no longer show the gender dependencies caused by the different spectral distributions.

For P.862.1 'PESQ' in narrowband mode, the results of the selected samples are closer to the average across all processed samples, suggesting a low speaker dependency for the selected sub-set.

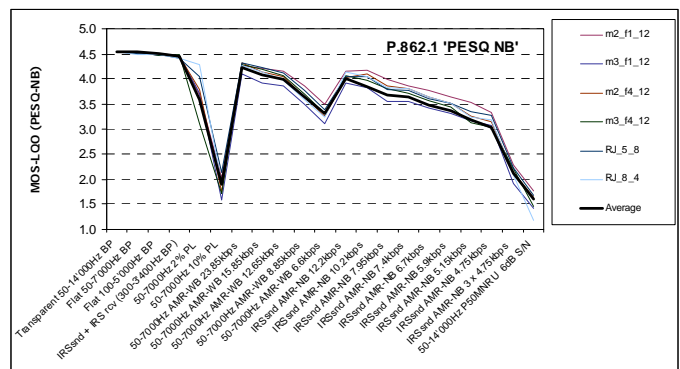


Figure 10: MOS-LQO (PESQ NB)

In comparison to the graphs given in the previous chapter it has to be considered that we have here single sentence pair results whilst before we had sub-averages across sentence pairs per speaker.

This is also the reason for the wide deviation of the 2% packet loss samples. The actual distortion of only 2% packet loss is always subject to the individual sentence structure and the distribution of the loss pattern. Thus, for individual sentence pairs we get a deviation of more than one MOS.

In case of Squad08 NB, the deviation is even smaller. For the common codec conditions, nearly every sample gives identical results to the average over all processed samples. This shows a very low speaker dependency for the chosen mixed male/female sentences.

1.1.1 Universal Speech Sample

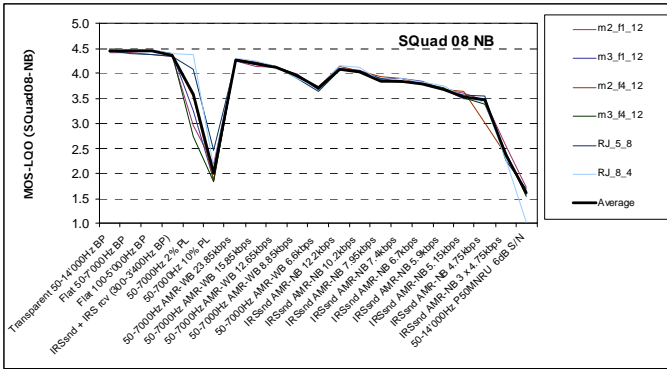


Figure 11: MOS-LQO (SQuad08 NB)

Analyzing P.862.2 'PESQ WB' we see again a wide range of scores depending on the speech sample used. It only takes little advantage of the selection of the best suiting samples and the male/female voice mixtures.

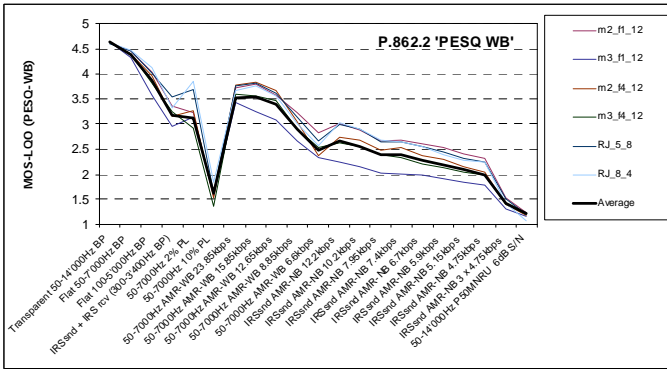


Figure 12: MOS-LQO (PESQ-WB)

Finally, SQuad08 SWB shows a very small sample dependency again. It is not as narrow as for the narrowband mode due to the stronger influence of the higher frequency ranges, which are not considered in narrowband mode.

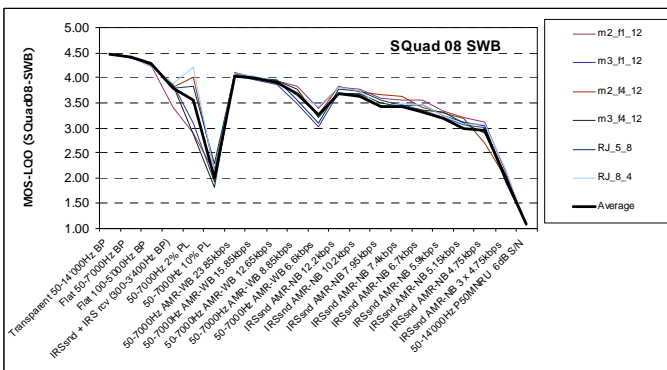


Figure 13: MOS-LQO (SQuad08 SWB)

7 Subjective listening experiment

7.1 Test design

The listening test will consider all 12 selected samples spoken by one speaker as well as the 6 male/female mixed samples. In

total 18 different source speech samples will be used for the experiment.

The tests were done in the listening room of DT in Berlin using Headphones:



Figure 14: Listening Test Set-Up

The playback device was a silent fan less PC with Solid State Drive. The PC is equipped with RME Fireface UC audio interface. The headphones were AKG K271 MKII. The user interface for listening tests is shown in Figure 15 was according ITU-T P.851.



