Automotive Hands-free Conversational Tests

Stefan Bleiholder, Nils Rohrer, Frank Kettler

HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail:Stefan.Bleiholder@head-acoustics.com

Introduction

An extensive conversational test (CT) has been conducted in a complex in-vehicle hands-free (HFT) scenario using an HFT development board, installed in an interactive driving simulator and configured with typical hands-free impairments. A cordless phone was connected via a VoIP network on the other end of the connection. Invited expert and naïve test subjects assessed different conversational quality aspects, e.g. listening quality, listening effort, double talk capability, echo, interactivity and overall quality [1]. These results represent the auditory quality ratings in each call as perceived by both test subjects in the end-to-end scenario.

During the CT, the conversations were recorded in the VoIP network and are used now for further analyses. Test subjects' voices are analyzed in terms of speech activity, active speech level (ASL), distribution of speech activity and occurrence of double talk. The analysis of the recorded conversations requires the application of noise reduction and voice activity as preprocessing. The accuracy of these assessments needs to be verified in a first step. This was realized using so called "pre-recorded conversations" (PRC), which were played back via two Head and Torso Simulator (HATS) substituting the CT test subjects.

Description of the Test Setup

The description of the test design, the setup and the test conduction can be found in [1], [2]. Test subjects were invited to discuss over a telecommunication connection consisting of a development board of an HFT implementation installed in a vehicle cabin of a driving simulator, a VoIP network and a cordless handset terminal (CHS) as depicted in Fig. 1. The driving simulator features an interactive driving mode.

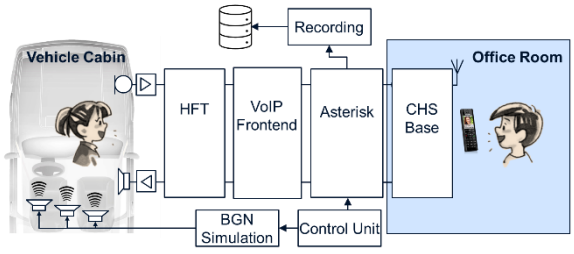

Figure 1: Conversational Test Setup, HFT to CHS
(BGN: noise simulation in the interactive driving simulator)

Signal processing relevant for the perception of communication aspects of both devices (HFT and CHS) was manipulated to create a wide range of communication impairments, such as delay, impaired speech sound quality by frequency response modifications and gain settings, echo disturbances and impairments during double talk. A total number of 21 different test conditions were used in the CT, see [1] for details.

To stimulate a realistic and lively discussion, the "Kandinsky" test method [3] was adapted to this HFT use. Both participants are presented with navigation maps in each CT run. The maps on both sides show slightly different sections, locations, and destinations. The objective for the conversation was to find a fitting meeting point. The driving condition in the simulator and the discussion about the meeting point introduced a realistic and significant cognitive load on both test subjects.

10 pairs of test subjects participated in the CT, resulting in a total number of 20 test subjects and 18 votes per condition for HFT and CHS each (two pairs only voted one side each). Of these 18 votes, 8 are gathered from pairs of expert listeners and 10 from naïve test persons. The pairs of tests subjects knew each other before. This facilitates the conversation and leads to natural discussions, initiated by the navigation task.

Both transmission directions ('HFTsend' for the HFT sending direction and 'CHSsend' for the handset sending direction) were recorded during the conversations inside the Asterisk private branch exchange software (see 'Recording' block in Figure 1), enabling detailed analysis of the conversation in terms of speech activity, the occurrence of double talk and speech levels, all in relation to the inserted impairments and test conditions, such as inserted delays, variation of gain settings and driving noise.

In Table 1, the impairments for each condition are described as measured in sending direction of the corresponding device, HFT or CHS. Additional impairments taking effect in receiving direction for the test subjects of the CT are not listed. For the impairments as they are perceived by the conversation participant, refer to [1].

