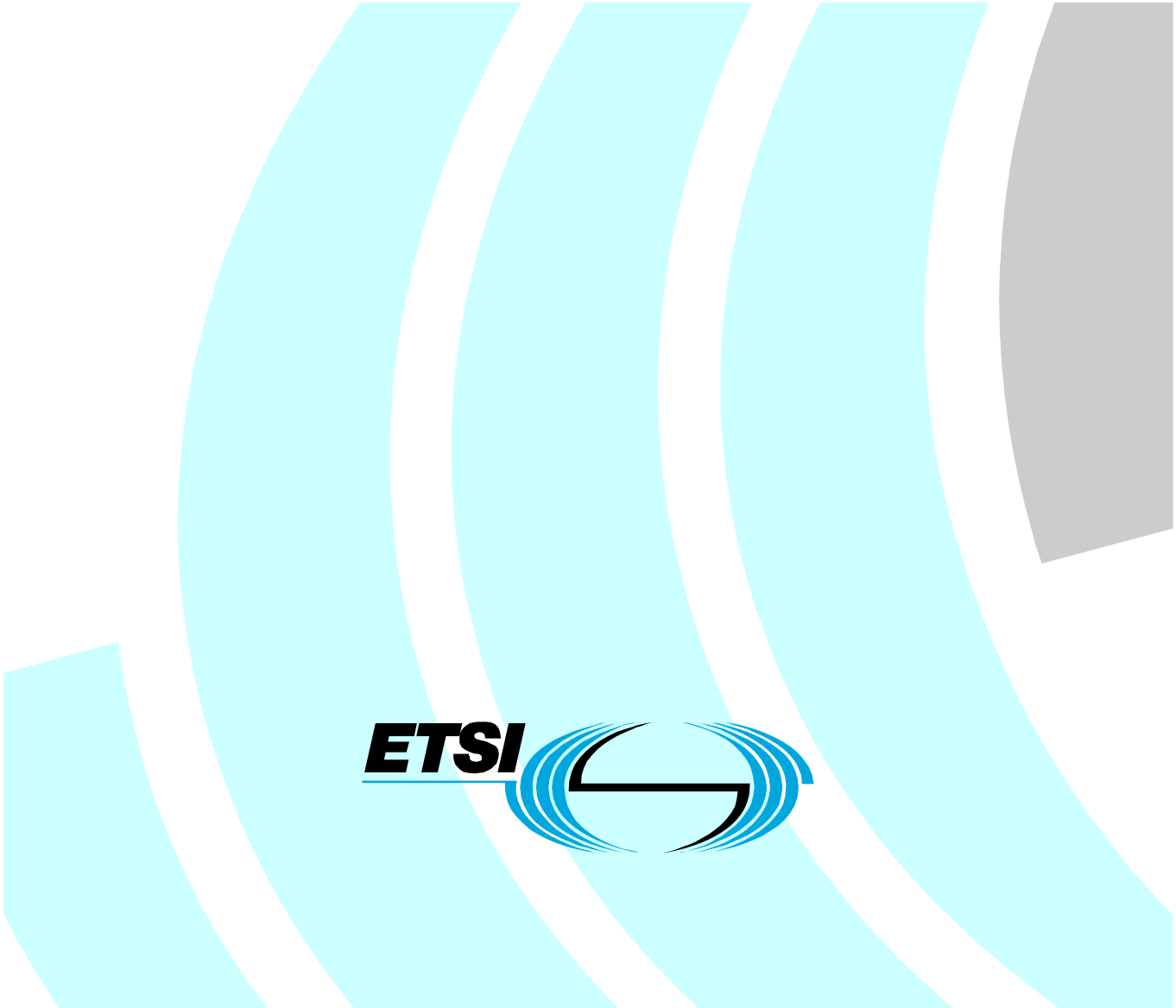
ETSI EG 202 396-3 V1.2.1 (2009-01)

ETSI Guide

Speech Processing, Transmission and Quality Aspects (STQ);
Speech Quality performance
in the presence of background noise
Part 3: Background noise transmission -
Objective test methods



Reference
REG/STQ-00124

Keywords
noise, QoS, quality, speech

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 ETSI Guide (EG) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

The present document is a deliverable of ETSI Specialized Task Force (STF) 294 entitled: "Improving the quality of eEurope wideband speech applications by developing a performance testing and evaluation methodology for background noise transmission".

The present document is part 3 of a multi-part deliverable covering speech quality performance in the presence of background noise, as identified below:

Part 1: "Background noise simulation technique and background noise database";

Part 2: "Background noise transmission - Network simulation - Subjective test database and results";

Part 3: "Background noise transmission - Objective test methods".

1 Scope

The present document aims to identify and define testing methodologies which can be used to objectively evaluate the performance of narrowband and wideband terminals and systems for speech communication in the presence of background noise.

Background noise is a problem in mostly all situations and conditions and need to be taken into account in both, terminals and networks. The present document provides information about the testing methods applicable to objectively evaluate the speech quality in the presence of background noise. The present document includes:

- The description of the experts post evaluation process chosen to select the subjective test data being within the scope of the objective methods.
- The results of the performance evaluation of the currently existing methods described in ITU-T Recommendation P.862 [i.16], [i.17] and in TOSQA2001 [i.19] which is chosen for the evaluation of terminals in the framework of ETSI VoIP speech quality test events [i.8], [i.9], [i.10] and [i.11].
- The method which is applicable to objectively determine the different parameters influencing the speech quality in the presence of background noise taking into account:
  - the speech quality;
  - the background noise transmission quality;
  - the overall quality.
- The document is to be used in conjunction with:
  - EG 202 396-1 [i.1] which describes a recording and reproduction setup for realistic simulation of background noise scenarios in lab-type environments for the performance evaluation of terminals and communication systems.
  - EG 202 396-2 [i.2] which describes the simulation of network impairments and how to simulate realistic transmission network scenarios and which contains the methodology and results of the subjective scoring for the data forming the basis of the present document.
  - French speech sentences as defined in ITU-T Recommendation P.501 [i.13] for wideband and English speech sentences as defined in ITU-T Recommendation P.501 [i.13] for narrowband.

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.

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

2.1 Normative references

The following referenced documents are indispensable for the application of the present document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

Not applicable.

2.2 Informative references

The following referenced documents are not essential to the use of the present document but they assist the user with regard to a particular subject area**.** For non-specific references, the latest version of the referenced document (including any amendments) applies.

[i.1] ETSI EG 202 396-1: "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise; Part 1: Background Noise Simulation Technique and Background Noise Database".

[i.2] ETSI EG 202 396-2: "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise; Part 2: Background Noise Transmission - Network Simulation - Subjective Test Database and Results".

[i.3] ITU-T Recommendation P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[i.4] ITU-T Recommendation P.800: "Methods for subjective determination of transmission quality".

[i.5] ITU-T Recommendation P.831: "Subjective performance evaluation of network echo cancellers".

[i.6] Genuit, K.: "Objective Evaluation of Acoustic Quality Based on a Relative Approach", InterNoise '96, Liverpool, UK.

[i.7] ITU-T Recommendation SG 12 Contribution 34: "Evaluation of the quality of background noise transmission using the "Relative Approach"".

[i.8] ETSI 2nd Speech Quality Test Event: "Anonymized Test Report", ETSI Plugtests, HEAD acoustics, T-Systems Nova.

NOTE: Available at: http://www.etsi.org/WebSite/OurServices/Plugtests/History.aspx.
Also available as ETSI TR 102 648-3.

[i.9] ETSI 3rd Speech Quality Test Event: "Anonymized Test Report "IP Gateways"".

NOTE: Available at: http://www.etsi.org/WebSite/OurServices/Plugtests/History.aspx.

[i.10] ETSI 3rd Speech Quality Test Event: "Anonymized Test Report "IP Phones"".

[i.11] ETSI 4th Speech Quality Test Event: "Anonymized Test Report "IP Gateways and IP Phones"".

NOTE: Available at: http://www.etsi.org/WebSite/OurServices/Plugtests/History.aspx.

[i.12] F. Kettler, H.W. Gierlich, F. Rosenberger: "Application of the Relative Approach to Optimize Packet Loss Concealment Implementations", DAGA, March 2003, Aachen, Germany.

[i.13] ITU-T Recommendation P.501: "Test Signals for Use in Telephonometry".

[i.14] R. Sottek, K. Genuit: "Models of Signal Processing in human hearing", International Journal of Electronics and Communications (AEÜ)" vol. 59, 2005, p. 157-165.

NOTE: Available at: http://www.elsevier.de/aeue.

[i.15] SAE International - Document 2005-01-2513: "Tools and Methods for Product Sound Design of Vehicles" R. Sottek, W. Krebber, G. Stanley.

[i.16] 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".

[i.17] ITU-T Recommendation P.862.1: "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[i.18] ITU-T Recommendation P.862.2: "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs".

[i.19] ITU-T Recommendation SG 12 Contribution 19: "Results of objective speech quality assessment of wideband speech using the Advanced TOSQA2001".

[i.20] ITU-T Recommendation G.722: "7 kHz audio-coding within 64 kbit/s".

[i.21] ITU-T Recommendation G.722.2: "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)".

[i.22] ITU-T Recommendation P.56: "Objective measurement of active speech level".

[i.23] ITU-T Recommendation P.57: "Artificial ears".

[i.24] M. Spiegel: "Theory and problems of statistics", McGraw Hill, 1998.

[i.25] R.A. Fisher: "Statistical methods and scientific inference", Oliver and Boyd, 1956.

[i.26] M. Kendall: "Rank correlation methods", Charles Griffin & Company Limited, 1948.

[i.27] Sottek, R.: "Modelle zur Signalverarbeitung im menschlichen Gehör, PHD thesis RWTH Aachen, 1993".

[i.28] ITU-T Recommendation P.830: "Subjective performance assessment of telephone-band and wideband digital codecs".

[i.29] ITU-T contribution COM 12-117, Study Period 1997-2000: "Report of the question 13/12 rapporteur's meeting (Solothurn, Germany, 6-10 March 2000)".

3 Abbreviations

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

ACR	Absolute Comparison Rating
AMR	Adaptive MultiRate
ASL	Active Speech Level

NOTE: According to ITU-T Recommendation P.56 [i.22].

BGN	BackGround Noise
CDF	Cumulative Density Function
CI	Confidence Interval
DB	Data Base
dB SPL	Sound Pressure Level re 20 µPa in dB
G-MOS	Global MOS

NOTE: MOS related to the overall sample.

HP	HighPass
IP	Internet Protocol
IRS	Intermediate Reference System
ITU	International Telecommunication Union
ITU-T	Telecom Standardization Body of ITU
MOS	Mean Opinion Score
MOS-LQSN	Mean Opinion Score - Listening Quality Subjective Noise
MRP	Mouth Reference Point
NI	Network I conditions
NII	Network II conditions
NIII	Network III conditions
NB	NarrowBand
N-MOS	Noise MOS

NOTE: MOS related to the noise transmission only.

NR	Noise Reduction
NR (filter)	Noise Reduction (filter)
PCM	Pulse Code Modulation
PLC	Packet Loss Concealment
RCV	ReCeiVe
RMSE	Random Mean Square Error
S-MOS	Speech MOS

NOTE: MOS related to the speech signal only.

SNR	Signal to Noise Ratio
STF	Specialized Task Force
TOR	Terms Of Reference
VAD	Voice Activity Detection
VoIP	Voice over IP
WB	WideBand

4 Speech signals to be used

As with any objective model, the prediction of speech quality depends on the conditions under which the model was tested and validated (see clauses 6.1 and 8). This dependency also applies to the speech material used in conjunction with the objective model.

The wideband version of the model uses French speech sentences. The near end speech signal (clean speech signal) consists of 8 sentences of speech (2 male and 2 female talkers, 2 sentences each). Appropriate speech samples can be taken from ITU-T Recommendation P.501 [i.13].

The narrowband version of the model uses English speech sentences. The near end speech signal (clean speech signal) consists of 8 sentences of speech (2 male and 2 female talkers, 2 sentences each). Appropriate speech samples can be taken from ITU-T Recommendation P.501 [i.13].

5 Selection of the data within the scope of the wideband objective model: Experts evaluation

5.1 Selection process

The aim of the selection process was to identify those data in the databases described in EG 202 396-2 [i.2] which are consistent with the scope of the objective models to be studied within the present document.

The experts were selected on the based on the definition found in ITU-T Recommendation e.g. P.831 [i.5]: experts are experienced in subjective testing. Experts are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are able to describe their subjective impressions in detail. They have a background in technical implementations of noise reduction systems and transmission impairments and do have detailed knowledge of the influence of particular implementations on subjective quality.

Their task was to select the relevant conditions within the scope of the model to be developed. Therefore they had to verify the consistency of the data with respect to the following selection criteria:

1) Artefacts others than the ones which should have been produced by the signal processing described in [i.2] e.g. due to the additional amplification required in order to provide a listening level of 79 dB SPL.

2) Inconsistencies within one condition due to the selection of the individual speech samples from the database for subjective evaluation.

3) Inconsistencies within one condition due to statistical variation of the signal processing described in [i.2] leading to non consistent judgements within this condition.

4) Inconsistencies due to ITU-T Recommendation P.56 [i.22] level adjustment process chosen for the complete files including the background noise.

5) Impact of the different listening levels used in the two databases - the French and the Czech database.

As a result of the experts listening test a set of data was selected which is used for the development of the objective model.

In the selection process five expert listeners (not native French/Czech speakers) were involved. Their task was not to produce new judgements, but to check all the samples in the database with respect to the possible artefacts described above.

A playback system with calibrated headphones was used for the test. The headphones used were Sennheiser HD 600 connected to the HEAD acoustics playback system HPS V. The equalization provided by the headphone manufacturer was used since this was the one used in the French and Czech test setup.

All samples could be heard by the experts as often as required in order to get final agreement about the applicability of the data within the terms of reference of the model. There was no limitation in comparing samples to the ones previously heard.

5.2 Results

In general it could be observed that the 4 seconds sample size chosen in the experiment according to ITU-T Recommendation P.835 [i.3] lead to a more difficult task even for expert listeners, especially in the case of non stationary background noises. It is more difficult to identify the nature of the noise itself and then identify in addition possible impairments introduced by the signal processing or by the network impairments. It is very likely that some comparatively high standard deviations seen in the data are caused by these effects.

5.3 French database

In general the French database is in line with the ToR except network condition NII. In network condition NII 1 % packet loss was chosen which is too low for the conditions to be evaluated. Due to the inhomogeneously distributed packet losses there are conditions where no packet loss is audible up to conditions where 5 out of 6 samples show packet loss. Furthermore the packet loss may occur during speech as well as during the noise periods. The impact of the different packet losses is not controlled with respect to their occurrence due to the statistical nature of the packet loss distribution, even within a set of 6 samples used for evaluating one condition. Since packet loss is clearly audible under NIII conditions (3 % packet loss) and much better distributed amongst the different samples the NII conditions are not used within the scope of the objective method. They are either covered by the NI condition (0 % packet loss) or by the NIII conditions. This results in 144 NII conditions which are not retained for the development of the model.

From the 288 NI and NIII conditions 28 conditions are not retained. The main reasons therefore are:

- Not consistent signal levels due to the amplification process.
- Insufficient S/N, speech almost inaudible.

The individual reasons for the samples of these conditions being not retained can be found in table A.1.

In total 260 out of 432 conditions are used as the reference for the objective model. In other words, 60,2 % of the data can be used for the model. The distribution of the ratings is between 1,2 and 4,96 MOS for S-/N-/G-MOS.

5.4 Czech database

For every combination of background noise and speaker gender, a single Czech sentence was used (see table 5.1). The 24 Czech listeners had to rate this single sentence, while the French ratings are a mean value of six different sentences (assessed by 4 listeners each).

Table 5.1: Sentences from the test corpus chosen for the different conditions

Condition	Sentence No.
Lux Car 130kmh Female2	S3
Lux Car 130kmh Male1	S2
Crossroads Female2	S4
Crossroads Male1	S3
Road Noise Female2	S5
Road Noise Male1	S4
Office Noise Female2	S6
Office Noise Male1	S5
Pub Noise Female2	S7
Pub Noise Male1	S6

This leads to a limited representation of the individual background noise conditions especially in the case of time varying background noises. Furthermore the NII conditions were even more critical in judgement compared to the French data since either there was no packet loss at all. Or if there was packet loss all listeners rated this particular packet loss because they all listened to the same sentence for one condition. In the French listening test 6 sentences were listened for one condition which provided a higher variance of the distributed packet loss.

The listening level variation in the Czech database, preserved from previous database processing adds another degree of complexity to the problem. The listening levels are generally lower as within the French database and as compared to the general rules laid down in ITU-Recommendations P.800 [i.4] and P.835 [i.3]. The listening level variation within the Czech database is up to 16 dB. In the experts tests the following conclusions were drawn:

- The conditions AMR NII and G.722 NII (1 % packet loss) were not selected, because in most cases, the sound files had too low packet loss. A distinction between and NI and NII conditions is hardly possible.
- The effect of packet loss in the samples should be audible in AMR NIII and G.722 NIII conditions. Because every single Czech condition consists just of one sentence, the packet loss may not be distributed uniformly in the sample. Therefore, only samples with at least one packet loss in speech *and* background noise (before or after speech) were selected.

- Due to the fact that every Czech sound file has a different level (which depends on codec, noise reduction algorithm, etc.), a minimum level of 69 dB SPL was set (10 dB below the recommended listening level of 79 dB SPL). All conditions below this limit were not retained.

- Analysis of NI conditions:

  a) AMR Codec:
  70 conditions were not retained based on the following selection criteria:

     1) Too low level (54).

     2) Inconsistent BGN level (12).

     3) Too low S/N (2).

     4) Too low overall level / given listening level not correct (2).

  b) G.722 Codec:
  19 conditions were not retained based on the following selection criteria:

     1) Too low level (15).

     2) MOS values irreproducible (4).

  c) Selected conditions dependent of BGN: see table 5.2.

Table 5.2: Selected Czech NI conditions

BGN-Condition	Total not retained	Total retained	Selected test samples / MOS available	Selected verification samples / no MOS available
Lux_Car	17	19	10	9
Crossroads	36	0	0	0
Road	17	1	1	0
Office	14	22	16	6
Pub	5	13	10	3

  d) Overall NI acceptance: 48 % of NI conditions are useful (22 % AMR, 65 % G.722).

- Analysis of NIII conditions:

  a) AMR Codec:
  76 conditions were not retained based on the following selection criteria:

     1) Too low level (43).

     2) Inconsistent packet loss (33).

  b) G.722 Codec:
  35 conditions were not retained based on the following selection criteria:

     1) Too low level (13).

     2) Inconsistent packet loss (22).

  c) Selected samples dependent of BGN: see table 5.3.

Table 5.3: Selected Czech NIII conditions

BGN-Condition	Total not retained	Total retained	Selected test samples / MOS available	Selected verification samples / no MOS available
Lux_Car	30	6	4	2
Crossroads	30	6	5	1
Road	16	2	2	0
Office	24	12	10	2
Pub	11	7	2	5

d) Overall NIII acceptance: 23 % of NIII conditions are useful (16 % AMR, 35 % G.722).

The list of the selected Czech conditions is found in table A.1.

In total 88 conditions out of 432 (20,4 %) are suited to be used in a further step for checking language dependencies.

5.5 General differences between the databases

The most important differences between the French and the Czech database can be summarized as follows:

- The French and Czech listening samples of one condition do not have the same levels. The French sound files are louder than the Czech ones, in some random tests, the mean of these level differences is given in table A.2, of EG 202 396-2 [i.2]. This may have lead to different ratings for the Czech samples compared to the French samples. This has be regarded especially for further processing of the sound files.

- For every background noise condition, a single Czech sentence was used (see table 5.1). To quantify the last point, the correlation between French and Czech ratings (S-, N- and G-MOS) can be calculated. As shown below, this correlation is very low. It seems that the differences mentioned above are reflected here. Coefficients of correlation (Pearson's equation) are summarized in table 5.4.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

with:

x MOS Data (Czech)

$\bar{x}$ Mean of MOS Data (Czech)

y MOS Data (French)

$\bar{y}$ Mean of MOS Data (French)

Table 5.4: Comparison of correlation

Over all available ratings (French and Czech, 302 condition each)	**Only selected French MOS Data (NI and NIII conditions, ratings reviewed by experts)** (179 selected French conditions)	**Only Czech and French selected MOS Data (NI and NIII conditions, ratings reviewed by experts)** (59 conditions selected for French *and* Czech)
S-MOS: 0,703 N-MOS: 0,816 G-MOS: 0,668	S-MOS: 0,736 N-MOS: 0,822 G-MOS: 0,776	S-MOS: 0,830 N-MOS: 0,897 G-MOS: 0,871

As shown in the scatter plots below, a slight correlation for the French-optimized data can be noticed, but for a usable correlation, the measurement points are distributed too far away from a (virtual) regression line of best fit (see figures 5.1, 5.3 and 5.5).

If the calculation of the correlation is limited only to the selected data (86 conditions are selected for French *and* Czech speech), the correlation increases for all values, especially for the G-MOS data (see figures 5.2, 5.4 and 5.6).

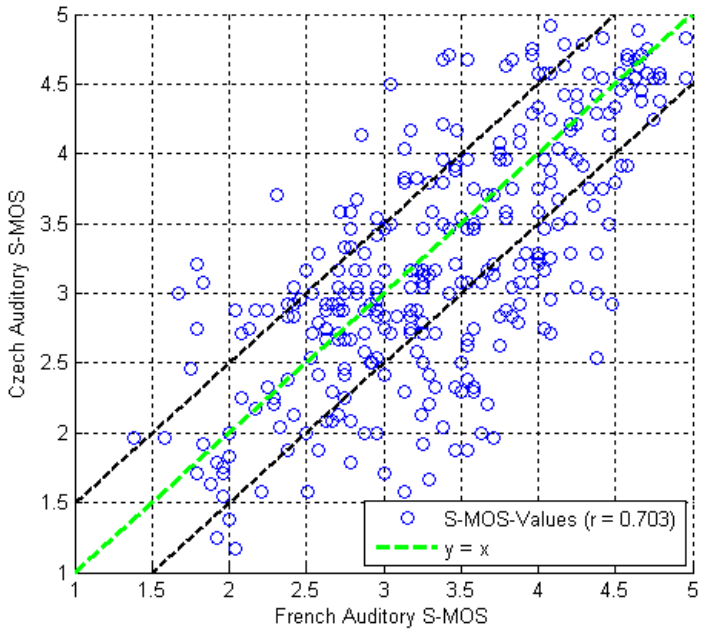


Figure 5.1: Scatter plot of the French data vs. the Czech data for the different conditions, S-MOS, *before* experts selection

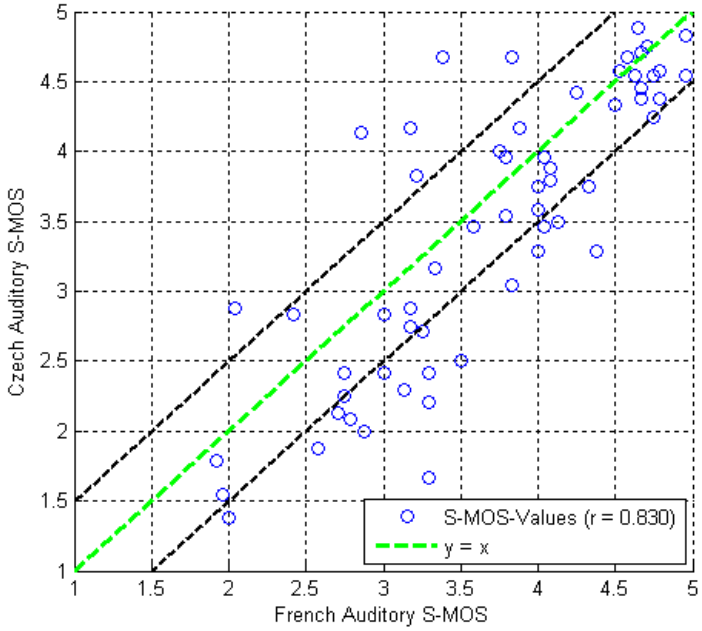


Figure 5.2: Scatter plot of the French data vs. the Czech data, S-MOS, *after* experts selection (only data selected for both languages)

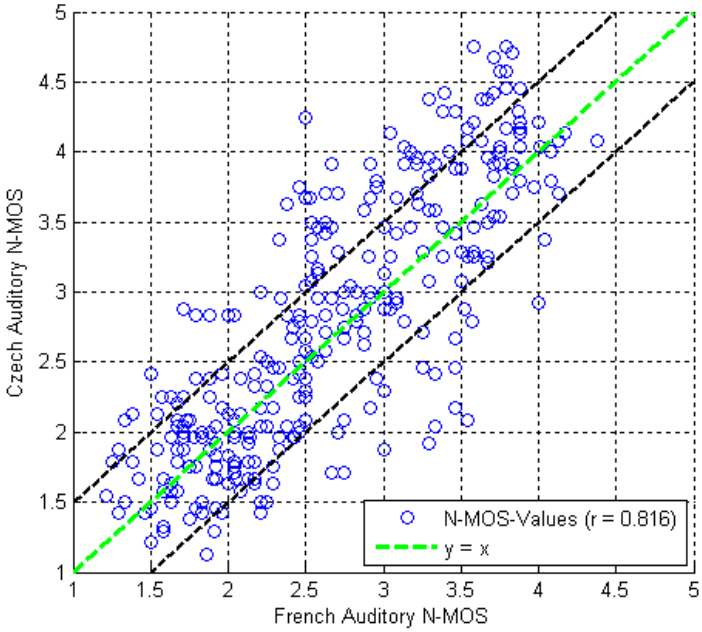


Figure 5.3: Scatter plot of the French data vs. the Czech data for the different conditions, N-MOS, *before* experts selection

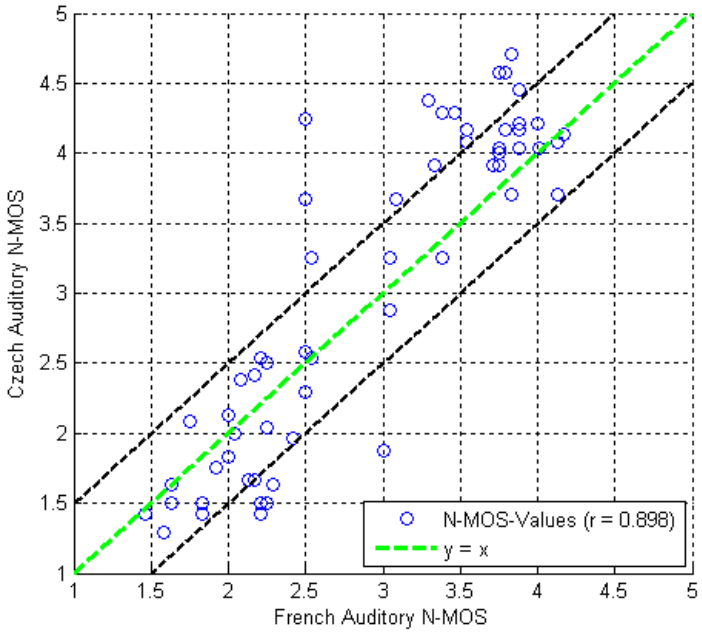


Figure 5.4: Scatter plot of the French data vs. the Czech data, N-MOS, *after* experts selection (only data selected for both languages)

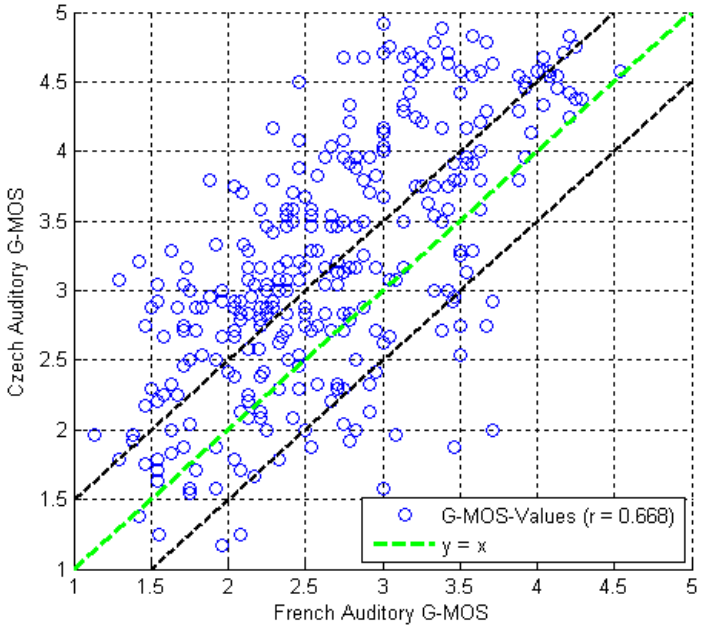


Figure 5.5: Scatter plot of the French data vs. the Czech data for the different conditions, G-MOS, *before* experts selection

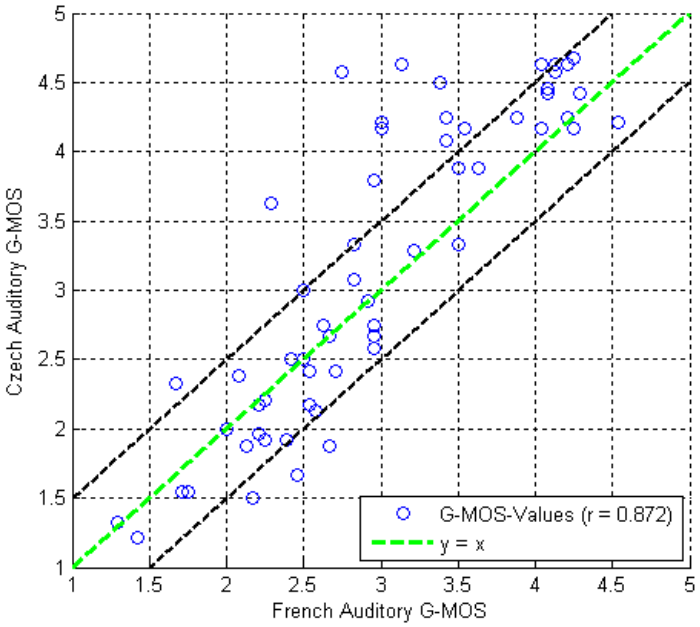


Figure 5.6: Scatter plot of the French data vs. the Czech data, G-MOS, *after* experts selection (only data selected for both languages)

6 Description of the wideband objective test method

6.1 Introduction

The present objective test method is developed in order to calculate objective MOS for speech, noise and the overall quality of a transmitted signal containing speech and background noise, designated N-MOS, S-MOS and G-MOS in the following.

The new model is based on an aurally-adequate analysis in order to best cover the listener's perception based on the previously carried out listening test i.2.

The wideband objective mdel is applicable for:

- wideband handset and wideband hands-free devices (in sending direction);
- noisy environments (stationary or non-stationary noise);
- different noise reduction algorithms;
- AMR [i.21] and G.722 [i.20] wideband coders;
- VoIP networks introducing packet loss.

NOTE 1: For the NIII conditions jitter was introduced. Finally jitter was observed for less than 2 % of the selected conditions. The jitter consideration of the new objective method could therefore not be validated on an appropriate amount of data. Quality impairments typically introduced by different strategies of packet loss concealment and different adaptive jitter buffer control mechanisms were not considered in the listening test database and therefore also not in the objective method.

NOTE 2: The method is not applicable for such background situations where speech intelligibility is the major issue.

Due to the special sample generation process the new method is only applicable for *electrically* recorded signals. The quality of terminals can therefore only be determined in sending direction.

The method was developed by attaching importance to a high reliability. The results of the listening test (selected conditions, see clause 5) were best modelled. Furthermore mechanisms were implemented to provide high robustness also for other than the present samples.

Due to the high diversity between the Czech and the French listening test (see clause 5.5) the development of the objective model is based on the French database being within the ToR and such provides the higher amount of selected samples. The sample preparation and nomenclatures for the new method are described in clause 6.2.

The calculation of *N-MOS*, *SMOS* and *GMOS* is described in detail in clause 6.4 to 6.6. Finally clause 6.7 analyses the results of the new method for the selected French and Czech samples individually and in comparison to each other.

6.2 Speech sample preparation and nomenclature

6.2.1 Speech sample preparation

Based on the data selected in clause 5 an objective model is developed in order to determine:

- the Noise-MOS (N-MOS);
- the Speech-MOS (S-MOS); and
- the "Global"-MOS (G-MOS), the overall quality including speech *and* background noise.

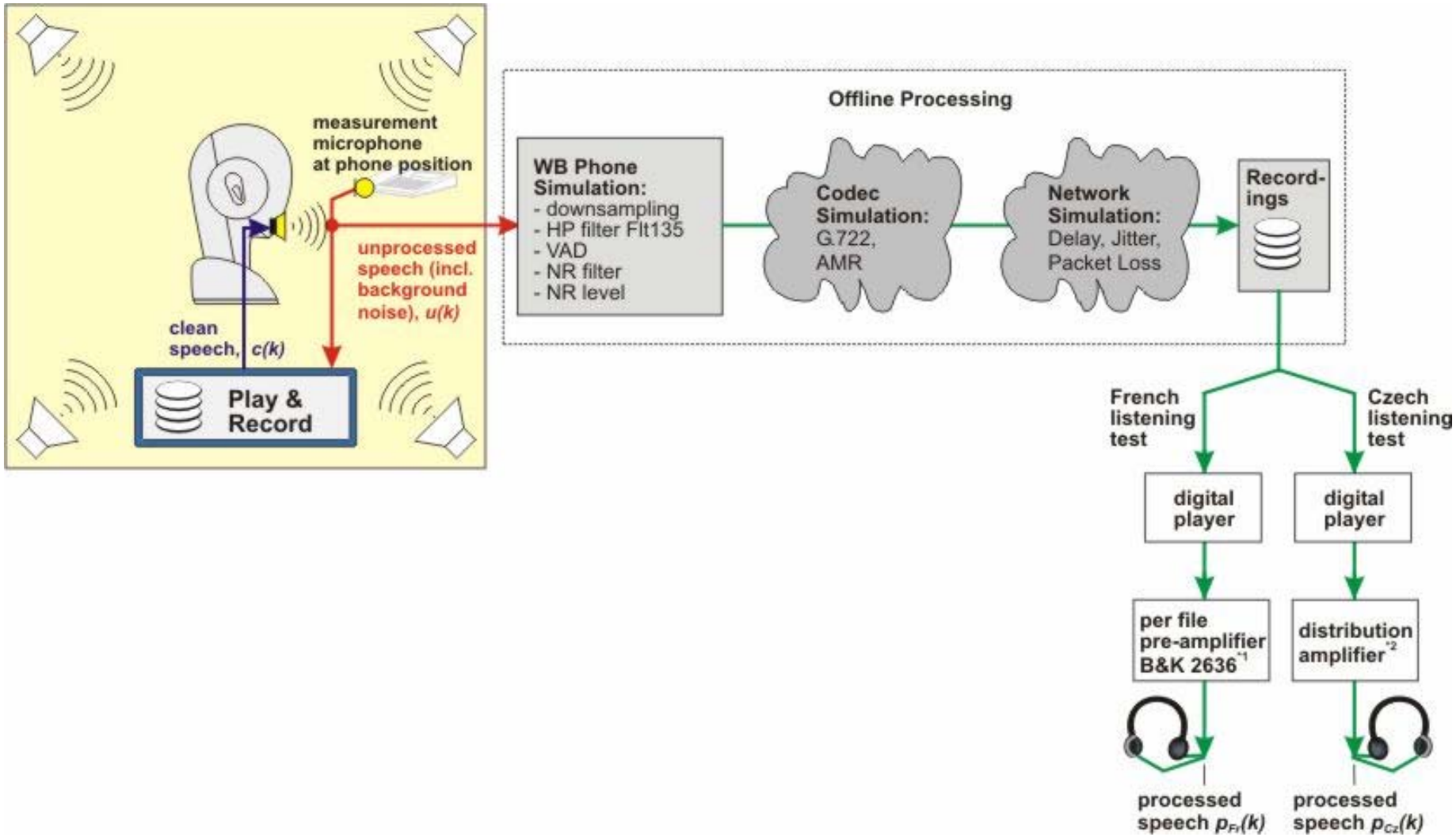
Different input signals can be accessed during the recording process and subsequently can be used for the calculation of N-MOS, S-MOS and G-MOS. Beside the signals used in the listening test ("processed signal"), two additional signals are used as a priori knowledge for the calculation:

1) The "clean speech" signal, which was played back via the artificial mouth at the beginning of the sample generation process.