Figure 15: User Interface for ACR Test

Since the experiment should not exceed 1 hour in duration; the number of conditions to be tested is limited. We have chosen six different conditions covering the whole range of quality.

Table 1: Test Conditions

Selection of Conditions	
Condition	Description
Transparent 50-14000 Hz band-pass (super-wideband)	Highest quality in the test
Flat 100-5000 Hz band-pass	Influence of band limitations
50-7000 Hz AMR-WB 12.65 kbps	Typical case for narrow-band cellular telephony
IRSSend AMR12.2 kbps	Typical case for narrow-band cellular telephony
IRSSend 3 x AMR-NB 4.75 kbps	Lower quality with typical codec distortions
50-7000 Hz 10% packet loss	Lower quality with interruptions

Since each source speech sample is processed by each condition, we have $18 \times 6 = 108$ individual files for testing. To increase the number of votes per file each file will be presented twice to each listener in the listening session. Thus, each listener will listen and score 216 files in total.

The experiment is designed as ACR LOT according to ITU T P.800 in a non-fractional design. The scale is using an extended 5-step labels according ITU-T P.851 with the possibility to score on an analogue slider with high precision.

The original outcomes of the subjective test were transformed linearly into the common 5-step ACR MOS scale by a simple equitation.

$$MOS_5 = \left(\frac{MOS_{RAW}}{1000} \cdot 4,5 \right) + 1$$

The complete test plan is available as a separate document.

7.2 Test Results

At first the samples spoke by one male or female speaker are analyzed regarding a speaker dependency. To minimize the dependency on a single sentence pair, the results of both samples spoken by a speaker are averaged.

It can be easily derived from the diagram, that the speaker 'male 3' of the Telekom Laboratories recordings is scored considerable higher than the others. All other speakers form a close group; the slight variation for the 10% packet loss is mainly caused by the individual error patterns hitting the sample structure.

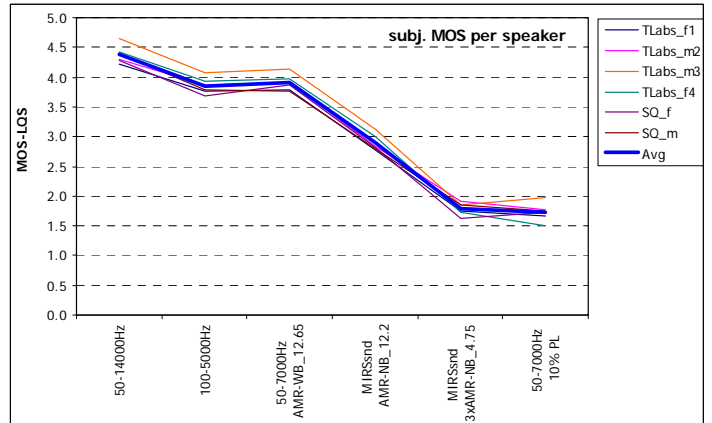


Figure 16: Speaker dependency in general

For the further selection of speech samples it can be assumed that 'male 3' is not getting considered.

The following graph shows the whole set of results for the individual samples tested. Since, there is no averaging anymore for each speaker; we have 12 individual data sets.

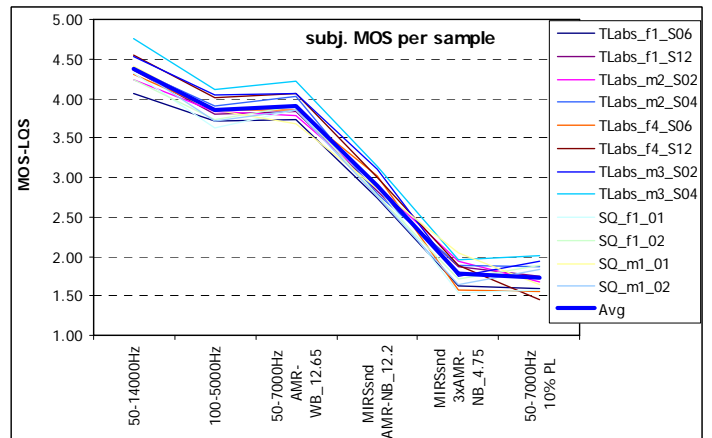


Figure 17: subjective MOS values per sample

The same type of evaluation for the mixed male/female samples is shown in Figure 17

It can be seen that the mixed samples have a much lower deviation to the targeted average value.

1.1.1 Universal Speech Sample

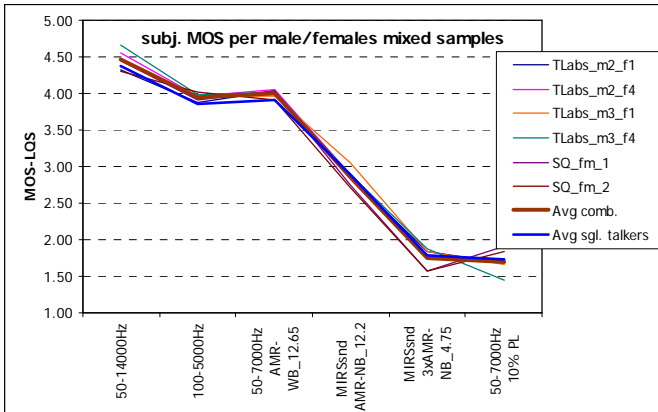


Figure 18: Subjective MOS values for mixed samples

For illustration the confidence of the obtained results are shown on two example sentences only. For illustration, the two most differing samples are used.

