ETSI EG 202 396-3 V1.5.1 (2015-10)

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

Reference
REG/STQ-229

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://ipr.etsi.org).

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 and multimedia Transmission Quality (STQ); 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".

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

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 Recommendations ITU-T P.862 [i.16] and P.862.1 [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 present document is to be used in conjunction with:
  - ETSI ES 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.
  - ETSI 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 Recommendation ITU-T P.501 [i.13] for wideband and English speech sentences as defined in Recommendation ITU-T P.501 [i.13] for narrowband.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the reference document (including any amendments) applies.

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.

The following referenced documents are necessary for the application of the present document.

Not applicable.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the reference document (including any amendments) applies.

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

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI ES 202 396-1: "Speech and multimedia Transmission Quality (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] Recommendation ITU-T P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

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

[i.5] Recommendation ITU-T 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] Recommendation ITU-T 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] Recommendation ITU-T 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) volume 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] Recommendation ITU-T 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] Recommendation ITU-T P.862.1: "Mapping function for transforming P.862 raw result scores to MOS-LQO".

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

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

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

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

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

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

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

[i.25] Void.

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

[i.29] Void.

[i.30] ANSI S1.1-1986 (ASA 65-1986): "Specifications for Octave-Band and Fractional-Octave-Band Analog and Digital Filters", 1993.

[i.31] Recommendation ITU-T G.160 Appendix II, Amendment 2: "Voice enhancement devices: Revised Appendix II - Objective measures for the characterization of the basic functioning of noise reduction algorithms".

[i.32] ETSI TS 103 106: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods".

[i.33] Hastie, T.; Tibshirani, R.; Friedman, J.: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", New York: Springer-Verlag, 2001.

3 Symbols and abbreviations

3.1 Symbols

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

σ^2 Variance

3.2 Abbreviations

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

AMR Adaptive MultiRate
ASL Active Speech Level

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

BGN	BackGround Noise
CDF	Cumulative Density Function
dB SPL	Sound Pressure Level re 20 µPa in dB
DB	Data Base
DUT	Device Under Test
EFR	Enhance Full Rate
FR	Full Rate
G-MOS	Global MOS

NOTE: MOS related to the overall sample.

GSM	Global System for Mobile Communication
HATS	Head And Torso Simulator
IP	Internet Protocol
IRS	Intermediate Reference System
ITU	International Telecommunication Union
ITU-T	Telecom Standardization Body of ITU
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MOS-LQSN	Mean Opinion Score - Listening Quality Subjective Noise
MRP	Mouth Reference Point
NB	NarrowBand
NI	Network I conditions
NII	Network II conditions
NIII	Network III conditions
N-MOS	Noise MOS

NOTE: MOS related to the noise transmission only.

NR	Noise Reduction
NR (filter)	Noise Reduction (filter)
NSA	Noise Suppression Algorithm
PESQ	Perceptual Evaluation of Speech Quality
PLC	Packet Loss Concealment
RCV	ReCeiVe
RMS	Root Mean Square
RMSE	Random Mean Square Error
S-MOS	Speech MOS

NOTE: MOS related to the speech signal only.

SND	Sending Direction
SNR	Signal to Noise Ratio
SQTE	Speech Quality Test Event
SPL	Sound Pressure Level
STD	STandard Deviation
STF	Specialized Task Force
TMOS	TOSQA Mean Opinion Score
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 Recommendation ITU-T 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 Recommendation ITU-T 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 ETSI 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 e.g. Recommendation ITU-T 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 ETSI EG 202 396-2 [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 ETSI EG 202 396-2 [i.2] leading to non consistent judgements within this condition.

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

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 (non-native French 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 PEQ V. The equalization provided by the headphone manufacturer was used since this was the one used in the auditory French 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 Recommendation ITU-T 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.

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 ETSI EG 202 396-2 [i.2].

The wideband objective model 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 Recommendation ITU-T G.722.2 [i.21] and Recommendation ITU-T 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.

The sample preparation and nomenclatures for the new method are described in clause 6.2.

The calculation of *N-MOS*, *S-MOS* and *G-MOS* is described in detail in clauses 6.5 to 6.7.

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.

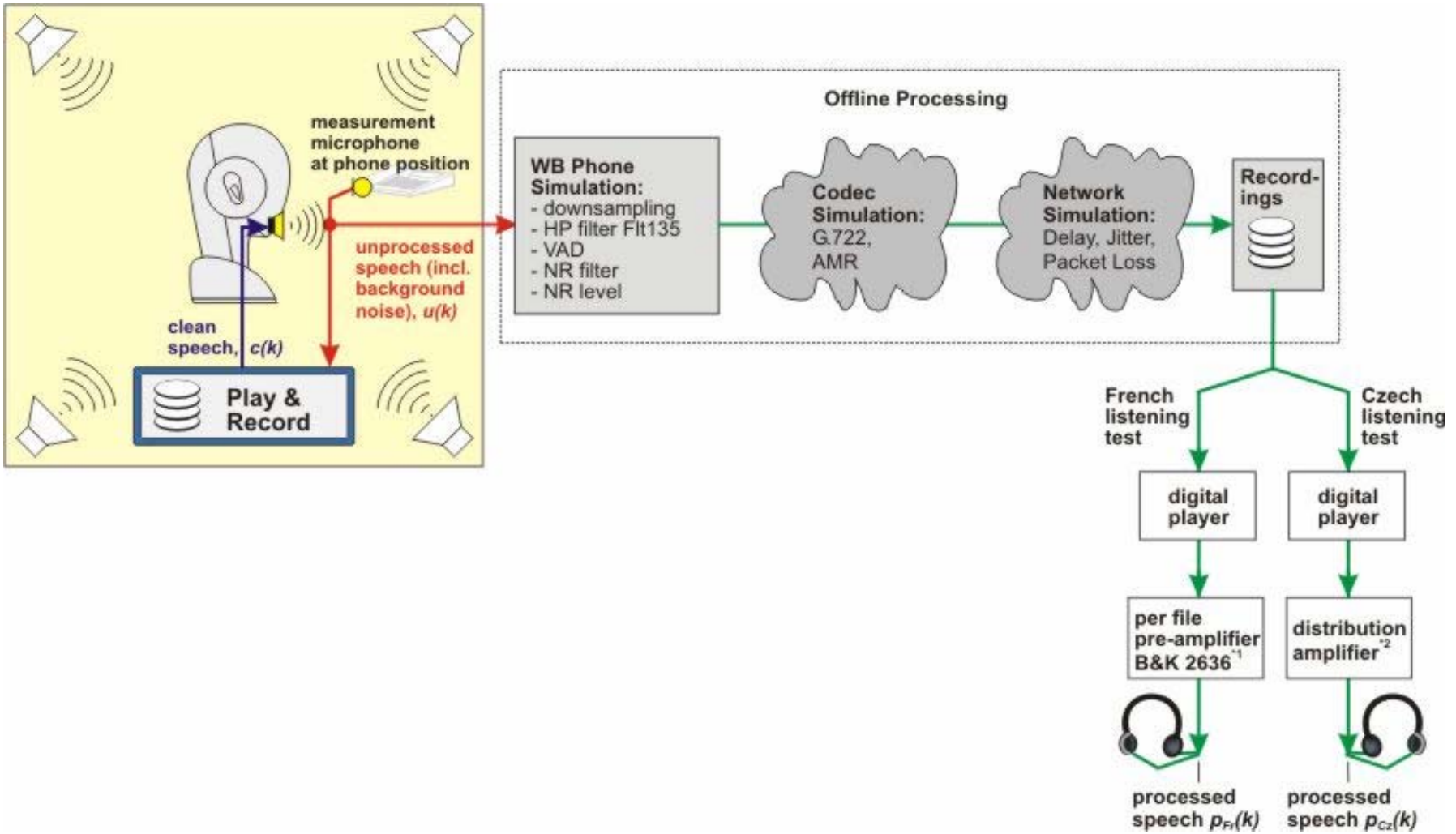
Different input signals can be accessed during the recording process and subsequently can be used for the calculation of N-MOS, S-MOS and G-MOS. Beside the signals used in the listening test ("processed signal"), two additional signals are used as a priori knowledge for the calculation:

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 ETSI EG 202 396-2 [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 (Recommendation ITU-T P.56 [i.22]).
NOTE 2: Once calibrated: -26 dBoV resulting to 79 dB SPL measured with a type 3.2 ear (Recommendation ITU-T P.57 [i.23]), 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 ETSI EG 202 396-2 [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 analyses lead to non-ambiguous results. E.g. by applying further processing algorithms in order to better separate between speech and noise parts.

For some of the following calculations, the information about speech and noise-only parts is needed. After the time alignment between the three signals as described above, the clean speech signal is segmented into frames and classified according to Recommendation ITU-T G.160 [i.31]. The signal parts classified as silence are assumed as background noise sections for unprocessed and processed signal. All other frames are assumed as active speech.

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 Recommendation ITU-T 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.

6.2.2 Nomenclature

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

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 ETSI EG 202 396-2 [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.

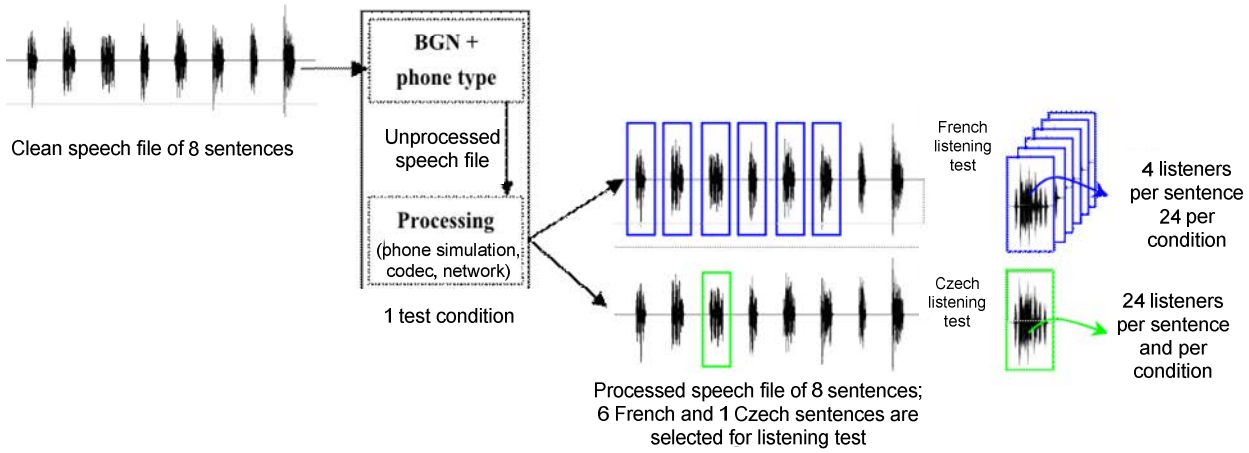


Figure 6.2: Nomenclature (file, condition, sentence)

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. The resulting auditory S-/N-/G-MOS per sentence were averaged to the condition MOS.

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. 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.

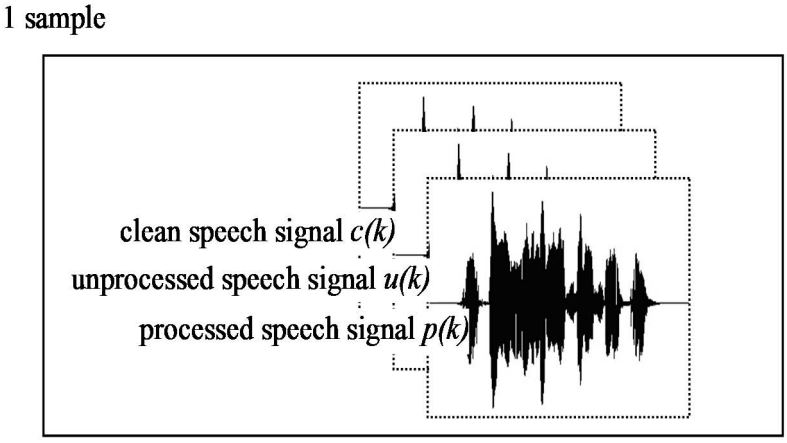


Figure 6.3: Nomenclature (sample)

All calculations in the following clauses 6.5 to 6.7 are always based on single sentences. The calculated objective MOS values of one condition are averaged to one objective condition MOS value. Comparisons with subjective MOS values are never conducted on a per-sample basis, only per-condition analyses are performed.

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

6.3 Additional Training data

In order to enlarge the training database regarding amount of conditions and real devices (the original work of ETSI EG 202 396-2 [i.2] only included simulated terminals), Orange kindly provided audio files and subjective results of a new auditory test. This new database was used for the development of ETSI TS 103 106 [i.32]. The database consists of 90 conditions with 12 sentences of 6 different talkers (3 male / 3 female), including the talkers presented in the experiments in ETSI EG 202 396-2 [i.2].

The focus of this additional database concentrates on state-of-the-art mobile devices (year 2012) in handset mode. Since the database in the original work ETSI EG 202 396-2 [i.2] also included many hands-free conditions, the bias between both datasets are different. All S-/N-/G-MOS values were known during the development phase of the model.

The overall training dataset then includes 179 + 90 = 269 conditions.

6.4 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.

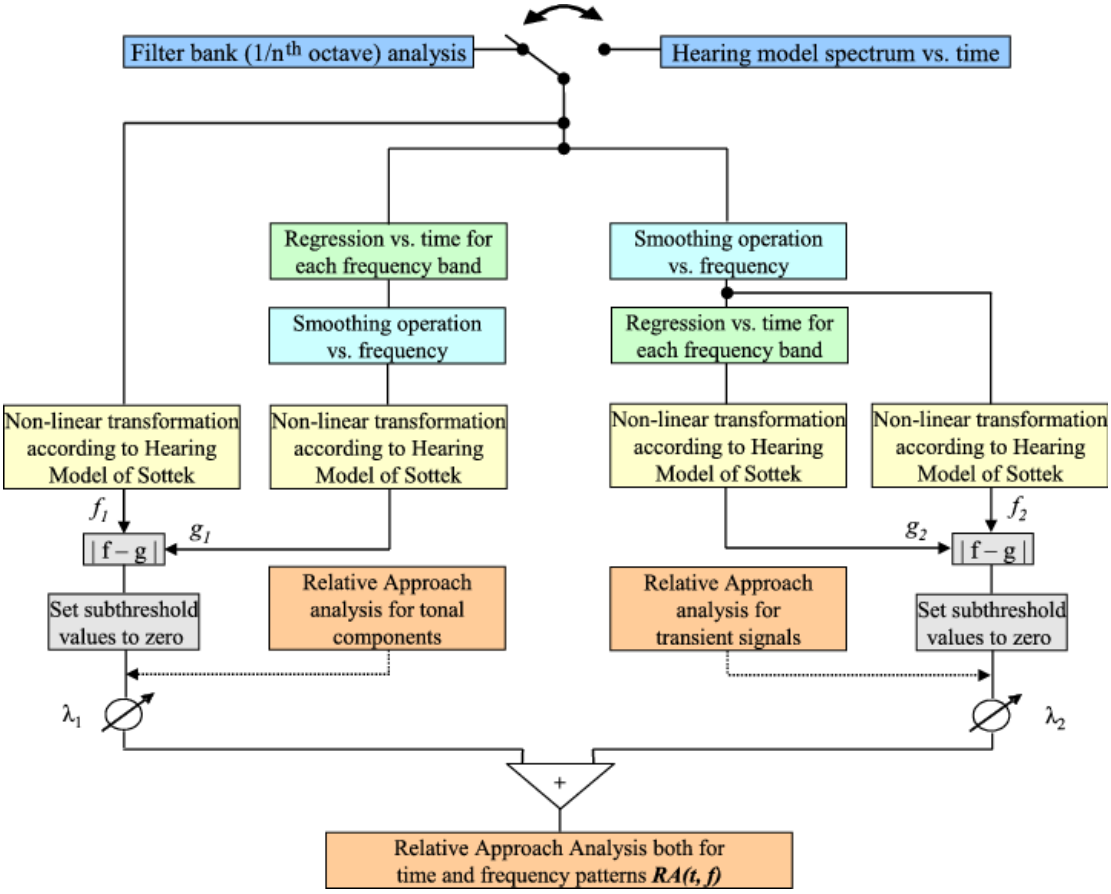


Figure 6.4: Block diagram of Relative Approach

The idea behind the Relative Approach analysis is based on the assumption that human hearing creates a continuous 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 patterns in both time and frequency. 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 (see also [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 pseudo-stationary 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 according to ANSI S1.1-1986 [i.30] or a spectral analysis based on a hearing model [i.27]. Both of them result in a spectral representation versus time. All calculations described in the following are based on the non-squared, but absolute magnitude of this spectrogram. All time-frequency bins are regarded in the physical unit [Pa] (not [Pa]^2).

The Relative Approach takes the absolute signal level into account. Therefore, the input data is calibrated to a realistic listening level and the physical unit of the input signals is pressure in Pascal (Pa). As input for either the filter bank or the Hearing Model signals adjusted to 79 dB SPL can be used (e.g. according to the French listening test) or signals with their original level after signal processing (e.g. according to the Czech listening test).

In the calculation, two regression / smoothing modes can be applied to the pre-processed signals (see figure 6.5), once versus time or versus frequency bands.

The estimation of the current magnitude $\hat{x}_t$ within a certain frequency band is calculated via a linear regression from a time window of the past 200 ms (see figure 6.5a).

The estimation of the current magnitude $\hat{x}_f$ within a certain time slot is calculated via a linear regression from neighbouring frequency bands (see figure 6.5b); 8 frequency bands above and 8 below are used for this calculation. For each time slot, this principle results into a sliding window of 17 frequency bands, including the current one.

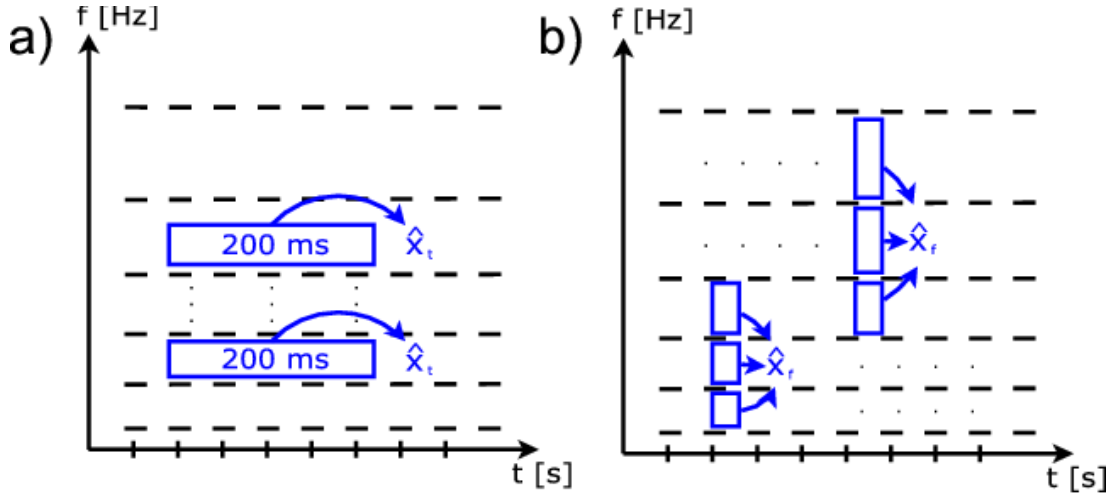


**Figure 6.5: Regression modes of Relative Approach:
a) versus time, b) versus frequency**

Another calculation method is the non-linear transformation according to the hearing model of Sottek [i.27]. No further hearing threshold or other spectral weighting is used. Due to the non-linear relationship between sound pressure and perceived loudness, the term "compressed pressure" in compressed Pascal (cPa) is used here as the physical unit of the output signal. Figure 6.6 compares this transformation against the transformation to a common sound pressure level (SPL).

