ETSI TR 126 931 V16.0.0 (2020-09)

Universal Mobile Telecommunications System (UMTS);
LTE;
Evaluation of Additional Acoustic Tests for Speech Telephony
(3GPP TR 26.931 version 16.0.0 Release 16)

Reference
RTR/TSGS-0426931vg00

Keywords
LTE,UMTS

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

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.

Legal Notice

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

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.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

Introduction

Mobile telephony devices and voice services continue to develop and evolve and their associated minimum performance requirements and test methodologies also need to stay relevant and representative of quality demands.

While many advances were made in Release 11 to the acoustic requirements and test specifications in TS 26.131 [13] and TS 26.132 [8] many items therein were left marked "for further study" and require a final decision by SA4. Additionally, there are new acoustic requirements and emerging tests worth considering in a future release, but that require further study before incorporation into specifications.

This technical report will, first and foremost, address the remaining items presently designated as "for further study" in TS 26.131 [13] and TS 26.132 [8].

The present document will also examine opportunities for new acoustic tests and requirements that help us to better characterize the UE acoustic experience, opportunities to replace existing test methods with others that are more accurate or more efficient and make specific recommendations for their inclusion in existing or new specifications.

1 Scope

The scope of the present document is to investigate, first and foremost, the existing items presently designated as "for further study" in TS 26.131 [13] and TS 26.132 [8].

The investigation will additionally identify, examine and evaluate opportunities for new acoustic tests and requirements that better help characterize the UE acoustic experience, opportunities to replace existing test methods with others that are more accurate or more efficient and to make specific recommendations for their inclusion in existing or new specifications.

While many advances were made in Release 11 to the acoustic requirements and test specifications in TS 26.131 [13] and TS 26.132 [8] many items therein were left marked "for further study" and require a final disposition by SA4 including:

- NB & WB Stability loss, Headset UE (TS 26.131 [13], subclauses 5.6 & 6.6).
- NB & WB Delay, Wireless Headset (TS 26.131 [13], subclauses 5.12.2.2 & 6.11.2.2).
- NB & WB Echo control ("double-talk") characteristics (TS 26.131 [13], subclauses 5.13 & 6.12, TS 26.132 [8], subclause 8.11).
- Handset, Headset, Handheld hands-free, Desktop and vehicle mounted hands-free are all marked FFS.
- NB& WB Free-field measurements for vehicle-mounted hands-free (TS 26.132 [8], subclauses 7.2.3 & 8.2.3).
- NB & WB Idle Channel Noise, Sending/Receiving of test signal (TS 26.132 [8], subclauses 7.3.1, 7.3.2, 8.3.1 & 8.3.2).

Additionally, there are new acoustic requirements and emerging tests that may be considered in a future release, but require further study before incorporation to our specifications. It has been anticipated that topics in this area would include, but would not be limited to, an evaluation of:

- Time-variant user behaviour.
- Additional UE usage environments.
- New or refined test methods for existing requirements.
- Acceptance of updates (if any) to existing ETSI and ITU-T dependencies.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] Recommendation ITU-T P.862 (02/2001): "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs".

[2] Recommendation ITU-T P.800 (08/1996): "Methods for subjective determination of transmission quality".

[3] Recommendation ITU-T P.862.3 (11/2007): "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2".

[4] Recommendation ITU-T P.835 (11/2003): "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[5] ETSI TS 103 106 (V1.3.1): "Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals - objective test methods".

[6] Recommendation ITU-T P.863 (09/2014): "Perceptual objective listening quality assessment".

[7] ETSI ES 202 396-1 (V1.4.1): "Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database".

[8] 3GPP TS 26.132: "Technical Specification Group Services and System Aspects; Speech and video telephony terminal acoustic test specification".

[9] Recommendation ITU-T P.862.1 (11/2003): "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[10] Recommendation ITU-T P.862.2 (11/2007): "Wideband extensions to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs".

[11] ETSI EG 202 396-3 (V1.5.1): "Speech quality performance in the presence of background noise; Part 3: Background noise transmission - Objective test methods".

[12] Recommendation ITU-T P.64, Annex E (11/2007): "Determination of sensitivity/frequency characteristics of local telephone systems".

[13] 3GPP TS 26.131: "Technical Specification Group Services and System Aspects; Terminal acoustic characteristics for telephony".

[14] ETSI TS 103 224 (V1.3.1): "Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database".

[15] Recommendation ITU-T P.58: "Head and Torso simulator for Telephonometry".

[16] Recommendation ITU-T P.502 (05/2000): "Objective test methods for speech communication systems using complex test signals".

[17] Recommendation ITU-T P.863.1 (09/2014): "Application guide for Recommendation ITU-T P.863".

[18] Recommendation ITU-T P.1401 (07/2012): "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models".

[19] Recommendation ITU-T P. 79 (11/2007): "Calculation of loudness ratings for telephone sets".

[20] ETSI TS 103 281 (V1.1.1): "Speech quality in the presence of background noise: objective test methods for SWB and FB terminals".

[21] Recommendation ITU-T P.501: "Test signals for use in telephonometry".

[22] ETSI TS 103 557 (V1.1.1): "Methods for reproducing reverberation for communication device measurements".

[23] Recommendation ITU-T P.862 (2001) Corrigendum 1 (10/2017).

[24] Recommendation ITU-T P.831: "Subjective performance evaluation of network echo cancellers" (12/1998).

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [1] and the following apply.
A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

3.2 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply.
An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

HHS Hand-held speakerphone (= Hand-held hands-free)

4 Release 11 "For Further Study" Items

4.1 Stability loss, Headset UE

Void

4.2 UE Delay, NB & WB Wireless Headset

Void

4.3 NB & WB Echo control ("double-talk") characteristics

4.3.1 Results from a study on NB Echo control ("double-talk") characteristics using P.835 methodology

4.3.1.1 Background

In Release 11 of TS 26.132 [8], new methods for evaluation of echo control characteristics were introduced in Clauses 7.11 and 8.11. However, corresponding requirements were not defined in TS 26.131 [13].

A subjective listening test based on methods from Recommendation ITU-T P.835 [4] was conducted in order to provide some data for purposes of investigating possible requirements.

Instead of a conversational test, or talking and listening test, the present document provides results from listening only test, so participants did not experience echo while talking, only while passively listening. Below are presented results of the subjective evaluation of real speech double talk test, for 12 devices, in both handset and handheld speakerphone for narrow band.

4.3.1.2 Test Method & Results

4.3.1.2.1 Methods

The categories defined in Clauses 7.11, Figure 17b5, (copied below in Figure 1 for convenience) and Table 1 (Table 2c, and 8.11, Figure 19b5, and Table 2g), are described in perceptually-relevant terms.

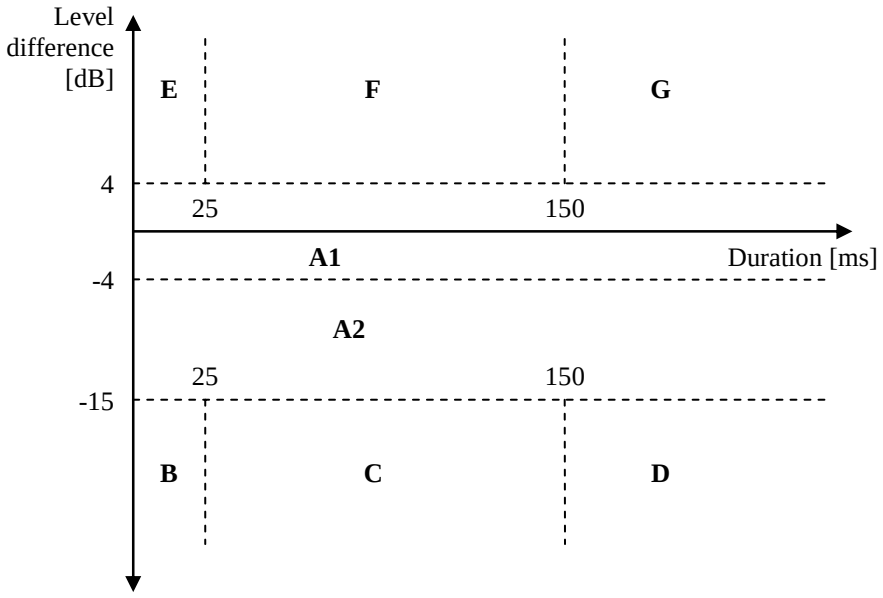


Figure 1: Classification of echo canceller performance

Table 1: Categories for echo canceller performance classification

Category	Level difference (ΔL)	Duration (D)	Description
A1	-4 dB ≤ ΔL < 4 dB		Full-duplex and full transparency
A2	-15 dB ≤ ΔL < -4 dB		Full-duplex with level loss in Tx
B	ΔL < -15 dB	D < 25 ms	Very short clipping
C	ΔL < -15 dB	25 ms ≤ D < 150 ms	Short clipping resulting in loss of syllables
D	ΔL < -15 dB	D ≥ 150 ms	Clipping resulting in loss of words
E	ΔL ≥ 4 dB	D < 25 ms	Very short residual echo
F	ΔL ≥ 4 dB	25 ms ≤ D < 150 ms	Echo bursts
G	ΔL ≥ 4 dB	D ≥ 150 ms	Continuous echo

4.3.1.2.1.1 Rating scales

The impairments in categories A2, B, C, and D can be understood as distortions of the uplink speech. In contrast, the impairments in categories E, F, and G can be understood as intrusions of residual or continuous echo. Based on these observations, the rating scales of P.835 [4], SIG, BAK, and OVRL, as shown in Figures 2 below, were adopted for this listening evaluation. In this study, the BAK rating scale was used to quantify the level of intrusiveness of any echo. In other P.835 studies, the BAK rating scale has been more typically used to quantify the level of intrusiveness of background noise.

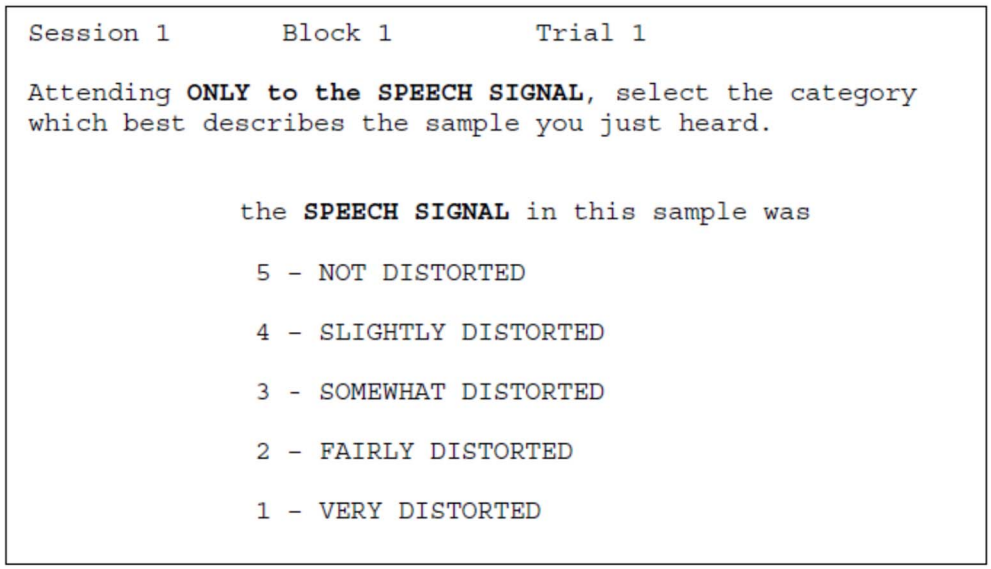


Figure 2a (Figure 5/P.835): Speech signal rating scale

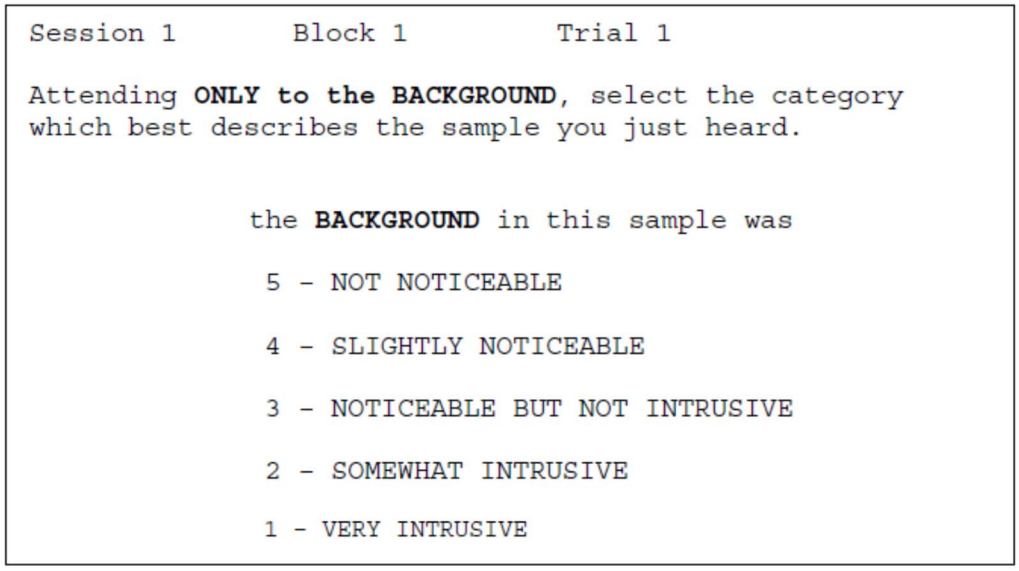


Figure 2b (Figure 6/P.835): Background rating scale

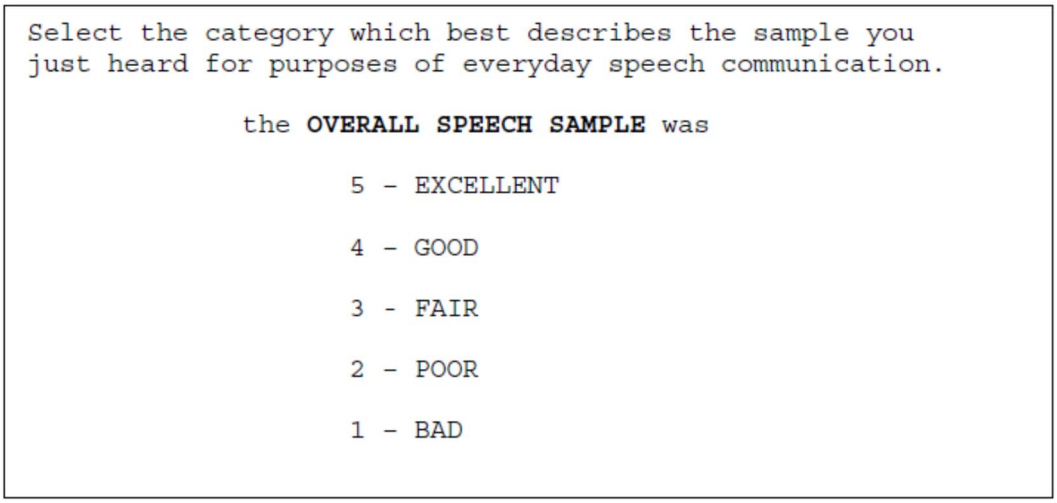


Figure 2c (Figure 7/P.835): Overall quality rating scale

It was anticipated that the impairments in categories A2, B, C and D, would be related to ratings on the SIG (speech distortion) scale, and that impairments in categories E, F, and G would be related to the ratings on the BAK background intrusiveness scale.

4.3.1.2.1.2 Speech Source

The speech source used is Segment 2 (four sentences) of the current double talk test, British English from Recommendation ITU-T P.501 [21]. This includes two male and two female talkers. While this is rather limited in comparison to some subjective tests, the exact signal and conditions are used to facilitate direct comparisons with the objective measures. For each presentation and rating, a single sentence was presented (total of four sentences, from two male and two female speakers).

4.3.1.2.1.3 Reference Signals

For the SIG dimension, the Wiener-filter based reference system proposed in [5] and used in [6] was used. While this reference system has been primarily used as a reference for noise suppression, as many echo control systems provide echo suppression using a multi-band attenuation mechanism, it seems reasonable to use that reference system in this context. Expert listening to the reference system and the distortions introduced by the devices exhibiting higher levels of impairments in the A2, B, C and D categories indicated qualitatively similar perceptions. Four levels of Wiener-filter-based distortion, similar to those used in P.835 tests for noise suppression and judged by expert listeners to span the range from 1 to 5, were used.

For the BAK dimension, recordings of echo were made on a device with the ability to disable the AEC system, and to capture signals at the microphone. To acquire the BAK echo component, the signal defined in TS 26.132 [8], Clause 7.11.1, for use in the receiving direction was injected into the input of a network simulator. The resulting echo signal from the device was recorded at the microphone of the device. The level of the receiving signal was adjusted to yield a range of echo levels. To construct the BAK references, the speech and echo signals were mixed at a range of Speech to Echo Ratios (SER): 0, 12, 24 and 36 dB SER. It is noted, in Figure 5, that the resulting range of BAK scales using these SER values is somewhat compressed to below 2.5 MOS. Use of an alternative range of SER (e.g. 0, 24, 36, 48 dB) may provide a more uniform range of BAK scores.

Three additional reference signals consisting of a combination of Wiener-filter distortion and echo, with increasing levels of both, were also constructed, resulting in a total of 11 impaired references (4 Wiener-filter, 4 echo-only, 3 Wiener-filter and echo). Clean speech was used as the twelfth reference signal.

4.3.1.2.1.4 Listening mode and level

Presentation was made monaurally, at 79 dB SPL, using closed-back Sennheiser HD-280 Pro headphones, without any additional equalization.

4.3.1.2.1.5 Listening Panel

Results are reported for a listening panel that consisted of 32 naïve listeners, native speakers of American English, all with self-reported normal hearing.

Listeners were presented with a practice block of signals including references for familiarization with the task. The rating scales described in Clause 4.3.1.2.1.1 were used.

4.3.1.2.1.6 Measurement Set up

As noted above, the measurements were taken corresponding to the method defined in Clause 7.11 of TS 26.132 [8]. To clarify the relationships between input signals and measurements used in the listening-only test, schematic depictions of the input signal paths and measurements are provided below, with the single talk configuration shown in Figure 3a and the double-talk configuration shown in Figure 3b.

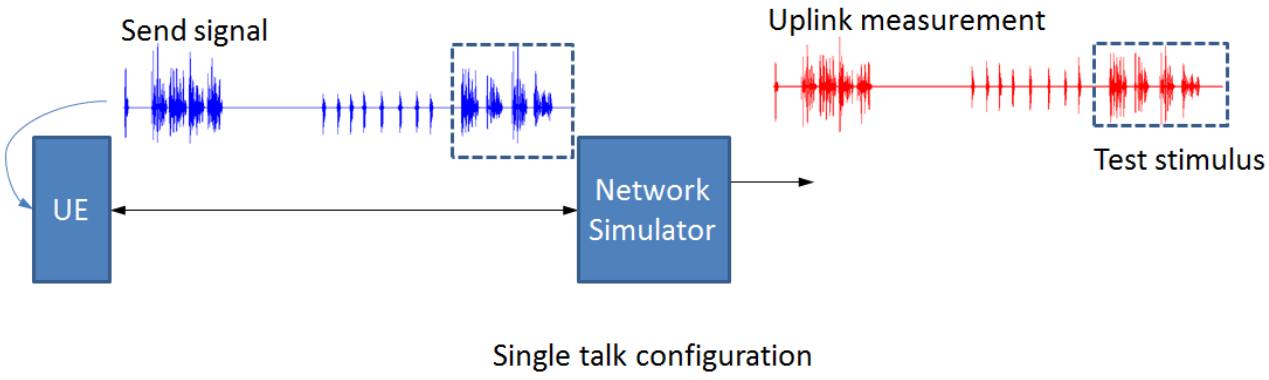


Figure 3a: Signal paths for single talk configuration

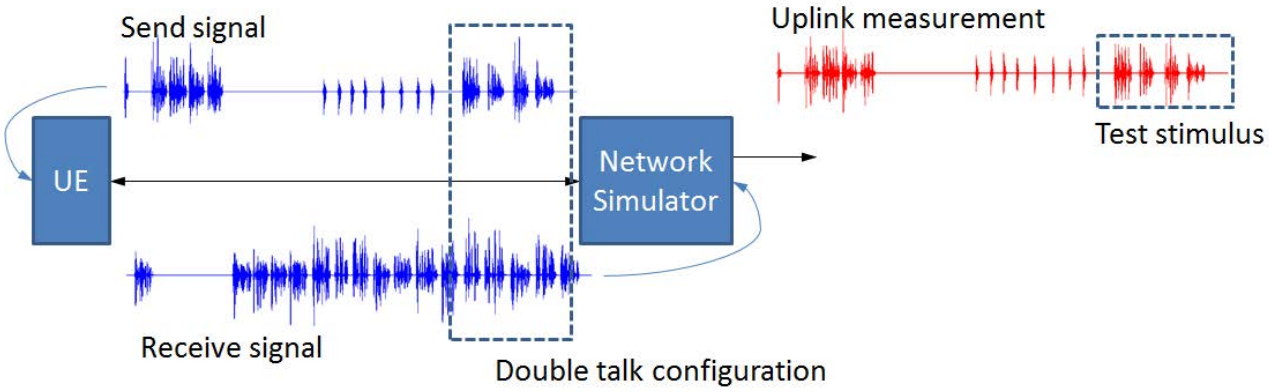


Figure 3b: Signal paths for double talk configuration

In the double talk configuration, there is a double talk situation at the terminal, as indicated by the dotted line enclosing the signals in sending and receiving direction (note that the methods in TS 26.132 [8] include provisions to time-align the acoustic signals at the terminal).