2) The "unprocessed signal", which was recorded close to the microphone position of the simulated handset device / hands-free telephone (see figure 6.1 and [i.2]). Note that no real phone / hands-free device was used. Phones and handsfree devices were simulated by a free-field microphone and an offline simulation for filtering, VAD, noise reduction, etc.

Both signals are used in order to determine the degradation of speech and background noise due to the signal processing as the listeners did during the listening tests.

The sample generation process is shown in figure 6.1.



NOTE 1: Calibrated for each file with B&K HATS (3.3 ears) to 79 dB SPL ASL (P.56).
NOTE 2: Once calibrated: -26 dBoV resulting to 79 dB SPL measured with a type 3.2 ear (P.57), 5N application force.

Figure 6.1: Sample generation process, indicating "clean speech", "unprocessed speech" and "processed speech"

The processed signal consists of the unprocessed signal after being processed via noise reduction algorithms, voice coder, network simulation, etc. This signal was subjectively rated in the previously carried out listening test (see [i.2] and figure 6.1).

In order to calculate S-MOS, N-MOS and G-MOS, all three signals are required for each sample. The a priori signals (clean speech and unprocessed) were extracted for each processed signal used in the listening tests.

The following preparation steps are required to be carried out for all three files:

1) The clean and unprocessed speech signals were shortened to 4 seconds in order to match the length of the processed signal in the listening tests.

2) The signals were time-aligned. This was achieved after pre-processing followed by a cross-correlation analysis.

NOTE: For samples with an instationary background noise or including packet loss and jitter it should be ensured that the cross-correlation analysis lead to non-ambiguous results. E.g. by applying further processing algorithms in order to better separate between speech and noise parts.

The signals are expected to be in a 48 kHz, 16 bit wave format. The clean speech signals are expected to have an Active Speech Level (ASL, see ITU-T Recommendation P.56 [i.22]) of -4,7 dBPa at the mouth reference point (MRP). For the unprocessed signal the ASL has to remain unchanged compared to the recording close to the phone's microphone. This ensures that the influence of phone position and test room is fully obtained. The processed French signals had an ASL of 79 dB SPL similar to the listening test. The ASL of the Czech processed signals varies between 56 dB SPL and 78 dB SPL and remained unchanged compared to the output of the transmission chain. For further use the speech signals can have either 79 dB SPL ASL or the originally level after the transmission. Care should be taken that the corresponding coefficient sets are used (see clauses 6.4 to 6.6).

6.2.2 Nomenclature

In order to provide a consistent nomenclature within the present document, the relevant terms are briefly described in the following.

The combination of speech sequences, a background noise, a phone type and simulation (filtering, NR level and aggressiveness), a speech codec and a network scenario leads to one **condition** in the terms of the present document and [i.2].

Each condition was generated by processing the clean speech **file** containing eight **sentences** per language via the corresponding scenario, see figure 6.2.

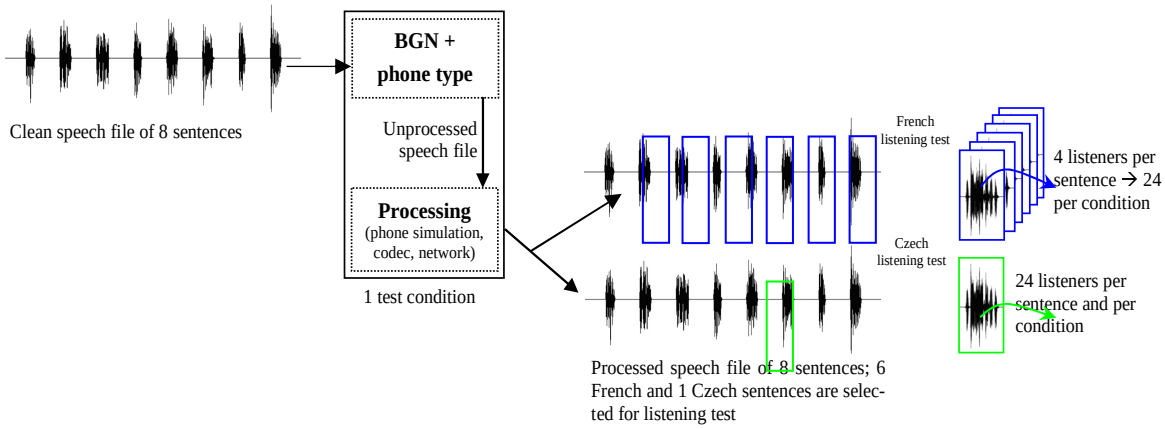


Figure 6.2: Nomenclature (file, condition, sentence)

For the listening tests different parts of the resulting processed files were used. Six of the French sentences per condition were chosen and assessed by 4 persons each. One of the Czech sentences per condition (randomly, see table 5.1) was presented to 24 Czech listeners. The resulting auditory S-/N-/G-MOS were averaged in each case separately.

The consecutively described algorithms calculate the S-/N-/G-MOS sentence-wise. For the French database the MOS scores for one condition were calculated based on 6 sentences, whereas for the Czech database one sentence is used. Beside the processed signal *p(k)* also the a priori signals (clean speech *c(k)* and unprocessed *u(k)*) are necessary (see figure 6.1). The bundle of those three **signals** for one sentence is called a **sample** in the following, see figure 6.3.

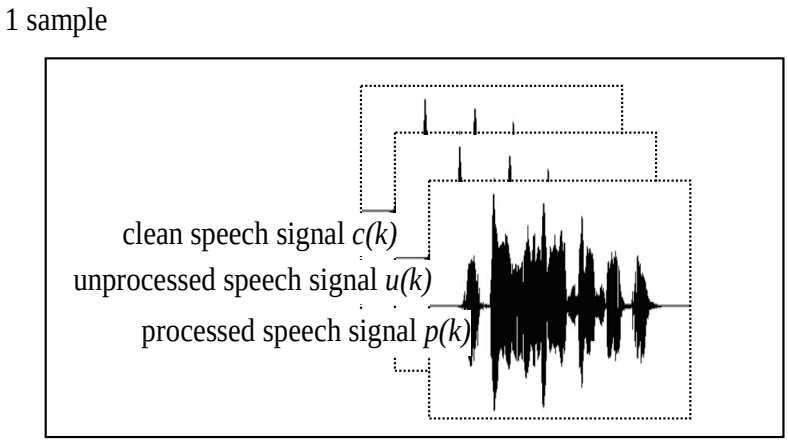


Figure 6.3: Nomenclature (sample)

6.3 Principles of Relative Approach and Δ Relative Approach

The **Relative Approach** [i.6] is an analysis method developed to model a major characteristic of human hearing. This characteristic is the much stronger subjective response to distinct patterns (tones and/or relatively rapid time-varying structure) than to slowly changing levels and loudnesses.

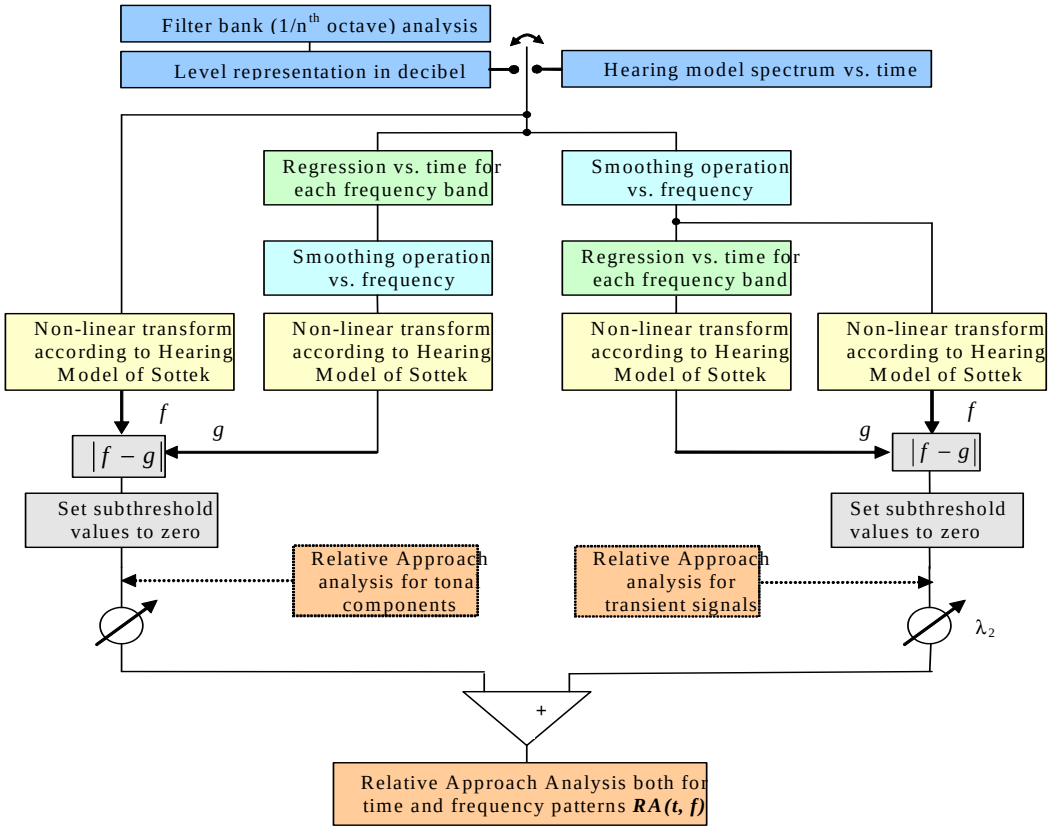


Figure 6.4: Block diagram of Relative Approach

The idea behind the Relative Approach analysis is based on the assumption that human hearing creates a running reference sound (an "anchor signal") for its automatic recognition process against which it classifies tonal or temporal pattern information moment-by-moment. It evaluates the difference between the instantaneous pattern in both time and frequency and the "smooth" or less-structured content in similar time and frequency ranges. In evaluating the acoustic quality of a complex "patterned" signal, the absolute level or loudness is almost without any significance. Temporal structures and spectral patterns are important factors in deciding whether a sound is judged as annoying or disturbing [i.12], [i.14], [i.15] and [i.27].

Similar to human hearing and in contrast to other analysis methods the Relative Approach algorithm does *not* require any reference signal for the calculation. Only the signal under test is analyzed. Comparable to the human experience and expectation, the algorithm generates an "internal reference" which can be best described as a forward estimation. The Relative Approach algorithm objectifies pattern(s) in accordance with human perception by resolving or extracting them while largely rejecting pseudostationary energy. At the same time, it considers the context of the relative difference of the "patterned" and "non-patterned" magnitudes.

Figure 6.4 shows a block diagram of the Relative Approach. The time-dependent spectral pre-processing can either be done by a filter bank analysis ($1/n^{th}$ octave, typically $1/12^{th}$ octave) or a Hearing Model spectrum versus time according to the Hearing Model of Sottek (see [i.27]). Both of them result in a spectral representation versus time. Both are calculating the spectrograph using only linear operation and their outputs are therefore directly comparable. The Hearing Model analysis parameters are fixed and based on the processing in human ears whereas the input parameters for the filter bank analysis can vary. The filter bank pre-processing approximates the Hearing Model version. As input for either the filter bank or the Hearing Model signals adjusted to 79 dB SPL can be used (according to the French listening test) or signals with their original level after signal processing (according to the Czech listening test).

Two different variants of Relative Approach can be applied to the pre-processed signal. The first one applies a regression versus time for each frequency band in order to cover human expectation for each band within the next short period of time. Afterwards for each time slot a smoothing versus frequency is performed. The next step is a non-linear transformation according to the Hearing Model of Sottek (see [i.27]). This output is compared to the source signal which is also Hearing Model transformed. Non-relevant components for human hearing are finally set to zero. This approach focuses on the detection of tonal components. The second version first smoothes versus frequency within a time slot and then applies the regression versus time. This output signal is transformed non-linear to the Hearing Model of Sottek. It is compared to the output of the smoothing versus frequency which is also non-linearly transformed according to the Hearing Model. Finally non-relevant components for human hearing are again set to zero. Thus more transient structures are detected.

Via the factors λ_1 and λ_2 the weighting of Relative Approach for tonal and transient signals can be set. For the new model $\lambda_1 = 0$ and $\lambda_2 = 1$ was chosen. Thus, the model is tuned to detect time-variant transient structures.

The result of the Relative Approach analysis is a 3D spectrograph displaying the deviation from the "close to the human expectation" between the estimated and the current signal is displayed versus time and frequency. Currently the Relative Approach uses a time resolution of $\Delta t = 6{,}66$ ms. The frequency range from 15 Hz to 24 kHz is divided into 128 frequency bands Δf_m which corresponds to a $1/12^{th}$ octave resolution. Due to the nonlinearity in the relationship between sound pressure and perceived loudness, the term "compressed pressure" in compressed Pascal (cPa) is used to describe the result of applying the nonlinear transform.

The N-MOS (and also the S-MOS) calculation of the present objective model is based on the Relative Approach. Due to the time variant characteristic of speech and most of the background noise signals, the 3D Relative Approach spectrograph always shows a deviation between the expected and the current signal which is indicated by patterns in the time-variant signal. A first attempt using Relative Approach for analyzing time variant background noises was submitted as a contribution in ITU-T 2001 [i.7]. For time variant signals this "estimation error" can best be interpreted as the "attention" which is attracted by the patterns of the particular signal on human perception. The 3D spectrograph of a time variant signal therefore provides some information for the N-MOS (and also S-MOS) determination. But it needs additionally be considered what humans expect if they think of a "good" sound quality for time variant background noise and speech signals. The unprocessed signal and the clean speech signal respectively (see clause 6.2) can be seen as such a "good quality reference". The knowledge about "good" or "poor" quality is not yet covered by Relative Approach. Relative Approach can only determine how "close to the human expectation" a signal is, but not if this expectation is of a high or a low quality origin.

The 3D Relative Approach spectrograph is therefore calculated for the processed as well as for the unprocessed signal. Both spectrographs are then subtracted from each other in order to determine what has *changed* due to the transmission. This differential analysis, the **Δ Relative Approach**, between the transmitted processed signal and the undisturbed unprocessed signal provides the information how "close to the human expectation" the processed signal still is compared to the unprocessed signal. The calculation is carried out using equation 6.1.

$$\Delta RA(\Delta t_i, \Delta f_j) = RA_p(\Delta t_i, \Delta f_j) - RA_u(\Delta t_i, \Delta f_j) \quad (6.1)$$

$$\forall\ \Delta t_i, \Delta f_j \text{ within } \quad \Delta f_{min} \le \Delta f_j \le \Delta f_{max},$$

Δt_i = 6,66 ms between t_{min} and t_{max} given by the beginning and the end of the sample.

An undisturbed transmission would lead to a homogeneous differential spectrograph indicating a "close to the original" transmission. A transmission leading to highly modulated background noises will result to an inhomogeneous differential spectrograph showing distinct patterns (time and frequency wise). They are caused by the signal processing during the transmission and raise compared to the original, unprocessed signal. They are aurally-adequate detected by the Δ Relative Approach. Those kinds of transmissions typically lead to a low N-MOS.

The Δ Relative Approach analysis was already successfully applied during the 4th SQTE [i.11] for VoIP transmission evaluating "transparency" of background noise transmission influenced, e.g. by VAD or comfort noise.

6.4 Objective N-MOS

6.4.1 Introduction

The N-MOS calculation is based on three principles:

1) Choice of a hearing-adequate analysis in order to reproduce human perception.

2) Tuning to the database in order to provide in a high correlation between auditory and objective N-MOS.

3) Ensure robustness for scenarios outside the database.

The present database contains 179 (French) conditions which were selected according to clause 4. Their S-/N-/G-MOS scores were known during the development phase of the model.

The objective N-MOS algorithm is based on the results of the subjective listening test and conclusions drawn from the consecutive expert listening analysis. Expert analysis led the extraction of the main parameters leading to the subjective N-MOS:

- Absolute background noise level.
- Modulation of background noise, e.g. musical tones.
- "Naturalness" of the background noise.
- Lost packets (minor influence).

6.4.2 Description of N-MOS algorithm

The aim of the N-MOS calculation is to reproduce the relevant parameters influencing subject's assessment by a technically analysis. These parameters are the absolute level, disturbing "modulations" and the "naturalness" as derived by the experts listening test. Simple analyses like A-weighted sound pressure level, 3rd octave analyses and also even most of the known psychoacoustic analyses were not capable to fully describe human listening perception in such complex listening situations. Besides level analyses, an analysis which is capable to adequately analyze the acoustic quality as typically perceived by humans is the Relative Approach [i.8], an aurally-adequate analysis.

The N-MOS is calculated as shown in figure 6.5. Scalar signal paths are shown with thin solid lines, vector signals are shown with dashed lines and 3D spectrographs are given with thick solid lines. Note that in advance of the N-MOS calculation the pre-processing steps described in clause 6.2 have to be carried out.

The N-MOS is calculated on basis of the Relative Approach and the absolute level of the processed background noise. High background noise levels were typically judged with low N-MOS in the listening test. This background noise **level** N_{BGN} is calculated for those sections of the processed signal *p(k)* which contain only background noise and no speech. The clean speech signal *c(k)* is used as a **mask** in order to determine the beginning and end of these sections.

The level N_{BGN} is then calculated in dB Pa for the extracted background noise sections in the processed signal $p_{BGN}(k)$ by using equations 6.2 and 6.3. The French subjects listened to the signal *p(k)*, which was adjusted to an acoustic level of 79 dB SPL active speech level. The level N_{BGN} is therefore also calculated as an acoustics level. 79 dB SPL corresponds to -15 dB Pa. This is furthermore necessary since the Relative Approach analysis requires a dB Pa calibrated signal.

$$N'_{BGN} = \frac{1}{K}\sum_{k} p^2_{BGN}(k) \tag{6.2}$$

$$N_{BGN} = 10 \cdot \log\left(\frac{N'_{BGN}}{1Pa}\right) \tag{6.3}$$

Where: *k* are the sample bins during the background noise sections of the processed signal *p(k)*.

The **3D Relative Approach** spectrograph is calculated for the unprocessed signal *u(k)* and the processed signal *p(k)* ($RA_u(t, f)$, $RA_p(t, f)$). In these spectrographs the background noise sections are again extracted using the clean speech signal as a mask resulting in $RA_{BGN,p}(t, f)$ and $RA_{BGN,u}(t, f)$. Note that the Relative Approach calculation is carried out for the whole 4 s duration *before* the noise sections are extracted and in order to guarantee a fully adapted Relative Approach, an adaptation time of 420 ms is considered.

In the next step the 3D spectrographs are **subtracted** from each other ($RA_p(t, f)$ - $RA_u(t, f)$) in order to assess the similarity between the processed versus the unprocessed background noise for human perception. The resulting 3D spectrograph is designated as $\Delta RA_{BGN,p\text{-}u}(t, f)$ in the following. In order classify these spectrographs with numerical values the **variance σ^2** for $RA_p(t, f)$, $RA_u(t, f)$ and $\Delta RA_{BGN,p\text{-}u}(t, f)$ and the **mean μ** for $RA_p(t, f)$ and $\Delta RA_{BGN,p\text{-}u}(t, f)$ are calculated according to equation 6.4 and 6.5. Note that the calculation of σ^2 and μ is again started after the adaptation time of Relative Approach (420 ms).

$$\mu = \frac{1}{A_{ges}} \cdot \sum_{t_i = t_{\min}}^{t_{\max}} \sum_{\Delta f_m = \Delta f_{\min}}^{\Delta f_{\max}} RA_{BGN}(t_i, f_m) \cdot dA(\Delta f_m) \tag{6.4}$$

and

$$\sigma^2 = \left(\frac{1}{A_{ges}} \cdot \sum_{t_i = t_{\min}}^{t_{\max}} \sum_{\Delta f_m = \Delta f_{\min}}^{\Delta f_{\max}} RA^2_{BGN}(t_i, f_m) \cdot dA(\Delta f_m)\right) - \mu^2 \tag{6.5}$$

with:

$$A_{ges} = \frac{1}{(t_{\max} - t_{\min})(f_{\max} - f_{\min})},$$

$$dA(\Delta f_m) = \Delta t \cdot \Delta f_m,$$

Δt = 6,66 ms.

$\Delta f_m \neq$ constant (1/12th octave frequency band resolution).

f_{min} = 50 Hz, lower frequency of band Δf_{min},.

f_{max} = 8 kHz, upper frequency of band Δf_{max},.

f_m centre frequency of band Δf_m.

t_{min} + 420 ms and t_{max} given by the background noise section extracted before.

Mean $(m\Delta RA_{BGN,p\text{-}u})$ and variance $(v\Delta RA_{BGN,p\text{-}u})$ are calculated for the $\Delta RA_{BGN,p\text{-}u}(t, f)$ spectrograph in order to determine the similarity between unprocessed and processed signal ("close to original"). For a high similarity both parameters should be low leading to a high N-MOS.

If the variance is high - independent of the mean - the processed signal is e.g. highly modulated compared to the unprocessed signal. A typical reason are musical tones. These modulations lead to patterns in the Relative Approach spectrographs $RA_{BGN,p}(t, f)$ and $\Delta RA_{BGN,p\text{-}u}(t, f)$. These indicate a high "attraction" on human perception, because these components are unexpected. They were not present in the unprocessed signal. These patterns appear typically only temporarily in $\Delta RA_{BGN,p\text{-}u}(t, f)$ and also only for distinct frequencies. They indicate which parts of the signal have changed compared to the unprocessed signal.

A high mean of $\Delta RA_{BGN,p\text{-}u}(t, f)$ typically indicates a low "naturalness" of the processed signal compared to the unprocessed signal. This might be caused by a high level difference between unprocessed and processed signal. Consequently a low N-MOS can be expected independent of the variance.

Mean and variance of $\Delta RA_{BGN,p\text{-}u}(t, f)$ alone are still not sufficient to predict the N-MOS reliable, because they are derived from a *differential* spectrograph. "Anchors" to the unprocessed and the processed signal are needed in order to judge this mean and variance for the N-MOS calculation correctly. For the processed signal therefore the mean value ($mRA_{BGN,p}$) is calculated in order to get references for the signal level, the potential SNR improvement (e.g. due to a noise reduction) and the degree of the "attention" attracted. The mean of the unprocessed signal is redundant due to the linearity of the operations (Δ Relative Approach and mean).

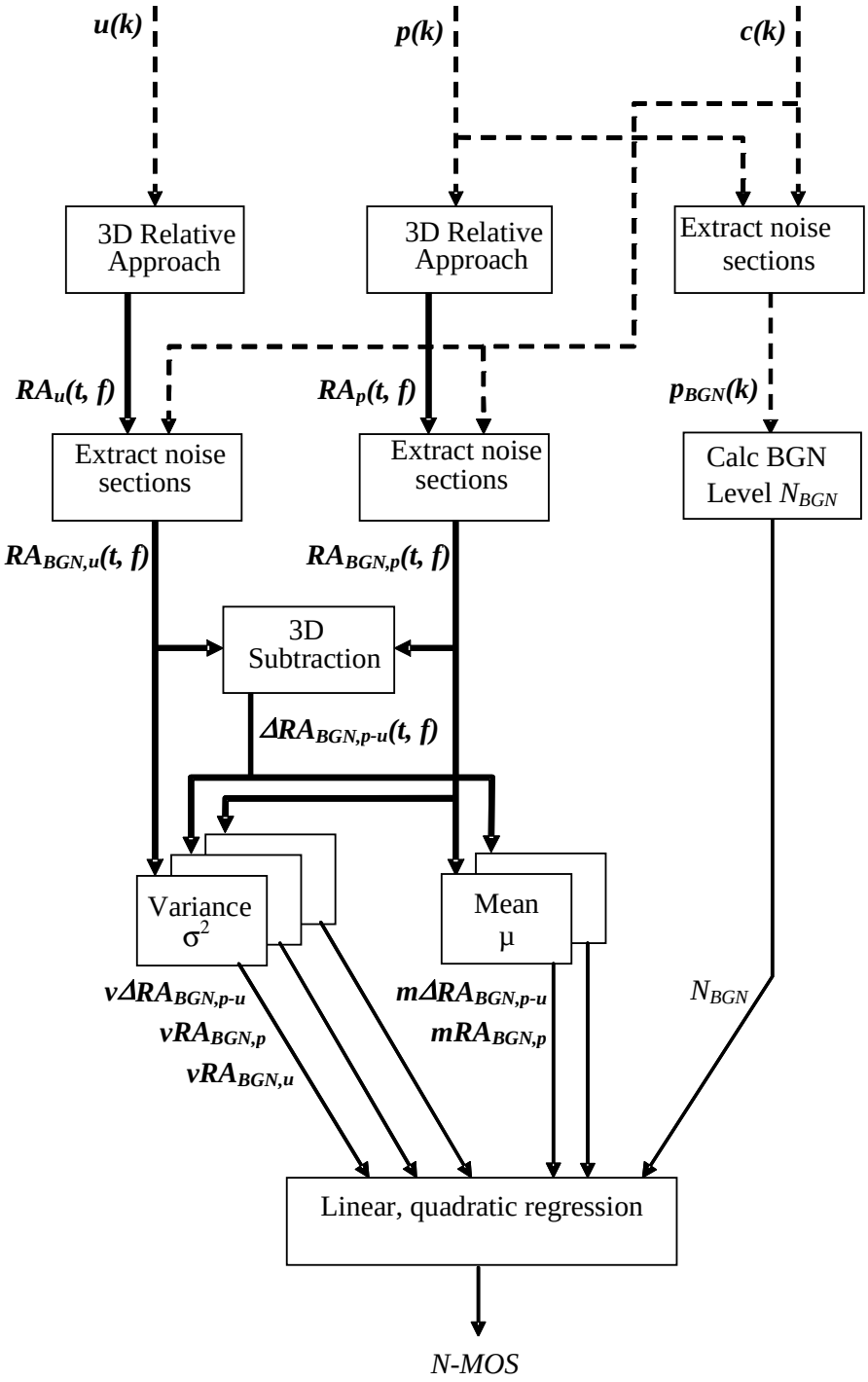


Figure 6.5: Block diagram of N-MOS calculation algorithm; *u(k)* unprocessed signal, *p(k)* processed signal, *c(k)* clean speech signal

Therefore the variance is calculated for both, the unprocessed ($vRA_{BGN,u}$) and the processed ($vRA_{BGN,p}$) signal in order to provide a measure for the "attention" attracted by each of the signals on human perception. In case of the unprocessed signal this is mainly depending on the structure of the background noise. Stationary noises lead to low variance values, whereas non-stationary noises lead to high variances corresponding to a high "attention" attracted. For the processed signal the variance is not only influenced by the structure of the background noise, but also by the *changes* noise reduction algorithms and other signal processing components introduce to the signal.

Finally the *N-MOS* is the result of a **linear, quadratic regression** algorithm applied to all six parameters (N_{BGN}, $m\Delta RA_{BGN,p\text{-}u}$, $v\Delta RA_{BGN,p\text{-}u}$, $mRA_{BGN,p}$, $vRA_{BGN,p}$ and $vRA_{BGN,u}$):

$$NMOS = c_0 + c_{BGN} \cdot N_{BGN} + \sum_{j=1}^{2} \sum_{i=1}^{5} c_{ji} \cdot P_i^{\,j} \qquad (6.6)$$

where:

c_0, c_{BGN} and c_{ji} are the coefficients for the linear regression;

j is the regression order index;

P_i are the Relative Approach related parameters $m\Delta RA_{BGN,p\text{-}u}$, $v\Delta RA_{BGN,p\text{-}u}$, $mRA_{BGN,p}$, $vRA_{BGN,p}$; and $vRA_{BGN,u}$.

NOTE: The influence of **packet loss** is *not* considered separately, but indirectly by the Relative Approach. A lost packet is typically a simple gap in the signal. The phase information is also completely lost. Gaps and phase errors sound very unpleasant and are detected by the Relative Approach as a highly disturbing wideband pattern or, in other words, as a high "attention" attracted at human perception. In case of a lost packet during the background noise sections the mean and the variance of the Δ Relative Approach and the 3D Relative Approach spectrograph of the processed signal are effected and will increase. This decreases the *N-MOS* accordingly. The influence of jitter is so far not considered. A maximum jitter of 20 ms was applied within the present data. But only for a very few conditions jitter could be observed. Jitter could therefore not be covered reliable by the model. Higher amounts of jitter and adaptive jitter buffers are not found in the present database and were therefore not yet investigated.

It should be noted that the expert study of the processed signals used in the listening tests (see [i.2]) showed that packet loss during the background noise sections only slightly decreased the *N-MOS*. Furthermore "real packet losses" occur only rarely in today's networks because VoIP devices like gateways and IP-phone are typically equipped with packet loss concealment (PLC) algorithms. Those PLC algorithms were not applied during the sample generation process of the present database used in the listening tests. In principle the Relative Approach algorithm was already successfully applied in the past to scenarios using different PLC and jitter buffer implementations [i.8], [i.9], [i.10], [i.11] and [i.12]. The N-MOS algorithm is therefore expected to work properly also for PLC scenarios.

Training and validation of the model were carried out using the regression coefficients for the N-MOS calculation summarized in table 6.1.

Table 6.1: Coefficients for linear, quadratic N-MOS regression algorithm

Order	c_0	c_{BGN} (N_{BGN})	c_{j1} ($vRA_{BGN,u}$)	c_{j2} ($vRA_{BGN,p}$)	c_{j3} ($v\Delta RA_{BGN,p\text{-}u}$)	c_{j4} ($m\Delta RA_{BGN,p\text{-}u}$)	c_{j5} ($mRA_{BGN,p}$)
1	2,1533	-0,0600	1,5715	0,2822	-0,2707	-3,6258	-0,7605
2	-	-	-0,0503	-0,0275	0,0263	0,9220	0,1560

6.4.3 Comparing subjective and objective N-MOS results

The coefficients for the linear quadratic regression were determined during the training of the algorithm by averaging the six contributing parameters (N_{BGN}, $m\Delta RA_{BGN,p\text{-}u}$, $v\Delta RA_{BGN,p\text{-}u}$, $mRA_{BGN,p}$, $vRA_{BGN,p}$ and $vRA_{BGN,u}$) for the six French sentences of one condition. In the second step these averaged parameters were mapped by the regression formula to the auditory N-MOS derived in the listening test.

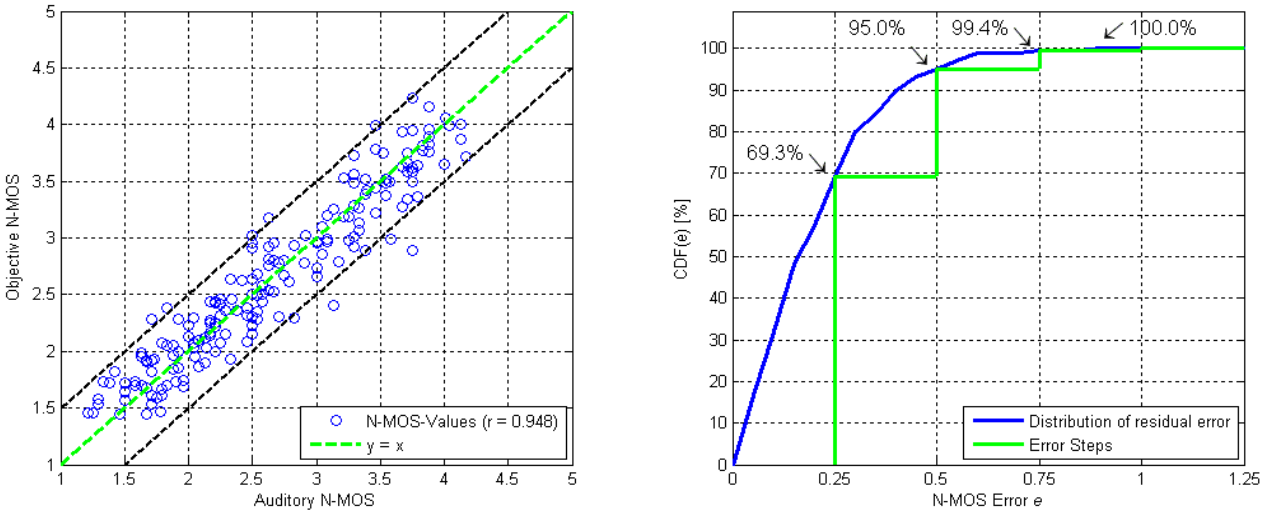


Figure 6.6: Left: Objectively calculated N-MOS versus auditory N-MOS; Right: CDF of residual error versus N-MOS error e

All selected (French) conditions according to clause 4 were used for this mapping - independent of the network condition.

The left hand graph in figure 6.6 shows that the per sample deviation between the subjective and objective N-MOS is less than 0,5 MOS for nearly all (179) conditions. This results in an overall correlation of 94,8 %.

The right graph in figure 6.6 shows the cumulative density function *CDF(e)* versus the N-MOS error *e*.

$$e = \left| NMOS_{auditory} - NMOS_{objective} \right| \tag{6.7}$$

Based on the cumulated density function the right hand graph in figure 6.6 shows additionally an adaptive tolerance scheme indicating the *CDF(e)* values for *e = 0,25*, *e = 0,5*, *e = 0,75* and *e = 1*. For example is the N-MOS error e lower than 0,25 for 69 % of the conditions and lower than 0,75 for 99 % of all conditions.

6.5 Objective S-MOS

6.5.1 Introduction

The objective S-MOS is also aimed to reproduce the listening impression of the test persons in the listening test, to provide a high correlation to the given database and also a high robustness for other databases. The experts group verified the subjective S-MOS values and in combination with their listening impression they extracted the parameters relevant for the S-MOS:

- Level and quality of processed background noise.
- Signal to noise ratio (SNR) between speech and noise in the processed signal.
- Improvement or impairment of SNR between unprocessed and processed signal.
- Packet loss.
- Modulation of speech / speech sound.
- "Naturalness".