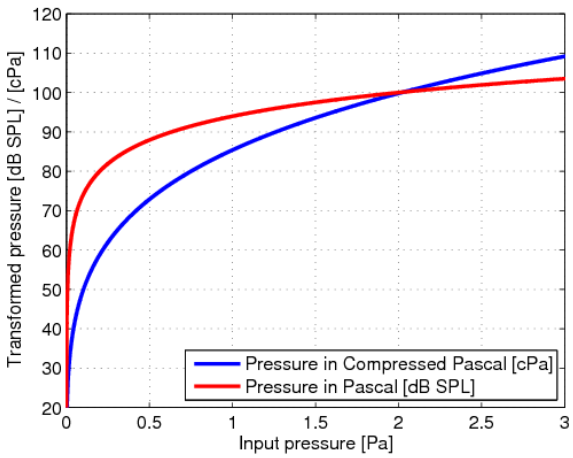


Figure 6.6: Non-linear transformation of sound pressure according to hearing model of Sottek

For the first branch of the Relative Approach, a regression versus time is applied for each frequency band (see figure 6.5a). Afterwards, a smoothing versus frequency is performed for each time slot (see figure 6.5b), which gives the intermediate result g_1 (see figure 6.4).

Its output is subtracted from the source signal which is transformed in the same way (result f_1, figure 6.4). Finally, all negative and non-relevant components (for human hearing) are set to zero by a threshold (experientially determined to 0,53 cPa). This variant focuses on the detection of tonal components.

The second variant first smoothes versus frequency within a time slot and then applies the regression versus time. This output signal is transformed non-linearly through the hearing model and leads to the intermediate result g_2. It is compared to the result f_2, which is determined as the output of the smoothing versus frequency only. Again all negative and non-relevant components are set to zero. Thus more transient structures are detected by this branch.

At this point, two representations are available. The final result of the Relative Approach is then a weighted sum of both. The positive weights λ_1 and λ_2 for each representation are chosen so that $\lambda_1 + \lambda_2 = 1$. In general, the factors λ_1 and λ_2 are used to describe the weighting of the Relative Approach for tonal and transient signals.

The result of the Relative Approach analysis is typically a 3D spectrogram displaying the deviation from the "close to the human expectation" between the estimated and the current signal. Because of the non-linear transformation, the physical unit of the magnitude remains cPa.

In order to adopt the Relative Approach for speech analysis, the following simplifications were applied:

- For our new objective speech quality models, $\lambda_1 = 0$ and $\lambda_2 = 1$ was chosen. Thus, the model is tuned to detect time-variant transient structures.
- To reduce computational complexity, a $1/12^{th}$ octave filter bank representation according to ANSI S1.1-1986 [i.30] is used as the input of the algorithm. The filterbank may compensate for the delays introduced by the filters, but it is not required. Compensation, or non-compensation, will have no significant effect on the Relative Approach results. (Filter delay is most significant at low frequencies which are not used for the analysis.).

Currently the Relative Approach uses a time resolution of Δt = 6,66 ms (150 blocks per second). Depending on the level calculation of the preceding filter bank, the resulting representation may provide more samples per second. Thus the output block size is achieved by applying a maximum filter over all samples within one frame.

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. A detailed explanation of the non-linear transformation used by the Relative Approach is given in annex K.

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 spectrogram 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 Recommendation ITU-T SG 12 [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 spectrogram 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 spectrogram is therefore calculated for the processed as well as for the unprocessed signal. Both spectrograms 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.

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

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

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 spectrogram indicating a "close to the original" transmission. A transmission leading to highly modulated background noises will result to an inhomogeneous differential spectrogram 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 4^{th} SQTE [i.11] for VoIP transmission evaluating "transparency" of background noise transmission influenced, e.g. by VAD or comfort noise.

6.5 Objective N-MOS

6.5.1 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 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).

6.5.2 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 spectrograms 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** spectrogram is calculated for the unprocessed signal *u(k)* and the processed signal *p(k)* ($Ra_u(t, f)$, $Ra_p(t, f)$). In these spectrograms 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 spectrograms 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 spectrogram is designated as $\Delta RA_{BGN,p-u}(t, f)$ in the following. In order classify these spectrograms with numerical values the **variance σ^2** for $Ra_p(t, f)$, $Ra_u(t, f)$ and $\Delta RA_{BGN,p-u}(t, f)$ and the **mean μ** for $Ra_p(t, f)$ and $\Delta RA_{BGN,p-u}(t, f)$ are calculated according to equations 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

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

with:
$$A_{ges} = (t_{max} - t_{min})(f_{max} - f_{min}),$$
$$dA(\Delta f_m) = \Delta t \cdot \Delta f_m,$$

Δt = 6,66 ms (= 1/150 s).

$\Delta f_m \neq$ 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 ($m\Delta RA_{BGN,p-u}$) and variance ($v\Delta RA_{BGN,p-u}$) are calculated for the $\Delta RA_{BGN,p-u}(t, f)$ spectrogram 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 is musical tones. These modulations lead to patterns in the Relative Approach spectrograms $RA_{BGN,p}(t, f)$ and $\Delta RA_{BGN,p-u}(t, f)$. 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 $\Delta RA_{BGN,p-u}(t, f)$ and also only for distinct frequencies. They indicate which parts of the signal have changed compared to the unprocessed signal.

A high mean of $\Delta RA_{BGN,p-u}(t, f)$ 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 $\Delta RA_{BGN,p-u}(t, f)$ alone are still not sufficient to predict the N-MOS reliable, because they are derived from a *differential* spectrogram. "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).

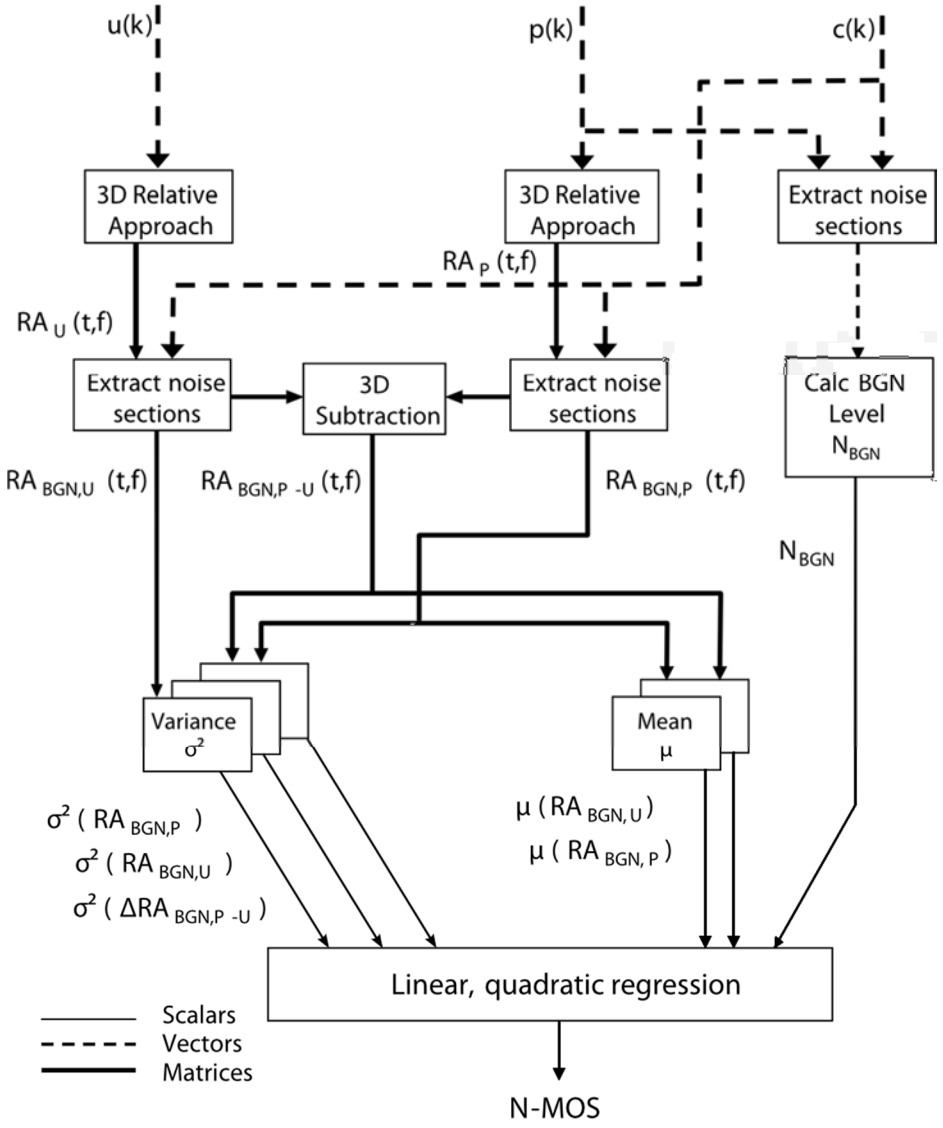


Figure 6.7: Block diagram of N-MOS calculation algorithm; *u(k)* unprocessed signal, *p(k)* processed signal, *c(k)* clean speech signal

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. Table 6.1 gives an overview about the extracted parameters.

Table 6.1: Extracted parameters for N-MOS

P_0	$N_{BGN, P}$	P_3	$\mu(RA_{BGN, U})$
P_1	$\mu(RA_{BGN, P})$	P_4	$\sigma^2(RA_{BGN, U})$
P_2	$\sigma^2(RA_{BGN, P})$	P_5	$\sigma^2(\Delta RA_{BGN, P-U})$

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 $vRA_{BGN,u}$ according to table 6.1.

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 spectrogram 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 ETSI EG 202 396-2 [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 6.1a.

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

Order	c_0	c_{BGN}	c_{j1}	c_{j2}	c_{j3}	c_{j4}	c_{j5}
1	1.8486	-0.0499	0.0094	0.2505	-0.1053	-0.9413	-0.9543
2	-	-	-0.0039	-0.0059	0.0037	0.6353	0.0098

6.5.3 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.

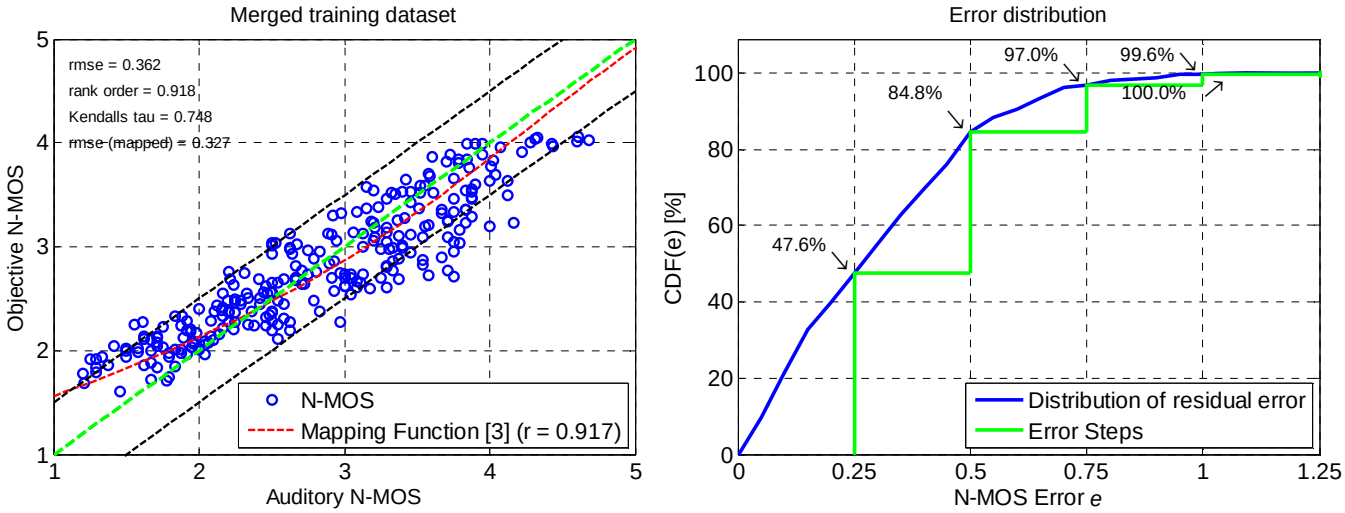


Figure 6.8: Left: Objectively calculated N-MOS versus auditory N-MOS; Right: CDF of residual error versus N-MOS Error e

All selected (French) training conditions according to clause 4 (independent of the network condition) and the new training database provided by Orange according to clause 6.3 were used for this mapping. All per-sample predictions belonging to one condition are averaged to a per-condition N-MOS which is used for comparison with the subjective per-condition N-MOS.

The left hand graph in figure 6.8 shows that the per sample deviation between the subjective and objective N-MOS is less than 0,5 MOS for nearly all (269) conditions. This results in an overall correlation of 91,7 %.

The right graph in figure 6.8 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 0,5 for 85 % of the conditions and lower than 0,75 for 97 % of all conditions.

6.6 Objective S-MOS

6.6.1 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.

6.6.2 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 6.9. 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 spectrograms** 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 spectrograms 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 spectrograms** 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 variance $v\Delta RA_{SP,p\text{-}c}$ is a measure for the amount of patterns in the differential spectrogram 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 $m\Delta RA_{SP,p-c}$ 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 $m\Delta RA_{SP,p-u}$ 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.

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).

The resulting values ΔSNR, $mRA_{SP,p}$, $v\Delta RA_{SP,p-u}$, $v\Delta RA_{SP,p-c}$, $m\Delta RA_{SP,p-u}$ and $m\Delta RA_{SP,p-c}$ are used as input parameters P_i for a feed forward neural network as described e.g. in [i.33]. Table 6.2 shows an overview over the extracted parameters.

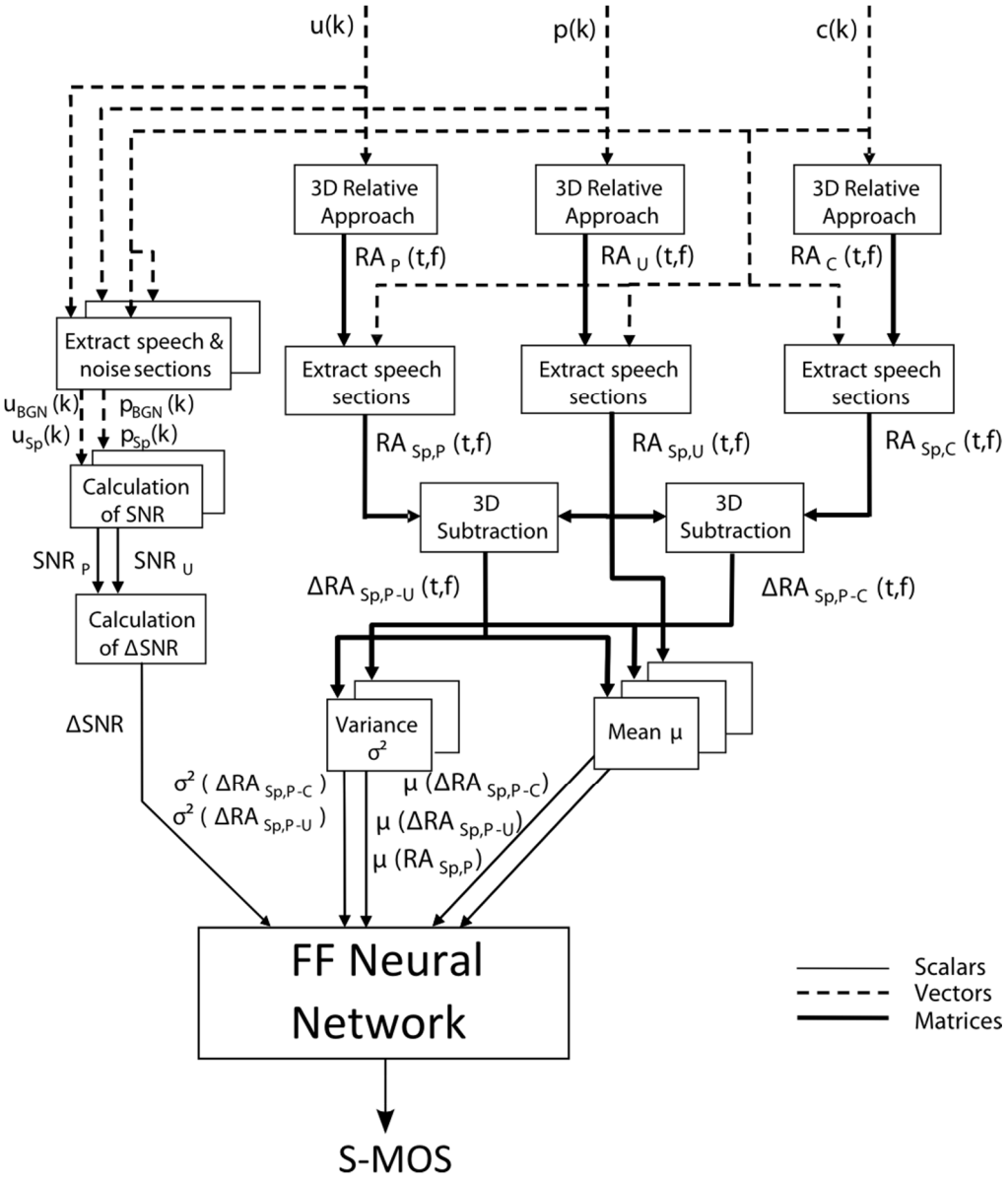


Figure 6.9: Block diagram of S-MOS calculation algorithm; *u(k)* unprocessed signal, *p(k)* processed signal, *c(k)* clean speech signal

Table 6.2: Extracted Parameters for S-MOS

P_1	ΔSNR	P_4	$\mu(\Delta RA_{Sp,\,P\text{-}C})$
P_2	$\mu(RA_{Sp,\,P})$	P_5	$\sigma^2(\Delta RA_{Sp,\,P\text{-}C})$
P_3	$\mu(\Delta RA_{Sp,\,P\text{-}U})$	P_6	$\sigma^2(\Delta RA_{Sp,\,P\text{-}U})$

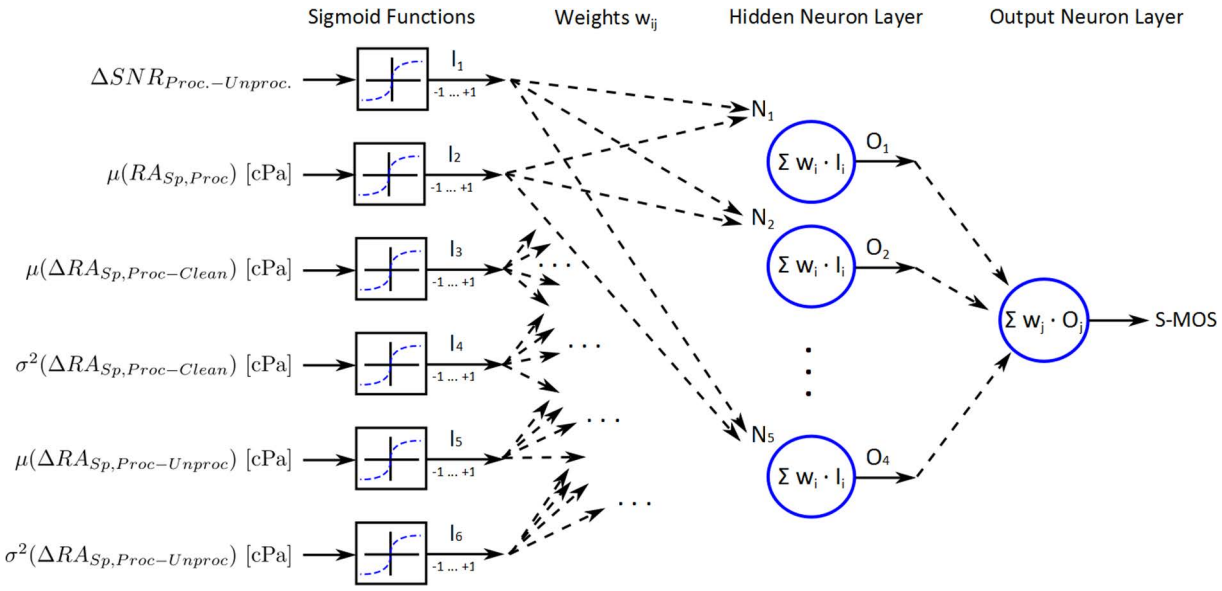


Figure 6.10: Structure of neural network for S-MOS

The setup of the neural network is shown in figure 6.10. It consists of 5 hidden layers; each layer N_j includes a connection from each transformed input parameter I_i. The output O_j of each layer is calculated as the weighted sum of each input I_i using the weights w_{ij}. The outputs O_j are then weighted by w_j and summed up to the output S-MOS. Both, w_{ij} and w_j are the result of the training of the network.