In both Figures 3a and 3b, the dotted box enclosing a portion of the Uplink measurement indicates that portion of the test signal that is used in the listening-only experiments.

In the single talk configuration, there is no possibility of impairment of the uplink measurement/listening file due to the presence of the downlink. In contrast, in the double talk configuration, the signal in the receiving direction is present simultaneously with the signal in the sending direction. There is the possibility of impairment of the speech in the sending direction due the action of the echo canceller in the terminal. Also, there is the possibility of the presence of residual echo induced by the receiving signal in the uplink measurement and listening file.

4.3.1.2.1.7 Test Conditions

Twelve commercially-available devices were tested using the methods of TS 26.132 [8], Clause 7.11, in both handset and hand-held speakerphone modes to obtain test stimuli. Recordings of Segment 2 in both single-talk (ST) and double-talk (DT) conditions were collected and used for testing.

A total of 48 listening conditions, comprised of 12 UEs, 2 use cases (handset and hand-held speakerphone) and 2 echo conditions (single-talk and double-talk), were tested. A fully balanced design using all four sentences was used, resulting in 192 votes per condition.

4.3.1.2.2 Results

4.3.1.2.2.0 General

Results for the reference signals are shown in Figures 4, 5, and 6. The error bars indicated 95th percentile confidence intervals, based on 34 participants.

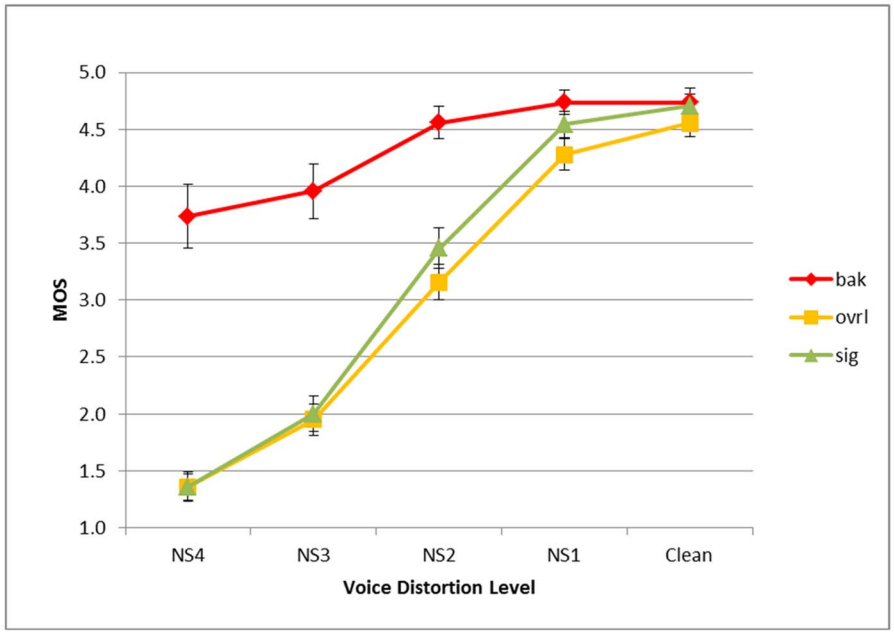


Figure 4: Speech distortion varies, no echo

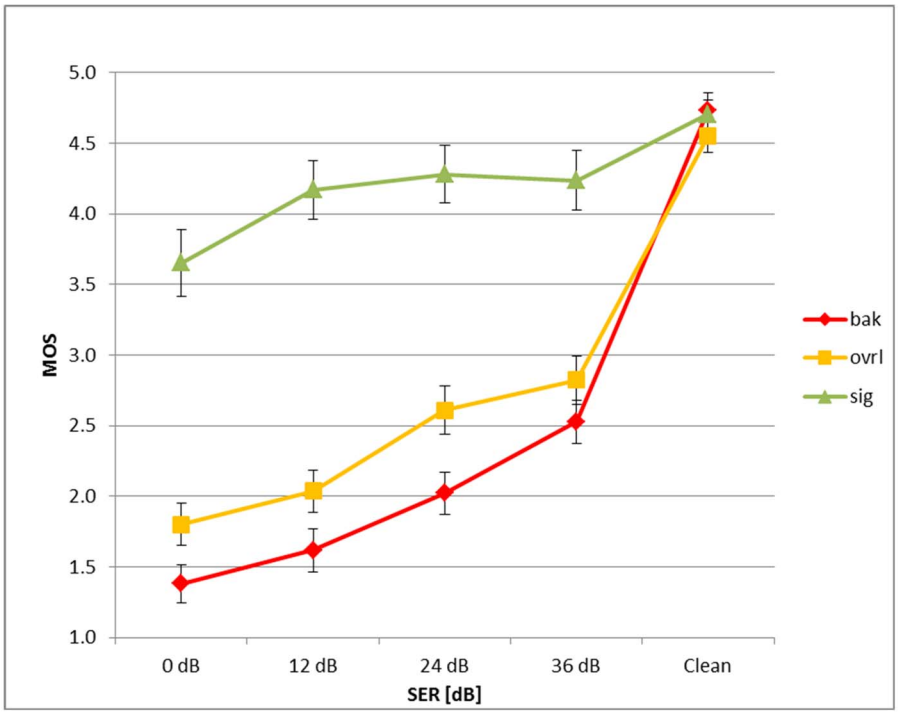


Figure 5: SER varies, no speech distortion

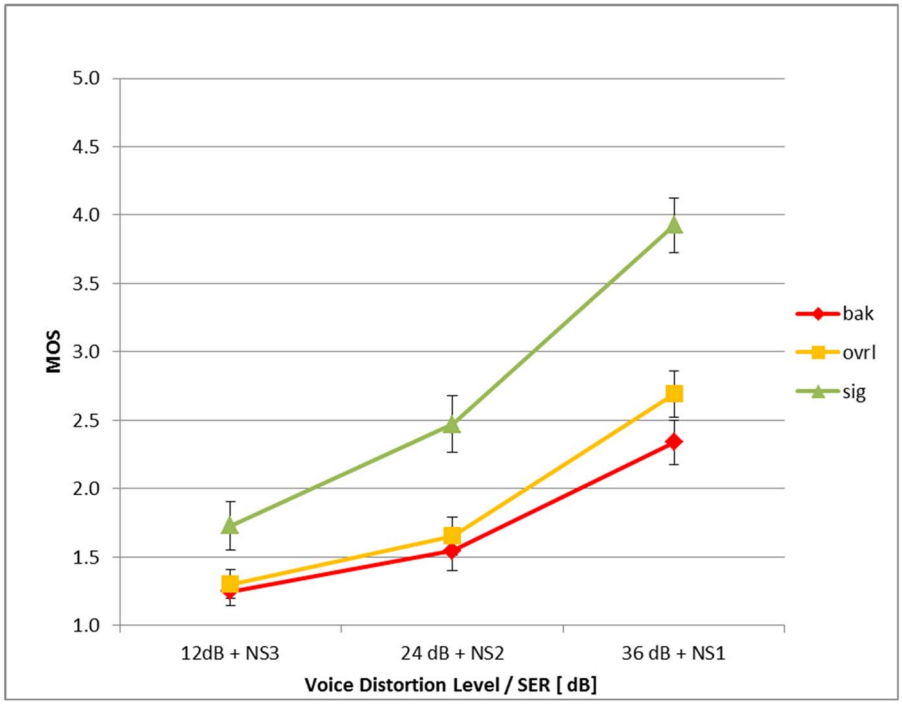


Figure 6: Speech distortion and SER vary

From Figures 4, 5, and 6, it can be seen that the selected reference systems result in listeners' using the full range of the scales, with good separation of the SIG and BAK scales for the NS and SER references respectively.

4.3.1.2.2.1 Results for test conditions, Handset mode

Figures 7, 8 and 9 show the ratings for 12 devices in handset, SIG, BAK, and OVRL, respectively. The error bars show 95% confidence intervals. Blue bars show single-talk (no echo) while red bars show ratings for double-talk (with echo). The test conditions are as defined in Clause 7.11.

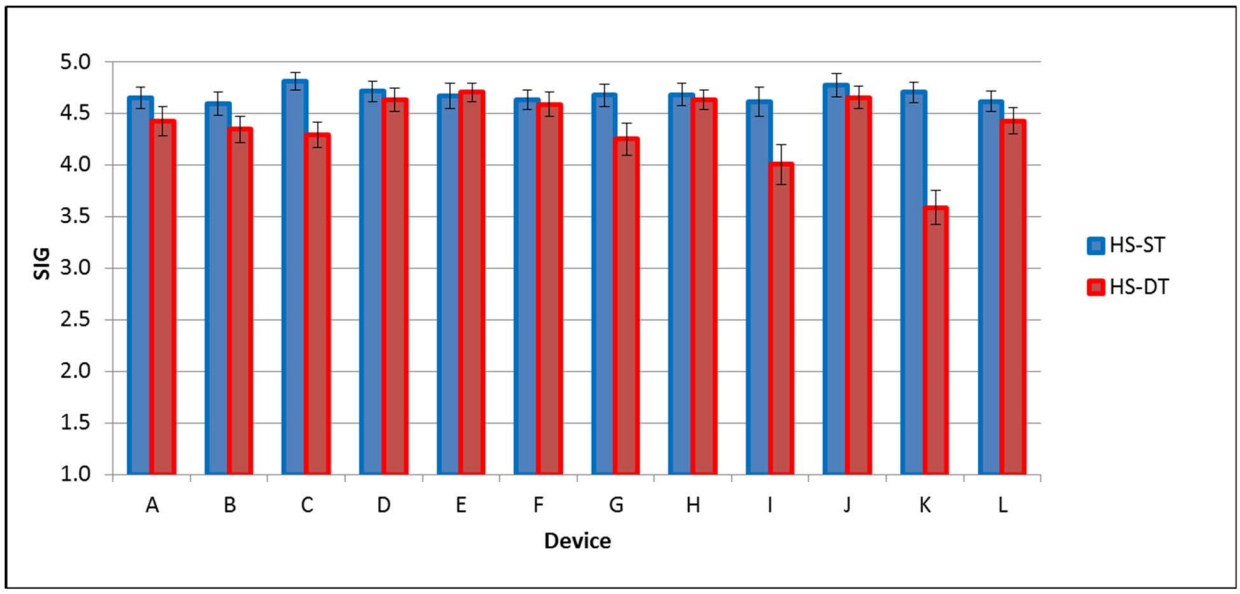


Figure 7: SIG ratings for Handset mode

The SIG ratings for all devices in ST are uniformly high, as might be expected. A few devices, I and K, show significant SIG degradations in double talk.

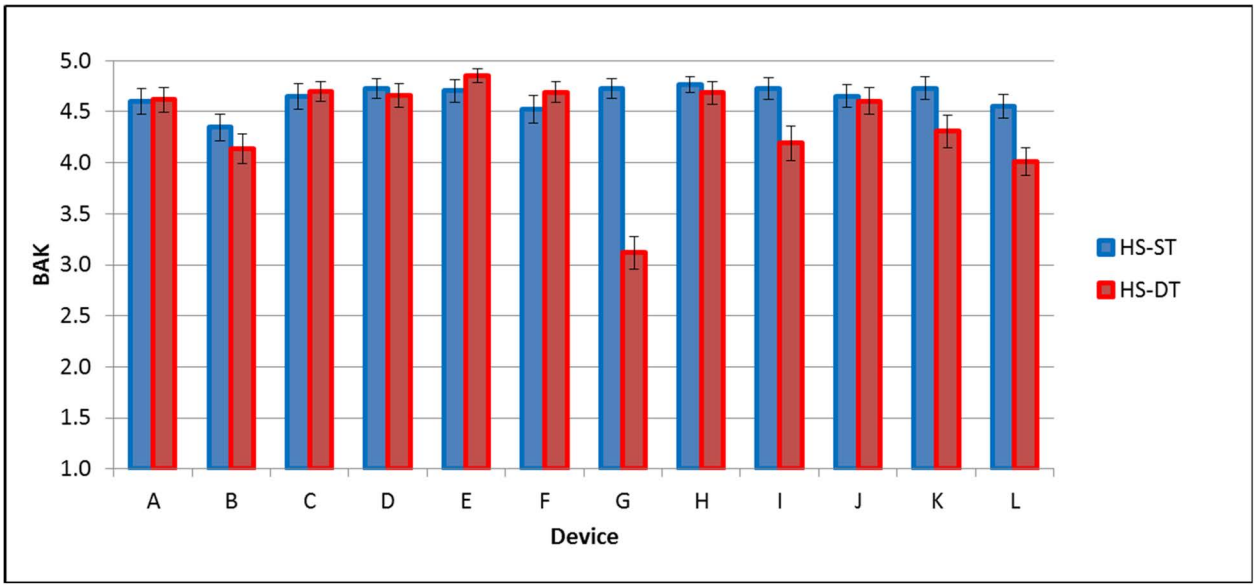


Figure 8: BAK ratings for Handset mode

The BAK ratings for all devices in ST are also uniformly high (except possibly device B). A few devices, G and L, show significant BAK degradations in double talk.

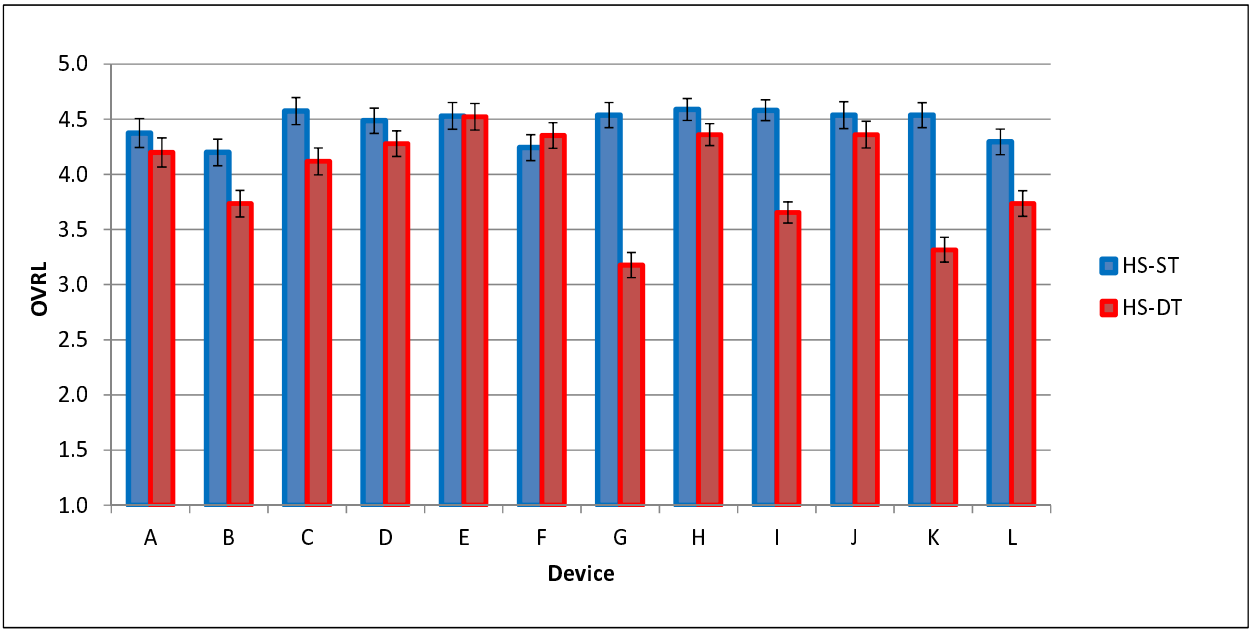


Figure 9: OVRL ratings for Handset mode

The OVRL ratings for all devices in ST are also uniformly high. A few devices, G, I, K, and L, show significant OVRL degradations in double talk, consistent with the results in Figures 7 and 8.

4.3.1.2.2.2 Results for test conditions, Hand-held Hands-free mode

Figures 10, 11 and 12 show the ratings for 12 devices in hand-held hands-free, SIG, BAK, and OVRL, respectively. The error bars show 95 % confidence intervals. Blue bars show single-talk (no echo) while red bars show ratings for double-talk (with echo).

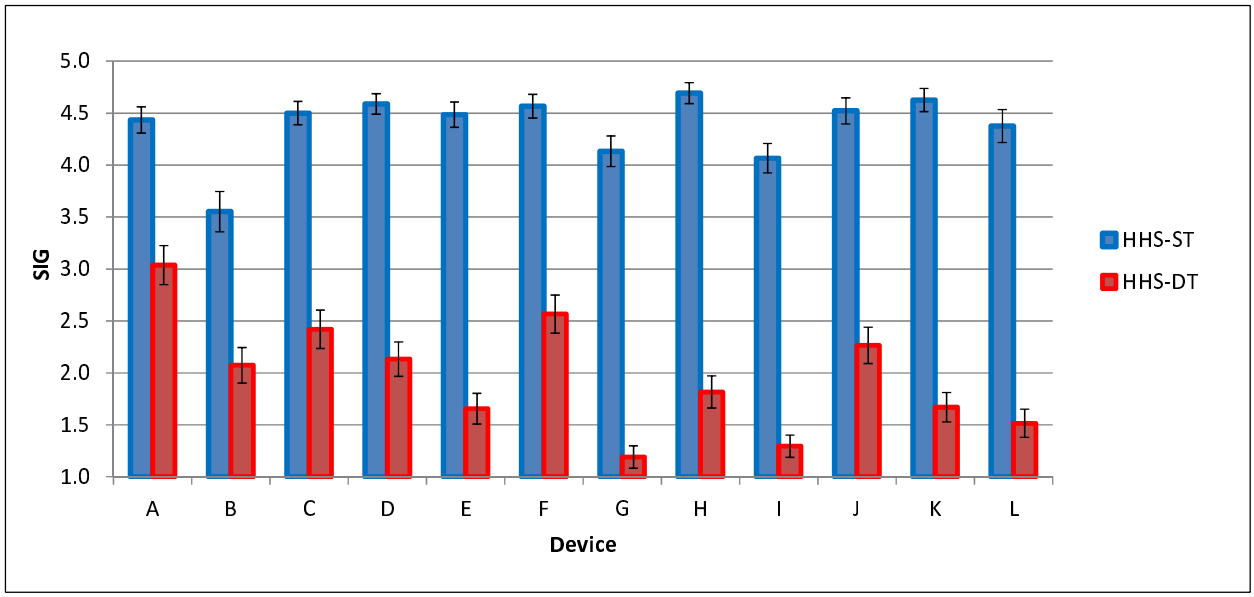


Figure 10: SIG ratings for Hand-held Hands-free mode

The SIG ratings for all devices in ST are generally high with the possible exception of devices B, G, and I. As might be expected, the SIG results in doubletalk for hand-held hands-free mode show substantial impairments, with a relatively large range from 3.0 (device A) to nearly 1.0 (Device G).

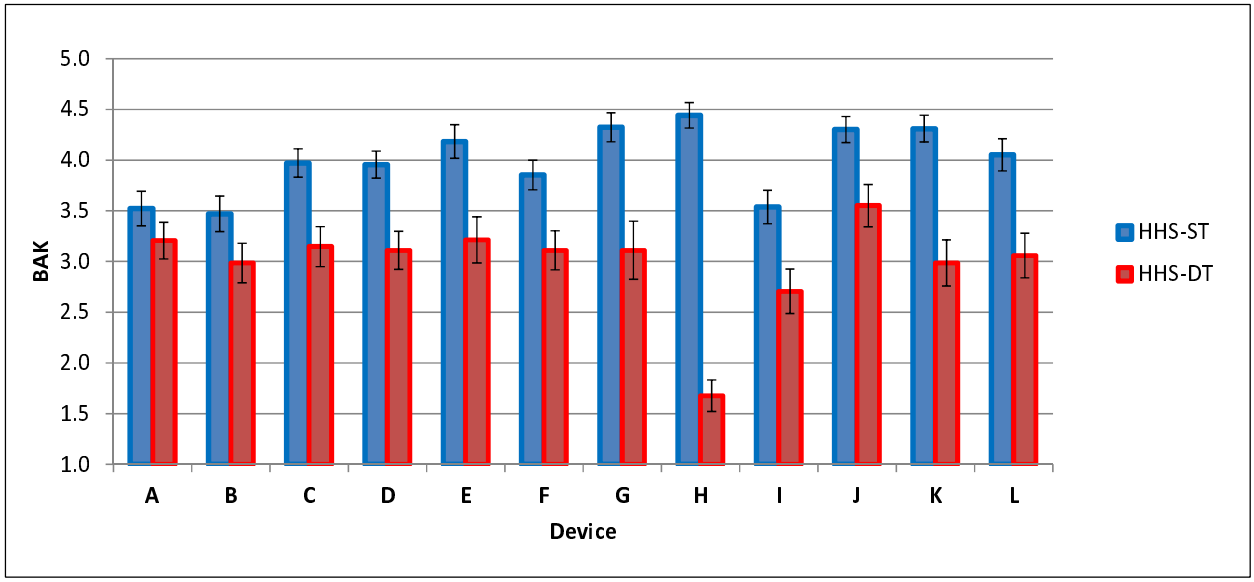


Figure 11: BAK ratings for Hand-held Hands-free mode

The BAK ratings for all devices in ST show some reduction over handset mode, with more devices showing performance below 4.0 (devices A, B, and I. For double talk, the BAK results are fairly consistent, above 3.0, with the exception of relatively poor performance of device H and relatively good performance of device J.

