ETSI EG 202 396-3 V1.7.1 (2018-03)

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-270

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

Essential patents

IPRs essential or potentially essential to normative deliverables 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 (https://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.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

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 "**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

Normative references are not applicable in the present document.

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 referenced 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. and Friedman J.: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", New York: Springer-Verlag, 2001.

[i.34] ETSI EG 202 396-3 (V1.1.1 to V1.3.1): "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise; Part 3: Background noise transmission - Objective test methods".

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
HiQ	High Quality (codec mode)
IP	Internet Protocol
IRS	Intermediate Reference System
ITU	International Telecommunication Union
ITU-T	Telecom Standardization Body of ITU
LQ	Low Quality (codec mode)
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MOS-LQSN	Mean Opinion Score - Listening Quality Subjective Noise
MRP	Mouth Reference Point
NB	NarrowBand
NBGN	Level of background noise
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
SG	Study Group
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 equalization provided by the headphone manufacturer was used since this was the one used in the auditory French test setup.

NOTE: These headphones and headphone amplifiers were used in the tests since they provide the performance required. Other products providing the equivalent performance could be used if such an experiment should be repeated by others. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of these products.

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.

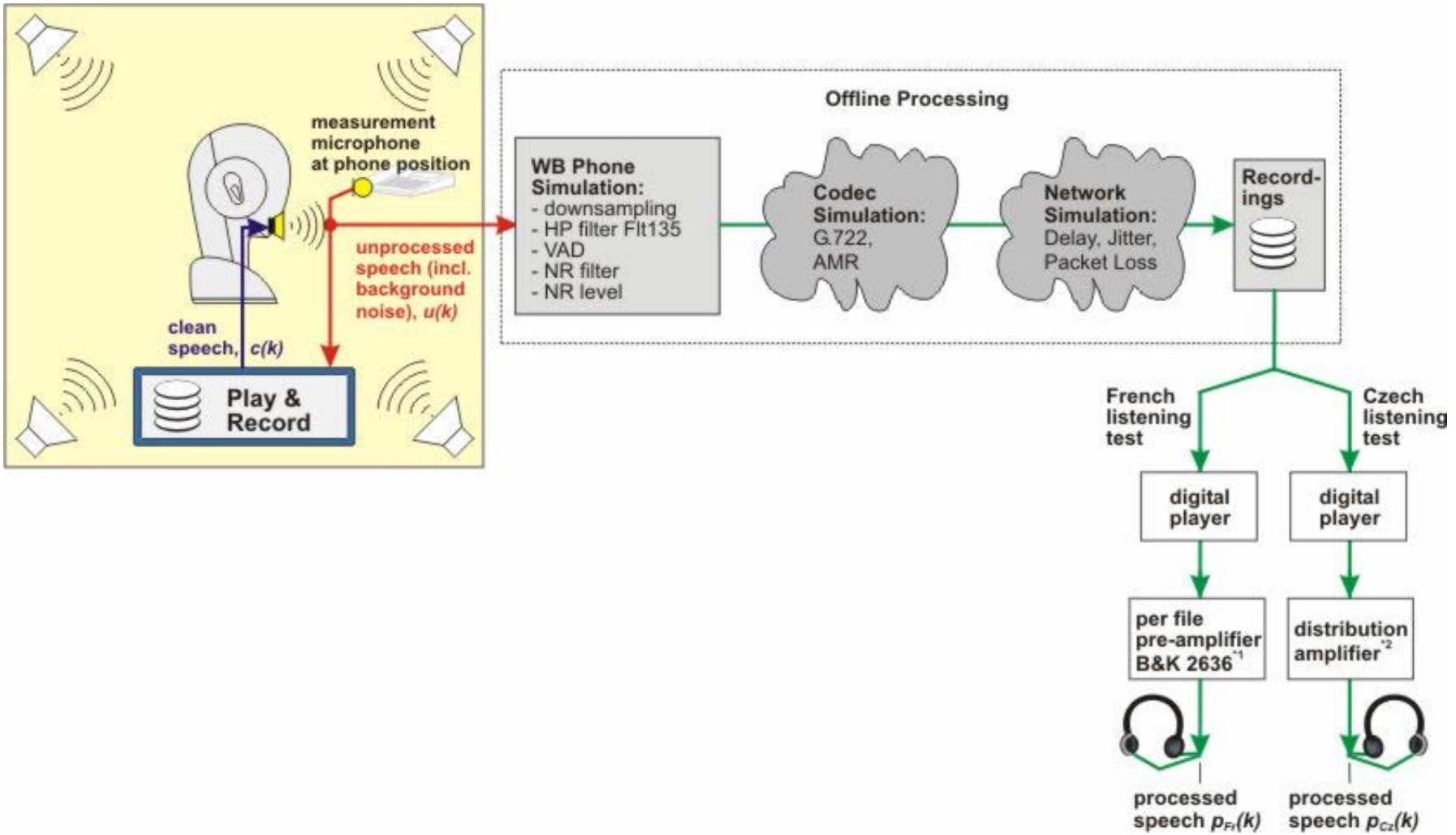
Different input signals can be accessed during the recording process and subsequently can be used for the calculation of N-MOS, S-MOS and G-MOS. Beside the signals used in the listening test ("processed signal"), two additional signals are used as a priori knowledge for the calculation:

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 1: 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.

Due to time alignment, several parts in signals may be obtained, where no corresponding part exists in the other signals. Thus these segments are discarded. Figure 6.1a illustrates the strategy of signal cropping after time alignment.

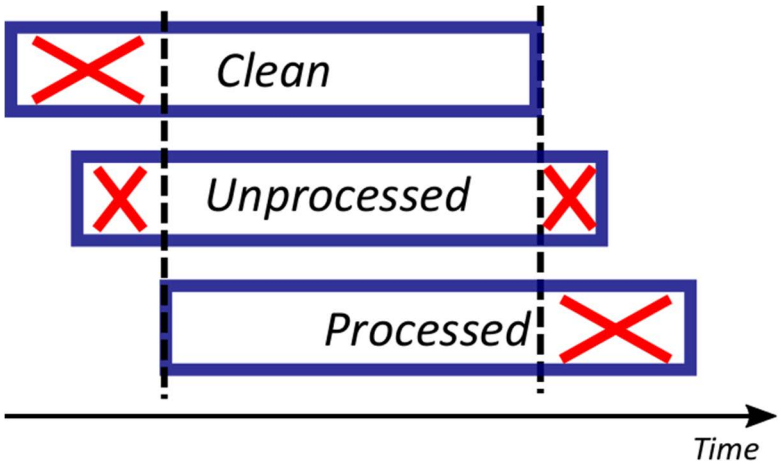


Figure 6.1a: Signal alignment

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 method described in [i.31] performs a frame categorization on the clean input signal (see section II.4.1 in Recommendation ITU-T G.160 [i.31]). It first transforms the signal into a level-vs-time transformation based on 10 ms frames. Each frame is then categorized as either silence, pause, uncertain, low/mid/high speech activity. The signal parts classified as silence are assumed as background noise/silence sections for unprocessed and processed signal. All other frames are considered as active speech.

For the recording procedure, the clean speech signals are expected to have an Active Speech Level (ASL, see Recommendation ITU-T P.56 [i.22]) of -4,7 dB Pa at the mouth reference point (MRP). Additional level increments may be added for compensating Lombard effect (typically +3 dB), i.e. obtaining a more realistic signal-to-noise ratio.

For the instrumental prediction method, all three input signals are scaled to an active speech level of either 73 dB SPL (narrowband mode) respectively 79 dB SPL (wideband mode). These levels correspond to the scaling used in the underlying listening test databases.

