Third-Party Listening Test as an Alternative to Conversational Tests

Stefan Bleiholder[1], Martin Spiertz[2], Frank Kettler[1]

[1] HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail: stefan.bleiholder@head-acoustics.com

[2] Technische Hochschule Würzburg-Schweinfurt, 97421 Schweinfurt, E-Mail: martin.spiertz@thws.de

Introduction

A Conversational Test (CT) was carried out in a complex in-vehicle hands-free (HFT) environment. A cordless handset terminal (CHS) is connected via a VoIP network on the other side of the connection. Expert and naïve test subjects evaluated various aspects of conversational quality, including listening quality, double talk impairment, echo interference and overall quality [1]. In the same scenario, two Head and Torso Simulators (HATS) were used to play back and record pre-recorded conversations (PRC). The PRC includes single talk sequences in both directions, as well as parts where both voices interact and are talking at the same time. The binaural recordings at both ends were then assessed in a Third-Party Listening Test (TPLT) using the same quality aspects as in the CT. The great advantage of TPLT compared to CT is the much lower effort to set up and control the conditions. The auditory results of both test designs are discussed and compared for individual quality dimensions, thus, enabling the comparison of a sophisticated, highly immersive CT and a simplified, but highly efficient TPLT.

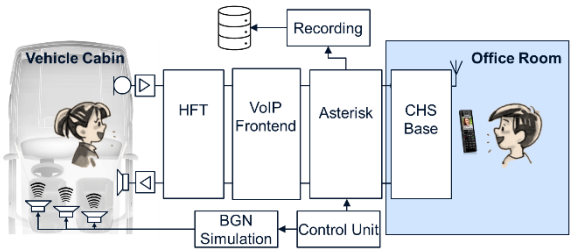

Figure 1: Conversational Test Setup, HFT to CHS
(BGN: noise simulation in the interactive driving simulator)

Conversational Test

The test design, test setup, and conduction are given in detail in [1] and [2]. Test subjects talked over a connection involving an HFT development board in a vehicle cabin of a driving simulator, a VoIP network, and a CHS (Fig. 1). The simulator features an interactive driving mode.

Signal processing in both terminals (HFT, CHS) was manipulated to create various impairments to trigger different conversational quality dimensions. A total number of 21 test conditions were used [1]. To stimulate a lively discussion, the "Kandinsky" test method [3] was adapted to a navigational task with diverging routing information on GPS maps. The driving condition in the simulator in conjunction with the discussion about the meeting point introduced a realistic cognitive load on both test subjects. 10 pairs of naïve and expert test subjects participated, resulting in 20 subjects and 10 naïve and 8 expert votes per condition for the HFT and CHS side each. The pairs of tests subjects knew each other before, facilitating natural discussions, initiated by the navigation task. The participants started with 2 initial anchor conditions to get familiar with the test setup, conduction, quality dimensions, attributes and rating scales.

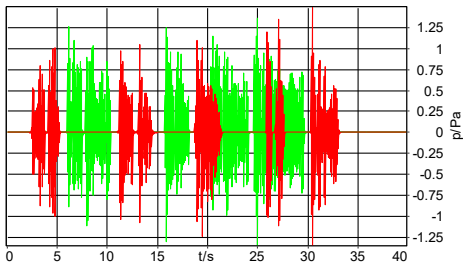

Figure 2: PRC, waveform, english, female voice HFT (green), male voice CHS (red), double talk activity (black)

In order to conduct a TPLT [4] with the same conditions, test subjects were replaced by two HATS [5], one positioned on each side in the conversational test setup, see Figure 1. A pre-recorded conversation [6] with an overall length of 40 s including about 10 seconds of initial and final silence was played back via both HATS mouths and binaurally recorded on each side, see Figure 2.

Table 1: Conversational Test Conditions (Subset)

No.	Perceived Impairment on HFT Side	Perceived Impairment on CHS Side
C1	LP1, N120	Sil120
C2	L+6, N120	NR0, Sil120
C4	D700	D700
C6	REF	NR0, HP2
C9	D1200	D1200
C10	EA18, N120	AEC27, Sil120
C13	BtHft	DT14
C14	Mob, D700, N120	DT24, D700, Sil120
C16	EA27	EA27
C17	N120	Sil120
C18	REF	REF
C19	HP1, N120	Sil120
C21	L+6	AEC27-6