Table 1: Conversational test conditions with impairments as measured on the VoIP network in sending direction of each device

No.	Impairment measured in 'HFTsend' Direction	Impairment measured in 'CHSsend' Direction
C1	N120	REF
C2	N120, NR0	L+6
C3	EA36	EA36
C4	D700	D700
C5	EA27, N60	EA27
C6	REF	NR0
C7	NR7, N120	L-6
C8	EA27, N120	EA27
C9	D1200	D1200
C10	N120, AEC27	EA18
C11	REF2	REF2
C12	BtHft, D700, DT14	D700
C13	BtHft, DT14	REF
C14	Mob, D700, N120, DT24	D700, sil120
C15	Mob, DT24	REF
C16	EA27	EA27
C17	N120	REF
C18	REF	REF
C19	N120	REF
C20	N120	REF
C21	AEC27-6	L+6

- L+6, L-6: volume level increase/decrease by 6 dB.
- EA"xx": echo with an echo loss of "xx" dB.
- D"xxx": one-way delay of "xxx" ms, baseline 200 ms.
- AEC27: echo with an average echo loss of 27 dB.
- N"xxx": "xxx" 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, NR7: noise reduction reduced to 0 dB or 7 dB
- 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

In order to generate suitable recordings for the optional conduction of a Third-Party Listening Test (TPLT, [4]) as an alternative to the CT [5], the test subjects were substituted by two HATS, one positioned in the driving simulator on the HFT side, one on the CHS side. A pre-recorded conversation (PRC acc. to [6]) was played back via both HATS artificial mouths and binaurally recorded on each side. These binaural recorded samples can be used for the TPLT as indicated in [5]. The PRC were also recorded in IP network, in the same way as the CT conversations.

The time signal of the PRC is shown in Figure 2. Contrary to the CT conversations, the characteristics of the PRC in terms of speech levels incl. potential Lombard gain, the time structure incl turn-takings, and the occurrence of double talk is known. The time signal structure of the PRC is generated individually for each connection and test condition, as the delay of the communication channel needs to be appropriately considered. This one-way delay is inserted between subsequent utterances of both HATS. This guarantees that the PRC time-structure, turn-taking and occurrence of double talk is kept identical independent of the connection delay. The synchronization plane is the acoustic interface of the hands-free terminal. These PRC are analyzed first to verify the accuracy of the analysis of the CT recordings.

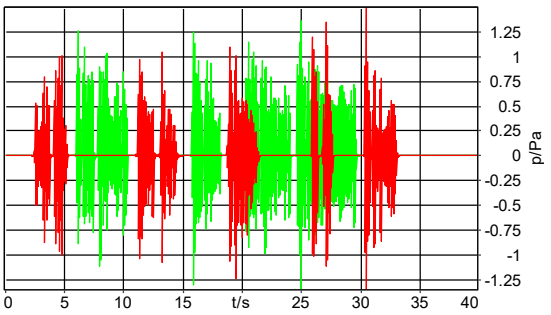

Figure 2: Pre-recorded Conversation, time signal, Engl language, HATS in driving simulator on HFT side (green), HATS in office room on CHS side (red), double talk activity (black)

Measurement of Audio Parameters

Since most of the test conditions incorporate considerable impairments including driving noise, double talk attenuation, variation of frequency responses or misaligned echo cancellers (see Table 1), the recordings need to be pre-processed before analysis (see Figure 3). Each recording is subjected to a noise reduction (NR) algorithm [7], a voice activity detection (VAD) algorithm [8] to identify the portions of speech activity. The active speech parts are then analyzed ("Analysis") in terms of terms of speech activity, Active Speech Level (ASL acc. to [9]) and double talk activity.

Figure 3: Pre-processing for analysis of recordings

Verification of the Conversational Task

The navigation task for the test subjects was designed to initiate a lively discussion with – ideally – equally distributed speech activity. Relevant parameters are the speech activity, the occurrence of double talk and ASL in both transmission directions measured in the IP network.

Speech Activity: The original sequence as shown in Figure 2 emulates a speech activity of 40.1% on the HFT side (green signal in Figure 2) and 32.5% on the CHS side (red signal in Figure 2).

Figure 4 shows the speech activity measured in sending direction of the HFT and CHS side analyzed for the PRC using HATS. The blue bars represent the speech activity on the HFT side (HFT), the red bars on the CHS side. Conditions C11 and C18, indicated on the x-axis in Figure 4, represent reference conditions without impairments.