The parameters according to table 6.2 are composed to a vector **P** according to equation 6.10 including a bias as the first element.

$$P = \begin{pmatrix} 1 & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{pmatrix} \tag{6.10}$$

The output calculation of the neural network shown in figure 6.10 can be described as concatenated matrix operations as shown in equation 6.11.

$$\text{S-MOS}_{objective,raw} = f_{sigmoid}\left(f_{sigmoid}\left(\frac{P - M_{in}}{S_{in}}\right) \times H\right) \times O \tag{6.11}$$

First the parameter vector **P** is normalized to mean 0.0 and standard deviation 1.0. This is done by subtracting the average of all training data for each parameter from each item of the input parameter vector. The averages for each parameter P_i can be described as a vector, which is different for narrow- and wideband mode:

$$M_{in,WB} = \begin{pmatrix} 0 & 11.2059 & 3.5049 & -1.4115 & 0.90054 & 13.1402 & 13.2832 \end{pmatrix} \tag{6.12}$$

NOTE 1: The first element is set to zero to be compatible with the bias element in **P**.

A similar approach can be made for the input standard deviation S_{in} for each parameter P_i, also separated for wide and narrowband in equation :

$$S_{in,WB} = \begin{pmatrix} 1 & 10.5212 & 1.3348 & 1.1011 & 0.83575 & 5.4454 & 10.2952 \end{pmatrix} \tag{6.13}$$

NOTE 2: The first element is set to one to be compatible with the bias element in **P**.

After normalizing the input data, the sigmoid function $f_{sigmoid}(x)$ is applied to the each normalized parameter P_i.

This ensures that each input of each neuron of the hidden layer is soft-limited to the range ± 1,0 and guarantees that parameters out of the training range cannot produce an overflow which results in eventually unreasonable scores.

For the current model, the hyperbolic tangent was chosen to a sigmoid function:

$$f_{sigmoid}(x) = \tanh(x) \tag{6.14}$$

Thus the input of the hidden neuron layers can also be given as a transformed parameter vector $\tilde{P}$:

$$\widetilde{P} = f_{sigmoid}\left(\frac{P - M_{in}}{S_{in}}\right) = \begin{pmatrix} 1 & \widetilde{P_1} & \widetilde{P_2} & \widetilde{P_3} & \widetilde{P_4} & \widetilde{P_5} & \widetilde{P_6} \end{pmatrix} \tag{6.15}$$

NOTE 3: The sigmoid function is not applied to the bias component.

The output of the hidden layer is calculated with a matrix multiplication of $\tilde{P}$ and **H**. **H** describes all weights from each input parameter to each neuron in the hidden layer. These weights are the results of the training with the back-propagation algorithm. In consequence, **H** is different for each bandwidth mode:

$$H_{WB} = \begin{pmatrix} -0.39721 & -0.50013 & -0.15194 & 0.52774 & 1.946 \\ 0.69961 & 1.6117 & -0.15658 & -0.040337 & 5.7951 \\ 0.77363 & -1.1763 & -0.70999 & -0.44794 & -0.58914 \\ -1.1668 & 0.27301 & 1.1257 & 0.4015 & -0.8096 \\ -0.8113 & -1.4355 & -0.2341 & 1.5061 & 0.35826 \\ 1.2961 & 0.81908 & 0.28889 & -1.5259 & -25.0298 \\ -2.1736 & 1.0789 & -1.4558 & 2.457 & -21.4014 \end{pmatrix} \tag{6.16}$$

The outputs of the hidden layer are then again soft-limited with the same sigmoid function to assure a valid range (±1,0) for the output neuron layer. The five transformed output values of the hidden layer are then given to the output layer. Here the output of the neural network is calculated with another matrix multiplication with the matrix **O**, which weights the outputs of the hidden layers to an output score $SMOS_{objective,\,raw}$. This output layer matrix **O** is also given for wide and narrowband mode independently:

$$O_{WB} = \begin{pmatrix} -0.4454 & 0.31827 & -0.46555 & -0.46436 & 0.18345 \end{pmatrix} \tag{6.17}$$

Another part of the back-propagation algorithm is also to normalize the output data to mean 0,0 and standard deviation 1,0. To revise this step and transform the output of the neural network back to the MOS scale, the objective S-MOS is calculated from the raw score:

$$\text{S-MOS}_{objective} = \max(1.0, \min(S_{out} \cdot (\text{S-MOS}_{objective,raw} + M_{out}), 5.0)) \tag{6.18}$$

The objective S-MOS is finally calculated with M_{out} = 3.0 and S_{out} = 2.0 and a hard limiter [1.0; 5.0].

6.6.3 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 sentences of a condition and then mapped to the auditory S-MOS.

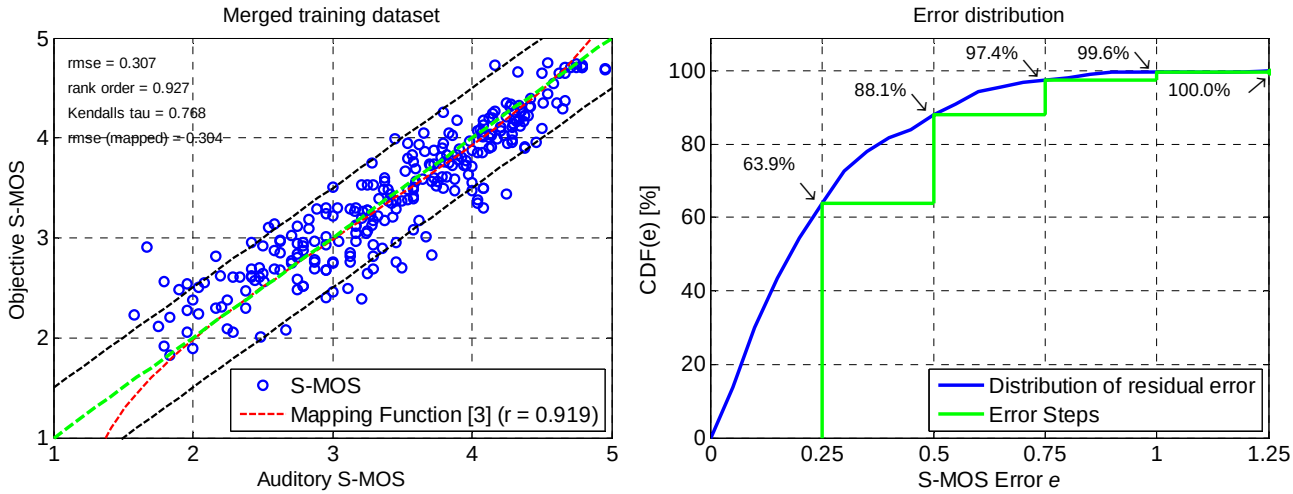


Figure 6.11: Left: Objectively calculated S-MOS versus auditory S-MOS; Right: CDF of residual error versus S-MOS Error e

Similar to the N-MOS training all training samples -were used. All per-sample predictions belonging to one condition are averaged to a per-condition S-MOS which is used for comparison with the subjective per-condition S-MOS.

The left hand graph in figure 6.11 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 (269) conditions. This results in an overall correlation of 91,9 %.

The right hand graph in figure 6.11 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 88,1 % of all conditions.

6.7 Objective G-MOS

6.7.1 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:

$$GMOS = c_0 + \sum_{j=1}^{2} c_{Sj} \cdot SMOS^{\,j} + \sum_{j=1}^{2} c_{Nj} \cdot NMOS^{\,j} \qquad (6.19)$$

where:

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

j is the regression order index.

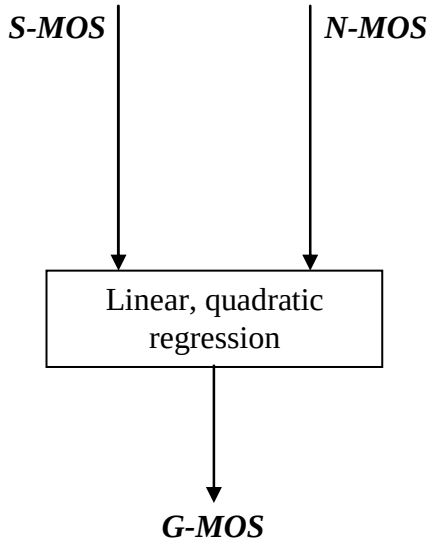


Figure 6.12: Block diagram of G-MOS calculation algorithm

Training and validation of the S-MOS regression were carried out using the regression coefficients in table 6.3.

Table 6.3: Coefficients for linear, quadratic G-MOS regression algorithm

Order	c_0	c_{Sj} (S-MOS)	c_{Nj} (N-MOS)
1	-1.1175	0.5805	0.6697
2	-	0.0217	-0.0262

6.7.2 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 6.13. All per-sample predictions belonging to one condition are averaged to a per-condition G-MOS which is used for comparison with the subjective per-condition G-MOS.

The left hand graph in figure 6.13 shows that the per sample deviation between objective and auditory G-MOS is less than 0,5 MOS for most of the (269) conditions. The overall correlation is determined to 95,1 %.

The cumulated density function *CDF (e)* versus the G-MOS Error e (see also equation 6.7) is shown on the right in figure 6.13. The CDF indicates that for 68 % 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.

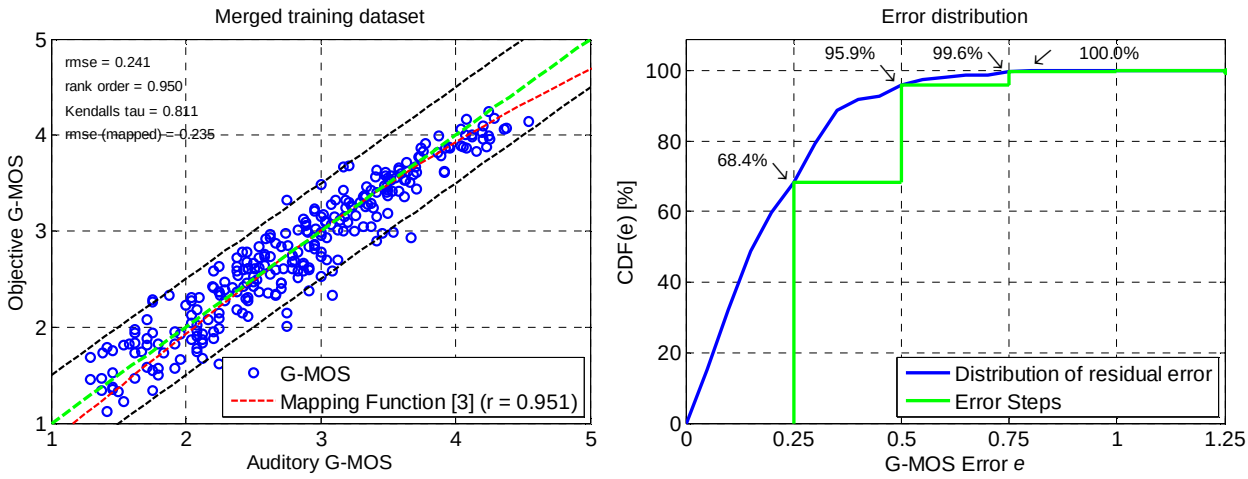


Figure 6.13: Left: Objectively calculated G-MOS versus auditory G-MOS; Right: CDF of residual error versus G-MOS Error e

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. These condition results are shown in annex F. Additionally, another subjective database provided by Orange with 18 conditions was provided to validate the test method (see clause 7.3).

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 6.5, 6.6 and 6.7 (N-MOS, S-MOS and G-MOS).

2) Comparison between obtained objective and the subjective results (see ETSI 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, the following statistical metrics will be used:

Root Mean Square Error (RMSE)

The root mean square error according to [i.24] 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

The Pearson Correlation coefficient according to [i.24] measures the linear relationship between the algorithm performance and the subjective data. This coefficient varies from -1,0 to 1,0; a value of 1,0 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 -1,0 shows that all data points lie on a single line but having opposite behaviour; a value of 0,0 shows that a linear model is inappropriate and that there is no linear relationship between the variables.

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

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

Spearman's Rank Correlation

The Spearman's Rank Correlation coefficient according to [i.24] is a non-parametric measure of correlation. It assesses how well an arbitrary monotonic function could describe the relationship between two variables. This parameter varies from -1,0 to +1,0, similar as the Pearson Correlation:

$$\rho = 1 - \frac{6 \cdot \sum \frac{d_i^2}{N}}{N(N^2 - 1)} \quad (7.4)$$

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

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

$$\tau = \frac{4 \sum \frac{q_i}{N}}{N(N-1)} - 1 \quad (7.5)$$

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

The residual error distribution according to [i.24] evaluates the consistency of the model using the Cumulative Density Function (CDF) applied to the Error e:

$$e = |MOS_{auditory} - MOS_{objective}| \quad (7.6)$$

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

NOTE: The prediction results for training and validation reported in previous versions of the present document (up to V1.3.1) reported somewhat better performance of the model than in this current one. This apparently lower prediction accuracy is caused by several reasons:

- In previous versions, it was assumed that the extracted parameters for regression and neural network according to clauses 6.5.2 and 6.6.2 are averaged for one condition and then mapped against subjective data. The average subjective condition MOS values are much more reliable than the average sample MOS values.
- In the new version, training is applied on a per-sample base. Due to the low amount of votes per sample (only 4), the training may obtain lower precision. On the other hand, with the new approach it is now possible to obtain also per-sample scores (with eventually lower significance), which was not recommended before.
- The new training set consists of two similar, but in general different databases. Even though the merged datasets seem to fit together, it may also decrease the prediction performance on the single databases. But with this additional data, which introduces real device recordings, the neural network gets more robust against new, unknown data.

- In almost all presented prediction results, it seems like that a mapping function could additionally improve the performance. Especially for the training conditions, it is expected that due to least-mean-square-mapping (regression and neural network use these algorithms), no offset or shift is present. This assumption is only valid for the per-sample training. After averaging, this principle is not necessarily observable.

7.2 ETSI EG 202 396-2 Database Results Analysis

7.2.1 Comparing subjective and objective N-MOS results

Only validation recordings from the original ETSI EG 202 396-2 [i.2] were used for the following results.

Figures 7.1 and 7.2 show that the per sample deviation between the subjective and the objective N-MOS is less than 0,5 MOS for most conditions. This results in an overall Pearson correlation of 92,3 %. The Spearman Correlation Coefficient is 0,931 and the Kendall Tau is 0,782, both of them are near to 1.

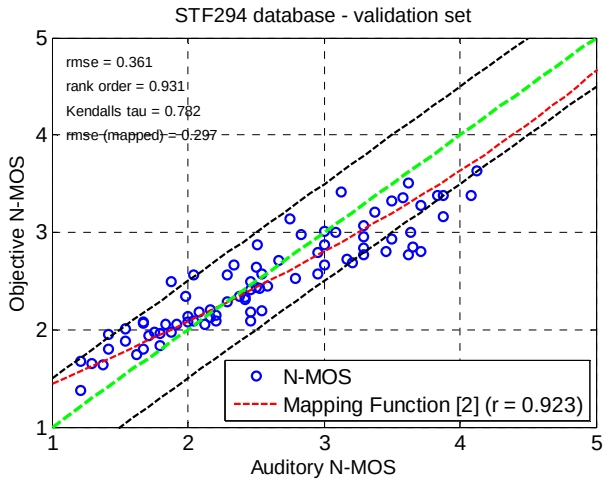


Figure 7.1: Left: Objectively calculated N-MOS vs. auditory N-MOS for validation conditions

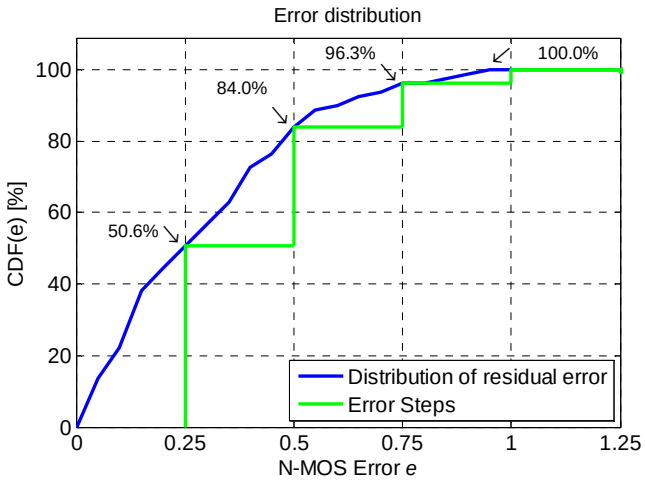


Figure 7.2: Objectively CDF of residual error vs. N-MOS Error e for validation conditions

For this situation, the RMSE value is 0,36 and the distribution of the residual error is shown in figure 7.2 where the N-MOS Error e is lower than 0,25 for approximately 50 % of the conditions and lower than 0,75 for 96 % for all conditions.

7.2.2 Comparing subjective and objective S-MOS results

Only validation recordings from the original ETSI EG 202 396-2 [i.2] were used for the following results.

Figures 7.3 and 7.4 show that the per sample deviation between the subjective and the objective S-MOS is less than 0,5 MOS for a large amount of conditions. This results in an overall correlation of 89,3 %. The Spearman Correlation Coefficient is 0,880 and the Kendall Tau is 0,716, both of them are near to 1.

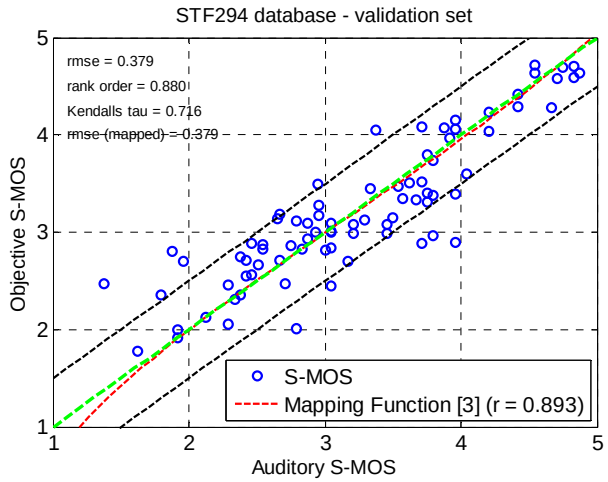


Figure 7.3: Left: Objectively calculated S-MOS vs. auditory S-MOS for validation conditions

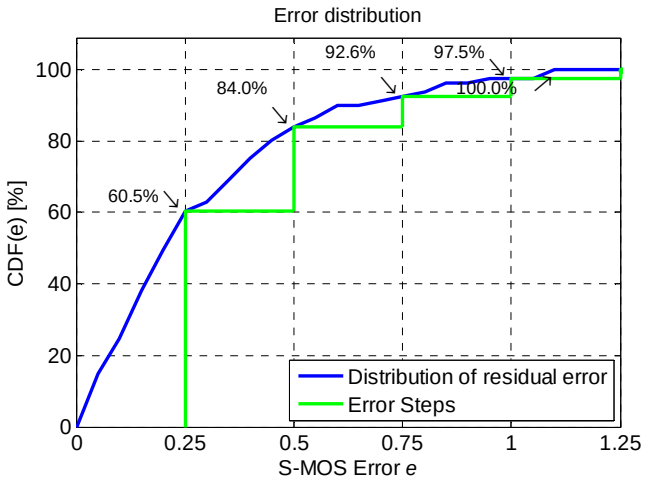


Figure 7.4: Objectively CDF of residual error vs. S-MOS Error e for validation conditions

For this situation, the RMSE value is 0,37 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 60 % of the conditions and lower than 0,75 for 93 % for all conditions.