- L+6: volume level increase by 6 dB.
- EA18, EA27: echo with an echo loss of 18 or 27 dB.
- D"xxx": one-way delay of "xxx" ms, baseline 200 ms.
- AEC27: echo with an average echo loss of 27 dB.
- N120: 120 km/h driving background noise (BGN).
- Sil120: residual noise of the HFT driving noise is audible.
- LP1, HP2: a lowpass (LP1) or high-pass (HP2) filter.
- NR0: noise reduction is turned off
- DT14, DT24: send double talk attenuation of 14 or 24 dB.
- Mob: mobile phone in hands-free mode
- BtHft: after-market Bluetooth speakerphone
- REF: Reference without degradation

Third-Party Listening Tests

The TPLT was first conducted with 45 naïve test subjects to evaluate the perception of the PRC recorded under the same conditions as the CT. 40 test subjects were male, 5 female, the

age ranges from 20 to 44 with a mean of 24 years. The participants were asked to imagine themselves as one of the conversation's participants while listening to the recordings ("ear witness"). The playback headsets were equalized and calibrated, thus, the signals received by the listeners matched exactly with those recorded at the HATS ears in the setup of the CT. They also started with initial anchor conditions to get familiar with the PRC test stimuli, test conduction, quality dimensions and scales. 10 PRC are used for the anchor phase. A subset of CT conditions was used, 10 conditions from the CHS side and 10 from the HFT side in the vehicle as shown in Table 1. Additionally, the TPLT was conducted with six expert listeners, to be able to compare TPLT and CT results individually for the two groups of test subjects.

Comparison of Auditory Results

The CT was realized in a sophisticated and interesting environment for the participants, provided a high level of immersion, a significant cognitive load, and allowed unlimited evaluation time – also already during the initial training and anchor phase. Conversely, the TPLT test featured a fixed training and anchor phase, low immersion, and short, condensed listening samples for evaluation. Considering these limitations for the TPLT, the accordance between the auditory results of the CT and TPLT is quite encouraging as shown by the following analysis figures.

Exemplarily, Figures 3 to 6 present the auditory MOS results for overall conversational quality as a bar plot with the 95% confidence intervals (CI95) comparing CT and TPLT for HFT and CHS side. For the TPLT the expert's ratings are plotted in blue, the naïve subject's ratings in red. The CT results are plotted in grey in each figure.

The results are further analyzed in detail for the individual quality dimensions in figures 7 to 14 as scatter plots. The MOS scores of the CT are given on the x-axis, the scores of the TPLT on the y-axis. For the sake of clarity, the representation of the confidence intervals is omitted in these plots. Table 2 shows the Pearson correlation coefficient (PCC) [7] and Table 3 the root-mean-square error (RMSE) values calculated for CT and TPLT MOS scores.

Table 2: Pearson Correlation Coefficient (PCC)

	Expert-HFT	Expert-CHS	Naive-HFT	Naive-CHS
PCC_G	0.82	0.80	0.83	0.80
PCC_{LQO}	0.95	0.66	0.82	0.76
PCC_{DT}	-0.06	0.96	-0.06	0.72
PCC_{Echo}	0.93	0.89	0.91	0.84

Table 3: Root-Mean-Square Error (RMSE)

	Expert-HFT	Expert-CHS	Naive-HFT	Naive-CHS
$RMSE_G$	0.45	0.40	0.57	0.42
$RMSE_{LQO}$	0.33	0.51	0.55	0.43
$RMSE_{DT}$	0.31	0.36	1.28	0.84
$RMSE_{Echo}$	0.41	0.73	0.21	0.50

Overall Speech Quality Results, $MOS_{E2E,G}$

Figures 3 and 4 show the overall MOS scores for the HFT and CHS sides for expert listeners. The TPLT scores align closely with CT results for both sides, except for condition C16 on the HFT side (Figure 3) and C2 for the CHS side (Figure 4). Nonetheless, the confidence intervals (CI95) for the CT and TPLT overlap for all conditions on both sides.

The C16 condition (Figure 3) comprises a 27 dB echo disturbance on the HFT side in silence (no driving noise, thus, no masking). In the CT, the expert subjects perceived this impairment as less annoying (but with high CI95) compared to the TPLT. This is likely caused by the high immersion in the driving simulator and the cognitive load of the navigational task. Both aspects are less prominent in the TPLT.