For the HFT side, speech activity is determined to 39% in average with variations between 42 and 32%. The 39% average speech activity also matches the speech activity of the reference conditions C11 and C18 and is very comparable to the 40.1% speech activity in original PRC sequence.Conditions C14 and C15 with 33% and 32% speech activity deviate more significantly from the average value. These conditions implement an attenuation of the sending level during double talk activity (see Table 2). The reduced speech level reduces the measured speech activity due to a threshold operation in the analysis. Thus, speech activity is slightly underestimated.

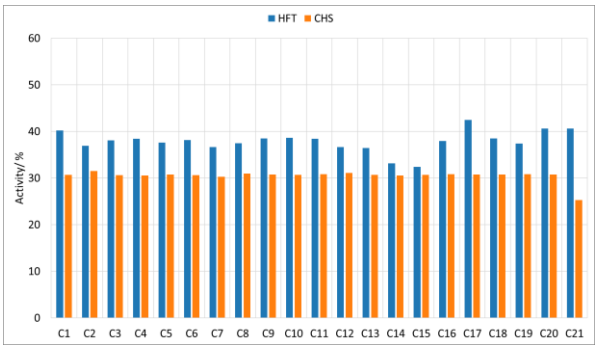

Figure 4: Speech activity in both transmission directions, PRC recordings

On the CHS side, speech activity is determined to 31% for almost all conditions. This 31% also corresponds to the 32.5% speech activity of the original PRC sequence. There is only one outlier in C21 which deviates more from the value of the

reference condition. These results give confidence in the analyses to identify the speech activity with reasonable accuracy also for the human voices CT recordings.

Figure 5 shows the comparisons of speech activity in both directions as monitored in the VoIP connection for the human test subjects averaged over all 18 measurements for 21 test conditions including the 95% confidence intervals (CI95). The blue bars represent the speech activity of human voices in the CT on the HFT side, the red bars on the CHS side.

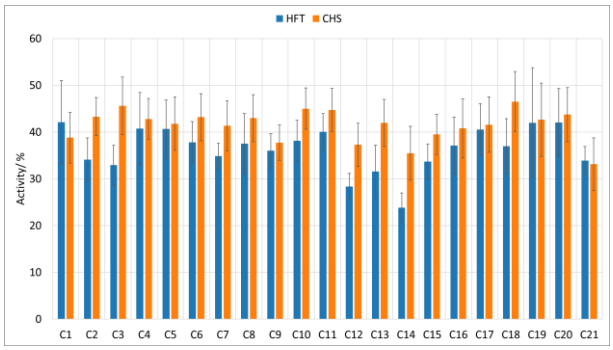

Figure 5: Speech activity in both transmission directions for CT recordings of human test subjects

Speech activity is almost equally balanced for both sides and for all test conditions. This is a very satisfying result as the task for the test subjects (navigation task) was designed to initiate a lively and – ideally – balanced discussion in terms of speech activity. The strategy to invite test participants as pairs of two, who knew each other before, seems to be successful to initiate an active discussion and facilitate the conversation.

Some conditions (C2, C3, C12, C13, C14) show speech activity results indicating that the HFT side is slightly less active than the CHS side. Interestingly, all these conditions except C3, which is a condition with moderate echo disturbance, include strong impairments with time-varying disturbances. E.g. there is the Bluetooth speaker in conjunction with a mobile phone having less-than-perfect noise reduction in the presence of background noise. Conditions C12, C13 and C14 all include an additional sending level attenuation during double talk activity, which leads to underestimation of the speech activity in sending direction measured on the network.

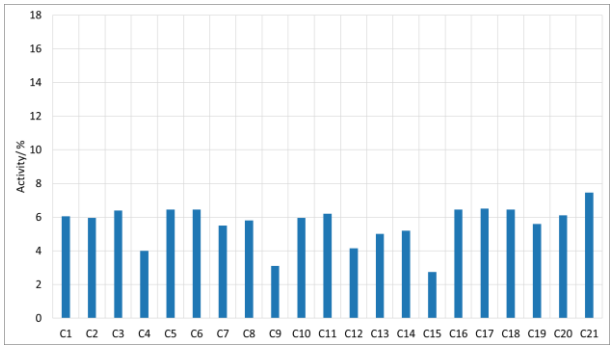

Figure 6: Occurance of double talk, PRC recordings

Double Talk Activity: Figure 6 shows the double talk activity results for all conditions for the PRC. The reference conditions C11 and C18 show a double talk activity of 6.5 %, and 2.6 s respectively. The PRC features two well-defined occurrences of double talk at around 20 s and 26 s as can be seen in Figure 2. The total double talk duration in the original sequence is 2.65 s. Thus, the double talk analysis gives proof of accuracy for undisturbed conditions.

