Listening Effort vs. Speech Intelligibility in Car Environments

Jan Reimes, Günter Mauer, H.-W. Gierlich
HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: telecom@head-acoustics.de

Introduction

The in-car listening situation is often impacted by low signal-to-noise ratio (SNR) which lead to reduced speech intelligibility and higher listening effort, respectively. This applies to hands-free communication but also to in-car communication (ICC) between driver and passengers in a similar way. Instrumental as well as auditory assessment of intelligibility in such scenarios is still a challenging task. The auditory test procedures are time consuming and thus expensive. Currently, also no suitable instrumental analyses are available or applicable on real measurement setups. For example the instrumental speech intelligibility index (SII) according to [1] requires degraded speech and the additive noise in two separate signals, which often cannot be provided in real environments.

On the other hand, the auditory assessment of listening effort can be derived from listening tests according to ITU-T recommendation P.800 [2], which is often more efficient to conduct. In the literature already several hints concerning the correlation between listening effort and intelligibility can be found. However, so far the comprehensions of car cabin acoustics or communication devices are not taken into account.

This contribution presents two auditory experiments. The first one evaluates the speech intelligibility in certain scenarios with "classical" methods. The second experiment uses real speech as stimuli and is carried out according to ITU-T recommendation P.800 evaluating the listening effort on a 5-point scale. The results of both auditory experiments are compared to investigate the relation between listening effort and intelligibility.

Recording Scenarios

All conducted auditory experiments were pure listening-only situations. However, test subjects were introduced to imagine that they are attending to a conversation. The next sections describe the different types of listening scenarios which were part of the evaluation.

Handset Mode

The first condition corresponds to the situation in which the co-driver conducts a mobile phone call and listens to the far-end talker. All recordings were obtained in a semi-anechoic measurement chamber with background noise (BGN) playback system according to [3]. The devices were mounted with a handset positioner to an artificial head and torso simulator (HATS). Additionally, two different mobile phones were selected. The first device under test (DUT1) is a state-of-the-art smartphone with latest signal processing capabilities. DUT2 is an older, much more simple device but which is also able to establish a wideband (WB) connection. Also narrowband (NB) mode was evaluated. The downlink signal was fed into the devices by an UMTS radio tester.

The application scenarios according to table 1 were evaluated.

No.	Device	Mode	Driving Noise
1	DUT1 (2014)	NB	no
2	DUT1 (2014)	NB	yes
3	DUT1 (2014)	WB	no
4	DUT1 (2014)	WB	yes
5	DUT2 (2010)	NB	yes
6	DUT2 (2010)	WB	yes

Table 1: Evaluated scenarios for handset

Hands-free Mode

The hands-free scenario was recorded in a driving simulator with a background noise playback system which consists of four loudspeaker and one sub-woofer. Here the setup is binaurally equalized to the driver's position. For the evaluation, a real NB-only hands-free system was mounted inside the car, the down-link signal was again inserted via an UMTS radio tester. Due to rarely available WB-capable hands-free systems, an additional loudspeaker was used in order to simulate this case (no real mobile connection was established). The speech signal and additive noise were binaurally recorded at the driver's position.

The application scenarios according to table 2 were evaluated.

No.	Mode	Driving Noise	SNR
1	NB	no	–
2	WB	no	–
3	NB	yes	+3 dB
4	WB	yes	+3 dB
5	NB	yes	−3 dB
6	WB	yes	−3 dB

Table 2: Evaluated scenarios for hands-free

Note: The SNR here is determined according to equation 1.

$$\mathrm{SNR} = \mathrm{ASL}\ [\mathrm{dB\ SPL}] - L_{\mathrm{BGN}}\ [\mathrm{dB(A)\ SPL}] \quad (1)$$

The active speech level (ASL) is calculated according to [4], the noise level L_{BGN} by applying A-weighting only on

the noise signal. In contrast to most real measurement setups, speech and noise signals were measured separately. For the auditory evaluation, both signals were mixed to obtain a noisy speech recording. Since the noise signal was assumed to be given, the SNR was adapted by modifying the speech level. In practice, this refers to either a hands-free system with higher/lower loudspeaker volume or to two different receiving loudness ratings according to [5].

In-Car Communication

In this listening condition, the speech is not originated from a down-link signal from a mobile network application, but from an artificial head mounted at the driver position. Its mouth output is transmitted to a microphone similar to the hands-free one. This signal is then processed by a customizable ICC system and then played back over two loudspeakers to the backseats.