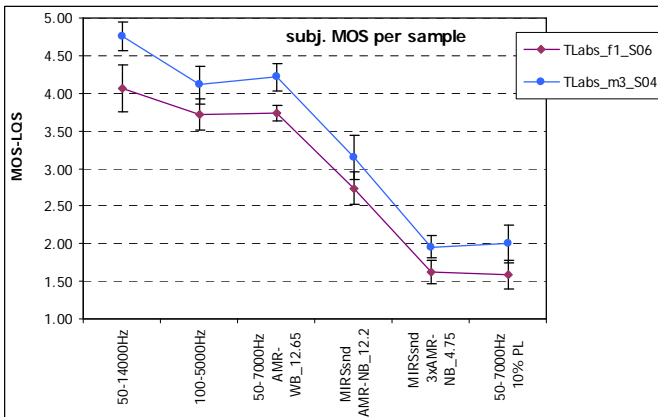


Figure 19: Subjective MOS per sample

It can be seen that from the formal point of view even these samples are statistically equivalent in most of the cases. It can be assumed that the other more narrow results – especially for the male/female mixed samples - can be considered as equivalent. Nevertheless, the best fitting samples to the average should be selected.

8 Selection of speech samples

The selection criterion of the best fitting male, female and mixed (male & female) samples is the smallest deviation to the average across all data by means of r.m.s.e.²

At first the r.m.s.e. values for the individual male and female samples are calculated and presented in Table 2.

² A correlation coefficient doesn't appear as appropriate method, since it removes the offset that is an important figure in our metrics.

Except the sample TLabs_m3_S04 – as already assumed in the previous section – all samples are relatively close to the targeted average.

Nevertheless, the samples TLabs_f1_S12, TLabs_m2_S02 and TLabs_m2_S04 from the Telekom Laboratories recordings as well as SQ_f1_02 and SQ_m1_02 are fitting at best and could be considered as pre-selected reference samples for pure male or female speech.

Table 2: Selection of samples

Selection of Samples	
Sample	r.m.s.e.
TLabs_f1_S06	0.19
TLabs_f1_S12	0.06
TLabs_m2_S02	0.10
TLabs_m2_S04	0.10
TLabs_f4_S06	0.13
TLabs_f4_S12	0.17
TLabs_m3_S02	0.17
TLabs_m3_S04	0.28
SQ_f1_01	0.15
SQ_f1_02	0.11
SQ_m1_01	0.14
SQ_m1_02	0.11

The same evaluation is made for the male/female mixed samples (Table 2).

Table 3: selection of mixed samples

Selection of mixed samples	
Sample	r.m.s.e.
TLabs_m2_f1	0.07
TLabs_m2_f4	0.10
TLabs_m3_f1	0.10
TLabs_m3_f4	0.18
SQ_fm_1	0.15
SQ_fm_2	0.14

Consequently, the mixed sample consist of the two best individual speakers shows also the best fit to the targeted average (TLabs_m2_f1).

9 Comparison to objective scores

Finally, it should be confirmed that the selected speech samples don't show abnormal behavior by use of objective measures.

For all three measures the r.m.s.e. to the average objective score across all samples is calculated. This evaluation should show that the selected samples don't show abnormal behavior in contrast to others.

Table 4: Objective results for selected samples

Objective results for selected samples	r.m.s.e.				
	Sample	SQuad08-SWB	PESQ-WB	SQuad08-NB	PESQ-NB
TLabs_f1_S06	0.06	0.12	0.08	0.18	
TLabs_f1_S12	0.07	0.09	0.11	0.13	
TLabs_m2_S02	0.06	0.04	0.10	0.05	
TLabs_m2_S04	0.18	0.07	0.14	0.11	
TLabs_f4_S06	0.15	0.11	0.07	0.16	
TLabs_f4_S12	0.12	0.09	0.13	0.13	
TLabs_m3_S02	0.17	0.14	0.09	0.21	
TLabs_m3_S04	0.16	0.10	0.12	0.14	
SQ_f1_01	0.08	0.07	0.08	0.11	
SQ_f1_02	0.14	0.03	0.12	0.05	
SQ_m1_01	0.11	0.07	0.05	0.11	
SQ_m1_02	0.08	0.07	0.07	0.10	

For this kind of comparison we have to take into account that only SQuad08-SWB is a full super-wideband measure that considers the entire range of conditions in the subjective test. All other measures apply internal band-passes either to 8 kHz (PESQ-WB) or even to 4 kHz (SQuad08-NB, PESQ-NB). For that reason they can't differentiate between narrowband and wide-band conditions. The wideband and super-wideband conditions are scored mostly in the higher saturation of the scale (see Figure 5 area1 in chapter 5). The r.m.s.e. values are influenced by this saturation and drawn in grey for information only in Table 4.

Based on these results, the samples TLabs_f1_S12 and TLabs_m2_S02 are selected as the best fitting samples to the overall averages in the subjective test as well as by objective measures.

The following Table 5 shows the analysis for the mixed male / female speech samples.