At a first glance it seems surprisingly that one of the main influences on the S-MOS seems to be the background noise quality. The experts found out that if the quality of the background noise at the beginning of the sample is good, the speech quality is also expected to be good. And if the processed background noise sounds unpleasant - for whatever reason - also the speech quality is expected to be low. Between both extremes a sliding crossover area can be observed.

The Δ Relative Approach is again chosen to determine parameters like "modulation" or "naturalness" and also in order to cover packet loss effects.

6.5.2 Description of S-MOS Algorithm

Similar to the N-MOS calculation also the S-MOS algorithm is also designed to reproduce the parameters which were extracted by the experts analysis.

The principle of the S-MOS calculation is shown in the block diagram in figure 5.7. Again it should be noted that the clean speech *c(k)*, the unprocessed *u(k)* and the processed signal *p(k)* have to be pre-processing along the steps described in clause 6.2. The input for the linear quadratic regression algorithm leading to the objective S-MOS are ΔSNR, five Relative Approach related parameters and the N-MOS for this particular sample.

The difference between the SNR of the unprocessed and the processed signal (***ΔSNR***) is one of the extracted parameters by the experts. In order to determine the SNR in each signal, the clean speech signal is again used as a mask in order to separate the speech sections ($u_{SP}(k)$ and $p_{SP}(k)$) and the noise sections ($u_{BGN}(k)$ and $p_{BGN}(k)$). The level is then calculated along equation (6.3), which results in the speech *and* noise level for those sections without ($(S+N)''_{SP,u}$ and $(S+N)''_{SP,p}$) and in the noise level during only background noise sections ($N''_{BGN,u}$ and $N''_{BGN,p}$). For the unprocessed and the processed signal SNR_u *and* SNR_p are then calculated in dB according to equation 6.8:

$$SNR = 10 \cdot \log\left(\frac{(S+N)'_{SP} - N'_{BGN}}{N'_{BGN}}\right) \tag{6.8}$$

The ΔSNR is the simple difference between SNR_u and SNR_p:

$$\Delta SNR = SNR_p - SNR_u \tag{6.9}$$

In order to cover the influence signal processing on the sound of the transmitted signal, the modulation and "naturalness" (potentially impaired e.g. by noise reduction algorithms) the Relative Approach and the Δ Relative Approach are used.

The **3D Relative Approach spectrographs** are calculated for all three signals, the unprocessed, the processed and for the clean speech signal ($RA_u(t,f)$, $RA_p(t,f)$ and $RA_c(t,f)$). With the clean speech as **mask** the speech sections of the 3D spectrographs are extracted ($RA_{SP,u}(t,f)$, $RA_{SP,p}(t,f)$ and $RA_{SP,c}(t,f)$).

In the next step two **Δ Relative Approach spectrographs** are calculated between the processed and the unprocessed signal ($\Delta RA_{SP,p\text{-}u}(t,f)$) and between the processed and the clean speech signal ($\Delta RA_{SP,p\text{-}c}(t,f)$).

The **variance σ^2** and the **mean μ** are calculated for both using the equations (6.4) and (6.5) ($v\Delta RA_{SP,p\text{-}u}$, $v\Delta RA_{SP,p\text{-}c}$, $m\Delta RA_{SP,p\text{-}u}$ and $m\Delta RA_{SP,p\text{-}c}$). Additionally the mean is calculated for $RA_{SP,p}(t,f)$ ($mRA_{SP,p}$).

The resulting values ΔSNR, $mRA_{SP,p}$, $v\Delta RA_{SP,p\text{-}u}$, $v\Delta RA_{SP,p\text{-}c}$, $m\Delta RA_{SP,p\text{-}u}$ and $m\Delta RA_{SP,p\text{-}c}$ are used as input parameters for a linear quadratic regression. A seventh indirect input parameter for the regression is the ***N-MOS***. As mentioned above the results of the experts listening test indicated that test persons tend to expect high quality speech if the background noise sounds pleasant at the beginning of the sample. And also vice versa: if the background noise sounds unpleasant, the speech sound is also expected to be impaired. During the algorithm training the selected French samples were therefore divided in three groups based on this finding:

- High N-MOS → high speech quality expected (N-MOS>N-MOS_{high} in figure 5.7).
- Average N-MOS → no clear conclusion can be drawn, several influences need to be considered (N-MOS_{low} <N-MOS<N-MOS_{high} in figure 5.7).

- Low N-MOS → low speech quality expected (N-MOS<N-MOS$_{low}$ in figure 5.7).

For the group with the high N-MOS results (low background noise level, no artefacts, natural sound) test persons most likely compare the speech quality to the speech sound without any background noise. They internally mask the background noise. This aspect is covered by the calculation of $\Delta RA_{SP,p\text{-}c}(t, f)$. Similar than in the N-MOS algorithm the mean of this differential Relative Approach spectrograph covers the average amount of difference between the processed and the clean speech (only during speech sections). If the speech in the processed signal is still similar to the clean speech signal, the differential spectrograph is flat and homogeneous versus time and frequency. It shows no patterns introduced by the transmission. In this case the transmission can be regarded as "close to the original". The mean value of this differential spectrograph will be low. Note that the differential spectrograph compares the processed signal consisting of speech *and* background noise and the clean speech signal which only consists of speech. The influence of the background noise in the processed signal is expected to be low. This can be concluded due to the high N-MOS (e.g. caused by a low background noise level).

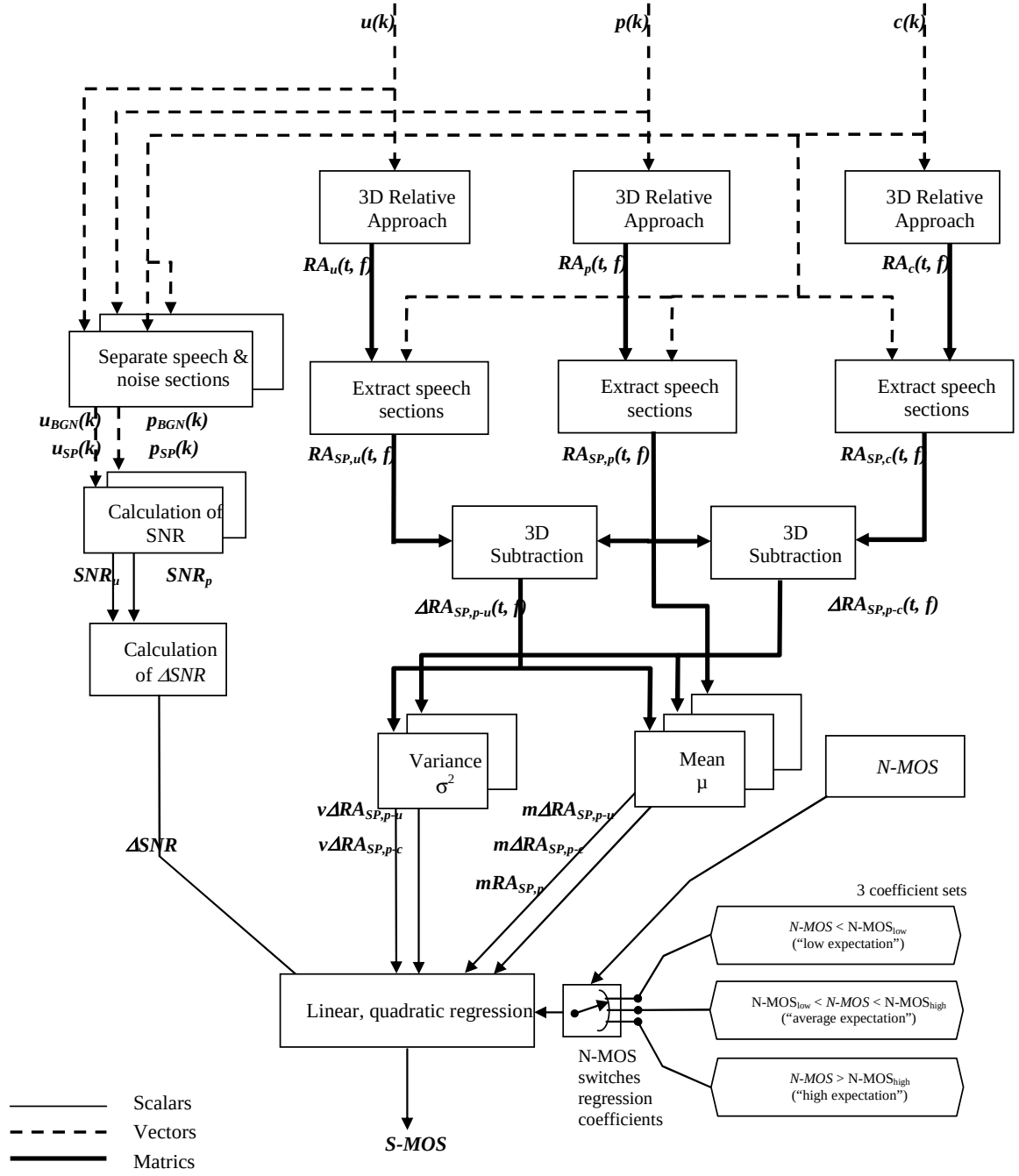


Figure 6.7: Block diagram of S-MOS calculation algorithm; *u(k)* unprocessed signal, *p(k)* processed signal, *c(k)* clean speech signal

The variance $v\Delta RA_{SP,p\text{-}c}$ is a measure for the amount of patterns in the differential spectrograph between processed and clean speech signal. Patterns may occur due to e.g. musical tones or modulations introduced by noise reductions or other signal processing components. Those patterns attract the listeners attention. The variance $v\Delta RA_{SP,p\text{-}c}$ can therefore also be seen as a measure for the amount of "attention" attracted.

A similar effect could be observed for those listening examples providing low N-MOS scores: if the quality of the background noise is poor at the beginning of the sample, subjects expect a poor speech quality. They compare the actual speech to a signal containing speech *and* background noise. Mean and variance are therefore calculated for the Δ Relative Approach between the processed and the unprocessed signal ($\Delta RA_{SP,p\text{-}u}(t, f)$).

The mean $mRA_{SP,p}$ is used in both cases in order to characterize the absolute "attention" attracted by the processed signal. The comparison of $mRA_{SP,p}$ and $m\Delta RA_{SP,p\text{-}c}$ covers the influence of added or removed patterns introduced by room acoustics, background noise, the phone and the signal processing during the transmission. Similarly $mRA_{SP,p}$ and $m\Delta RA_{SP,p\text{-}u}$ can be compared in order to assess only the influence of the terminal and the transmission. The combination of these three parameters indicates whether the speech quality was impaired or improved.

Depending on the N-MOS of a sample the parameters $v\Delta RA_{SP,p\text{-}u}$, $m\Delta RA_{SP,p\text{-}u}$ or $v\Delta RA_{SP,p\text{-}c}$, $m\Delta RA_{SP,p\text{-}c}$ are more or less important. In order to cover this and before starting the regression algorithm the N-MOS of a sample is compared to two thresholds $\text{N-MOS}_{\text{low}}$ and $\text{N-MOS}_{\text{high}}$. If the actual N-MOS is lower than $\text{N-MOS}_{\text{low}}$, a set of regression coefficients is loaded which stronger weights the results (mean and variance) of $\Delta RA_{SP,p\text{-}u}(t, f)$. If N-MOS is higher than $\text{N-MOS}_{\text{high}}$, the regression coefficient set emphasis the result of $\Delta RA_{SP,p\text{-}c}(t, f)$. This decision stronger weights either the comparison of the processed signal to the clean speech or to the unprocessed signal.

In case the N-MOS is between both thresholds a third set of regression coefficients is chosen, which has no preferable comparison base. This again is a result of the expert analysis of the listening test results. One reason for that is that the six sentences of one condition are often very different in terms of speech quality (due to different packet loss rates, different background noise parts, etc). The results of all six sentences were averaged to one S-MOS. The N-MOS of each of the six sentences also may vary, some sentences belong to the upper N-MOS group and some to the lower N-MOS group. This high diversity between the sentence-based results of one condition requires a "crossover-area" between the other two groups ($N\text{-}MOS < \text{N-MOS}_{\text{low}}$ and $N\text{-}MOS > \text{N-MOS}_{\text{high}}$).

Another influence is that some subjects may compare a processed "average quality" signal to unprocessed signals, some to clean speech signals. This depends on individual expectation of "good speech quality".

Based on the expert analysis and the amount and distribution of the conditions (selected, French, trainings set) in the actual version of the objective model $\text{N-MOS}_{\text{low}}$ is set to 2,25 and $\text{N-MOS}_{\text{high}}$ to 3,0.

Note that beside the two variances and means also ΔSNR is always used as one of the **regression** input parameters.

The final S-MOS equation is:

$$SMOS = {}_{R}c_0 + \sum_{j=1}^{2}\sum_{n=1}^{6} {}_{R}c_{jn} \cdot P_n^j \qquad (6.10)$$

where: j is the regression order index;

P_n are the parameters ΔSNR, $v\Delta RA_{SP,p\text{-}u}$, $m\Delta RA_{SP,p\text{-}u}$, $v\Delta RA_{SP,p\text{-}c}$, $m\Delta RA_{SP,p\text{-}c}$, $mRA_{SP,p}$; and

${}_{R}c_0$, ${}_{R}c_{jn}$ are the regression coefficients with R = 1, 2, 3 choosing the coefficient set depending on *N-MOS*.

Note that again the influence of **packet loss** is not covered separately but implicitly in the variance and the mean of the Δ Relative Approach (see also end of clause 6.4.2).

Tables 6.2 to 6.4 summarize the coefficients for the linear quadratic S-MOS regression algorithm depending on the previously calculated N-MOS used for training and validation of the algorithm.

Table 6.2: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS \leq N\text{-}MOS_{low} = 2{,}25$

Order	$_1c_{j0}$	$_1c_{j1}$ (ΔSNR)	$_1c_{j2}$ ($mRA_{SP,p}$)	$_1c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	$_1c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	$_1c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	$_1c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	6,4866	-0,0063	2,8784	3,5063	-0,0966	0,0767	-0,3738
2	-	-	-0,5483	0,4540	-0,3377	-0,0014	0,0168

Table 6.3: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS_{low} < N\text{-}MOS < N\text{-}MOS_{high}$

Order	$_2c_{j0}$	$_2c_{j1}$ (ΔSNR)	$_2c_{j2}$ ($mRA_{SP,p}$)	$_2c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	$_2c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	$_2c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	$_2c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	3,7991	0,0081	-0,0397	-0,4669	-0,5838	0,0862	-0,2850
2	-	-	0,0755	-0,1395	-0,0933	-0,0026	0,0086

Table 6.4: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS \geq N\text{-}MOS_{high} = 3{,}0$

Order	$_3c_{j0}$	$_3c_{j1}$ (ΔSNR)	$_3c_{j2}$ ($mRA_{SP,p}$)	$_3c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	$_3c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	$_3c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	$_3c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	5,4499	-0,0239	-1,4397	-2,2538	0,0256	-0,0097	-0,1391
2	-	-	0,2044	-0,4539	-0,0037	-0,0022	0,0043

6.5.3 Comparing Subjective and Objective S-MOS Results

The coefficients for the linear quadratic regression were determined in a similar way as for the N-MOS: the contributing parameters (ΔSNR, $mRA_{SP,p}$, $v\Delta RA_{SP,p\text{-}u}$, $m\Delta RA_{SP,p\text{-}u}$, $v\Delta RA_{SP,p\text{-}c}$, $m\Delta RA_{SP,p\text{-}c}$) were averaged for the six French sentences of a condition and then mapped to the auditory S-MOS.

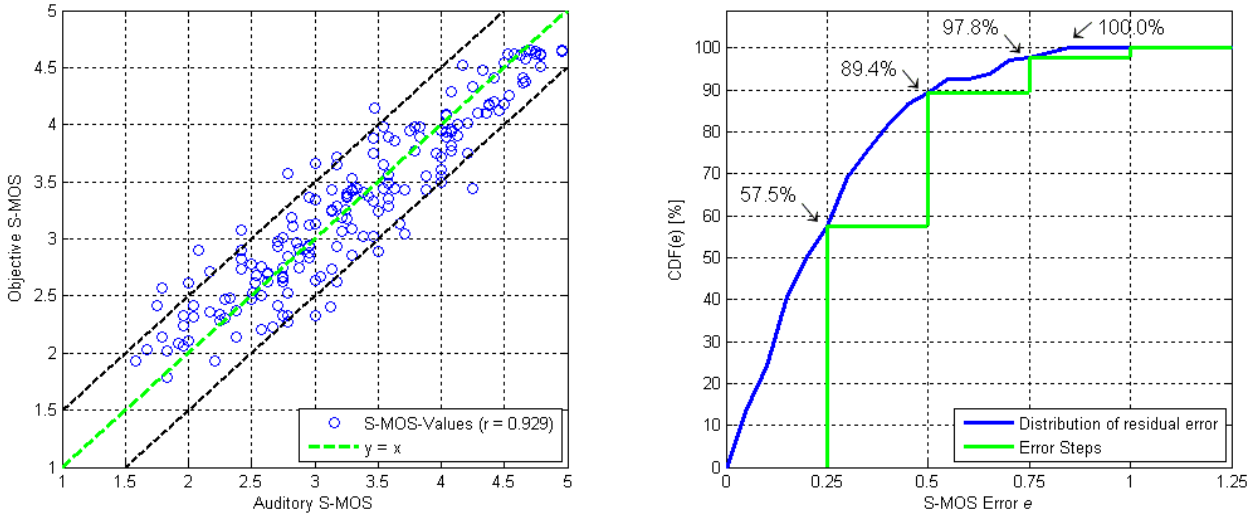


Figure 6.8: Left: Objectively calculated S-MOS versus auditory S-MOS; Right: CDF of residual error versus S-MOS error e

Similar to the N-MOS training all samples - independent of the network condition - were used.

The left hand graph in figure 6.8 shows that the per sample deviation between the subjective and objective S-MOS is higher than 0,5 MOS only for about 10 % of all (179) conditions. This results in an overall correlation of 92,9 %.

The right hand graph in figure 6.8 indicates the cumulated density function *CDF(e)* versus the S-MOS error *e* (see also equation 6.7). It also give an adaptive tolerance scheme indicating the *CDF(e)* values for *e = 0,25*, *e = 0,5*, *e = 0,75* and *e = 1*. The S-MOS error e is e.g. lower than 0,5 for 89 % of all conditions.

6.6 Objective G-MOS

6.6.1 Description of G-MOS Algorithm

The subjectively derived global quality is expected to be a combination of speech quality and noise quality. The expert analysis did not only extract those conditions of both languages which were somehow inconsistent. This test was also carried out to extract the main influencing parameters during the subjective ratings of N- and S-MOS. These parameters were then reproduced by the N-MOS and S-MOS calculation described in clauses 6.4 and 6.5 in order to model the human perception concerning speech and noise quality during the listening test.

Both, N-MOS and S-MOS calculation are optimized on the reproduction of the perceptual effects during the listening test. They were not optimized for "artificial" conditions like a highly modulated background noise together with a clean speech signal or vice versa. Those kinds of data were not considered in the listening test and were therefore also not considered by the objective model.

In accordance to the human perception, the new model first calculates the noise and speech quality. In a second step the overall quality is modelled. The G-MOS is therefore calculated by applying a linear, quadratic regression algorithm to *N-MOS* and *S-MOS*. The principle is shown in figure 6.9.

The corresponding *G-MOS* calculation equation is:

$$GMOS = c_0 + \sum_{j=1}^{2} c_{Sj} \cdot SMOS^{\,j} + \sum_{j=1}^{2} c_{Nj} \cdot NMOS^{\,j} \qquad (6.11)$$

where:

c_0, c_{Sj} and c_{Nj} are the coefficients for the linear quadratic regression;

j is the regression order index.

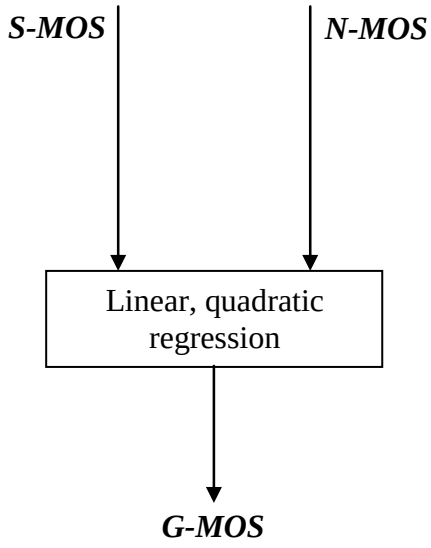


Figure 6.9: Block diagram of G-MOS calculation algorithm

Training and validation of the S-MOS regression were carried out using the regression coefficients in table 6.5.

Table 6.5: Coefficients for linear, quadratic G-MOS regression algorithm

Order	c_0	c_{Nj} (N-MOS)	c_{Sj} (S-MOS)
1	0,4539	0,5981	-0,1603
2	-	-0,0242	0,1143

6.6.2 Comparing subjective and objective G-MOS results

The coefficients for the G-MOS regression were derived by mapping the previously calculated objective N-MOS and S-MOS to the G-MOS results collected in the listening test using the linear, quadratic regression. The result compared to the auditory G-MOS is shown in figure 6.10.

The left hand graph in figure 6.10 shows that the per sample deviation between objective and auditory G-MOS is less than 0,5 MOS for most of the (179) conditions. The overall correlation is determined to 95,4 %.

The cumulated density function *CDF (e)* versus the G-MOS error *e* (see also equation 6.7) is shown on the right in figure 6.10. The CDF indicates that for 72 % of all conditions the G-MOS error e is less than 0,25 MOS and for nearly all conditions e is less than 0,5 MOS.

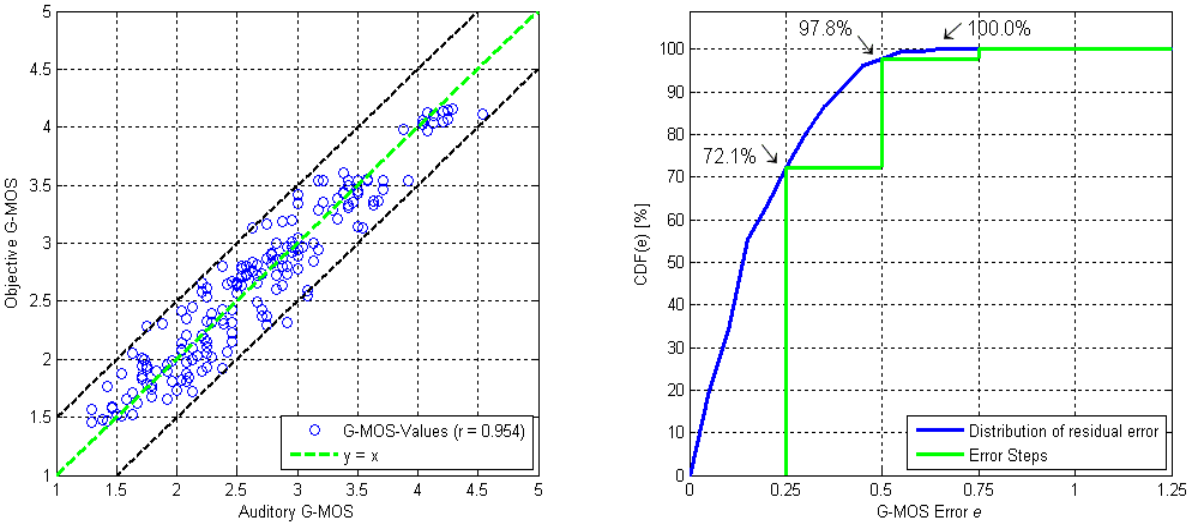


Figure 6.10: Left: Objectively calculated G-MOS versus auditory G-MOS; Right: CDF of residual error versus G-MOS error e

6.7 Comparison of the objective method results for Czech and French samples

Due to the differences between the Czech and the French listening tests already described in clause 5.5 the datasets for the model generation and validation were completely different in terms of level. While the level of the processed French signals was adjusted to 79 dB SPL, the level of the processed Czech signals was left unmodified. Therefore also the characteristic of the listening tests is different. The processed French signals are much louder (up to 16 dB) than the Czech ones - but all French samples are equal in terms of level: French listeners probably have not taken into account the absolute overall active speech level of the processed signal. It is very likely that in contrary Czech listeners took into account the different absolute overall active speech levels.

This also affects the results of the objectively calculated N-MOS, S-MOS and G-MOS values. As shown in figure 6.5 the level of the processed background noise is one influencing factor for the N-MOS calculation. This level is relatively high for all French samples. If the N-MOS is now calculated for the Czech samples using the regression coefficients acquired for the French sentences the resulting objective N-MOS scores are higher than the auditory scores. This is due to the lower background noise level of the Czech sentences. This could be expected: if a French listener would have listened to the Czech sentences among the French ones, he would have probably rated them with a higher N-MOS - due to the lower background noise level.

Figures 6.11 and 6.12 show the scatter plots for the objectively calculated N-MOS (for the selected French and Czech samples) versus the auditory N-MOS derived in the corresponding listening tests. The regression coefficients were optimized for the **French** dataset in both plots.

As already analysed in clause 6.4.3 the objective N-MOS correlates with 94,8 % to the results of the French listening test. Figure 6.12 shows that the objective N-MOS calculated for the Czech data using the French coefficients do not sufficiently correlate to the auditory results (correlation of 88,4 %). The results tends to be too good, which is mainly caused by the lower background noise level of the Czech samples. They would be assessed better by French listeners than the French samples with the higher level.

For another "cross check" the N-MOS regression algorithm is tuned on the *Czech* data, and the N-MOS scores are again calculated for the French and the Czech samples.

Note that for this training of the Czech data not only the selected (60) conditions were used, but also the selected conditions of network condition 1 (clean network). The disadvantage of this approach is that also conditions with very low signal levels and irreproducible ratings were considered. The big advantage is that the number of conditions increases from 60 to 120. This allows a higher numerical stability, especially for the S-MOS calculation, where the amount of conditions is separated in three groups according to the N-MOS. Using only a total of 60 Czech conditions would lead to a non-stable regression for the S-MOS due to the splitting in three groups. Only 20 conditions per group are too few to reliably calculate the 11 S-MOS regression coefficients.

The scatter plots are given in figures 6.13 and 6.14. They show that the objective results for the French data (figure 6.13) tend to be about 1 MOS lower than the auditory results (correlation of 82,1 %) whereas the objective N-MOS scores for the Czech samples correlate with 98 % to the auditory results (figure 6.14). Figure 6.13 indicates that a Czech listener would assess all French sample with a lower N-MOS - probably caused by the higher background noise level.

The conclusion of the scatter plot analysis is that:

- The new objective model is in principle applicable for both databases.
- Different regression coefficient sets are needed in order to reproduce the different level strategies used in the two datasets and listening tests.

Comparable analyses are carried out for S-MOS and G-MOS. The analyses results for the objective S-MOS are given in figure 6.15 to 6.18. Figures 6.15 and 6.18 show that if the regression coefficient set matching to the input data is used, the correlation is high (92,9 % for French data and 96,4 % for Czech data).

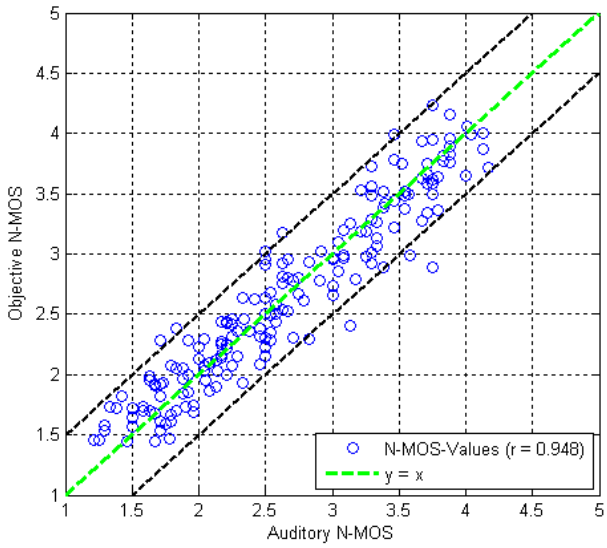


Figure 6.11: Objective vs. auditory N-MOS for French samples calculated with regression coefficients optimized for French data

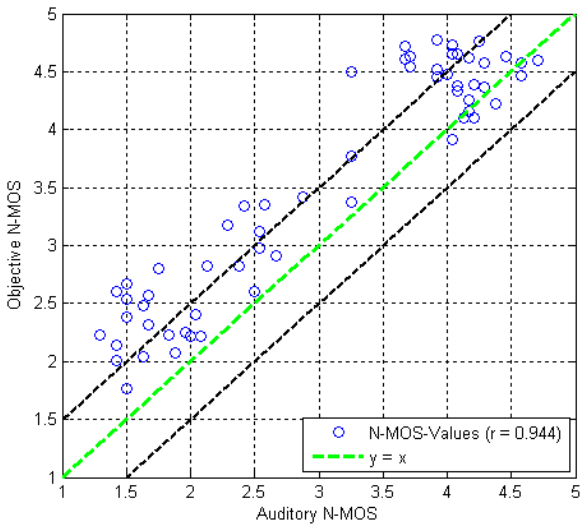


Figure 6.12: Objective vs. auditory N-MOS for Czech samples calculated with regression coefficients optimized for French data

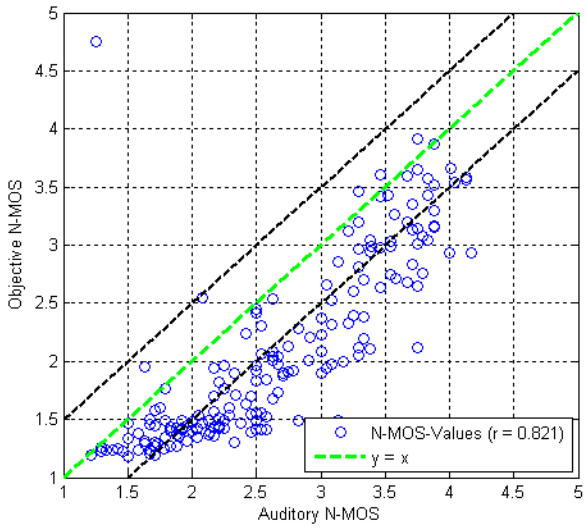


Figure 6.13: Objective vs. auditory N-MOS for French samples calculated with regression coefficients optimized for Czech data

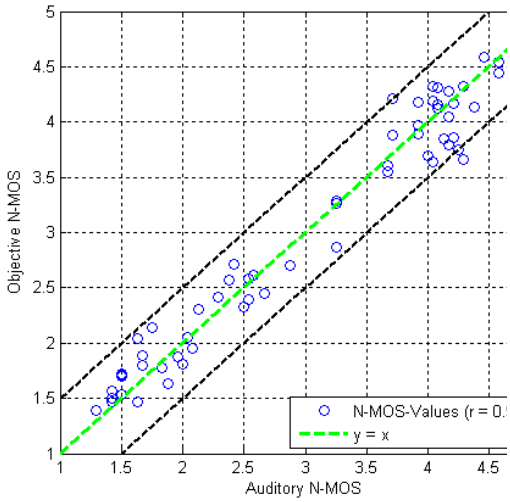


Figure 6.14: Objective vs. auditory N-MOS for Czech samples calculated with regression coefficients optimized for Czech data

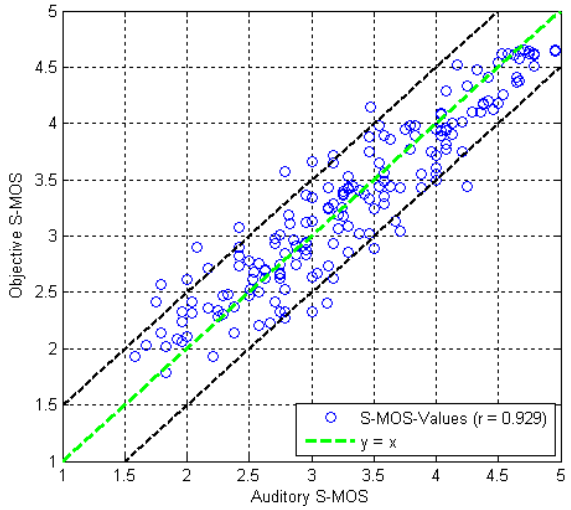


Figure 6.15: Objective vs. auditory S-MOS for French samples calculated with regression coefficients optimized for French data

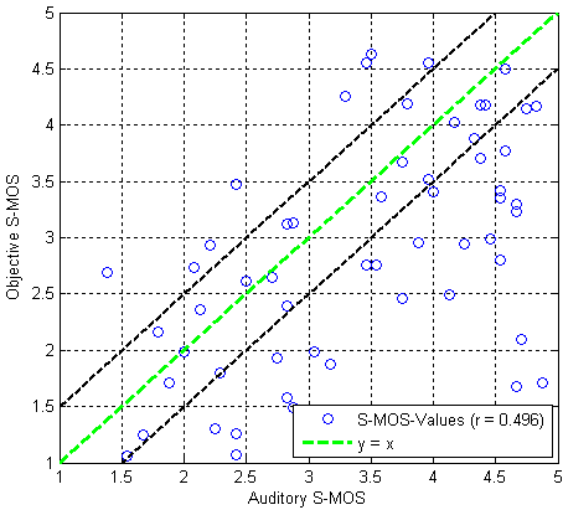


Figure 6.16: Objective vs. auditory S-MOS for Czech samples calculated with regression coefficients optimized for French data

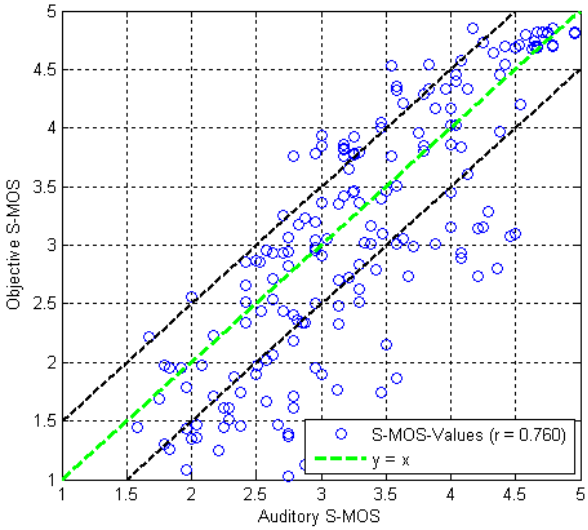


Figure 6.17: Objective vs. auditory S-MOS for French samples calculated with regression coefficients optimized for Czech data

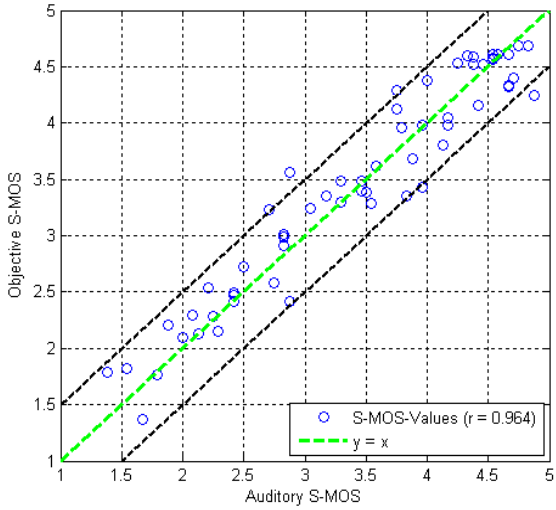


Figure 6.18: Objective vs. auditory S-MOS for Czech samples calculated with regression coefficients optimized for Czech data

If vice versa the coefficients of the other language are used, the correlation for the S-MOS decreases down to 46 %. Note that the objective S-MOS shown in figures 6.16 and 6.17 are based on the objective N-MOS which are also calculated using the "wrong" coefficient set of the other language. This "wrong" N-MOS may be the reason for ambiguous distribution of the objective S-MOS calculated for the Czech samples using the French coefficient compared to the auditory S-MOS. The objective S-MOS calculated for the French data using the Czech coefficients tend to be lower for auditory S-MOS lower than 3,5. For auditory S-MOS higher than 3,5 the objective S-MOS leads again to ambiguous results. One reason may again be the higher level of the French data.