Note that this system was set up for demonstration purposes. It was tuned to achieve a maximum attenuation (about +5 dB) without causing a feedback loop. Since this was the only optimization criterion, the quality of the transmitted speech was quite poor.

During the recording procedure, no background noise playback was used. All noise signals used in the different scenarios were real in-car binaural recordings at the corresponding position (co-driver, backseat right).

The application scenarios according to table 3 were evaluated. Note: The condition *ICC on/BGN off* includes the identical (degraded) speech signal as in *ICC on/BGN on*. In practice, an ICC system would either be not active in silent environments or would perform different than with driving noise.

No.	Position	ICC active	Driving Noise
1	Back seat right	no	no
2	Back seat right	no	yes
3	Back seat right	yes	no
4	Back seat right	yes	yes
5	Co-driver	no	no
6	Co-driver	no	yes

Table 3: Evaluated scenarios for ICC system

Recordings

Irrespective of the application scenario, all recordings were performed binaurally with diffuse field equalization. Even if in some cases no speech signal was present on the left channel (handset recordings), binaural presentation was always applied in the auditory experiments.

Within each scenario, driving noise at approx. 130 km/h was added the speech recording. However, since the recording environment was different for each scenario, also the noise playback and the noise signal differ in some details:

- Handset mode: The recording *FullSizeCar_130* provided in [3] was used for the playback.
- Hands-free mode: Recording of driving noise of luxury car was used for playback.
- ICC mode: Recording of driving noise of a compact minivan was used for the mixture of binaural signals.

Auditory Testing

In this study, typical speech intelligibility and listening effort scores were investigated. Besides that, the comparison of these two completely different auditory experiments should also be analyzed. Altogether, 18 German test subjects participated in both auditory tests judging all 18 test conditions presented in the previous section.

Speech Intelligibility

For the intelligibility test, a "classical" test design was chosen. The test subject listens to the (noisy) stimuli and writes down all understood words. Then the speech intelligibility per participant is determined as the sum of correctly understood words divided by the sum of all listened words. The overall intelligibility per condition is the average over all participants. All results presented in the next sections refer to these per-condition scores.

In this evaluation, 432 German monosyllabic words of *Oldenburg Logatome Corpus* (OLLO, [6]) by 6 different talkers were selected. One listening sequence consists of 24 randomly selected words (4 words per talker). The duration per condition is approx. 130 s (one word each 5 s, plus initial phase). 18 unique sets of words for each condition were created in this way, which means that none of the 432 appeared twice. However, due to random distribution of the words, the same word (same spoken content) from different talkers within one condition occurred in a few cases.

The overall test duration (with introduction and pause) was 60 minutes for every participant.

This test design on the one hand is a suitable method for the evaluation of the listening situations previously described, where speech and noise are given and cannot be modified (e.g. like in audiology applications). On the other hand, also several difficulties occur. Since participants should not listen to identical words (of same talkers), the sequences used for each condition must be unique within the auditory test. This either limits the number of words per condition or the number of conditions which can be evaluated (with a given number of words per condition).

Additionally, using this approach on testing intelligibility causes an enormous effort in order to obtain the final scores. Each written word must be reviewed manually by proofreading and marking. Even for this small evaluation reading the handwriting of the participants and making a decision if the word was understood correctly (regarding the pronunciation) leads to more than 7500 single items for reviewing.

Several other approaches for obtaining auditory intelligibility scores are known in literature (e.g. rhyme test,

sentence tests, etc.) which can also decrease effort and costs. Pros and cons of these methods are not discussed in this contribution.

Listening Effort

The second auditory experiment was designed as an absolute category rating (ACR) test. A common five-point scale acc. to ITU-T Rec. P.800 [2] is available for listening effort and was used for this test:

(5) Complete relaxation possible; no effort required

(4) Attention necessary; no appreciable effort required

(3) Moderate effort required

(2) Considerable effort required

(1) No meaning understood with any feasible effort

The main difference to the intelligibility assessment is that instead of a "measurement" the "impression" (or self-assessment) of the participant is asked for.

In this evaluation, the German ITU-T P.501 [7] speech material was used (4 talkers with 2 sentences each). Every stimulus of 8 s consists of two sentences of the same talker. Thus 4 samples per condition were obtained. All votes of all samples belonging to each condition are averaged to mean opinion score (MOS). All results presented in the next sections refer to these per-condition scores.