C2 on the CHS side (Figure 4) represents a test condition with driving noise in the simulator and deactivated noise reduction in the HFT. Thus, the driving noise is clearly audible on the CHS side. C2 was rated higher in the CT and lower by the TPLT subjects, most likely because TPLT subjects are less aware of the driving situation compared to the CT subjects, thus driving noise appears more annoying in the 'passive' TPLT situation (higher immersion for both participants in CT scenarios).

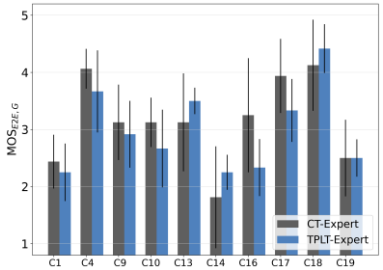

Figure 3: $MOS_{E2E,G}$ for HFT expert test subjects

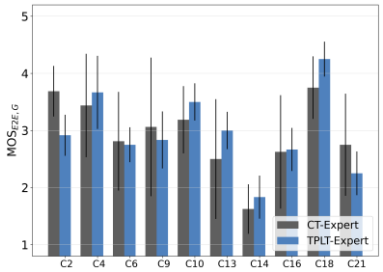

Figure 4: $MOS_{E2E,G}$ for CHS expert test subjects

Figures 5 and 6 show the corresponding scores for the naive test subjects. Again, the scores from both tests align closely and the CI95 overlap for all conditions, except C10 on the HFT side (Figure 5) and C2 on the CHS side (Figure 6). C10 on the HFT side (Figure 5) incorporates 120 km/h driving noise and a strong 18 dB echo interference. The sidetone of the PRC female voice on the HFT side is quite high due to the Lombard gain in presence of driving noise. This is combined with a strong, audible echo disturbance (18 dB echo loss). TPLT participants seem to react sensitively to these two effects, the scores are lower compared to the CT.

Furthermore, it is possible that the CT participants instinctively lower their voice level to not be annoyed by such a strong echo disturbance - additionally to the mentioned high immersion and cognitive load. The same tendency of lower TPLT ratings is observed on the HFT side for other driving noise conditions C1, C10 and C17.

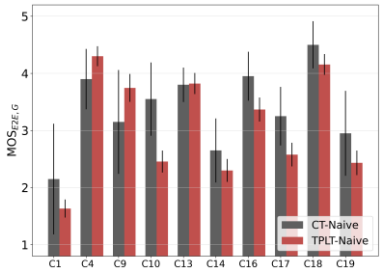

Figure 5: $MOS_{E2E,G}$ for HFT naive test subjects

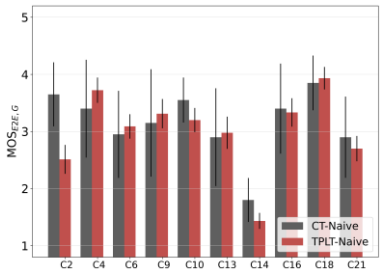

Figure 6: $MOS_{E2E,G}$ for CHS naïve test subjects

For condition C2 on the CHS side (Figure 6, lower TPLT results compared to CT), the same explanation may be valid as for the expert listeners above: The driving noise in the simulator and the deactivated HFT noise reduction is more disturbing on the CHS side in the TPLT than in the CT. The missing awareness of the driving situation on the HFT side leads to a more critical rating in the TPLT.

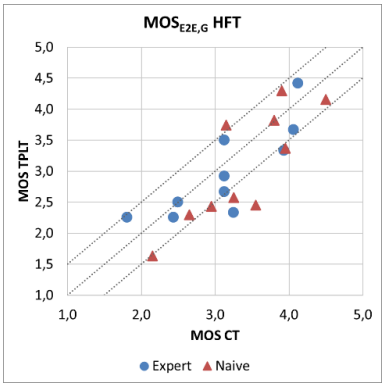

Figure 7: $MOS_{E2E,G}$ scatter plots HFT

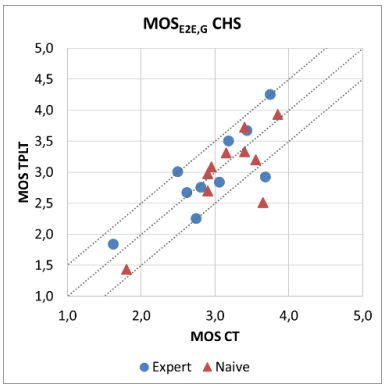

Figure 8: $MOS_{E2E,G}$ scatter plots CHS

The scatter plots for overall conversational quality in Figures 7 and 8 again show the high accordance between CT and TPLT results. This is supported by a PCC between 0.8 and 0.83 and an RMSE between 0.4 and 0.57 (Table 2 and 3). These values give confidence that a complex CT can be replaced by a TPLT, even for the assessment of complex overall quality, if both tests are carefully designed.