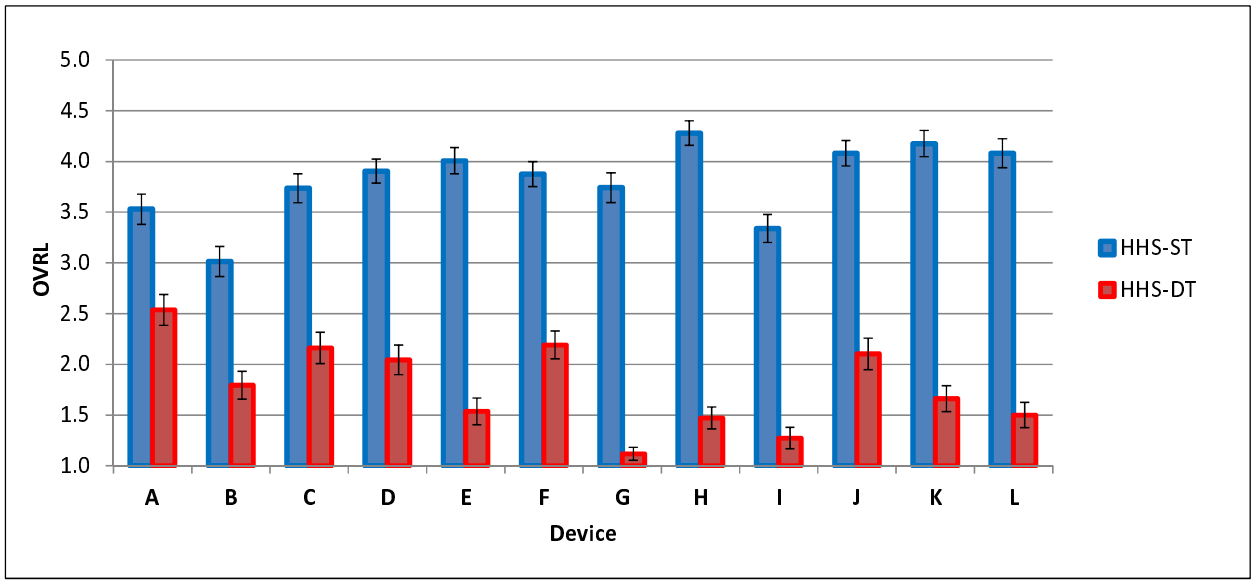


Figure 12: OVRL ratings for Handheld Hands-free mode

The OVRL ratings for all devices in ST show some reduction over handset mode, with more devices showing performance below 4.0 (devices A, B, G, and I). The OVRL results for double talk have more variability, driven primarily by the large variation in SIG scores.

The correlation between sending only tests based on recorded distortions and the perception in live conversation is not known (Recommendation ITU-T P.831 [24]). Hence, the correlation of the results from this listening test and the category classification needs further study.

4.3.2 Comparison of predictions from objective metrics to subjective double talk ratings

4.3.2.1 Background and metrics

In TS 26.132 [8], computational methods are provided to compute the categories and levels described in Clause 4.3.1.2.1 from real-speech recordings from terminals. In this clause, the eight categories and corresponding levels defined in Table 1 for echo canceller performance are compared to the subjective ratings reported in Clause 4.3.1.2.2. In addition to the categories and levels from TS 26.132 [8], two additional metrics are considered. The first is found in Recommendation ITU-T P.502, Amendment II [16], defining a method for computing the attenuation in the sending direction under double-talk conditions, $A_{s,DT}$. The second is found in Recommendation ITU-T P.863 [6]. While the application to impairments resulting from double-talk is not explicitly within scope of P.863, it has been used for this purpose in some instances.

4.3.2.1.1 Method for P.863 computations

For P.863 computations, guidance from P.863.1 [17] was followed. The four sentences of segment 2 of the double-talk test signal from TS 26.132 [8] were grouped sequentially into two sentence-pairs. The reference signal for the P.863 NB mode was the full-band source, filtered by a NB filter according to P.863.1 Table 3. The reference signal for the P.863 SWB mode was the full band source, filtered by a SWB filter according to P.863.1 Table 3. The validation tests defined in P.863.1 Clause 8.8 were passed, with both sentence-pair references scoring 4.50 for MOS-LQOn and 4.75 for MOS-LQOw. For each test condition, the P.863 scores MOS-LQOn and MOS-LQOw were computed using the uplink measurements from that test condition, grouped into two sentence-pairs. The MOS-LQOn and MOS-LQOw for the test condition are reported as the average of the scores for each of the two sentence-pairs from the measurements for that condition.

4.3.2.2 Comparison of metrics to subjective SIG ratings

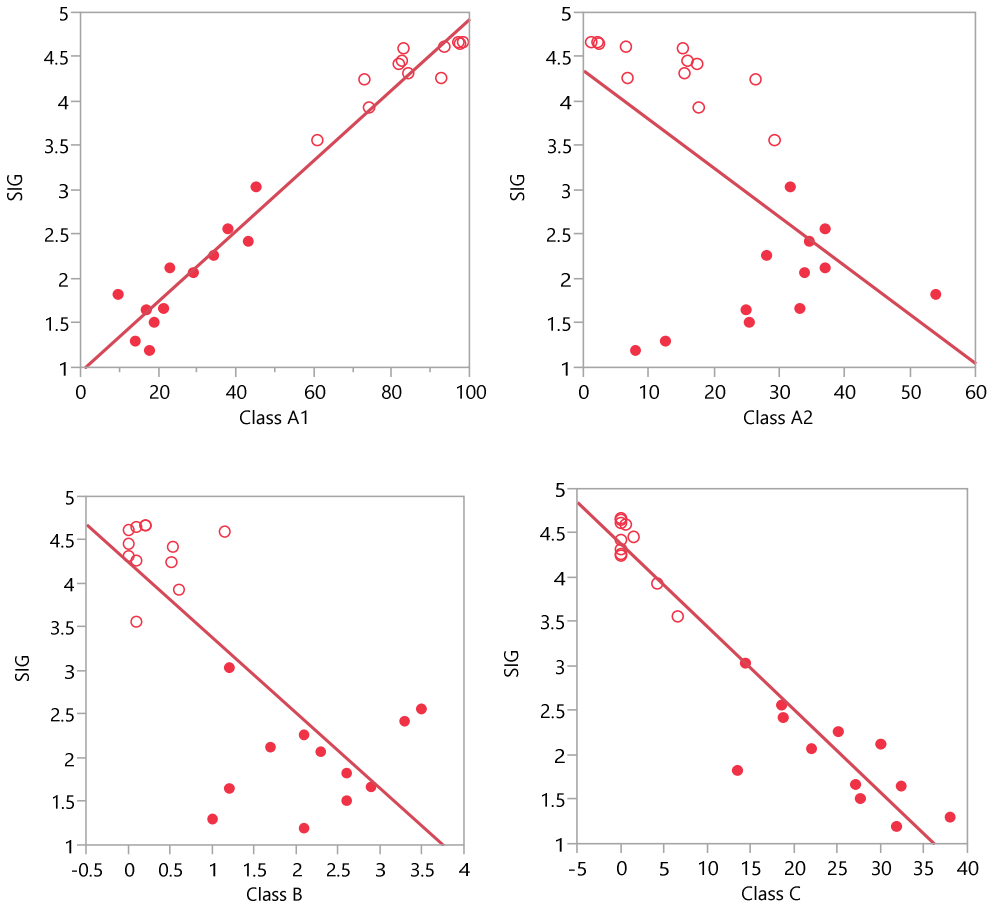
Table 2 contains summary values for the fit of the above-described metrics to the SIG DT ratings for both HS and HHHF combined.

Table 2: Summary of model fits for SIG DT

SIG			R^2	correl.	rmse	ANOVA F
3GPP	Atten Class	DT class A1	0,963	0,981	0,256	*<0,0001*
		DT class A2	0,329	0,574	1,091	0,0034
		DT class B	0,591	0,769	0,851	*<0,0001*
		DT class C	0,921	0,960	0,375	*<0,0001*
		DT class D	0,693	0,832	0,739	*<0,0001*
	Echo Class	DT class E	0,000	0,021	0,577	0,9216
		DT class F	0,153	0,391	1,226	0,0590
		DT class G	0,000	0,000	1,303	na
	Atten Level	DT level A1	0,680	0,825	0,754	*<0,0001*
		DT level A2	0,668	0,817	0,767	*<0,0001*
		DT level B	0,589	0,767	0,854	*<0,0001*
		DT level C	0,718	0,847	0,797	*<0,0001*
		DT level D	0,811	0,901	0,580	*<0,0001*
	Echo Level	DT level E	0,000	0,004	1,332	0,9837
		DT level F	0,199	0,446	1,192	0,0289
		DT level G	0,000	0,000	1,303	na
P.502		$A_{s,DT}$	0,812	0,901	0,578	*<0,0001*
P.863		MOS-LQOn	0,989	0,994	0,189	*<0,0001*
		MOS-LQOw	0,848	0,921	0,518	*<0,0001*

Table 2 reports the R^2, correlation, rmse, and ANOVA results for the Class (% frames) and Level (dB atten) according to the 3GPP analysis, the Sending attenuation analysis $A_{s,DT}$ of updated Appendix III of P.502, and both MOS-LQOn and MOS-LQOw according to P.863.

These results show that the best single predictor of the SIG results for both HS and HHHF taken together is the P.863 MOS-LQOn (correlation 0,994, rmse 0,189) followed by the 3GPP DT Class A1 (correlation 0,981, rmse 0,256). Scatter plots are shown in Figures 13, 14, 15 and 16, where filled symbols are for HHHF and open symbols are for HS.


SIG
Class A1
Class A2
Class B
Class C

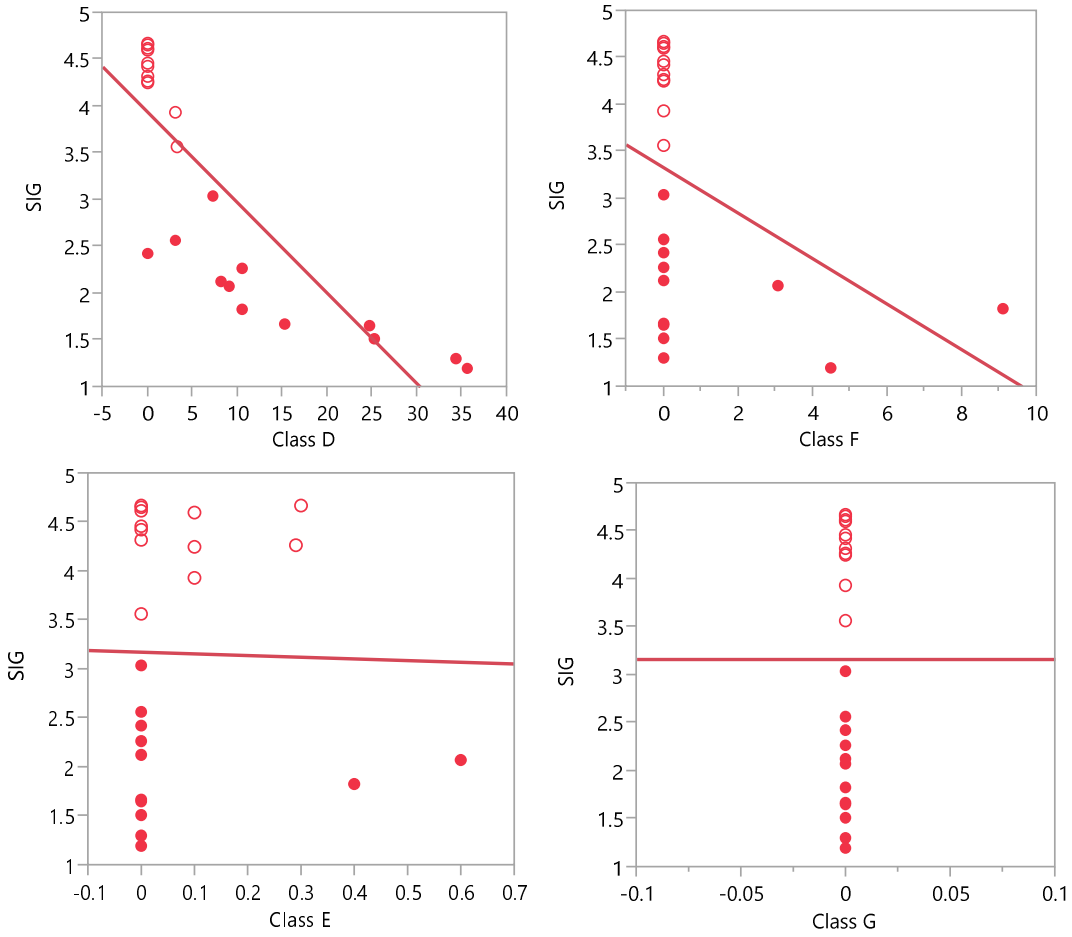


Figure 13: Scatter plots, SIG DT by 3GPP DT Class

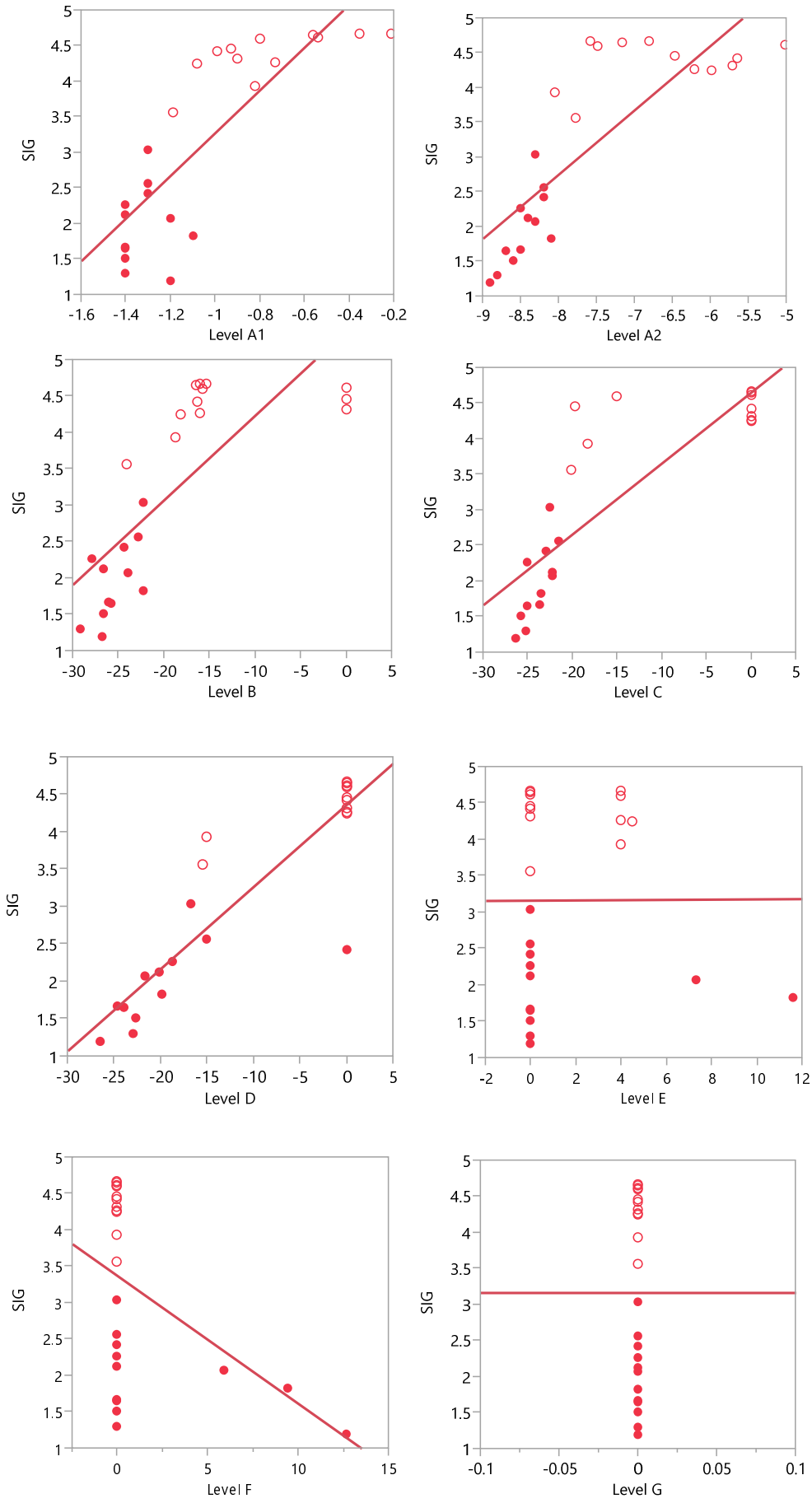


Figure 14: Scatter plots, SIG DT by 3GPP DT Level

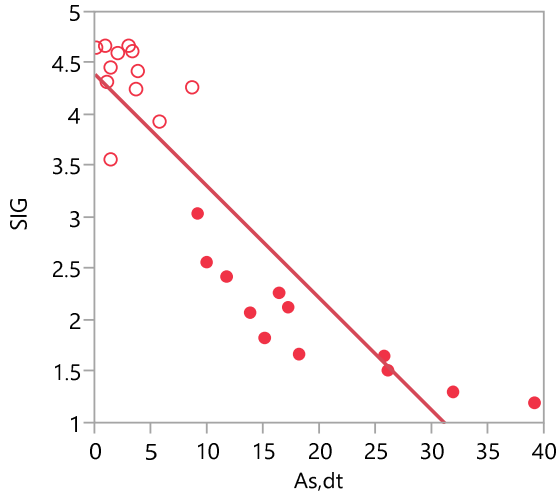


Figure 15: Scatter plot, SIG DT by P.502 As,dt

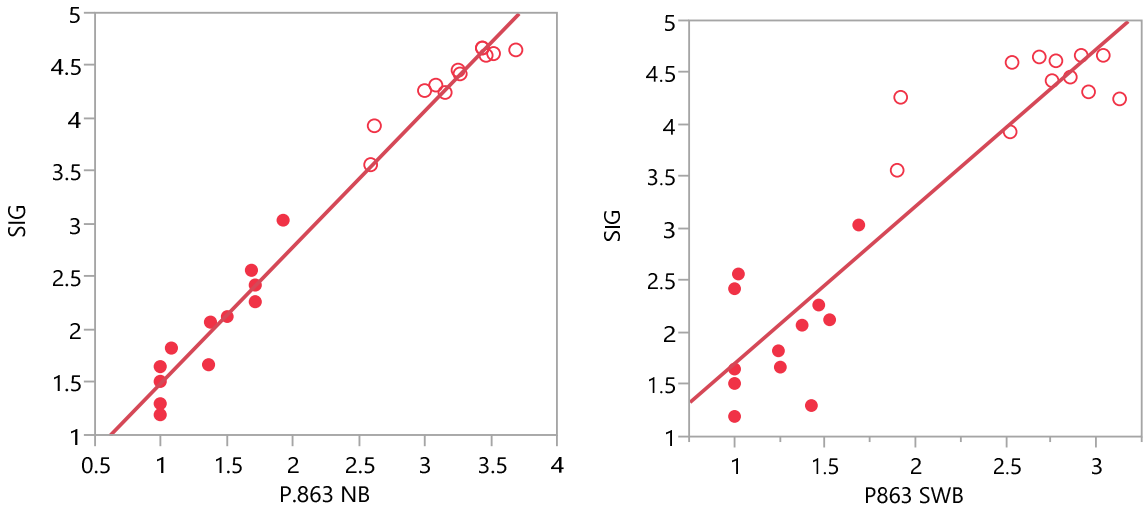


Figure 16: Scatter plots, SIG DT by P.863

Since several of the 3GPP metrics appear to have some predictive capability, the analysis was extended to include the SIG ratings in Single Talk (ST) as well. An optimal (in Akaike Information Criterion sense) linear combination of the 3GPP metrics is compared to P.863 MOS-LQOn for the combined SIG ratings. The linear combination of Class B, C, and D was AIC optimal (AIC=21,0098), with scatter plot and fit for this model shown in Figure 17.

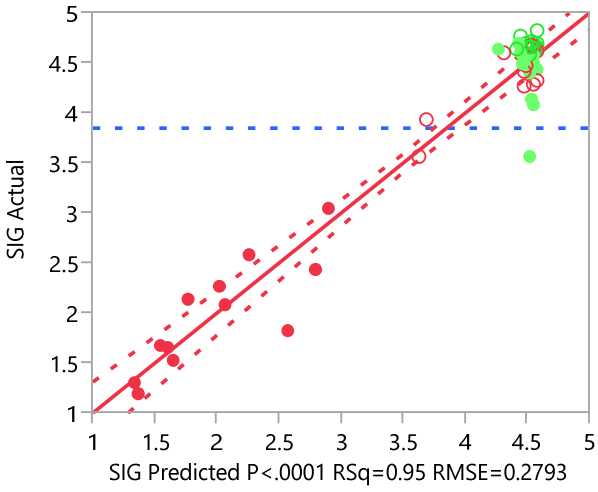


Figure 17: Scatter plot and fit to SIG DT & ST, by 3GPP

Figure 18 provides a scatter plot and fit to the combined SIG DT & ST ratings by P.863 MOS-LQOn. In Figures 17 and 18, red symbols are for DT and green symbols are for ST, while filled symbols are for HHHF and open symbols are for HS.