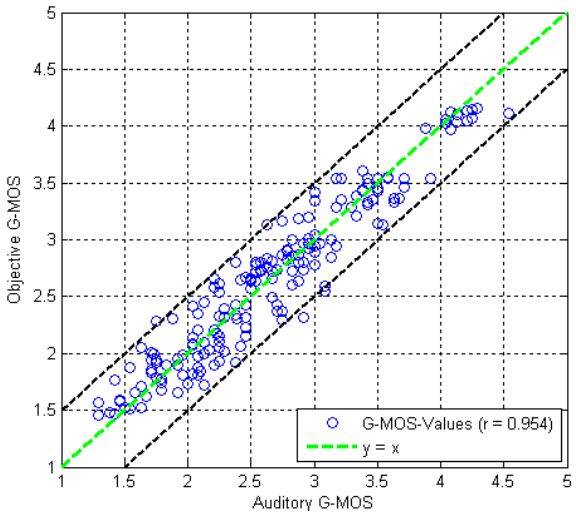


Figure 6.19: Objective vs. auditory G-MOS for French samples calculated with regression coefficients optimized for French data

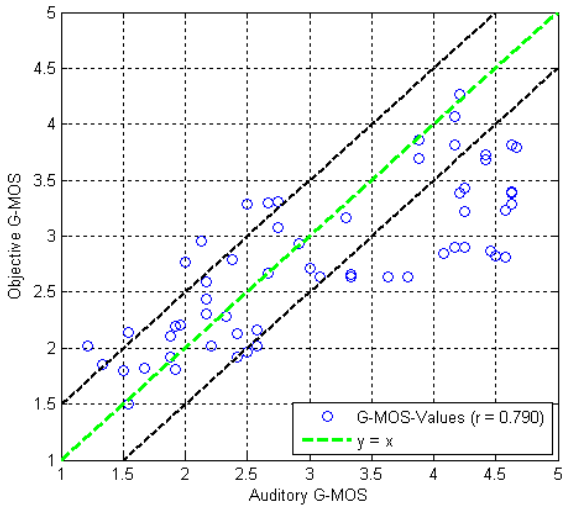


Figure 6.20: Objective vs. auditory G-MOS for Czech samples calculated with regression coefficients optimized for French data (N-MOS and S-MOS optimized for French data)

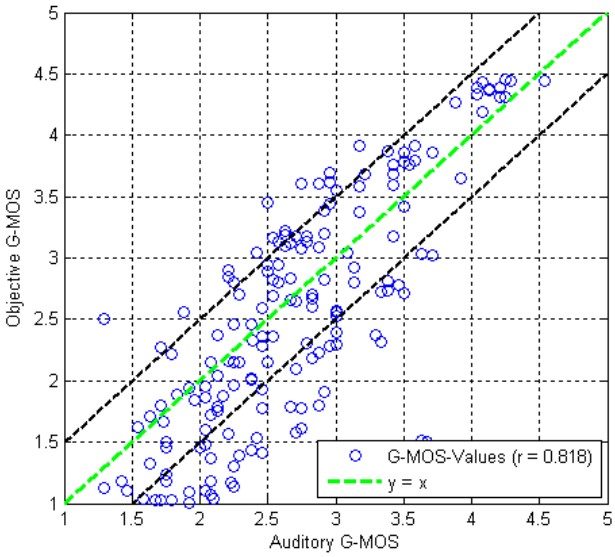


Figure 6.21: Objective vs. auditory G-MOS for French samples calculated with regression coefficients optimized for Czech data (N-MOS and S-MOS optimized for Czech data)

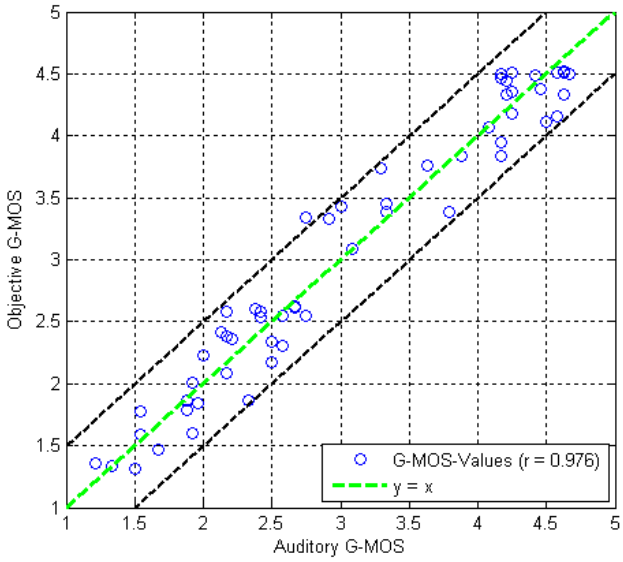


Figure 6.22: Objective vs. auditory G-MOS for Czech samples calculated with regression coefficients optimized for Czech data

The analysis for the objective G-MOS are shown with the same principle in figures 6.19 to 6.22. For both datasets using their optimized coefficient set the correlation is higher than 95 %. Note that the objective G-MOS calculation using the "wrong" coefficients was based on also the wrong N-MOS and S-MOS coefficients. This cumulated error leads to correlations of only 79 % and 81 % respectively.

6.8 Language Dependent Robustness of G-MOS

The listening tests carried out with French and Czech subjects used in principle the same database, but different level strategies. The French listening examples were all played back with the same active speech level of 79 dB SPL (see [i.22]), whereas the Czech listening examples had different play back levels reflecting the level and level differences after the processing (see also clause 5.5).

The listening tests in two different languages were originally carried out in order to verify language dependencies for the new objective method. Due to the different level strategies it is not possible to use the same regression coefficients of the new model for calculating N-MOS and S-MOS for both languages (see clause 5.5). However the G-MOS regressions for both, Czech and French data, can be used in order to verify, whether Czech and French listeners perhaps combined speech and noise quality to a "global" quality in the same way or if there are significant differences.

The G-MOS is therefore again calculated for Czech and French data. As input parameters N-MOS and S-MOS are used based on the individual ("correct") coefficient set. In other words, S-MOS and N-MOS for the French data are calculated using the corresponding French coefficients and vice versa. The G-MOS is then finally calculated using the coefficients of the other language each.

The results are given in figures 6.23 and 6.24. They show that the correlation between objective and auditory G-MOS is still higher than 94 % in both cases. This means, the final calculation of the G-MOS is very similar for both datasets and level strategies - if N-MOS and S-MOS consider all listening perception influences *including* levels. This indicates that - independent of the listening level strategy - Czech and French listeners combined speech and noise quality in a similar manner to the global quality.

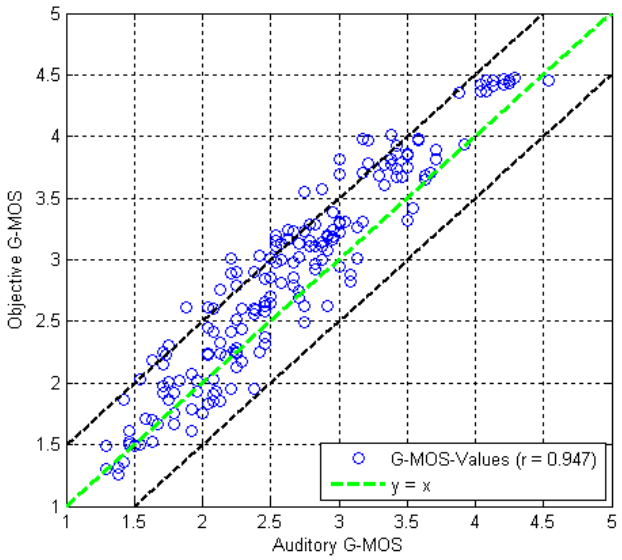


Figure 6.23: Objective vs. auditory G-MOS for French samples calculated with regression coefficients optimized for Czech data (N-MOS and S-MOS optimized for French data)

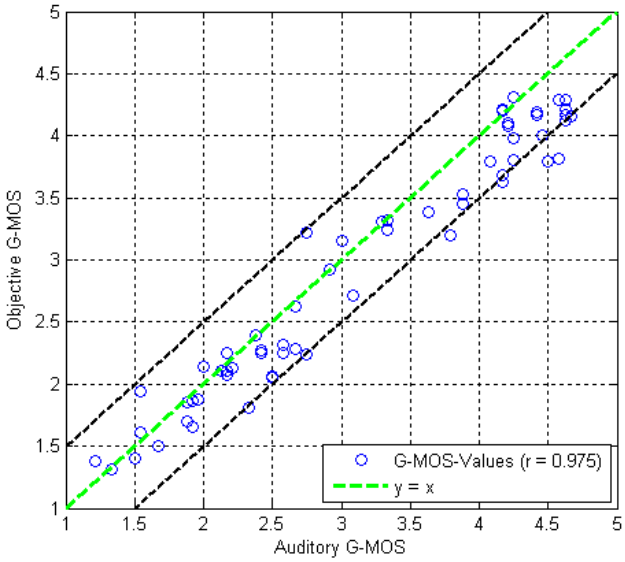


Figure 6.24: Objective vs. auditory G-MOS for Czech samples calculated with regression coefficients optimized for French data (N-MOS and S-MOS optimized for Czech data)

This effect can also be proved by comparing the G-MOS regression planes for the Czech and French coefficients as given in figures 6.25 and 6.26. The G-MOS regression planes for French and Czech coefficients are very similar. This indicates that the G-MOS dependency of S-MOS and N-MOS is similar for both languages.

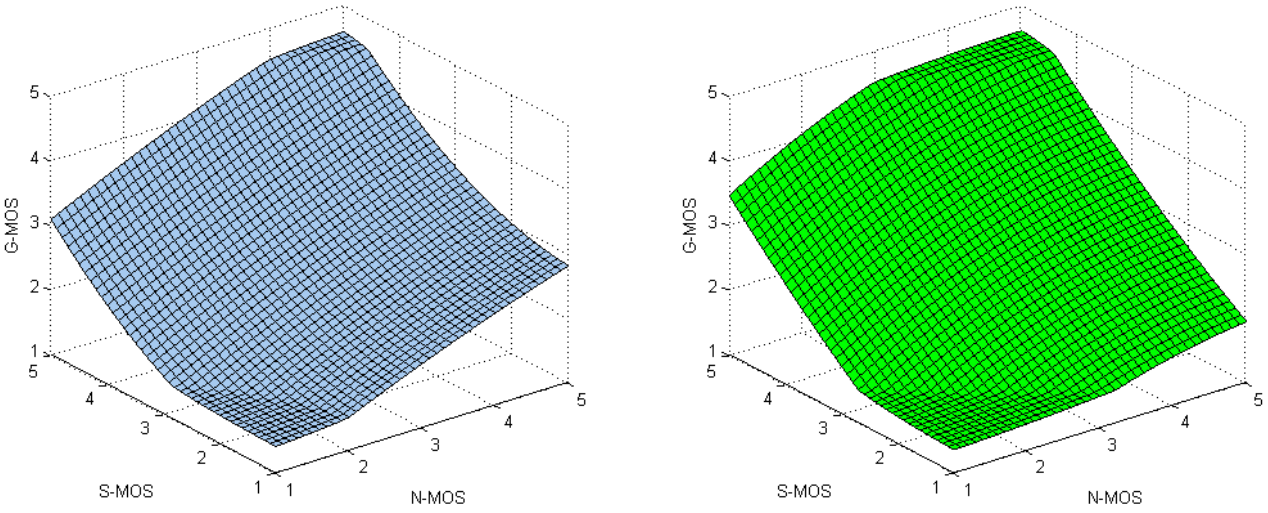


Figure 6.25: Comparison of French (left, blue) and Czech (right, green) regression plane

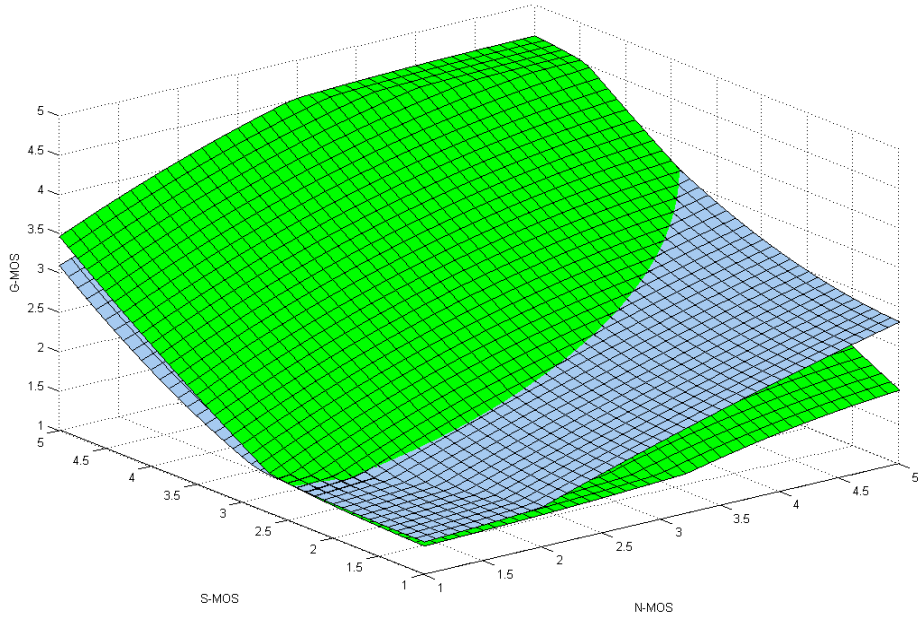


Figure 6.26: Comparison of French (blue) and Czech (green) regression plane

7 Validation of the Wideband Objective Test Method

7.1 Introduction

In order to validate the Objective Test Method results, 130 out of the 432 initial conditions per language were reserved to the validation activity. Due to the consistent problems related in clauses 4.3 and 4.4, the final validation conditions retained were 81 considering the French Database and 28 considering the Czech one. These conditions results are shown in annex F.

The process carried out to validate the Objective Test Method had the following steps:

1) Objective results obtaining: using the developed calculation algorithms, described in clauses 6.4, 6.5 and 6.6 (N-MOS, S-MOS and G-MOS) and the validation condition samples considering the language differentiation (coefficients for the linear, quadratic X-MOS regression algorithm).

2) Comparison between previously obtained objective results and the subjective results (see EG 202 396-2 [i.2]) considering all the validation condition samples and statistical evaluation. This evaluation will consist on the accuracy, monotonicity and consistency Test Method characterization. To carry out this characterization it will be used the statistical metrics:

   - Root Mean Square Error [i.24]: which measures the difference between values predicted by the algorithm and the auditory values to evaluate its accuracy,

$$RMSE = \sqrt{\frac{1}{N}\sum_{N} Perror[i]^2} \tag{7.1}$$

$$Perror(i) = MOS(i) - MOS_p(i) \tag{7.2}$$

   where:

   N is the number of samples, MOS(i) is the subjective MOS and MOS_p is the predicted MOS.

   - Pearson Correlation [i.24]: which measures the linear relationship between the algorithm performance and the subjective data, this coefficient varies from -1 to 1; a value of 1 shows that a linear equation describes the relationship perfectly and positively, with all data points lying on the same line and having the same behaviour; a score of -1 shows that all data points lie on a single line but having opposite behaviour; a value of 0 shows that a linear model is inappropriate - that there is no linear relationship between the variables,

$$R = \frac{\sum_{i=1}^{N} (Xi - \overline{X}) * (Yi - Y)}{\sqrt{\sum (Xi - \overline{X}^2) * \sqrt{\sum (Yi - \overline{Y}^2)}}} \tag{7.3}$$

   where:

   N is the number of samples, Xi denotes the subjective score MOS and Yi the objective one.

   The 95 % confidence interval for the correlation coefficient is determined using the Gaussian distribution which characterized the variable z (also called Fisher Z Transformation) [i.24] and its given by:

$$z \pm 2 \cdot \sigma_z \tag{7.4}$$

   where:

$$z = 0.5 \cdot \ln\left(\frac{1+R}{1-R}\right) \tag{7.5}$$

   and:

$$\sigma_z = \sqrt{\frac{1}{N-3}} \tag{7.6}$$

Otherwise, to calculate the 95 % confidence interval it is used the inverse Fisher Z Transformation [i.24]:

$$InverseZ = \frac{\exp(2z) - 1}{\exp(z) + 1} \tag{7.7}$$

The 95 confidence interval represents values for the Pearson correlation coefficient for which the difference between the parameter and the observed estimate is not statistically significant at the 5 % level [i.25].

- Spearman's Rank Correlation Coefficient [i.24]: which is a non-parametric measure of correlation - i.e. it assesses how well an arbitrary monotonic function could describe the relationship between two variables. This parameter varies from -1 to 1, as the Pearson Correlation:

$$\rho = 1 - \frac{6 \cdot \sum \frac{d_i^2}{N}}{N(N^2 - 1)} \tag{7.8}$$

  where:

  N is the number of samples and d the difference between each rank (position in an ordered table of conditions) of corresponding values of *x* and *y*.

- Kendall Tau Rank Correlation Coefficient [i.26]: which is used to measure the degree of correspondence between two rankings. If the agreement is perfect the coefficient value is 1, on the other hand if the disagreement is perfect the value is -1, if the rankings are completely independent , the coefficient has value 0:

$$\tau = \frac{4 \sum \frac{q_i}{N}}{N(N-1)} - 1 \tag{7.9}$$

  where:

  N is the number of samples and q_i the sum, over all samples, of samples ranked after the given sample by both rankings.

  - Residual Error Distribution [i.24]: which evaluates the consistency of the model using the Cumulative Density Function (CDF) applied to the error e:

$$e = |MOS_{auditory} - MOS_{objective}| \tag{7.10}$$

    The graphical representation of the CDF will show the number of conditions which yields a maximum residual error.

3) Results comparison per language.

The following clauses will be centred on the three different analyses.

7.2 All conditions results analysis

7.2.1 Comparing subjective and objective N-MOS results

All selected French and Czech conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective N-MOS is less than 0,5 MOS for nearly all (104 out of 109) conditions. This results in an overall Pearson correlation of 95,4 % (**R=0,954** very near to 1 with a confidence interval [0,933, 0,969]). The Spearman Correlation Coefficient is 0,952 and the Kendall Tau is 0,821, both of them are near to 1.

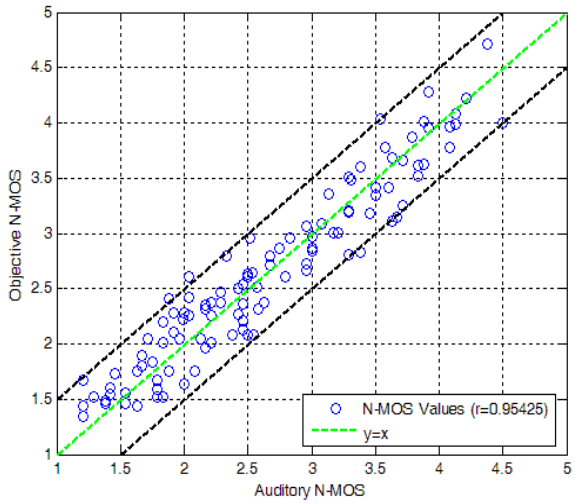


Figure 7.1: Objectively calculated N-MOS versus auditory N-MOS for validation conditions

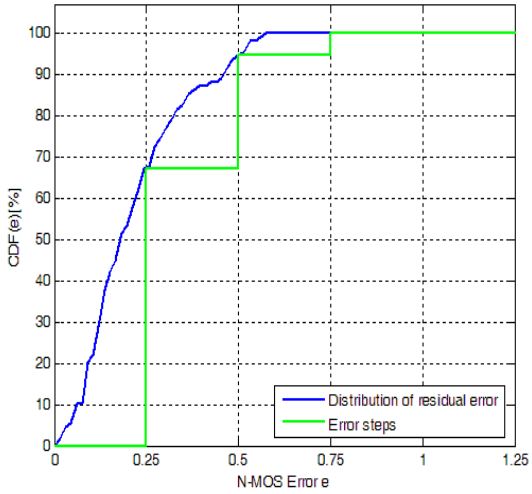


Figure 7.2: Objectively CDF of residual error versus N-MOS error e for validation conditions

For this situation, the **RMSE value is 0,255** and the distribution of the residual error is shown in figure 6.2 where the N-MOS error e is lower than 0,25 for approximately 67 % of the conditions and lower than 0,6 for 99 % for all conditions.

7.2.2 Comparing subjective and objective S-MOS results

All selected French and Czech conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective S-MOS is less than 0,5 MOS for nearly all (95 out of 109) conditions. This results in an overall correlation of 92 % (**R=0,920** near to 1 with a confidence interval [0,884, 0,945]). The Spearman Correlation Coefficient is 0,914 and the Kendall Tau is 0,749, both of them are near to 1.

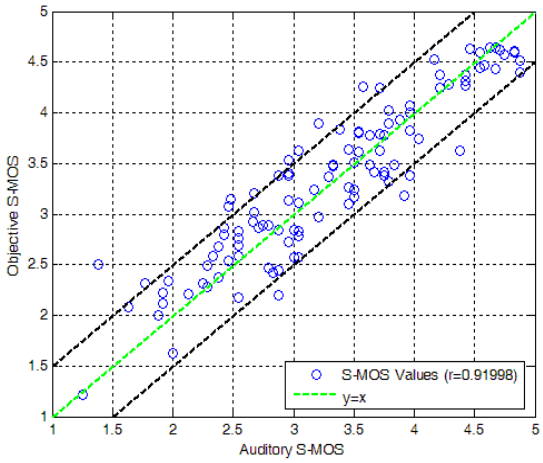


Figure 7.3: Objectively calculated S-MOS versus auditory S-MOS for validation conditions

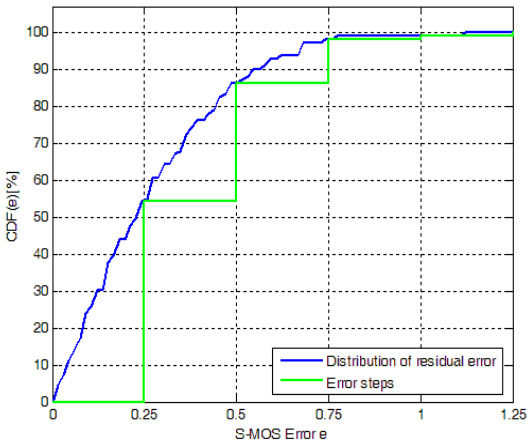


Figure 7.4: Objectively CDF of residual error versus S-MOS error e for validation conditions

For this situation, the **RMSE value is 0,338** and the distribution of the residual error is shown in figure 7.4 where the S-MOS error e is lower than 0,25 for approximately 55 % of the conditions and lower than 0,75 for 99 % for all conditions.

7.2.3 Comparing Subjective and Objective G-MOS Results

All selected French and Czech conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective G-MOS is less than 0,5 MOS for nearly all (102 out of 109) conditions. This results in an overall correlation of 94,5 % (**R=0,945** very near to 1 with a confidence interval [0,920, 0,962]). The Spearman Correlation Coefficient is 0,935 and the Kendall Tau is 0,793, both of them are near to 1.

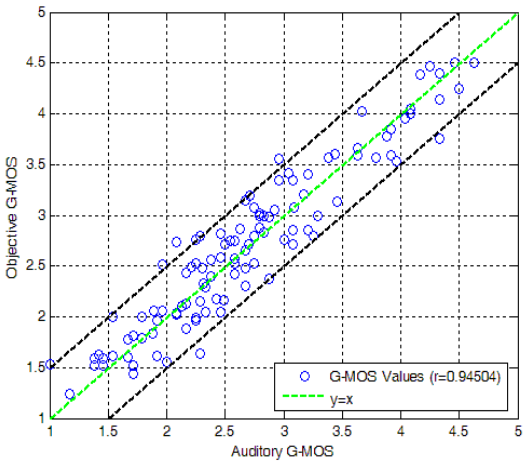


Figure 7.5: Objectively calculated G-MOS versus auditory G-MOS for validation conditions

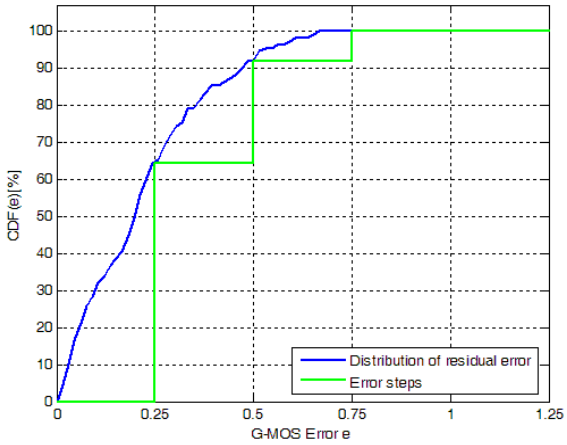


Figure 7.6: Objectively CDF of residual error versus G-MOS error e for validation conditions

For this situation, the **RMSE value is 0,272** and the distribution of the residual error is shown in figure 7.6 where the G-MOS error e is lower than 0,25 for approximately 65 % of the conditions and lower than 0,7 for 99 % for all conditions.

7.3 French Conditions Results Analysed

7.3.1 Comparing Subjective and Objective N-MOS Results

All selected French conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective N-MOS is less than 0,5 MOS for nearly all (79 out of 81) conditions. This results in an overall correlation of 95 % (**R=0,95** very near to 1 with a confidence interval [0,923, 0,968]). The Spearman Correlation Coefficient is 0,947 and the Kendall Tau is 0,810, both of them are near to 1.

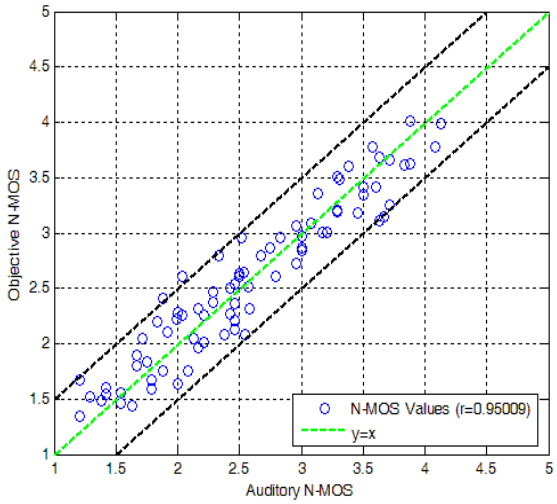


Figure 7.7: Objectively calculated N-MOS versus auditory N-MOS for French validation conditions

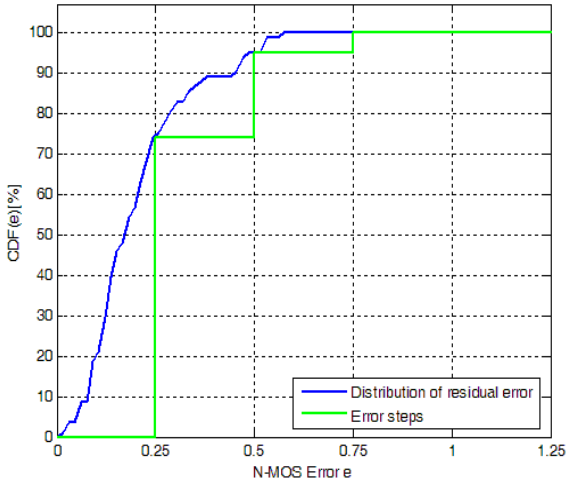


Figure 7.8: Objectively CDF of residual error versus N-MOS error e for French validation conditions

For this situation, the **RMSE value is 0,241** and the distribution of the residual error is shown in figure 7.8 where the N-MOS error e is lower than 0,25 for approximately 75 % of the conditions and lower than 0,6 for 99 % for all conditions.

7.3.2 Comparing Subjective and Objective S-MOS Results

All selected French conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective S-MOS is less than 0,5 MOS for nearly all (70 out of 81) conditions. This results in an overall correlation of 91,7 % (**R=0,917** near to 1 with a confidence interval [0,873, 0,946]). The Spearman Correlation Coefficient is 0,905 and the Kendall Tau is 0,747, both of them are near to 1.

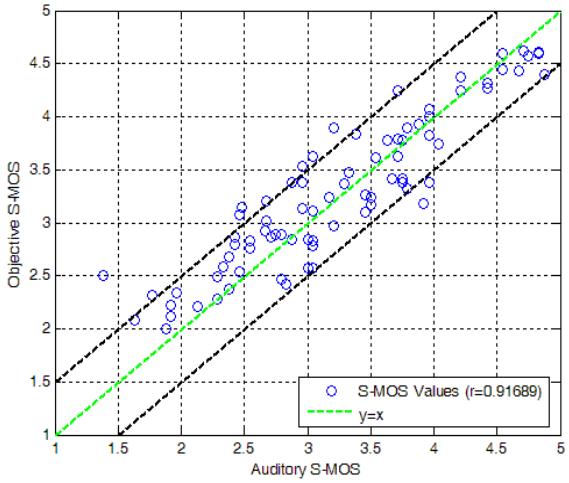


Figure 7.9: Objectively calculated S-MOS versus auditory S-MOS for French validation conditions

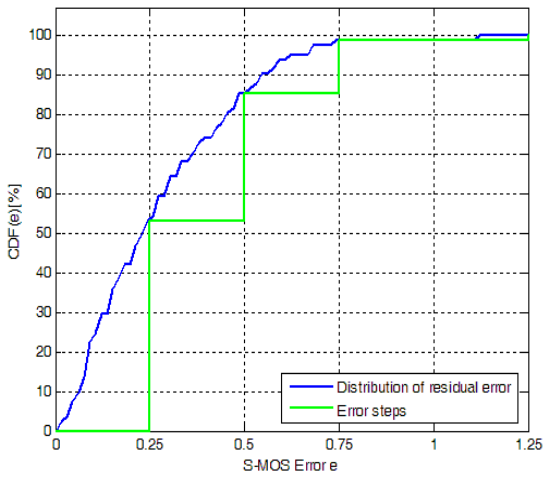


Figure 7.10: Objectively CDF of residual error versus S-MOS error e for French validation conditions

For this situation, the **RMSE value is 0,344** and the distribution of the residual error is shown in figure 7.10 where the S-MOS error e is lower than 0,25 for approximately 54 % of the conditions and lower than 0,75 for 99 % for all conditions.

7.3.3 Comparing subjective and objective G-MOS results

All selected French conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective G-MOS is less than 0,5 MOS for nearly all (79 out of 81) conditions. This results in an overall correlation of 93,9 % (**R=0,939** near to 1 with a confidence interval [0,906, 0,961]). The Spearman Correlation Coefficient is 0,925 and the Kendall Tau is 0,781, both of them are near to 1.

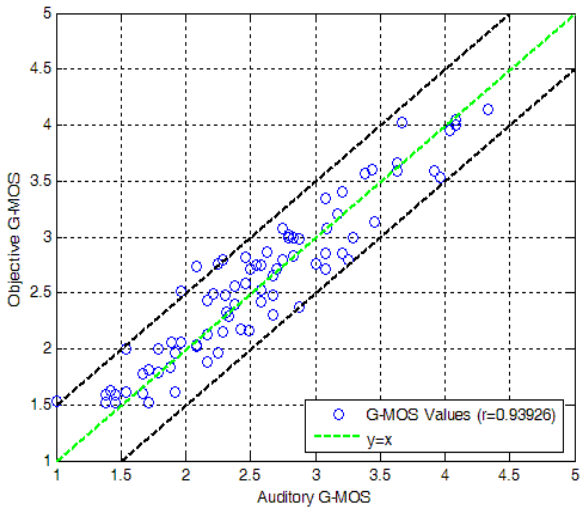


Figure 7.11: Objectively calculated G-MOS versus auditory G-MOS for French validation conditions

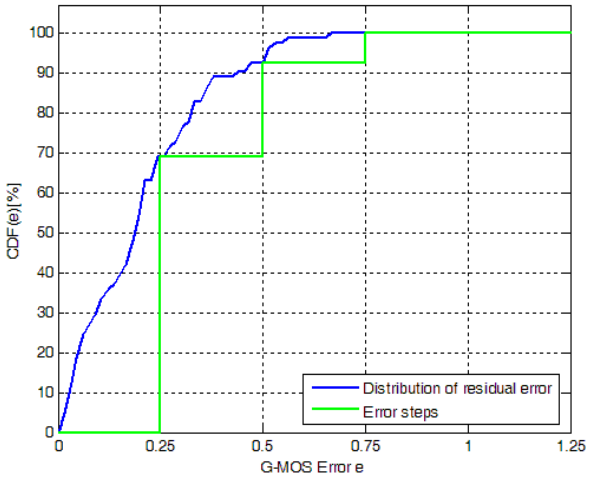


Figure 7.12: Objectively CDF of residual error versus G-MOS error e for French validation conditions

For this situation, the **RMSE value is 0,253** and the distribution of the residual error is shown in figure 7.12 where the G-MOS error e is lower than 0,25 for approximately 70 % of the conditions and lower than 0,65 for 99 % for all conditions.

7.4 Czech conditions results analysis

7.4.1 Comparing subjective and objective N-MOS results

All selected Czech conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective N-MOS is less than 0,5 MOS for nearly all (27 out of 28) conditions. This results in an overall correlation of 95,9 % (**R=0,959** very near to 1 with a confidence interval [0,912, 0,981]). The Spearman Correlation Coefficient is 0,961 and the Kendall Tau is 0,856, both of them are near to 1.