The overall test duration (with introduction and pause) was 35 minutes for every participant.

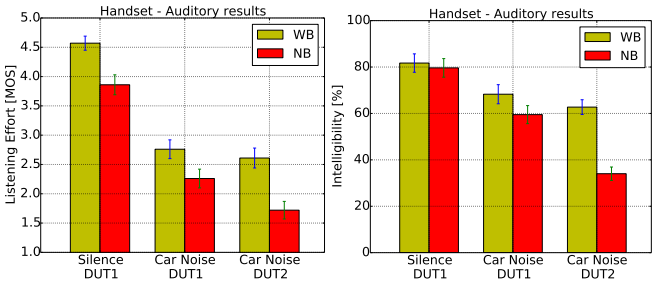
In contrary to the speech intelligibility evaluation, this method has some advantages in practice. First, it is obvious that the speech material may be originated from a much more limited corpus, because (within tolerable limits) repetitions of the stimulus do not corrupt or influence the test results. Another advantage is that this test design can be combined with additional scales (e.g. speech quality).

Auditory Results

The bar plots in the following illustrate the results of the auditory tests. Figure 1 shows the results of both auditory tests for the handset scenario. When comparing the auditory results of listening effort (LE) and speech intelligibility (SI), there is a consistent rank order between both quantities.

One remarkable result can be observed for the SI test in condition silence/WB/DUT1. The score here is rather low (about 80%), whereas almost the maximum score for LE is achieved. Since this condition refers almost to the best-possible listening situation in the whole test, the listening effort rating seems to be more plausible than the intelligibility score.

Figure 2 shows the results of both auditory tests for the hands-free scenario. The rank order between LE and SI seems to be consistent in most cases, some observations regarding consistency can be made.



(a) Listening Effort (b) Intelligibility

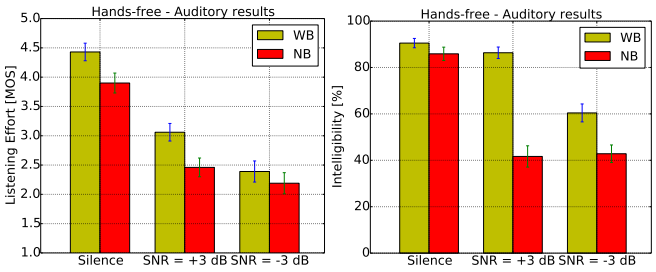
Figure 1: Auditory test results in handset mode

- For the SI results in condition SNR = 3 dB, the gap between NB and WB seems not reasonable. The score for WB appears to be too high compared to the corresponding condition silence/WB.
- Additionally, the SI score for NB does not decrease from SNR = 3 dB to SNR = −3 dB, which would be an expected behavior: Less speech level causes less intelligibility.

On the other hand, the corresponding listening effort ratings for these two cases seem to be more plausible.

- The gap between WB/silence and WB/SNR = 3 dB conditions amounts to about 1.5 MOS.
- A decrease of 0.3 MOS is observable for the transition SNR = 3 dB to SNR = −3 dB.

Possible explanations for these inconsistencies are given in the conclusions.

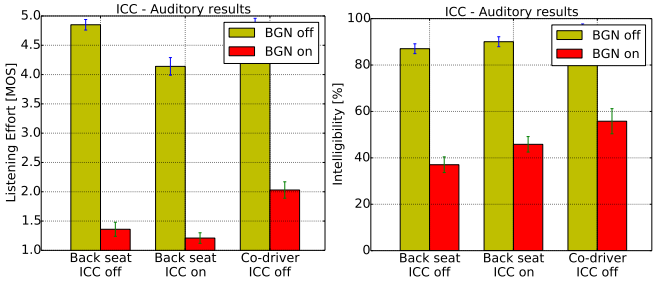


(a) Listening Effort (b) Intelligibility

Figure 2: Auditory test results in hands-free mode

Figure 3 shows the results of both auditory tests for the ICC scenario. In contrast to the previous scenarios, the rank order of the LE does not match the rank order of the SI.