NOTE 2: The unprocessed signal and also the processed signal as well may include too much noise for the proper calculation of active speech level according to Recommendation ITU-T P.56 [i.22]. In this case, the level of the noisy speech is calculated via the speech part detection previously described.

NOTE 3: Speech level calculations are carried out over speech including noise. The more noise is present in the processed or unprocessed signal, the less speech-only energy contributes to the overall level. In borderline cases this may result in an unreasonable biased estimate of speech-only level only, but this method corresponds to the level calibration used in the auditory experiments.

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.

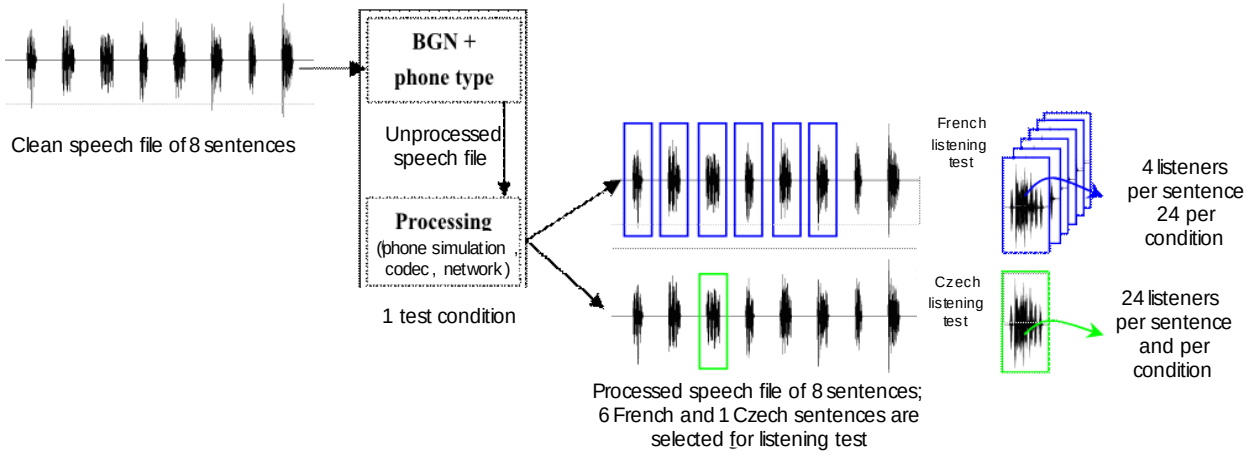


Figure 6.2: Nomenclature (file, condition, sentence)

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.

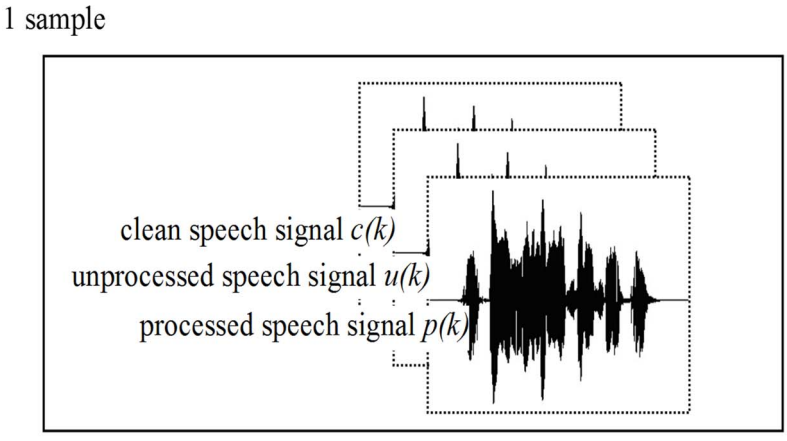


Figure 6.3: Nomenclature (sample)

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.

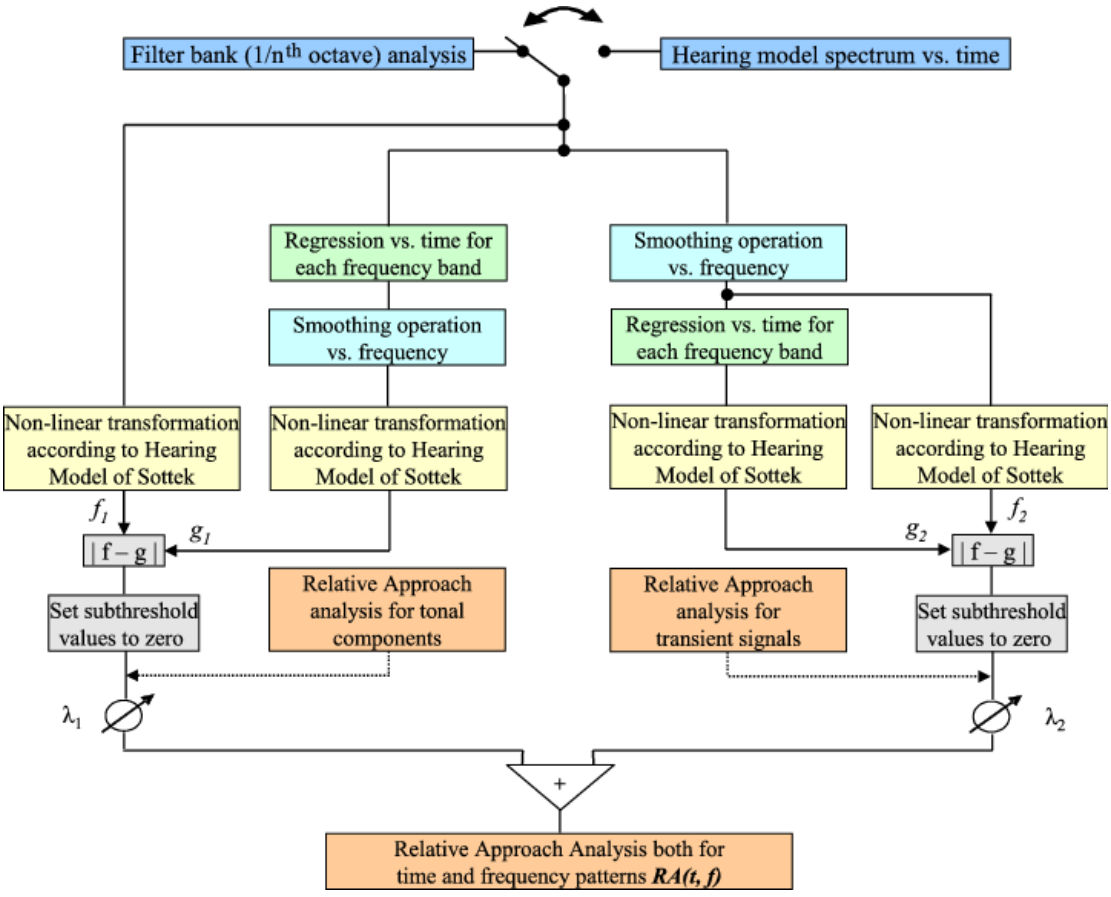


Figure 6.4: Block diagram of Relative Approach

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 6.5 a)).

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 6.5 b)); 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.