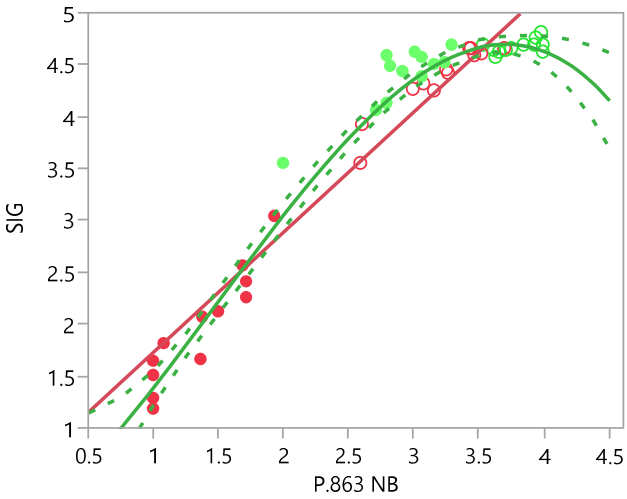


Figure 18: Scatter plot and fit to SIG DT & ST, by P.863

For P.863, linear fit (red line) and 3rd order polynomial fit (green line), according to Recommendation ITU-T P.1401 [18] are provided. Details of the fits shown in Figures 17 and 18 are provided in Table 3.

Table 3: Fits to SIG DT & ST

SIG DT & ST		R^2	correl.	rmse	ANOVA F	AIC
3GPP Optimal		0.946	0.973	0.279	*<0.0001*	21.010
P.863 MOS-LQOn	linear	0.915	0.957	0.341	*<0.0001*	
	3rd order	0.977	0.988	0.182	*<0.0001*	

The 3rd order mapping of MOS-LQOn appears to provide the best overall prediction of combined SIG ratings for combined ST and DT.

4.3.2.3 Comparison of metrics to subjective BAK ratings

Table 4 contains summary values for the fit of the above-described metrics to the BAK DT ratings.

Table 4: Summary of model fits for BAK DT

BAK DT			R^2	correl.	rmse	ANOVA F
3GPP	Atten Class	DT class A1	0.733	0.856	0.452	*<0.0001*
		DT class A2	0.441	0.664	0.654	*0.0004*
		DT class B	0.532	0.729	0.598	*<0.0001*
		DT class C	0.544	0.738	0.590	*<0.0001*
		DT class D	0.352	0.593	0.204	*0.0022*
	Echo Class	DT class E	0.073	0.270	0.841	0.2007
		DT class F	0.318	0.564	0.722	*0.0041*
		DT class G	0.000	0.000	0.855	na
	Atten Level	DT level A1	0.471	0.686	0.636	*0.0002*
		DT level A2	0.391	0.625	0.682	*0.0011*
		DT level B	0.393	0.627	0.681	*0.0010*
		DT level C	0.461	0.679	0.641	*0.0003*
		DT level D	0.523	0.723	0.604	*0.0001*
	Echo Level	DT level E	0.075	0.274	0.841	0.1941
		DT level F	0.223	0.472	0.187	*0.0199*
		DT level G	0.000	0.000	0.855	na
P.502		$A_{s,DT}$	0.510	0.714	0.612	*0.0001*
P.863		MOS-LQOn	0.782	0.884	0.408	*<0.0001*
		MOS-LQOw	0.738	0.859	0.447	*<0.0001*

For BAK (intrusiveness of echo), the correlations reported here are generally higher than in Recommendation ITU-T P.502, Amendment II revised Appendix 3 as in that report, the analysis was restricted to only HHHF, whereas the results in Table 4 include both HHHF and HS. As there can be no echo in the ST condition, an AIC-optimal linear combination of the 3GPP components was performed. Figure 19 shows scatter plots and fits to this optimal combination (Class B and C) and the P.863 MOS-LQOn.

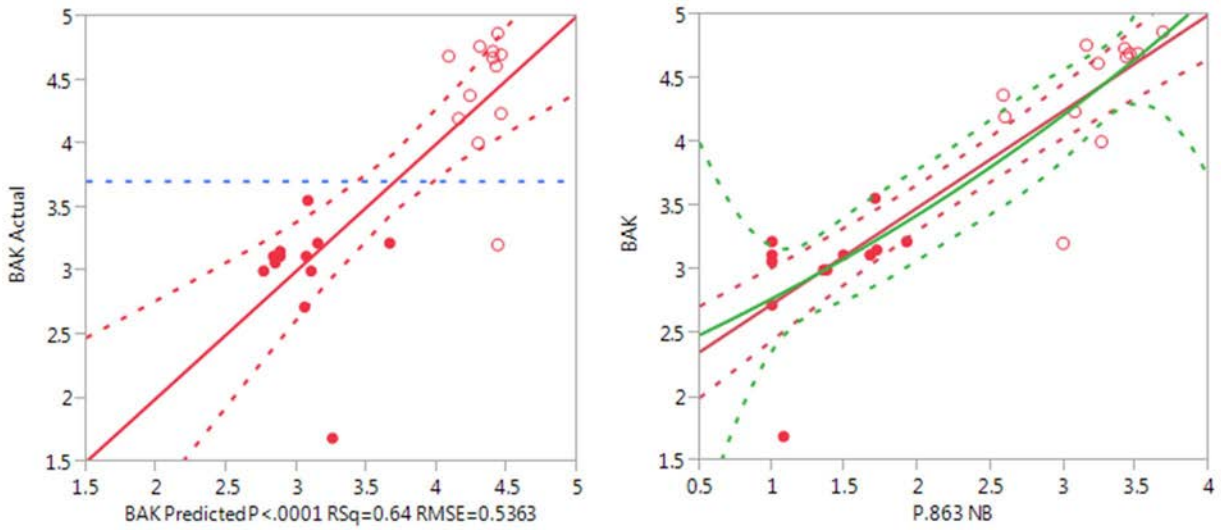


Figure 19: Scatter plot, BAK DT by 3GPP (left) and P.863 MOS-LQOn (right)

For P.863, linear fit (red line) and 3rd order polynomial fit (green line), according to P.1401 [18] are provided. Details of the fits shown in Figure 19 are provided in Table 5.

Table 5: Fits to BAK DT

BAK DT		R^2	correl.	rmse	ANOVA F
3GPP Optimal		0.641	0.801	0.536	<0.0001
P.863 MOS-LQOn	linear	0.772	0.879	0.408	<0.0001
	3rd order	0.784	0.885	0.426	<0.0001

The 3rd order mapping of MOS-LQOn appears to provide the best overall prediction of combined BAK ratings for DT, in terms of correlation. The rmse for the 3rd order mapping is somewhat higher than the rmse for the linear mapping.

4.3.2.4 Comparison of metrics to subjective OVRL ratings

Table 6 contains summary values of the fit of the above-described metrics to the OVRL DT ratings.

Table 6: Summary of model fits for OVRL DT

OVRL DT			R^2	correl.	rmse	ANOVA F
3GPP	Atten Class	DT class A1	0.929	0.964	0.324	*<0.0001*
		DT class A2	0.347	0.589	0.988	*0.0024*
		DT class B	0.588	0.785	0.598	*<0.0001*
		DT class C	0.860	0.927	0.458	*<0.0001*
		DT class D	0.635	0.797	0.739	*<0.0001*
	Echo Class	DT class E	0.006	0.077	1.219	0.7190
		DT class F	0.174	0.417	1.111	0.0424
		DT class G	0.000	0.000	1.196	na
	Atten Level	DT level A1	0.678	0.823	0.694	*0.0001*
		DT level A2	0.604	0.777	0.770	*0.0001*
		DT level B	0.569	0.754	0.803	*0.0001*
		DT level C	0.664	0.815	0.709	*0.0001*
		DT level D	0.767	0.876	0.590	*0.0001*
	Echo Level	DT level E	0.001	0.032	1.222	0.8684
		DT level F	0.212	0.460	1.086	*0.0235*
		DT level G	0.000	0.000	0.855	na
P.502		$A_{s,DT}$	0.785	0.886	0.566	*0.0001*
P.863		MOS-LQOn	0.976	0.988	0.188	*<0.0001*
		MOS-LQOw	0.879	0.938	0.425	*<0.0001*

For OVRL (overall listening quality), the correlations reported here are generally slightly lower than those reported for SIG in Table 2, but higher than those reported for BAK in Table 4. Values in Table 6 include the R^2, correlation, rmse, and ANOVA results for the Class (% frames) and Level (dB atten) according to the 3GPP analysis, the Sending attenuation analysis $A_{s,DT}$ of updated Appendix3 of P.502, and both MOS-LQOn and MOS-LQOw according to P.863, as for Tables 2 and 4.

Similar to results for SIG in Table 2, these results show that the best single predictor of the OVRL results for both HS and HHHF taken together is the P.863 MOS-LQOn (correlation 0.988, rmse 0.188) followed by the 3GPP DT Class A1 (correlation 0.964, rmse 0.324). As the scatter plots for OVRL are generally similar to those shown above in Figures 13, 14, 15 and 16, they will not be included here, as very little additional information is provided.

4.3.2.5 Summary of comparison of metrics to subjective ratings for double-talk

Clause 4.3.2 compares subjective results for double talk to three classes of metrics. For speech distortion, SIG, the best single-value predictor among the set considered is the MOS-LQOn according to P.863, with a 3rd order polynomial remapping.

In contrast to SIG, for echo intrusiveness, BAK, the metrics investigated here do not perform as well. This may be due to the specifics of the UEs used, their echo behaviour, or it may be due to spectral aspects of the echo signal that are not captured in the metrics examined here. However, the best performing metric is again MOS-LQOn according to P.863. The reasons for the superiority of MOS-LQOn over MOS-LQOw are not examined herein.

Note that:

- The evaluated 3GPP metrics (categories, attenuations) described in Clause 4.3.2.1 were not designed to be quality predictors, rather they are technical parameters that characterize the behaviour of echo controllers.
- The ITU-T metric described in Clause 4.3.2.1 is intended to predict only the attenuation aspect of DT behaviour,
- P.863 is not designed for predicting the subjective tests used here (P.835).

- The P.835 test methodology was not designed for echo evaluation
- For practical reasons, the subjective evaluation used listening-only experiments which does not include all aspects of a conversational situation.

Nevertheless, the results presented are of interest when evaluating echo controller related performance.

4.4 Free-field measurements for vehicle-mounted hands-free

Void

4.5 Idle Channel Noise, Sending/Receiving of test signal

Void

5 New Acoustic Tests

5.1 Time-variant user behaviour

In current terminal testing specifications for handset mode, e.g. according to TS 26.132 [8], only one default positioning of the device is issued. The study presented in clause 5.3 considered multiple positions of a DUT and identified critical test cases.

5.2 Additional UE usage environments

The background noise simulation system according to ETSI ES 202 396-1 [7] is used for terminal testing in TS 26.132 [8]. Especially for the handheld hands-free mode, the reproduction of the sound field method may be inaccurate at the position of the DUT. For this reason, the background noise playback system according to ETSI TS 103 224 [14] was introduced here as an additional method. It provides an automated equalization procedure. An example for the usage and comparison of this reproduction system is given in clause 5.3.

Another UE usage environment for consideration in terminal testing specifications are reverberant rooms. The upcoming specification ETSI TS 103 557 [22] describes methods for the simulation of reverberation in laboratories.

5.3 Results from a study on positional robustness tests and background noise simulations

5.3.1 General

Following current standards for acoustic handset testing, a mobile phone is positioned according to [12]. The annex of [12] defines how to position a handset for artificial ears of type 3.3 which should correspond to typical human usage of a handset. It furthermore defines how to select different positions.

This position does not cover all possible human holding positions of a mobile phone handset today.

More extreme positioning are not covered by existing 3GPP specifications. This clause presents an evaluation of an alternative "down position". This position where the phone is tilted downwards might be used by people when not really concentrating on positioning the phone the proper way. Users may have little awareness of the impact on speech quality when using the phone in this way since there is no acoustical feedback in sending which would help them to relocate the phone back in a proper position. In some cases the other party on the call may provide feedback that the other user is hard to understand which may indicate to re-position the phone. But this is not intuitively understood by all users.

The positioning effects in sending are illustrated using analyses in silence and in background noise situations. Additionally, two types of background noise simulations are compared for the evaluation

5.3.2 Setup

5.3.2.0 General

A calibrated system consisting of HATS HMS II.3 (head and torso simulator), MFE VI.1 (analogue/digital reference interface) and a 3G radio tester (Rohde & Schwarz CMW 500 with analogue audio output / audio board) according to Figure 20 was used. Four modern mobile phones from four different manufacturers in wideband operational mode (AMR-WB, 12,65 kbit/s) were used. Table 7 provides some technical data on these devices under test (DUT). It is not known if any device uses one, two or more microphones e.g. for noise cancellation.

Table 7: Information on different DUTs

Device	Year of production	Size
A	2013	12,4 cm x 5,9 cm x 0,8 cm
B	2012	13,7 cm x 7,1 cm x 0,9 cm
C	2011	10,0 cm x 5,0 cm x 1,0 cm
D	2014	13,9 cm x 7,1 cm x 0,9 cm

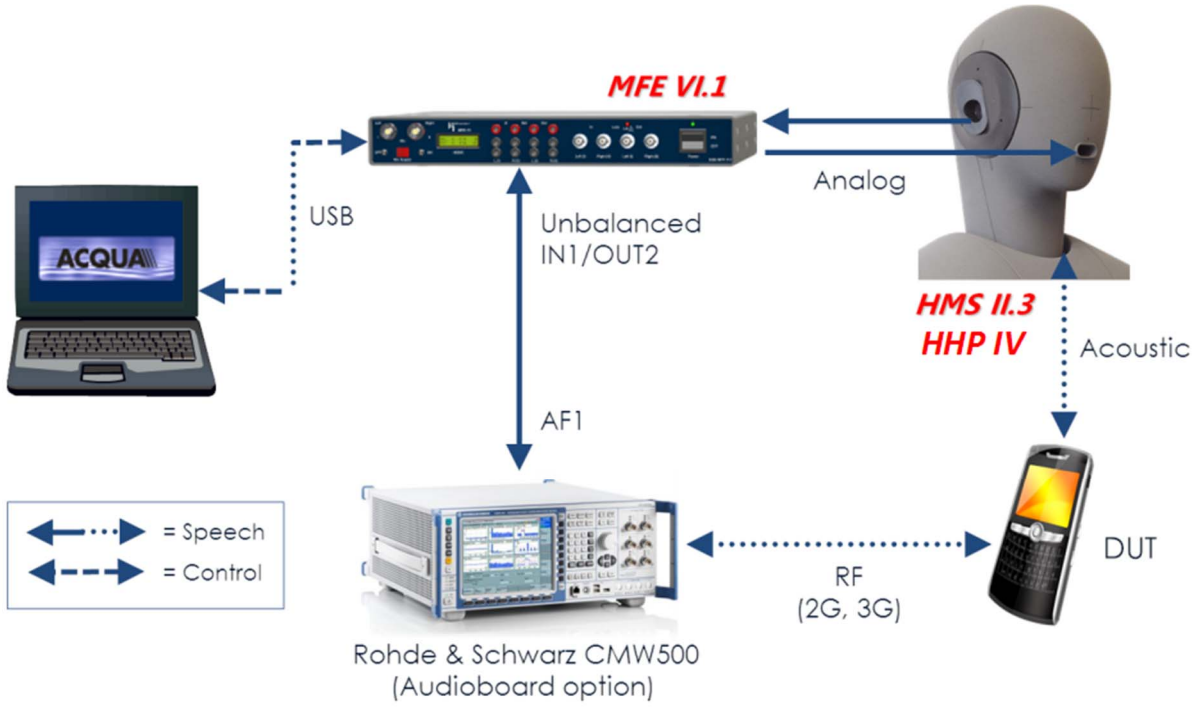
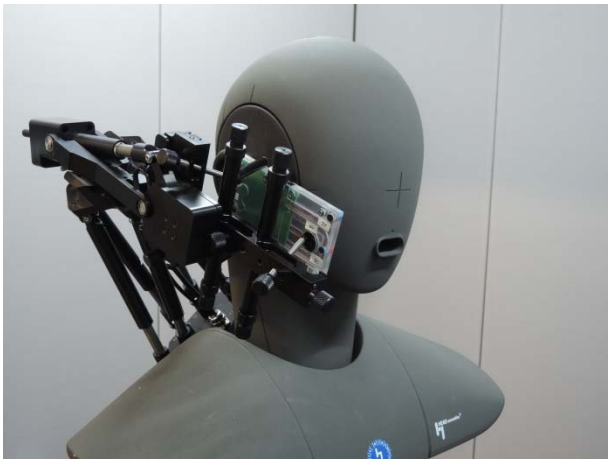


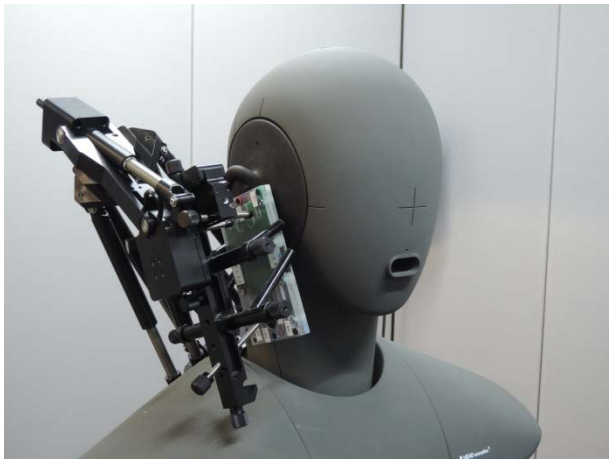
Figure 20: Test setup for evaluation

5.3.2.1 Handset Mounting

The DUT was mounted with *HEAD Handset Positioner IV* (HHP IV), capable of automatically placing a clamped-in handset in a wide range of standard or user-defined positions. The different mounting positions are illustrated with a mockup phone (which was not part of the evaluation) in the following photos.



a) Position I: Handset position for flat handsets (see text)



b) Position II: Alternative 'Vertical' position (see text)

Figure 21: Mounting of handset

Figure 21a shows the mockup phone with a handset positioning for flat handsets (0°, 0°, 5°) denoting the positioning delta from the standard HATS position for (A, B, C) in ITU-T P.64 [12].

Position I:

Table 7a

MECRP (delta from actual ECRP)

Axis	Delta [mm]
ye	0
ze	0

Table 7b

Angle settings

Angle	Delta from standard angle [°]
A	0
B	0
C	5

This mounting is named "norm" in the following clauses.

Figure 21b shows an alternative version of mounting. The handheld positioner was set to the delta angles (A, B, C) = (45°, 0°, 5°) . Additionally, a larger distance between mouth and DUT was simulated; this was realized by shifting the device 1 cm higher than the default mounting. This shift takes into account that the dimensions of HATS as defined in [15] are felt to be too small compared to the head size of many people today and such the distance from mouth to microphone and the cheek shadow effect might not be properly taken into account in tests. To some extent this shadowing effect can be compensated in sending by shifting the DUT. In order to provide a better separation to the existing positions, this new alternative position is named "vertical" in the following clauses.

Position II:

Table 7c

MECRP (delta from actual ECRP)

Axis	Delta [mm]
ye	0
ze	10

Table 7d

Angle settings

Angle	Delta from standard angle [°]
A	45
B	0
C	5

For both mountings, an application force of 8N was used.

5.3.2.2 Background Noise Systems

5.3.2.2.0 General

For the generation of background noise during the measurements two different simulation approaches were used: the new simulation technique described in ETSI TS 103 224 [14] and the present industry standard described in ETSI ES 202 396-1 [7]. Both systems were set up and equalized in a small, mostly non-reverberant room (RT60 < 100 ms). The next two paragraphs summarize the properties of these two systems.

5.3.2.2.1 Equalization process according to ETSI ES 202 396-1

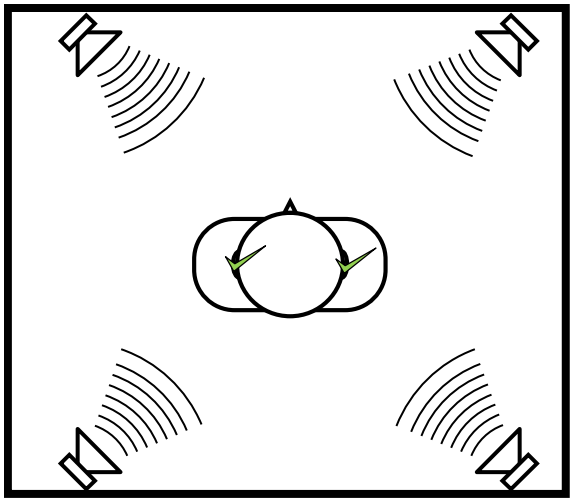


Figure 22: Setup for equalization according to ETSI ES 202 396-1 [7]
The green checkmarks show where the frequency response is correctly equalized
(note that only the magnitude is taken into account)

Figure 22 shows the setup used for the system described in ETSI ES 202 396-1 [7]. It uses four loudspeakers placed around the HATS and additionally a subwoofer. The goal is that the frequency response of the background noise corresponds between 50 Hz to 10 kHz at the two ears of the HATS to the original recording.

Both left and right loudspeakers respectively are handled as a group and one FIR-Filter is calculated for each side automatically. To compensate for the crosstalk between both sides IIR filters are adjusted for each side manually. It has to be emphasized that the phase of the signal is not equalized and that in addition to that every loudspeaker introduces an individual delay. This creates a more diffuse sound field in the vicinity of the HATS.

With the equalized system binaural recordings can be played back and used as background noise during measurements. When doing hands-free measurements the system has to be equalized with the HATS and after that the DUT has to be positioned at the position of the HATS.

The HEAD acoustics implementation *HEAD Automated Equalization for Background Noise Simulation in Laboratories* (HAE-BGN) of the present document was used for this evaluation.

