Source: **Audience, Inc.**

Title: **Additional inter-lab comparison of test results according to EG 202-396-3**

Document for: **Discussion**

Agenda Item: **8.5**

1. Introduction

Contribution S4-110942 [1] reports on results obtained according to ETSI EG 202 396-3 for two WB terminals from three labs, using eight different types of noise. This contribution reports additional results obtained according to ETSI EG 202 396-3 for eight NB terminals from two labs, using six different types of noise. The objective of both sets of data is to examine the inter-lab repeatability of the ETSI EG 202 396-3 method.

2. Measurement conditions

The two labs each implemented a background noise simulation consistent with ETSI EG 202 396-1, following the calibration and equalization procedures provided by the test equipment manufacturer. Measurements were taken to ensure that the acoustic requirements on ambient noise and reverberation time were met. Additional measurements of the two test rooms according to IEEE-269 were followed to ensure uniformity of diffuse field within a 15-cm radius sphere centered on the HATS reference point.

Six background noise types were selected, as listed in Table 1.

Noise Type	Source	File
Car	EG 202 396-1	Fullsize_Car1_130kmh_binaural
Pub	EG 202 396-1	Pub_binaural_V2
Street	EG 202 396-1	Outside_Traffic_Road_binaural
Music	G.160Amd2 [2]	NoiseMusic
Voice	G.160Amd2 [2]	NoiseVoice
Pink	N/A	2-ch uncorrelated

Table 1: Noise Types and Files

All noise types were presented at their nominal level, and also at levels 6-dB higher and 6-dB lower. The three signals from EG 202 396-1 were presented at the nominal level listed therein. The nominal level for the three additional noise types was set to 70 dB(A).

The speech signal is the single-talk sequence comprised of single sentences from each of six male and six female talks as described in the update of P.501, Section 7.3.2.1, preceded with a 15-second conditioning sequence that included sentences in sending and receiving direction for activation of the device [3]. The speech was presented through an equalized HATS mouth at a level of -4.7dBPa at MRP.

All devices used are commercially available NB terminals. Four are UMTS and used AMR-NB 12.2kbps; four are CDMA and used EVRC. A base-station simulator (CMU-200) was used for all devices.

The computations according to EG 202 396-3 in narrowband mode were made on the portion of the recordings that included the single-talk sequence, without including the 15-second initial interval.

3. Results

The summary results of the study are shown below:

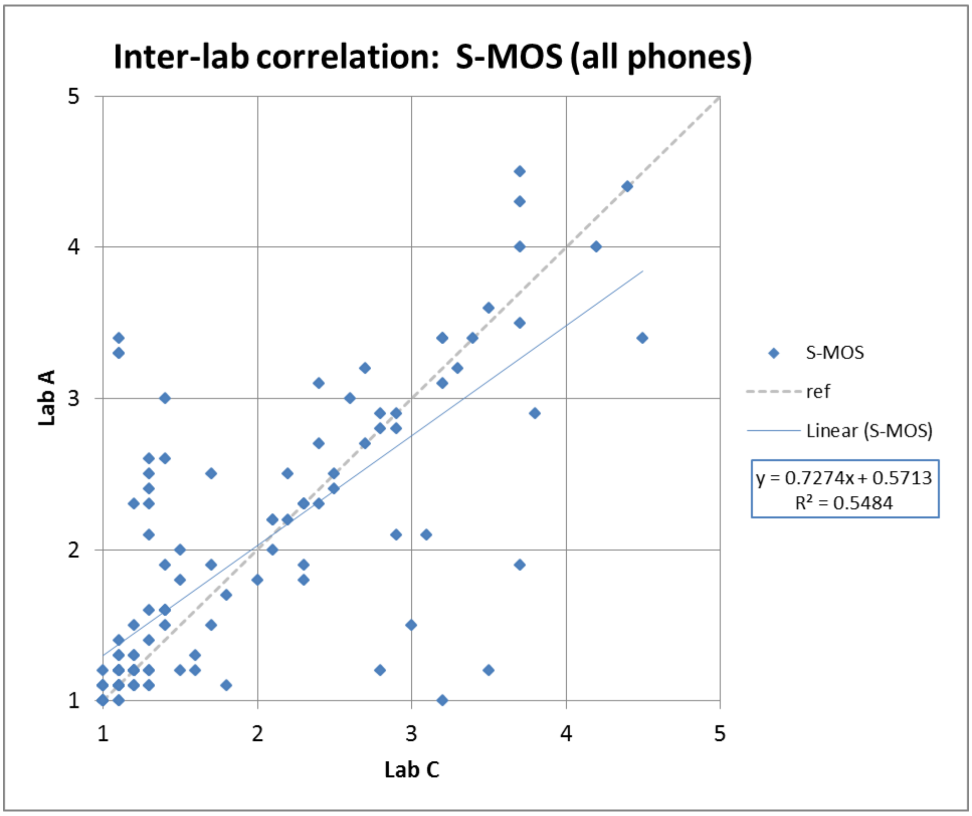


Figure 1: Inter-lab correlation for S-MOS (all phones)