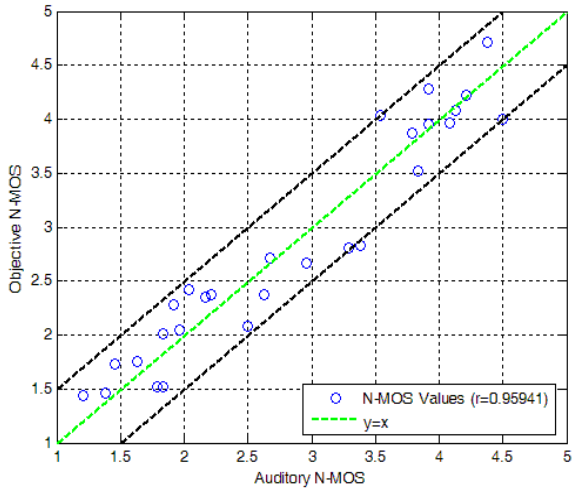


Figure 7.13: Objectively calculated N-MOS versus auditory N-MOS for Czech validation conditions

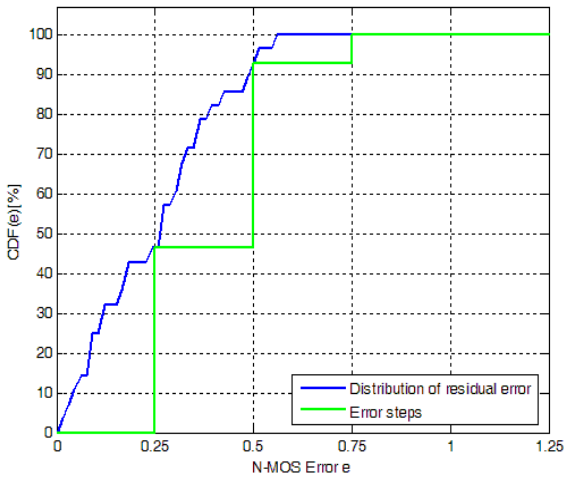


Figure 7.14: Objectively CDF of residual error versus N-MOS error e for Czech validation conditions

For this situation, the **RMSE value is 0,293** and the distribution of the residual error is shown in figure 7.14 where the N-MOS error e is lower than 0,25 for approximately 47 % of the conditions and lower than 0,55 for 99 % for all conditions.

7.4.2 Comparing subjective and objective S-MOS results

All selected Czech conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective S-MOS is less than 0,5 MOS for nearly all (25 out of 28) conditions. This results in an overall correlation of 94,3 % (**R=0,943** near to 1 with a confidence interval [0,879, 0,974]). The Spearman Correlation Coefficient is 0,930 and the Kendall Tau is 0,808, both of them are near to 1.

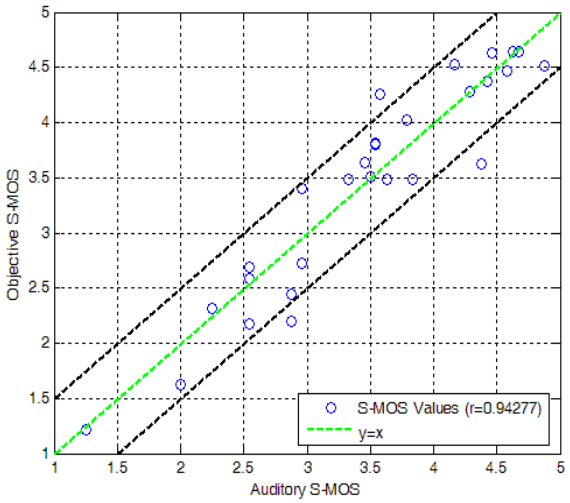


Figure 7.15: Objectively calculated S-MOS versus auditory S-MOS for Czech validation conditions

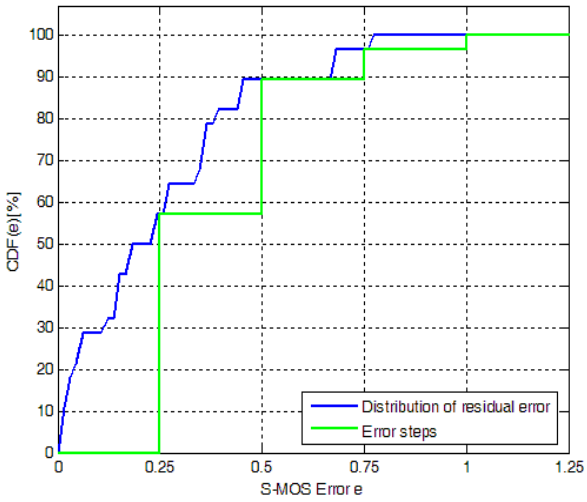


Figure 7.16: Objectively CDF of residual error versus S-MOS error e for Czech validation conditions

For this situation, the **RMSE value is 0,22** and the distribution of the residual error is shown in figure 7.16 where the N-MOS error e is lower than 0,25 for approximately 58 % of the conditions and lower than 0,77 for 99 % for all conditions.

7.4.3 Comparing Subjective and Objective G-MOS Results

All selected Czech conditions were used for this mapping - independent of the language and the network condition.

The following figure shows that the per sample deviation between the subjective and the objective G-MOS is less than 0,5 MOS for nearly all (25 out of 28) conditions. This results in an overall correlation of 94,9 % (**R=0,949** near to 1 with a confidence interval [0,892, 0,976]). The Spearman Correlation Coefficient is 0,935 and the Kendall Tau is 0,793, both of them are near to 1.

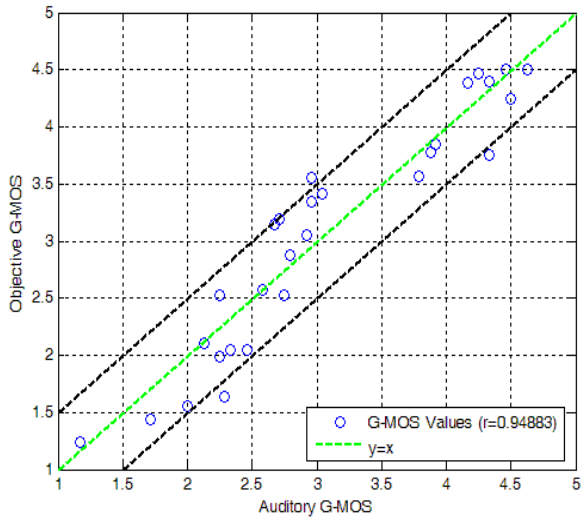


Figure 7.17: Objectively calculated S-MOS versus auditory G-MOS for Czech validation conditions

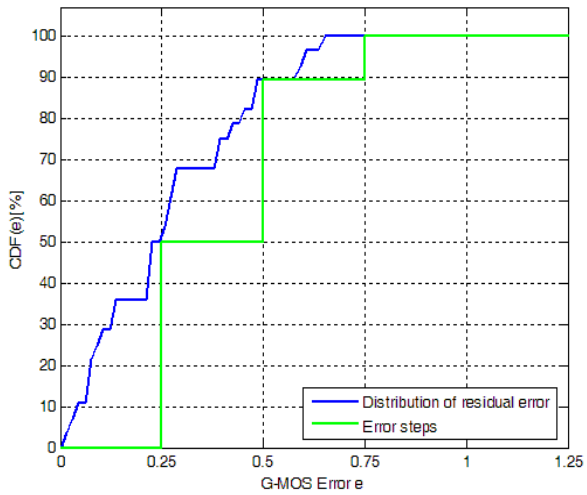


Figure 7.18: Objectively CDF of residual error versus G-MOS error e for Czech validation conditions

For this situation, the **RMSE value is 0,21** and the distribution of the residual error is shown in figure 7.18 where the G-MOS error e is lower than 0,25 for approximately 50 % of the conditions and lower than 0,65 for 99 % for all conditions.

8 Objective Model for Narrowband Applications

The objective model described in the clauses before in general is also applicable for narrowband scenarios. However some modifications have to be made in order to address the narrowband case which are described below.

The narrowband version of the model is based on an aurally-adequate analysis in order to best cover the listener's perception based on the previously carried out listening tests.

The test method is applicable for:

- narrowband handset and narrowband hands-free devices (in sending direction);
- noisy environments (stationary or non-stationary noise);
- different noise reduction algorithms;
- G.711, G.726, G.729A, iLBC, Speex HiQ / LQ and GSM FR, GSM EFR, and AMR narrowband coders;
- VoIP networks introducing packet loss.

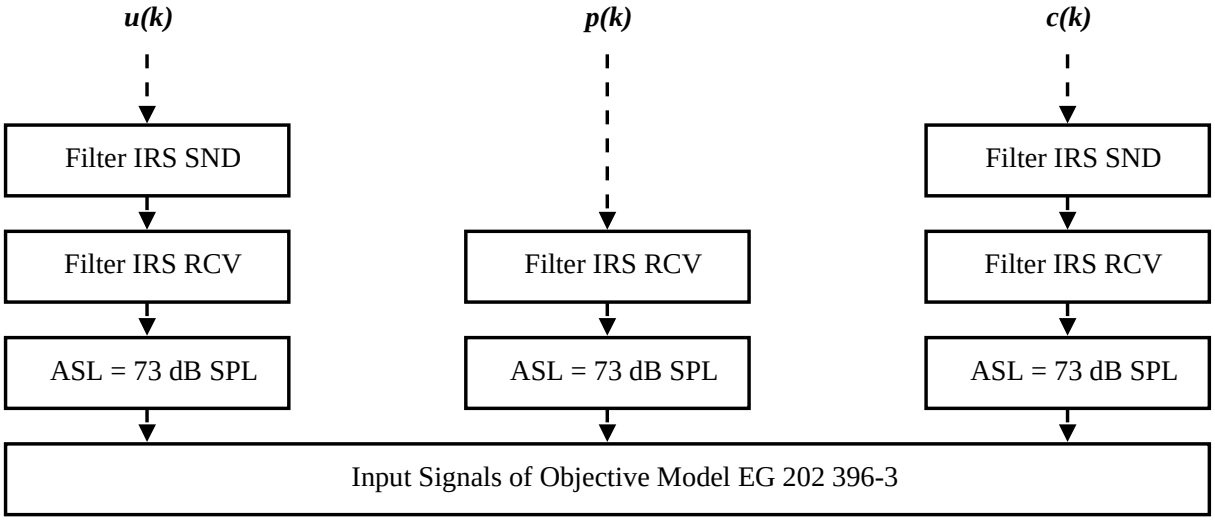
Due to the special sample generation process the method is only applicable for *electrically* recorded signals. The quality of terminals can therefore only be determined in sending direction.

8.1 File pre-processing

The processed signal ***p(k)*** is already calibrated to the active speech level (ASL) of -21 dB Pa / 73 dB SPL and filtered with an modified intermediate reference system (IRS) according to ITU-T Recommendation P.830 [i.28] in receiving direction for the presentation in the listening test. Exactly this signal is used in the objective model.

For the new narrowband mode, the clean speech and the unprocessed signal (***c(k)*** and ***u(k)***) are filtered with an modified IRS filter according to ITU-T Recommendation P. 830 [i.28] in sending and receiving direction. With this pre-processing step, all following analyses refer to a perfect transmission over a typical narrowband telephony network.

After filtering, both reference files are calibrated to the same active speech level like the processed signal. This refers to the acoustical presentation of the listening test. The overall pre-processing steps result in the following diagram.



8.2 Adaptation of the Calculations

The input parameters for the narrowband adapted model are the same as in the wideband mode. In the calculation of mean and variance from (Delta-) Relative Approach spectrograms, the limits of the frequency range are also adapted to the narrowband mode.

Table 8.1: Comparison of frequency ranges narrowband/wideband

	WB Data	**NB Data**
f_{min}	50 Hz	200 Hz
f_{max}	7 000 Hz	3 600 Hz

The three output MOS scores of the objective Model are calculated with a second order regression. The modified objective model needs to be mapped to the subjective data. The regression coefficients for the S-MOS are switched by the N-MOS value. For the narrowband model the switching thresholds for the N-MOS are modified slightly:

- $N\text{-}MOS_{low} = 2{,}48$.
- $N\text{-}MOS_{high} = 3{,}30$.

The new coefficients for S-, N- and G-MOS regression are given in the following tables.

Table 8.2: Coefficients for linear, quadratic N-MOS regression algorithm

Order	c_0	c_{BGN} (N_{BGN})	c_{j1} ($vRA_{BGN,\ u}$)	c_j ($vRA_{BGN,\ p}$)	c_{j3} ($v\Delta RA_{BGN,p-u}$)	c_{j4} ($m\Delta RA_{BGN,p-u}$)	c_{j5} ($mRA_{BGN,p}$)
1	0,1577	-0,0856	0,4040	1,6501	-1,2438	-1,5426	-3,0451
2	-	-	-0,1953	-0,2472	0,3400	2,1182	0,0277

Table 8.3: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS \leq N\text{-}MOS_{low} = 2{,}48$

order	$_1c_{j0}$	$_1c_{j1}$ (ΔSNR)	$_1c_{j2}$ ($mRA_{SP,p}$)	$_1c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	$_1c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	$_1c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	$_1c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	0,9875	-0,0253	5,6488	1,9201	-0,4286	0,1960	-1,3501
2	-	-	-1,5095	0,5245	1,9321	-0,0100	0,1565

Table 8.4: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS_{low} < N\text{-}MOS < N\text{-}MOS_{high}$

order	$_2c_{j0}$	$_2c_{j1}$ (ΔSNR)	$_2c_{j2}$ ($mRA_{SP,p}$)	$_2c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	$_2c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	$_2c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	$_2c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	2,6416	-0,0138	2,6584	1,4259	0,4318	0,1727	-0,5499
2	-	-	-0,5577	0,4662	1,0629	-0,0160	0,0227

Table 8.5: Coefficients for linear, quadratic S-MOS regression algorithm, N-MOS ≥ N-MOShigh = 3,30

order	$_3c_{j0}$	$_3c_{j1}$ (ΔSNR)	$_3c_{j2}$ ($mRA_{SP,p}$)	$_3c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	$_3c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	$_3c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	$_3c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	6,0014	-0,0094	0,5662	3,3369	0,3627	0,5344	-0,4843
2	-	-	-0,2103	1,1546	0,6943	-0,0544	0,0323

Table 8.6: Coefficients for linear, quadratic G-MOS regression algorithm

order	C_0	c_{Nj} (N-MOS)	c_{Sj} (S-MOS)
1	-1,0558	0,5215	0,8058
2	-	-0,0167	-0,0112

Annex A:
Detailed post evaluation of listening test results

Tables A.1 and A.2 contain the conditions and related auditory S-MOS, N-MOS and G-MOS for two tested languages. Also standard deviations for all MOS scores are given. The results for validation purposes are blinded.

Table A.1: Result of subjective experiment results -experts listening:
Samples *not retained* from the French database in addition to the NII condition (hs - handset, hf - hands-free, f - female, m - male speaker)

Extension French	Condition	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	FRENCH						Comment
									MOS Speech	MOS Noise	MOS Global	STD Speech	STD Noise	STD Global	
19	19	Lux_Car	hs	f	AMR _NI	yes	Smooth	18	4,08	3,42	3,46	0,58	0,58	0,59	Wideband noise
145	145	Crossroads	hf	f	AMR _NI	no	Sharp	9							Not consistent, Sample 4 loud Samples 3 and 6 too low speech level
151	151	Crossroads	hf	f	AMR _NI	yes	Smooth	9	2,96	1,54	1,71	1,37	0,66	0,81	Inconsistent Levels of Samples
157	157	Crossroads	hf	f	AMR _NI	yes	Sharp	9							Not consistent, Sample 4 loud Samples 3 and 6 too low speech level
160	160	Crossroads	hf	f	AMR _NI	yes	Sharp	18	1,88	1,63	1,54	1,03	0,71	0,78	Inconsistent Levels of samples
162	162	Crossroads	hf	f	AMR _NIII	yes	Sharp	18	1,38	1,54	1,13	0,71	0,93	0,45	Inconsistent, amplification 2and 6 too high
168	168	Crossroads	hs	m	AMR _NIII	no	Smooth	9	2,96	2,42	2,29	1,27	0,88	0,91	Inconsistent, noise 2 and 6 too high, not visible in the gains but audible
169	169	Crossroads	hs	m	AMR _NI	no	Smooth	18	3,08	2,92	2,75	1,06	1,18	1,11	Inconsistent Levels of samples
175	175	Crossroads	hs	m	AMR _NI	no	Sharp	18	3,21	3,17	2,88	1,06	1,05	0,85	Inconsistent Levels of samples
178	178	Crossroads	hs	m	AMR _NI	yes	Smooth	9	3,96	2,92	3,13	0,81	0,93	1,03	Inconsistent Levels of samples
180	180	Crossroads	hs	m	AMR _NIII	yes	Smooth	9	2,83	2,63	2,5	1,17	0,97	0,98	Inconsistent, noise 2 and 6 too high, visible in the gains (up to 5 dB)
183	183	Crossroads	hs	m	AMR _NIII	yes	Smooth	18	3,25	3	2,79	1,15	1,29	1,22	Inconsistent, noise 2 and 6 too high, visible in the gains (up to 5 dB)

Extension French	Condition	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	FRENCH						Comment
									MOS Speech	MOS Noise	MOS Global	STD Speech	STD Noise	STD Global	
189	189	Crossroads	hs	m	AMR _NIII	yes	Sharp	18	3,25	3,46	2,67	1,15	0,93	0,87	Inconsistent, noise 2 and 6 too high, visible in the gains (up to 5 dB)
193	193	Crossroads	hf	m	AMR _NI	no	Smooth	9							Bad S/N sounds unprocessed speech low 3 and 6, not intelligible
199	199	Crossroads	hf	m	AMR _NI	no	Sharp	9							Bad S/N sounds unprocessed speech low 3 and 6, not intelligible
208	208	Crossroads	hf	m	AMR _NI	yes	Smooth	18	2,67	1,96	2,04	1,2	0,91	0,86	Inconsistent Levels of samples
211	211	Crossroads	hf	m	AMR _NI	yes	Sharp	9	2,88	1,75	2,13	1,33	0,94	0,9	Inconsistent Levels of samples
214	214	Crossroads	hf	m	AMR _NI	yes	Sharp	18	1,92	2,13	1,55	1,02	1,12	0,71	Inconsistent Levels of samples
216	216	Crossroads	hf	m	AMR _NIII	yes	Sharp	18	1,92	1,67	1,54	0,88	0,7	0,59	Example 2 too loud
279	252	Road	hs	m	AMR _NIII	no	Smooth	18	2,31	2,21	2,09	0,8	0,98	0,78	Example 2 too loud
357	303	Office	hf	f	G722_NIII	no	Smooth	9							Poor S/N, packet loss determines speech quality, processing errors in sample 6
373	319	Office	hf	f	G722_NI	yes	Sharp	9							Processing noise, processing errors in sample 4
406	352	Office	hf	m	G722_NI	no NSA	no NSA	no NSA							Fair S/N processing errors in sample 6
423	369	Office	hf	m	G722_NIII	yes	Smooth	9	4,25	2,53	2,79	0,99	0,77	0,88	6 examples with packet loss, Result Speech and noise influenced by packet loss, processing noise
447	393	Pub	hs	f	G722_NIII	no	Sharp	18							Packet loss during speech determines speech quality, highly modulated BGN, processing errors in sample 4
478	424	Pub	hs	m	G722_NI	yes	Smooth	18	3,17	2,41	2,5	1,13	0,66	0,78	Strong amplification difference

Extension French	Condition	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	FRENCH						Comment
									MOS Speech	MOS Noise	MOS Global	STD Speech	STD Noise	STD Global	
480	426	Pub	hs	m	G722_NIII	yes	Smooth	18	2,58	2,33	2,08	1,02	0,87	0,88	Inconsistent levels
484	430	Pub	hs	m	G722_NI	yes	Sharp	18	2,92	2	1,96	1,06	0,83	0,81	Strong amplification difference

Table A.2: Result of subjective experiment results -experts listening: Samples *selected* from the Czech database (hs - handset, hf - hands-free, f - female, m - male speaker)

Condition	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	CZECH						Listening level dB SPL
								MOS Speech	MOS Noise	MOS Global	STD Speech	STD Noise	STD Global	
1	Lux_Car	hs	f	AMR _NI	no NSA	no NSA	no NSA							72,8
10	Lux_Car	hs	f	AMR _NI	no	Sharp	9							69,33
18	Lux_Car	hs	f	AMR _NIII	yes	Smooth	9	2,42	3,25	2,58	0,72	0,53	0,65	69,02
22	Lux_Car	hs	f	AMR _NI	yes	Sharp	9							70,18
24	Lux_Car	hs	f	AMR _NIII	yes	Sharp	9							71,41
25	Lux_Car	hs	f	AMR _NI	yes	Sharp	18	3,29	3,92	3,33	0,86	0,58	0,82	71,85
28	Lux_Car	hf	f	AMR _NI	no NSA	no NSA	no NSA	3,54	1,5	2,17	0,88	0,66	0,87	78,06
31	Lux_Car	hf	f	AMR _NI	no	Smooth	9							70,3
37	Lux_Car	hf	f	AMR _NI	no	Sharp	9							71,44
40	Lux_Car	hf	f	AMR _NI	no	Sharp	18	2,83	2,42	2,38	0,64	0,72	0,49	71,5
43	Lux_Car	hf	f	AMR _NI	yes	Smooth	9							69,85
49	Lux_Car	hf	f	AMR _NI	yes	Sharp	9							70,79
51	Lux_Car	hf	f	AMR _NIII	yes	Sharp	9	2,25	1,75	1,88	0,61	0,61	0,54	70,74
55	Lux_Car	hs	m	AMR _NI	no NSA	no NSA	no NSA	3,75	2,88	3,29	0,61	0,9	0,55	74,86
61	Lux_Car	hs	m	AMR _NI	no	Smooth	18	3,79	4,17	3,88	0,78	0,48	0,54	72,34
73	Lux_Car	hs	m	AMR _NI	yes	Smooth	18	4,17	4,08	4,17	0,76	0,41	0,38	70,59
76	Lux_Car	hs	m	AMR _NI	yes	Sharp	9	4,42	3,25	3,88	0,5	0,61	0,61	69,24
79	Lux_Car	hs	m	AMR _NI	yes	Sharp	18							73,81
81	Lux_Car	hs	m	AMR _NIII	yes	Sharp	18							71,64
82	Lux_Car	hf	m	AMR _NI	no NSA	no NSA	no NSA	3,58	1,42	2,17	1,14	0,58	0,82	78,13
84	Lux_Car	hf	m	AMR _NIII	no NSA	no NSA	no NSA	2,29	1,5	1,67	0,86	0,59	0,56	77,71
85	Lux_Car	hf	m	AMR _NI	no	Smooth	9	3,96	2,54	2,92	0,62	0,66	0,65	69,77
87	Lux_Car	hf	m	AMR _NIII	no	Smooth	9	2,13	2,13	1,96	0,74	0,74	0,62	70,16
97	Lux_Car	hf	m	AMR _NI	yes	Smooth	9	3,88	2,29	3,08	0,8	0,69	0,72	69,08
103	Lux_Car	hf	m	AMR _NI	yes	Sharp	9							69,71
111	Crossroads	hs	f	AMR _NIII	no NSA	no NSA	no NSA	2,21	1,88	1,88	0,78	0,61	0,61	71,23
120	Crossroads	hs	f	AMR _NIII	no	Sharp	9	2	1,96	1,92	0,72	0,55	0,41	69,34
138	Crossroads	hf	f	AMR _NIII	no NSA	no NSA	no NSA	1,79	1,29	1,33	0,88	0,46	0,56	73,3
174	Crossroads	hs	m	AMR _NIII	no	Sharp	9	2,42	2,38	2	0,93	0,58	0,66	72,27
195	Crossroads	hf	m	AMR _NIII	no	Smooth	9	1,38	1,42	1,21	0,65	0,58	0,41	69,57
201	Crossroads	hf	m	AMR _NIII	no	Sharp	9							70,94

Condition	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	CZECH						Listening level dB SPL
								MOS Speech	MOS Noise	MOS Global	STD Speech	STD Noise	STD Global	
217	Road	hs	f	AMR _NI	no NSA	no NSA	no NSA	2,5	1,67	1,92	0,83	0,64	0,5	72
219	Road	hs	f	AMR _NIII	no NSA	no NSA	no NSA	1,67	1,5	1,5	0,64	0,51	0,59	72,26
243	Road	hs	f	AMR _NIII	yes	Sharp	18	1,54	2,58	1,54	0,66	0,88	0,59	70,91
271	Office	hs	f	G722_NI	no NSA	no NSA	no NSA	4,54	4	4,25	0,59	0	0,44	74,23
274	Office	hs	f	G722_NI	no	Smooth	9	4,58	4,17	4,42	0,58	0,38	0,5	72,49
276	Office	hs	f	G722_NIII	no	Smooth	9							73,68
277	Office	hs	f	G722_NI	no	Smooth	18							73,06
280	Office	hs	f	G722_NI	no	Sharp	9	4,58	3,71	4,21	0,58	0,46	0,66	75,22
282	Office	hs	f	G722_NIII	no	Sharp	9	3,83	3,92	3,79	0,87	0,5	0,78	73,6
283	Office	hs	f	G722_NI	no	Sharp	18	4,33	4,04	4,17	0,48	0,36	0,56	72,64
285	Office	hs	f	G722_NIII	no	Sharp	18	2,71	3,71	2,75	1,12	0,46	1,03	74,77
286	Office	hs	f	G722_NI	yes	Smooth	9	4,38	4,08	4,42	0,58	0,28	0,58	74,81
289	Office	hs	f	G722_NI	yes	Smooth	18							73,77
291	Office	hs	f	G722_NIII	yes	Smooth	18	2,42	4,29	2,67	1,25	0,55	0,92	74,05
292	Office	hs	f	G722_NI	yes	Sharp	9							75,57
295	Office	hs	f	G722_NI	yes	Sharp	18	4,38	4,04	4,17	0,71	0,46	0,56	75,24
297	Office	hs	f	G722_NIII	yes	Sharp	18							72,38
325	Office	hs	m	G722_NI	no NSA	no NSA	no NSA	4,54	4,04	4,46	0,72	0,55	0,72	75,74
328	Office	hs	m	G722_NI	no	Smooth	9	4,54	4,58	4,63	0,72	0,5	0,49	74,1
331	Office	hs	m	G722_NI	no	Smooth	18							72
334	Office	hs	m	G722_NI	no	Sharp	9							75,41
336	Office	hs	m	G722_NIII	no	Sharp	9	3,75	4,38	4,08	0,94	0,49	0,83	74,73
337	Office	hs	m	G722_NI	no	Sharp	18	4,67	4,21	4,63	0,64	0,41	0,49	71,98
339	Office	hs	m	G722_NIII	no	Sharp	18	4,13	4,08	4,17	0,8	0,41	0,64	73,17
340	Office	hs	m	G722_NI	yes	Smooth	9	4,75	4,13	4,67	0,44	0,45	0,48	75,37
342	Office	hs	m	G722_NIII	yes	Smooth	9	4	4,29	4,21	0,88	0,46	0,51	74,51
343	Office	hs	m	G722_NI	yes	Smooth	18	4,25	4,46	4,25	0,68	0,72	0,94	74,52
346	Office	hs	m	G722_NI	yes	Sharp	9	4,83	4,21	4,63	0,48	0,51	0,58	75,38
348	Office	hs	m	G722_NIII	yes	Sharp	9	3,17	4,17	3,33	1,05	0,38	0,92	74,36
349	Office	hs	m	G722_NI	yes	Sharp	18	4,46	4,71	4,58	0,59	0,46	0,5	74,55
351	Office	hs	m	G722_NIII	yes	Sharp	18	4,67	4,58	4,63	0,48	0,5	0,49	75,26
354	Office	hf	m	G722_NIII	no NSA	no NSA	no NSA	4,17	3,25	3,63	0,64	0,68	0,71	69,13
361	Office	hf	m	G722_NI	no	Sharp	9	4,71	3,67	4,25	0,46	0,56	0,53	70,54
367	Office	hf	m	G722_NI	yes	Smooth	9	4,88	3,92	4,5	0,34	0,5	0,51	69,88
373	Office	hf	m	G722_NI	yes	Sharp	9							70,68
375	Office	hf	m	G722_NIII	yes	Sharp	9	2,88	3,67	3	0,85	0,7	0,83	70,53
376	Office	hf	m	G722_NI	yes	Sharp	18	4,67	4,25	4,58	0,56	0,61	0,58	69,67
379	Pub	hs	f	G722_NI	no NSA	no NSA	no NSA							69,94
384	Pub	hs	f	G722_NIII	no	Smooth	9							70,95
385	Pub	hs	f	G722_NI	no	Smooth	18	2,75	2,5	2,5	0,68	0,59	0,51	70,71
387	Pub	hs	f	G722_NIII	no	Smooth	18	2,88	2,08	2,33	0,8	0,58	0,7	69,22
388	Pub	hs	f	G722_NI	no	Sharp	9	3,29	1,42	2,13	0,95	0,58	0,61	74,31

Condition	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	CZECH						Listening level dB SPL
								MOS Speech	MOS Noise	MOS Global	STD Speech	STD Noise	STD Global	
390	Pub	hs	f	G722_NIII	no	Sharp	9							72,13
391	Pub	hs	f	G722_NI	no	Sharp	18	2,83	2,04	2,21	0,82	0,62	0,72	70,61
393	Pub	hs	f	G722_NIII	no	Sharp	18							72,13
394	Pub	hs	f	G722_NI	yes	Smooth	9	3,46	1,67	2,42	0,83	0,56	0,58	72,84
396	Pub	hs	f	G722_NIII	yes	Smooth	9							69,49
400	Pub	hs	f	G722_NI	yes	Sharp	9	3,04	1,63	2,42	0,69	0,58	0,72	73,24
403	Pub	hs	f	G722_NI	yes	Sharp	18	2,08	2,54	2,17	0,83	0,93	0,64	75,43
406	Pub	hs	m	G722_NI	no NSA	no NSA	no NSA	3,5	1,63	2,5	0,66	0,58	0,72	70,97
408	Pub	hs	m	G722_NIII	no NSA	no NSA	no NSA	1,88	1,5	1,54	0,74	0,51	0,59	70,62
409	Pub	hs	m	G722_NI	no	Smooth	9	3,46	2	2,67	0,66	0,72	0,48	69,39
415	Pub	hs	m	G722_NI	no	Sharp	9							72
421	Pub	hs	m	G722_NI	yes	Smooth	9	3,96	1,83	2,75	0,62	0,48	0,68	70,45
424	Pub	hs	m	G722_NI	yes	Smooth	18	2,83	2,67	2,58	0,82	0,7	0,58	69,35
427	Pub	hs	m	G722_NI	yes	Sharp	9							70,89
432	Pub	hs	m	G722_NIII	yes	Sharp	18							69,19

Annex B:
Results of PESQ and TOSQA2001 - Analysis of EG 202-396-2 database

Although it is known that neither PESQ (ITU-T Recommendation P.862.2 [i.18]) nor TOSQA2001 [i.19] are capable to predict MOS values for scenarios with speech being transmitted and processed together with background noise some data were analyzed in order to document these limitations. This data set consists of 32 conditions (out of 179 overall selected conditions with known MOS values) with French speech, different types of packet loss, voice coders, background noise and noise reduction.

Table B.1: Test set chosen from EG 202-396-2 database to be analysed with PESQ and TOSQA2001

Extension French	Noise	Recording	Speaker	Network	NSA	Sharp/ smooth	dB	MOS Speech	MOS Noise	MOS Global
3	Lux_Car	hs	f	AMR_NIII	no NSA	no NSA	no NSA	3,63	3,13	3,08
7	Lux_Car	hs	f	AMR_NI	no	Smooth	18	4,21	3,71	3,63
28	Lux_Car	hf	f	AMR_NI	no NSA	no NSA	no NSA	3,79	2,25	2,54
54	Lux_Car	hf	f	AMR_NIII	yes	Sharp	18	2	1,92	1,63
55	Lux_Car	hs	m	AMR_NI	no NSA	no NSA	no NSA	4,33	3,04	3,21
57	Lux_Car	hs	m	AMR_NIII	no NSA	no NSA	no NSA	3,46	3	2,79
82	Lux_Car	hf	m	AMR_NI	no NSA	no NSA	no NSA	4	2,21	2,54
87	Lux_Car	hf	m	AMR_NIII	no	Smooth	9	2,71	2	2,21
109	Crossroads	hs	f	AMR_NI	no NSA	no NSA	no NSA	4,38	3,29	3,42
120	Crossroads	hs	f	AMR_NIII	no	Sharp	9	2,88	2,42	2,25
138	Crossroads	hf	f	AMR_NIII	no NSA	no NSA	no NSA	1,92	1,58	1,29
151	Crossroads	hf	f	AMR_NI	yes	Smooth	9	2,96	1,54	1,71
166	Crossroads	hs	m	AMR_NI	no	Smooth	9	4,13	2,83	3
174	Crossroads	hs	m	AMR_NIII	no	Sharp	9	2,75	2,08	2
205	Crossroads	hf	m	AMR_NI	yes	Smooth	9	3	1,67	1,71
207	Crossroads	hf	m	AMR_NIII	yes	Smooth	9	2,67	1,29	1,5
231	Road	hs	f	AMR_NIII	no	Sharp	18	2,21	2,25	1,92
232	Road	hs	f	AMR_NI	yes	Smooth	9	4	2,29	2,88
291	Road	hs	m	AMR_NIII	yes	Smooth	18	2,38	2,46	2,08
295	Road	hs	m	AMR_NI	yes	Sharp	18	2,54	2,92	2,38
328	Office	hs	f	G722_NI	no	Smooth	9	4,53	3,88	4,08
339	Office	hs	f	G722_NIII	no	Sharp	18	3,25	3,83	2,96
361	Office	hf	f	G722_NI	no	Sharp	9	4,08	2,67	3,21
369	Office	hf	f	G722_NIII	yes	Smooth	9	3,46	2,33	2,46
382	Office	hs	m	G722_NI	no	Smooth	9	4,75	3,79	4,13
393	Office	hs	m	G722_NIII	no	Sharp	18	2,86	3,54	3
414	Office	hf	m	G722_NIII	no	Smooth	18	2,75	2,54	2,25
418	Office	hf	m	G722_NI	no	Sharp	18	3,54	2,67	2,88
445	Pub	hs	f	G722_NI	no	Sharp	18	3	2,25	2,25
456	Pub	hs	f	G722_NIII	yes	Sharp	9	2,71	1,9	2,25
466	Pub	hs	m	G722_NI	no	Smooth	18	3,25	2,21	2,71
483	Pub	hs	m	G722_NIII	yes	Sharp	9	2,75	1,58	1,96

As shown in table B.1, the data set combines the various conditions and is somehow representative for the full database i.2.

Only French samples were chosen since these are the only ones which were judged with a listening level of approximately 79 dB SPL.

NOTE:

- The sample length is less than 3,6 seconds for all samples listed above. Both algorithms, TOSQA2001 and PESQ, require a sample length of 8 seconds to 32 seconds.
- None of the methods was originally designed to work on files recorded in presence of background noise.

Analysis Description

Each condition consists of six different sentences (French language). In the listening test, the resulting MOS values are the mean over these sentences. Both PESQ [i.18] and TOSQA2001 [i.19] were therefore tested with all sentences; the mean of these measurements is finally compared to the auditory S-MOS values.

Since both algorithms are known to be very sensitive to background noise, a modified version of each sample was analysed in addition. The sequences were cut in order to minimize the noisy parts. The original test samples have a length of exactly 4 seconds; the speech part is active between 0,750 seconds and 3,250 seconds for all conditions. Thus only 2,5 seconds of speech with background noise were analysed by PESQ and TOSQA2001 in this test case.

PESQ and TOSQA2001 usually use a clean speech signal as the reference in order to estimate the degradation of a processed speech sample. For the present database both, a clean speech as well as unprocessed signal with (unprocessed) background noise are available as reference signals. Due to the fact, that the algorithms were not tested with noisy speech signals yet, both types of references, clean speech and the unprocessed signal, were analysed.

Altogether, the four test cases are summarized in table B.2.