5.3.2.2.2 Equalization process according to ETSI TS 103 224

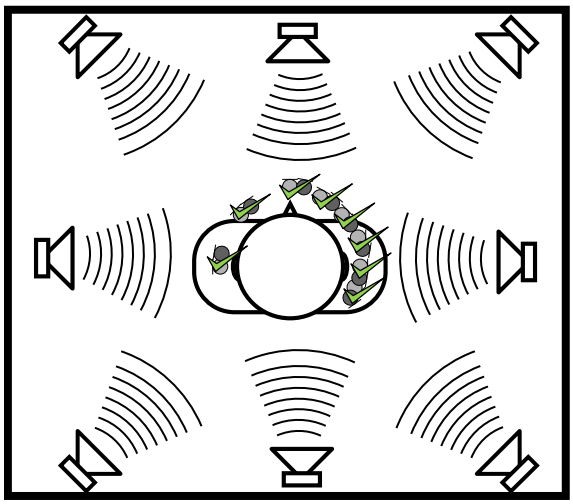


Figure 23: Setup for equalization according to ETSI TS 103 224 [14]
The green checkmarks show where the frequency response is correctly equalized
(note that also the phase is taken into account up to a frequency of about 2 kHz)

As it can be seen in Figure 23 eight loudspeakers are used for generating the background noise. This system achieves a close-to-reality simulation of the frequency responses of the background noise at a minimum of eight positions around the HATS. For this purpose 64 FIR filters are calculated automatically from the impulse responses between every loudspeaker and every microphone (8x8). These filters also ensure that the characteristics of the simulated sound field correspond to the original situation also in-between the microphone positions up to 2 kHz regarding magnitude and phase. The locations of the microphones are selected such that a close to real sound field is generated close to the HATS which makes the sound field less diffuse and brings it closer to reality.

After equalization, recordings which were made using the same 8 positions can be reproduced. For hands-free measurements, the microphone array has to be positioned at the position of the DUT in hands-free position (cf. ETSI TS 103 224 [14]).

The HEAD acoustics implementation *HEAD 3-dimensional Playback of Acoustic Sound Scenarios* (3PASS) of the present document was used for this evaluation.

5.3.2.2.3 Background noises

Four different background noises were selected for this evaluation:

- Road noise (~71 dBSPL(A))
- Train Station (~78 dBSPL(A))
- Full-size Car 130 km/h (~68 dBSPL(A))
- Cafeteria noise (~69 dBSPL(A))

Those background noises were recorded simultaneously with the microphone array described in ETSI TS 103 224 [14] for handset setup as well as for hands-free setup. In addition to that, a binaural recording was also made for the system from ETSI ES 202 396-1 [7]. Thus the same noises can be played back on both systems.

5.3.3 Measurement Results

5.3.3.1 Measurements in Silence Condition

Figure 24 shows the frequency responses ($1/12^{th}$ octave resolution) for each device. The green curves indicate the transfer functions in normal position, red curves were measured with the alternative vertical position. Beside a constant offset for devices A-C, all devices show issues in the upper frequency region (> 2 kHz). Device D at least tries to compensate the absolute level difference, the lower frequency content almost matches in both positions.

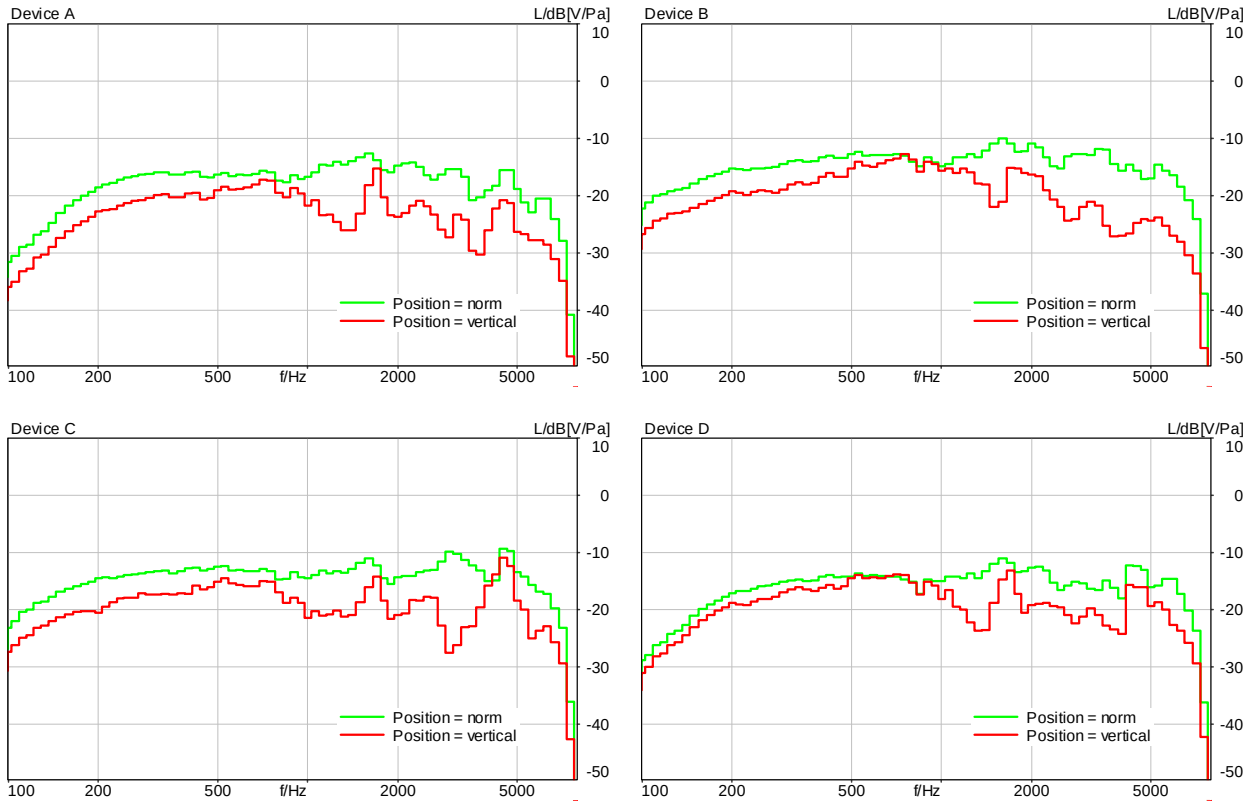


Figure 24: Frequency Responses for different devices and positioning

Table 8 shows some more typical metrics for measurements under silence conditions for each device which can describe these differences by single values:

- Sending Loudness Rating (acc. to Recommendation ITU-T P. 79 [19])
- TOSQA 2001 (WB mode, electrical recording)
- POLQA according to Recommendation ITU-T P.863 [6], Version 2.4 (fixed active speech level of 73 dB SPL)

The absolute values are given in the column "norm", the difference to the vertical position is given as "vert. - norm.".

Table 8: Metrics for sending direction per device

		DUT A		DUT B		DUT C		DUT D	
		norm.	vert.-norm.	norm.	vert.-norm.	norm.	vert.-norm.	norm.	vert.-norm.
Sending Loudness Rating	[dB]	10,15	+4,43	7,44	+3,31	7,36	+4,62	8,64	+3,14
TOSQA2001 (WB)	[MOS]	3,61	-0,12	3,61	-0,37	3,74	-0,22	3,37	-0,23
P.863 (POLQA)	[MOS]	3,90	-0,23	3,89	-0,07	3,75	-0,24	3,79	+0,06

All loudness ratings increase by at least 3,1 dB (up to 4.6), which is mainly caused by the modified frequency responses shown in Figure 24. The large differences in the spectral domain are not leading to huge differences in the speech quality measures TOSQA and POLQA. In fact, all MOS scores decrease for POLQA as well as for TOSQA when comparing normal vs. vertical position. The maximum difference for POLQA is -0,24 for Device C, whereas the largest difference for TOSQA is found for device B (-0,37). Please note that the effect of decreased loudness is not captured by TOSQA and POLQA, because a constant listening level of 73 dB (A) SPL was set.

5.3.3.2 Measurements with Ambient Noise

The following tables provide the results according to ETSI TS 103 106 [5] for each device and each positioning. Each MOS value is determined as the average over 16 American English test sentences taken from annex C of [5]. Note that MOS values as well as the calculated differences are round to one decimal place.

Table 9: TS 103 106 S-/N-/G-MOS values for device A

			HAE		3PASS		3PASS - HAE	
DUT	BGN	Value	norm.	vert.-norm.	norm.	vert.-norm.	norm.	vert.
A	Roadnoise	G-MOS	3.6	-0.5	3.7	-0.4	0.1	0.2
		N-MOS	4.0	-0.4	4.0	-0.3	0.1	0.2
		S-MOS	3.9	-0.4	4.0	-0.4	0.1	0.1
	TrainStation	G-MOS	2.8	-0.5	2.9	-0.5	0.1	0.1
		N-MOS	3.4	-0.3	3.4	-0.2	0.1	0.1
		S-MOS	3.1	-0.5	3.2	-0.6	0.1	0.0
	FullSizeCar_130	G-MOS	3.8	-0.3	3.9	-0.3	0.1	0.1
		N-MOS	4.0	-0.3	4.2	-0.3	0.2	0.2
		S-MOS	4.1	-0.2	4.2	-0.2	0.0	0.0
	Cafeteria	G-MOS	3.7	-0.4	3.8	-0.4	0.1	0.0
		N-MOS	3.6	-0.5	3.8	-0.5	0.2	0.1
		S-MOS	4.1	-0.2	4.2	-0.2	0.0	0.0

Table 10: TS 103 106 S-/N-/G-MOS values for device B

			HAE		3PASS		3PASS - HAE	
DUT	BGN	Value	norm.	vert.-norm.	norm.	vert.-norm.	norm.	vert.
B	Roadnoise	G-MOS	3.3	-0.9	3.5	-0.8	0.1	0.2
		N-MOS	3.2	-0.4	3.4	-0.4	0.2	0.2
		S-MOS	3.9	-1.0	4.0	-0.9	0.1	0.2
	TrainStation	G-MOS	2.6	-0.7	2.8	-0.8	0.2	0.1
		N-MOS	2.6	-0.4	2.8	-0.4	0.2	0.2
		S-MOS	3.4	-0.9	3.5	-1.0	0.1	0.0
	FullSizeCar_130	G-MOS	3.6	-0.4	3.7	-0.4	0.1	0.1
		N-MOS	3.3	-0.3	3.4	-0.2	0.1	0.2
		S-MOS	4.1	-0.4	4.2	-0.4	0.1	0.1
	Cafeteria	G-MOS	3.2	-0.6	3.3	-0.7	0.1	0.0
		N-MOS	2.5	-0.3	2.6	-0.3	0.1	0.1
		S-MOS	4.1	-0.7	4.1	-0.8	0.0	-0.1

Table 11: TS 103 106 S-/N-/G-MOS values for device C

			HAE		3PASS		3PASS - HAE	
DUT	BGN	Value	norm.	vert.-norm.	norm.	vert.-norm.	norm.	vert.
C	Roadnoise	G-MOS	3.7	-0.6	3.8	-0.6	0.1	0.1
		N-MOS	4.5	-0.5	4.6	-0.5	0.0	0.0
		S-MOS	3.7	-0.6	3.8	-0.6	0.1	0.1
	TrainStation	G-MOS	2.9	-0.7	3.0	-0.8	0.1	0.0
		N-MOS	3.8	-0.8	3.9	-0.9	0.1	0.1
		S-MOS	3.0	-0.5	3.1	-0.6	0.1	0.0
	FullSizeCar_130	G-MOS	4.0	-0.5	4.1	-0.5	0.1	0.1
		N-MOS	4.6	-0.5	4.7	-0.3	0.1	0.2
		S-MOS	4.0	-0.4	4.1	-0.5	0.0	0.0
	Cafeteria	G-MOS	3.9	-0.5	3.9	-0.6	0.0	0.0
		N-MOS	4.6	-0.4	4.6	-0.5	0.0	-0.1
		S-MOS	3.9	-0.5	4.0	-0.4	0.0	0.1

Table 12: TS 103 106 S-/N-/G-MOS values for device D

			HAE		3PASS		3PASS - HAE	
DUT	BGN	Value	norm.	vert.-norm.	norm.	vert.-norm.	norm.	vert.
D	Roadnoise	G-MOS	3.8	-0.5	3.9	-0.4	0.1	0.1
		N-MOS	4.8	-0.7	4.8	-0.7	0.0	0.0
		S-MOS	3.7	-0.2	3.8	-0.1	0.1	0.2
	TrainStation	G-MOS	3.0	-0.7	3.2	-0.7	0.2	0.2
		N-MOS	3.6	-0.8	4.1	-0.8	0.6	0.6
		S-MOS	3.3	-0.6	3.3	-0.5	-0.1	0.0
	FullSizeCar_130	G-MOS	3.8	-0.3	4.0	-0.3	0.2	0.2
		N-MOS	4.0	-0.5	4.8	-0.4	0.9	1.0
		S-MOS	4.1	-0.1	3.9	-0.1	-0.2	-0.2
	Cafeteria	G-MOS	3.6	-0.4	3.7	-0.5	0.1	0.1
		N-MOS	3.6	-0.7	4.0	-0.9	0.4	0.1
		S-MOS	4.0	-0.1	4.0	-0.1	0.0	0.0

The decrease of S-/N-/G-MOS for the vertical position is observable for all devices. The minimum and maximum differences between vertical and normal position for each category is shown in Table 13. In overall, scores decreases at least by 0,1 and up to 1,0 MOS for S-, N- and G-MOS.

When using the two different background noise simulation systems very similar scores are obtained for S-, N- and G-MOS in the normal position with a tendency to slightly bigger differences in the vertical position. An exception here is device D. This device shows significant differences between the playback systems, for both positions. N-MOS scores increase up to 1,0 for the background noise simulation according to ETSI TS 103 224 [14]. Since the sound field reproduction system according to ETSI TS 103 224 [14] provides an accurate sound field reproduction at the location of the hand-held terminal (in contrast to an almost diffuse noise field generated when using ETSI ES 202 396-1 [7] which does not reproduce the physical characteristics of the real sound field at the location of the terminal) it can be assumed that these results might represent more accurately the "real life" behaviour of this device.

Table 13: Minimum/maximum difference

	Min. Diff.	Device	BGN	Max. Diff	Device	BGN
G-MOS	-0,3	A	Car	-0,9	B	Road
N-MOS	-0,2	A	Train Stat.	-0,9	D	Cafeteria
S-MOS	-0,1	C	Car	-1,0	B	Train Stat.

5.3.4 Summary

The study shows that degradations of device performance can be expected when using different positioning as used in the study. The degradations depend on device implementations. Different behaviour can be observed for some devices with different background noise simulations.

5.4 Results from a study on objective measures with noise suppression and background noise

5.4.1 Comparison of P.862 to subjective results for noise suppression

Recommendation ITU-T P.862, *Perceptual Evaluation of Speech Quality* (PESQ) [1], was published in 2001 to predict subjective scores as obtained in Recommendation ITU-T P.800 Absolute Category Rating (ACR) tests of Overall Speech Quality [2]. P.862 [1] is a very useful tool for assessing the speech quality of devices in many situations; however it has significant limitations for assessing speech quality when noise suppression is used, primarily because the model was developed and trained before modern UE noise suppression evolved. Since almost all modern UE's contain some form of noise suppression algorithm, P.862 [1] should not be used for testing UE's in background noise as it produces misleading results as demonstrated in the following clauses.

The inadequacy of P.862 [1] for devices incorporating noise suppression is well documented. ITU-T Recommendation P.862.3 [3], *Application Guide for Objective Quality Measurement Based on Recommendations ITU-T P.862 [1], P.862.1 [9], and P.862.2 [10]* provides unambiguous guidance regarding usage with noise suppression as can be seen in the following excerpt [p3-4]:

> **6.1 Testing factors**
>
> [ITU-T P.862] is validated for the evaluation of test factors, coding technologies and applications, which are listed in Table 1 of [ITU-T P.862]. In particular, care should be taken when one carries out live network testing since there might be some equipment that causes degradation that [ITU-T P.862] cannot handle, e.g., artefacts caused by noise reduction systems, in between the signal insertion point and signal capture point. It is also known that PESQ underestimates severe linear frequency response distortions. This applies especially to, e.g., bandwidth limitations narrower than 300 Hz ... 3.4 kHz.
>
> Use of [ITU-T P.862] with systems that include noise suppression algorithms between the signal insertion point and the signal capture point is not recommended.

Figure 24a: See Figure 1 from ITU-T P.862.3 [3]

ITU-T Recommendation P.835 is used for subjective evaluation of systems with noise suppression [4]. P.835 [4] evaluations use three rating scales: SIG assessing the amount of speech distortion; BAK assessing the degree of intrusiveness of any background noise; OVRL providing an overall speech quality assessment. ETSI TS 103 106 [5] was designed specifically to predict the three ratings from a P.835 [4] test and was explicitly trained on noise reduction. In contrast, both P.862 [1] and P.863 [6] were designed to predict results from P.800 [2] listening tests, which only have one rating scale, overall speech quality, MOS-LQS. In fact, P.835 [4] was developed partly in reaction to a high degree of uncertainty observed in P.800 [2] tests on conditions with noise reduction. Studies directly comparing P.800 [2] MOS-LQS scores to P.835 [4] OVRL scores are predominantly proprietary, but experience has shown that there is generally good correlation between these two metrics.

An objective quality predictor is normally trained on a wide variety of conditions, while a single subjective experiment includes a more limited set of conditions which potentially changes the experimental context. Therefore results from a specific subjective experiment may deviate from objective predictions, hence the results below should be considered as case studies, not as exhaustive and definitive results. However, these case studies can be considered as indicative of the nature of issues that could arise when predictors are applied outside their scope.

5.4.2 Experiment 1 - NB P.835 versus P.862.1

5.4.2.1 Setup

In Experiment 1, a recording of American English speech consisting of four sentences from each of two male and two female native talkers, as used in ETSI TS 103 106 [5], was reproduced through an equalized HATS artificial at 92,3 dB SPL active speech level at the MRP of HATS. The background noise generation method described in ETSI ES 202 396-1 [7] was used to reproduce eight noises described in Table 2d of Clause 7.12 of 3GPP TS 26.132 [8]. Six handsets were used, denoted as A, B, C, D, E, and F in the following plots. Each device was mounted on the HATS in standard position and recordings made of clean speech, and speech mixed with the eight noises described above, at the output of a UMTS base-station simulator with AMR speech encoding at 12.2 kbit/s.

Recommendation ITU-T P.835 [4] is the recommended methodology for assessing the listening quality of systems incorporating noise suppression. A group of 32 naïve listeners, all native speakers of American English, using the P.835 [4] methodology, rated all sentences in all conditions in a partially balanced randomized blocks design, resulting in 128 votes per condition. Results were used as training data for ETSI TS 103 106 [5] with further details in Clause 7.2.1 of that document.

For computing metrics, all recordings were taken at 48 kHz sample rate. For P.862.1 [9], the clean source file was filtered to NB for use as reference and the scores were computed separately on sentence pairs then averaged to obtain per-condition values. For ETSI TS 103 106 [5], the NB operational mode was selected and scores were computed on individual sentences and then averaged to obtain per-condition values.

5.4.2.2 Results

The results of this experiment are shown below, for four noise types.

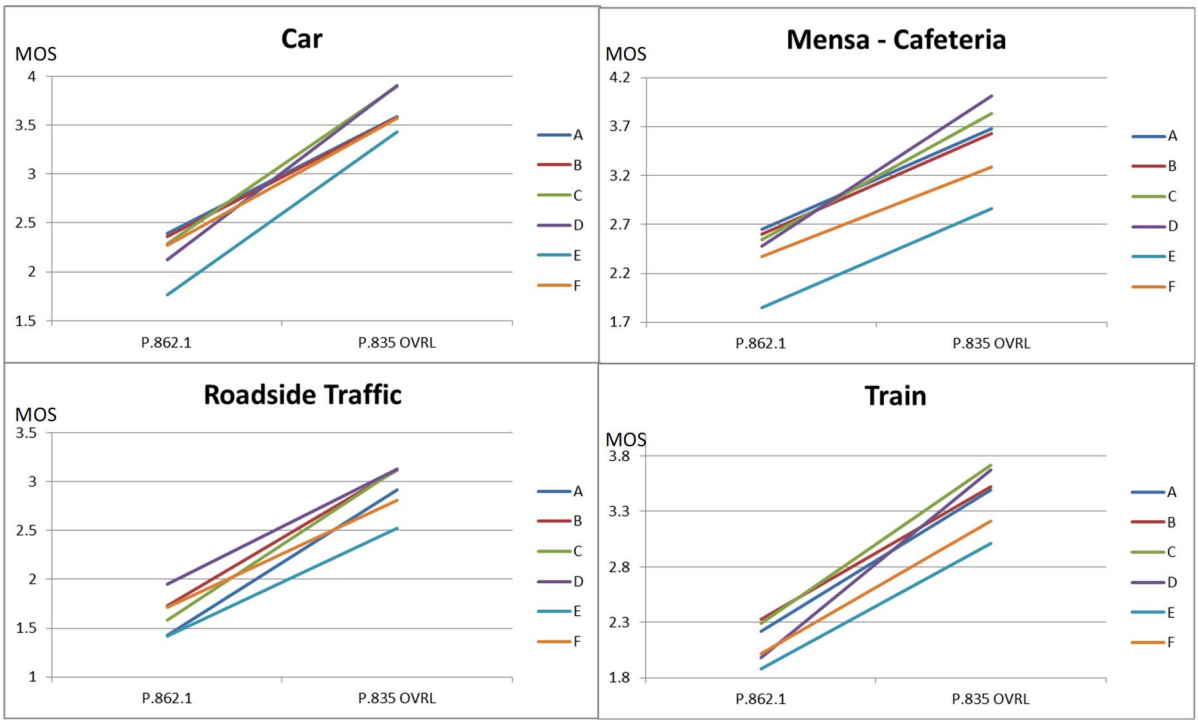


Figure 24b: Results of Experiment 1: comparing scores from P.835 [4] and P.862.1 [9] in 4 noise types

It can be seen from these graphs that P.862.1 [9] significantly underestimates the performance of the various handsets when compared to the results obtained from the P.835 [4] test with human listeners. In addition the crossing of the lines shows that the results obtained using P.862.1 [9] do not preserve the rank order that is obtained with human listeners. For example handset D is the top or almost top performing handset for all noise types when human listeners are used, with a score of fair to good, however when tested with P.862.1 [9], it is the 4th or 5th ranking handset in 3 of the noise types with a score of only poor.