- When driving noise is present, the results show a decrease of LE from *ICC off* to *ICC on* at back seat position . Even though both scores are at the lower end of the scale ($\leq$ 1.5 MOS), an improvement would be expected for an active ICC system. On the other hand, scores for SI increase with active ICC as expected.
- In silence conditions, again a decrease of LE from *ICC off* to *ICC on* is observable. In contrast, an increase between these two conditions can be noted for



(a) Listening Effort (b) Intelligibility

Figure 3: Auditory test results of ICC system

SI, which is an expected behavior: A higher speech level caused by the active ICC system increases intelligibility.

Possible explanations for these inconsistencies are given in the conclusion.

Figure 4 shows the comparison for both auditory experiments. Here all conditions are presented within one single scatter plot. After applying a 3rd order mapping between both scales, the Pearson correlation coefficient reaches r=92.0 %. However, rank order shifts cannot be compensated with this transformation. Possible reasons for all these shifts are discussed in the conclusion.

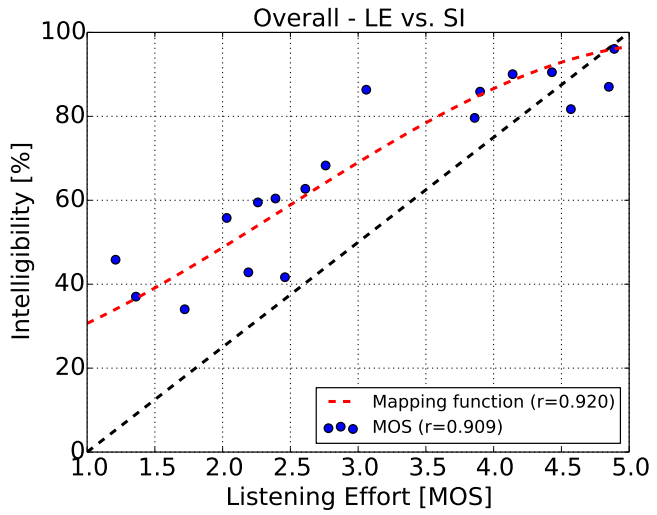


Figure 4: Comparison of auditory tests

Conclusion

Two different types of auditory experiments were conducted and typical intelligibility scores for several speech communication situations in car cabins were obtained.

In general a correlation between speech intelligibility and listening effort can be found. The overall correlation in terms of merging all listening conditions, shows an accurate correlation coefficient after 3rd order mapping, but also shows several constraints regarding the rank order of certain conditions.

Several explanations for these rank order shifts can be hypothesized:

(1) Due to the random allocation of words to the conditions, there may be an unbalanced distribution of words in every condition regarding the "a-priori intelligibility" of each word. Some words may be recognized better even in noisy environments than other words, which are harder to understand even in silent environment. Other words may also be less comprehensible when applying a band-limitation (NB/WB) than other words which may be understood almost independent of their bandwidth. Especially in mixed bandwidth scenarios like presented in this contribution, this effect can complicate the composition of unique word sequences.

(2) The rating of the listening effort may also be influenced by other features. For example, bad speech quality may also lead to a higher listening effort because it is more annoying to listen to a extremely degraded/distorted voice. At least for the considered ICC system here, this is definitely the case.

On the other hand, assessment of listening effort instead of intelligibility has several advantages. As a whole, this test shows more consistent results, even if there may be constraints regarding the impact of bad speech quality. Since the speech material may be repeated within the test, results can be reproduced by the same listener group without a noticeable training effect. Due to the reduced testing time, either more test conditions per listener group can be obtained or the same amount of conditions can be evaluated in less time. Finally, category tests in general can be fully automated regarding the assessment of aggregated auditory scores and thus cause less costs.

References

[1] ANSI S3.5-1997. *Methods for Calculation of the Speech Intelligibility Index*, 1997.

[2] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality*, Aug. 1996.

[3] European Telecommunications Standards Institute. *Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database*, August 2014.

[4] ITU-T Recommendation P.56. *Objective measurement of active speech level*, Dec. 2011.

[5] ITU-T Recommendation P.79. *Calculation of loudness ratings for telephone sets*, Nov. 2007.

[6] Thorsten Wesker, Bernd T. Meyer, Kirsten Wagener, Jörn Anemüller, Alfred Mertins, and Birger Kollmeier. Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. In *INTERSPEECH*, pages 1273–1276. ISCA, 2005.

[7] ITU-T Recommendation P.501. *Test signals for use in telephonometry*, Jan. 2012.