Perceived Listening Effort for In-car Communication Systems

Jan Reimes

HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: telecom@head-acoustics.de

Abstract

Communication inside a car cabin can be quite difficult depending on the driving situation, e.g. due to a low signal-to-noise ratio. Up to a certain degree, in-car communication (ICC) systems serve as a remedy for this situation. ICC systems aim to lower the necessary listening effort by recording the talker's speech signal and reproducing it over loudspeakers close to the listener's ears.

Concerning ICC systems in particular, a clear trade-off has to be made between listening effort and speech quality. Accordingly, a tailor-made auditory test design that assesses both attributes simultaneously has been developed recently. The test procedure and the corresponding test conditions are described in the upcoming specification ETSI TS 103 558.

Results of recently conducted auditory tests based on this design are presented and analyzed in this work. The technical specification ETSI TS 103 558 will also contain an instrumental evaluation approach for listening effort, which is based on the same binaural recordings being used as stimuli in the auditory evaluation. This contribution presents an algorithmic overview of the model's first version, as well as initial prediction results.

Introduction

The in-car listening situation is often impacted by a low signal-to-noise ratio (SNR), which leads to reduced speech intelligibility and higher listening effort, respectively. This applies in particular to the communication between driver and passengers. Several ICC systems have been recently introduced in the market, aiming at improving this situation as well as at decreasing driver distraction.

The current ITU-T work item *P.ICC* [1] deals with the performance evaluation of ICC. Beside numerous technical parameters like e.g., delay of reinforcement path, SNR or dynamic behavior of such systems, the improvement as perceived by the user still is a challenging task. Common auditory test procedures for speech in noisy environments are intelligibility tests. But as already discussed in e.g., [2], there are several drawbacks of existing auditory as well as instrumental test procedures.

An alternative approach for auditory evaluations is the perceived *Listening Effort* (LE). Here test subjects provide a self-assessment on a five-point categorical scale, similar to well-known speech quality testing methods. Several recent studies (e.g., [3, 4]) indicate that a wider range of SNRs can be evaluated regarding speech enhancement benefits, without reaching positive or negative saturation of intelligibility tests.

Score	Listening Effort	Speech Quality
5	Complete relaxation possible; No effort required	Excellent
4	No appreciable effort required	Good
3	Attention necessary; Moderate effort required	Fair
2	Considerable effort required	Poor
1	No meaning understood with any feasible effort	Bad

Table 1: Auditory scales for combined assessment

In contrast to speech intelligibility evaluations, 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.

The usage of an optional *Speech Quality* (SQ) attribute supports test subjects in differentiating between the noise component (major impact on listening effort) of the signal and possible speech degradation (minor to medium impact on listening effort), which are included in a stimulus. Scales of both attributes were taken from ITU-T P.800 [5], as provided in Table 1.

Recent Work on Auditory Evaluation

The investigations presented in the following sections were already shown in [2]. Thus, only a brief overview is given, with focus on auditory results.

Simulation Environment

Impulse response measurements in the cabin of two different vehicles were conducted (one mid- and one full-size car), which are considered as two devices under test (DUTs).

The talker (driver position) and listener (behind driver) are realized by two head and torso simulators (HATS), equipped with artificial mouths and ears.

In order to accurately simulate the whole ICC system offline, several impulse responses were measured with white noise signals:

- Driver's mouth to all ICC microphones (input path)
- Driver's mouth to listener's ears (direct path)
- Loudspeakers to ICC microphones (feedback path)
- Loudspeakers to listener's ears (reinforcement path)

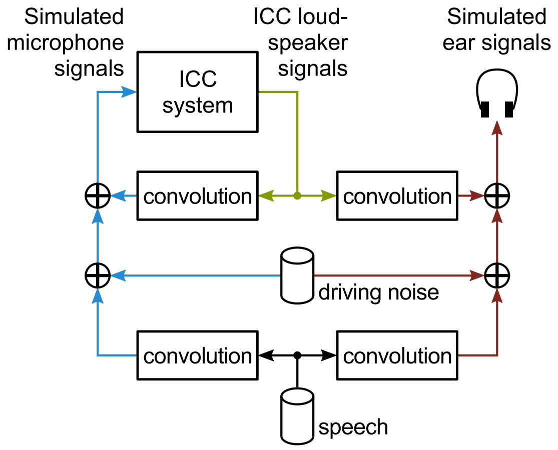


Figure 1: Structure of the simulation environment (taken from [2])

Two driving noise conditions (medium and maximum speed) were recorded synchronously at the ICC microphones and the listener's ears in both DUTs. The structure of the simulation environment that was used to obtain simulated binaural ear signals is shown in Figure 1.