Figure 25 shows a scatter plot of the subjective P.835 [4] OVRL ratings against objective scores. The red-coloured symbols are for P.862.1 [9] while the blue-coloured symbols are for the GMOS output from the P.835 [4] predictor of TS 103 106 [5]. As noted above, the range of the P.862.1 [9] predictions is compressed relative to the range of subjective ratings. For example, subjective ratings in the range of 3.5, or mid-way between 3 "Fair" and 4 "Good" are predicted by P.862.1 [9] in the region of 2 "Poor". In contrast, for GMOS, the predicted scores span the full range of the subjective ratings, and fall very close to the diagonal line of slope 1.

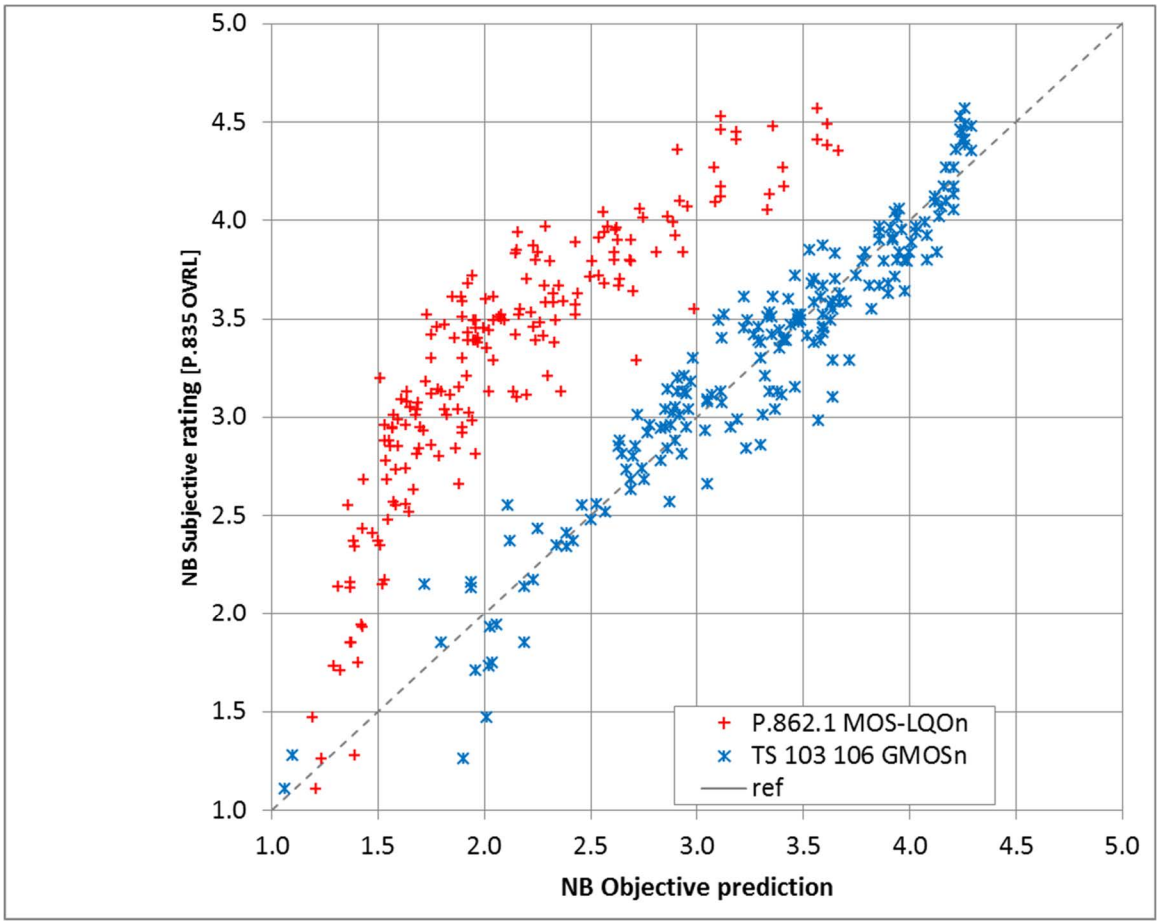


Figure 25: Results of Experiment 1 - correlation of P.835 [4] with P.862.1 [9] and ETSI TS 103 106 [5] (NB). Absolute maximum error for P.862.1 [9] is 1.630, absolute maximum error for ETSI TS 103 106 [5] is 0,880. Spearman rank-order correlation, accounting for 95 % confidence interval, for P.862.1 [9] is 0.914 and for ETSI TS 103 106 [5] is 0,984

5.4.3 Experiment 2 - Problems with tuning for P.862.1

5.4.3.1 Setup

In Experiment 2, a recording of American English speech consisting of four sentences from each of two male and two female native talkers, as used in ETSI TS 103 106 [5], was reproduced through an equalized artificial mouth. The active speech level was -1.7 dBPa at the MRP of HATS. The background noise generation method described in ETSI ES 202 396-1 [7] was used to reproduce eight noises described in Table 2d of Clause 7.12 of 3GPP TS 26.132 [8]. In this case, a single handset was used, but with noise suppression parameters adjusted in two ways. The tuning labelled "Tuned for P.862.1" was defined so as to provide high values of Recommendation ITU-T P.862.1 [9]. The tuning labelled "Alternative Tuning" was defined in general accordance with the requirements of the marketplace, based on network operator requirements. The device was mounted on the HATS in standard position, and recordings made of clean speech, and speech mixed with the eight noises described above, at the output of a UMTS base-station simulator with AMR speech encoding at 12.2 kbit/s.

A group of 32 naïve listeners, all native speakers of American English, using the Recommendation ITU-T P.835 [4] methodology, rated all sentences in all conditions in a partially balanced randomized blocks design, resulting in 128 votes per condition.

5.4.3.2 Results

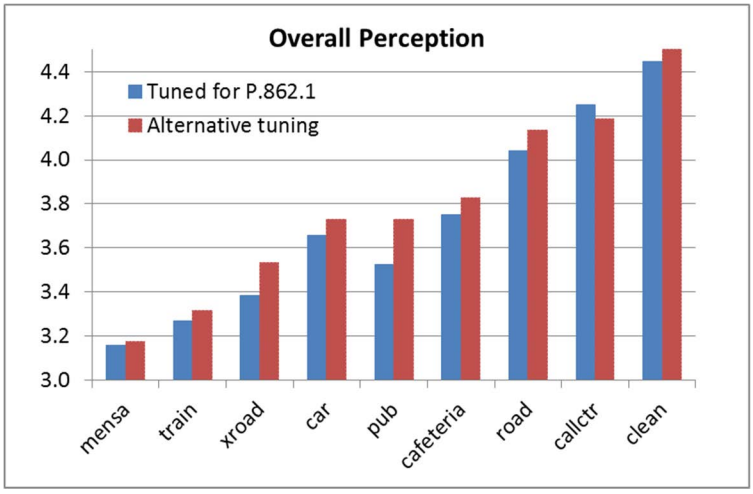


Figure 26: Results of Experiment 2 - P.835 [4] Overall Score

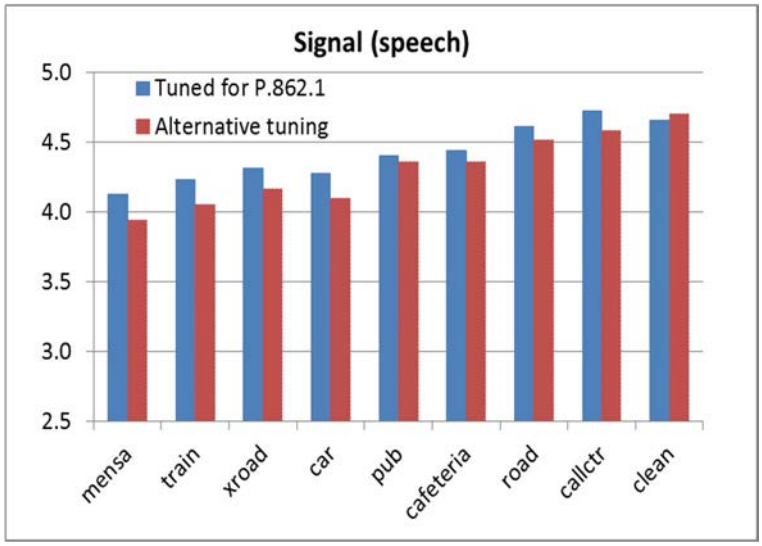


Figure 27: Results of Experiment 2 - P.835 [4] Signal Score

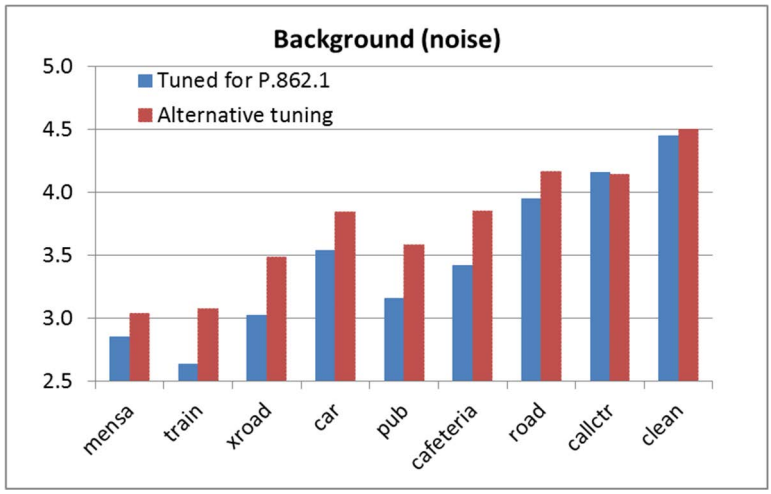


Figure 28: Results of Experiment 2 - P.835 [4] Background Score

The results for the overall performance in the P.835 [4] listening test of figure 26 show that tuning an algorithm to get the best P.862.1 [9] score does not produce an optimum result. The version of the algorithm tuned for P.862.1 [9] significantly underperforms in 6 of the 9 noise types. The graphs in Figures 27 and 28 break down the performance in terms of the signal and background noise scores. These show that while tuning an algorithm to maximize the P.862.1 [9] score slightly improves the perceived quality of the speech signal, Figure 27, it significantly degrades the perception of the background noise signal, figure 28, leading to the degraded overall score as shown Figure 26.

5.4.4 Experiment 3: WB P.835 v P.862.2, P.863 and TS 103 106

5.4.4.1 Setup

The method used was very similar to the one described for narrow-band in Clause 5.4.2.1. However, the UMTS base-station simulator was set to use speech encoding with AMR-WB at 12.65 kbit/s.

For computing metrics, all recordings were taken at 48 kHz sample rate. For P.862.2 [10] the clean source file was filtered to WB for use as reference and the scores were computed separately on sentence pairs then averaged to obtain per-condition values. For P.863 [6], version 2.4 was used; the clean source was filtered to SWB for use as reference and the scores were computed using the SWB mode, separately on sentence pairs then averaged to obtain per-condition values. For ETSI TS 103 106 [5], scores were computed on individual sentences and then averaged to obtain per-condition values.

5.4.4.2 WB Correlation Results

Figure 29 shows a scatter plot of the subjective P.835 OVRL [4] ratings against objective scores. The red-coloured (+) symbols are for P.862.2 (MOS-LQOw) [10], while the blue-coloured (x) symbols are for the GMOS output from the P.835 [4] predictor of [5] and the green-coloured (o) symbols are for P.863 Version 2.4 [6].

As for narrowband, the range of the P.862.2 [10] predictions is compressed relative to the range of the subjective ratings. Subjective ratings of about 3.5, or midway between "Fair" and "Good" are scored by P.862.2 [10] at about 1.7, or below "Poor", and the RMSE is 1,659.

For P.863 [6] (MOS-LQOs), there is some compression and offset of the predictions relative to the subjective ratings, but the compression is not as severe as for P.862.2 [10]. Subjective ratings of about 3.5 are predicted as about 2.0 by P.863, and the RMSE is 1.108.

The ETSI TS 103 106-WB [5] scores generally span the same range as the subjective ratings, and fall close to the reference line, without the compression observed in the P.862.2 [10] predictions. The RMSE is 0.196; however for some cases, particularly at lower scores, the error can again be large. These larger errors are not unexpected, since during the training phase of ETSI TS 103 106 [5] there was relatively little data available with low scores.

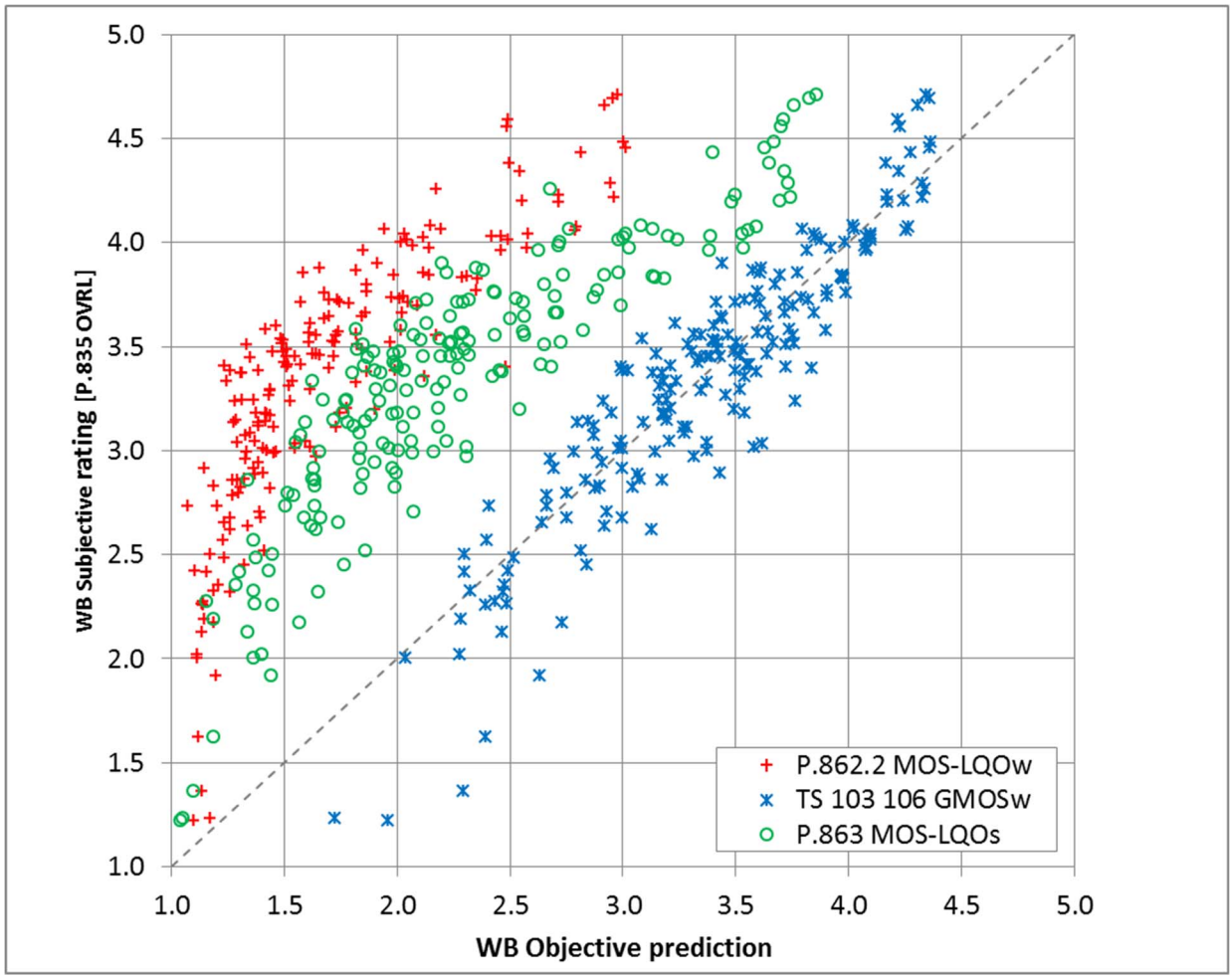


Figure 29: Correlation of P.835 Overall with P.862.2 [10] (MOS-LQOw), ETSI TS 103 106 [5] (WB) GMOSw and P.863 [6] (MOS-LQOs). Maximum error of P.862.2 [10] is 2,192, of P.863 [6] is 1,459, and of ETSI TS 103 106 [5] is 0,424. Spearman rank-order correlation, accounting for 95% confidence interval, for P.862.2 [10] is 0,890, for P.863 [6] is 0,897, and for ETSI TS 103 106 [5] is 0,975

5.4.4.3 WB Rank Order Results

The graphs in figure 30 compare the scores of the various measures for each of the six phones (A-F), in each noise type. Figure 31 then shows the number of absolute rank order errors for each metric in each noise type when compared to the results from the listening test. These results do not take account of overlapping confidence intervals; hence the errors may be exaggerated, especially for the clean condition where the listening tests results were very similar.

For P.862.2 [10] the rank order is not preserved, again there are frequent shifts of one and two positions, as well as shifts of three positions in cafeteria and car noise. For P.863 [6] rank order errors are observed, but not as many as for P.862.2 [10]. There are still frequent single rank switches, but fewer two-rank switches, although there are larger errors observed in the clean condition where the confidence intervals overlap. TS 103 106 [5] has the best performance; however single rank switches are still common and occasional two and three-rank switches also occur in clean, cafeteria and pub noise.

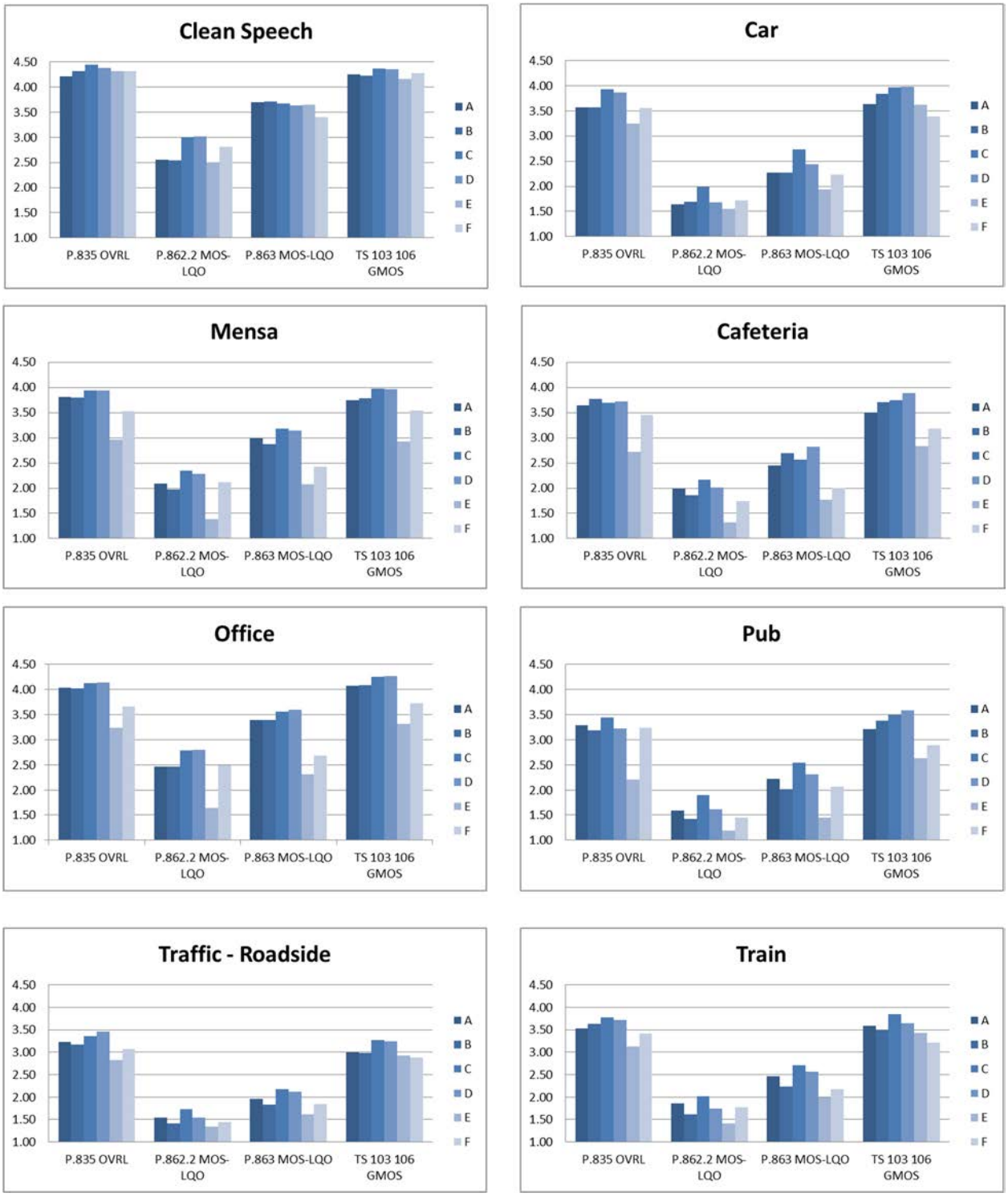


Figure 30: Scores per device and noise type for P.835 Overall [4], P.862. 2[10] (MOS-LQOw), ETSI TS 103 106 [5] (WB) GMOSw and P.863 [6] (WB) (MOS-LQOw)