7.2.3 Comparing Subjective and Objective G-MOS Results

Only validation recordings from the original ETSI EG 202 396-2 [i.2] were used for the following results.

Figures 7.5 and 7.6 show that the per sample deviation between the subjective and the objective G-MOS is less than 0,5 MOS for nearly all conditions. This results in an overall correlation of 91,2%. The Spearman Correlation Coefficient is 0,892 and the Kendall Tau is 0,735, both of them are near to 1.

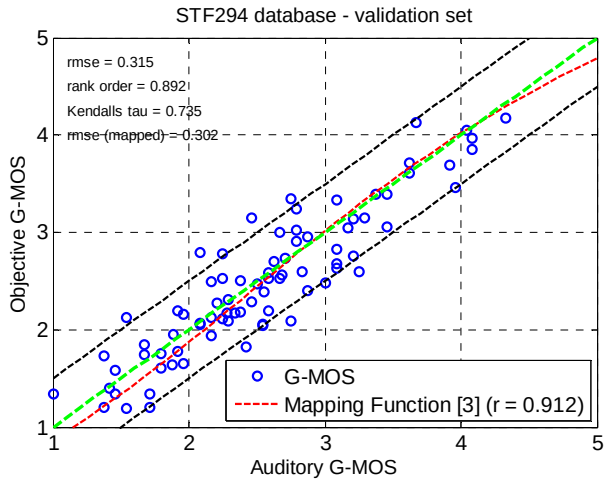


Figure 7.5: Left: Objectively calculated G-MOS vs. auditory G-MOS for validation conditions

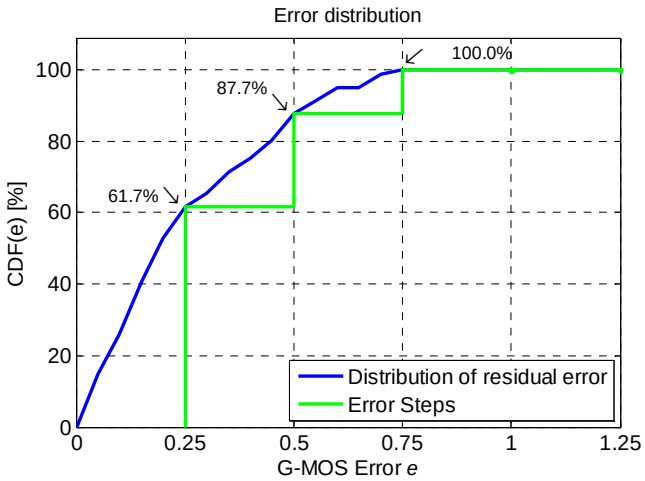


Figure 7.6: Objectively CDF of residual error vs. G-MOS Error e for validation conditions

For this situation, the RMSE value is 0,31 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 62 % of the conditions and lower than 0,75 for 100 % for all conditions.

7.3 Orange Validation Database results Analysed

7.3.0 Introduction

In addition to the new training database described in clause 6.3, another validation database from Orange was provided. This database was used for the validation process of ETSI TS 103 106 [i.32] and also included the same speech material as the Orange training database. The database consists of 18 conditions with 8 sentences each (2 male / 2 female talkers). It is used for a full external validation in terms of validating conditions which were not part of the same listening test as the training set.

Since this additional validation database is completely unknown, an eventual mapping function can be applied in order to improve the prediction performance. A third order mapping function is always printed within the result plots, but is not applied to the results, also for the error distribution plots. Only the RMSE after mapping is reported for informational purposes.

7.3.1 Comparing subjective and objective N-MOS results

Figures 7.7 and 7.8 show that the per sample deviation between the subjective and the objective N-MOS is less than 0,5 MOS for almost all conditions. This results in an overall Pearson correlation of 97,4 %. The Spearman Correlation Coefficient is 0,937 and the Kendall Tau is 0,818, both of them are near to 1.

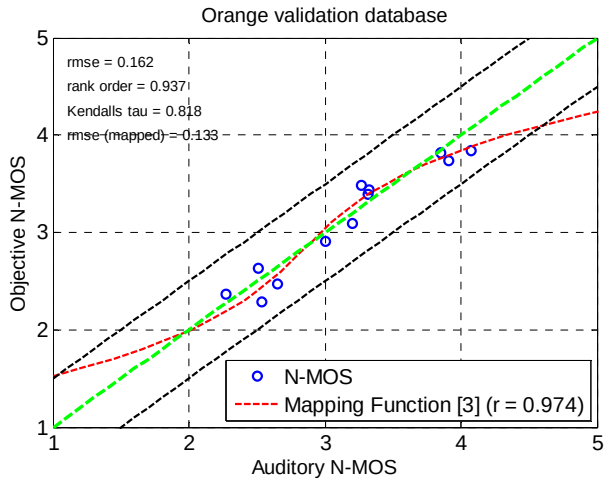


Figure 7.7: Left: Objectively calculated N-MOS vs. auditory N-MOS for validation conditions

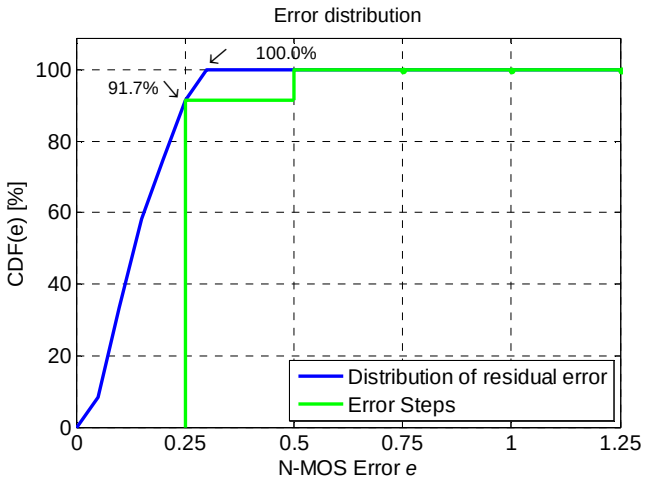


Figure 7.8: Objectively CDF of residual error vs. N-MOS Error e for validation conditions

For this situation, the RMSE value is 0,16 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 92 % of the conditions and lower than 0,5 for 100 % for all conditions. An additional mapping can be applied (RMSE after mapping is 0,13), but does not change the overall prediction behaviour.

7.3.2 Comparing subjective and objective S-MOS results

Figures 7.9 and 7.10 show that the per sample deviation between the subjective and the objective S-MOS is less than 0,5 MOS for a large amount of conditions. This results in an overall correlation of 70,7 %. The Spearman Correlation Coefficient is 0,818 and the Kendall Tau is 0,606.

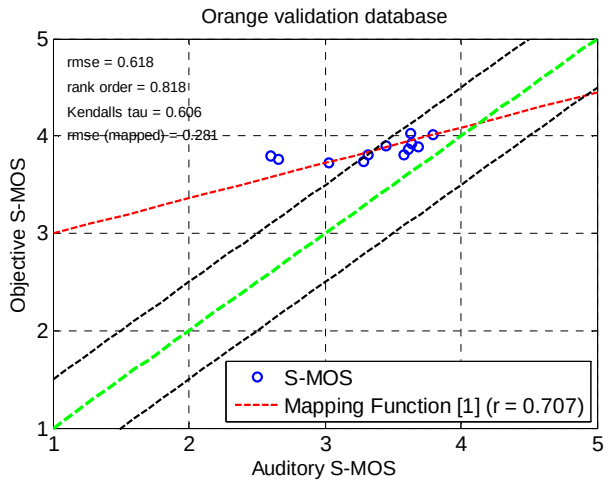


Figure 7.9: Left: Objectively calculated S-MOS vs. auditory S-MOS for validation conditions

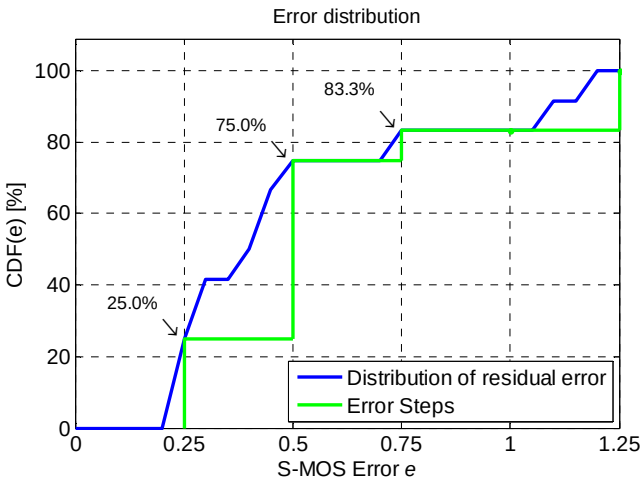


Figure 7.10: Objectively CDF of residual error vs. S-MOS Error e for validation conditions

For this situation, the RMSE value is 0,62 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 25 % of the conditions and lower than 0,75 for 83 % for all conditions.

An additional mapping would dramatically increase the prediction performance; the RMSE would decrease to 0,28.

7.3.3 Comparing Subjective and Objective G-MOS Results

Figures 7.11 and 7.12 show that the per sample deviation between the subjective and the objective G-MOS is less than 0,5 MOS for nearly all conditions. This results in an overall correlation of 81,8%. The Spearman Correlation Coefficient is 0,811 and the Kendall Tau is 0,667.

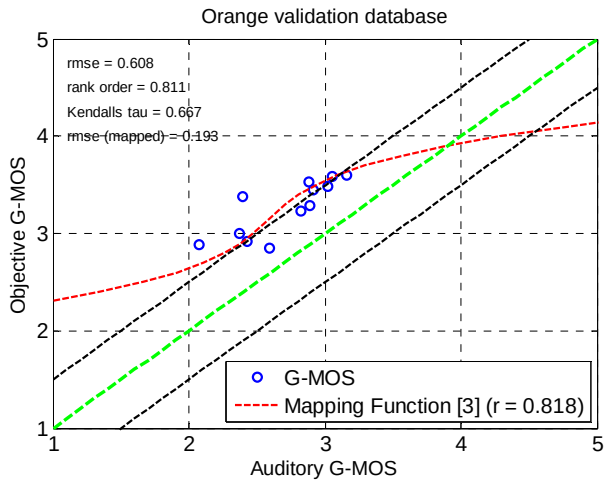


Figure 7.11: Left: Objectively calculated G-MOS vs. auditory G-MOS for validation conditions

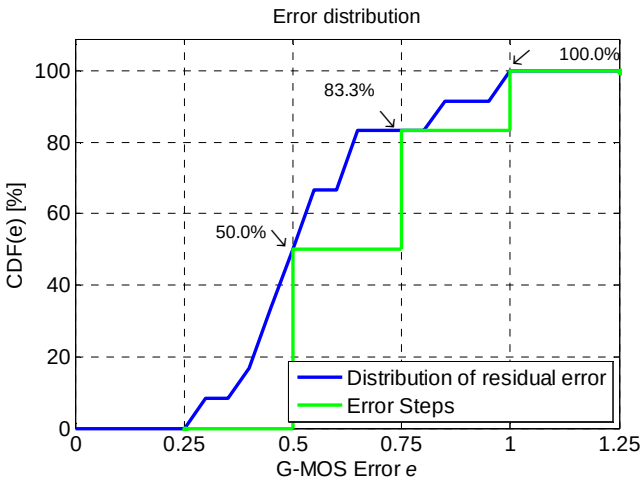


Figure 7.12: Objectively CDF of residual error vs. G-MOS Error e for validation conditions

For this situation, the RMSE value is 0,61 and the distribution of the residual error is shown in figure 7.12 where the G-MOS Error e is lower than 0,5 for approximately 50 % of the conditions and lower than 1 for 100 % for all conditions.

Again, an additional mapping would dramatically increase the G-MOS prediction performance; the RMSE would decrease to 0,19. This is mainly the influence of the also overrated S-MOS, since G-MOS is just a combination of N- and S-MOS.

8 Objective Model for Narrowband Applications

8.0 Introduction

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. These modifications 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 Recommendation ITU-T 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 a modified IRS filter according to Recommendation ITU-T 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 figure 8.1.

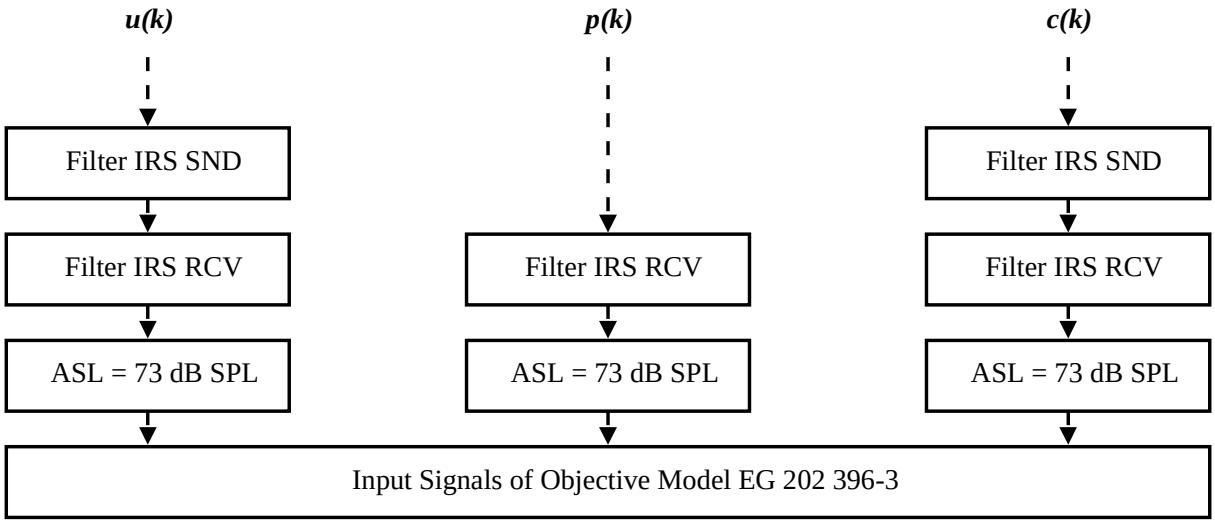


Figure 8.1: Modified pre-processing for narrowband mode

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 new coefficients for 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}	c_{j1}	c_{j2}	c_{j3}	c_{j4}	c_{j5}
1	2.1778	-0.0673	0.2517	0.2157	-0.1066	-2.9044	-1.4480
2	-	-	-0.0009	0.0179	-0.0071	0.6378	-0.1753

Table 8.3: Coefficients for linear, quadratic G-MOS regression algorithm

Order	c_0	c_{Sj} (S-MOS)	c_{Nj} (N-MOS)
1	-0.6298	0.5070	0.5443
2	-	0.0335	-0.0176

The new values (vectors M_{in}, S_{in}, O and matrix **H**) according to clause 6.6 for the neural network configuration to calculate objective S-MOS are given in the following equations.

$$M_{in,NB} = \begin{pmatrix} 0 & 6.5615 & 1.7518 & -0.34849 & 0.080803 & 4.8439 & 2.7659 \end{pmatrix} \quad (8.1)$$

$$S_{in,NB} = \begin{pmatrix} 1 & 8.2533 & 0.27953 & 0.22865 & 0.18403 & 2.1831 & 1.232 \end{pmatrix} \quad (8.2)$$

$$H_{NB} = \begin{pmatrix} -0.19712 & 0.16831 & 1.2911 & 0.25815 & 0.61799 \\ -1.6076 & -0.90138 & -0.15011 & 0.43588 & 0.59045 \\ -0.12558 & -0.33731 & 0.8453 & -0.37592 & -0.2913 \\ 0.81989 & 1.7359 & -0.29084 & -0.74025 & 0.084253 \\ -0.75444 & 1.1972 & 2.0637 & 0.97744 & 0.41328 \\ 1.23 & 1.0684 & -0.77656 & -0.33681 & -2.0019 \\ -3.0518 & 0.090804 & -2.0868 & 1.2275 & -1.227 \end{pmatrix} \quad (8.3)$$

$$O_{NB} = \begin{pmatrix} -0.35713 & -0.20793 & -0.22151 & -0.30572 & 0.26762 \end{pmatrix} \quad (8.4)$$

8.3 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 (see ETSI EG 202 396-2 [i.2]).

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 scatter plots below (see figure 8.2).

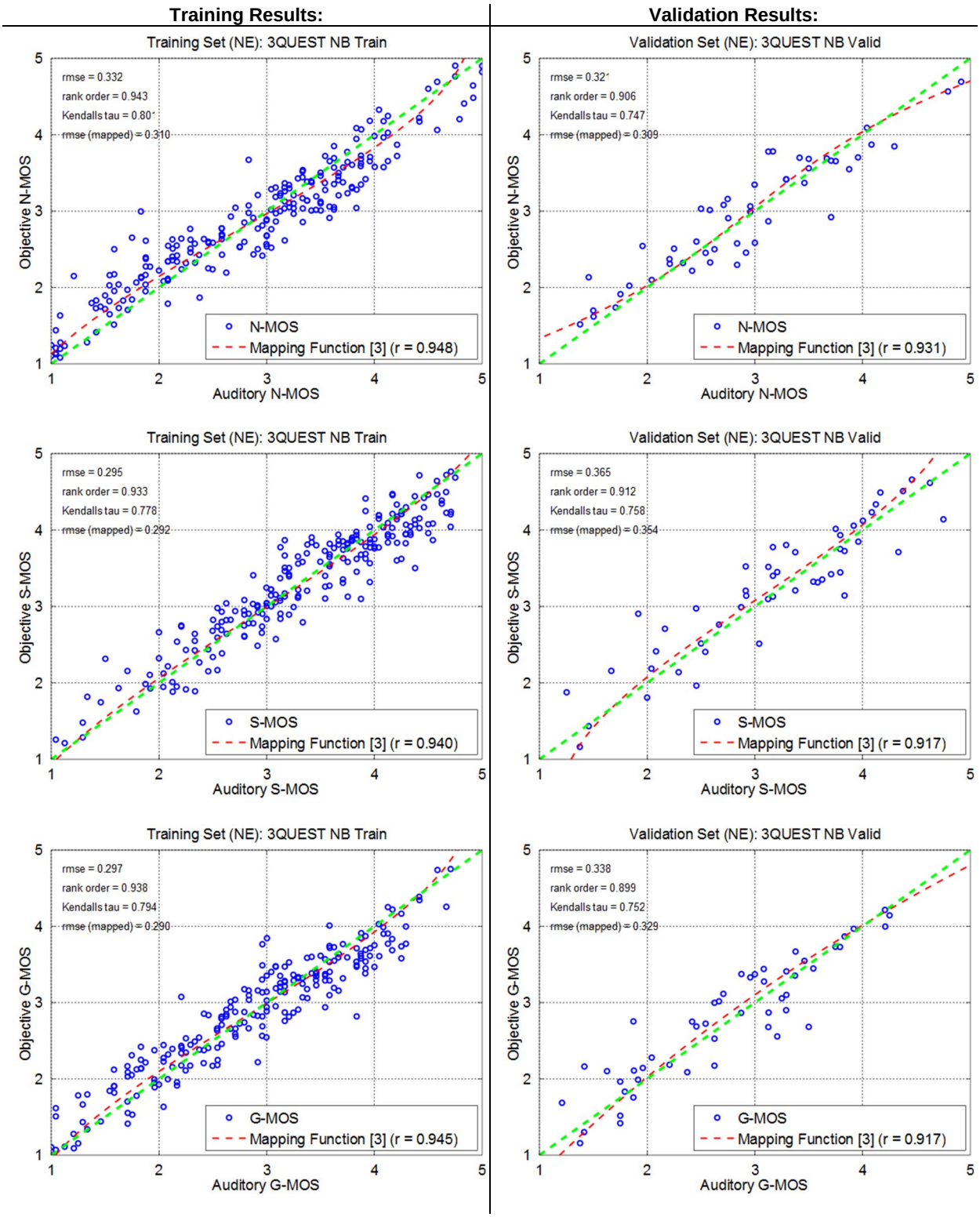


Figure 8.2: HEAD acoustics NB Database - Comparison subjective versus Objective data

Annex A:
Detailed post evaluation of listening test results

Table A.1 contains the conditions and related auditory S-MOS, N-MOS and G-MOS for French. 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

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

Although it is known that neither PESQ (Recommendation ITU-T 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 ETSI EG 202 396-2 [i.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 ETSI EG 202 396-2 [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 1: The sample length is less than 3,6 seconds for all samples listed above. Both algorithms, TOSQA2001 [i.19] and PESQ [i.18], require a sample length of 8 seconds to 32 seconds.