Listening Quality Dimension, $MOS_{E2E,LQO}$

Figure 9 shows a high correlation between the $MOS_{E2E,LQO}$ scores for CT and TPLT expert subjects on the HFT side, indicated also by a high PCC of 0.95 and a low $RMSE_{LQO}$ of 0.33. On the CHS side (Figure 10), the correlation is moderate due to two outliers (C2, C10), where TPLT expert subjects rated the conditions lower than CT expert subjects. However, the CI95 still overlap for both conditions. Both conditions include driving noise on the HFT side, experts tend to judge the residual noise slightly harsher in the less immersive TPLT.

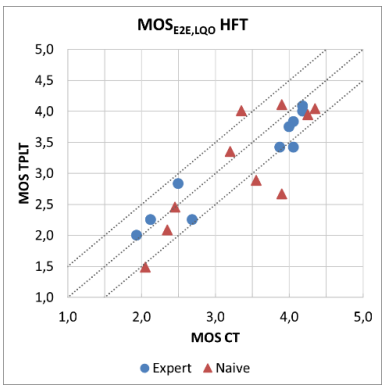

Figure 9: $MOS_{E2E,LQO}$ scatter plot HFT

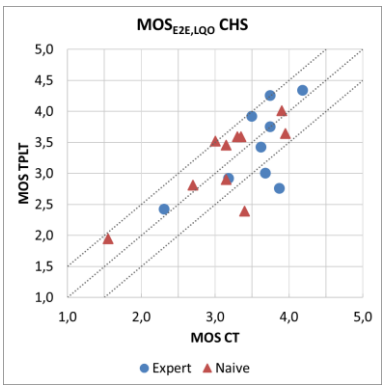

Figure 10: $MOS_{E2E,LQO}$ scatter plot CHS

For naïve listeners, the correlation remains high with a PCC of 0.82 on the HFT side (Figure 9), despite one particular outlier C17, where TPLT subjects seem to struggle to differentiate between the driving noise in the car cabin and the quality of the transmitted speech. On the CHS side (Figure 10) the correlation is moderately high with a PCC of 0.76 and a low $RMSE_{LQO}$ of 0.43 MOS. Notably, one outlier in condition C2 affects the correlation similarly to the expert listeners.

Double Talk Disturbance Dimension, $MOS_{E2E,DT}$:

The scatter plot for $MOS_{E2E,DT}$ in Figure 11 shows that expert listeners on the HFT side strongly agree, that all conditions were free of double talk impairments. The moderate outlier in Figure 11 represents condition C21. The low $RMSE_{DT}$ value of 0.31 MOS validates this good agreement. The very low PCC value of -0.06 is caused by the limited used MOS range from the 5 points scale. None of the conditions were disturbed by double talk impairments. For the CHS side (Figure 12), there is a high correlation between the CT and TPLT scores

for the experts, indicated by a high PCC of 0.96 and a low $RMSE_{DT}$ of 0.36 MOS.

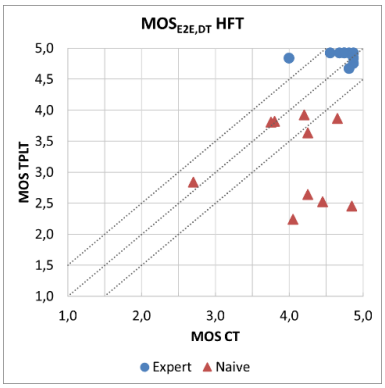

Figure 11: $MOS_{E2E,DT}$ scatter plot HFT

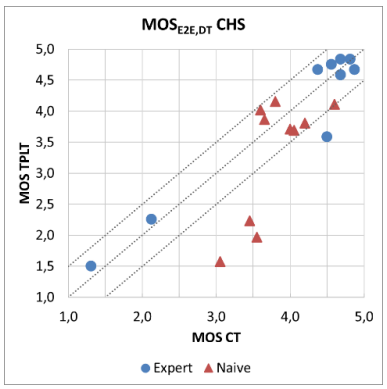

Figure 12: $MOS_{E2E,DT}$ scatter plot CHS

Surprisingly, naïve listeners on the HFT side rated the conditions inconsistently (Figure 11). The conditions were not disturbed by double talk impairments, and the CT scores are still in the upper MOS range (3.7 to 4.9). However, the TPLT scores are only between 2.2 and 3.9 $MOS_{E2E,DT}$. Obviously, naïve subjects were unable to identify and assess this impairment reliably in the TPLT. The same can be derived from Figure 12 for the CHS side. Although TPLT subjects could identify conditions with double talk impairments (C13, C14), there is a large difference in judgment between CT and TPLT. In C21, featuring a strong permanent echo disturbance, but no specific double talk impairment, TPLT subjects seem to attribute the permanent echo to a double talk degradation.