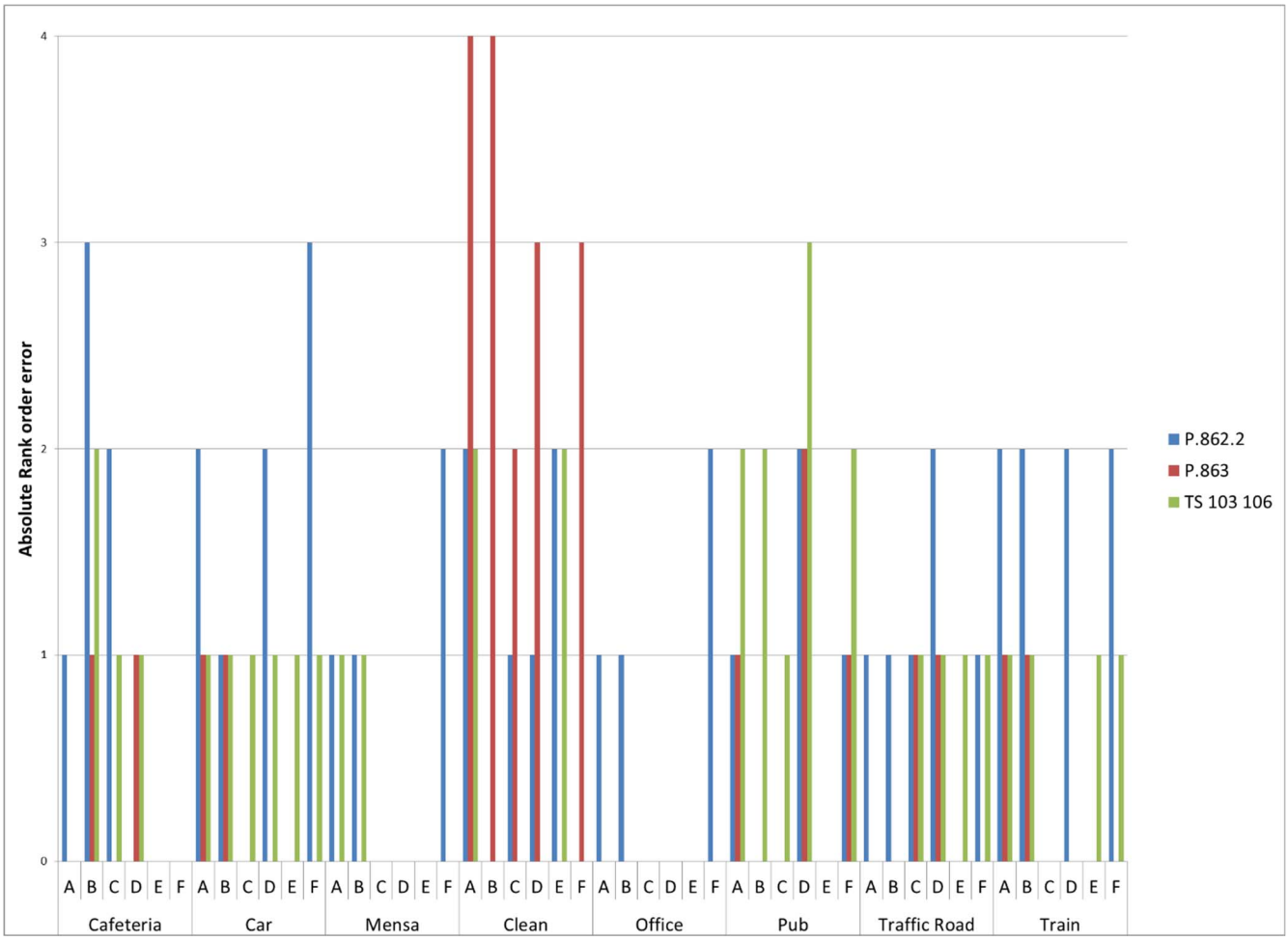


Figure 31: Rank Order Errors for each Objective Measure in WB

5.4.5 Experiment 4 SWB P.835 v P.862.2, P.863 and TS 103 106

5.4.5.1 Setup

The method used was similar to the one described in Clause 5.4.2.1. However, the following differences should be noted. As commercial super-wideband terminals are not generally available, a mock- up of a handset was used. The mock-up was the size and shape of a typical mobile handset, and was equipped with several microphones, as in some current commercially available wideband terminals. The same speech and background noise generation as for NB and WB was used, but recordings were made from the microphones on the mock-up. The signals were processed with offline processing to produce the noise-reduced signals. No speech encoding was used.

For computing metrics, all recordings were taken at 48 kHz sample rate. For P.862.2 [10], the source file was filtered to WB and the scores were computed separately on sentence pairs then averaged to obtain per-condition values. For P.863 [6] version 2.4 was used; the source was filtered to SWB and the scores were computed separately on sentence pairs then averaged to obtain per-condition values. For ETSI TS 103 106 [5], the source file was used as SWB and scores were computed on individual sentences using the WB mode of the tool, and then averaged to obtain per-condition values.

5.4.5.2 SWB Correlation Results

Figure 32 shows a scatter plot of the subjective P.835 [4] OVRL ratings against objective scores. The red-coloured (+) symbols are for P.862.2 (MOS-LQOw) [10], while the blue-coloured (x) symbols are for the GMOS output from the P.835 predictor of [5] and the green-coloured (o) symbols are for P.863 (ITU-T P.863) V2.4 [6].

As expected the P.862.2 [10] results show a poor correlation with the SWB listening scores with an RMSE of 1.458. SWB P.863 [6] scores are more consistent and have an RMSE of 0.815, however they show a similar compression of the range of scores as was observed for NB and WB. Also as expected, the ETSI TS 103 106-WB [5] scores are less well correlated in SWB than in WB with an RMSE of 0.345.

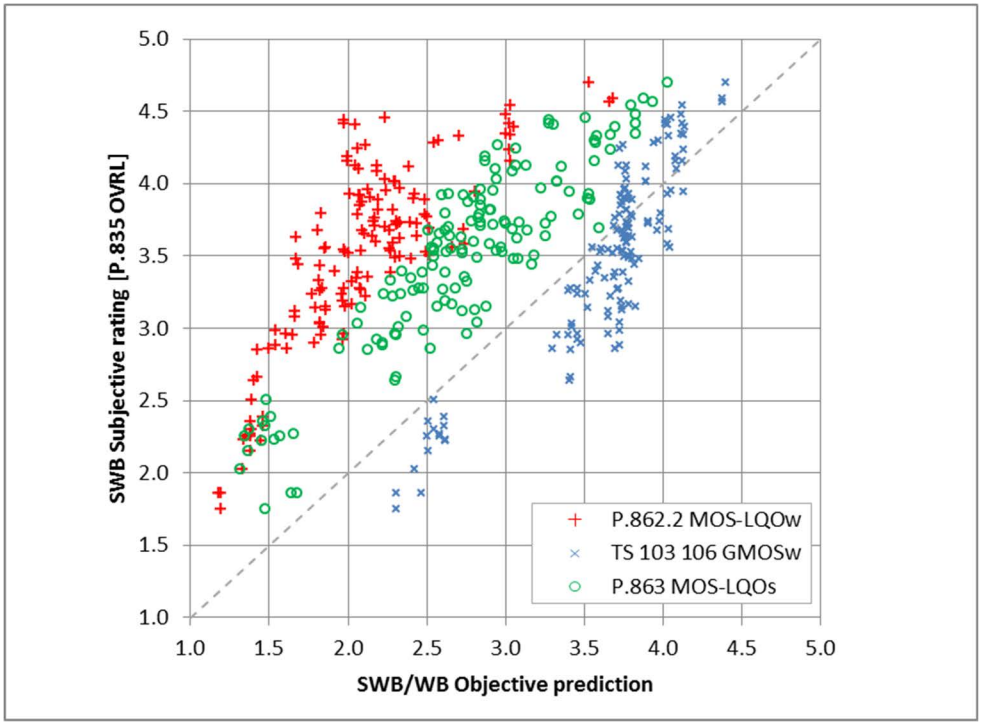


Figure 32: Correlation of P.835 Overall with P.862.2 [10](MOS-LQOw), ETSI TS 103 106 [5] (WB) GMOSw and P.863 [6] (SWB) (MOS-LQOs). Maximum absolute error for P.862.2 [10] is 2,463, for P.863 [6] is 1,326, and for ETSI TS 103 106 [5] is 0,842. Spearman rank-order correlation, accounting for 95 % confidence interval, for P.862.2 [10] is 0,755, for P.863 [6] is 0,855, and for ETSI TS 103 106 [5] is 0,956

5.4.5.3 SWB Rank Order Results

The graphs in figure 33 compare the scores of the various measures for the mock-up phones with various levels of noise suppression (0-15), in each noise type. Figure 34 then shows the number of absolute rank order errors for each metric in each noise type when compared to the results from the listening test.

None of the objective measures preserve the rank order of the listening tests for SWB. All three measures have frequent shifts of up to three ranks and occasional larger shifts. P.863 [6] also commonly shifts by four positions.

It can also be seen from the graphs in figure 33 that if these scores were used to try and tune the noise suppression parameter in this algorithm, different results would be obtained depending on which objective measure was used, and in addition none of the measures would reliably select the same optimization as that determined by the human listeners.

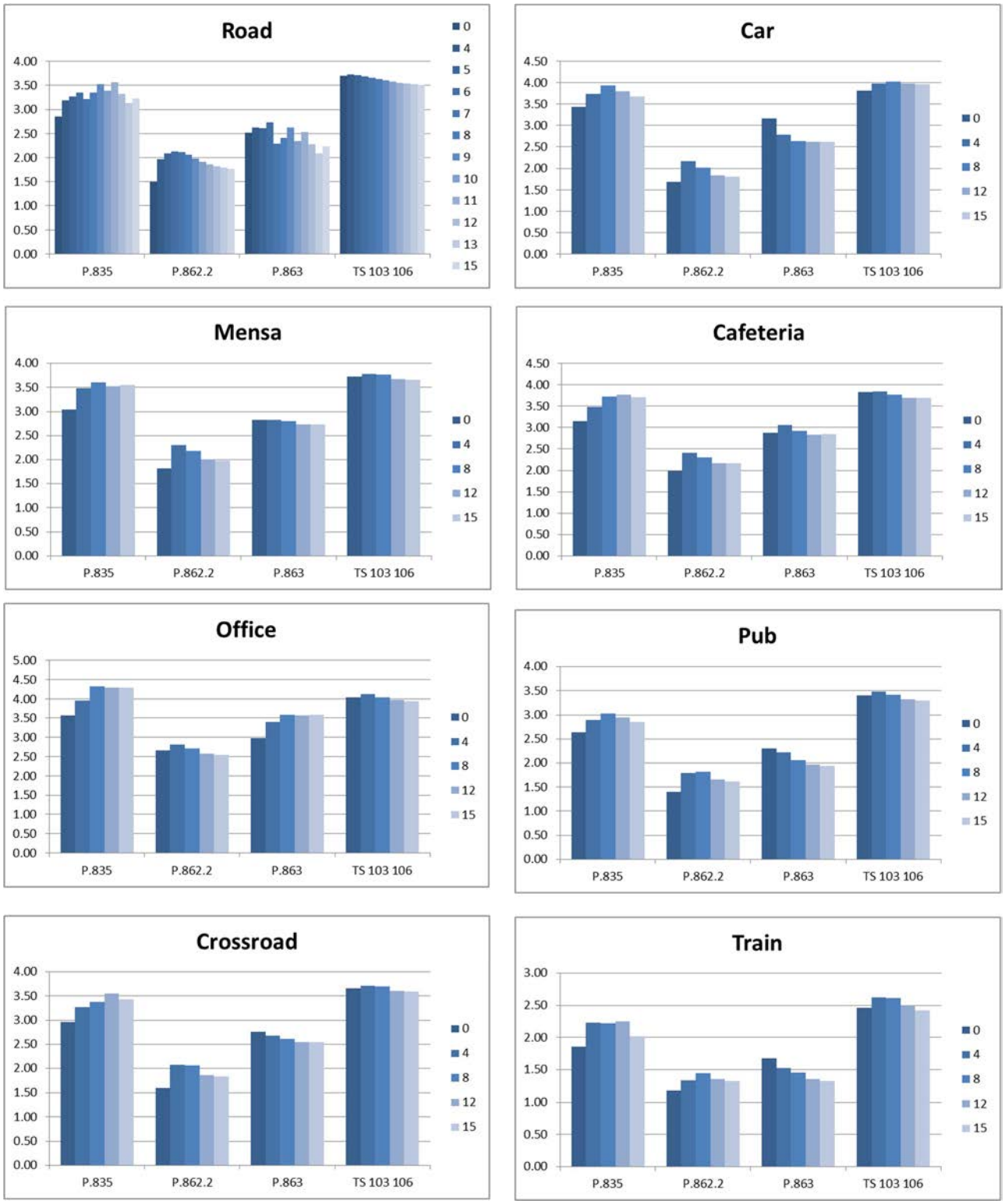


Figure 33: Scores per device and noise type for P.835 [4] Overall, P.862.2 [10] (MOS-LQOw), ETSI TS 103 106 [5] (WB) GMOSw and P.863 [6] (SWB) (MOS-LQOs)

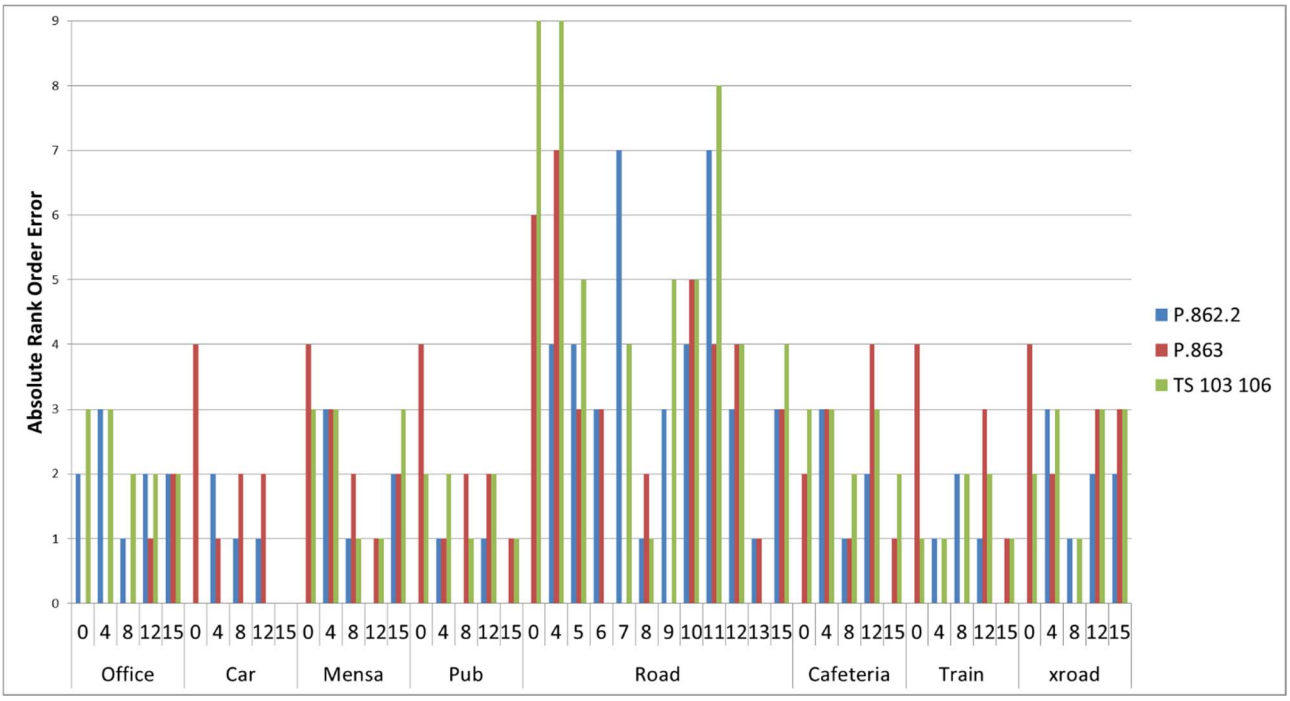


Figure 34: Rank Order Errors for each Objective Measure in SWB

5.4.6 Conclusions

The use of P.862.1 [9] and P.862.2 [10] for assessing the performance of handsets in noise should be avoided. Almost all modern UEs now provide noise suppression as part of their default operation and it is shown that P.862.1 [9] and P.862.2 [10] produce misleading results when used in conjunction with noise suppression algorithms. Using P.862.1 [9] and P.862.2 [10] to compare UE's is unreliable since a comparison of the results with actual listening tests shows a different rank order between the two tests, i.e. the best P.862.1 [9] and P.862.2 [10] score does not always produce the best overall score from a listening test. The P.862.1 [9] and P.862.2 [10] scores (MOS_LQO) also substantially underpredict the P.835 [4] OVRL scores obtained from the listening test. In addition, in Annex B of ETSI EG 202 396-3 [11], results for P.835 [4] listening tests are compared to predictions from P.862.1 [9] and P.862.2 [10] and are shown to not correlate well with the subjective results.

Finally, the use of P.862.1 [9] and P.862.2 [10] for comparing handsets may steer manufacturers to tune their algorithms to maximize the P.862.1 [9] and P.862.2 [10] score. For the reasons exposed, such tuning may actually degrade the speech quality as perceived by human listeners.

For WB it is not too surprising that the objective tools that are intended to predict the perceived quality of telephone speech with noise reduction (e.g. ETSI TS 103 106 [5]) perform that task better than tools which were not initially designed to do so. However there is still room for improvement, as even ETSI TS 103 106 [5] does not consistently preserve the rank order, which can make comparative evaluations unreliable.

For SWB none of the three predictors performed particularly well. Again this is to be expected since P.862.2 [10] and ETSI TS 103 106 [5] were not designed to be used on SWB speech, and P.863 [6] was not designed for use with modern terminal noise suppression algorithms.

Further work has since been done to develop more effective objective measures especially for wider bandwidths ETSI TS 103 281 [20]. It is particularly important to ensure that maximum error and rank order are taken into account as well as just RMSE, which would enable more reliable comparative evaluations of solutions across a range of operation scenarios including different background noises and noise suppression technologies.

5.5 Validation results for combination of model A and B according to ETSI TS 103 281

5.5.1 Introduction

In ETSI TS 103 281 [20], two models for predicting results of ITU-T P.835 [4] evaluations of the sending speech quality in noise are described. Due to a desire by 3GPP SA4 to reference a single model in 3GPP TS 26.132 [8], a model consisting of the combination of the predictions from Model A and Model B of [20] was proposed in [8].

Below are described the combination of the models and an analysis of the performance of the combined model on three validation databases of [20] based on the input from ETSI TC STQ.

5.5.2 Description of combination of model predictions

As defined in [8], the predictions from Model A and Model B of [20] are combined as follows:

$S\text{-}MOS\text{-}LQO_{fb} = (S\text{-}MOS\text{-}LQO_{fb\_modelA} + S\text{-}MOS\text{-}LQO_{fb\_modelB})/2$

$N\text{-}MOS\text{-}LQO_{fb} = (1.438*N\text{-}MOS\text{-}LQO_{fb\_modelA} – 1.959 + N\text{-}MOS\text{-}LQO_{fb\_modelB})/2$

$G\text{-}MOS\text{-}LQO_{fb} = (G\text{-}MOS\text{-}LQO_{fb\_modelA} + G\text{-}MOS\text{-}LQO_{fb\_modelB})/2$

Except for a linear post-mapping of N-MOS from Model A, the combination is simply the average of the corresponding predictions from each Model. The linear post-mapping of N-MOS from Model A was derived from validation databases 3 (DES-25) and 4 (DES-26). Figure 35 shows a scatter plot of unmapped model N-MOS predictions and subjective BAK ratings for both databases separately. In addition, linear regression lines are also shown separately for each database.

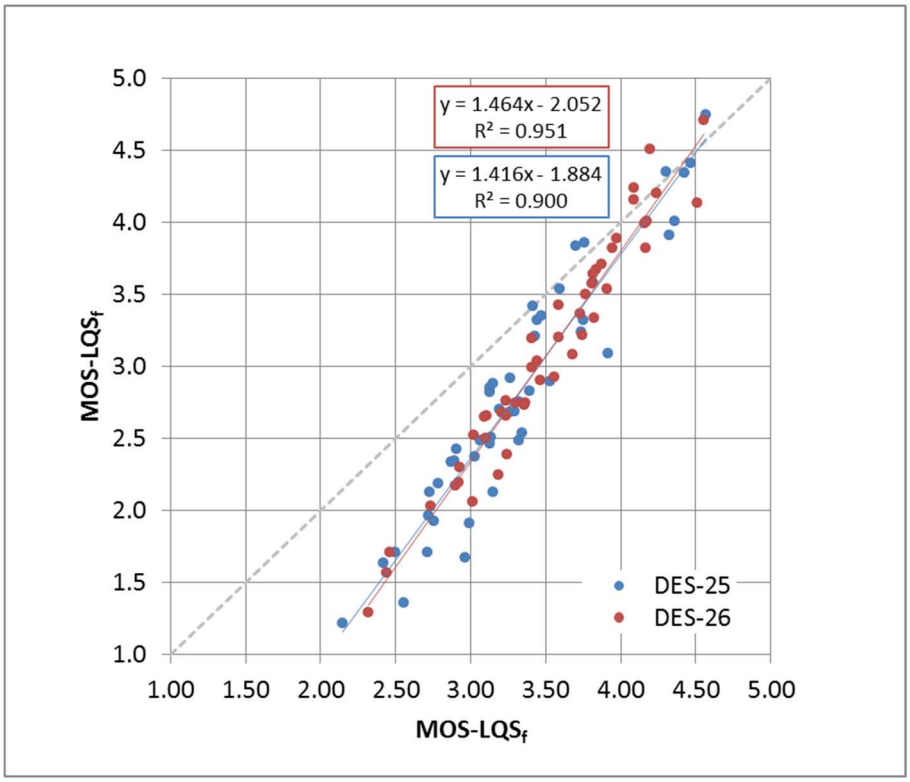


Figure 35: Scatter plot of N-MOS from Model A versus subjective BAK ratings for validation databases DES-25 and DES-26

Note that the scatter plots for both databases are nearly the same, as are the linear regression lines. From this, it was proposed to combine the results from both databases and compute a single regression line, as shown in Figure 36.