NOTE 2: 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 in ETSI EG 202 396-2 [i.2] 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 tables 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				

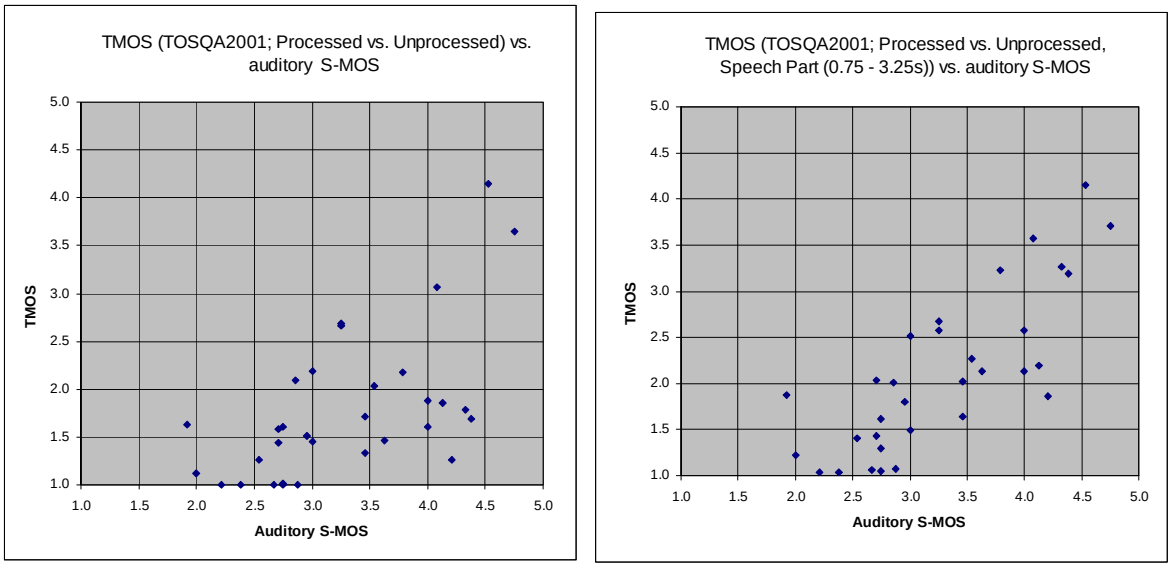


Figure B.1: TOSQA2001 results (TMOS) of processed data versus auditory S-MOS (unprocessed signal used as TOSQA2001 reference)

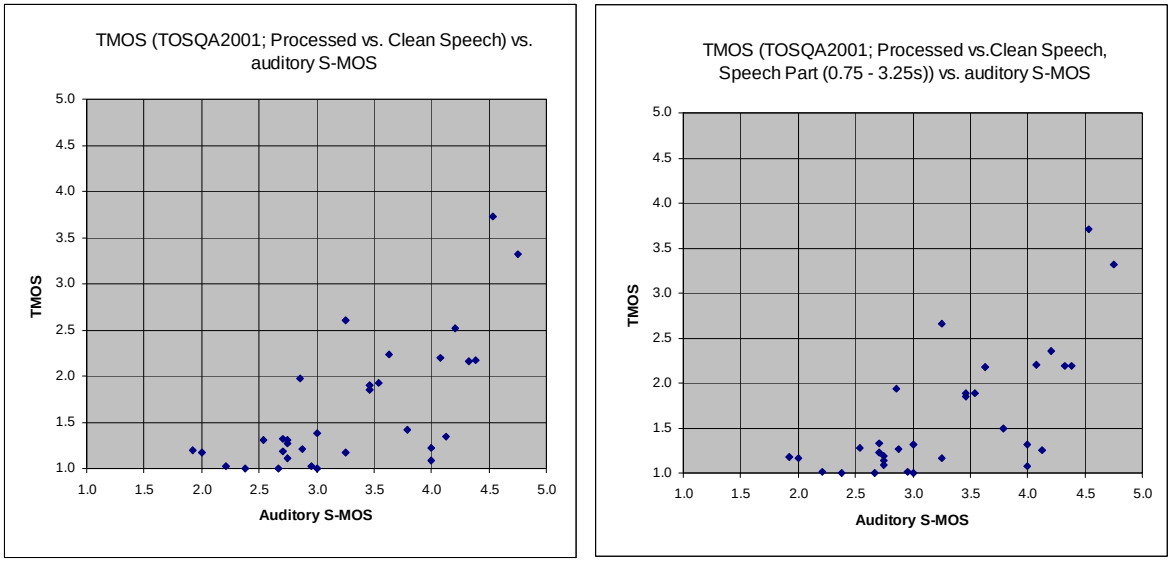


Figure B.2: TOSQA2001 results (TMOS) of processed data versus auditory S-MOS (clean speech signal used as TOSQA2001 reference)

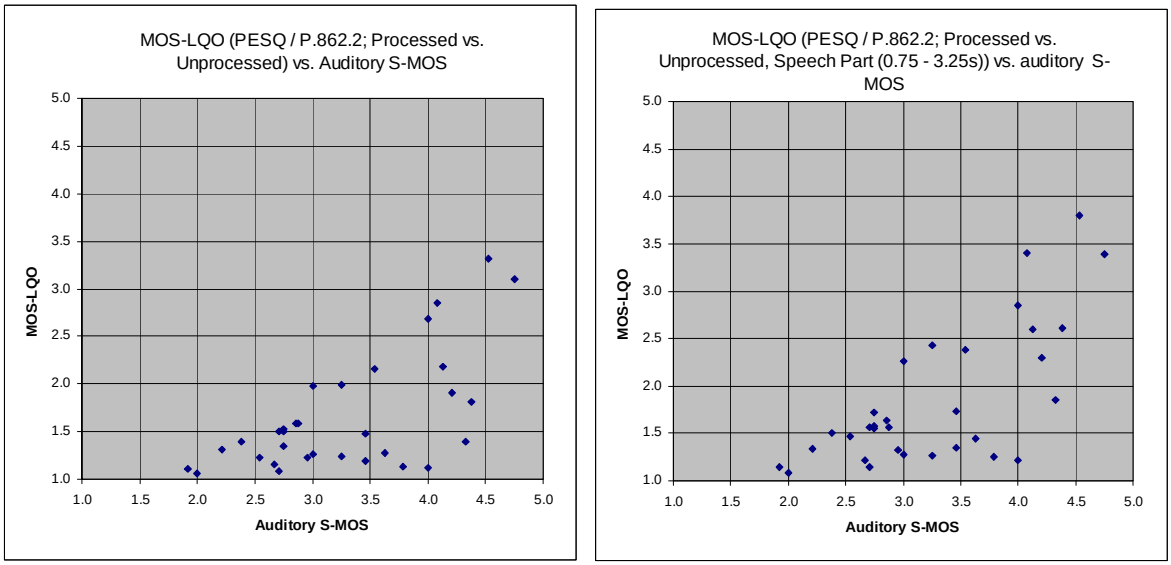


Figure B.3: PESQ (MOS-LQO, P.862.2) results of processed data versus auditory S-MOS (unprocessed signal used as PESQ reference)

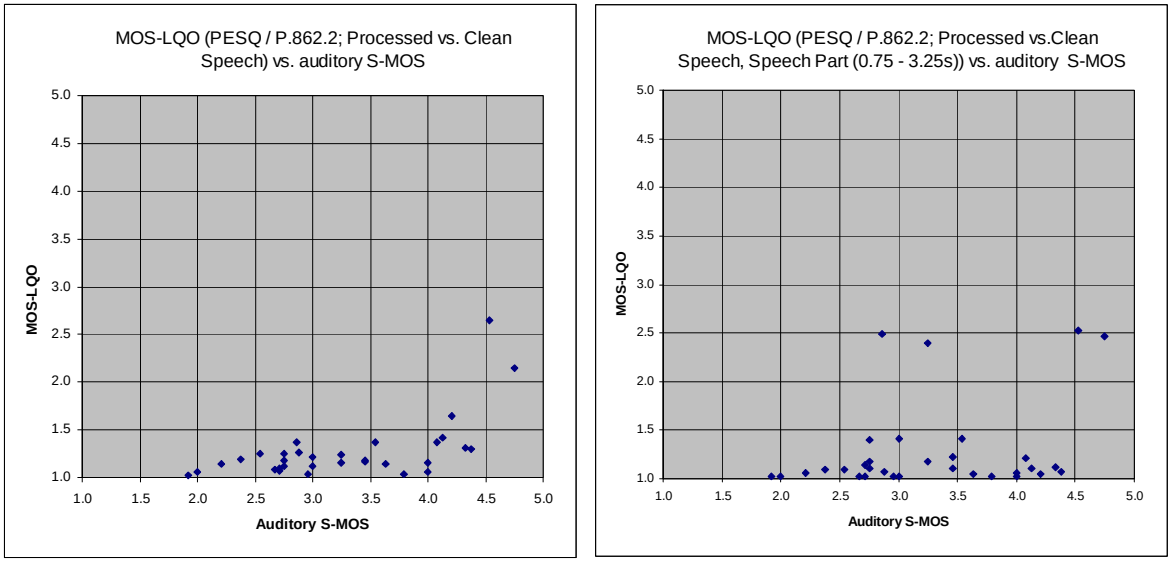


Figure B.4: PESQ results (MOS-LQO, P.862.2) of processed data versus auditory S-MOS (clean speech signal used as PESQ reference)

Annex C:
Comparison of objective MOS versus auditory MOS for the complete STF 294 database

This annex shows the correlation plots between the objective and the auditory S-/N-/G-MOS for all French data used during the training of the new method. Note that the MOS scores for ***all*** conditions of the STF 294 database (see ETSI EG 202 396-2 [i.2]) are compared to the listening test results, including also rejected samples as described in clause 5.

Figures C.1, C.3 and C.5 show the results for the French training data and figures C.2, C.4 and C.6 or the French validation 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 "+" (red) and the accepted (acc.) by a "○" (blue). 123 training and 49 validation conditions were rejected for the development of the model.

For the training data, the correlation for the objective N-MOS decreases to 89,6 %. This lower performance could be compensated with a mapping function. The correlation of the objective N-MOS to the auditory N-MOS for the validation data even increases slightly to 93,0 %.

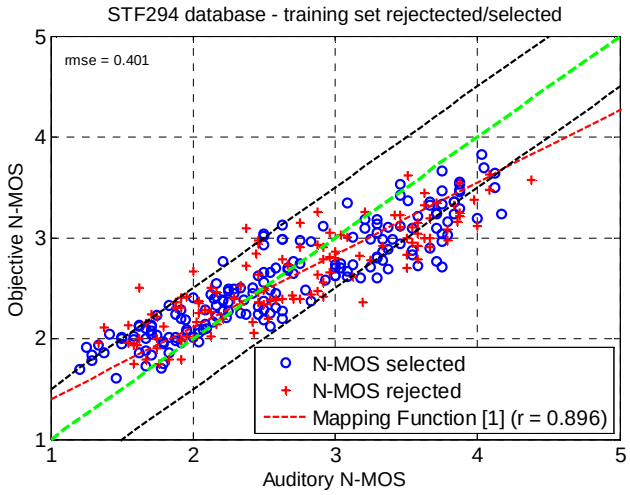


Figure C.1: Objective vs. auditory N-MOS for all French training conditions

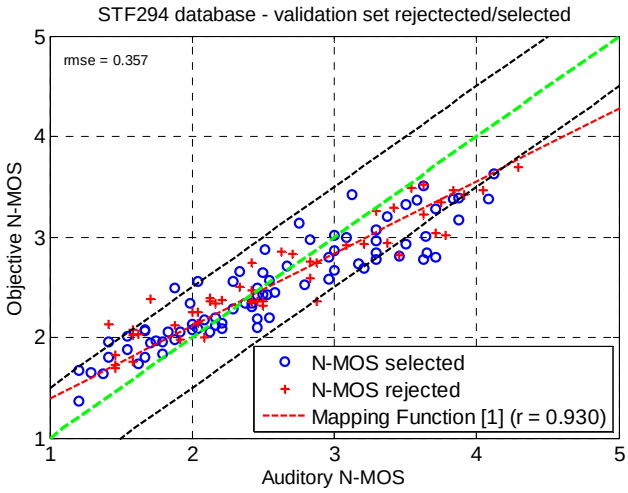


Figure C.2: Objective vs. auditory N-MOS for all French validation conditions

The correlation of the objective to the auditory S-MOS data decreases to 89,6 for the training data and but increases to 93,0 % for the validation data.

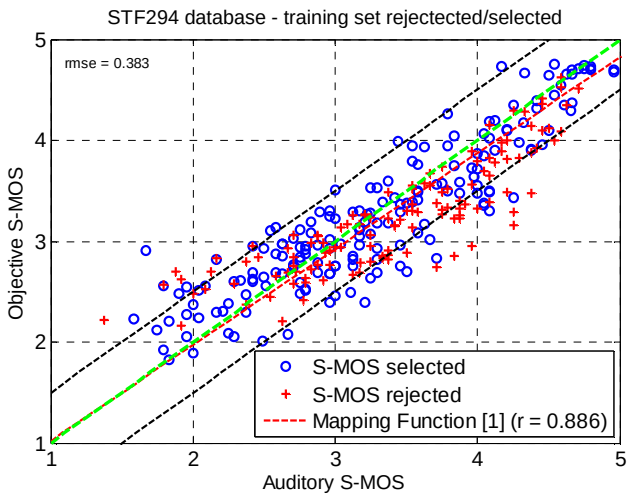


Figure C.3: Objective vs. auditory S-MOS for all French training conditions

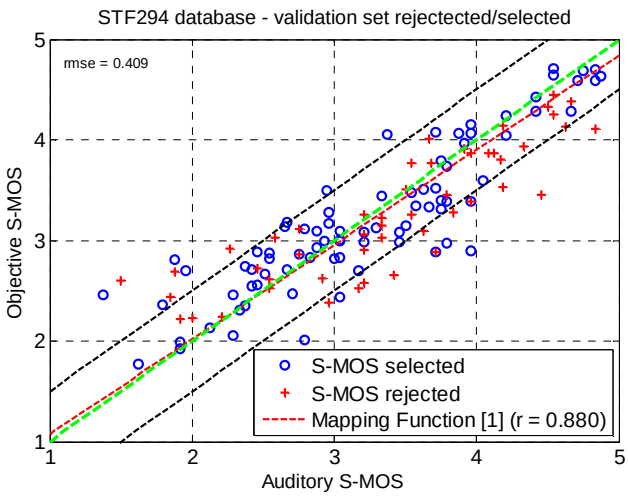


Figure C.4: Objective vs. auditory S-MOS for all French validation conditions

In consequence, the correlation between the objective and auditory G-MOS decreases also only slightly decreases to 92,6 % for the training data. Again for the validation data the correlation increases to 92,2 % as shown in figure C.6.

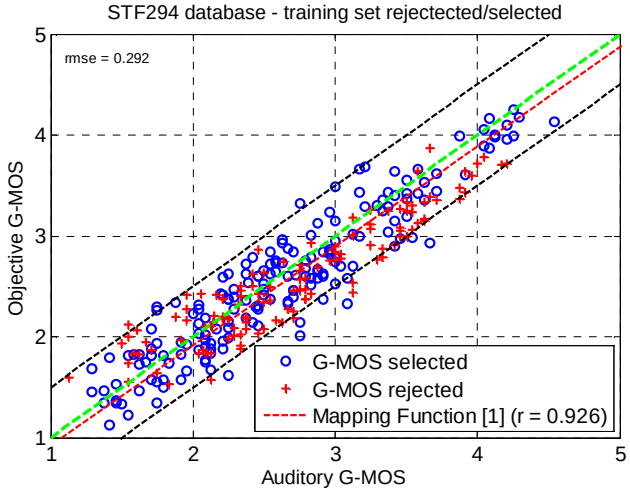


Figure C.5: Objective vs. auditory G-MOS for all French training conditions

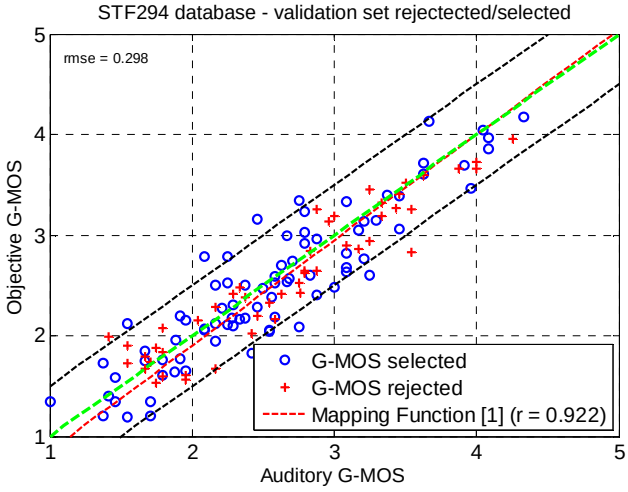


Figure C.6: Objective vs. auditory G-MOS for all French validation conditions

Generally it can be concluded that the selection process on the French data made the training set more consistent. Training prediction results on the accepted samples showed to be more accurate. On the other hand, including the rejected validation data would have even increased the prediction performance.

Annex D:
Comparison of objective MOS versus auditory MOS for rejected conditions

For information purpose, figures 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 training and validation set (see also annex C). Again the data not used for the model development (123 training, 49 validation conditions) are indicated by a "+" in the scatter plots.