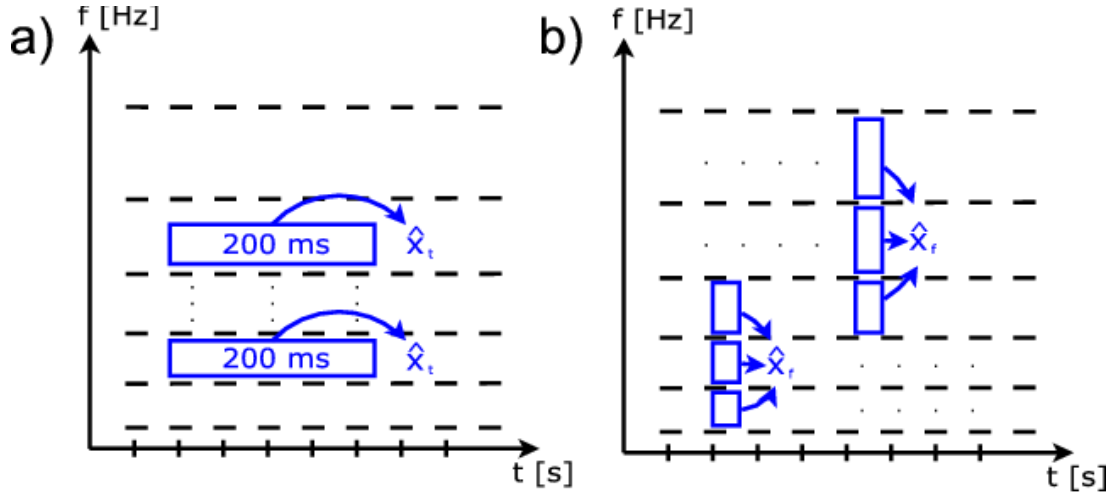


Figure 6.5: Regression modes of Relative Approach:
a) versus time, b) versus frequency

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).

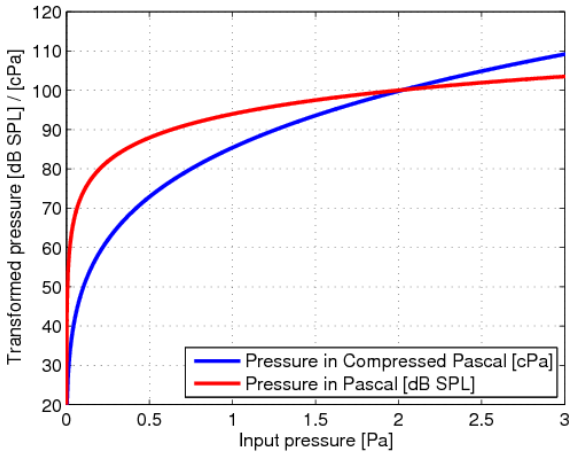


Figure 6.6: Non-linear transformation of sound pressure according to hearing model of Sottek

For the first branch of the Relative Approach, a regression versus time is applied for each frequency band (see figure 6.5 a)). Afterwards, a smoothing versus frequency is performed for each time slot (see figure 6.5 b)), 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 $\Delta t = 6{,}66$ 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 analysing 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.

$$\Delta RA(\Delta t_i, \Delta f_j) = RA_p(\Delta t_i, \Delta f_j) - RA_u(\Delta t_i, \Delta f_j) \tag{6.1}$$

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

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 4th 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 analyse 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 250 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 $\Delta RA_{BGN,p\text{-}u}(t, f)$ 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 $\Delta RA_{BGN,p\text{-}u}(t, f)$ and the **mean μ** for $RA_p(t, f)$ and $\Delta RA_{BGN,p\text{-}u}(t, f)$ are calculated according to 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 (250 ms).

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

and

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

with: $A_{ges}=(t_{max}-t_{min})(f_{max}-f_{min})$.

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

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} + 250 ms and t_{max} given by the background noise section extracted before.

The mean values $\mu(RA_{BGN,U})$ and $\mu(RA_{BGN,P})$ as well as the variance $\sigma^2(\Delta RA_{BGN,P-U})$ are calculated for the spectrogram $\Delta RA_{BGN,p-u}(t, f)$ 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 and variance 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 $\mu(RA_{BGN,P})$ 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).

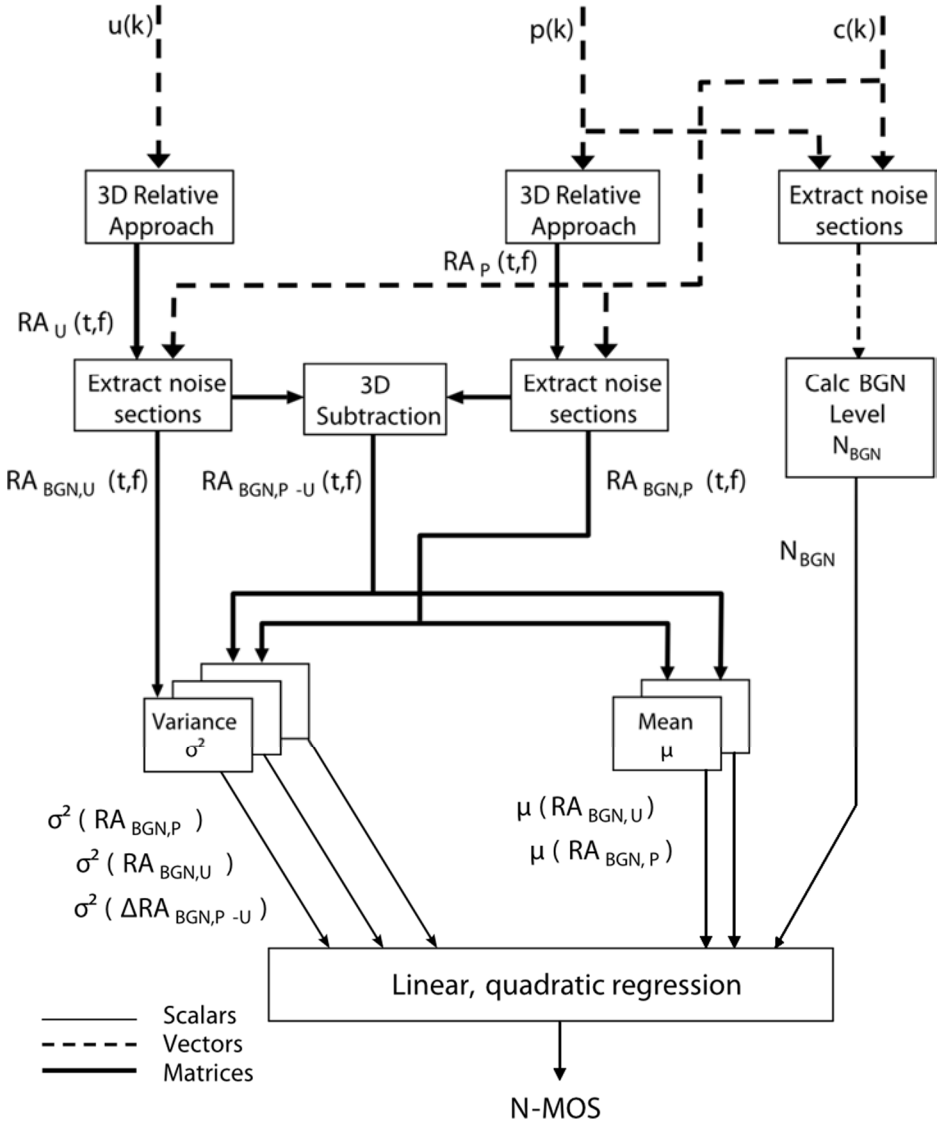


Figure 6.7: Block diagram of N-MOS calculation algorithm; *u(k)* unprocessed signal, *p(k)* processed signal, *c(k)* clean speech signal

Therefore the variances $\sigma(RA_{BGN,U})$ and $\sigma(RA_{BGN,P})$ are calculated for the unprocessed and the processed 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. Note that the variance parameters are square-rooted ($\sigma = \mathrm{sqrt}(\sigma^2)$) for better conditioning of the linear regression.