Table 5: Selection of male/female mixed samples

Selection of Conditions	r.m.s.e.			
	Sample	SQuad08 SWB	PESQ-WB	SQuad08 NB
TLabs_m2_f1	0.07	0.09	0.13	0.14
TLabs_m2_f4	0.09	0.03	0.06	0.04
TLabs_m3_f1	0.04	0.12	0.05	0.17
TLabs_m3_f4	0.08	0.05	0.07	0.07
SQ_fm_1	0.16	0.10	0.22	0.15
SQ_fm_2	0.10	0.04	0.09	0.06

Also here the pre-selected TLabs_f1_m2 shows a good compromise for the objective measures.

Thus, the mixed sample combines the two talkers selected for the male and the female sample too. This gives also a high grade of consistency in the selection process.

10 Limitations due to experimental design

Super-wideband listening tests combine usually multiple quality dimensions for scoring. In comparison to narrow-band tests where usually only coding distortions (and background noises) are in the focus, in super-wideband tests also various types of band-width limitations have to be scored.

The more individual quality dimensions are in the subjective experiment the more important becomes a balanced test design. That means there should be no over- or under-representation of an individual distortion. ITU-T recommended strict constraints for those super-wideband experiments within the P.OLQA project. The first experiments were conducted and discussed in the last meeting of ITU-T SG12 (November 2009). A simple narrow-band telephony band-pass is scored in these P.OLQA tests with around 3.6.

The experiment conducted here could not fully meet these constraints due to the few conditions tested. It has to be stated that the amount of narrow-band conditions (2) is too low in contrast to wide-band and super-wideband signals (4). The band-width limitation is the most clear perceptible distortion in this test. It dominates the quality perception. That can cause a more pessimistic score of the narrow-band conditions in this test.

It should be noted that the narrow-band conditions (AMR-NB as well as band-pass 100...5000Hz) are rated lower in the subjective listening test as by SQuad08-SWB. SQuad08-SWB is trained on the P.OLQA experiments conducted by ITU-T and predicts closer to these values.

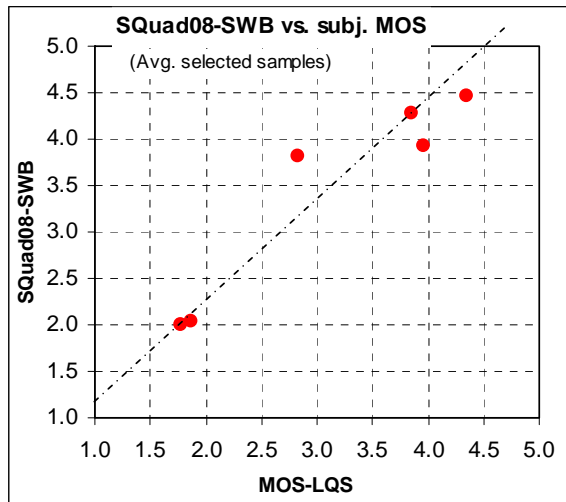


Figure 20: SQuad08-SWB vs. subjective MOS

11 Objective example scores for the selected speech samples

For illustration, all of the objective scores of the selected speech samples are shown. The following graphs show a sub-set of the results drawn in sections 5 and 6.

Those conditions tested in the subjective experiment too are marked by arrows.

The average (in bold) gives the average over all tested samples for each condition. The individual lines show the compliance to that average. It means how representative the individual samples in contrast to the average are across a higher number of samples.

The SQuad08-SWB that is the only super-wideband model in this investigation shows a very narrow distribution and almost no dependency on the individual samples. Only in case of AMR-NB the male sample appears a bit advantaged. Consequently, the mixed sample consisting of one sentence of that talker too is also bit advantaged compared to the average. However, the rank-order of the individual bit-rates of all codec's can be reproduced pretty well by the objective scores.

As already discussed, the AMR-NB samples are scored higher than in the subjective test.

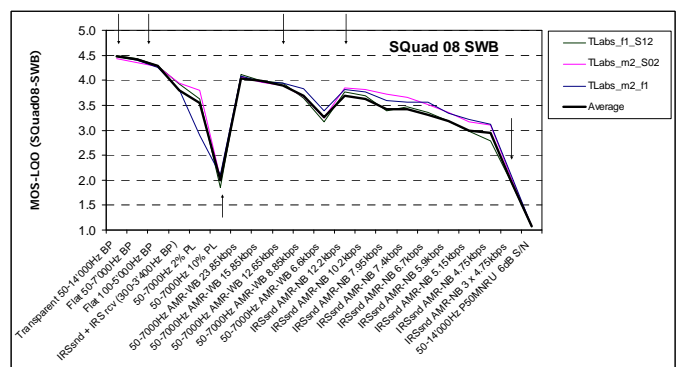


Figure 21: SQuad08 values

1.1.1 Universal Speech Sample

The other objective models have restrictions in their analysis bandwidth. Therefore bandwidth-limitations will be less counted than for super-wideband models those compare to a wider reference signal.

The method according to P.862.2 'PESQ-WB' is using a bandwidth up to 8 kHz. Thus, the super-wideband condition and the 100...5000 band-pass can still be differentiated. However, the AMR-WB conditions appear a bit low, in the subjective (super-wideband) test a MOS of around 4.0 was reached for AMR 12.65kbps while PESQ-WB shows only 3.5 even the band-width limitation is not counted (PESQ-WB compares only to a 8 kHz reference, which is almost the same as the 7kHz AMR-WB).