Table B.2: Test cases

Number	Cut / Full sample	Reference
1	Full	Unprocessed
2	Full	Clean Speech
3	Cut	Unprocessed
4	Cut	Clean Speech

After all, 4 different test cases were analysed for the 32 conditions with 6 sentences each. This results into 32 x 6 x 4 = 768 single values for PESQ and also for TOSQA2001, which can be considered as a reliable base to draw conclusions. The PESQ and TOSQA2001 settings listed in table B.3 were used for testing.

Table B.3: Settings of PESQ/TOSQA2001

PESQ	Sampling rate 16 kHz Wideband extension (P862.2)
TOSQA2001	Electrical measurement, Compare to Headphone (Wideband) No fixed delay (all samples were exactly realigned in a prior step) Variable delay up to 62 ms (due to packet loss and jitter)

In order to provide a better overview of the results, the analysis was split into the two different network conditions NI and NIII. The results are listed separately for both algorithms and network conditions in table B.4 to B.7.

As expected, the results clearly indicate, that neither PESQ nor TOSQA2001 is able to estimate S-MOS values reliable. As expected, almost all calculated MOS values are lower than the corresponding auditory S-, N- and G-MOS values.

There is no linear relationship between the S- or G-MOS values and the PESQ/TOSQA2001 results, as the Pearson correlation coefficient shows. The correlation of the S-MOS data is always below 0,8, the G-MOS data correlate up to 0,89 with the calculated data (TOSQA2001 measurements for Network I + III, cut sample, clean speech as reference). The assumption of a relationship between G-MOS and calculated data cannot be verified when analyzing the scatter plot of this condition. It is obvious that too many TOSQA2001 MOS values are mapped to 1,0, a value close to a virtual, but meaningless regression line.

The results of both algorithms show MOS values less than 1,5, often close or equal to 1,0 for a lot of conditions. It can be assumed, that the algorithms completely fail and return a kind of a mapped minimum value for these samples.

The stochastic character of these measurements also arises, when comparing the auditory N-MOS values to these calculated by PESQ/TOSQA2001. The correlation between N-MOS and TOSQA2001 / PESQ MOS is often higher than between TOSQA2001 / PESQ MOS and S- or G-MOS, which should originally be approximated with these algorithms.

In order to show that there is also no non-linear relationship between the PESQ/TOSQA2001 scores and auditory S-MOS values, the scatter plots for all test cases are shown below in figures B.1 to B.4 (Network NI and NIII conditions).

On the other hand, the calculated MOS value seemed to be close to the subjective results for a lot conditions. For these the standard deviation (STD) of the calculated MOS averaged over the six sentences is high. This could not be expected because the same voice, background noise and processing were used for the recording.

These itemized points and the scatter plots given below show that the MOS values calculated by PESQ and TOSQA2001 measurements do not correlate at all with the results of the listening test.

Table B.4: TOSQA2001 results for NI conditions (clean network)

TOSQA2001, Network NI											
	MOS	Var.	MOS	Var.	MOS	Var.	MOS	Var.	Auditory MOS		
Reference	Unprocessed		Clean Speech		unprocessed		Clean Speech		S-MOS	N-MOS	G-MOS
Full/Cut	full		full		cut		cut				
Condition											
7	1,26	0,20	2,52	0,30	1,87	0,43	2,35	0,34	4,21	3,71	3,63
28	2,17	0,28	1,42	0,17	3,23	0,23	1,50	0,20	3,79	2,25	2,54
55	1,79	0,57	2,16	0,52	3,27	0,42	2,19	0,55	4,33	3,04	3,21
82	1,88	0,46	1,22	0,19	2,58	0,23	1,32	0,22	4,00	2,21	2,54
109	1,69	0,32	2,18	0,34	3,19	0,67	2,18	0,35	4,38	3,29	3,42
151	1,52	0,37	1,02	0,04	1,80	0,29	1,02	0,03	2,96	1,54	1,71
166	1,86	0,50	1,35	0,33	2,19	0,28	1,25	0,27	4,13	2,83	3,00
205	1,45	0,29	1,00	0,00	1,49	0,33	1,00	0,00	3,00	1,67	1,71
232	1,60	0,24	1,09	0,11	2,13	0,28	1,08	0,10	4,00	2,29	2,88
295	1,26	0,46	1,31	0,24	1,41	0,64	1,28	0,29	2,54	2,92	2,38
328	4,15	0,12	3,73	0,27	4,15	0,11	3,71	0,29	4,53	3,88	4,08
361	3,06	0,34	2,20	0,28	3,57	0,23	2,21	0,27	4,08	2,67	3,21
382	3,64	0,53	3,32	0,29	3,71	0,39	3,32	0,27	4,75	3,79	4,13
418	2,03	0,29	1,93	0,34	2,27	0,31	1,89	0,32	3,54	2,67	2,88
445	2,19	0,28	1,38	0,23	2,51	0,18	1,32	0,26	3,00	2,25	2,25
466	2,66	0,10	1,17	0,15	2,57	0,33	1,17	0,16	3,25	2,21	2,71
Correlation:											
S-MOS	0,48		0,72		0,73		0,73				
N-MOS	0,44		0,88		0,52		0,87				
G-MOS	0,60		0,89		0,70		0,89				

Table B.5: TOSQA2001 results for NIII conditions (3 % packet loss, 20 ms jitter)

TOSQA2001, Network NIII											
	MOS	Var.	MOS	Var.	MOS	Var.	MOS	Var.	Auditory MOS		
Reference	Unprocessed		Clean Speech		Unprocessed		Clean Speech		S-MOS	N-MOS	G-MOS
Full/Cut	full		full		cut		cut				
Condition											
3	1,46	0,34	2,24	0,44	2,13	0,83	2,18	0,33	3,63	3,13	3,08
54	1,11	0,18	1,17	0,20	1,22	0,18	1,17	0,19	2,00	1,92	1,63
57	1,33	0,15	1,90	0,30	2,03	0,25	1,89	0,32	3,46	3,00	2,79
87	1,44	0,28	1,32	0,26	1,43	0,22	1,33	0,26	2,71	2,00	2,21
120	1,00	0,00	1,22	0,26	1,08	0,09	1,27	0,29	2,88	2,42	2,25
138	1,62	0,16	1,19	0,25	1,87	0,20	1,17	0,21	1,92	1,58	1,29
174	1,01	0,02	1,31	0,38	1,29	0,44	1,19	0,31	2,75	2,08	2,00
207	1,00	0,00	1,00	0,00	1,06	0,08	1,00	0,00	2,67	1,29	1,50
231	1,00	0,00	1,02	0,06	1,04	0,09	1,02	0,04	2,21	2,25	1,92
291	1,00	0,00	1,00	0,00	1,04	0,09	1,00	0,00	2,38	2,46	2,08
339	2,69	0,63	2,60	0,56	2,67	0,62	2,66	0,61	3,25	3,83	2,96
369	1,71	0,34	1,85	0,34	1,63	0,43	1,85	0,33	3,46	2,33	2,46
393	2,09	0,46	1,97	0,53	2,01	0,51	1,94	0,56	2,86	3,54	3,00
414	1,00	0,00	1,11	0,14	1,05	0,11	1,09	0,15	2,75	2,54	2,25
456	1,59	0,28	1,19	0,11	2,03	0,54	1,23	0,12	2,71	1,90	2,25
483	1,60	0,24	1,27	0,27	1,61	0,42	1,14	0,19	2,75	1,58	1,96
Correlation:											
S-MOS	0,37		0,75		0,51		0,74				
N-MOS	0,56		0,81		0,57		0,83				
G-MOS	0,53		0,75		0,62		0,83				

Table B.6: PESQ results for NI conditions (clean network)

PESQ, Network NI											
	MOS	Var.	MOS	Var.	MOS	Var.	MOS	Var.	Auditory MOS		
Reference	Unprocessed		Clean Speech		Unprocessed		Clean Speech		S-MOS	N-MOS	G-MOS
Full/Cut	full		full		cut		cut				
Condition											
7	1,91	0,05	1,65	0,24	2,30	0,11	1,05	0,01	4,21	3,71	3,63
28	1,14	0,03	1,03	0,00	1,25	0,06	1,02	0,00	3,79	2,25	2,54
55	1,40	0,16	1,31	0,12	1,86	0,50	1,12	0,05	4,33	3,04	3,21
82	1,12	0,05	1,06	0,02	1,22	0,10	1,02	0,01	4,00	2,21	2,54
109	1,81	0,13	1,30	0,08	2,61	0,37	1,08	0,02	4,38	3,29	3,42
151	1,23	0,12	1,04	0,02	1,32	0,16	1,02	0,00	2,96	1,54	1,71
166	2,19	0,27	1,41	0,23	2,60	0,44	1,10	0,07	4,13	2,83	3,00
205	1,27	0,09	1,12	0,06	1,28	0,06	1,03	0,01	3,00	1,67	1,71
232	2,69	0,37	1,15	0,07	2,86	0,46	1,06	0,02	4,00	2,29	2,88
295	1,23	0,12	1,25	0,19	1,47	0,22	1,09	0,09	2,54	2,92	2,38
328	3,32	0,20	2,64	0,20	3,80	0,18	2,53	0,12	4,53	3,88	4,08
361	2,85	0,31	1,38	0,13	3,41	0,26	1,21	0,05	4,08	2,67	3,21
382	3,11	0,24	2,15	0,27	3,39	0,25	2,46	0,24	4,75	3,79	4,13
418	2,16	0,19	1,37	0,11	2,38	0,27	1,41	0,09	3,54	2,67	2,88
445	1,99	0,11	1,22	0,10	2,27	0,16	1,41	0,09	3,00	2,25	2,25
466	2,00	0,30	1,15	0,04	2,43	0,16	1,18	0,07	3,25	2,21	2,71
Correlation:											
S-MOS	0,56		0,59		0,61		0,45				
N-MOS	0,57		0,81		0,65		0,62				
G-MOS	0,73		0,80		0,79		0,65				

Table B.7: PESQ results for NIII conditions (3 % packet loss, 20 ms jitter)

PESQ, Network NIII											
	MOS	Var.	MOS	Var.	MOS	Var.	MOS	Var.	Auditory MOS		
Reference	Unprocessed		Clean Speech		Unprocessed		Clean Speech		S-MOS	N-MOS	G-MOS
Full/Cut	full		full		cut		cut				
Condition											
3	1,27	0,12	1,15	0,04	1,44	0,25	1,05	0,01	3,63	3,13	3,08
54	1,06	0,02	1,07	0,03	1,08	0,02	1,02	0,00	2,00	1,92	1,63
57	1,19	0,05	1,17	0,03	1,34	0,09	1,11	0,04	3,46	3,00	2,79
87	1,08	0,02	1,08	0,03	1,15	0,07	1,03	0,01	2,71	2,00	2,21
120	1,58	0,15	1,26	0,10	1,57	0,23	1,07	0,02	2,88	2,42	2,25
138	1,11	0,03	1,03	0,01	1,14	0,04	1,02	0,00	1,92	1,58	1,29
174	1,35	0,13	1,26	0,15	1,58	0,36	1,11	0,07	2,75	2,08	2,00
207	1,15	0,06	1,09	0,03	1,22	0,09	1,03	0,01	2,67	1,29	1,50
231	1,31	0,09	1,15	0,05	1,34	0,12	1,06	0,02	2,21	2,25	1,92
291	1,39	0,24	1,19	0,09	1,50	0,34	1,09	0,09	2,38	2,46	2,08
339	1,24	0,07	1,24	0,09	1,26	0,08	2,40	0,21	3,25	3,83	2,96
369	1,48	0,13	1,17	0,09	1,73	0,26	1,22	0,06	3,46	2,33	2,46
393	1,58	0,11	1,37	0,19	1,64	0,12	2,49	0,25	2,86	3,54	3,00
414	1,51	0,20	1,18	0,10	1,72	0,37	1,40	0,09	2,75	2,54	2,25
456	1,50	0,16	1,10	0,02	1,56	0,19	1,14	0,05	2,71	1,90	2,25
483	1,52	0,12	1,13	0,02	1,55	0,27	1,18	0,07	2,75	1,58	1,96
Correlation:											
S-MOS	0,26		0,41		0,42		0,27				
N-MOS	0,22		0,70		0,24		0,74				
G-MOS	0,34		0,63		0,40		0,59				

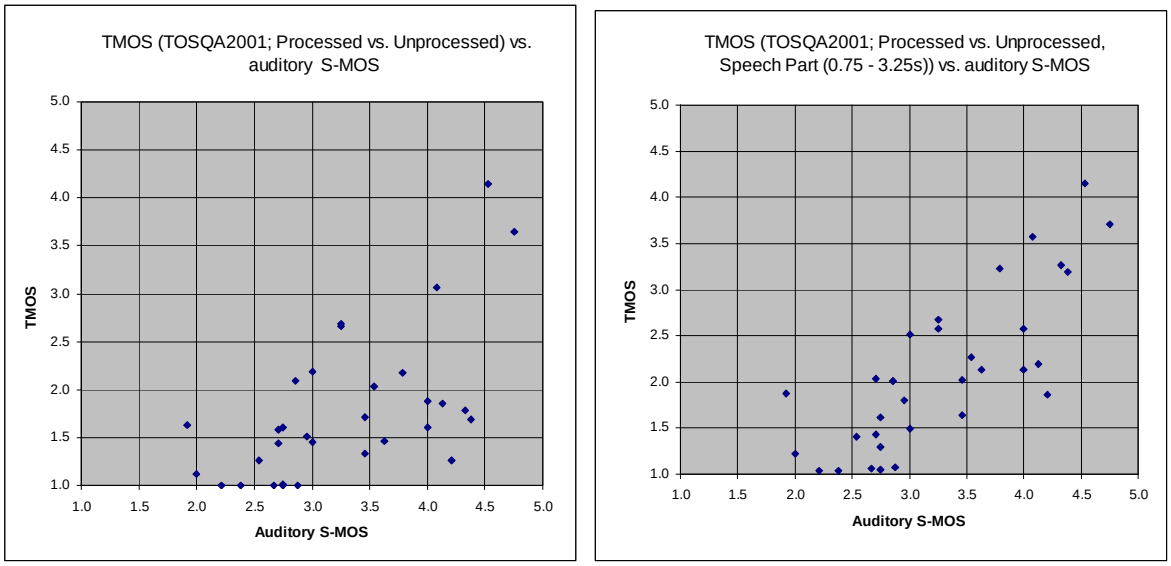


Figure B.1: TOSQA2001 results (TMOS) of processed data versus auditory S-MOS (unprocessed signal used as TOSQA2001 reference)

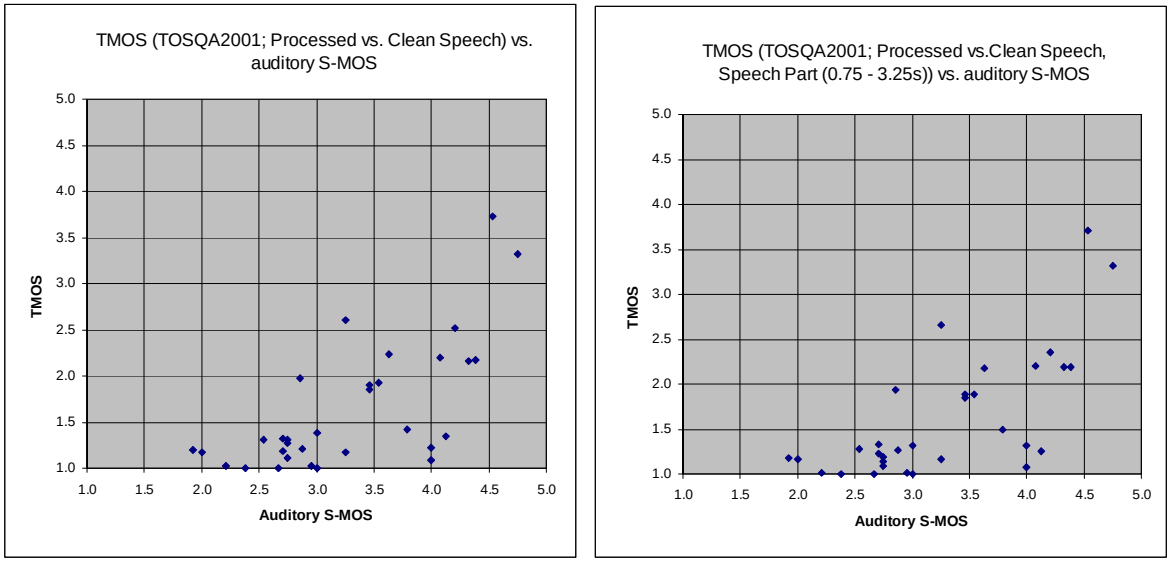


Figure B.2: TOSQA2001 results (TMOS) of processed data versus auditory S-MOS (clean speech signal used as TOSQA2001 reference)

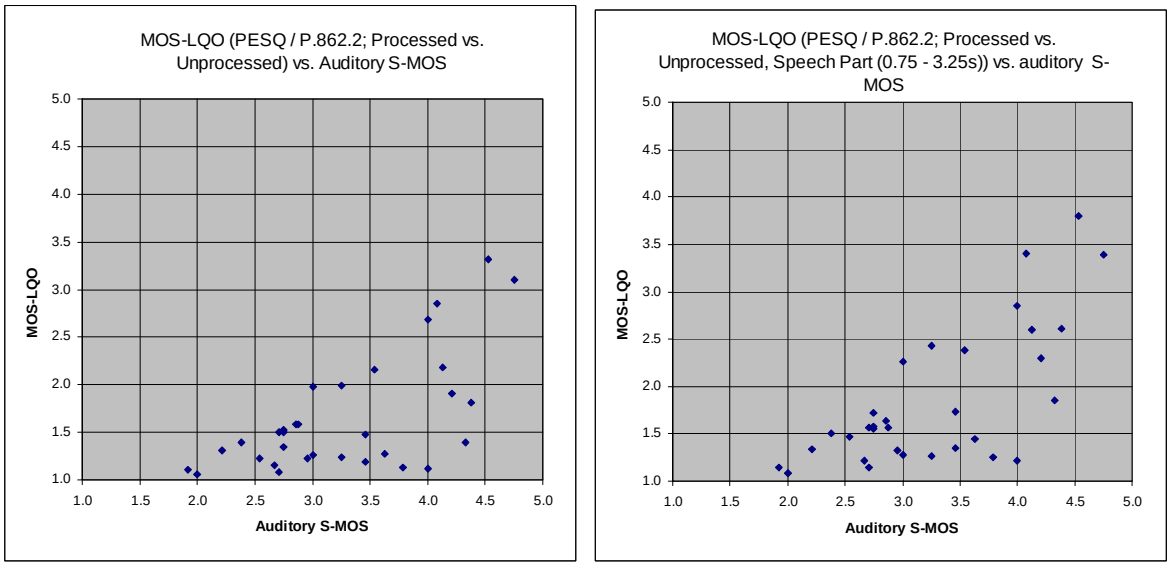


Figure B.3: PESQ (MOS-LQO, P.862.2) results of processed data versus auditory S-MOS (unprocessed signal used as PESQ reference)

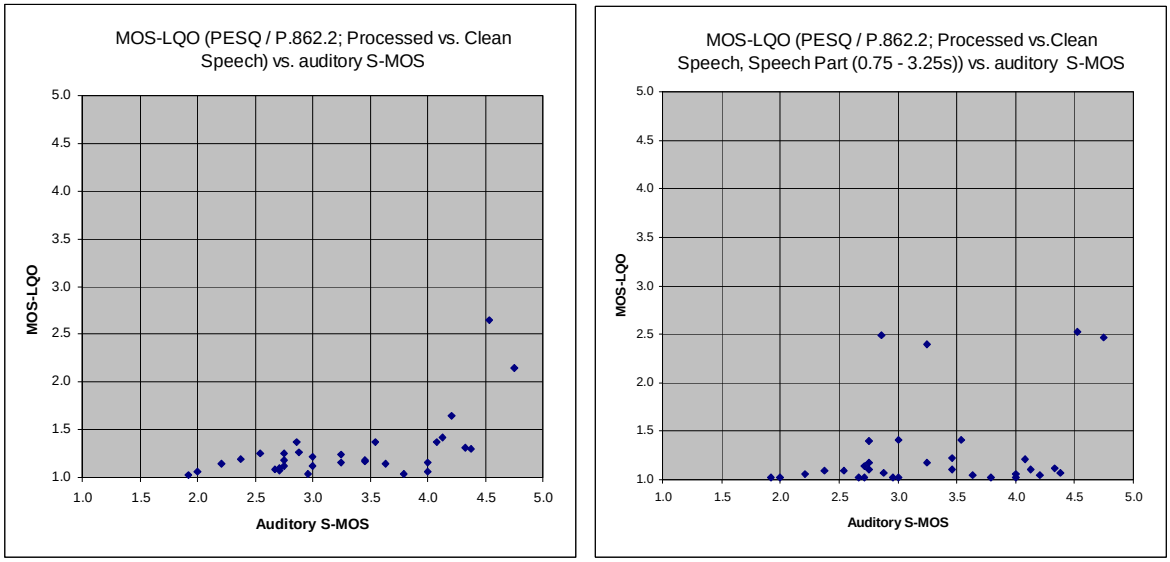


Figure B.4: PESQ results (MOS-LQO, P.862.2) of processed data versus auditory S-MOS (clean speech signal used as PESQ reference)

Annex C:
Comparison of objective MOS versus auditory MOS for the All Data of Training Period

This annex shows the correlation plots between the objective and the auditory S-/N-/G-MOS for all French and Czech data used during the training of the new method. Note that the MOS scores for ***all*** conditions were compared to the listening test results. For the Czech data again all selected conditions ***including*** the NI conditions were used for the training.

Figures C.1, C.3 and C.5 show the results for the French data and figures C.2, C.4 and C.6 or the Czech data. In order to distinguish between the selected data and the ones which were not used for the model development, the conditions not used (rej.) are indicated by a "+" and the selected (acc.) by a "○".

For the French data the correlation for the objective N-MOS decreases only slightly from 94,8 % to 93,9 %. This can be expected because the unused French samples were mainly influenced by the speech and not by the background noise. The correlation of the objective N-MOS to the auditory N-MOS for the Czech data decreases more (from 98 % to 92,2 %). This can also be expected because some of the unused samples had very low background noise level compared to others.

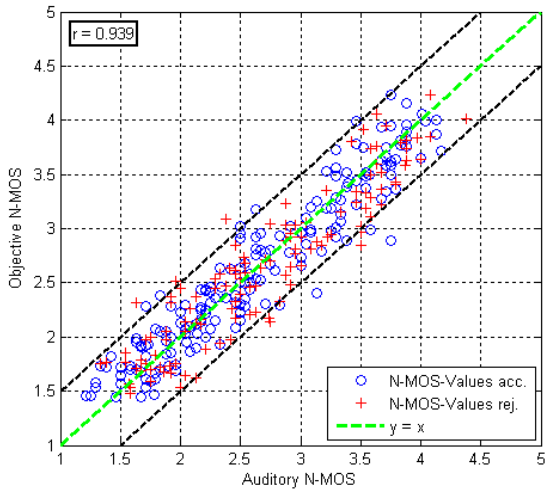


Figure C.1: Objective versus auditory N-MOS for all French data used in listening test

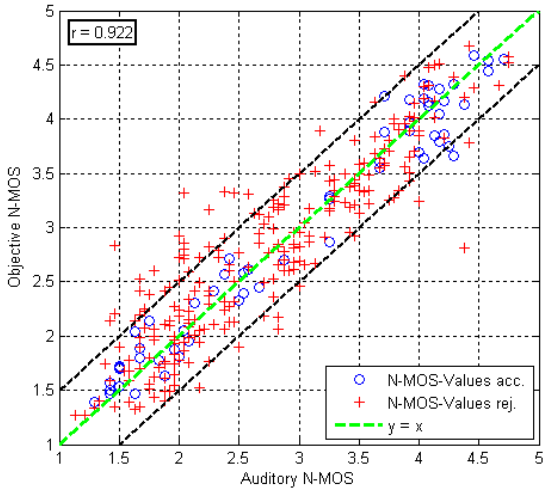


Figure C.2: Objective versus auditory N-MOS for all Czech data used in listening test

The correlation of the objective to the auditory S-MOS decreases from 92,9 % to 88,6 % for the French data and from 96,4 % to 82,9 % for the Czech data. Within the French data a per sample deviation of 0,5 MOS or higher between objective and auditory S-MOS can be observed for some selected as well as for some unused conditions (see figure C.3). As shown in figure C.4 the conditions with the lowest correlation between objective and auditory S-MOS are calculated for the unused conditions of the Czech sample. One of the main issues is probably again the high variation of overall levels within the Czech data. Nevertheless the deviation between auditory and objective S-MOS is less 0,5 MOS for most of the conditions not used for the model development.

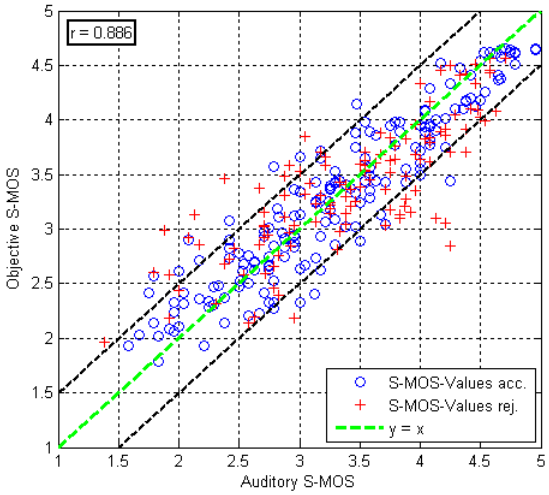


Figure C.3: Objective versus auditory S-MOS for all French data used in listening test

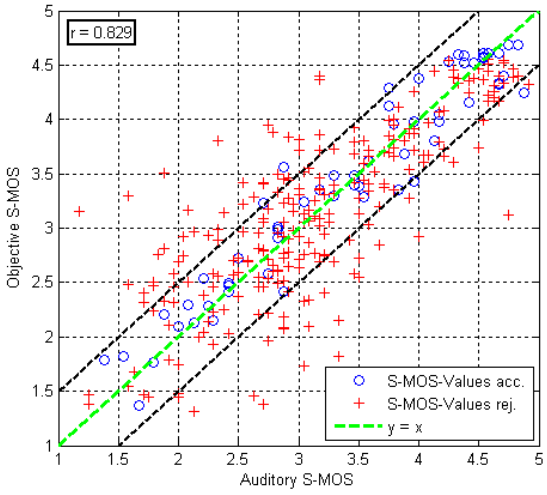


Figure C.4: Objective versus auditory S-MOS for all Czech data used in listening test

The correlation between the objective and auditory G-MOS decreases also only slightly from 95,4 % to 94 % for the French data. The per sample deviation is higher as 0,5 MOS for only a very few conditions. Again for the Czech data the correlation decreases more from 97,6 % to 90,1 %. As shown in figure C.6 the highest per sample deviations between objective and auditory G-MOS occur for the conditions not used for the model development.

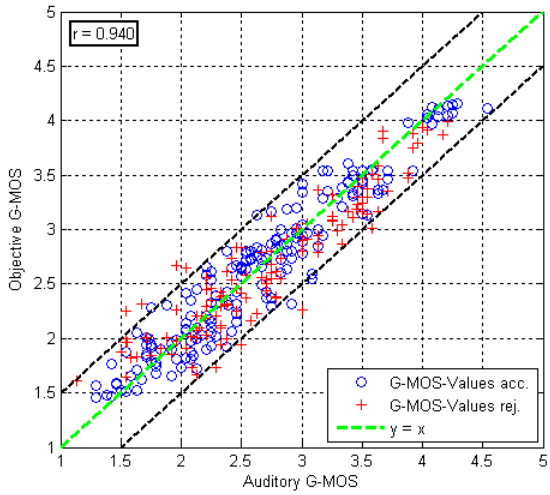


Figure C.5: Objective versus auditory G-MOS for all French data used in listening test

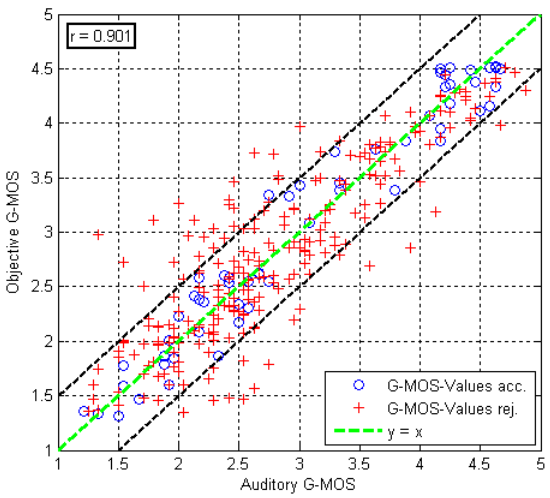


Figure C.6: Objective versus auditory G-MOS for all Czech data used in listening test

Generally it can be concluded that the new model is more applicable on the French data than on the Czechs if all conditions are considered. The main reasons are:

- the higher number of selected French samples leading to higher numerical stability;
- the high variety of overall level within the Czech data and thus the lower number of selected data.

Annex D:
Comparison of objective MOS versus auditory MOS for the Data not used during the Training Period

For information purpose figure D.1 to D.6 show the correlation plots for the objective and auditory S-/N-/G-MOS only for the rejected conditions of both languages (see clause 5.5) with were not used during the development of the method for the French Samples (Due to the limited number of selected Czech data, the N1 conditions were included). Again the data not used for the model development are indicated by a "+" in the scatter plots. For the Czech data all selected conditions plus all NI conditions were used for the training.

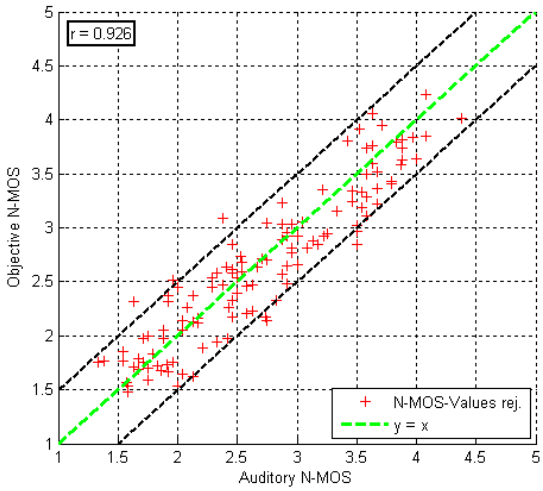


Figure D.1: Objective versus auditory N-MOS only for French data not used for the model development

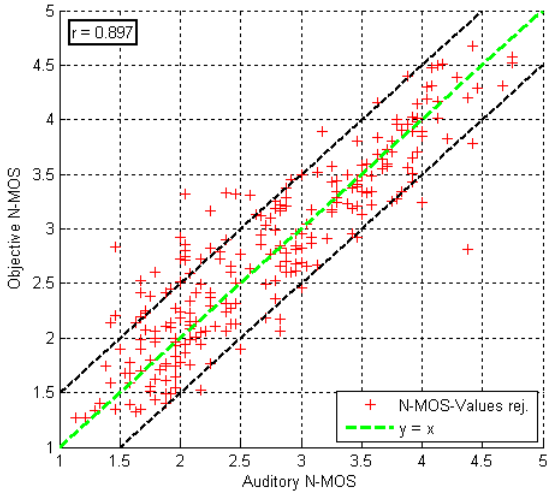


Figure D.2: Objective versus auditory N-MOS only for Czech data not used for the model development

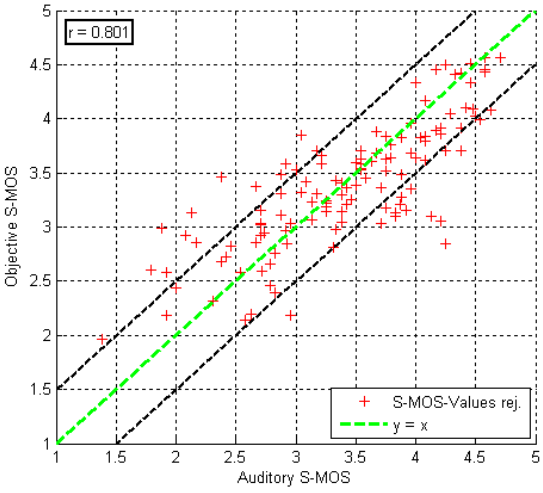


Figure D.3: Objective versus auditory S-MOS only for French data not used for the model development

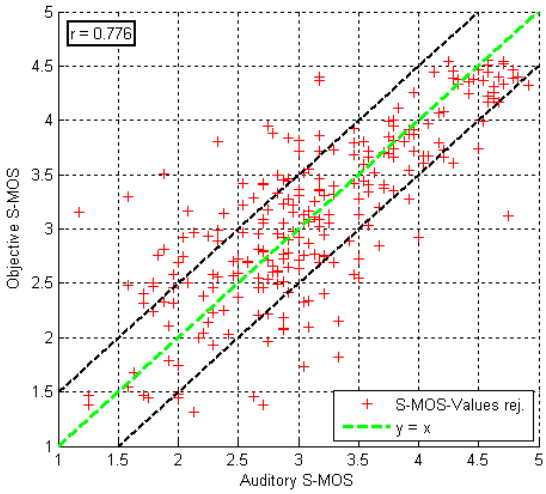


Figure D.4: Objective versus auditory S-MOS only for Czech data not used for the model development

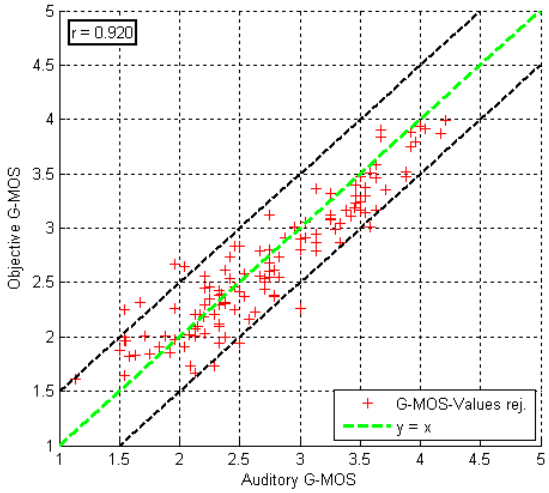


Figure D.5: Objective versus auditory G-MOS only for French data not used for the model development

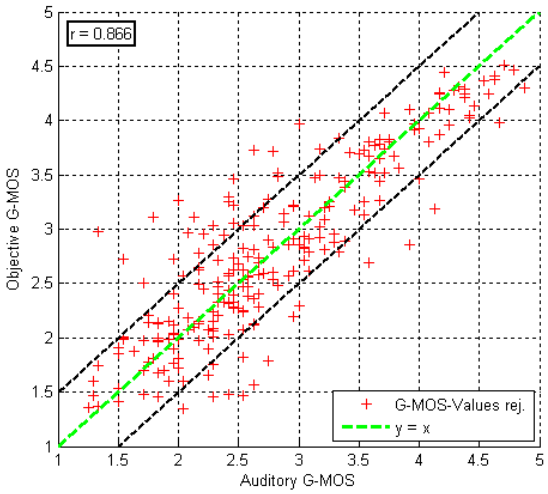


Figure D.6: Objective versus auditory G-MOS only for Czech data not used for the model development

Annex E:
Regression Coefficients for Czech data

This annex summarizes the regression coefficients for the S-, N- and G-MOS calculation of the Czech data. The coefficients for the French data can be found in tables 6.1 to 6.5.