Table 6.1: Extracted parameters for N-MOS

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

Finally the *N-MOS* is the result of a **linear, quadratic regression** algorithm applied to all six parameters P_0 to P_5:

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

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 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 ($N_{BGN,P}$, $\sigma(RA_{BGN,U})$, $\sigma(RA_{BGN,P})$, $\sigma(\Delta RA_{BGN,P-U})$, $\mu(RA_{BGN,U})$ and $\mu(RA_{BGN,P})$) 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.

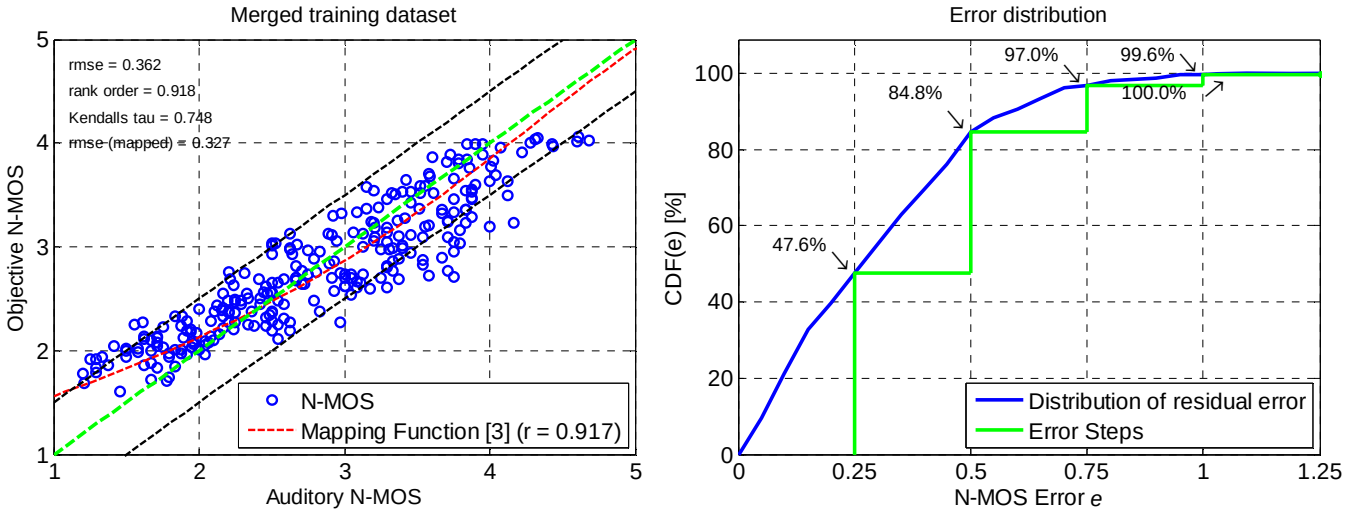


Figure 6.8: Left: Objectively calculated N-MOS versus auditory N-MOS; Right: CDF of residual error versus N-MOS Error e

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.

$$e = \left| NMOS_{auditory} - NMOS_{objective} \right| \quad (6.7)$$

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 neural network leading to the objective S-MOS are ΔSNR and five Relative Approach related parameters.

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 deltaq spectra using the equations (6.4) and (6.5) resulting in $\sigma^2(\Delta RA_{Sp,\,P\text{-}C})$, $\sigma^2(\Delta RA_{Sp,\,P\text{-}U})$, $\mu(\Delta RA_{Sp,\,P\text{-}C})$ and $\mu(\Delta RA_{Sp,\,P\text{-}U})$. Additionally the mean $\mu(RA_{Sp,\,P})$ is calculated for $RA_{SP,p}(t, f)$.

The variance $\sigma^2(\Delta RA_{Sp,\,P\text{-}C})$ 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 $\sigma^2(\Delta RA_{Sp,\,P\text{-}C})$ 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 $\mu(RA_{Sp,\,P})$ is used in both cases in order to characterize the absolute "attention" attracted by the processed signal. The comparison of $\mu(RA_{Sp,\,P})$ and $\mu(\Delta RA_{Sp,\,P\text{-}C})$ covers the influence of added or removed patterns introduced by room acoustics, background noise, the phone and the signal processing during the transmission. Similarly $\mu(RA_{Sp,\,P})$ and $\mu(\Delta RA_{Sp,\,P\text{-}U})$ can be compared in order to assess only the influence of the terminal and the transmission. The combination of these three parameters indicates whether the speech quality was impaired or improved.

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 6.5.2).

The resulting values *ΔSNR*, $\mu(RA_{Sp,\,P})$, $\sigma^2(\Delta RA_{Sp,\,P\text{-}U})$, $\sigma^2(\Delta RA_{Sp,\,P\text{-}C})$, $\mu(\Delta RA_{Sp,\,P\text{-}U})$ and $\mu(\Delta RA_{Sp,\,P\text{-}C})$ 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. Note that again the square-rooted values of variances are used as the input of the neural network for better conditioning of the calculation.