Comparing the double talk activity for all other conditions it is evident that most conditions show the same value with only slight deviations. This indicates that the pre-processing works successfully to identify the relevant speech parts for the analysis in the network recordings for most conditions.

Conditions C4, C9 (delay impairment), and C12, C13, C14 and C15 represent combinations of double talk impairments with other impairments, also caused by additional devices in the setup (BtHFT, Mob.). As the IP recordings are done with the Asterisk private branch exchange software in the IP network, the delays from both endpoints to the Asterisk are different compared to the reference conditions. The PRC signals are time-synchronized on the HFT interface in the vehicle cabin. Thus, for conditions with additional delays, the measured double talk activity in the IP network differs, but at the acoustic interface of the HFT, the double talk activity remains identical to the reference condition.

Furthermore, conditions C12 to C15 all implement a sending level attenuation during double talk sequence. As indicated above, this leads to further underestimation of speech activity, and consequently double talk activity, for these conditions.

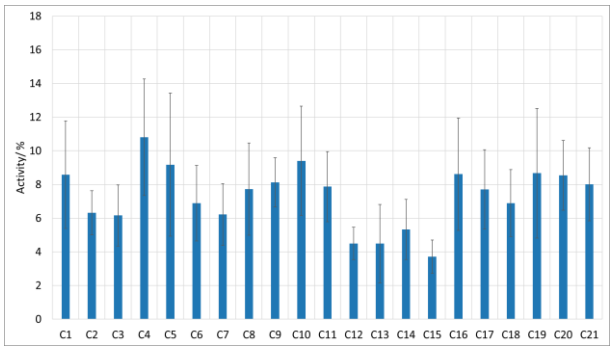

Figure 7: Occurance of double talk for human test subjects

Figure 7 shows the double talk activity measured for the human test subjects with the CI95 values for all conditions. The double talk activity values ranging from 4 % to 11 % (average 7.3 %) show that for all conditions, reasonable double talk activity can be measured. This can be regarded as an additional proof that the conversational test design as developed in [2] facilitates natural conversations and initiates double talk between the participants.

Reversely, the amount of 7.3 % double talk activity averaged over all conversations over the connection between a hands-free implementation and a handset on the other end subsequently validates the structure of the PRC as first suggested in [6].

Active speech level: The ASL in sending direction of a hands-free terminal is determined by the source speech level (i.e. -4.7 dBPa [10]), a hands-free offset of 3 dB according to [11], a Lombard gain according to [10] depending on the driving noise level in the vehicle and the sending loudness rating (SLR) of the HFT implementation itself [11], [10].

For the handset in sending direction the ASL depends on the source level (i.e. -4.7 dBPa [13]) and the SLR of the device [12].

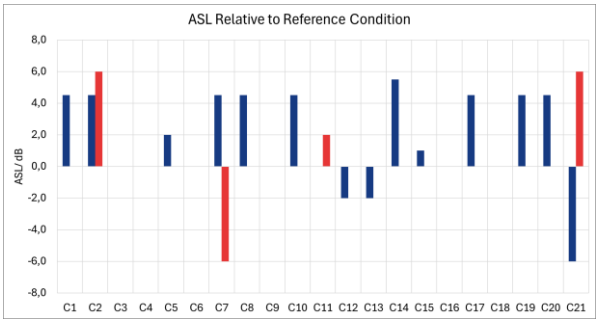

Figure 8: Calculated ASL relative to the reference condition C18 in both transmission directions