In addition the AMR-NB is clearly lower. It matches with the results in this test, however, in a wide-band context it should be rated significantly higher.

Finally, there is still a talker dependency for all AMR conditions.

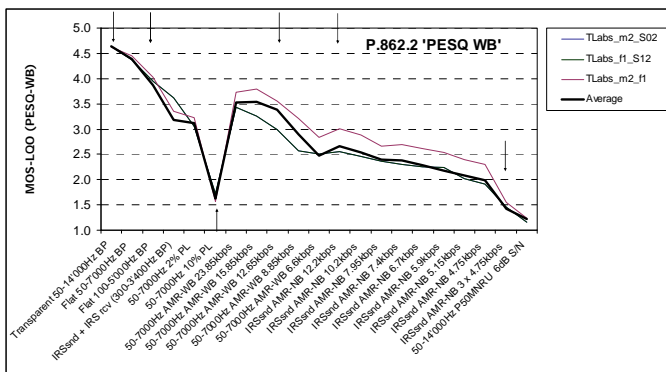


Figure 22: Talker dependency for AMR

The following two methods, P.862.1 PESQ-NB' as well as Squad08-NB compare the signals to be tested only to a 4 kHz reference as typical for traditional telephony.

Consequently, there is no differentiation between the 14 kHz and the 5 kHz conditions anymore.

The following graph for Squad08-NB shows almost no sample dependency. All values are widely identical with the average across all tested samples.

As usual for narrow-band tests, the AMR-NB 12.2 condition is scored slightly above MOS = 4.0. The result is almost the same as for AMR-WB 12.65 after imitation to 8 kHz.

The qualitative rank-order for the individual bitrates can be reproduced clearly for AMR codec types.

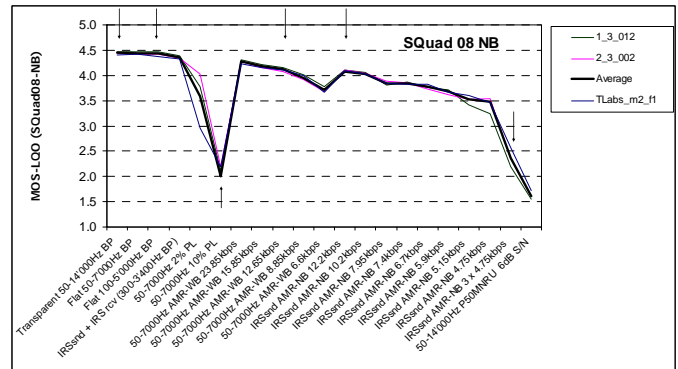


Figure 23: Sample dependency measured with Squad08

Finally, the common narrow-band version of P.862 'PESQ-NB' is analyzed in the same way as well.

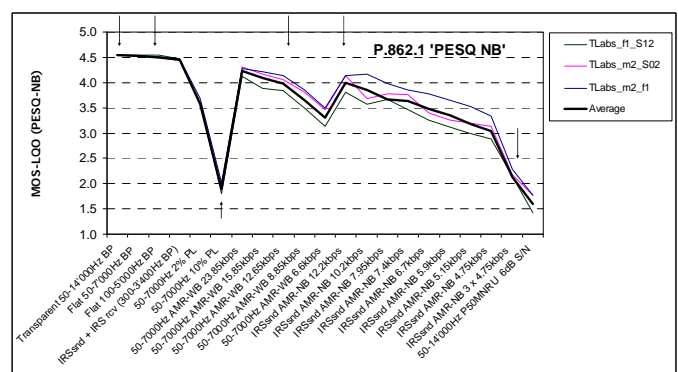


Figure 24: Speaker dependency measured with PESQ-NB

Also here the differentiation between super-wideband and 100...500Hz is not possible anymore. In average the AMR-NB 12.2 also reaches the MOS = 4.0 as usual for narrow-band investigations. The same is for the AMR-WB 12.65. In average the qualitative rank-order of the AMR-bitrates can be reproduced as well. However, the sample dependency in case of AMR coding is higher than for Squad08-NB.

12 Post-processing of the selected file(s)

The final processing of the sample was done after the listening tests according requirements of ITU-T P.862.3 and ETSI TR 102 506.

There are stated several requirements which should be fulfilled by the sample:

- Length of signal: 8 ...30 sec
- Minimum amount of active speech: 3.2 sec
- Silent period between (two) sentences: > 1 sec; < 2s³

³ The most important reason for this requirement is the method that P.862 uses for setting the silence thresholds. P.862 only considers the pause in the middle for the threshold adjustment. A pause that is too short leads to miss-adjustment of the speech-pause threshold and may affect the quality prediction.

1.1.1 Universal Speech Sample

- Leading silence: 0,5 ... 2 sec
- Trailing silence: 0,5 ... 2 sec
- Active speech: 40 ... 80 %
(includes leading and trailing silence)⁴
- Active speech level: -26 dBov (-30 dBov)
- Noise floor: -75 dBov
- Pre-Filtering: according to application listed below

Because of restrictions of some test systems which are used at T-Mobile, the maximum length of the sequence should not exceed 10 sec. However, for minimizing the speaker dependency, a sample with mixed male and female talkers is desired.