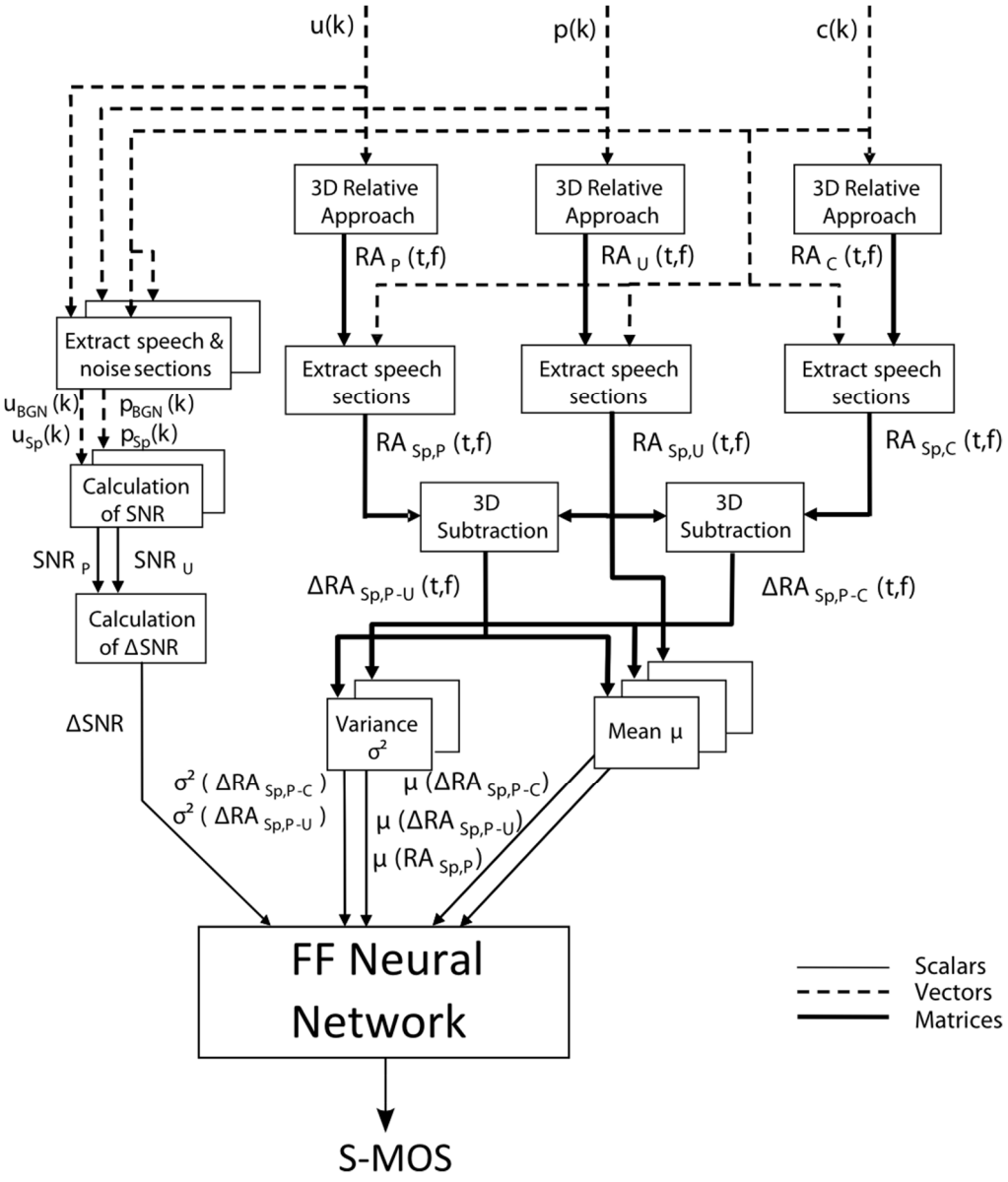


Figure 6.9: Block diagram of S-MOS calculation algorithm;
***u(k)* unprocessed signal, *p(k)* processed signal, *c(k)* clean speech signal**

Table 6.2: Extracted Parameters for S-MOS

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

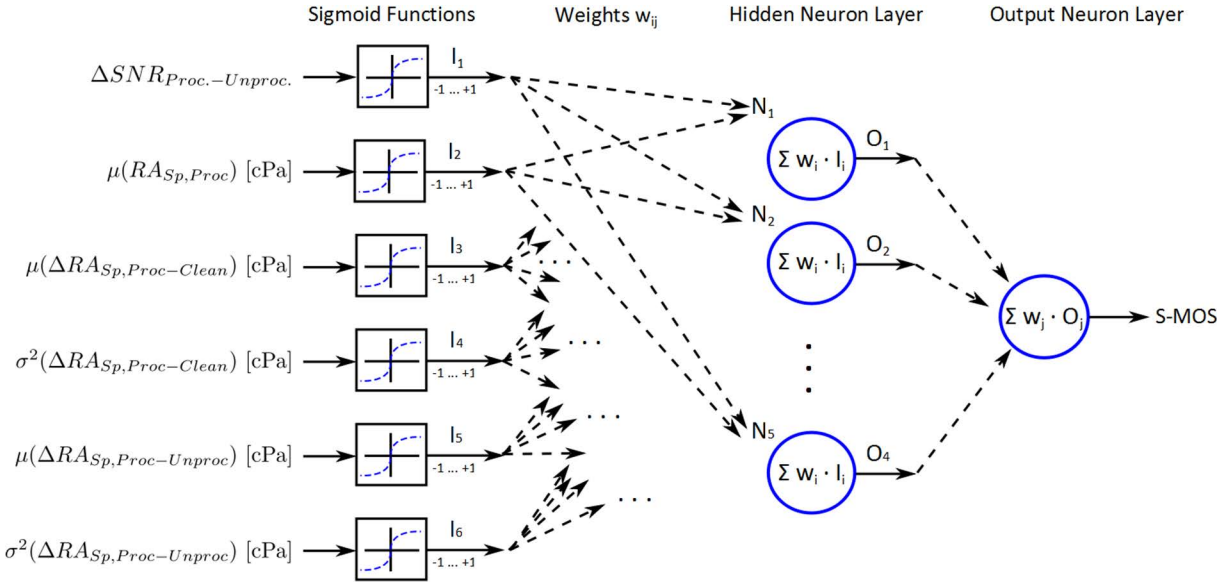


Figure 6.10: Structure of neural network for S-MOS

The setup of the neural network is shown in figure 6.10. It consists of 5 units (also known as "node" or "neuron") in one hidden layer; 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.

$$P = (-1 \quad P_1 \quad P_2 \quad P_3 \quad P_4 \quad P_5 \quad P_6) \tag{6.10}$$

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.

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

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:

$$M_{in,WB} = (0 \quad 11.2059 \quad 3.5049 \quad -1.4115 \quad 0.90054 \quad 13.1402 \quad 13.2832) \tag{6.12}$$

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:

$$S_{in,WB} = (1 \quad 10.5212 \quad 1.3348 \quad 1.1011 \quad 0.83575 \quad 5.4454 \quad 10.2952) \tag{6.13}$$

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 $\tilde{P}$:

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

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 $\tilde{P}\tilde{P}$ 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 five transformed output values of the hidden layer are then passed 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:

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

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

6.6.3 Comparing Subjective and Objective S-MOS Results

In contrast to the N-MOS calculation, the contributing parameters ΔSNR, $\mu(RA_{Sp,\ P})$, $\sigma^2(\Delta RA_{Sp,\ P\text{-}U})$, $\sigma^2(\Delta RA_{Sp,\ P\text{-}C})$, $\mu(\Delta RA_{Sp,\ P\text{-}U})$ and $\mu(\Delta RA_{Sp,\ P\text{-}C})$ are evaluated per sample with the neural network described above. The six S-MOS values for each samples are then averaged to the per-condition result.

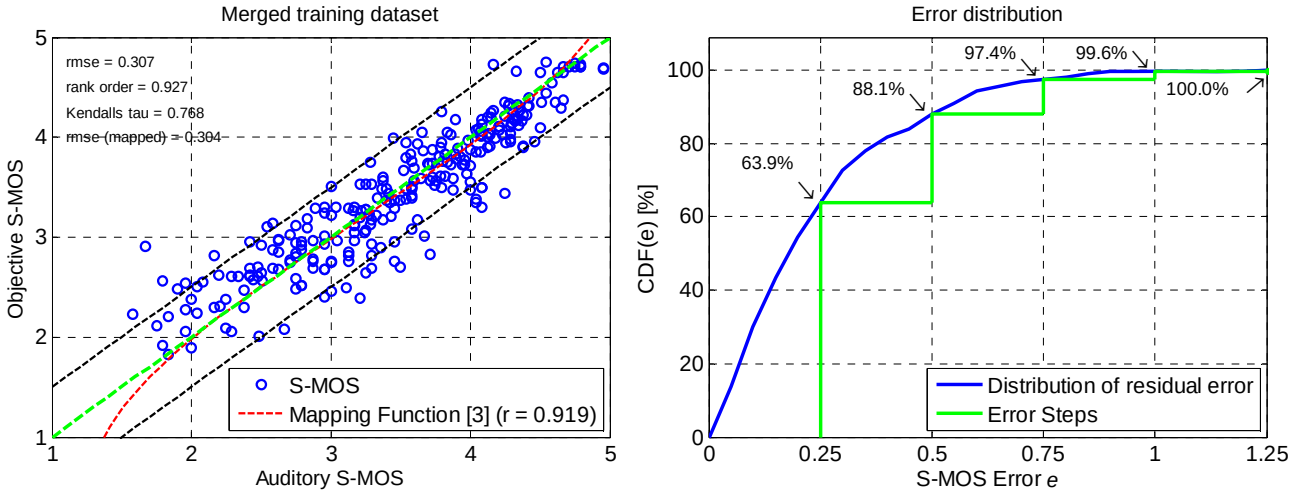


Figure 6.11: Left: Objectively calculated S-MOS versus auditory S-MOS; Right: CDF of residual error versus S-MOS Error e

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.