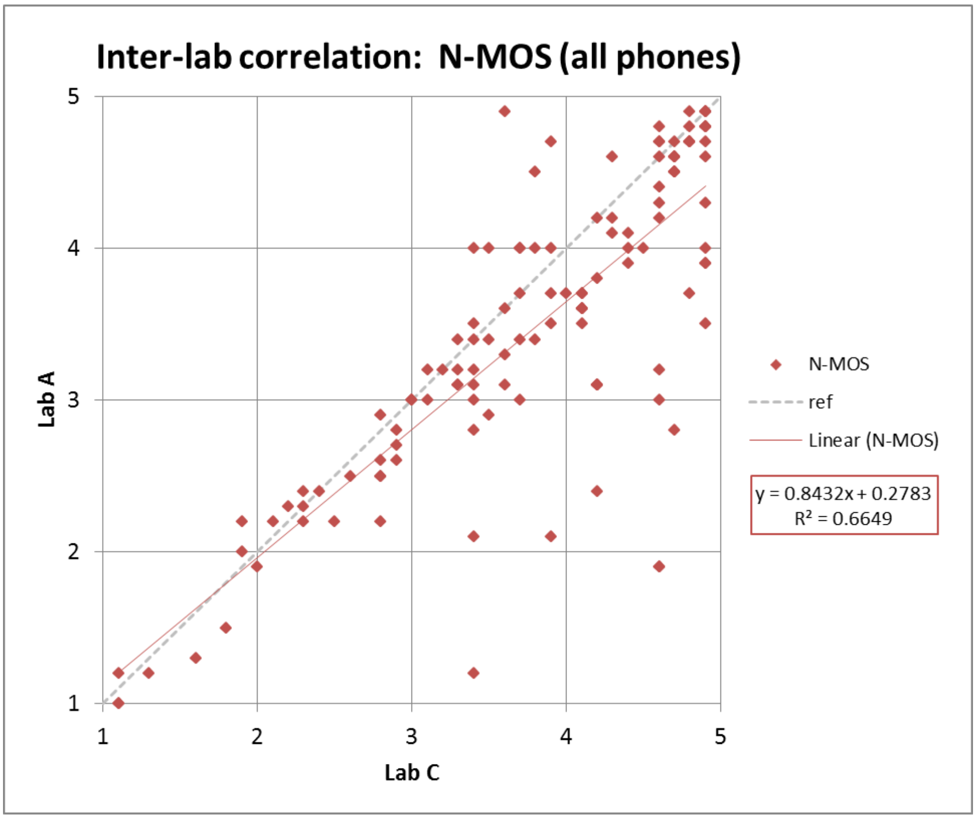


Figure 2: Inter-lab correlation for N-MOS (all phones)

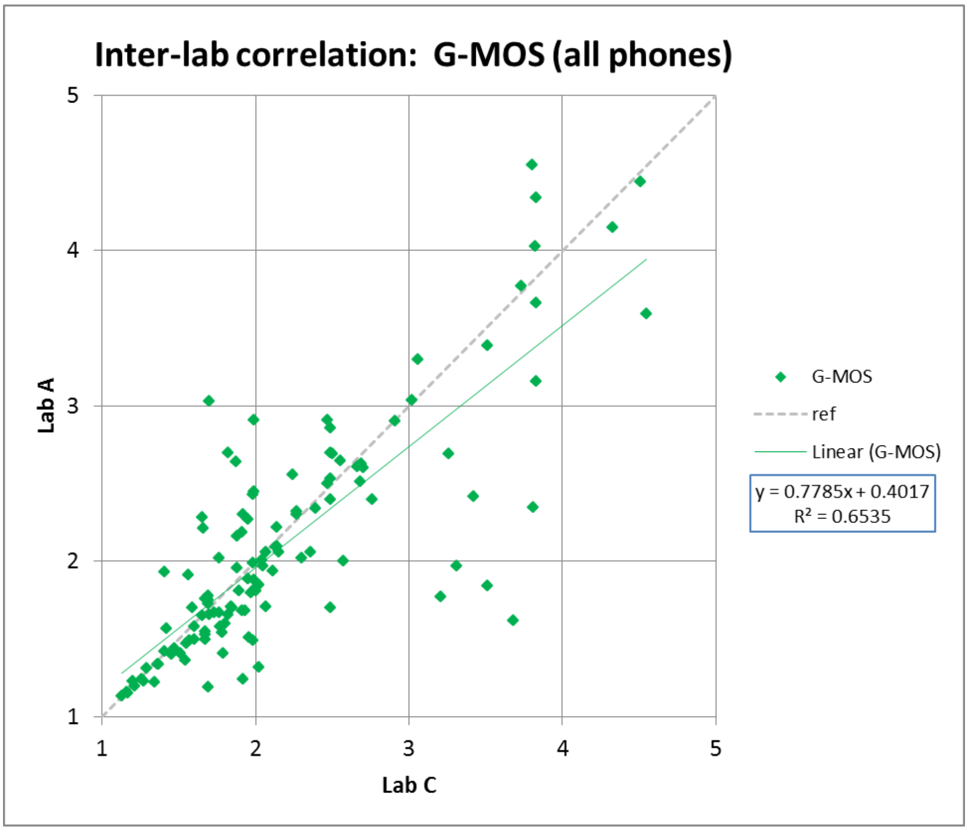


Figure 3: Inter-lab correlation for G-MOS (all phones)

The poor correlation across all eight phones for ETSI 202-396-3 is very similar to the poor correlation for G.160 SNRI for the same data set, as shown below (and similar to the poor correlation seen in [4, figure 1]):