To meet the different requirements regarding the sample length, we will provide the following sample combinations:

- 1) Short sample for automated devices
 - a. One sentence male / One sentence female
 - b. Sample length 6s
- 2) Short sample for listening tests (male) (P.800)
 - a. Two sentences male
 - b. Sample length 8s
- 3) Short sample for listening tests (female) (P.800)
 - a. Two sentences female
 - b. Sample length 8s
- 4) Long sample for automated devices
 - a. Two sentences male / two sentences female
 - b. Sample length 10s

Each of the sample combinations will be provided in different formats to be used in different measurement applications. The targeted measurement applications are:

- 1) Full-band applications (to 20kHz)
 - a. Sampling frequency 48kHz
 - b. No band limitation applied except very low frequency cut-off
- 2) Super-wideband 50...14'000Hz application acc. to ITU-T P.OLQA
 - a. Sampling frequencies 48 kHz and 32 kHz
 - b. 50...14'000 Hz high-quality band-pass
(acc. to P:OLQA specification for SWB mode)

⁴ The speech activity is widely irrelevant, since it depends highly on the leading and trailing silences. Silent periods will neither be considered in subjective test nor by P.862.

- 3) Common wide-band measures 50 ... 7000 Hz⁵
 - a. Sampling frequency 16 kHz
 - b. Wide-band channel filter acc. P.341
 - c. IRS(send) mod acc. P.830 + wide-band channel filter acc. P.341⁶
- 4) Narrow-band telephony
 - a. Sampling frequency 8 kHz
 - b. Only PCM channel filter acc. P.341 (equivalent to TMD_German_5s_8kHz_16bit.wav)
 - c. IRS(send) mod acc. P.830 + PCM channel filter acc. G.712
(as specified in P.862.3)

Each sample is provided in PCM raw format as well as with WAV header. The narrowband signals (item 4) are further available in A-Law and μ -Law PCM coding acc. to G.711.

All flavors of the sample will be derived stepwise from the same high-quality raw recording.

The processing was done by means of the standard ITU-T tools which are described and published as Recommendation ITU-T G.191. For some format conversions the Afsp library was used. The checksums were calculated with the Microsoft tool "File Checksum Integrity Verifier" (FCIV). The xml file with md5 checksums is delivered together with the audio files.

⁵ The continuation of test in the common wide-band mode is under discussion in ITU-T. This mode might be superseded by measurements in super-wideband mode.

⁶ Basically a flat band-pass 50 ... 7000Hz.

1.1.1 Universal Speech Sample

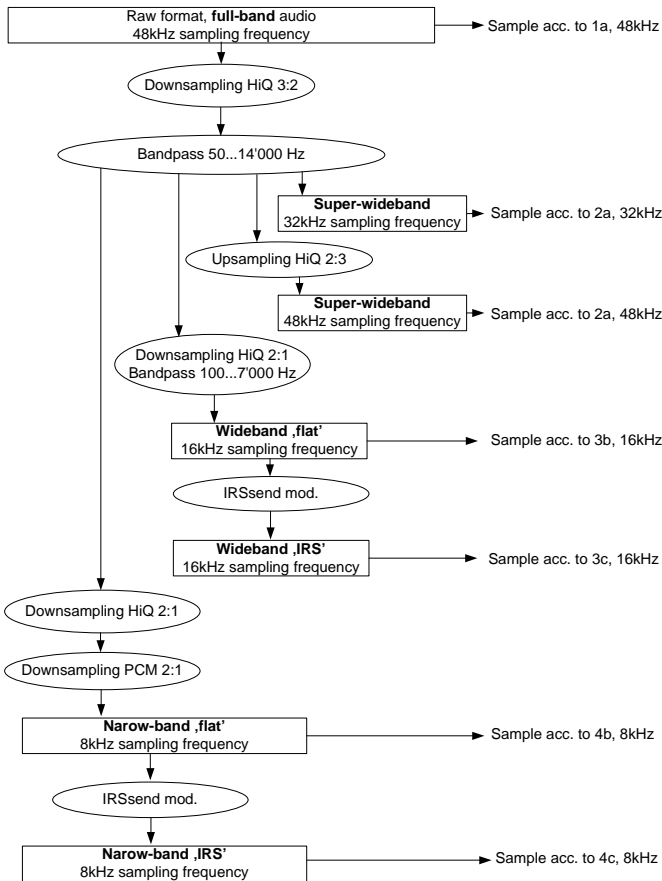


Figure 25: Processing steps

The output files which are available for certain application scenarios are as shown in the following table.

13 File naming convention

List of samples which are delivered as appendix to this paper:

- German_male_2010
- German_female_2010
- German_mixed_6s_2010
- German_mixed_10s_2010

In Table 6, there are shown the filename and appropriate use cases for the several files.