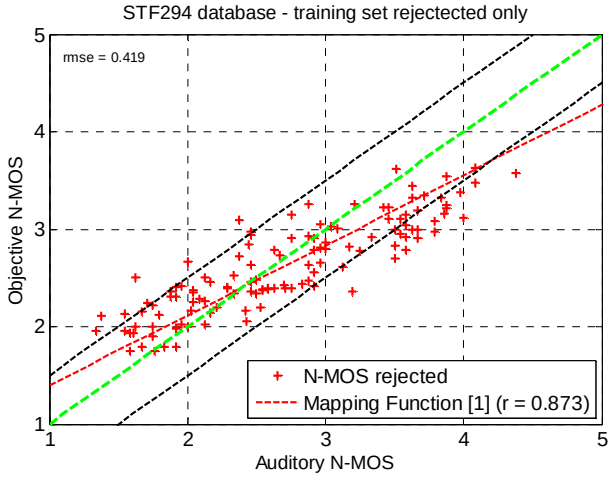


Figure D.1: Objective vs. auditory N-MOS only for training data not used for the model development

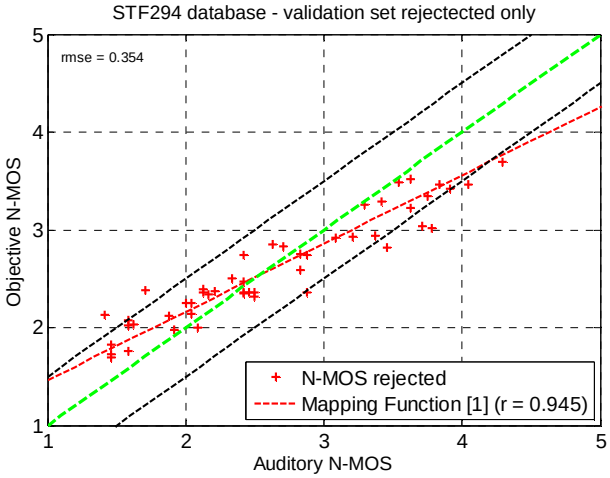


Figure D.2: Objective vs. auditory N-MOS only for validation data not used for the model development

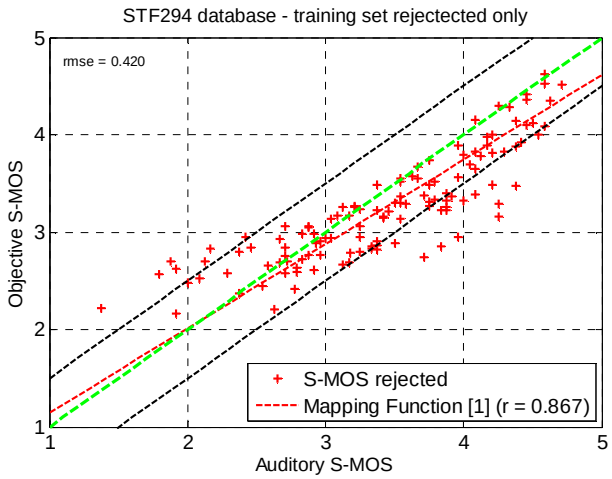


Figure D.3: Objective vs. auditory S-MOS only for training data not used for the model development

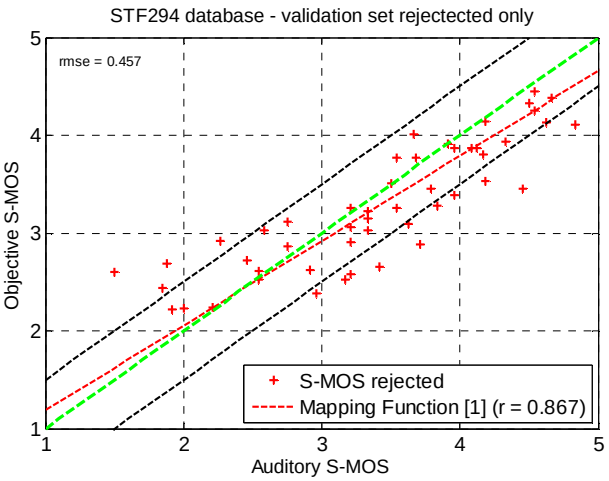


Figure D.4: Objective vs. auditory S-MOS only for validation data not used for the model development

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

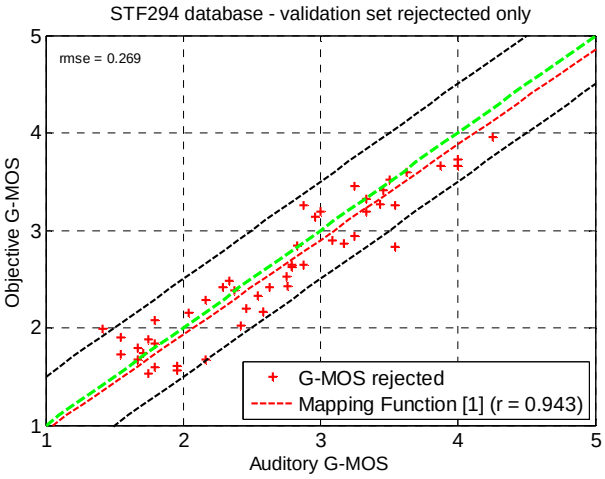


Figure D.6: Objective vs. auditory G-MOS only for validation data not used for the model development

Annex E:
Void

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

Table F.1 contains the conditions and related auditory and objective S-MOS, N-MOS and G-MOS for French language. 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	VAD	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	4,29	2,85	3,46
4	4	Lux_Car	hs	f	AMR _NI	no	Smooth	9	4,88	3,63	3,92	0,45	0,71	0,65	4,63	2,77	3,69
6	6	Lux_Car	hs	f	AMR _NIII	no	Smooth	9	3,50	3,21	3,08	1,07	0,72	0,88	3,34	2,69	2,68
9	9	Lux_Car	hs	f	AMR _NIII	no	Smooth	18	3,46	3,31	3,08	1,18	0,62	0,93	3,08	2,96	2,63
10	10	Lux_Car	hs	f	AMR _NI	no	Sharp	9	4,54	3,46	3,63	0,59	0,66	0,71	4,64	2,80	3,72
22	22	Lux_Car	hs	f	AMR _NI	yes	Sharp	9	4,42	3,50	3,63	0,58	0,72	0,71	4,42	2,93	3,61
24	24	Lux_Car	hs	f	AMR _NIII	yes	Sharp	9	3,79	3,29	3,21	0,83	0,69	0,78	3,38	2,77	2,76
31	31	Lux_Car	hf	f	AMR _NI	no	Smooth	9	3,79	2,21	2,75	1,02	0,88	0,74	2,97	2,09	2,09
34	34	Lux_Car	hf	f	AMR _NI	no	Smooth	18	3,04	2,29	2,42	1,00	0,75	0,78	2,44	2,28	1,82
37	37	Lux_Car	hf	f	AMR _NI	no	Sharp	9	3,21	2,04	2,29	0,98	0,69	0,55	2,99	2,09	2,09
39	39	Lux_Car	hf	f	AMR _NIII	no	Sharp	9	2,71	1,71	1,96	1,04	0,69	0,62	2,46	1,94	1,65
42	42	Lux_Car	hf	f	AMR _NIII	no	Sharp	18	1,92	1,92	1,71	0,88	0,72	0,62	1,92	2,05	1,34
43	43	Lux_Car	hf	f	AMR _NI	yes	Smooth	9	3,96	2,00	2,54	0,91	0,83	0,78	2,90	2,13	2,06
48	48	Lux_Car	hf	f	AMR _NIII	yes	Smooth	18	2,33	1,83	1,79	0,76	0,70	0,66	2,31	2,06	1,61
49	49	Lux_Car	hf	f	AMR _NI	yes	Sharp	9	3,71	2,17	2,54	0,86	0,76	0,66	2,88	2,12	2,04
52	52	Lux_Car	hf	f	AMR _NI	yes	Sharp	18	2,67	2,04	1,96	0,92	0,62	0,69	2,71	2,56	2,16
63	63	Lux_Car	hs	m	AMR _NIII	no	Smooth	18	3,21	3,29	2,67	1,02	0,81	0,81	3,08	2,84	2,57
67	67	Lux_Car	hs	m	AMR _NI	no	Sharp	18	3,71	3,60	3,44	1,08	0,71	0,65	4,08	3,00	3,39
78	78	Lux_Car	hs	m	AMR _NIII	yes	Sharp	9	2,96	2,79	2,50	1,00	0,66	0,78	3,17	2,52	2,47
79	79	Lux_Car	hs	m	AMR _NI	yes	Sharp	18	3,96	3,58	3,46	0,75	0,78	0,59	3,39	3,36	3,06
81	81	Lux_Car	hs	m	AMR _NIII	yes	Sharp	18	3,04	3,29	2,63	1,00	0,81	0,88	3,09	3,07	2,70
90	90	Lux_Car	hf	m	AMR _NIII	no	Smooth	18	2,38	2,13	1,88	1,17	1,03	0,90	2,35	2,05	1,64
94	94	Lux_Car	hf	m	AMR _NI	no	Sharp	18	2,88	2,42	2,38	0,90	0,88	0,88	2,93	2,31	2,18
99	99	Lux_Car	hf	m	AMR _NIII	yes	Smooth	9	3,04	2,46	2,25	0,86	0,72	0,74	3,00	2,09	2,11
102	102	Lux_Car	hf	m	AMR _NIII	yes	Smooth	18	2,54	2,21	2,08	1,18	0,93	0,97	2,88	2,15	2,05
103	103	Lux_Car	hf	m	AMR _NI	yes	Sharp	9	3,63	2,46	2,58	0,71	0,88	0,78	3,51	2,19	2,53
114	114	Crossroads	hs	f	AMR _NIII	no	Smooth	9	3,04	2,53	2,31	1,04	0,88	0,80	2,83	2,42	2,17
132	132	Crossroads	hs	f	AMR _NIII	yes	Sharp	9	3,46	2,96	3,00	0,83	0,75	0,72	2,99	2,79	2,48
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	3,14	2,07	2,19

Conditions	Id real	Noise	Recording	Speaker	Network	VAD	Smooth	dB	FRENCH								
									Subjective						Objective		
									MOS			Standard deviation			MOS		
									Speech	Noise	Global	Speech	Noise	Global	Speech	Noise	Global
141	141	Crossroads	hf	f	AMR _NIII	no	Smooth	9	2,79	1,29	1,54	1,28	0,46	0,72	2,01	1,65	1,20
150	150	Crossroads	hf	f	AMR _NIII	no	Sharp	18	1,92	1,67	1,42	0,78	0,82	0,58	1,99	2,06	1,40
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	3,97	2,41	3,00
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	2,83	2,34	2,12
171	171	Crossroads	hs	m	AMR _NIII	no	Smooth	18	2,42	3,00	2,21	1,02	1,18	0,93	2,55	3,01	2,28
172	172	Crossroads	hs	m	AMR _NI	no	Sharp	9	3,88	2,42	2,83	0,85	0,93	0,92	4,07	2,34	3,03
177	177	Crossroads	hs	m	AMR _NIII	no	Sharp	18	2,48	3,00	2,46	0,88	1,25	0,88	2,67	2,87	2,28
181	181	Crossroads	hs	m	AMR _NI	yes	Smooth	18	3,04	3,08	2,83	1,08	1,14	0,92	3,01	3,00	2,60
184	184	Crossroads	hs	m	AMR _NI	yes	Sharp	9	3,96	2,57	3,29	1,08	1,06	0,75	4,06	2,57	3,15
186	186	Crossroads	hs	m	AMR _NIII	yes	Sharp	9	3,75	2,58	2,67	1,22	1,10	1,13	3,31	2,45	2,53
187	187	Crossroads	hs	m	AMR _NI	yes	Sharp	18	2,96	3,50	2,88	1,27	1,10	1,15	3,28	3,33	2,96
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	2,36	1,95	1,58
201	201	Crossroads	hf	m	AMR _NIII	no	Sharp	9	2,29	1,38	1,71	1,23	0,58	0,69	2,05	1,63	1,21
204	204	Crossroads	hf	m	AMR _NIII	no	Sharp	18	1,63	1,88	1,38	0,97	1,15	0,65	1,77	1,98	1,20
213	213	Crossroads	hf	m	AMR _NIII	yes	Sharp	9	2,13	1,42	1,46	0,95	0,58	0,59	2,13	1,80	1,34
225	225	Road_Noise	hs	f	AMR _NIII	no	Smooth	18	2,29	2,46	2,17	0,62	0,78	0,64	2,46	2,49	1,94
226	226	Road_Noise	hs	f	AMR _NI	no	Sharp	9	3,67	2,54	2,88	0,96	0,72	0,80	3,34	2,19	2,41
229	229	Road_Noise	hs	f	AMR _NI	no	Sharp	18	3,17	2,50	2,58	1,09	0,88	0,72	2,70	2,64	2,19
238	238	Road_Noise	hs	f	AMR _NI	yes	Sharp	9	3,71	2,29	2,70	0,95	0,62	0,75	3,52	2,56	2,74
241	241	Road_Noise	hs	f	AMR _NI	yes	Sharp	18	2,96	3,13	2,46	1,30	0,85	0,83	3,49	3,42	3,16
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	3,40	2,08	2,39
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	2,56	2,01	1,75
247	274	Road_Noise	hs	m	AMR _NI	no	Smooth	9	3,33	1,54	2,29	1,09	0,66	0,75	3,44	1,88	2,31
249	276	Road_Noise	hs	m	AMR _NIII	no	Smooth	9	1,38	1,21	1,00	0,97	0,83	0,00	2,46	1,37	1,34
256	283	Road_Noise	hs	m	AMR _NI	no	Sharp	18	2,75	2,17	2,08	0,99	0,82	0,88	2,86	2,20	2,07
276	330	Office_Noise	hs	f	G722_NIII	no	Smooth	9	3,29	3,71	3,09	0,95	0,69	0,72	3,13	3,28	2,82
277	331	Office_Noise	hs	f	G722_NI	no	Smooth	18	4,75	3,83	4,04	0,53	0,82	0,81	4,69	3,38	4,05
289	343	Office_Noise	hs	f	G722_NI	yes	Smooth	18	4,54	3,63	3,67	0,59	0,65	0,76	4,71	3,51	4,13
292	346	Office_Noise	hs	f	G722_NI	yes	Sharp	9	4,83	4,13	4,33	0,48	0,45	0,64	4,70	3,63	4,17
294	348	Office_Noise	hs	f	G722_NIII	yes	Sharp	9	3,54	3,38	3,17	1,14	0,71	0,76	3,48	3,21	3,04
297	351	Office_Noise	hs	f	G722_NIII	yes	Sharp	18	2,67	3,88	2,79	0,96	0,61	0,72	3,18	3,38	2,91
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	4,04	2,58	3,13
301	355	Office_Noise	hf	f	G722_NI	no	Smooth	9	4,21	3,00	3,08	0,93	0,78	0,65	4,24	2,66	3,33
309	363	Office_Noise	hf	f	G722_NIII	no	Sharp	9	2,54	2,50	2,33	0,78	0,51	0,48	2,82	2,43	2,17
310	364	Office_Noise	hf	f	G722_NI	no	Sharp	18	3,38	2,83	2,75	1,17	0,87	0,79	4,05	2,97	3,35

Conditions	Id real	Noise	Recording	Speaker	Network	VAD	Smooth	dB	FRENCH								
									Subjective						Objective		
									MOS			Standard deviation			MOS		
									Speech	Noise	Global	Speech	Noise	Global	Speech	Noise	Global
316	370	Office_Noise	hf	f	G722_NI	yes	Smooth	18	3,75	2,75	2,79	1,03	0,79	0,98	3,79	3,14	3,24
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	3,15	2,80	2,60
331	385	Office_Noise	hs	m	G722_NI	no	Smooth	18	4,83	4,08	4,08	0,48	0,65	0,65	4,59	3,38	3,96
334	388	Office_Noise	hs	m	G722_NI	no	Sharp	9	4,71	3,88	4,08	0,69	0,54	0,88	4,58	3,17	3,86
366	420	Office_Noise	hf	m	G722_NIII	no	Sharp	18	2,79	2,33	2,17	1,25	0,76	0,76	3,12	2,66	2,50
372	426	Office_Noise	hf	m	G722_NIII	yes	Smooth	18	3,00	2,52	2,30	1,14	0,77	0,80	3,00	2,88	2,52
373	427	Office_Noise	hf	m	G722_NI	yes	Sharp	9	4,67	3,17	3,38	0,48	0,56	0,65	4,28	2,73	3,40
378	432	Office_Noise	hf	m	G722_NIII	yes	Sharp	18	2,88	2,67	2,38	0,95	0,96	0,82	3,09	2,71	2,50
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	3,60	2,18	2,59
384	438	Pub_Noise	hs	f	G722_NIII	no	Smooth	9	3,00	1,63	1,92	1,02	0,65	0,58	2,82	1,74	1,78
390	444	Pub_Noise	hs	f	G722_NIII	no	Sharp	9	2,42	1,79	1,79	1,02	1,06	0,66	2,71	1,84	1,76
396	450	Pub_Noise	hs	f	G722_NIII	yes	Smooth	9	2,38	1,79	1,67	0,88	0,66	0,87	2,74	1,96	1,85
415	469	Pub_Noise	hs	m	G722_NI	no	Sharp	9	3,96	1,67	2,08	1,08	0,96	0,88	4,15	1,80	2,79
420	474	Pub_Noise	hs	m	G722_NIII	no	Sharp	18	1,88	1,21	1,38	0,90	0,41	0,49	2,81	1,68	1,73
423	477	Pub_Noise	hs	m	G722_NIII	yes	Smooth	9	2,46	1,75	1,89	0,98	0,68	0,75	2,89	1,97	1,96
427	481	Pub_Noise	hs	m	G722_NI	yes	Sharp	9	3,79	1,99	2,25	1,06	0,92	0,79	3,74	2,34	2,79
432	486	Pub_Noise	hs	m	G722_NIII	yes	Sharp	18	1,96	1,88	1,54	0,95	0,68	0,66	2,70	2,50	2,12

Annex G:
Void

Annex H:
Extension of the Speech Quality Test Method to Narrowband: Adaptation, Training and Validation

The first version of the present document was restricted to wideband application only. Due to the lack of freely available databases containing narrowband speech and evaluated according to Recommendation ITU-T 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 (see ETSI EG 202 396-2 [i.2]) 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 Recommendation ITU-T P.835 [i.3] listening-only test.

Table H.1: Comparison of Databases

	ETSI STF WB Database	HEAD acoustics NB Database
Language		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 (see ETSI EG 202 396-2 [i.2]), 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 ETSI ES 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 Recommendation ITU-T 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 (see ETSI EG 202 396-2 [i.2]), 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 (see ETSI EG 202 396-2 [i.2]) and the new database are listed in table H.3.

Table H.3: Design differences between the STF 294 database (see ETSI EG 202 396-2 [i.2]) 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 Recommendation ITU-T P.800 [i.4].

Annex I:
Void

Annex J:
Summary of Czech samples not used for model training

J.0 Introduction

Almost all topics regarding the Czech database presented in ETSI EG 202 396-2 [i.2] were already marked as invalid or at least not recommended to use, mainly due to the insufficient audio material and the corresponding subjective database. Thus it was decided to move all related content regarding this topic from the main body of a previous version to this annex for informational purposes.

Due to the high diversity between the Czech and the French listening test the development of the objective model is based only on the French database being within the ToR and such provides the higher amount of selected samples.

J.1 Selection process - Czech database

For every combination of background noise and speaker gender, a single Czech sentence was used (see table J.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 J.1: Sentences from the test corpus chosen for the different conditions

Condition	Sentence No.
Lux Car 130 kmh Female2	S3
Lux Car 130 kmh 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. Since 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 J.2.

Table J.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 J.3.

Table J.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.

J.2 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 ETSI EG 202 396-2 [i.2]. This may have lead to different ratings for the Czech samples compared to the French samples. This has to be regarded especially for further processing of the sound files.

- For every background noise condition, a single Czech sentence was used (see table J.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 J.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 J.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 J.1, J.3 and J.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 J.2, J.4 and J.6).