In both DUTs, an ICC implementation of sonoware GmbH was used for the signal processing and the generation of binaural signals. The following operational modes simulated in order to obtain a wide range of impairments and reinforcement sounds:

- *ICC Off*: system is deactivated.
- *Default*: is tuned for optimum/balanced execution.
- *High Gain*: similar to *Default* mode, but with additional output gain and artifacts.
- *Extra Delay 15*: same as *Default*, but processing delay artificially increased by 15 ms.
- *Extra Delay 25*: same as *Default*, but processing delay artificially increased by 25 ms.

The German speech material according to TS 103 281 [6] was used for the simulation, which includes two sentences of four male and four female talkers each.

Auditory Testing and Results

With two DUTs, five ICC modes, three noise scenarios (including silence), 30 conditions were obtained by the offline processing and were available for the listening test. In addition, 12 reference (or sometimes referred as *anchor*) conditions according to [6] were included in the experiment.

The listening test procedure is a combination of the well-known methodologies according to ITU-T P.800 [5] and ITU-T P.835 [7]. Test subjects evaluate each presented sample twice. After the first presentation, a rating for Listening Effort (LE) was given. In a second trial, SQ was assessed.

A total of 48 naïve German test subjects participated in the auditory test, which contained 672 samples (42 conditions, 16 sentences each). The stimuli were presented via diffuse-field equalized headphone playback, compatible to the equipment used for recording.

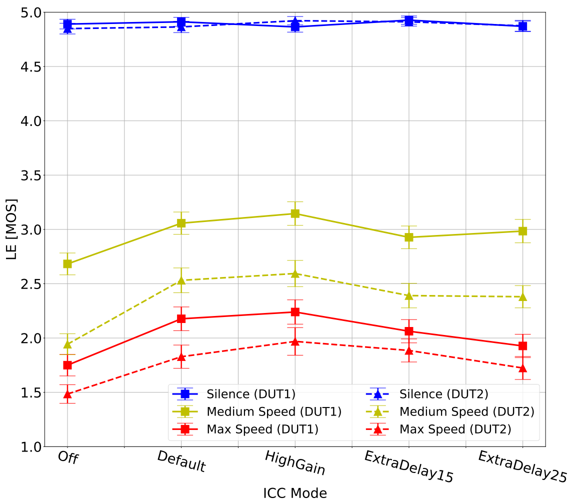


Figure 2: Auditory results for $\mathrm{MOS}_{\mathrm{LE}}$ (taken from [2])

The key findings were already presented in [2] and can be summarized as follows (Results for SQ are out of scope for the current work), see also Figure 2:

- DUT 1 (full-size) performs better than DUT 2 (mid-size) due to superior acoustic properties.
- Silence: $\mathrm{MOS}_{\mathrm{LE}}$ close to maximum of ≈ 5.0. No further impairment of LE due to artifacts.
- Medium Driving: $\mathrm{MOS}_{\mathrm{LE}}$ significantly decreases. As expected, *High Gain* obtains the best results for both DUTs, but also introduces most artifacts. Even the most degraded ICC settings performs much better than *ICC Off.*
- Maximum Driving: $\mathrm{MOS}_{\mathrm{LE}}$ decreases even more, improvement of ICC becomes smaller. Again, *High Gain* obtains the best results and the most degraded ICC settings perform still better than *ICC Off.*

Prediction Model

The model for predicted listening effort introduced in the following sections is currently under development. The work shown here represents an intermediate status and results are preliminary. Thus, only a brief high-level description of the algorithmic parts is provided here.

Figure 3 provides an overview of the prediction algorithm, which consists of different stages. Binaural recordings are used as an input and can be considered as a tuple of two single-channel signals (example given by equation 1). The clean speech reference $r\,(k)$ is a single-channel signal.

$$\mathbf{d}(k) = \langle\, d_{\mathrm{L}}(k),\, d_{\mathrm{R}}(k)\,\rangle \qquad (1)$$

The pre-processing stage aligns all input signals with regard to differences in temporal shifts and levels:

- The delay between degraded signals $\mathbf{d}(k)$ and the reference $r(k)$ is calculated by a cross-correlation analysis for both ears. The delay is selected from the ear providing the higher magnitude of the cross-correlation.
- The reference signal $r(k)$ is scaled to a fixed active speech level (ASL) of 79 $\mathrm{dB}_{\mathrm{SPL}}$ ($r_{\mathrm{Opt}}(k)$), which is assumed to be optimal regarding listening effort [8].