Table 6: Description of delivered samples

Files	
Filename	Description
*_full_48k	Full Bandwidth 20...20000 Hz, 48 kHz sampling frequency, to be used for full-band audio testing and as source material for further processing
*_SWB_48k	Band-pass to 50...14000 Hz, 48 kHz Sampling frequency according SWB specification to be used as source and reference sample for SWB testing as well as for P.OLQA SWB mode
*_SWB_32k	Band-pass to 50...14000 Hz, 32 kHz Sampling frequency according SWB specification to be used as source and reference sample for SWB testing as well as for P.OLQA SWB mode (if actual model supports 32 kHz sampling frequency)
*_WB_16k	Band-pass 150...7000 Hz, 16 kHz Sampling frequency according to P.341 'Transmission characteristics for wideband', to be used as source signal for WB testing.
*_WB_IRSm_16k	Band-pass 150...7000 Hz + IRSmod filter, 16 kHz Sampling frequency according to P.341 'Transmission characteristics for wideband', to be used as source signal for WB testing if IRS prefiltering is required.
*_NB_G712_08k	Band-pass 150...3500 Hz, 8 kHz Sampling frequency according to G.712 'Channel filter', to be used as source signal for NB testing. This signal should be used if the terminal or terminal model is part of transmission chain.
*_NB_IRS_08k	Band-pass 150...3500 Hz + IRS filter, 8 kHz Sampling frequency according to G.712 'Channel filter', to be used as source signal for NB testing. This signal should be used if no terminal or terminal model is part of transmission chain i.e. connection to network termination points or equivalent digital interfaces.

14 Appendix 1 Batch procedure for file processing

The Batch procedure for file processing was as follows. The routines are originated from ITU-T G.191 STL and AfsP (Audio File Programs and Routines 8.2, by Peter Kabal 2006)

```
@echo on
: Reference Files universelles Sprachsample erzeugen...
: Input file ist 48 kHz, 16 bit, mono im wav format

: 1. Full Band applications,
copyaudio -F "noheader" .\audio\%1.wav .\audio\%1_temp1.raw
filter DC .\audio\%1_temp1.raw .\audio\%1_temp2.raw
sv56demo -lev -26 -sf 48000 .\audio\%1_temp2.raw .\audio\%1_full_48k.raw
copyaudio -t "noheader" -P "integer16, 0, 48000, native, 1, default" -F "WAVE" -D
"integer16" .\audio\%1_full_48k.raw .\audio\%1_full_48k.wav
del .\audio\%1_temp1.raw
del .\audio\%1_temp2.raw

: 2. Superwideband applications
copyaudio -F "noheader" .\audio\%1.wav .\audio\%1_temp1.raw
filter DC .\audio\%1_temp1.raw .\audio\%1_temp2.raw
filter -up HQ2 .\audio\%1_temp2.raw .\audio\%1_temp3.raw
filter -down HQ3 .\audio\%1_temp3.raw .\audio\%1_temp4.raw
filter 14kbp .\audio\%1_temp4.raw .\audio\%1_temp5.raw
sv56demo -lev -26 -sf 32000 .\audio\%1_temp5.raw .\audio\%1_SWB_32k.raw
copyaudio -t "noheader" -P "integer16, 0, 32000, native, 1, default" -F "WAVE" -D "inte-
ger16" .\audio\%1_SWB_32k.raw .\audio\%1_SWB_32k.wav
filter -up HQ3 .\audio\%1_SWB_32k.raw .\audio\%1_temp6.raw
filter -down HQ2 .\audio\%1_temp6.raw .\audio\%1_temp7.raw
sv56demo -lev -26 -sf 48000 .\audio\%1_temp7.raw .\audio\%1_SWB_48k.raw
copyaudio -t "noheader" -P "integer16, 0, 48000, native, 1, default" -F "WAVE" -D "inte-
ger16" .\audio\%1_SWB_48k.raw .\audio\%1_SWB_48k.wav
del .\audio\%1_temp1.raw
del .\audio\%1_temp2.raw
del .\audio\%1_temp3.raw
del .\audio\%1_temp4.raw
del .\audio\%1_temp5.raw
del .\audio\%1_temp6.raw
del .\audio\%1_temp7.raw

: 3. Wideband applications
filter -down HQ2 .\audio\%1_SWB_32k.raw .\audio\%1_temp1.raw
filter P341 .\audio\%1_temp1.raw .\audio\%1_temp2.raw
sv56demo -lev -26 -sf 16000 .\audio\%1_temp2.raw .\audio\%1_WB_16k.raw
copyaudio -t "noheader" -P "integer16, 0, 16000, native, 1, default" -F "WAVE" -D
"integer16" .\audio\%1_WB_16k.raw .\audio\%1_WB_16k.wav
filter -mod IRS16 .\audio\%1_temp2.raw .\audio\%1_temp3.raw
sv56demo -lev -26 -sf 16000 .\audio\%1_temp3.raw .\audio\%1_WB_IRSm_16k.raw
copyaudio -t "noheader" -P "integer16, 0, 16000, native, 1, default" -F "WAVE" -D
"integer16" .\audio\%1_WB_IRSm_16k.raw .\audio\%1_WB_IRSm_16k.wav
del .\audio\%1_temp1.raw
del .\audio\%1_temp2.raw
del .\audio\%1_temp3.raw
```

: 4. Narrowband applications

```
filter -down HQ2 .\audio\%1_SWB_32k.raw .\audio\%1_temp1.raw
filter -down PCM .\audio\%1_temp1.raw .\audio\%1_temp2.raw
sv56demo -lev -26 -sf 8000 .\audio\%1_temp2.raw .\audio\%1_NB_G712_08k.raw
copyaudio -t "noheader" -P "integer16, 0, 8000, native, 1, default" -F "WAVE" -D "integer16" .\audio\%1_NB_G712_08k.raw .\audio\%1_NB_G712_08k.wav
filter IRS8 .\audio\%1_temp2.raw .\audio\%1_temp3.raw
sv56demo -lev -26 -sf 8000 .\audio\%1_temp3.raw .\audio\%1_NB_IRS_08k.raw
copyaudio -t "noheader" -P "integer16, 0, 8000, native, 1, default" -F "WAVE" -D "integer16" .\audio\%1_NB_IRS_08k.raw .\audio\%1_NB_IRS_08k.wav
del .\audio\%1_temp1.raw
del .\audio\%1_temp2.raw
del .\audio\%1_temp3.raw
```

15 Appendix 2 Recording Conditions at Telekom Laboratories

The recordings were made in the big anechoic room of Technical University Berlin. As shown in Figure 26 there were used 2 two different microphones, an omni-directional and a cardioid condenser microphone by Schoeps.