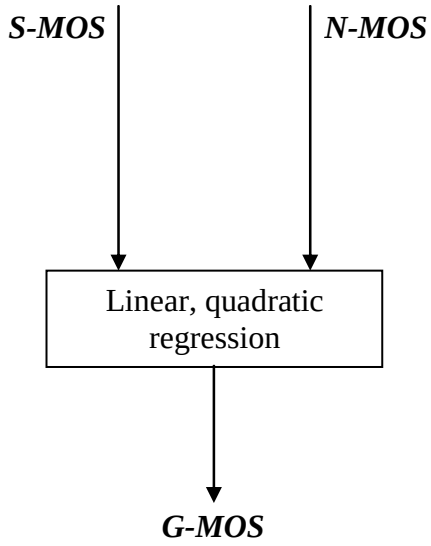


Figure 6.12: Block diagram of G-MOS calculation algorithm

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.

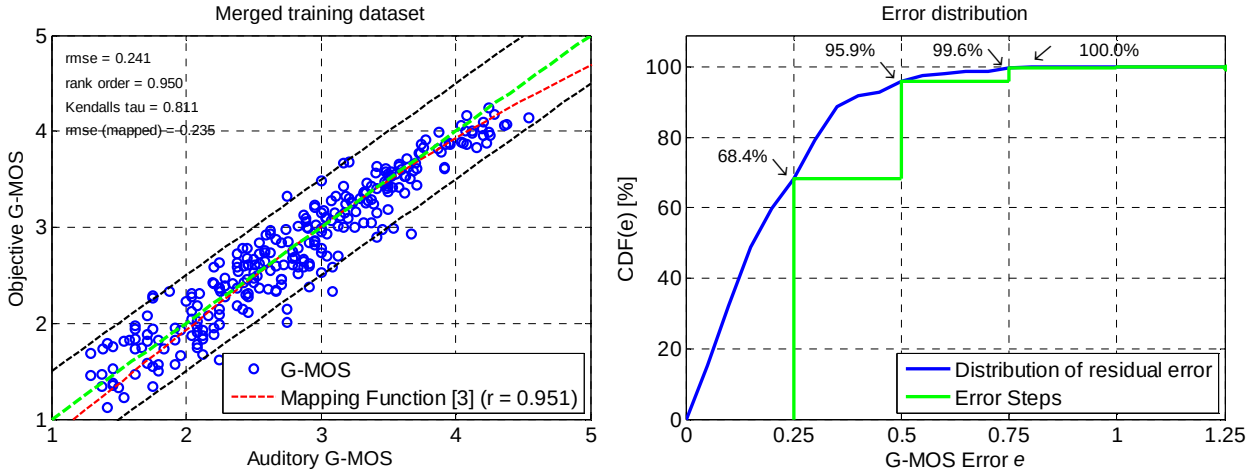


Figure 6.13: Left: Objectively calculated G-MOS versus auditory G-MOS; Right: CDF of residual error versus G-MOS Error e

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 5.3 and J.1, 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:

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

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:

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

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) [i.34] reported somewhat better performance of the model than in this present document. 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 present document, 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.

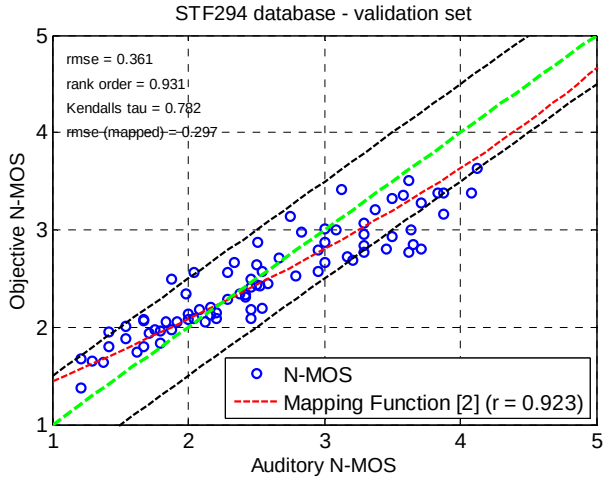


Figure 7.1: Left: Objectively calculated N-MOS vs. auditory N-MOS for validation conditions

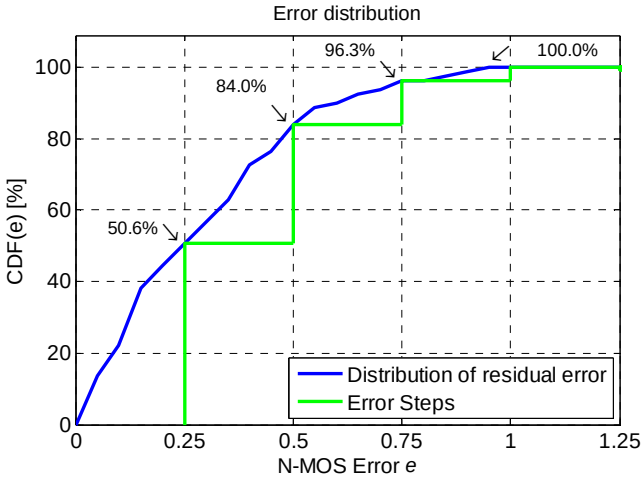


Figure 7.2: Objectively CDF of residual error vs. N-MOS Error e for validation conditions

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.

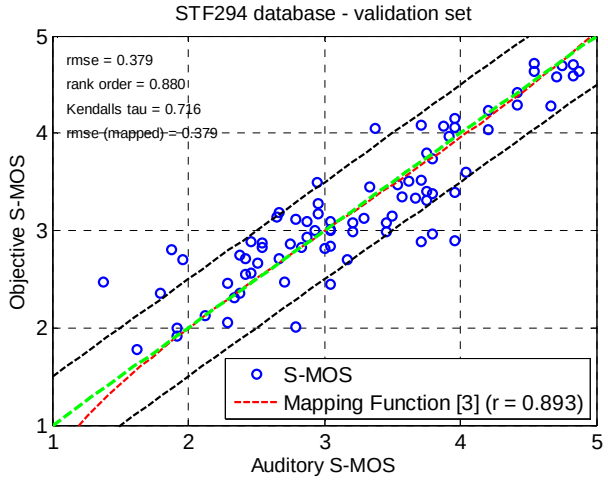


Figure 7.3: Left: Objectively calculated S-MOS vs. auditory S-MOS for validation conditions

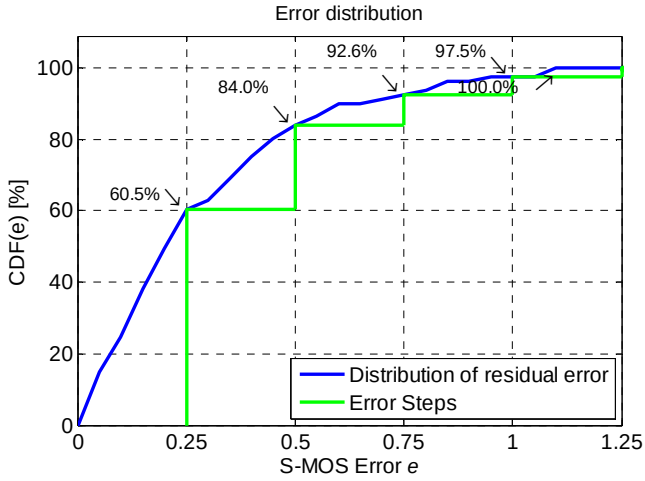


Figure 7.4: Objectively CDF of residual error vs. S-MOS Error e for validation conditions

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.