This is an important "lesson learned". Future TPLT assessments of double talk impairments obviously need to include more specific, easy-to-identify double talk impairments in the training and anchor conditions.

Echo Disturbance Dimension, $MOS_{E2E,ECHO}$

The correlation between expert listener CT and TPLT scores is very high for both sides (Figures 13 and 14), indicated by a PCC of 0.93 for the HFT side and 0.89 for the CHS side.

The $RMSE_{Echo}$ of 0.41 is reasonably low for the HFT side, but high for the CHS side (0.73), caused by one outlier representing condition C10, a complex echo condition. The echo was rated inaudible by TPLT listeners (MOS 4.9) but only 3.2 MOS in the CT. The condition was implemented by a misconfigured echo canceller and, thus, reacted very sensitive on the temporal dynamic of the received speech. The waveform of the PRC did obviously not adequately cover this.

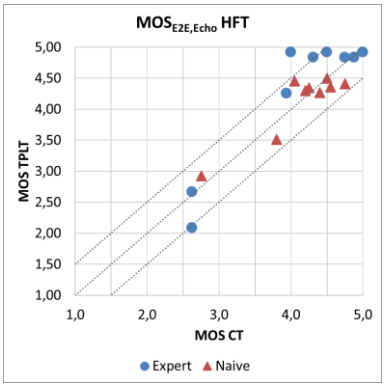

Figure 13: $MOS_{E2E,Echo}$ scatter plot HFT

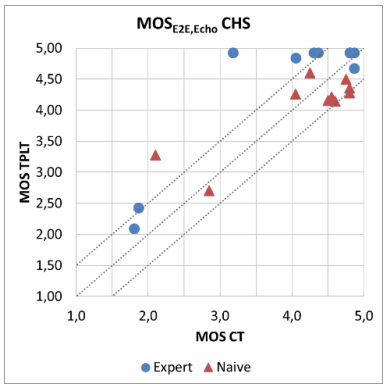

Figure 14: $MOS_{E2E,cho}$ scatter plot CHS

For naïve listeners, the correlation is also remarkably high, the $RMSE_{Echo}$ values are low for both sides (Figures 13 and 14).

Conclusion

The comparison of results from both tests give confidence that even a complex CT scenario can be replaced by a simplified TPLT, if both tests are carefully designed and conducted. For expert listeners, the results show a high overall accordance across all quality dimensions, with only some singular outliers for the listening speech quality $MOS_{E2E,LQO}$ on the CHS side. Similarly, naïve listeners showed high accordance except for the double talk quality dimension. Obviously, the test subjects were not sufficiently trained and anchored to reliably assess this quality dimension. More clearly detectable double talk impairments need to be included in the training phase.

References

[1] S. Bleiholder et al., "Overall End-to-End Conversational Speech Quality Prediction in an In-Vehicle Hands-free Scenario" in Fortschritte der Akustik – DAGA 2024, Hannover, DEGA e.V.

[2] S. Bleiholder, C. Kettler, N. Rohrer, S. Weyer, "Advances In End-to-End Conversational Speech Quality Prediction", in ITG Fachtagung, Aachen, 2023

[3] ITU-T Recommendation P.805, "Subjective evaluation of conversational quality", Apr. 2007

[4] ITU-T Recommendation P.832, "Subjective performance evaluation of hands-free terminals", May 2000

[5] ITU-T Recommendation P.58, "Head and torso simulator for telephonometry", March 2023

[6] F. Kettler, S. Poschen, R. Serafimov, "Design of Double Talk Sequences in Different Languages to Harmonize Third Party Listening Test Results", in ITG Fachtagung, Paderborn, 2016

[7] E. Cramer, U. Kamps, „Grundlagen der Wahrscheinlichkeits-rechnung und Statistik" Springer-Verlag, 2008