For further processing the recordings of the omni-directional microphone were used. The other components were:

- Microphone Preamplifier Studer D19,
- Sound Board RME Digi96 with digital input
- PC with Adobe Audition for postprocessing of the original recordings.



Figure 26: Recording at TU Berlin

16 List of Abbreviations

PESQ	Perceptual Evaluation of Speech Quality
P.OLQA	Objective Listening Quality Assessment
MOS	Mean Opinion Score
WB	Wideband
NB	Narrowband
SWB	Super Wideband
ACR	Absolute Category Rating
LOT	Listening Only Test
r.m.s.e.	root mean squared error

17 Index of figures

Figure 1: Phoneme distribution Berlin sequences	5
Figure 2: Phoneme distribution SwissQual sequences.....	5
Figure 3: MOS-LQO (P.862.1)	6
Figure 4: MOS-LQO.....	6
Figure 5: MOS-LQO (P.862.1)	6
Figure 6: MOS-LQO (SQuad 08 NB)	7
Figure 7: MOS-LQO (P.862.2 'PESQ-WB')	7
Figure 8: MOS-LQO (SQuad08 SWB)	7
Figure 9: Phoneme distribution female / male mixed pairs	8
Figure 10: MOS-LQO (PESQ NB)	8
Figure 11: MOS-LQO (SQuad08 NB)	9
Figure 12: MOS-LQO (PESQ-WB).....	9
Figure 13: MOS-LQO (SQuad08 SWB).....	9
Figure 14: Listening Test Set-Up	9
Figure 15: User Interface for ACR Test.....	9
Figure 16: Speaker dependency in general.....	10
Figure 17: subjective MOS values per sample	10
Figure 18: Subjective MOS values for mixed samples	11
Figure 19: Subjective MOS per sample	11
Figure 20: SQuad08-SWB vs. subjective MOS.....	13
Figure 21: SQuad08 values	13
Figure 22: Talker dependency for AMR	14
Figure 23: Sample dependency measured with SQuad08	14
Figure 24: Speaker dependency measured with PESQ-NB	14
Figure 25: Processing steps.....	16
Figure 26: Recording at TU Berlin	19

18 Index of tables

Table 1: Test Conditions.....	10
Table 2: Selection of samples.....	11
Table 3: selection of mixed samples	12
Table 4: Objective results for selected samples	12
Table 5: Selection of male/female mixed samples.....	13
Table 6: Description of delivered samples	16

19 References

ITU-T, 'Recommendation ITU-T P.851', Geneva 2003

ITU-T, 'Recommendation ITU-T P.800', Geneva 2001

ITU-T, 'Recommendation ITU-T P.862.3', Geneva 2003

ETSI, 'Technical Report TR 102 506'

ITU-T, Recommendation G.191 (09/05) "Software tools for speech and audio coding standardization"

P. Kabal, AFsp Library v8r2, programs and routines. <http://www-mmsp.ece.mcgill.ca/Documents/Downloads/AFsp/>

Microsoft, "File Checksum Integrity Verifier", <http://support.microsoft.com/kb/841290/de>

1.1.1 Universal Speech Sample

Publisher:

Deutsche Telekom AG
Laboratories
Ernst-Reuter -Platz 7
D-10587 Berlin
Telefon: +49 30 8353-58555
www.laboratories.telekom.com

Authors: Ulf Wüstenhagen ulf.wuestenhagen@telekom.de
Jens Berger jens.berger@swissqual.com

© 2010 Deutsche Telekom Laboratories

The information contained in this document represents the current view of the authors on the issues discussed as of the date of publication. This document should not be interpreted to be a commitment on the part of Deutsche Telekom Laboratories, and Deutsche Telekom Laboratories cannot guarantee the accuracy of any information presented after the date of publication.

This White Paper is for informational purposes only. Deutsche Telekom Laboratories makes no warranties - express, implied, or statutory - as to the information in this document.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording or otherwise), or for any purpose, without the express written permission of Deutsche Telekom Laboratories.

Deutsche Telekom Laboratories may have patents, patent applications, trademarks, copyrights or other intellectual property rights covering the subject matter in this document. Except as expressly provided in any written license agreement from Deutsche Telekom Laboratories, the furnishing of this document does not give you any license to these patents, trademarks, copyrights or other intellectual property.

SwissQual may have patents, patent applications, trademarks, copyrights or other intellectual property rights covering the subject matter in this document. When you refer to a SwissQual technology or product, you must acknowledge the respective text or logo trademark somewhere in your text.

SwissQual® and SQuad® as well as the following logos are registered trademarks of SwissQual AG.