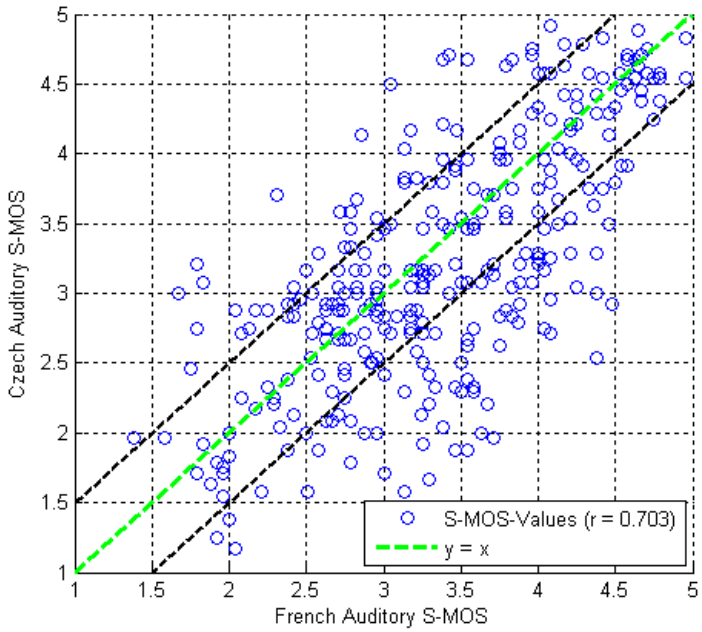


Figure J.1: Scatter plot of the French data versus the Czech data for the different conditions, S-MOS, *before* experts' selection

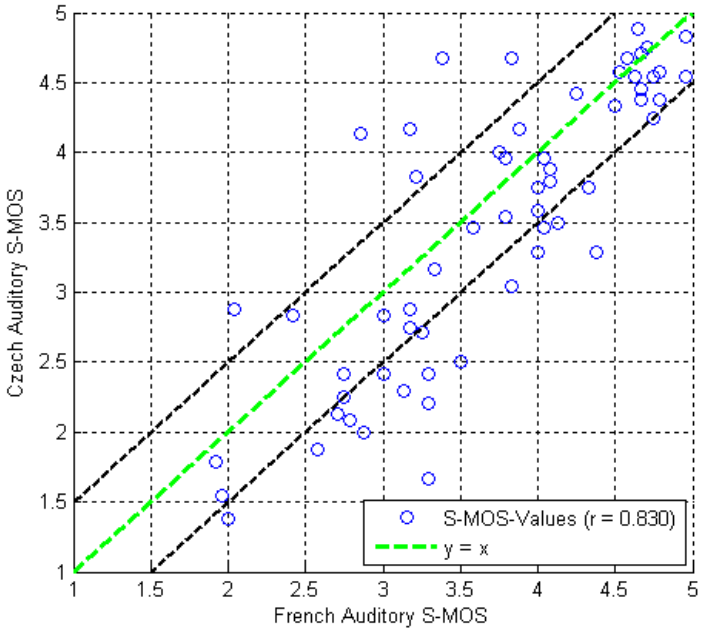


Figure J.2: Scatter plot of the French data versus the Czech data, S-MOS, *after* experts' selection (only data selected for both languages)

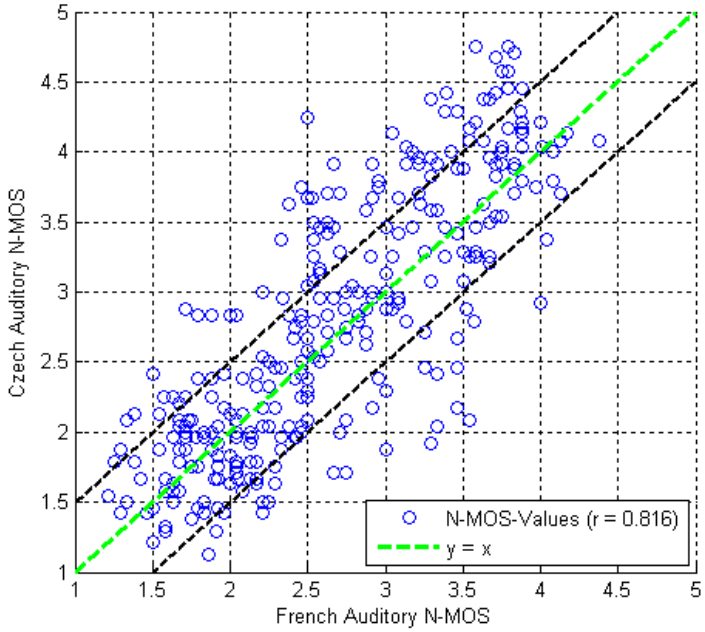


Figure J.3: Scatter plot of the French data versus the Czech data for the different conditions, N-MOS, *before* experts' selection

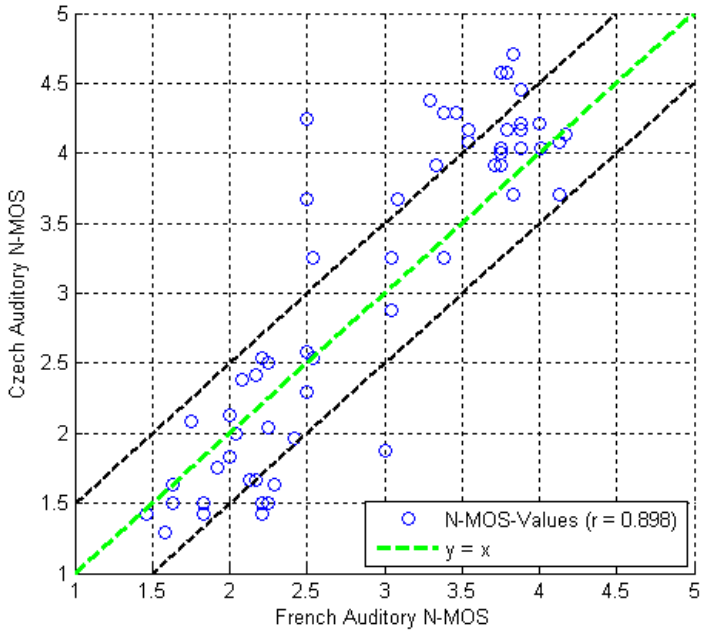


Figure J.4: Scatter plot of the French data versus the Czech data, N-MOS, *after* experts' selection (only data selected for both languages)

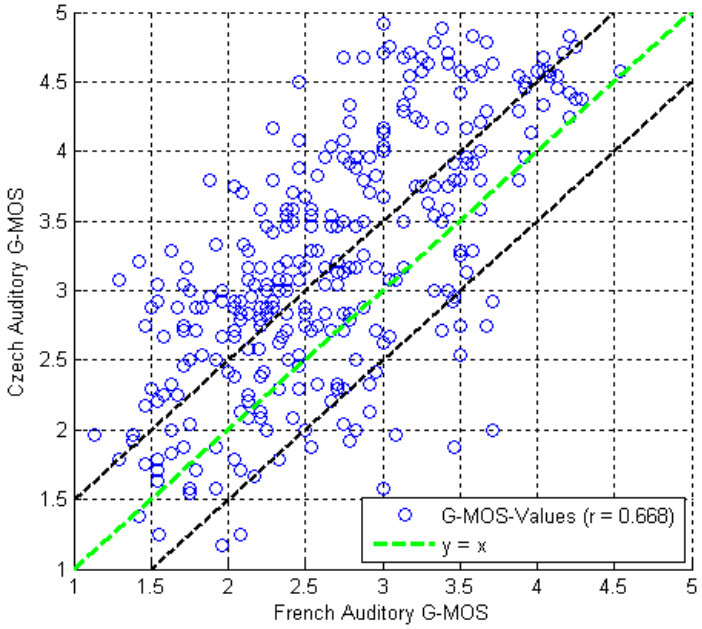


Figure J.5: Scatter plot of the French data versus the Czech data for the different conditions, G-MOS, *before* experts' selection

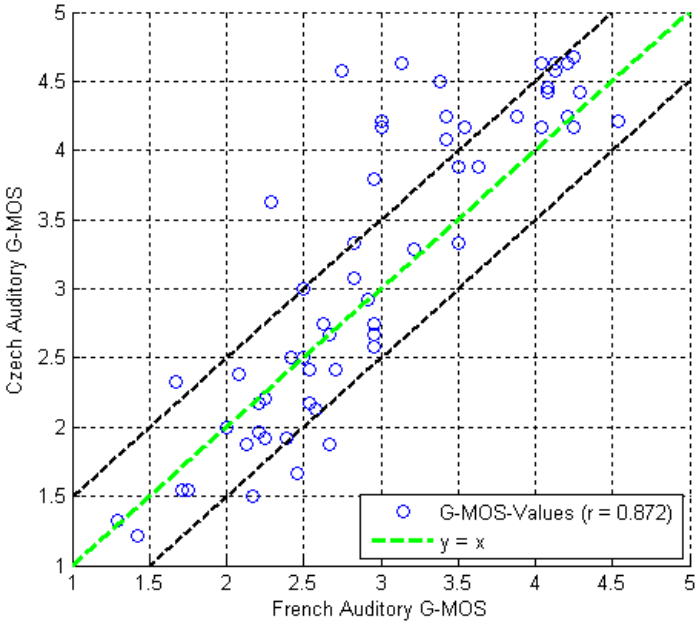


Figure J.6: Scatter plot of the French data versus the Czech data, G-MOS, *after* experts' selection (only data selected for both languages)

J.3 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 J.7 and J.8 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 6.5.3 the objective N-MOS correlates with 94,8 % to the results of the French listening test. Figure J.8 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 tend 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 J.9 and J.10. They show that the objective results for the French data (figure J.9) 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 J.9). Figure J.9 indicates that a Czech listener would assess all French samples 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 figures J.11 to J.14. Figures J.11 and J.14 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).

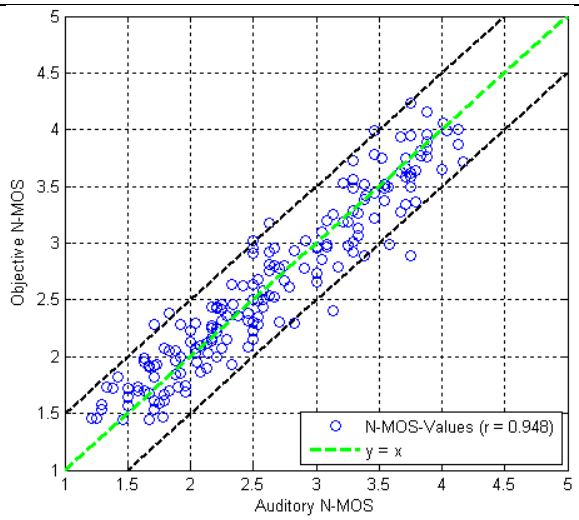


Figure J.7: Objective vs. Auditory N-MOS for French samples calculated with regression coefficients optimized for French data

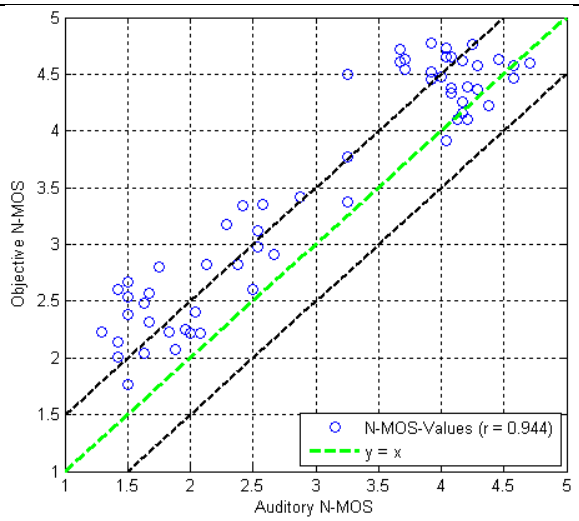


Figure J.8: Objective vs. Auditory N-MOS for Czech samples calculated with regression coefficients optimized for French data

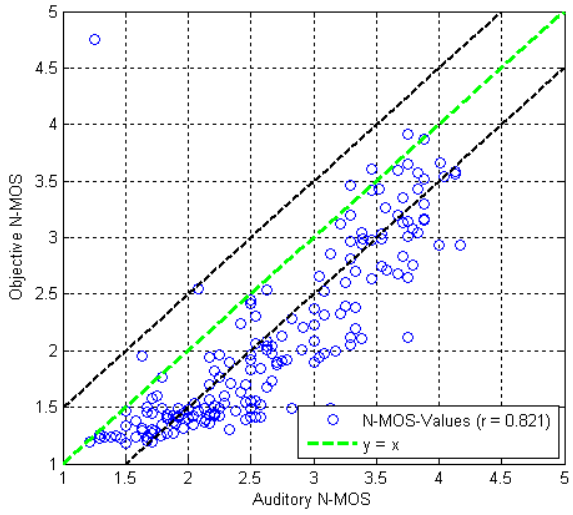


Figure J.9: Objective vs. Auditory N-MOS for French samples calculated with regression coefficients optimized for Czech data

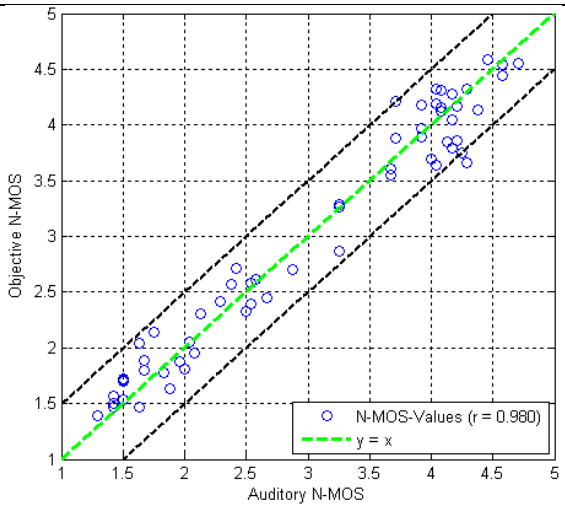


Figure J.10: Objective vs. Auditory N-MOS for Czech samples calculated with regression coefficients optimized for Czech data

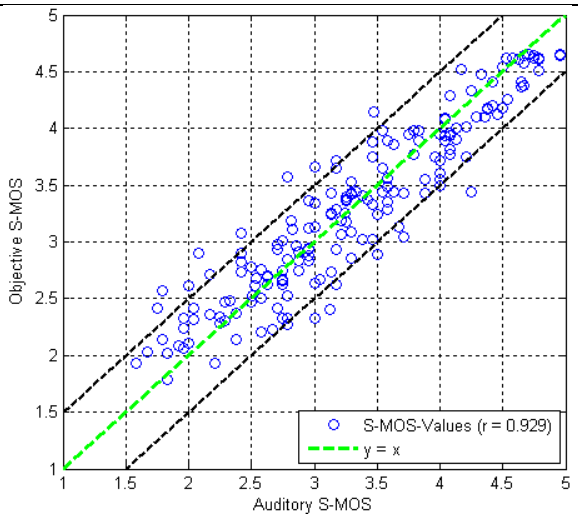


Figure J.11: Objective vs. Auditory S-MOS for French samples calculated with regression coefficients optimized for French data

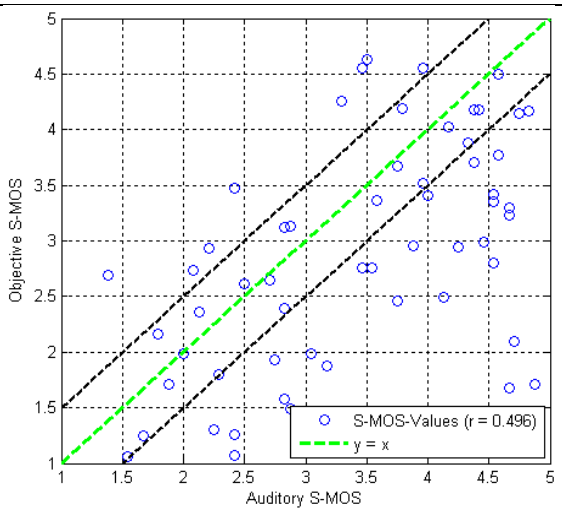


Figure J.12: Objective vs. Auditory S-MOS for Czech samples calculated with regression coefficients optimized for French data

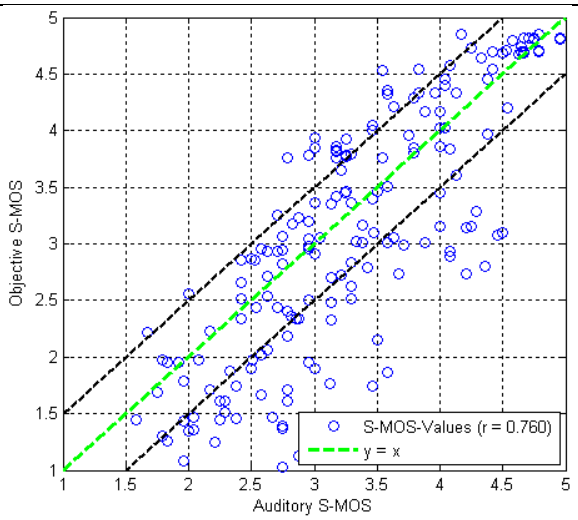


Figure J.13: Objective vs. Auditory S-MOS for French samples calculated with regression coefficients optimized for Czech data

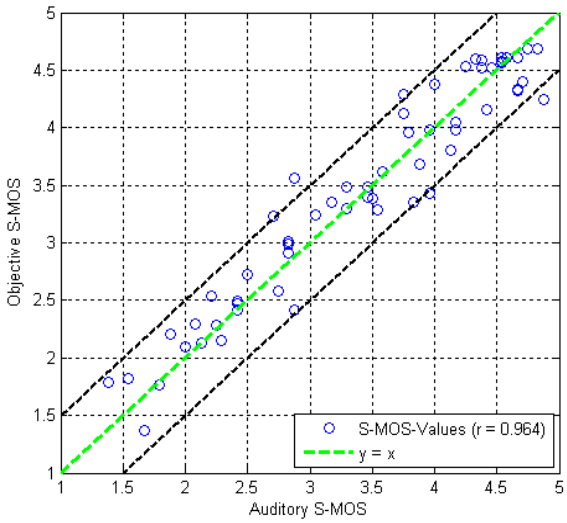


Figure J.14: Objective vs. Auditory S-MOS for Czech samples calculated with regression coefficients optimized for Czech data

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 J.9 and J.10 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.

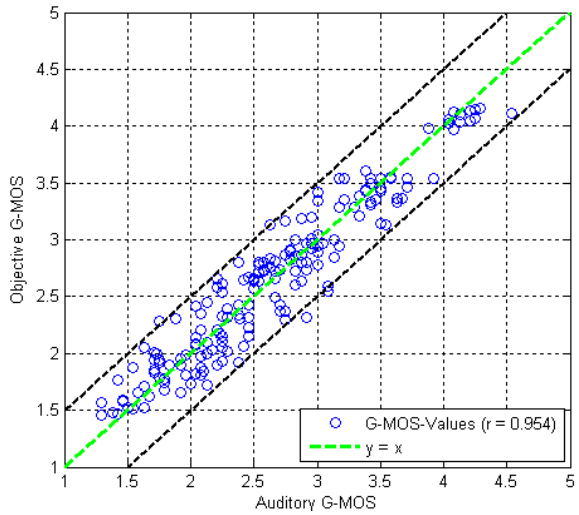


Figure J.15: Objective vs. Auditory G-MOS for French samples calculated with regression coefficients optimized for French data

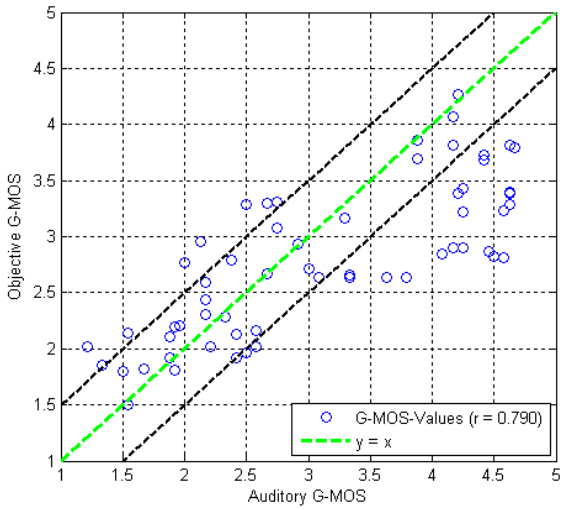


Figure J.16: Objective vs. Auditory G-MOS for Czech samples calculated with regression coefficients optimized for French data (N-MOS and S-MOS optimized for French data)

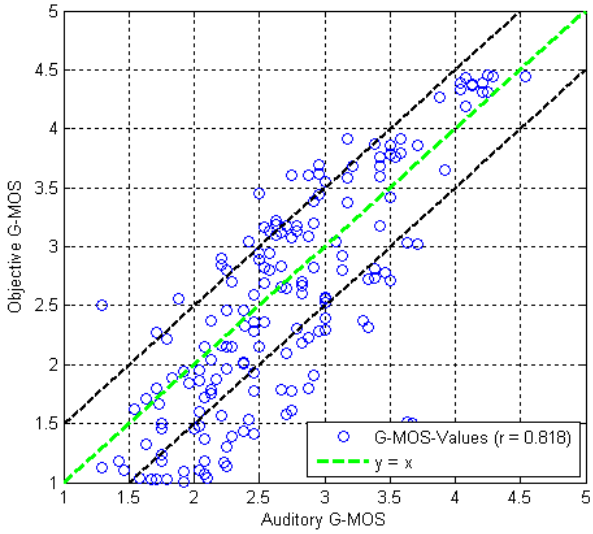


Figure J.17: Objective vs. Auditory G-MOS for French samples calculated with regression coefficients optimized for Czech data (N-MOS and S-MOS optimized for Czech data)

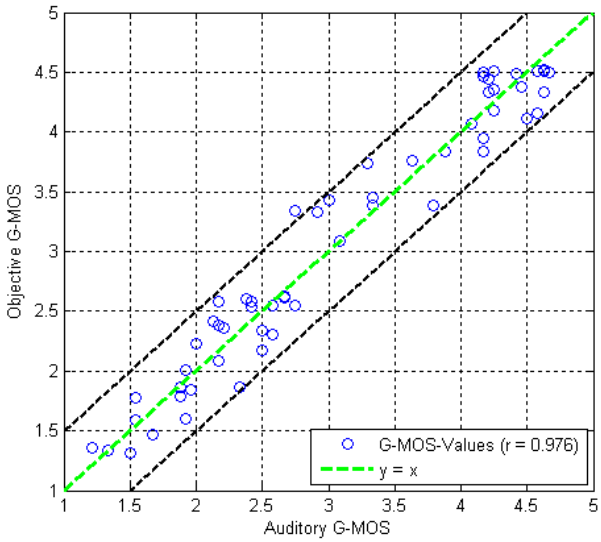


Figure J.18: Objective vs. Auditory G-MOS for Czech samples calculated with regression coefficients optimized for Czech data

The analysis for the objective G-MOS are shown with the same principle in figures J.15 to J.18. 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.