Table E.1: Coefficients for linear, quadratic N-MOS regression algorithm (Czech)

Order	c_0	c_{BGN} (N_{BGN})	c_{j1} ($vRA_{BGN,\,u}$)	c_{j2} ($vRA_{BGN,\,p}$)	c_{j3} ($v\Delta RA_{BGN,p-u}$)	c_{j4} ($m\Delta RA_{BGN,p-u}$)	c_{j5} ($mRA_{BGN,p}$)
1	0,6733	-0,0908	3,0159	0,2811	-0,3802	-6,2485	0,2150
2	-	-	-0,5760	-0,0334	0,0578	2,0176	-0,1686

Table E.2: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS \leq N\text{-}MOS_{low} = 2{,}25$ (Czech)

Order	${}_1c_{j0}$	${}_1c_{j1}$ (ΔSNR)	${}_1c_{j2}$ ($mRA_{SP,p}$)	${}_1c_{j3}$ ($m\Delta RA_{SP,p-c}$)	${}_1c_{j4}$ ($m\Delta RA_{SP,p-u}$)	${}_1c_{j5}$ ($v\Delta RA_{SP,p-c}$)	${}_1c_{j6}$ ($v\Delta RA_{SP,p-u}$)
1	9,7860	-0,0211	0,0835	1,7008	-1,1706	-0,0289	-0,2701
2	-	-	0,0936	0,0779	-0,4926	0,0008	0,0022

Table E.3: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS_{low} < N\text{-}MOS < N\text{-}MOS_{high}$ (Czech)

Order	${}_2c_{j0}$	${}_2c_{j1}$ (ΔSNR)	${}_2c_{j2}$ ($mRA_{SP,p}$)	${}_2c_{j3}$ ($m\Delta RA_{SP,p-c}$)	${}_2c_{j4}$ ($m\Delta RA_{SP,p-u}$)	${}_2c_{j5}$ ($v\Delta RA_{SP,p-c}$)	${}_2c_{j6}$ ($v\Delta RA_{SP,p-u}$)
1	2,2623	-0,0283	1,7981	-1,1318	-1,2940	-0,1389	-0,2207
2	-	-	-0,2587	-0,1753	-0,9927	0,0022	0,0051

Table E.4: Coefficients for linear, quadratic S-MOS regression algorithm, $N\text{-}MOS \geq N\text{-}MOS_{high} = 3{,}0$ (Czech)

Order	${}_3c_{j0}$	${}_3c_{j1}$ (ΔSNR)	${}_3c_{j2}$ ($mRA_{SP,p}$)	${}_3c_{j3}$ ($m\Delta RA_{SP,p-c}$)	${}_3c_{j4}$ ($m\Delta RA_{SP,p-u}$)	${}_3c_{j5}$ ($v\Delta RA_{SP,p-c}$)	${}_3c_{j6}$ ($v\Delta RA_{SP,p-u}$)
1	4,2104	-0,0371	1,9003	-0,2506	-0,5132	-0,2349	0,0428
2	-	-	-0,2983	-0,0167	-0,3223	0,0031	-0,0043

Table E.5: Coefficients for linear, quadratic G-MOS regression algorithm (Czech)

Order	c_0	c_{Nj} (N-MOS)	c_{Sj} (S-MOS)
1	-0,9326	0,8097	0,5074
2	-	-0,0696	0,0443

Annex F:
Detailed STF 294 subjective and objective validation test results

Tables F.1 and F.2 contain the conditions and related auditory and objective S-MOS, N-MOS and G-MOS for two tested languages. Also standard deviations for all MOS scores are given.

Table F.1: Subjective and objective experiment results - French validation part
(Recording: hs - handset, hf - hands-free. Speaker: f - female, m - male)

Conditions	Id real	Noise	Recording	Speaker	Network	DAV	Smooth	dB	FRENCH								
									Subjective						Objective		
									MOS			Standard deviation			MOS		
									Speech	Noise	Global	Speech	Noise	Global	Speech	Noise	Global
1	1	Lux_Car	hs	f	AMR _NI	no NSA	no NSA	no NSA	4,42	3,67	3,96	0,65	0,64	0,36	4,31	3,15	3,53
4	4	Lux_Car	hs	f	AMR _NI	no	Smooth	9	4,88	3,63	3,92	0,45	0,71	0,65	4,40	3,11	3,59
6	6	Lux_Car	hs	f	AMR _NIII	no	Smooth	9	3,50	3,21	3,08	1,07	0,72	0,88	3,24	3,01	2,71
9	9	Lux_Car	hs	f	AMR _NIII	no	Smooth	18	3,46	3,31	3,08	1,18	0,62	0,93	3,10	3,48	2,85
10	10	Lux_Car	hs	f	AMR _NI	no	Sharp	9	4,54	3,46	3,63	0,59	0,66	0,71	4,45	3,18	3,66
22	22	Lux_Car	hs	f	AMR _NI	yes	Sharp	9	4,42	3,50	3,63	0,58	0,72	0,71	4,27	3,35	3,59
24	24	Lux_Car	hs	f	AMR _NIII	yes	Sharp	9	3,79	3,29	3,21	0,83	0,69	0,78	3,32	3,21	2,85
31	31	Lux_Car	hf	f	AMR _NI	no	Smooth	9	3,79	2,21	2,75	1,02	0,88	0,74	3,90	2,26	2,80
34	34	Lux_Car	hf	f	AMR _NI	no	Smooth	18	3,04	2,29	2,42	1,00	0,75	0,78	2,78	2,38	2,18
37	37	Lux_Car	hf	f	AMR _NI	no	Sharp	9	3,21	2,04	2,29	0,98	0,69	0,55	3,89	2,26	2,79
39	39	Lux_Car	hf	f	AMR _NIII	no	Sharp	9	2,71	1,71	1,96	1,04	0,69	0,62	2,87	2,05	2,06
42	42	Lux_Car	hf	f	AMR _NIII	no	Sharp	18	1,92	1,92	1,71	0,88	0,72	0,62	2,22	2,10	1,81
43	43	Lux_Car	hf	f	AMR _NI	yes	Smooth	9	3,96	2,00	2,54	0,91	0,83	0,78	3,82	2,28	2,75
48	48	Lux_Car	hf	f	AMR _NIII	yes	Smooth	18	2,33	1,83	1,79	0,76	0,70	0,66	2,58	2,20	2,00
49	49	Lux_Car	hf	f	AMR _NI	yes	Sharp	9	3,71	2,17	2,54	0,86	0,76	0,66	3,79	2,32	2,75
52	52	Lux_Car	hf	f	AMR _NI	yes	Sharp	18	2,67	2,04	1,96	0,92	0,62	0,69	3,20	2,61	2,51
63	63	Lux_Car	hs	m	AMR _NIII	no	Smooth	18	3,21	3,29	2,67	1,02	0,81	0,81	2,97	3,19	2,65
67	67	Lux_Car	hs	m	AMR _NI	no	Sharp	18	3,71	3,60	3,44	1,08	0,71	0,65	4,25	3,41	3,60
78	78	Lux_Car	hs	m	AMR _NIII	yes	Sharp	9	2,96	2,79	2,50	1,00	0,66	0,78	3,53	2,61	2,71
79	79	Lux_Car	hs	m	AMR _NI	yes	Sharp	18	3,96	3,58	3,46	0,75	0,78	0,59	3,38	3,78	3,13
81	81	Lux_Car	hs	m	AMR _NIII	yes	Sharp	18	3,04	3,29	2,63	1,00	0,81	0,88	3,11	3,51	2,86
90	90	Lux_Car	hf	m	AMR _NIII	no	Smooth	18	2,38	2,13	1,88	1,17	1,03	0,90	2,37	2,05	1,84
94	94	Lux_Car	hf	m	AMR _NI	no	Sharp	18	2,88	2,42	2,38	0,90	0,88	0,88	3,38	2,50	2,56
99	99	Lux_Car	hf	m	AMR _NIII	yes	Smooth	9	3,04	2,46	2,25	0,86	0,72	0,74	2,57	2,13	1,96
102	102	Lux_Car	hf	m	AMR _NIII	yes	Smooth	18	2,54	2,21	2,08	1,18	0,93	0,97	2,83	2,01	2,02
103	103	Lux_Car	hf	m	AMR _NI	yes	Sharp	9	3,63	2,46	2,58	0,71	0,88	0,78	3,78	2,36	2,75
114	114	Crossroads	hs	f	AMR _NIII	no	Smooth	9	3,04	2,53	2,31	1,04	0,88	0,80	2,83	2,64	2,33
132	132	Crossroads	hs	f	AMR _NIII	yes	Sharp	9	3,46	2,96	3,00	0,83	0,75	0,72	3,26	3,07	2,76

Conditions	Id real	Noise	Recording	Speaker	Network	DAV	Smooth	dB	FRENCH								
									Subjective						Objective		
									MOS			Standard deviation			MOS		
									Speech	Noise	Global	Speech	Noise	Global	Speech	Noise	Global
136	136	Crossroads	hf	f	AMR _NI	no NSA	no NSA	no NSA	2,66	1,67	1,92	1,27	0,96	0,93	2,92	1,80	1,96
141	141	Crossroads	hf	f	AMR _NIII	no	Smooth	9	2,79	1,29	1,54	1,28	0,46	0,72	2,47	1,52	1,61
150	150	Crossroads	hf	f	AMR _NIII	no	Sharp	18	1,92	1,67	1,42	0,78	0,82	0,58	2,12	1,80	1,63
163	163	Crossroads	hs	m	AMR _NI	no NSA	no NSA	no NSA	3,92	2,46	2,67	1,14	0,93	0,87	3,18	2,21	2,30
165	165	Crossroads	hs	m	AMR _NIII	no NSA	no NSA	no NSA	2,83	2,38	2,17	1,27	0,98	1,01	2,42	2,08	1,88
171	171	Crossroads	hs	m	AMR _NIII	no	Smooth	18	2,42	3,00	2,21	1,02	1,18	0,93	2,86	2,97	2,49
172	172	Crossroads	hs	m	AMR _NI	no	Sharp	9	3,88	2,42	2,83	0,85	0,93	0,92	3,93	2,27	2,83
177	177	Crossroads	hs	m	AMR _NIII	no	Sharp	18	2,48	3,00	2,46	0,88	1,25	0,88	3,15	2,84	2,58
181	181	Crossroads	hs	m	AMR _NI	yes	Smooth	18	3,04	3,08	2,83	1,08	1,14	0,92	3,63	3,09	3,00
184	184	Crossroads	hs	m	AMR _NI	yes	Sharp	9	3,96	2,57	3,29	1,08	1,06	0,75	4,00	2,51	2,99
186	186	Crossroads	hs	m	AMR _NIII	yes	Sharp	9	3,75	2,58	2,67	1,22	1,10	1,13	3,38	2,32	2,48
187	187	Crossroads	hs	m	AMR _NI	yes	Sharp	18	2,96	3,50	2,88	1,27	1,10	1,15	3,38	3,41	2,98
192	192	Crossroads	hf	m	AMR _NIII	no NSA	no NSA	no NSA	1,77	1,42	1,46	0,98	0,72	0,59	2,31	1,60	1,59
201	201	Crossroads	hf	m	AMR _NIII	no	Sharp	9	2,29	1,38	1,71	1,23	0,58	0,69	2,28	1,49	1,52
204	204	Crossroads	hf	m	AMR _NIII	no	Sharp	18	1,63	1,88	1,38	0,97	1,15	0,65	2,08	1,75	1,59
213	213	Crossroads	hf	m	AMR _NIII	yes	Sharp	9	2,13	1,42	1,46	0,95	0,58	0,59	2,21	1,54	1,52
225	225	Road_Noise	hs	f	AMR _NIII	no	Smooth	18	2,29	2,46	2,17	0,62	0,78	0,64	2,49	2,54	2,13
226	226	Road_Noise	hs	f	AMR _NI	no	Sharp	9	3,67	2,54	2,88	0,96	0,72	0,80	3,41	2,08	2,38
229	229	Road_Noise	hs	f	AMR _NI	no	Sharp	18	3,17	2,50	2,58	1,09	0,88	0,72	3,24	2,61	2,53
238	238	Road_Noise	hs	f	AMR _NI	yes	Sharp	9	3,71	2,29	2,70	0,95	0,62	0,75	3,63	2,47	2,71
241	241	Road_Noise	hs	f	AMR _NI	yes	Sharp	18	2,96	3,13	2,46	1,30	0,85	0,83	3,14	3,36	2,82
244	271	Road_Noise	hs	m	AMR _NI	no NSA	no NSA	no NSA	3,75	2,00	2,49	0,85	0,93	0,84	3,42	1,64	2,16
246	273	Road_Noise	hs	m	AMR _NIII	no NSA	no NSA	no NSA	2,46	1,54	1,67	1,28	0,88	0,87	2,54	1,46	1,60
247	274	Road_Noise	hs	m	AMR _NI	no	Smooth	9	3,33	1,54	2,29	1,09	0,66	0,75	3,47	1,56	2,15
249	276	Road_Noise	hs	m	AMR _NIII	no	Smooth	9	1,38	1,21	1,00	0,97	0,83	0,00	2,50	1,35	1,53
256	283	Road_Noise	hs	m	AMR _NI	no	Sharp	18	2,75	2,17	2,08	0,99	0,82	0,88	2,89	1,96	2,03
276	330	Office_Noise	hs	f	G722_NIII	no	Smooth	9	3,29	3,71	3,09	0,95	0,69	0,72	3,37	3,66	3,08
277	331	Office_Noise	hs	f	G722_NI	no	Smooth	18	4,75	3,83	4,04	0,53	0,82	0,81	4,57	3,61	3,95
289	343	Office_Noise	hs	f	G722_NI	yes	Smooth	18	4,54	3,63	3,67	0,59	0,65	0,76	4,60	3,69	4,02
292	346	Office_Noise	hs	f	G722_NI	yes	Sharp	9	4,83	4,13	4,33	0,48	0,45	0,64	4,61	3,99	4,14
294	348	Office_Noise	hs	f	G722_NIII	yes	Sharp	9	3,54	3,38	3,17	1,14	0,71	0,76	3,61	3,60	3,21
297	351	Office_Noise	hs	f	G722_NIII	yes	Sharp	18	2,67	3,88	2,79	0,96	0,61	0,72	3,02	4,01	3,02
298	352	Office_Noise	hf	f	G722_NI	no NSA	no NSA	no NSA	4,21	2,96	3,21	0,72	0,69	0,78	4,38	2,73	3,40
301	355	Office_Noise	hf	f	G722_NI	no	Smooth	9	4,21	3,00	3,08	0,93	0,78	0,65	4,25	2,87	3,35
309	363	Office_Noise	hf	f	G722_NIII	no	Sharp	9	2,54	2,50	2,33	0,78	0,51	0,48	2,76	2,63	2,29

Conditions	Id real	Noise	Recording	Speaker	Network	DAV	Smooth	dB	FRENCH								
									Subjective						Objective		
									MOS			Standard deviation			MOS		
									Speech	Noise	Global	Speech	Noise	Global	Speech	Noise	Global
310	364	Office_Noise	hf	f	G722_NI	no	Sharp	18	3,38	2,83	2,75	1,17	0,87	0,79	3,84	2,96	3,08
316	370	Office_Noise	hf	f	G722_NI	yes	Smooth	18	3,75	2,75	2,79	1,03	0,79	0,98	3,78	2,87	3,00
327	381	Office_Noise	hs	m	G722_NIII	no NSA	no NSA	no NSA	3,50	3,71	3,25	1,10	0,62	0,94	3,17	3,25	2,79
331	385	Office_Noise	hs	m	G722_NI	no	Smooth	18	4,83	4,08	4,08	0,48	0,65	0,65	4,60	3,78	4,05
334	388	Office_Noise	hs	m	G722_NI	no	Sharp	9	4,71	3,88	4,08	0,69	0,54	0,88	4,62	3,62	4,00
366	420	Office_Noise	hf	m	G722_NIII	no	Sharp	18	2,79	2,33	2,17	1,25	0,76	0,76	2,89	2,80	2,43
372	426	Office_Noise	hf	m	G722_NIII	yes	Smooth	18	3,00	2,52	2,30	1,14	0,77	0,80	2,84	2,96	2,48
373	427	Office_Noise	hf	m	G722_NI	yes	Sharp	9	4,67	3,17	3,38	0,48	0,56	0,65	4,43	3,01	3,57
378	432	Office_Noise	hf	m	G722_NIII	yes	Sharp	18	2,88	2,67	2,38	0,95	0,96	0,82	2,84	2,80	2,40
379	433	Pub_Noise	hs	f	G722_NI	no NSA	no NSA	no NSA	4,04	2,08	2,58	0,91	0,88	0,65	3,74	1,75	2,42
384	438	Pub_Noise	hs	f	G722_NIII	no	Smooth	9	3,00	1,63	1,92	1,02	0,65	0,58	2,57	1,44	1,61
390	444	Pub_Noise	hs	f	G722_NIII	no	Sharp	9	2,42	1,79	1,79	1,02	1,06	0,66	2,79	1,59	1,79
396	450	Pub_Noise	hs	f	G722_NIII	yes	Smooth	9	2,38	1,79	1,67	0,88	0,66	0,87	2,68	1,67	1,78
415	469	Pub_Noise	hs	m	G722_NI	no	Sharp	9	3,96	1,67	2,08	1,08	0,96	0,88	4,07	1,90	2,74
420	474	Pub_Noise	hs	m	G722_NIII	no	Sharp	18	1,88	1,21	1,38	0,90	0,41	0,49	2,00	1,67	1,52
423	477	Pub_Noise	hs	m	G722_NIII	yes	Smooth	9	2,46	1,75	1,89	0,98	0,68	0,75	3,08	1,84	2,06
427	481	Pub_Noise	hs	m	G722_NI	yes	Sharp	9	3,79	1,99	2,25	1,06	0,92	0,79	3,89	2,22	2,76
432	486	Pub_Noise	hs	m	G722_NIII	yes	Sharp	18	1,96	1,88	1,54	0,95	0,68	0,66	2,34	2,41	2,00

Table F.2: Subjective and objective experiment results - Czech validation part (Recording: hs – handset, hf – hands-free. Speaker: f – female, m – male)

Conditions	Noise	Recording	Speaker	Network	DAV	Smooth	dB	CZECH								
								Subjective						Objective		
								MOS			Standard deviation			MOS		
								Speech	Noise	Global	Speech	Noise	Global	Speech	Noise	Global
1	Lux_Car	hs	f	AMR _NI	no NSA	no NSA	no NSA	3,33	2,67	2,92	0,48	0,76	0,50	3,48	2,71	3,05
10	Lux_Car	hs	f	AMR _NI	no	Sharp	9	4,42	3,29	3,92	0,65	0,86	0,65	4,37	2,81	3,85
22	Lux_Car	hs	f	AMR _NI	yes	Sharp	9	4,29	3,38	3,88	0,62	0,58	0,68	4,28	2,83	3,78
24	Lux_Car	hs	f	AMR _NIII	yes	Sharp	9	2,54	2,96	2,46	0,93	0,69	0,66	2,18	2,67	2,05
31	Lux_Car	hf	f	AMR _NI	no	Smooth	9	3,58	2,63	2,96	0,88	0,71	0,81	4,26	2,37	3,56
37	Lux_Car	hf	f	AMR _NI	no	Sharp	9	3,50	1,92	2,79	0,78	0,72	0,78	3,51	2,28	2,88
43	Lux_Car	hf	f	AMR _NI	yes	Smooth	9	3,79	2,21	2,96	0,78	0,72	0,69	4,02	2,37	3,35
49	Lux_Car	hf	f	AMR _NI	yes	Sharp	9	3,54	2,17	2,67	0,66	0,64	0,64	3,80	2,35	3,15
79	Lux_Car	hs	m	AMR _NI	yes	Sharp	18	3,83	3,92	3,79	0,82	0,41	0,78	3,49	4,28	3,57
81	Lux_Car	hs	m	AMR _NIII	yes	Sharp	18	2,88	3,54	2,75	0,85	0,59	0,74	2,20	4,03	2,53
103	Lux_Car	hf	m	AMR _NI	yes	Sharp	9	3,54	2,04	2,71	0,78	0,55	0,62	3,81	2,42	3,19
201	Crossroads	hf	m	AMR _NIII	no	Sharp	9	1,25	1,21	1,17	0,68	0,41	0,38	1,22	1,44	1,24
276	Office_Noise	hs	f	G722_NIII	no	Smooth	9	2,96	3,92	3,04	1,27	0,41	1,00	3,40	3,95	3,42
277	Office_Noise	hs	f	G722_NI	no	Smooth	18	4,58	4,21	4,33	0,65	0,41	0,64	4,47	4,22	4,40
289	Office_Noise	hs	f	G722_NI	yes	Smooth	18	4,17	4,08	4,17	0,92	0,58	0,76	4,53	3,97	4,39
292	Office_Noise	hs	f	G722_NI	yes	Sharp	9	4,63	3,79	4,25	0,58	0,41	0,61	4,64	3,87	4,47
297	Office_Noise	hs	f	G722_NIII	yes	Sharp	18	4,38	4,38	4,33	0,71	0,49	0,64	3,62	4,71	3,75
331	Office_Noise	hs	m	G722_NI	no	Smooth	18	4,67	4,50	4,63	0,56	0,51	0,49	4,64	4,00	4,50
334	Office_Noise	hs	m	G722_NI	no	Sharp	9	4,46	4,13	4,46	0,78	0,61	0,72	4,63	4,08	4,50
373	Office_Noise	hf	m	G722_NI	yes	Sharp	9	4,88	3,83	4,50	0,34	0,48	0,59	4,52	3,52	4,25
379	Pub_Noise	hs	f	G722_NI	no NSA	no NSA	no NSA	2,96	1,63	2,25	0,81	0,65	0,61	2,73	1,75	1,99
384	Pub_Noise	hs	f	G722_NIII	no	Smooth	9	2,54	1,83	2,13	0,66	0,56	0,61	2,69	2,01	2,10
390	Pub_Noise	hs	f	G722_NIII	no	Sharp	9	2,25	1,83	2,00	0,61	0,56	0,42	2,31	1,52	1,55
393	Pub_Noise	hs	f	G722_NIII	no	Sharp	18	2,88	1,79	2,29	0,61	0,51	0,75	2,44	1,52	1,64
396	Pub_Noise	hs	f	G722_NIII	yes	Smooth	9	2,00	1,96	1,71	0,66	0,36	0,46	1,62	2,05	1,44
415	Pub_Noise	hs	m	G722_NI	no	Sharp	9	3,46	1,38	2,25	0,78	0,58	0,61	3,64	1,46	2,53
427	Pub_Noise	hs	m	G722_NI	yes	Sharp	9	3,63	1,46	2,58	0,92	0,51	0,88	3,48	1,73	2,57
432	Pub_Noise	hs	m	G722_NIII	yes	Sharp	18	2,54	2,50	2,33	0,83	0,66	0,64	2,58	2,08	2,05

Annex G:
Void

Annex H:
Extension of the EG 202 396-3 Speech Quality Test Method to Narrowband: Adaptation, Training and Validation

The first version of EG 202 396-3 was restricted to wideband application only. Due to the lack of freely available databases containing narrowband speech and evaluated according to ITU-T Recommendation P.835 [i.3], a new database including 263 conditions was created. This database includes a wide variety of different impairments found in today's communication systems including mobile and stationary handset/hands-free terminals.

The annex describes the adaptation of the model to narrowband scenarios, as well as some adjustments in the calculation and pre-processing which had to be done but without modifying the main principles of the algorithm.

Design of the new database

The base for each objective model is a database, containing speech samples (with references) and subjective MOS-LQSN scores from listening tests. The output scores of the model are always related to these subjective ratings.

For an extension to narrowband mode, a new database had to be designed. The database from the ETSI STF 294 project allowed the prediction of wideband speech based on French (and Czech) speech sequences. Based on the good experience with the well-balanced distribution of background noises and handset/hands-free modes in this old database, the new database was designed in a similar way.

General design of ITU-T Recommendation P.835 [i.3] listening-only test.

Table H.1: Comparison of Databases

	ETSI STF WB Database	**HEAD acoustics NB Database**
Language	French	English
Speakers	1 male **or** 1 female per condition	2 male **and** 2 female per condition
Different speakers	2	8
Training Validation	179 81	216 50

Table H.2: Distribution of conditions according background noise and handset/hands-free mode

	ETSI STF WB Database		**HEAD acoustics NB Database**	
	Handset	Hands-free	Handset	Hands-free
Overall	116	63	200	66
Background Noises:				
Car	23	22	40	25
Crossroad	18	18	36	8
Road	25	0	43	9
Office	27	23	39	13
Pub / Café	23	0	42	10

In the ETSI STF 294 project, all conditions were simulated offline. In the new narrowband database, 184 of 266 conditions were recorded from real devices in sending direction, 82 conditions were also simulated offline in the same way like in the STF 294 project:

- Recording of "Unprocessed Signal" at position of DUT.
- Background Noise Simulation according EG 202 396-1 [i.1].
- Simulation Steps:
  - IRS SND Filter.
  - Speech Enhancements / Noise Reduction.

- Coder + Decoder.

To simulate typical communication systems, the following processing steps were used:

- "Speech Enhancement":
  - Different MMSE Algorithms.
  - Different Algorithms with spectral subtraction.
  - Without any processing.
- Coder + Decoder:
  - G.726, G.729A.
  - iLBC.
  - Speex HiQ / LQ.
  - Without any coding/decoding.

Presentation of speech material in listening test:

The listening test for the new database was performed according to ITU-T Recommendation P.835 [i.3], where naïve listeners give three different votes (S-, N- and G-MOS for speech, noise and global quality) to a single sample.

Compared to the STF 294 project, some moderate changes based on the experience from the STF 294 project were introduced in the procedure. The design differences between the STF 294 database and the new database are listed in table H.3.

Table H.3: Design differences between the STF 294 database and the new database mode

	ETSI STF WB Database	HEAD acoustics NB Database
Sentences / Sample	1	2
Duration of Sample	4s	8-9s
Samples / Condition	6	4
Votes / Sample	4	6
Votes / Condition	24	24
Diotic ASL	79 dB SPL	73 dB SPL
Pre-Filtering	None / flat	IRS RCV

The main differences are:

- The amount of speech and noise parts in each sample was increased, so that the vote is more reliable.
- The listening level of active speech was decreased from 79 dB SPL to 73 dB SPL - an expert test led to the conclusion that this diotic level is preferred by listeners over a large range of signal-to-noise ratios.
- The narrowband speech material was prefiltered with an IRS RCV - simulates a reference listening system according to ITU-T Recommendation P.800 [i.4].

File Processing / Calculations

For the new narrowband mode, the clean speech and the unprocessed signal are filtered with an intermediate reference system (IRS) in sending and receiving direction. With this pre-processing step, all further analyses refer to a perfect transmission over a typical narrowband telephony network.

In the calculation of mean and variance from (Delta-) Relative Approach spectrograms, the limits of the frequency range are also adopted to the narrowband mode:

Table H.4: Bandwidth of speech sequences

	ETSI STF WB Database	HEAD acoustics NB Database
f_{min}	50 Hz	200 Hz
f_{max}	7 000 Hz	3 600 Hz

Prediction Results

Overall, there are 263 conditions in the new narrowband database. The training of the model was done using 213 randomly chosen conditions; the remaining 50 conditions were used to test the model against unknown, retained data (in terms of data which were not used to train the model). This process of training and validation data was also used in the ETSI STF 294 project.

The correlation coefficients and root-mean-square error between the subjective data from the listening test and the prediction of the narrowband adapted model are shown in the following table and in the scatter plots below.

Table H.5: Correlation coefficients and root-mean-square error

	ETSI STF WB Database				HEAD acoustics NB Database			
	Training		Validation		Training.		Validation	
	Corr.	RMSE	Corr.	RMSE	Corr.	RMSE	Corr.	RMSE
S-MOS	91,2%	0,37	93,0%	0,33	91,6%	0,37	90,0%	0,45
N-MOS	94,3%	0,27	92,4%	0,32	94,2%	0,33	93,5%	0,35
G-MOS	94,6%	0,25	93,5%	0,28	94,3%	0,31	93,2%	0,36

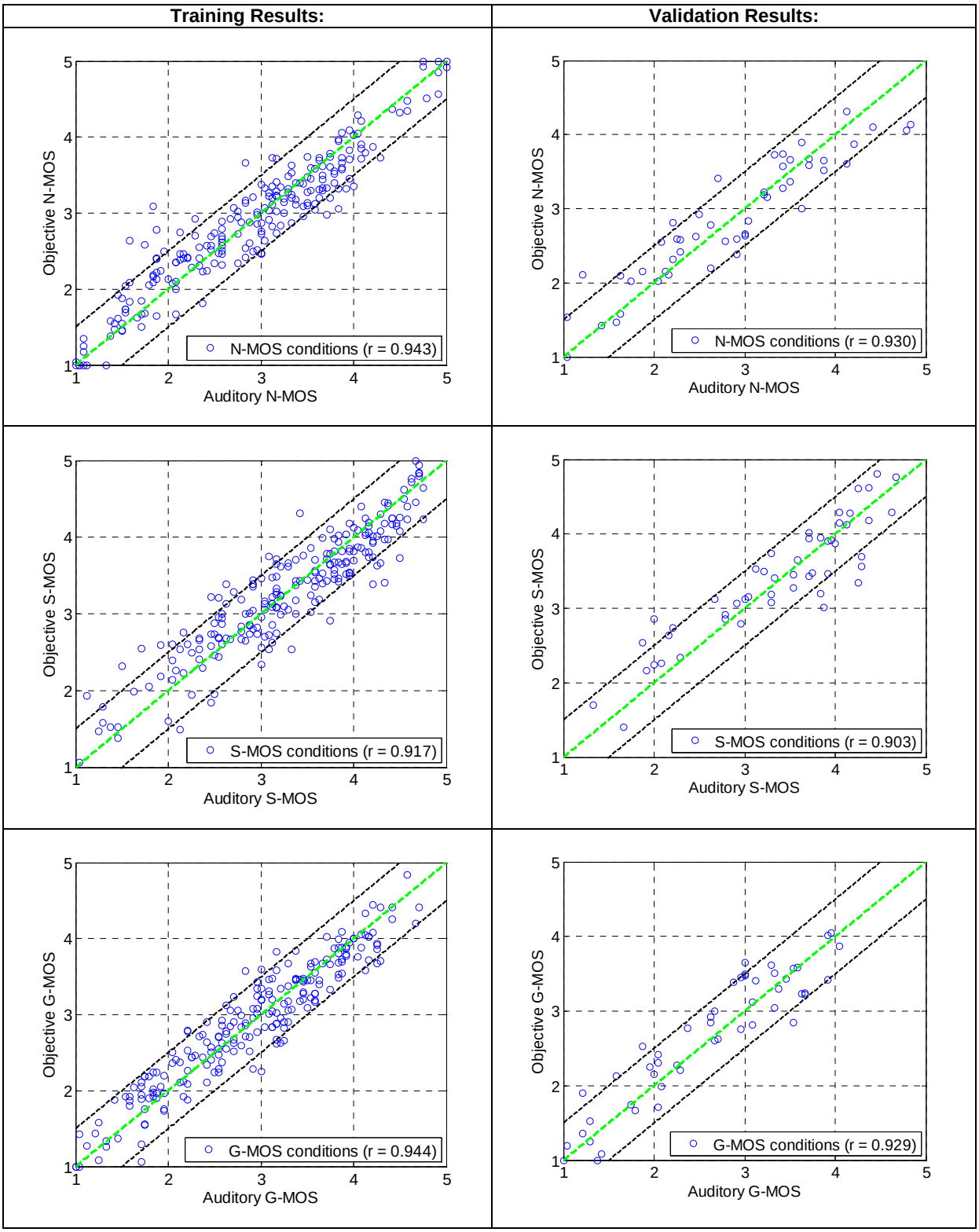


Figure H.1: HEAD acoustics NB Database - Comparison subjective vs. objective data

Annex I:
Validation results of the modified EG 202 396-3 objective speech quality model for narrowband data

I.1 Introduction

To evaluate the proposed narrowband extension of the objective test method described in EG 202 396-3, a set of ITU-T Recommendation P.835 [i.3] narrowband databases was provided by France Telecom R&D. A software tool with the implemented narrowband extension was developed by HEAD acoustics GmbH and made it available to France Telecom R&D. With this software tool, the analysis of the databases was carried out. In this contribution, the prediction results of this analysis are introduced for a discussion of the performance of the new extension.

I.2 Description of the Databases

The tested databases contain the result of applying some typical and realistic transmission network scenarios and traffic patterns to a selected group of speech recordings.

Four databases were built and evaluated during 4 different listening tests. For each database (DB), the speech samples consist of 4 talkers (2 males, 2 females) with up to 6 different double sentences per talker. The language is English for DB1 and French for DB2, DB3 and DB4.

The samples are narrow-band (NB), the sampling frequency is Fe=8 kHz. Typical noisy background ambiences are used: office, street, babble.

Table I.1: Description of DB1

Condition description	Number of conditions	Total
Language	English	1
Speakers	2 males 2 females	4
Noisy Background	Street Office Babble	3
SNR	2 levels for each noise	2
Noise Reduction	NR3, NR4, NR6	3
TOTAL		72

Table I.2: Description of DB2

Condition description	Number of conditions	Total
Language	French	1
Speakers	2 males 2 females	4
Noisy Background	Street Office Babble	3
SNR	2 levels for each noise	2
Noise Reduction	NR3, NR4, NR5, NR6	4
TOTAL		96

Table I.3: Description of DB3

Condition description	Number of conditions	Total
Language	French	1
Speakers	2 males 2 females	4
Noisy Background	Street Office Babble	3
SNR	2 levels for each noise	2
Noise Reduction	NR1, NR2	2
TOTAL		48

Table I.4: Description of DB4

Condition description	Number of conditions	Total
Language	French	1
Speakers	2 males 2 females	4
Noisy Background	Street Office Babble	3
SNR	1 level	1
Noise Reduction	NR1, NR2	2
TOTAL		24

I.3 Collection of the subjective scores

The methodology used for the 4 subjective tests is in accordance with ITU-T Recommendation P.835 [i.3]. Each trial contains three presentations of one sample, each presentation is followed by a silent voting period of 4 s. Each sample is 8 s in duration (two sentences per sample). For the two first presentations, listeners rate either the signal or the background depending on the rating scale order specified for that trial (for each listening test, half of the subjects rated first the signal and the other half rated first the noise). For the signal, subjects are instructed to attend *only* to the *speech signal* and rate the speech on the five-category distortion scale shown in figure 1 for French listeners and figure 4 for English listeners. For the background, subjects are instructed to attend *only* to the *background* and rate the background on the five-category intrusiveness scale shown in figure 2 for French listeners and figure 5 for English listeners. For the third presentation in each trial, subjects are instructed to listen to the speech + background and rate it on the five-category overall quality scale shown in figure 3 for French listeners and figure 6 for English listeners, the Mean Opinion Score (MOS) used with the ACR.