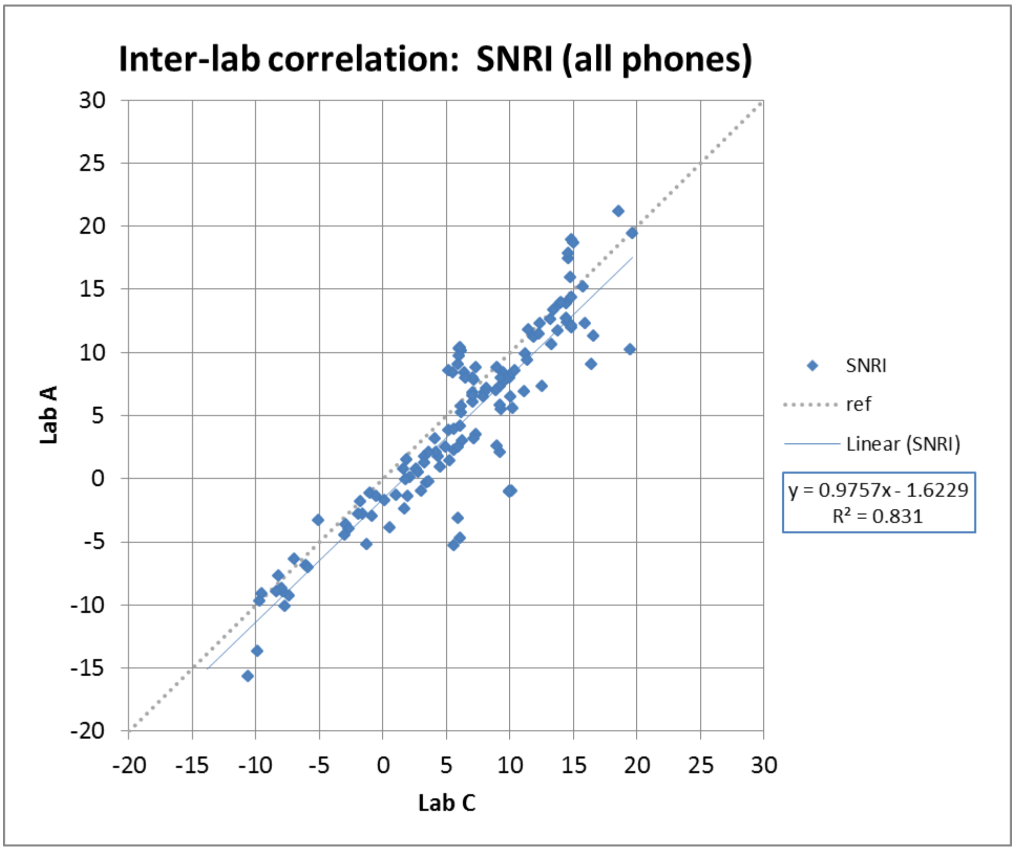


Figure 4: Inter-lab correlation for SNRI (all phones)

However, we noted that there are a few devices which show very good inter-lab correlation, similar to the original G.160 SNRI study [4, figure 3]. In particular, the correlation for device ID-007 is shown below (with one outlier removed, as described and justified in Appendix I):

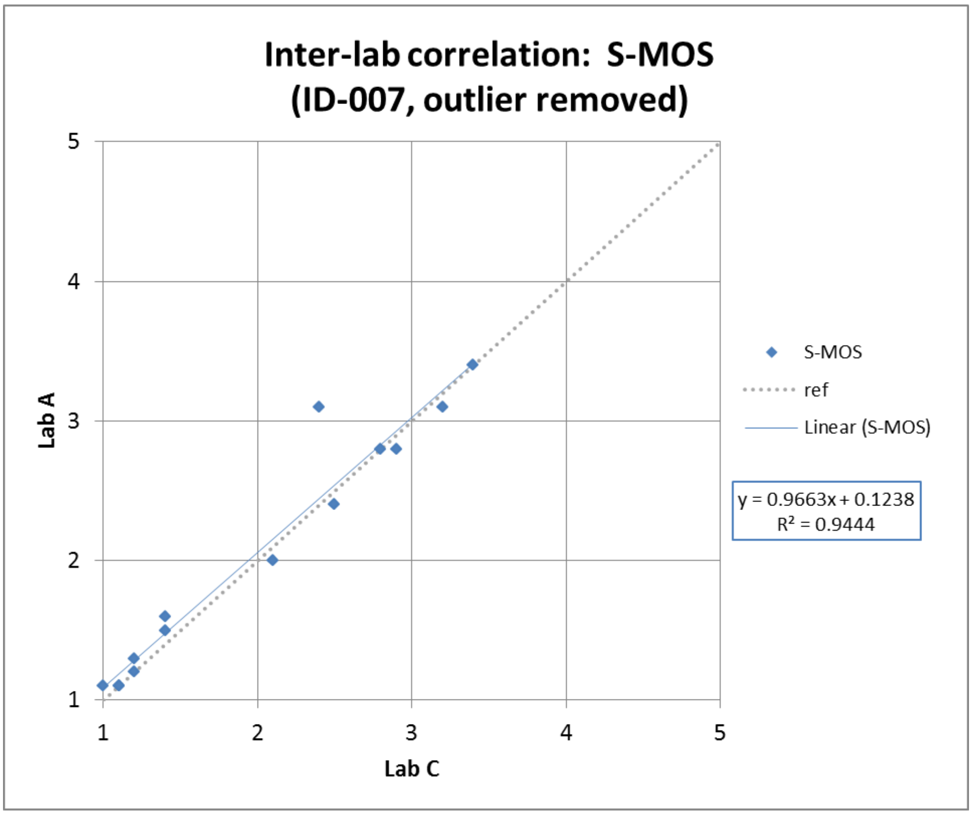


Figure 5: Inter-lab correlation for S-MOS (ID-007)