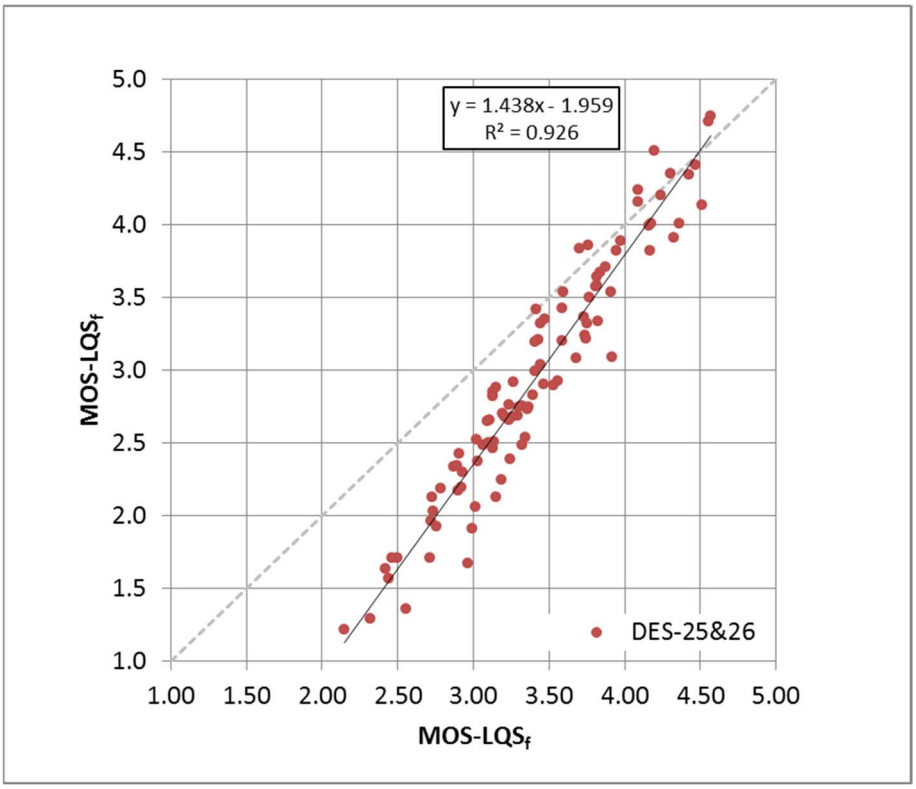


Figure 36: Scatter plot of N-MOS from Model A versus subjective BAK ratings for combined validation databases DES-25 and DES-26

The regression equation obtained from the combined databases was then used as post-mapping for the N-MOS predictions from Model A. Using the combined model, comparisons to validation databases 3, 4, and 5 are shown in the following clauses.

5.5.3 Validation database 3 (DES-25): Results for combined model

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from the combined model on validation database 3 for each of the three ratings, SIG, BAK, and OVRL, are shown in Figure 37. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic 3rd order mapping is applied (left column).

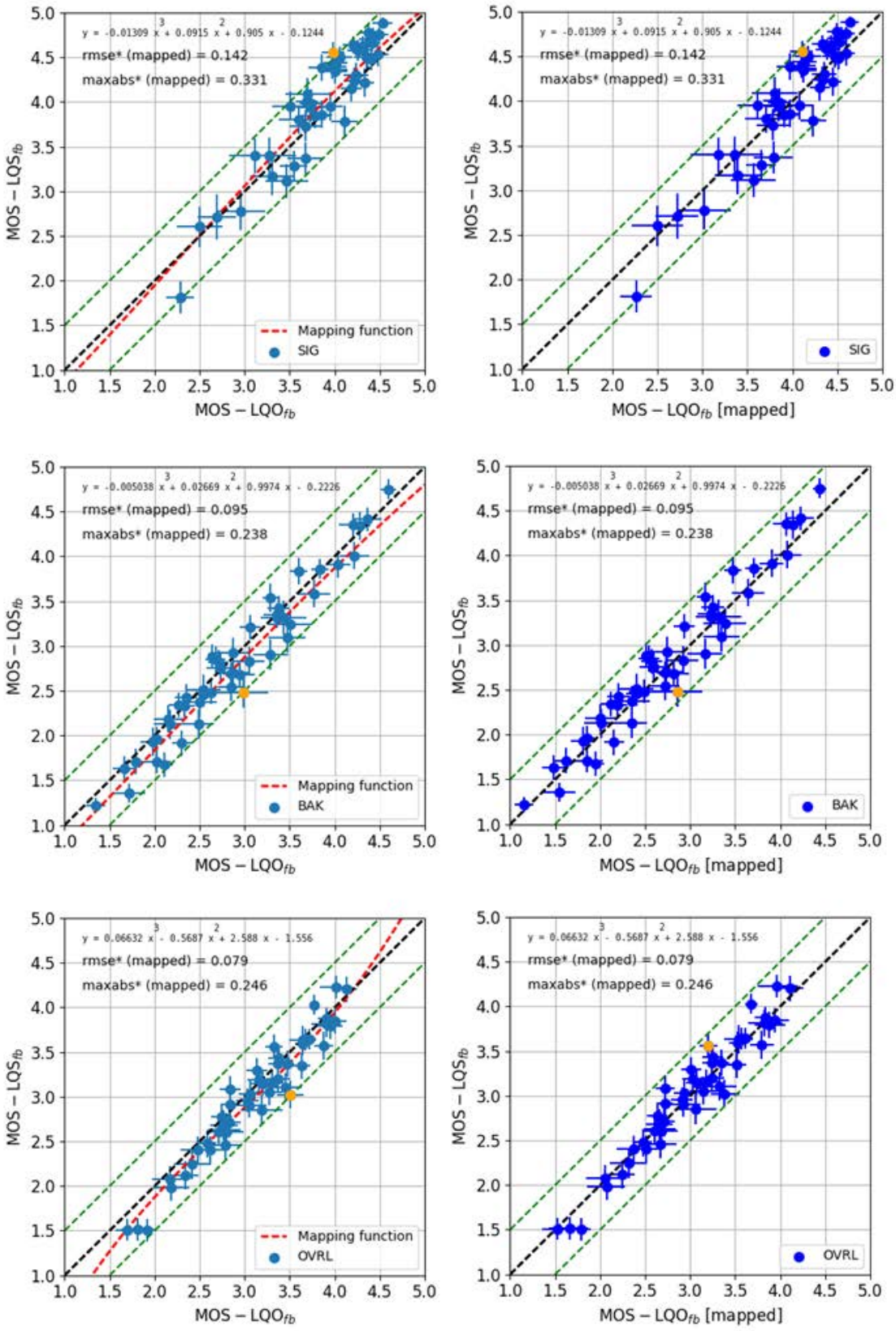


Figure 37: Scatter plots from combined model for validation database 3 (DES-25)

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with and orange-coloured symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ± 0.5 MOS. The error bars indicate the 95% confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in Table 14.

Table 14: Performance metrics for combined model on validation database 3 (DES-25)

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,299	0,256	0,194	0,142
	Max Abs Error	0,572	0,450	0,452	0,331
	Pearson's ρ	0,941	0,942	0,972	0,980
	Spearman's rank order ρ	0,902	0,902	0,942	0,978
	Kendall's τ	0,740	0,740	0,821	0,869
BAK	Rmse	0,210	0,207	0,111	0,095
	Max Abs Error	0,504	0,377	0,333	0,238
	Pearson's ρ	0,975	0,974	0,992	0,995
	Spearman's rank order ρ	0,968	0,968	0,987	0,988
	Kendall's τ	0,854	0,854	0,926	0,932
OVRL	Rmse	0,195	0,168	0,094	0,079
	Max Abs Error	0,480	0,359	0,325	0,246
	Pearson's ρ	0,972	0,972	0,991	0,993
	Spearman's rank order ρ	0,965	0,965	0,989	0,993
	Kendall's τ	0,854	0,854	0,940	0,952

5.5.4 Validation database 4 (DES-26): Results for combined model

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from the combined model on validation database 4 for each of the three ratings, SIG, BAK, and OVRL, are shown in Figure 38. As in the previous figure, for each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

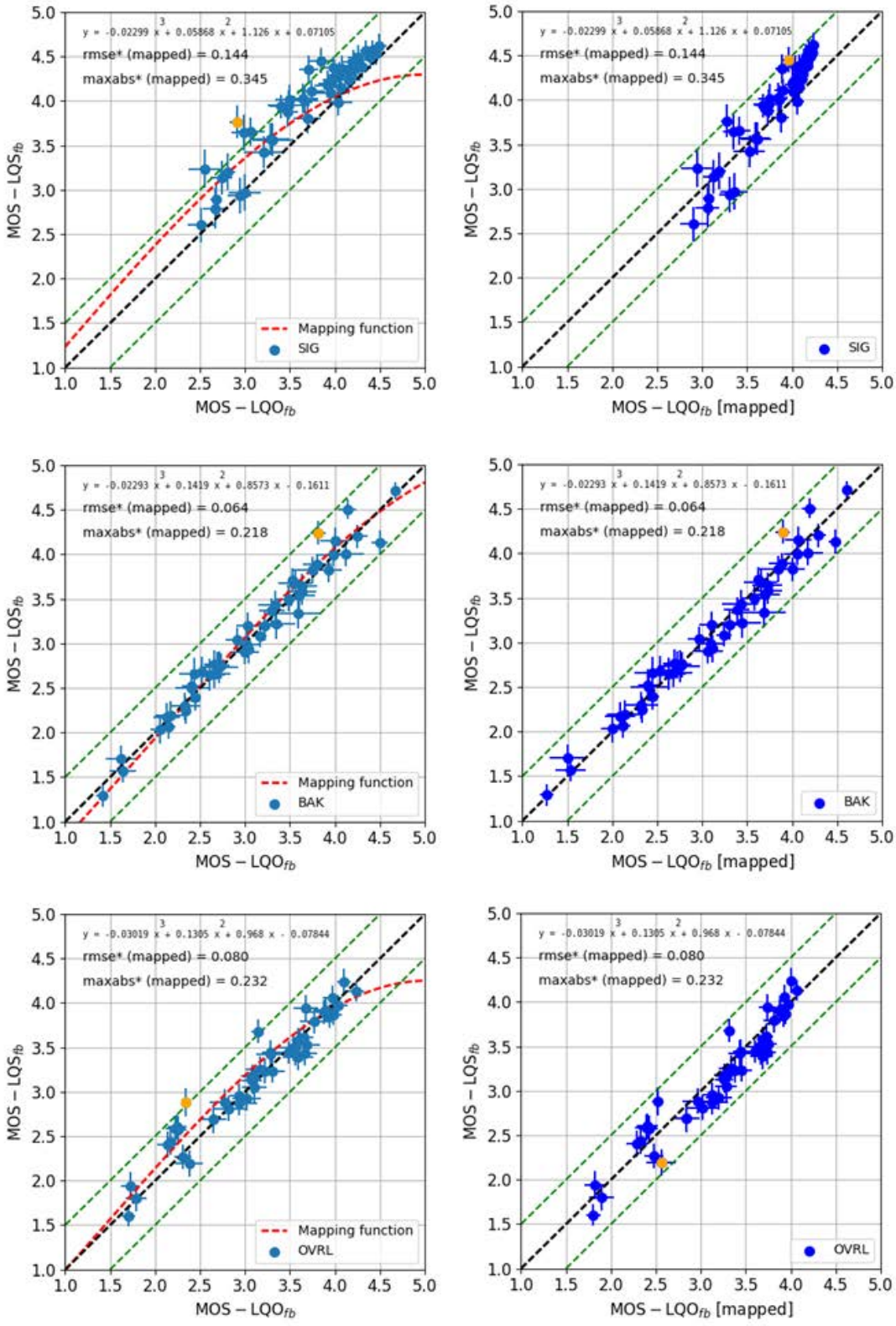


Figure 38: Scatter plots from combined model for validation database 4 (DES-26)

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with and orange-coloured symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ± 0.5 MOS. The error bars indicate the 95% confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in Table 15.

Table 15: Performance metrics for combined model on validation database 4

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,345	0,276	0,224	0,144
	Max Abs Error	0,858	0,488	0,669	0,345
	Pearson's ρ	0,938	0,945	0,957	0,976
	Spearman's rank order ρ	0,945	0,945	0,955	0,962
	Kendall's τ	0,758	0,758	0,783	0,881
BAK	Rmse	0,139	0,147	0,069	0,064
	Max Abs Error	0,434	0,347	0,305	0,218
	Pearson's ρ	0,985	0,985	0,996	0,997
	Spearman's rank order ρ	0,988	0,988	0,998	0,997
	Kendall's τ	0,917	0,917	0,982	0,973
OVRL	Rmse	0,183	0,188	0,102	0,080
	Max Abs Error	0,538	0,363	0,405	0,232
	Pearson's ρ	0,967	0,963	0,991	0,993
	Spearman's rank order ρ	0,968	0,968	0,992	0,989
	Kendall's τ	0,863	0,863	0,954	0,932

5.5.5 Validation database 5 (DES-27): Results for combined model

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from the combined model on validation database 5 for each of the three ratings, SIG, BAK, and OVRL, are shown in Figure 39. As in the previous figures, for each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

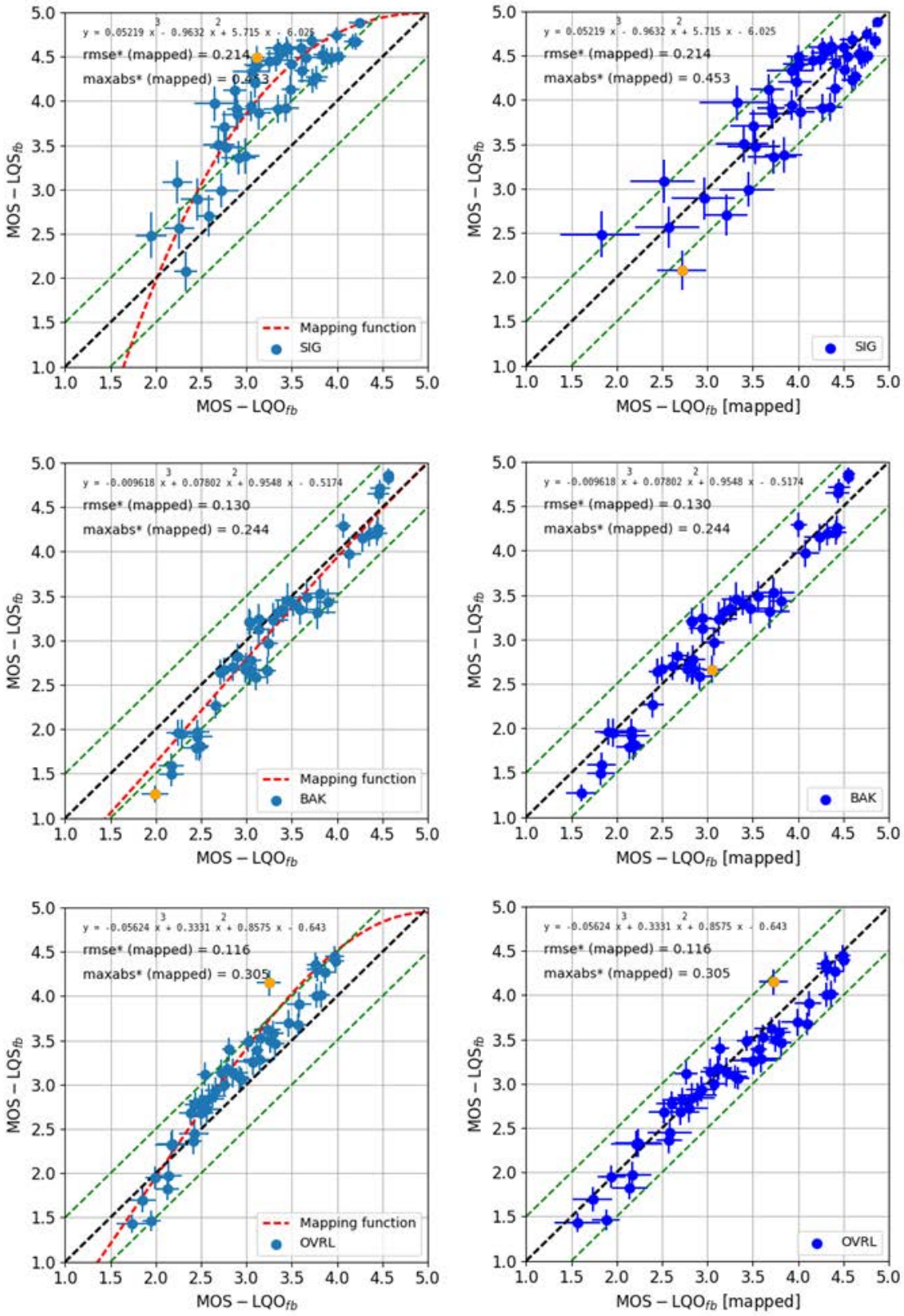


Figure 39: Scatter plots from combined model for validation database 5 (DES-26)

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with and orange-coloured symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ± 0.5 MOS. The error bars indicate the 95% confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in Table 16.

Table 16: Performance metrics for combined model on validation database 5

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,870	0,351	0,737	0,214
	Max Abs Error	1,383	0,648	1,260	0,453
	Pearson's ρ	0,849	0,888	0,867	0,960
	Spearman's rank order ρ	0,849	0,849	0,855	0,955
	Kendall's τ	0,672	0,672	0,681	0,835
BAK	Rmse	0,358	0,242	0,248	0,130
	Max Abs Error	0,713	0,388	0,616	0,244
	Pearson's ρ	0,971	0,971	0,984	0,992
	Spearman's rank order ρ	0,962	0,962	0,974	0,989
	Kendall's τ	0,857	0,857	0,890	0,940
OVRL	Rmse	0,343	0,217	0,233	0,116
	Max Abs Error	0,902	0,421	0,760	0,305
	Pearson's ρ	0,965	0,970	0,972	0,991
	Spearman's rank order ρ	0,975	0,975	0,973	0,994
	Kendall's τ	0,884	0,884	0,878	0,950

5.5.6 Conclusions

The provided data indicated that the combined model A and B performs better than the individual models on their own. The results support the use of the combined model in TS 26.132 [8].

6 Conclusions

The present document presented several further analyses and studies related to the terminal testing specification TS 26.132 [8].

A large auditory evaluation was conducted to investigate the relation between human perception and echo control characteristics ("double talk performance" according to clauses 7.11 and 8.11 of [8]). However, even though extensive analyses on instrumental and auditory data were carried out, adequate and reasonable requirements for TS 26.131 [13] could still not be derived for this measurement method.

Clause 5 of the present document presented several studies regarding new and/or advanced acoustic testing of terminals.

The usage of alternative and more challenging handset positions as well as the background noise simulation system in ETSI TS 103 224 [14] were evaluated for handset devices. Results on both methods are presented in this document.

Another study investigated the usage of several speech quality prediction models for noise-suppressed speech. Even though the methods according to ITU-T P.862 [9] and P.862.2 [10] explicitly excluded this application in their scope, they are widely used for this purpose. The systematic score under-prediction of P.862.2 was analysed in ITU-T SG12 and the incorrect transformation of the sound pressure level was found as a reason of the bias; the Corrigendum in [23] improves the performance. Substantial amounts of auditory tests were compared with such metrics, concluding that they provide poor correlation with the subjective data. Even the more recent speech quality prediction method according ITU-T P.863 [6] does not perform well for noise-suppressed speech.

The recently introduced method for assessing the quality of noise-suppressed SWB/FB speech according to ETSI TS 103 281 provides two models (A and B). In order to benefit from the performance of both models, a combined approach is presented as a separate analysis here. The data of two validation listening test databases indicate that the performance of an average model performs better than the single ones and confirms the usage of this approach in TS 26.132 [8]. Further terminal investigations based on ETSI TS 103 281 is expected to provide the basis for agreement on adequate and reasonable requirements/objectives for TS 26.131 [13].

Annex A:
Change history

Change history							
Date	**TSG #**	**TSG Doc.**	**CR**	**Rev**	**Subject/Comment**	**Old**	**New**
2017-09	77	SP-170620			Presented to TSG SA#77 for information		1.0.0
2017-12	78	SP-170829			Presented to TSG SA#78 for approval	1.0.0	2.0.0
2017-12	78				Approved at TSG SA#78 for Release 15	2.0.0	15.0.0
2020-07	-	-	-	-	Update to Rel-16 version (MCC)	15.0.0	**16.0.0**

History

Document history		
V16.0.0	September 2020	Publication