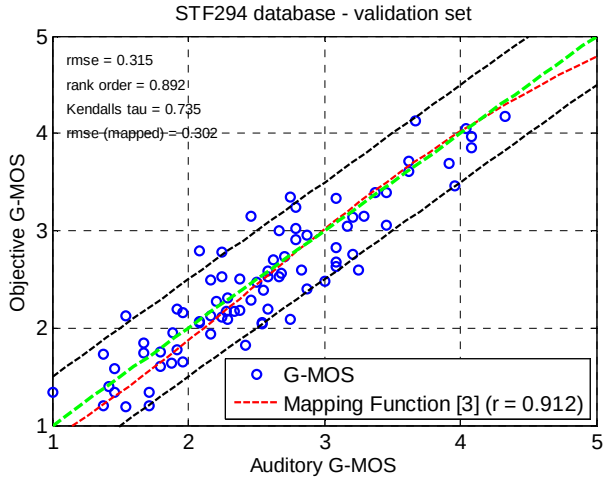


Figure 7.5: Left: Objectively calculated G-MOS vs. auditory G-MOS for validation conditions

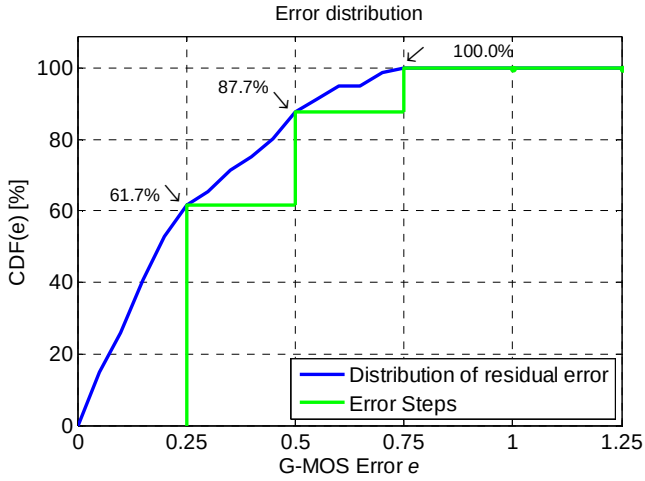


Figure 7.6: Objectively CDF of residual error vs. G-MOS Error e for validation conditions

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.

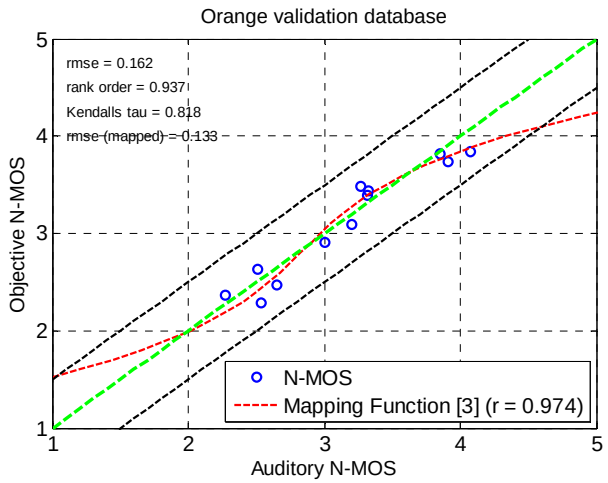


Figure 7.7: Left: Objectively calculated N-MOS vs. auditory N-MOS for validation conditions

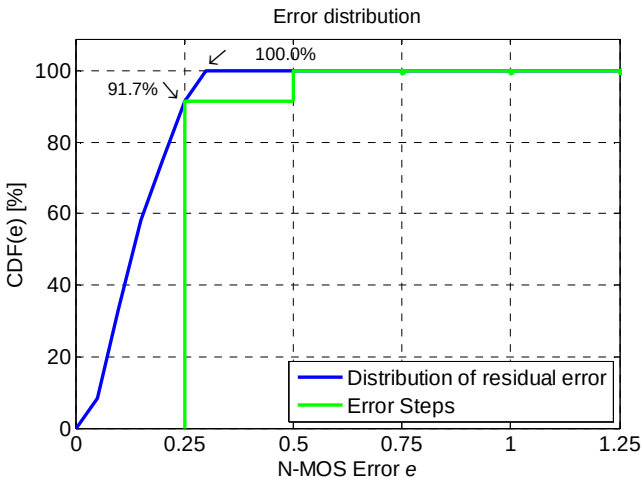


Figure 7.8: Objectively CDF of residual error vs. N-MOS Error e for validation conditions

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.

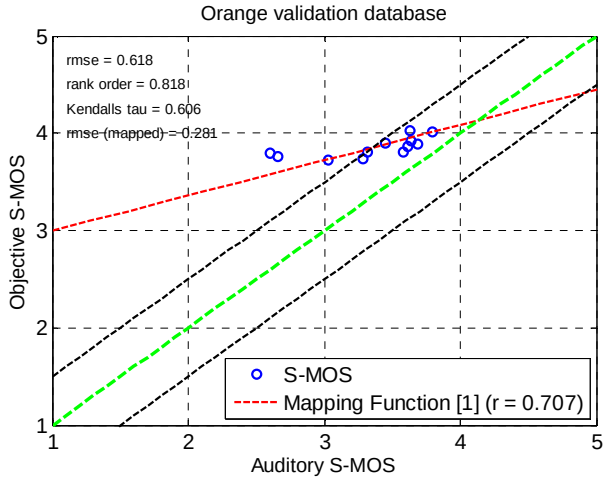


Figure 7.9: Left: Objectively calculated S-MOS vs. auditory S-MOS for validation conditions

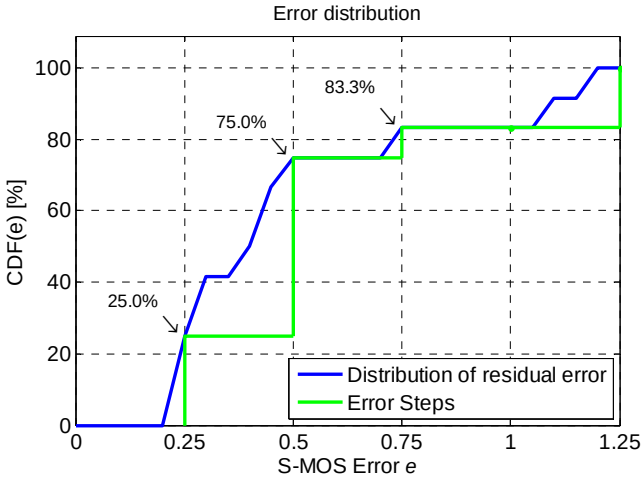


Figure 7.10: Objectively CDF of residual error vs. S-MOS Error e for validation conditions

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.