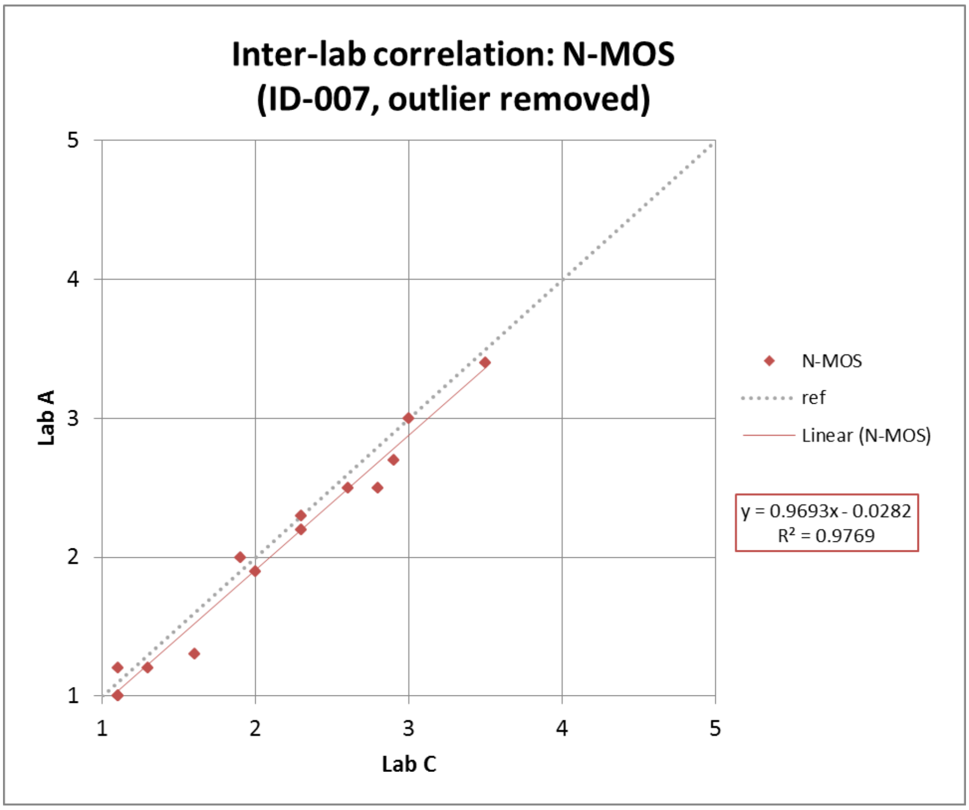


Figure 6: Inter-lab correlation for N-MOS (ID-007)

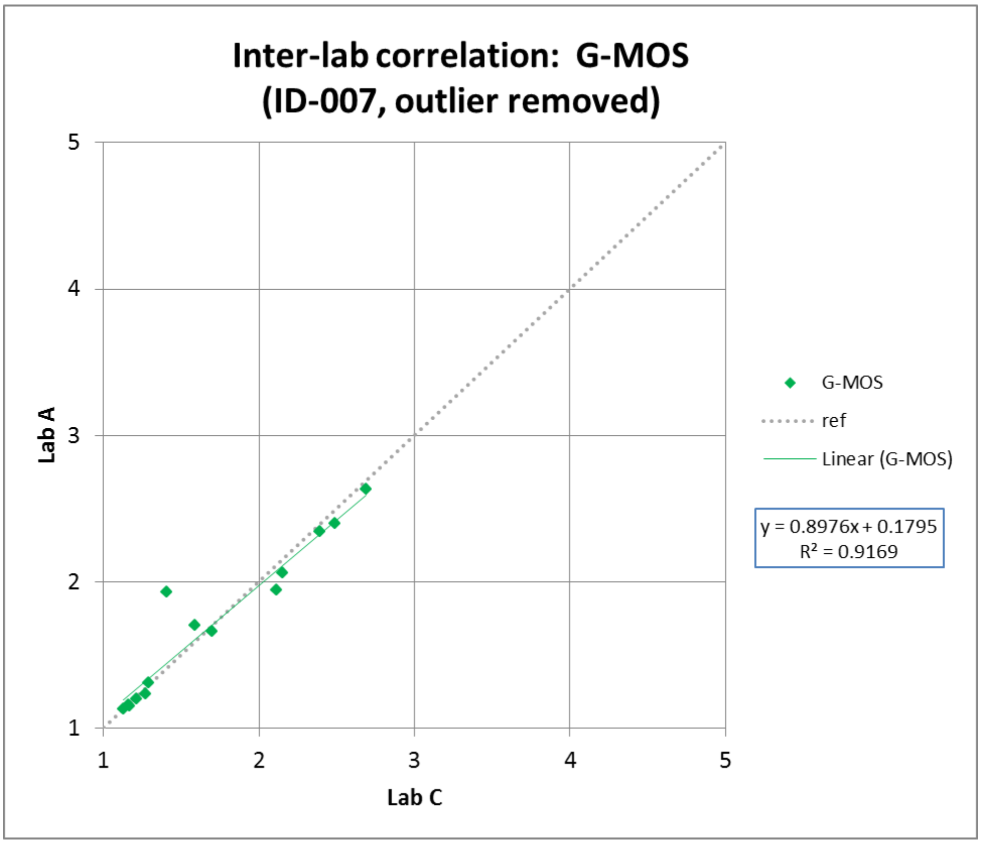


Figure 7: Inter-lab correlation for G-MOS (ID-007)

Similarly, it is possible to find a device (ID-008) for which there was a good inter-lab correlation in the SNRI measurement:

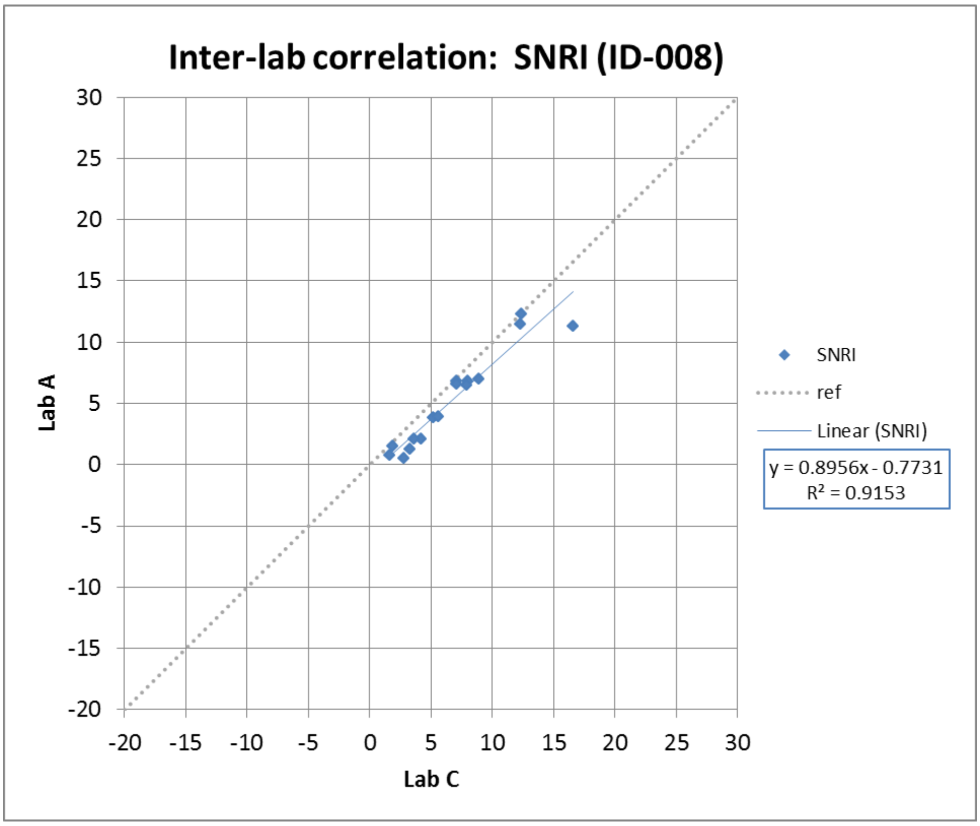


Figure 8: Inter-lab correlation for SNRI (ID-008)

4. Discussion

We notice that the observations reported here are very similar to the G.160 results reported in [4], namely that there can be generally poor inter-lab correlation for both G.160 SNRI and for ETSI EG 202-396-3 when using the recommended setup procedures from ETSI EG 202-396-1 and the test equipment manufacturer, but good correlation for at least one device.

We observe that EG 202 396-1 is common to both G.160 and EG 202 396-3 methods. With the available data, we cannot yet separately assign sources of variation to acoustic set-up (EG 202 396-1), computations (G.160 or EG 202 396-3, including pre-processing such as time alignment), terminal state behaviour (see Appendix), interactions between the above factors, or other factors.

Finally, these results would appear to differ from the positive result in S4-110942, in which generally good inter-lab correlation was found between three labs for two WB terminals using eight different types of noise [1]. Possible explanations for the different findings could include the small number of phones (only 2) reported in [1], possible different calibration procedures used in multiple iterations in [1].

5. Conclusions

It remains challenging to achieve good inter-lab correlation for either G.160 SNRI or ETSI EG 202 396-3 in a study with many phones. Sources of variability may include acoustic set-up (EG 202 396-1), computations (G.160 or EG 202 396-3, including pre-processing such as time alignment), terminal state behaviour (see Appendix), interactions between the above factors, or other factors.

More complete assessment of inter-lab repeatability of EG 202 396-3 or other methods such as G.160 that are based on EG 202 396-1, is needed before any conclusions on reliability of any of the methods can be made.

References

[1] S4-110942 Interlab comparison of tests results according to EG 202 396-3, 3GPP SA4#66, 7-11 November 2011, Jeju, Korea.

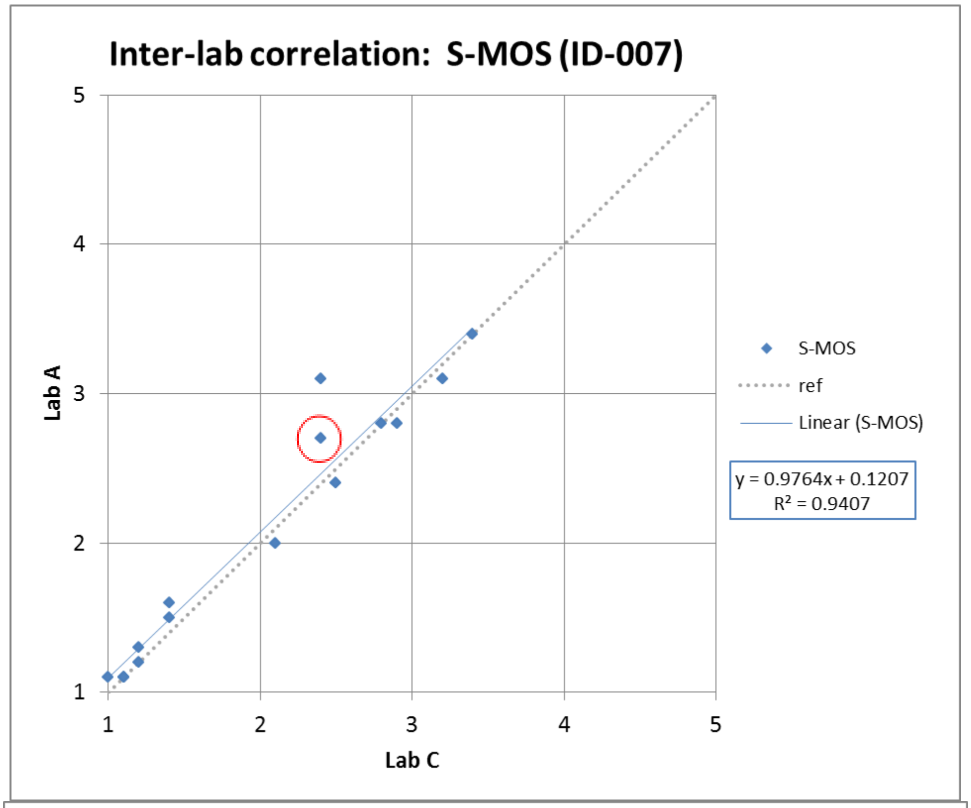
[2] ITU-T Rec. G.160 Amendment 2, Revised Appendix II, Software distribution (03/2011).