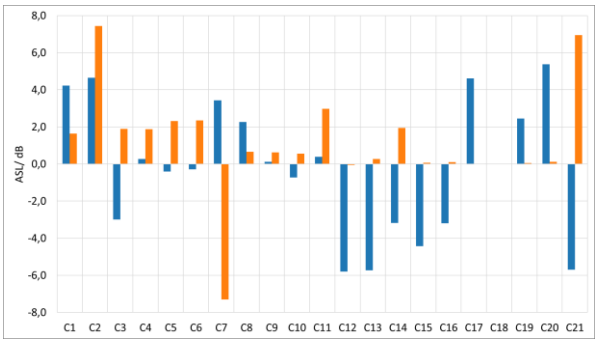

Figure 9: ASL relative to the reference condition C18, PRC recordings

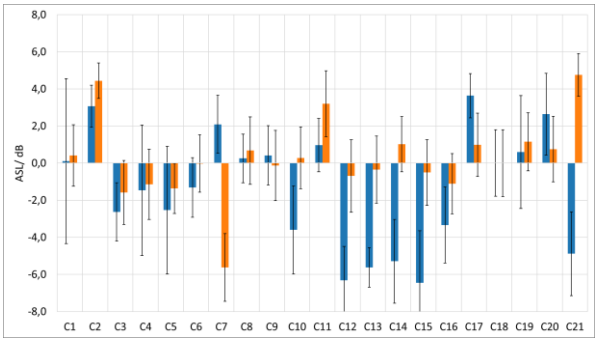

Figure 10: ASL relative to the reference condition C18, human test subjects

Figure 8 shows the calculated (not measured) ASL based on device properties and environmental conditions as indicated above. For reasons of clarity, the ASL is normalized on the ASL of the reference condition C18. Thus, the result for the C18 condition is 0 dB. The blue bars indicate an ASL change compared to C18 as calculated for the sending direction of the HFT, the red bars for the CHS. For the HFT side, where subjects are exposed to driving noise in condition C1, C2, C7, C8, C10 C17, C19 and C20, a 4.5 dB offset due to the Lombard effect can be seen. For C21 the ASL is 6 dB lower due to the lower sending sensitivity of the HFT implementation. On the CHS side, the SLR changes in condition C2 and C21 (+6 dB gain) and in condition C7 (-6 dB gain) can be seen.

Figure 9 displays measured ASL for all PRC conditions in both transmission directions in the IP network. The significant level offsets estimated in Figure 8 are also measurable in the network for the HFT in condition C1, C2, C7, C8, C17, C19, C20 and C21. The same can be analyzed for the CHS side in condition C2, C7 and C21. For some conditions (C3, C12 to C15 for HFT) the speech level calculation underestimates the ASL offsets due to double talk or echo impairments.

In general, this comparison validates ASL level measurements in the IP network. Figure 10 shows average ASL measurements for human test subjects, with confidence intervals reflecting varying speaker levels. Figures 9 and 10 share similarities for many conditions (C2, C7, C11, C17, C20, C21), demonstrating detailed investigation capabilities for human test subjects. The SLR influences can be analyzed, but also the Lombard effect. The driving noise in the driving simulator triggered the Lombard effect for the test persons as expected.

Conclusion

The analyses of the human voices CT recordings give interesting insights in the test subjects' behavior and validates the CT test design. The CT task stimulated a lively conversation between the participants, with almost equally balanced speech activity on both sides and a realistic, reasonable amount of double talk and interaction. The strategy to invite test participants as pairs of two, who knew each other before, was successful and is recommendable for a CT. The analysis also shows that known human behavior like the Lombard effect in presence of driving noise, was realistically triggered by the test design.

References

[1] 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., 2024

[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] S. Bleiholder, M. Spiertz, F. Kettler, "Third-Party Listening Test as an Alternative to Conversational Tests" in Fortschritte der Akustik – DAGA 2025, Copenhagen, DEGA e.V.

[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] Sainburg et Al., "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[8] GitHub - daanzu/py-webrtcvad-wheels: Python interface to the WebRTC Voice Activity Detector

[9] ITU-T Recommendation P.56, "Objective measurement of active speech level", 12/2011

[10] ITU-T Recommendation P.1110, "Wideband hands-free communication in motor vehicles", 01/2019

[11] ITU-T Recommendation P.340, "Transmission characteristics and speech quality parameters of hands-free terminals", 05/2000

[12] ITU-T Recommendation P.79, "Calculation of loudness ratings for telephone sets", 11/2007.

[13] ETSI ES 202 739 V1.8.1 Transmission requirements for wideband VoIP terminals from a QoS perspective as perceived by the user 03/2020