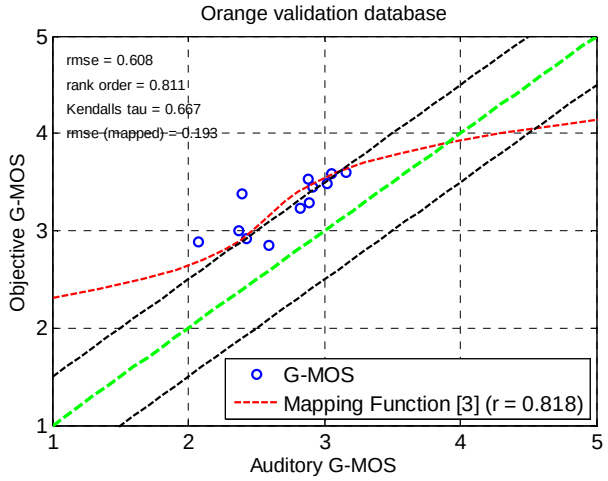


Figure 7.11: Left: Objectively calculated G-MOS vs. auditory G-MOS for validation conditions

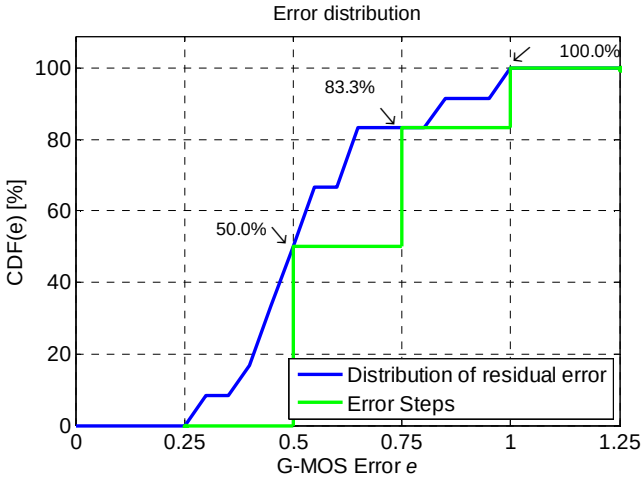


Figure 7.12: Objectively CDF of residual error vs. G-MOS Error e for validation conditions

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.

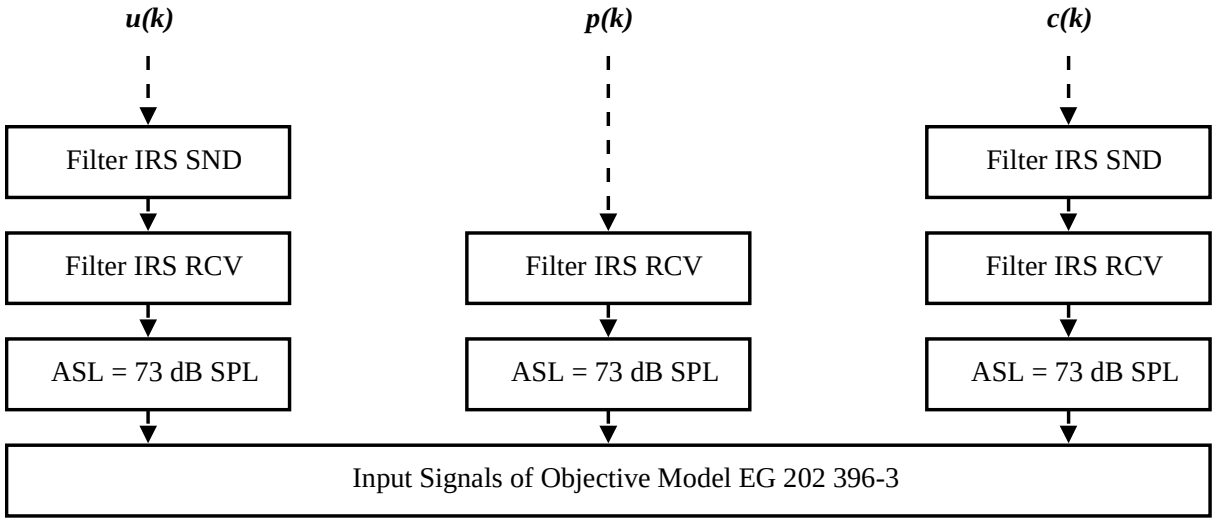


Figure 8.1: Modified pre-processing for narrowband mode

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).

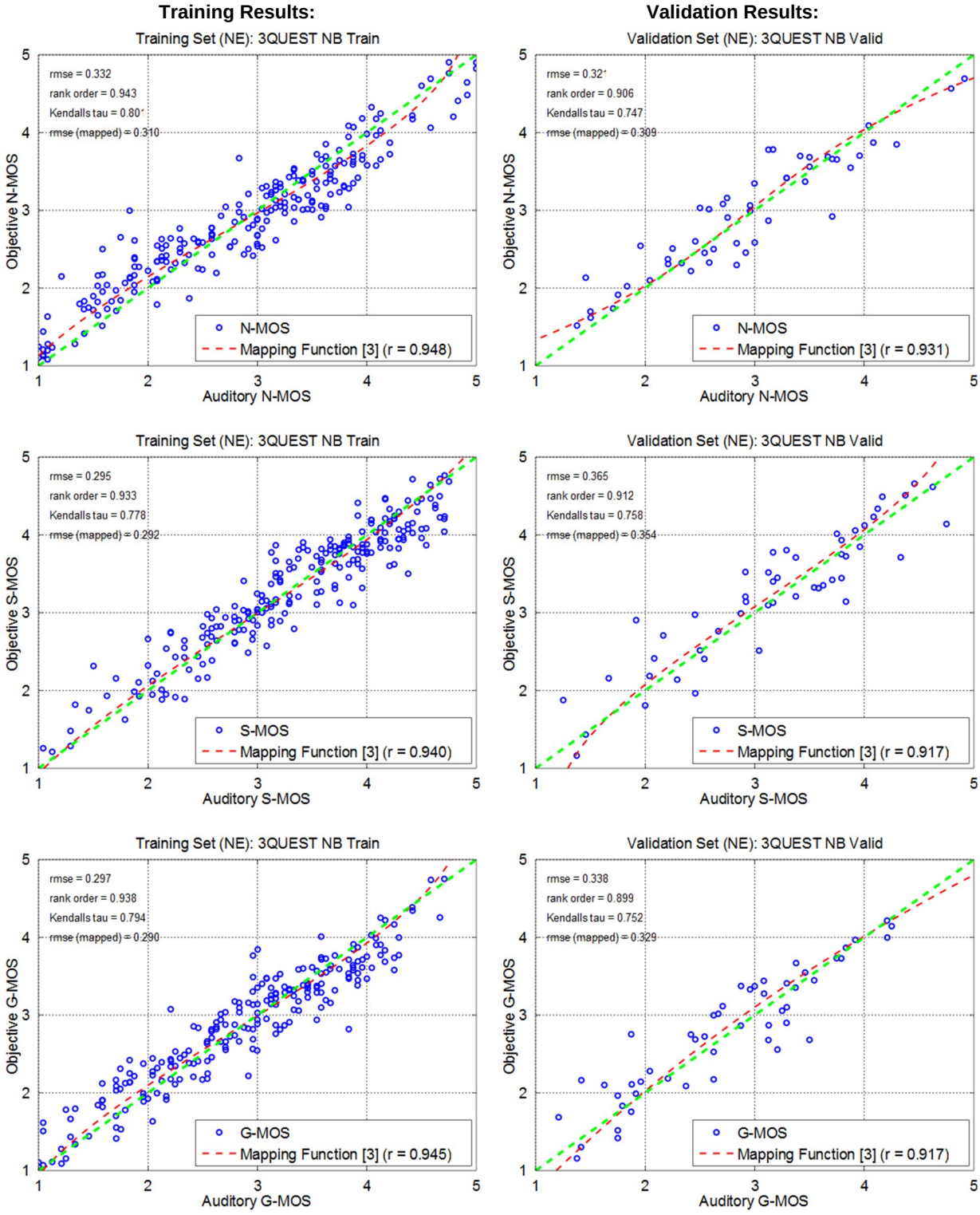


Figure 8.2: HEAD acoustics NB Database - Comparison subjective versus Objective data

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 2 and 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