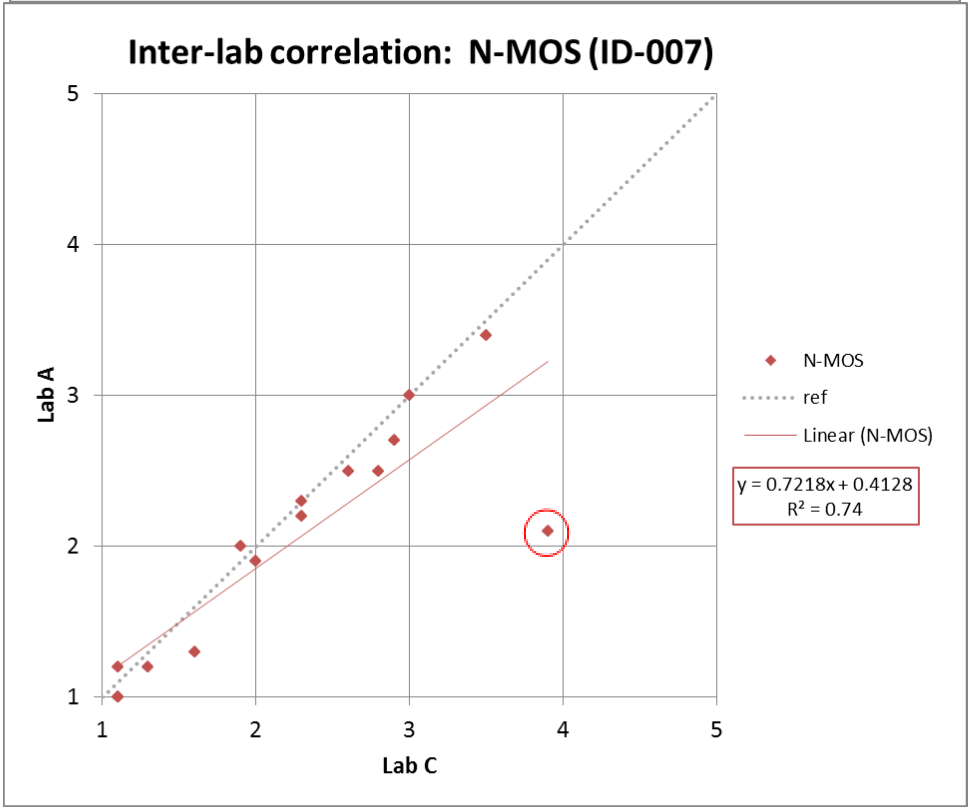
[3] COM 12 – C 255, Revision to Rec. ITU-T P.501, Test signals for use in telephonometry, October, 2011, Geneva, Switzerland.

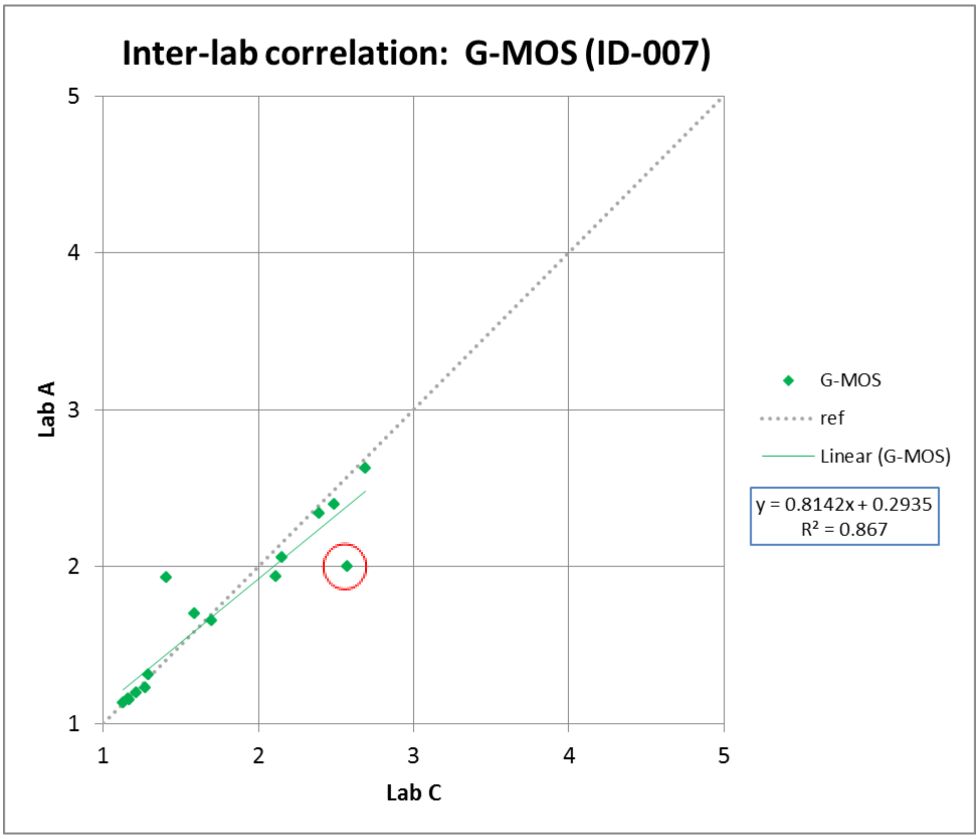
[4] S4-110598 Preliminary analysis of results of ANR round robin test, 3GPP SA4#65, 15-19 August, 2011, Kista, Sweden.