J.4 Czech conditions results analysis

J.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.

Figures J.19 and J.20 show 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.

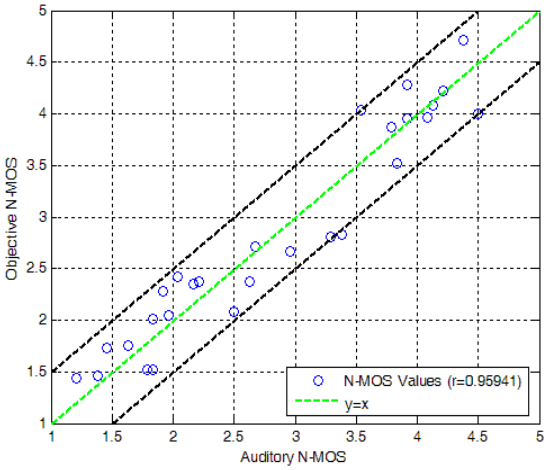


Figure J.19: Objectively calculated N-MOS vs. auditory N-MOS for Czech validation conditions

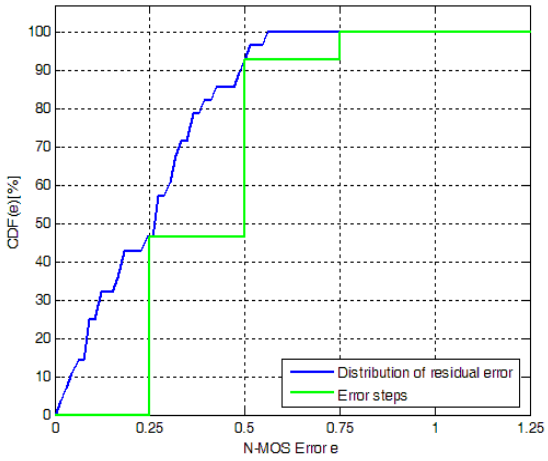


Figure J.20: Objectively CDF of residual error vs. N-MOS Error e for Czech validation conditions

For this situation, the **RMSE value is 0,293** and the distribution of the residual error is shown in figure J.20 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.

J.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.

Figures J.21 and J.22 show 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.

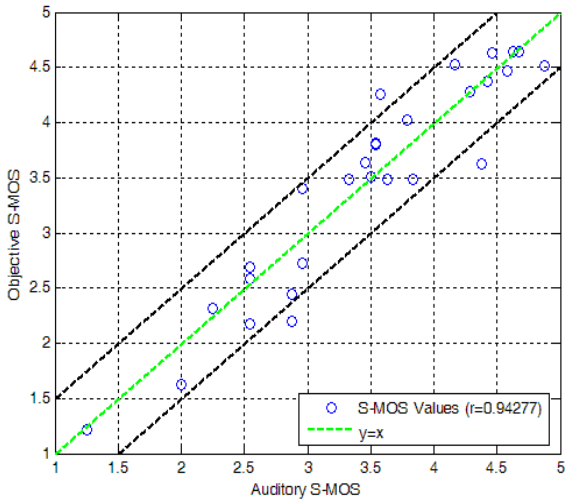


Figure J.21: Objectively calculated S-MOS vs. auditory S-MOS for Czech validation conditions

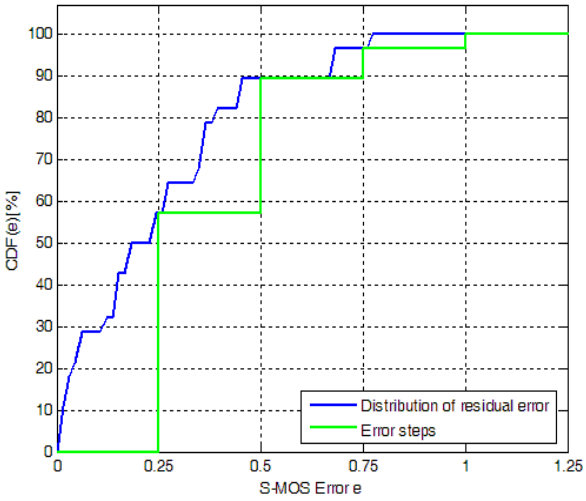


Figure J.22: Objectively CDF of residual error vs. S-MOS Error e for Czech validation conditions

For this situation, the **RMSE value is 0,22** and the distribution of the residual error is shown in figure J.22 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.

J.4.3 Comparing Subjective and Objective G-MOS Results

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

Figures J.23 and J.24 show 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.

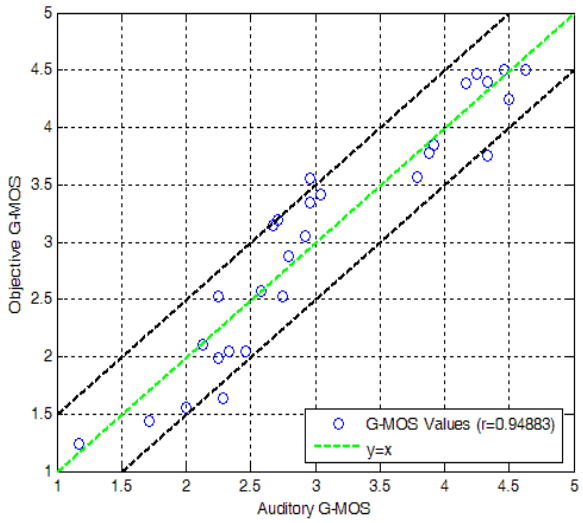


Figure J.23: Objectively calculated S-MOS vs. auditory G-MOS for Czech validation conditions

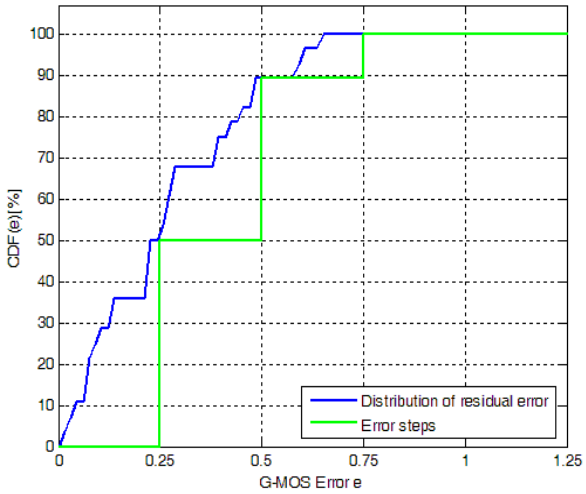


Figure J.24: Objectively CDF of residual error vs. G-MOS Error e for Czech validation conditions

For this situation, the **RMSE value is 0,21** and the distribution of the residual error is shown in figure J.24 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.

J.5 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 Recommendation ITU-T P.56 [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 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 J.25 and J.26.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.

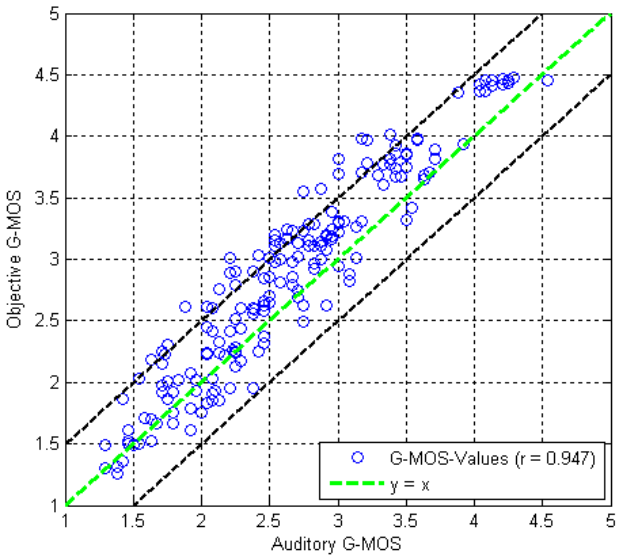


Figure J.25: Objective vs. Auditory G-MOS for French samples calculated with regression coefficients optimized for Czech data (N-MOS and S-MOS optimized for French data)

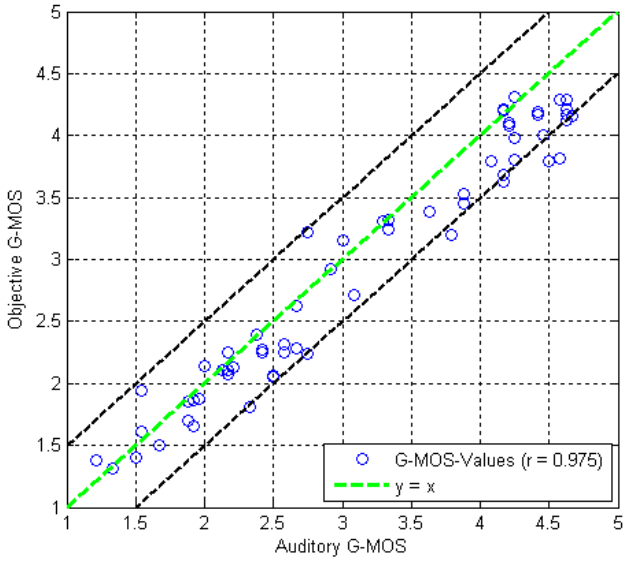


Figure J.26: Objective vs. Auditory G-MOS for Czech samples calculated with regression coefficients optimized for French data (N-MOS and S-MOS optimized for Czech data)

This effect can also be proved by comparing the G-MOS regression planes for the Czech and French coefficients as given in figures J.27 and J.28. 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.

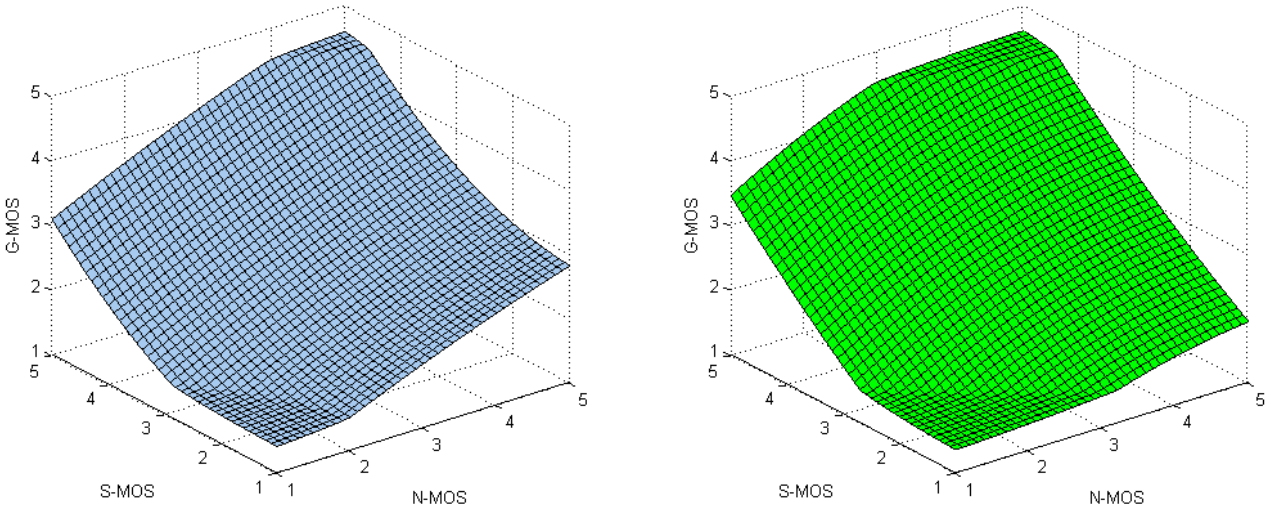


Figure J.27: Comparison of French (left, blue) and Czech (right, green) regression plane

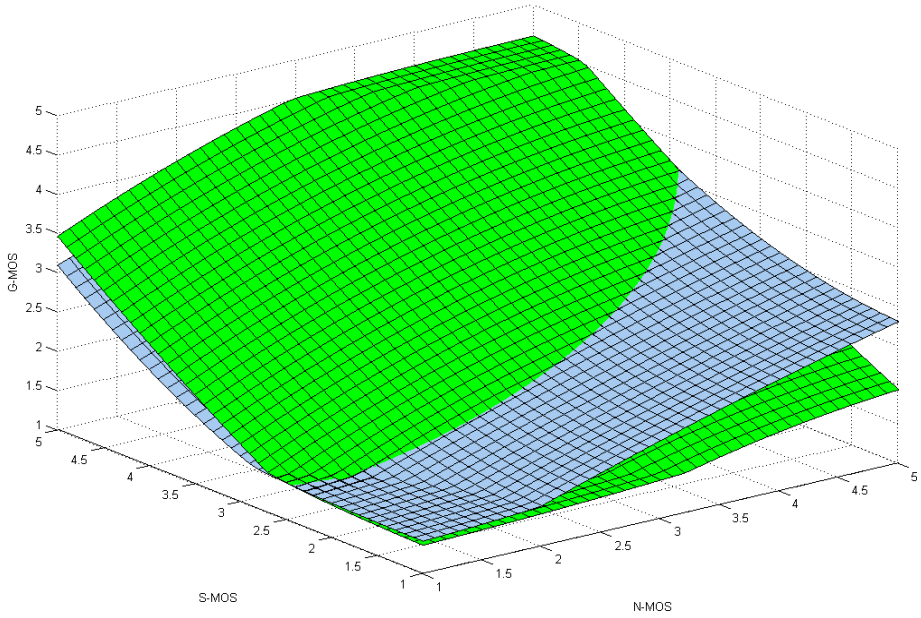


Figure J.28: Comparison of French (blue) and Czech (green) regression plane

J.6 Regression Coefficients for Czech data

This clause summarizes the regression coefficients for the S-, N- and G-MOS calculation of the Czech data.

Table J.5: 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 J.6: 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)	${}_1c_{j2}$ ($mRA_{SP,p}$)	${}_1c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	${}_1c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	${}_1c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	${}_1c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	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 J.7: 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)	${}_2c_{j2}$ ($mRA_{SP,p}$)	${}_2c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	${}_2c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	${}_2c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	${}_2c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	2,2623	-0,0283	1,7981	-1,1318	-1,2940	-0,1389	-0,2207
2	-	-	-0,2587	-0,1753	-0,9927	0,0022	0,0051

Table J.8: 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)	${}_3c_{j2}$ ($mRA_{SP,p}$)	${}_3c_{j3}$ ($m\Delta RA_{SP,p\text{-}c}$)	${}_3c_{j4}$ ($m\Delta RA_{SP,p\text{-}u}$)	${}_3c_{j5}$ ($v\Delta RA_{SP,p\text{-}c}$)	${}_3c_{j6}$ ($v\Delta RA_{SP,p\text{-}u}$)
1	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 J.9: 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

J.7 Post selection

Table J.10 contains the conditions and related auditory S-MOS, N-MOS and G-MOS for the Czech language. Also standard deviations for all MOS scores are given. The results for validation purposes are blinded.

**Table J.10: 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
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

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	
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
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

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	
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 K:
Relative Approach Non-Linear Transformation

The non-linear transformation can be described by a function which is defined in two intervals:

$$y_1 = \begin{cases} x - a \cdot x^2, & 0 \le x < x_S \\ b \cdot x^{\nu} - c, & x_S \le x \end{cases}$$

Where x is P_{eff} / P_0. P_{eff} is the RMS value of the sound pressure level and P_0 is the reference sound pressure level equivalent to 20 uPa. Then y_1 represents the normalized envelope value of the non-linear transformation.

When assuming a linear characteristic and applying a sinusoidal signal with a sound pressure level of 20 dB the value derived by (4,46) will be 10.

The constants a, b and c are predetermined for the variable v (exponent is approximately 0,20 to 0,25) and x_s (the threshold of nonlinearity is approximately 5,65) in such a manner that the function is twice continuously differentiable at the transition x_s.

$$\begin{aligned} a &= \frac{1}{2} \cdot \frac{1-\nu}{2-\nu} \cdot \frac{1}{x_S}, \\ b &= \frac{1}{\nu} \cdot \frac{1}{2-\nu} \cdot \frac{x_S}{x_S^{\nu}}, \\ c &= \frac{1}{2} \cdot \frac{1-\nu}{\nu} \cdot x_S . \end{aligned}$$

The inverse function of (4,46) is:

$$x = \begin{cases} \frac{1}{2a} + \sqrt{\frac{1}{(2a)^2} - \frac{y_1}{a}}, & 0 \le y_1 < y_S \\ \sqrt[\nu]{\frac{y_1 + c}{b}}, & y_S \le y_1 \end{cases},$$

In which $y_s = x_s - a \times x_s^2$ has been used.

Alternatively an additional function was used. However the inverse function cannot be represented as a closed form expression:

$$y_2 = x \cdot \left(\sqrt[\alpha]{\frac{x^{\alpha} + x_S^{\alpha}}{1 + x_S^{\alpha}}} \right)^{\nu - 1}$$

The asymptotes of the expression in (4,51) are $y_2 = x$ when $x << x_s$ and $y_2 = x_s \times (x/x_s)^v$ when $x >> x_s >> 1$. When plotted in a double-logarithmic graph two straight lines with a slope of 1 respective v and intersection at ($x = x_s$, $y_2 = x_s$).

The parameter alfa allows to adjust the size of the transition region between the asymptotic approximations.

The following values are used: $v = 0,2$; $x_s = 5,66$.

Annex L: Bibliography

R.A. Fisher: "Statistical methods and scientific inference", Oliver and Boyd, 1956.

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)".

3GPP S4-120542: "Common subjective testing framework for training of P.835 test predictors".

History

Document history		
V1.1.1	August 2007	Publication
V1.2.1	January 2009	Publication
V1.3.1	February 2011	Publication
V1.4.1	June 2014	Publication
V1.5.1	August 2015	Membership Approval Procedure MV 20151011: 2015-08-12 to 2015-10-12
V1.5.1	October 2015	Publication