Note that French people participated to the tests with French samples and English people to the test with English samples.

```
Séance 1 Bloc 1 Essai 1
En vous concentrant UNIQUEMENT sur le SIGNAL VOCAL,
choisissez la catégorie qui décrit le mieux l'échantillon
que vous venez d'écouter
Le SIGNAL VOCAL dans cet échantillon était
5 – NON DÉFORMÉ
4 – LEGEREMENT DÉFORMÉ
3 – MOYENNEMENT DÉFORMÉ
2 – DÉFORMÉ
1 – TRES DÉFORMÉ
```

Figure I.1: French speech signal rating scale

Séance 1 Bloc 1 Essai 1
En vous concentrant **UNIQUEMENT** sur le **BRUIT DE FOND**,
choisissez la catégorie qui décrit le mieux
l 'échantillon que vous venez d'écouter
Le **BRUIT DE FOND** dans cet échantillon était
5 – IMPERCEPTIBLE
4 – PERCEPTIBLE MAIS NON GÊNANT
3 – UN PEU GÊNANT
2 – GENANT
1 – TRÈS GENANT

Figure I.2: French background noise rating scale

Choisir la catégorie qui décrit le mieux l'échantillon que vous
venez d'écouter pour des communications vocales courantes
L'**ECHANTILLON VOCAL GLOBAL** était
5 – EXCELLENT
4 – BON
3 - PASSABLE
2 – MÉDIOCRE
1 – MAUVAIS

Figure I.3: French overall quality rating scale

Session 1 Block 1 Trial 1
Attending **ONLY to the SPEECH SIGNAL**, select the category
which best describes the sample you just heard.
the **SPEECH SIGNAL** in this sample was
5 – NOT DISTORTED
4 – SLIGHTLY DISTORTED
3 - SOMEWHAT DISTORTED
2 – FAIRLY DISTORTED
1 – VERY DISTORTED

Figure I.4: English speech signal rating scale

Session 1 Block 1 Trial 1
Attending **ONLY to the BACKGROUND**, select the category
which best describes the sample you just heard.
the **BACKGROUND** in this sample was
5 – NOT NOTICEABLE
4 – SLIGHTLY NOTICEABLE
3 – NOTICEABLE BUT NOT INTRUSIVE
2 – SOMEWHAT INTRUSIVE
1 – VERY INTRUSIVE

Figure I.5: English background noise rating scale

Select the category which best describes the sample you

just heard for purposes of everyday speech communication.

the **OVERALL SPEECH SAMPLE** was

5 – EXCELLENT

4 – GOOD

3 - FAIR

2 – POOR

1 – BAD

Figure I.6: English overall quality rating scale

I.4 Differences: HEAD acoustics training database vs. France Telecom validation databases

In principle, the listening tests in the HEAD acoustics database and the introduced France Telecom R&D databases were applied according to the test procedure of ITU-T Recommendation P.835 [i.3]. Anyhow, there are some important differences between the training and the validation databases, which may account for possible deviations between the subjective data and the objective predictions:

- Distance metrics were optimized to a special set of English P.501 sentences.
- French test sentences were completely unknown to the trained model.
- Language differences: distribution of phonems differs from English to French.
- HEAD acoustics database was designed to cover the full range of S-/N-/G-MOS (approximately MOS values from 1.0 to 5.0); France Telecom databases were designed to test certain problems (different parameters of noise reduction types, SNRs, etc.) and not necessarily cover the whole quality range (this is the case for DB4).
- Recordings of HEAD acoustics database included the acoustical influence between speaker & DUT (handset, hands-free & handheld mode) and the processing of complete terminals; databases from France Telecom were processed completely electrically.
- The proposed narrowband as well as the already standardized wideband model switches regression coefficients for S-MOS dependent on the N-MOS. The two N-MOS thresholds which are part of the model and which control the S-MOS ideally also should be extracted from the database; this was ignored here for simplification (was hardcoded set to 2.48 and 3.30), but this may explain deviations for S-MOS and G-MOS.

I.5 Results

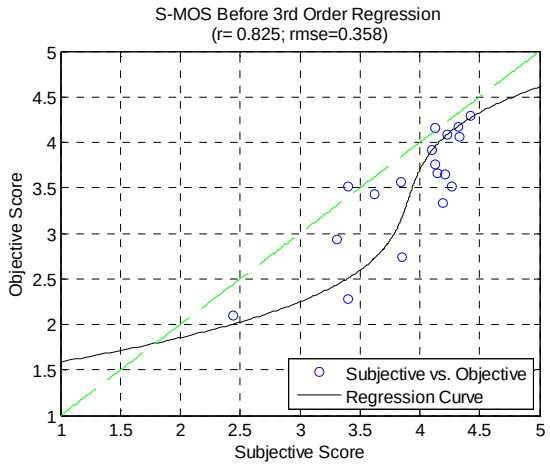
The databases of France Telecom R&D consist of several amounts of conditions (from 6 to 24). Within each condition, there are 4 subsamples (spoken by 2 male and 2 female speakers), which were also presented in the listening tests of the databases. These subsamples were used as an input for the modified prediction model. The output scores of one condition are then averaged to a prediction result of one condition. In all further comparisons, the subjective and objective MOS are shown and discussed per condition.

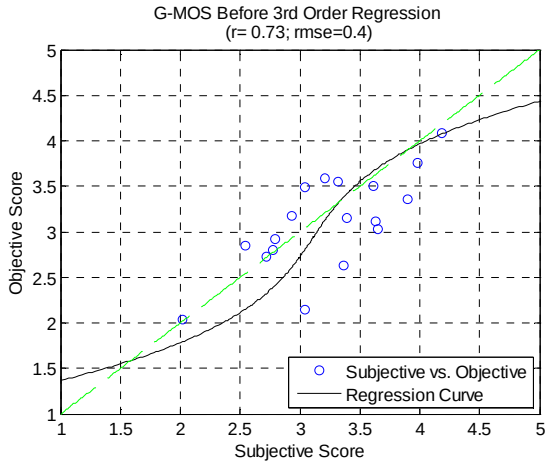
As comparison metrics between the objective and subjective MOS, the Pearson correlation coefficient, the root-mean-square error (RMSE) and Spearman's rank correlation coefficient are calculated. To give a complete overview of the distribution of the subjective MOS and its estimation, the scatter plots are also depicted.

I.6 Unmapped Results

In a first step, the results of the different databases are compared against the subjective data resulting from listening tests. Although deviations between the objective and subjective data can be expected when using the raw output scores of the model extension, this comparison gives first indications about the performance.

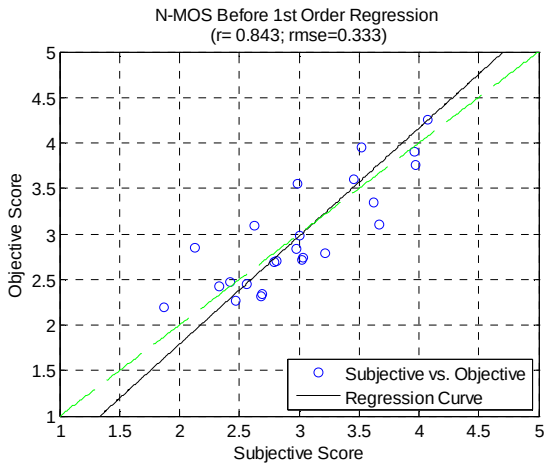
Database #1

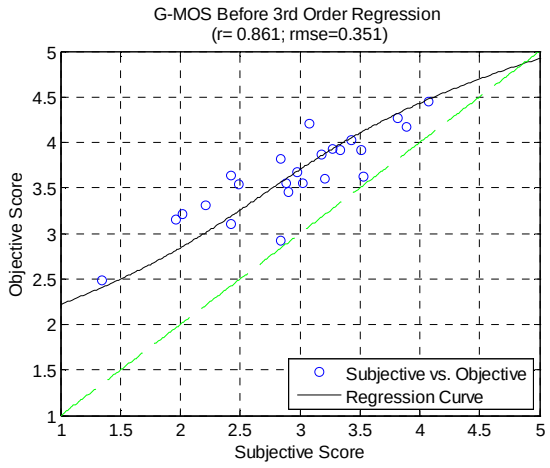




Metrics without mapping:			
	Correlation	**RMSE**	**Rank Corr.**
S-MOS	0,825	0,358	0,730
N-MOS	0,931	0,331	0,920
G-MOS	0,730	0,400	0,637

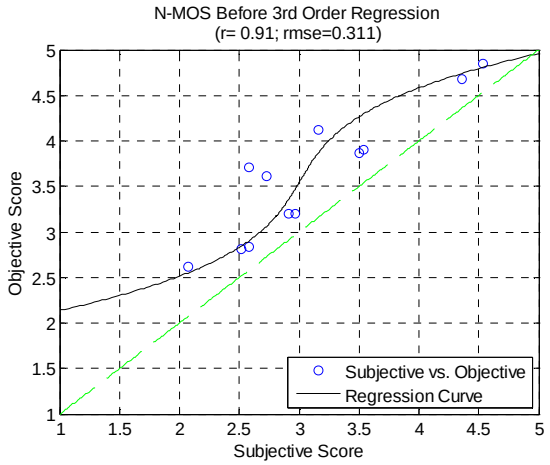
Database #2

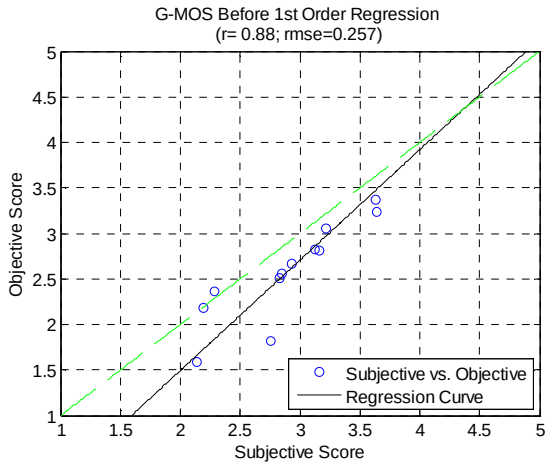




Metrics without mapping:			
	Correlation	**RMSE**	**Rank Corr.**
S-MOS	0,840	0,453	0,822
N-MOS	0,843	0,333	0,799
G-MOS	0,861	0,351	0,834

Database #3

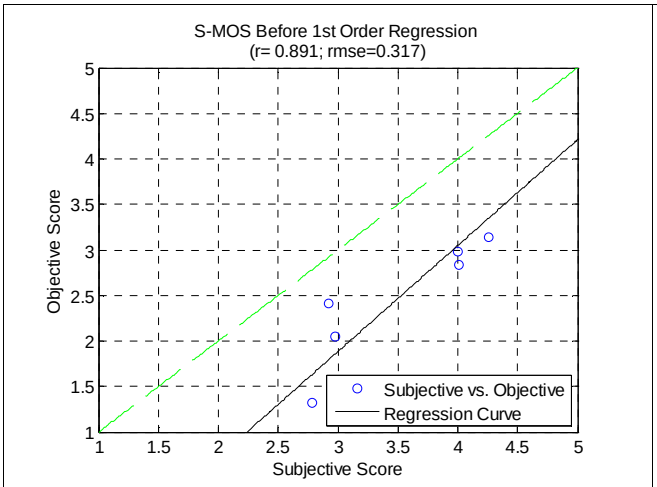


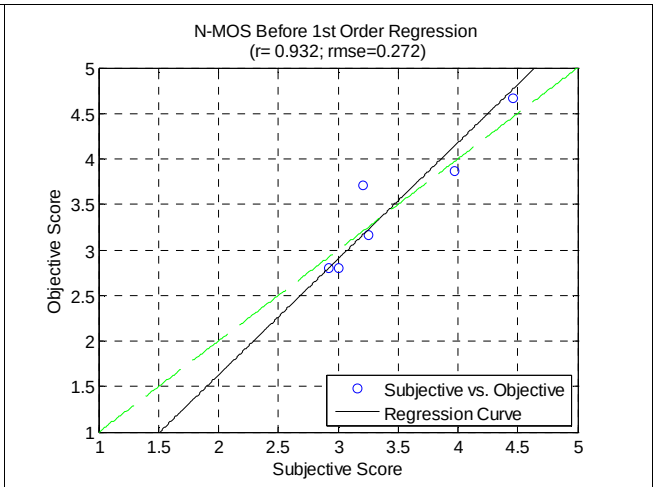


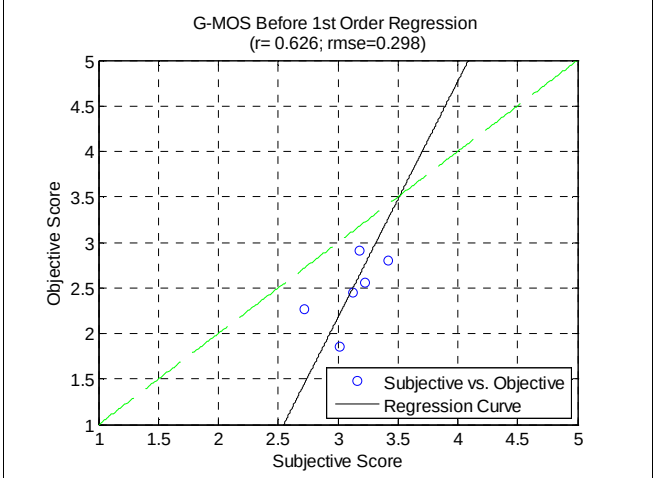
Metrics without mapping:

	Correlation	RMSE	Rank Corr.
S-MOS	0,958	0,234	0,99
N-MOS	0,910	0,311	0,90
G-MOS	0,880	0,257	0,97

Database #4







Metrics without mapping:			
	Correlation	**RMSE**	**Rank Corr.**
S-MOS	0,891	0,317	0,89
N-MOS	0,932	0,272	0,89
G-MOS	0,626	0,298	0,77

I.7 Mapped Results

I.7.1 Use of mapping functions

The ITU-T Recommendation P.835 [i.3] narrowband database by HEAD acoustics GmbH was designed to cover a wide range of communication systems. In this context, also the quality range was widely scattered. The new narrowband mode of the model was trained with this data.

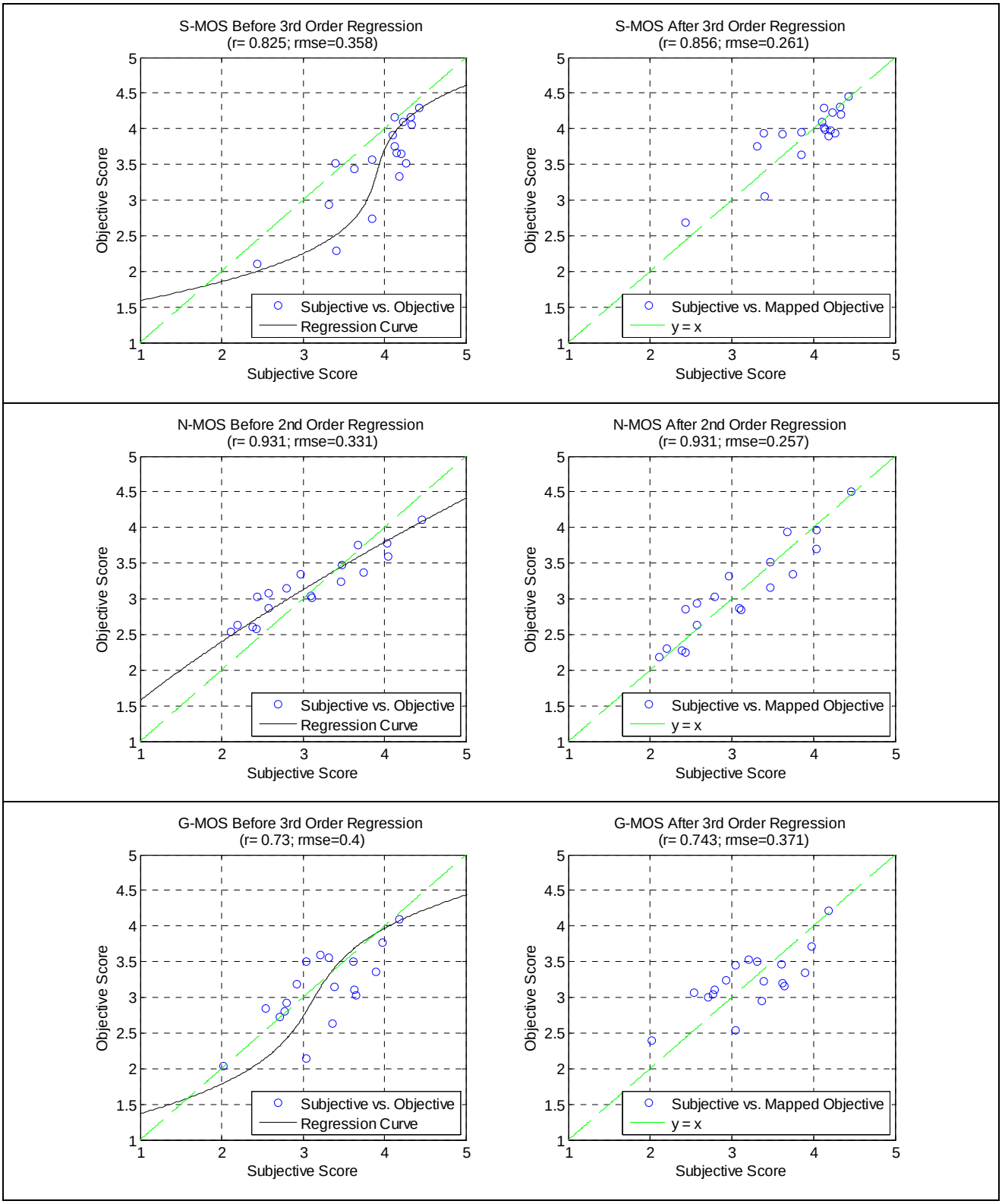
When output scores are calculated on data, which comes from other databases / listening tests, where the conditions are designed differently as the ones of the HEAD acoustics narrowband database, the prediction results need to be adapted to the listening test in order to take into account differences in the listening tests. This is usually done with a third order regression curve (see ITU-T e.g. developments of P.862 and 862.1).

Although it is always allowed to use a third order mapping, it is not always reasonable to use this approach. The calculated mapping curves should be applied to compensate effects like different designs of databases (including different quality ranges and parameters). Mathematically spoken, these curves should compensate compression, expansions, shifts and constant offsets. As a consequence, only monotonically increasing mapping functions are meaningful.

Dependent on the amount of conditions and the distribution of the subjective / objective data, it is not always possible to find such a mapping function of 3rd order. But to find a "maximum order mapping function", the following algorithm was used:

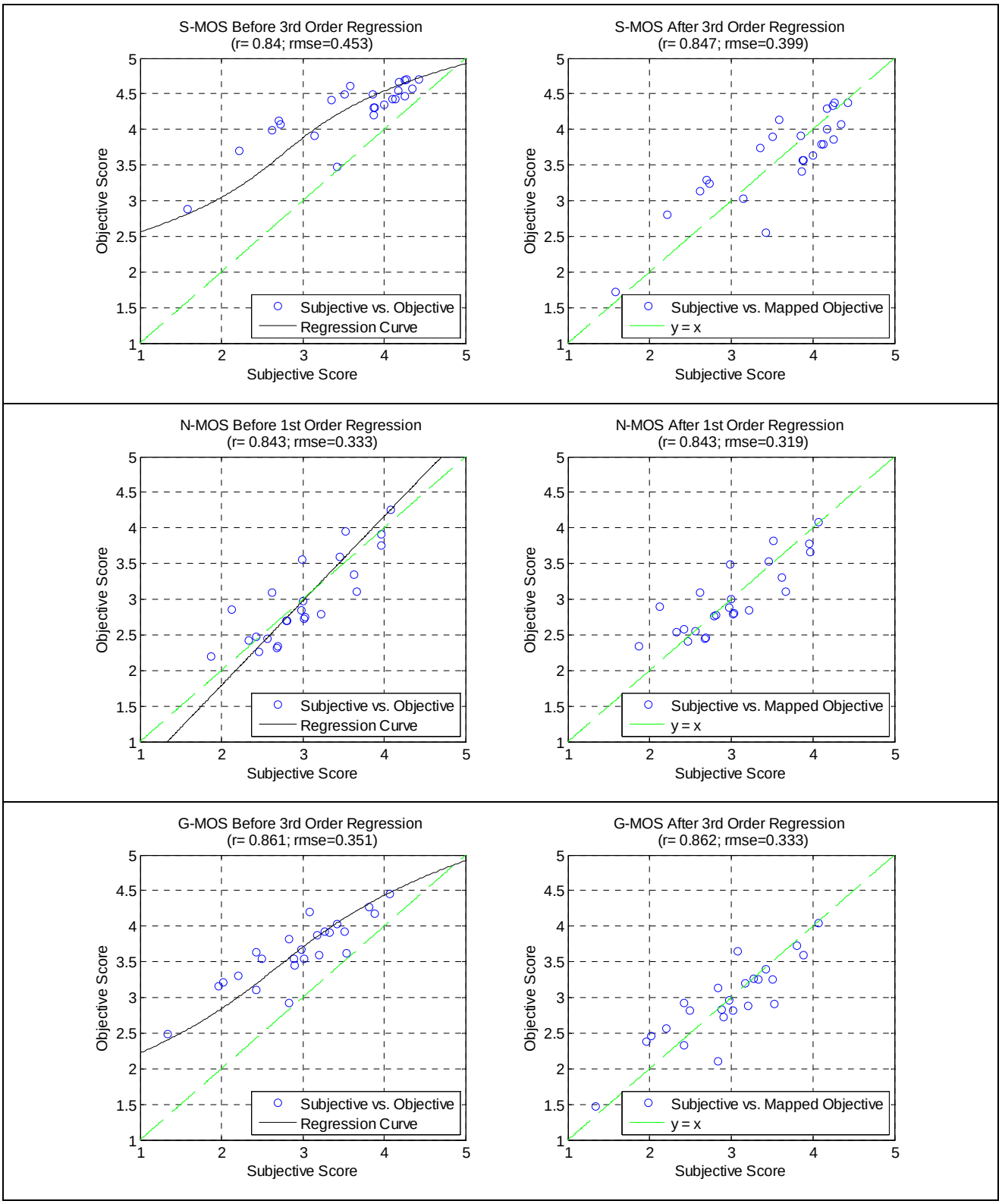
- Apply 3rd order mapping to the objective data.
- Check: Mapping function monotonically increasing in the MOS range 1.0 to 5.0?
  - If not: Apply 2nd order mapping to the objective data.
  - Check: Mapping function monotonically increasing in the MOS range?
    - If not: Apply only 1st order mapping (always monotonically increasing).

Database #1



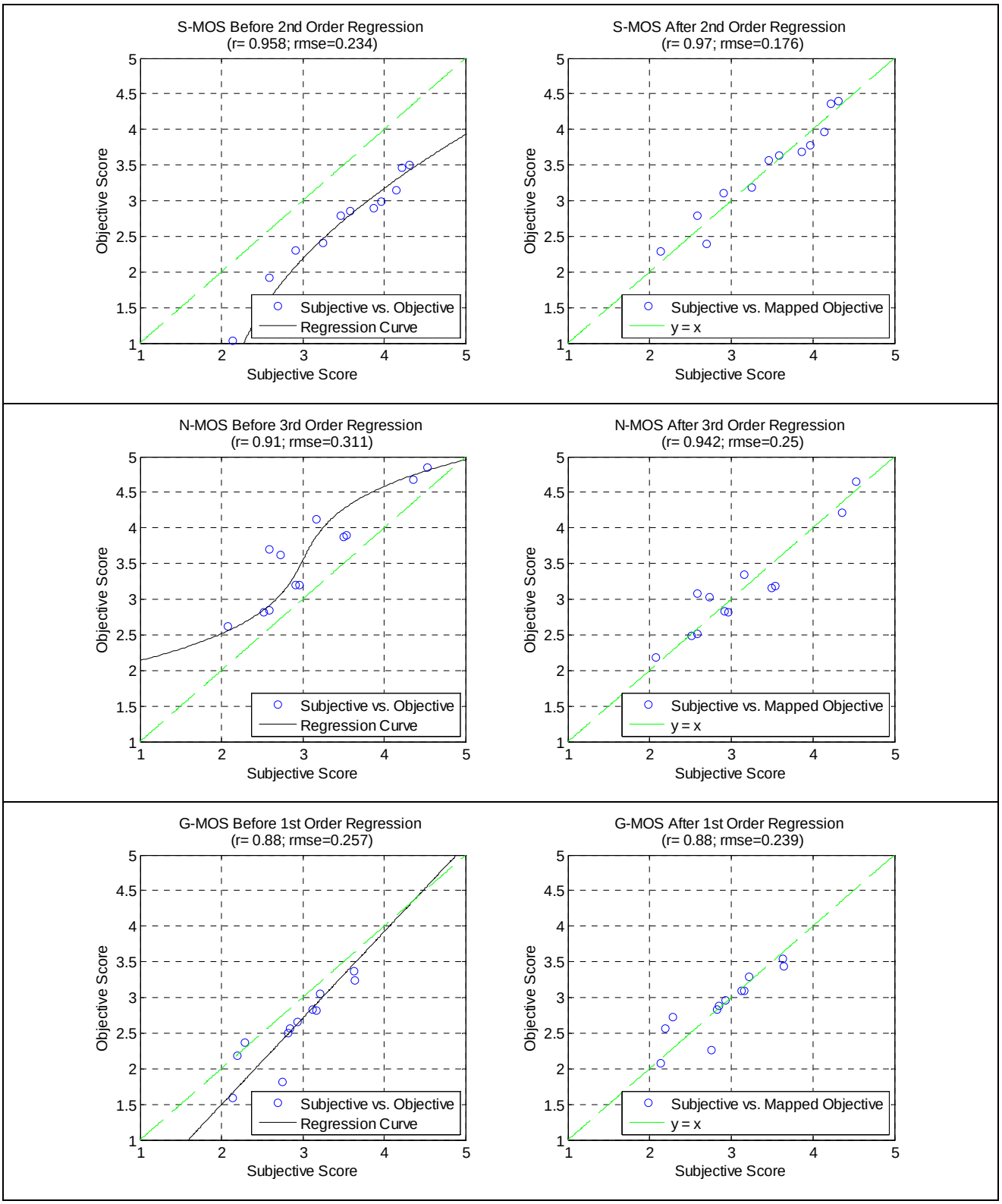
Metrics:	Without Mapping			After Regression	
	Correlation	RMSE	Rank Corr.	Correlation	RMSE
S-MOS	0,825	0,358	0,730	0,856	0,261
N-MOS	0,931	0,331	0,920	0,931	0,257
G-MOS	0,730	0,400	0,637	0,743	0,371

Database #2



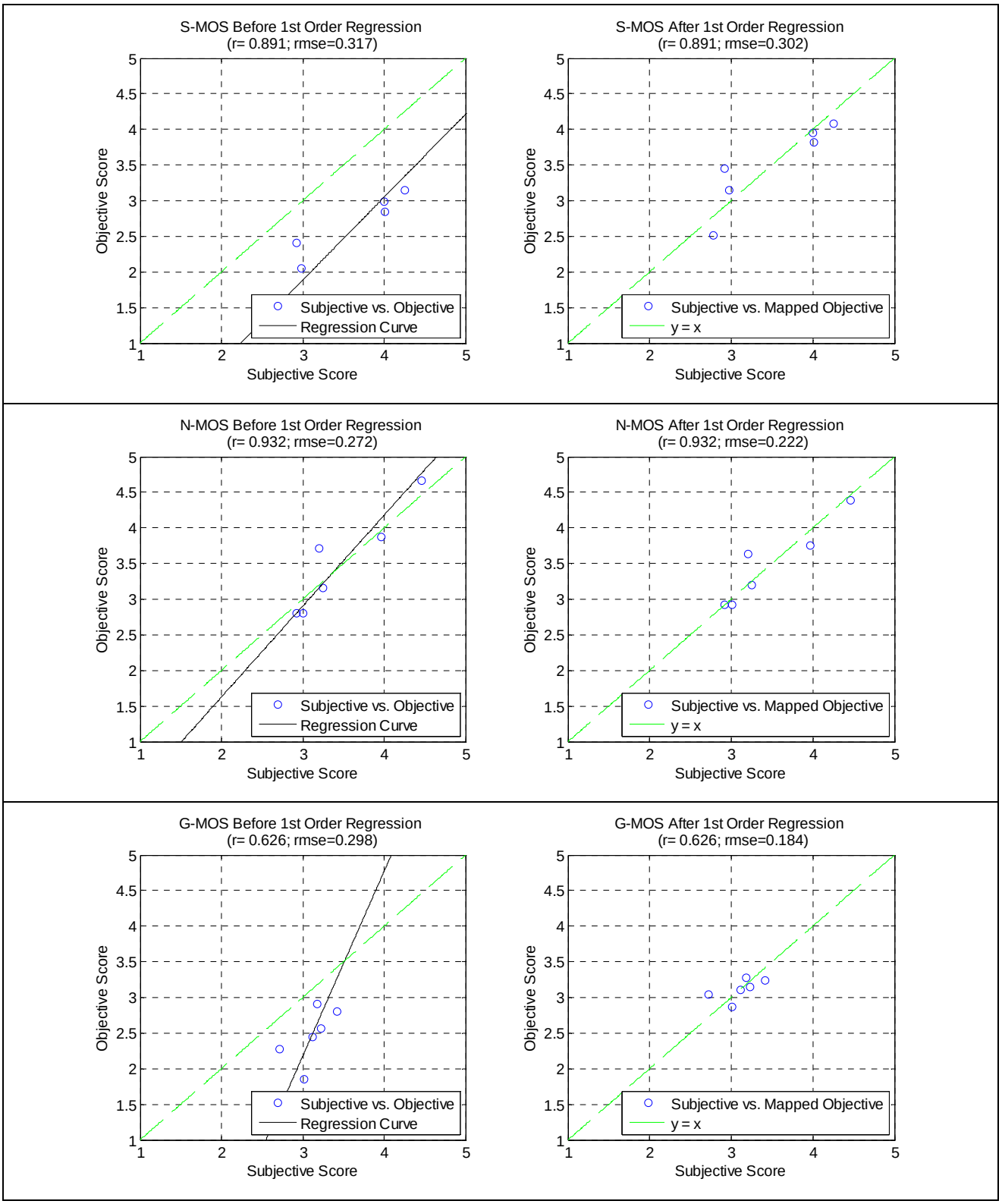
Metrics:	Without Mapping			After Regression	
	Correlation	RMSE	Rank Corr.	Correlation	RMSE
S-MOS	0,840	0,453	0,822	0,847	0,399
N-MOS	0,843	0,333	0,799	0,843	0,319
G-MOS	0,861	0,351	0,834	0,862	0,333

Database #3



Metrics:	Without Mapping			After Regression	
	Correlation	RMSE	Rank Corr.	Correlation	RMSE
S-MOS	0,958	0,234	0,99	0,970	0,176
N-MOS	0,910	0,311	0,90	0,942	0,250
G-MOS	0,880	0,257	0,97	0,880	0,239

Database #4



Metrics:	Without Mapping			After Regression	
	Correlation	RMSE	Rank Corr.	Correlation	RMSE
S-MOS	0,891	0,317	0,89	0,891	0,302
N-MOS	0,932	0,272	0,89	0,932	0,222
G-MOS	0,626	0,298	0,77	0,626	0,184

I.8 Conclusions

As mentioned above, every database was created on a separate listening test, with different types and amount of conditions. Therefore, also the results have to be reviewed individually per database.

In database #1, the unmapped prediction of the S-MOS has a low RMSE, which highly increases after the 3rd order mapping. The distribution of the S-MOS is concentrated on scores higher than 3,5, so this compression has to be compensated.

The quality range of the background noises in this database is comparable to the one in the HEAD acoustics database. A 2nd order regression is applied, which slightly improves the RMSE. But a high rank order correlation indicates a high consistency of the predicted scores.

The objective G-MOS calculated for this database has a similar mapping function like the S-MOS, which results from the composition of the G-MOS out of the S-MOS and the N-MOS. After a 3rd order mapping, the RMSE and the scatter plot indicate a stable prediction. The correlation coefficient remains nevertheless rather low (0,743). This means almost no progress compared to the one before mapping, due to a distribution of the subjective G-MOS mainly in the middle of the scale.

The objective S-MOS of database #2 is predicted too optimistic compared to the subjective data. This shift and slight compression at the upper end of the scale can be compensated with a 3rd order regression, which mainly improves the RMSE.

Like in the first database, the estimation of the N-MOS is comparable with the subjective N-MOS even without a mapping. To compensate the small shift, a linear (1st order) regression is applied.

The prediction of the G-MOS is comparable to one of the S-MOS; hence a nearly identical 3rd order regression is used to transform the results to the scale of this listening test.

Database #3 shows the highest correlation and lowest RMSE for all scores. This indicates a high similarity to the HEAD acoustics database on the one side, but may also be a result of the low amount of conditions in this database. Nevertheless, very high rank order correlation coefficients refer to a highly stable prediction quality.

In contrast to database #2, here the S-MOS is estimated always too low with a nearly constant offset. This offset and a slight compression at the lower end of the subjective S-MOS scale can be compensated with a 2nd order mapping.

For the N-MOS, a positive offset and a slight compression in the middle of the scale can be observed and is compensated with a 3rd order regression curve.

Nearly no additional mapping is needed for the objective G-MOS to match the subjective data. This results from the effect, that the N-MOS is predicted too high and the S-MOS too low. Only a slight 1st order transformation is applied to correct the small resulting shift.

In Database #4, again the objective S-MOS is predicted too low, but this time only with a constant offset. Therefore, only a 1st order regression line needs to be applied.

For the objective N-MOS, almost no mapping needed which is also indicated by the very high correlation coefficient. Thus also a 1st order regression is used to remove the slight shift.

The objective G-MOS obtains a very low correlation at the first look, even after an additional mapping, due to the clouded distribution of subjective data in the middle of the G-MOS scale (contrary to the S-MOS and the N-MOS, which are widely scattered and well predicted). That the G-MOS is estimated adequately is indicated by the low RMSE and a rank order coefficient, which is higher than the Pearson correlation coefficient. The subjective data show a wide distribution of the individual N-MOS and S-MOS values but only small differences in the overall score G-MOS. This effect which shows really the benefit of using the P.835 approach for measuring the speech quality in the presence of background noise is well modelled by the objective model.

The results of this evaluation on unknown and new databases show that the model exhibits quite good performance for the N-MOS estimation, slightly decreased performance for the S-MOS, the weakest point being the overall score estimation (G-MOS). The model was tested on French as well as on English new databases and the results do not show any language dependency. However, although the model was trained on English samples, the results on the only new English database (DB1) show performance of the model inferior to those observed during the training phase.

For information, the existing standardized psychoacoustic models which were designed to estimate the overall quality score as specified in ITU-T Recommendation P.800 [i.4] exhibit the following overall performance:

- a correlation of 0,935 (in narrow-band only) for the P.862 model for known databases.
- a correlation of 0,89 for P.563 model (however, it is well known that this model does not work well).
- for comparison, the performance of the models participating in the ITU-T competition for P.862 on known and unknown databases can be found in ITU-T contribution COM 12-117 [i.29].

History

Document history		
V1.1.1	August 2007	Publication
V1.2.1	November 2008	Membership Approval Procedure MV 20090116: 2008-11-18 to 2009-01-16
V1.2.1	January 2009	Publication