Appendix I:

It was observed for device ID-007 that the correlation for both S-MOS and G-MOS was quite good (R^2 = 0.9407 and 0.867, respectively), but the correlation for N-MOS was much poorer (R^2 = 0.74), as shown:

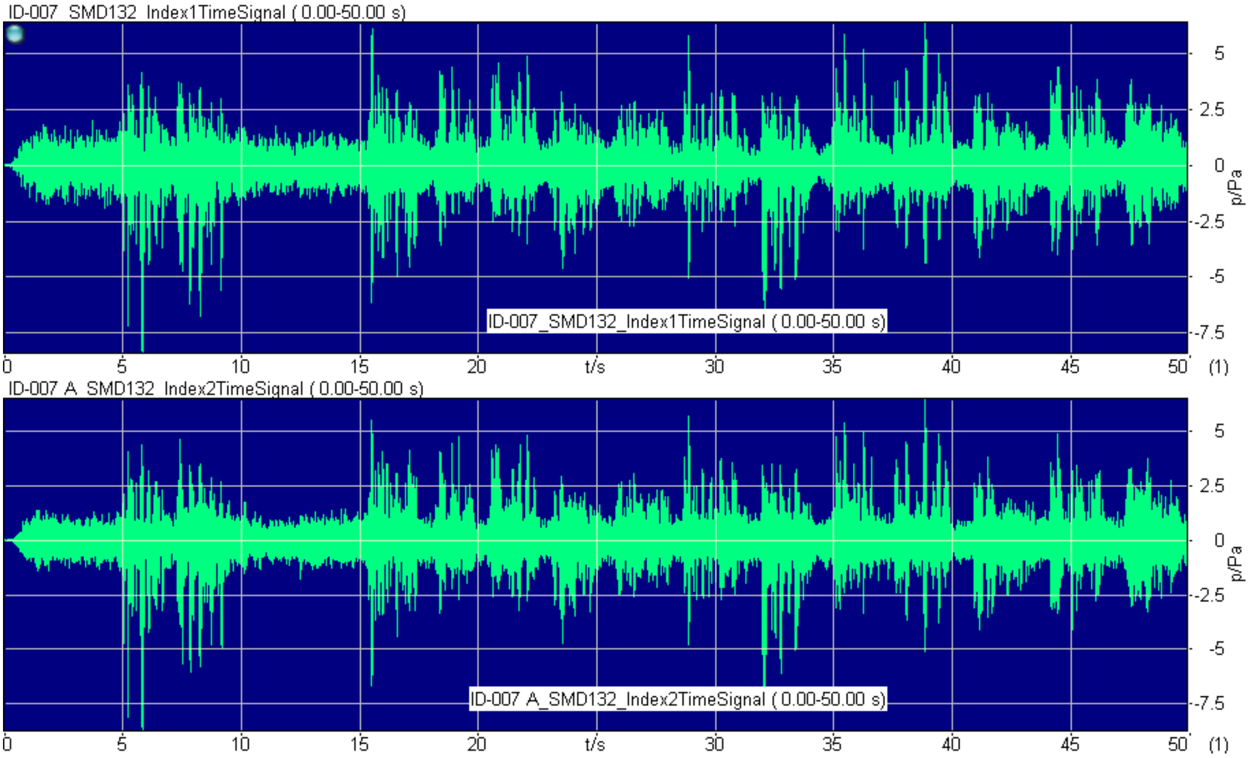




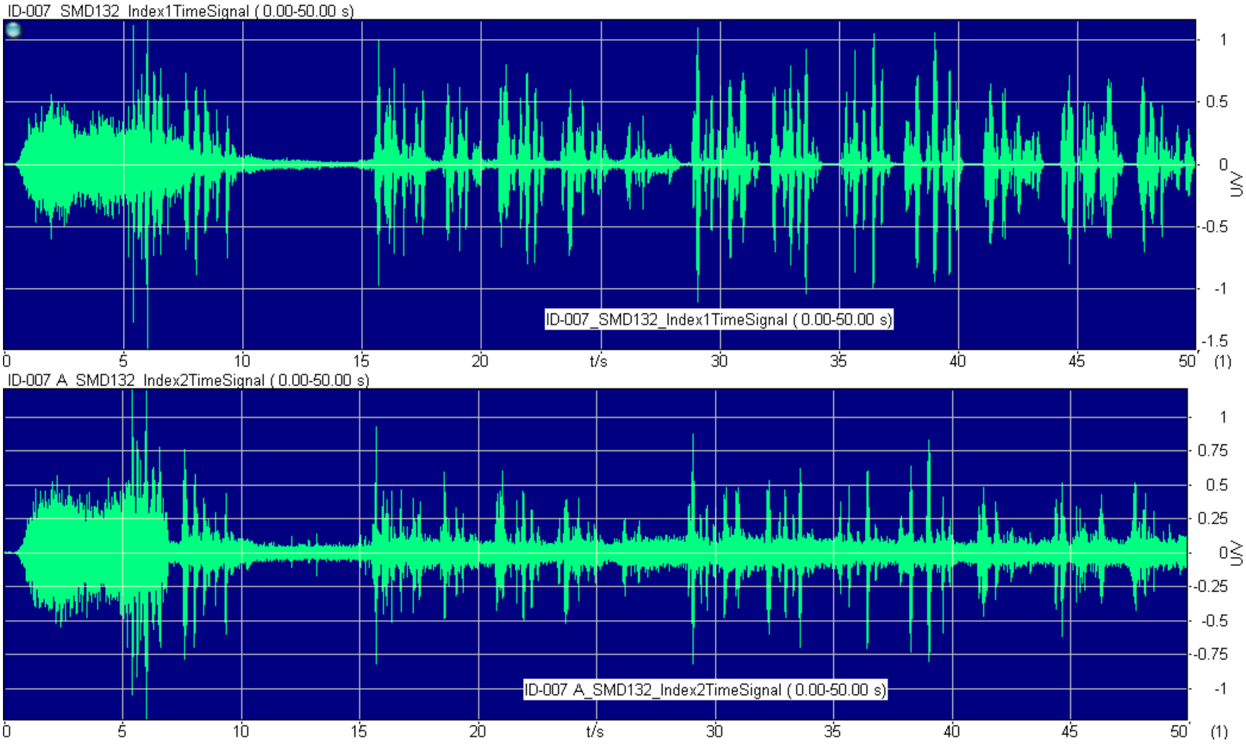


It was clear that the N-MOS correlation was being significantly degraded by a single outlier point, as circled in red. We examined the waveforms for this particular measurement, and found the input files as recorded by the two labs to be very similar, but the output files to be very different, suggestive that the device was in different states in the two different labs. This was confirmed by a detailed examination and listening test of the recorded waveforms:

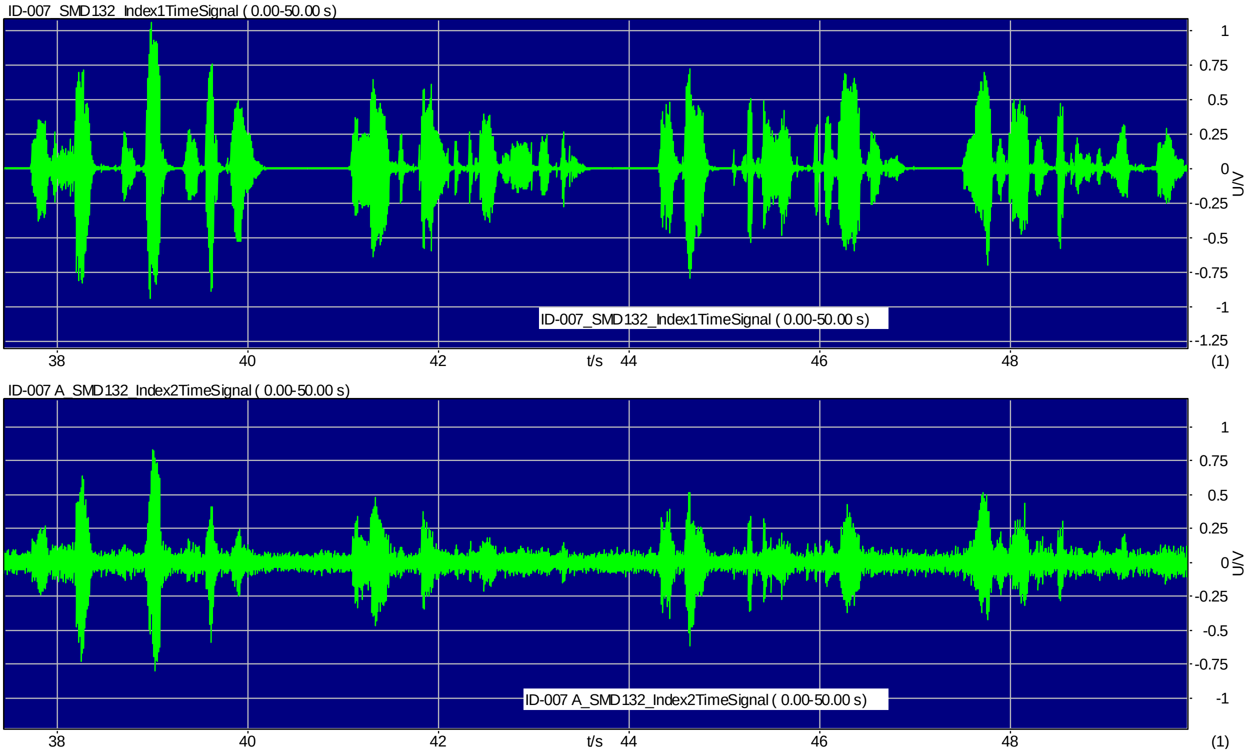
Noisy Speech input signals: Upper panel for Lab C, lower panel for Lab B



Output signals: panel assignment as above.



Last 12 seconds of output where it is easier to see differences:



This suggests that further precautions must be taken to ensure that the devices remain in the same states in the two labs during the experiments, in order to get repeatable measurements from either method (G.160 or ETSI EG 202 396-3). This is beyond the scope of this document, but the observation that the device was in different states in the two labs is sufficient to justify treating the point as an outlier and considering the inter-lab correlation without it.