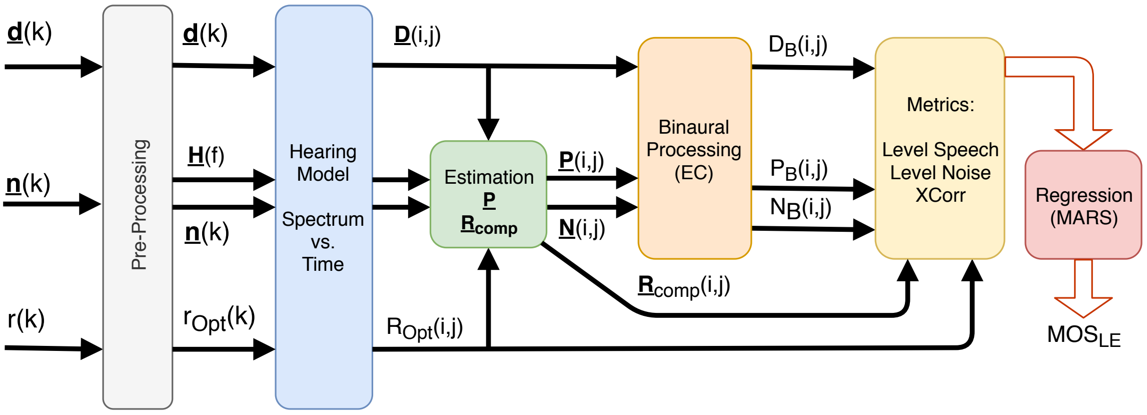


Figure 3: Flow chart of prediction algorithm

- Time ranges of active speech in $r_{\mathrm{Opt}}(k)$ are detected based on the activity classification according to ITU-T G.160 [9].
- The transfer functions $\mathbf{H}(f)$ between degraded and reference signal are calculated by the H1 methodology (cross-power spectral density) in order to exclude non-correlated noise components.
- The processed (but mostly noise-free) signal is determined by $\mathbf{p}(k) = \mathbf{d}(k) - \mathbf{n}(k)$.

Similar to related speech quality metrics, an aurally-motivated transformation to the time-frequency domain of the signal is applied here (time index i, frequency bin j). The hearing model according to Sottek [10] is used for this purpose. The resulting signal representations are denoted with capital letters $\mathbf{D}$(egraded), $\mathbf{N}$(oise), $\mathbf{P}$(rocessed) and $\mathbf{R}$(eference), which can again be considered as a tuple of spectra for the left and right ear (example given by equation 2).

$$\mathbf{D}(i,j) = \langle\, D_{\mathrm{L}}(i,j),\; D_{\mathrm{R}}(i,j)\,\rangle \tag{2}$$

In order to address the capability of human hearing to improve SNR compared to monaural listening, a binaural processing stage is included in the prediction model. The spectral components for left and right ears are combined by a short-term equalization-cancellation (stec) according to [11]. This extension of the well-known model of Durlach [12] requires the availability of the isolated speech and *masker* (noise-only) components. As a result of this stage, combined and enhanced hearing model spectra vs time are provided, like e.g., $D_{\mathrm{B}}(i,j)$ for the degraded signal.

Based on the binaural spectra, several comparisons and single value metrics can be considered. A common method is to compare the degraded and/or processed signal to the reference, but also single-ended metrics like e.g., level of (active) speech and noise-only are possible. For the current investigation, six level- and correlation-based metrics (often also referred as "features") are determined, which are combined to an instrumental overall listening effort Mean opinion score (MOS). However, since the development of the model is still ongoing, detailed descriptions of metric calculations are beyond the scope of this work.

Instrumental Results

Prediction models as introduced in the previous section usually require a certain amount of training data in order to reliably work on unknown data. Since comparable listening test databases are rarely available (especially for the application in ICC), the previously presented auditory study is used for several cross-validation experiments. As previously described, the listening test database provides a reasonable amount of material for such an analysis.

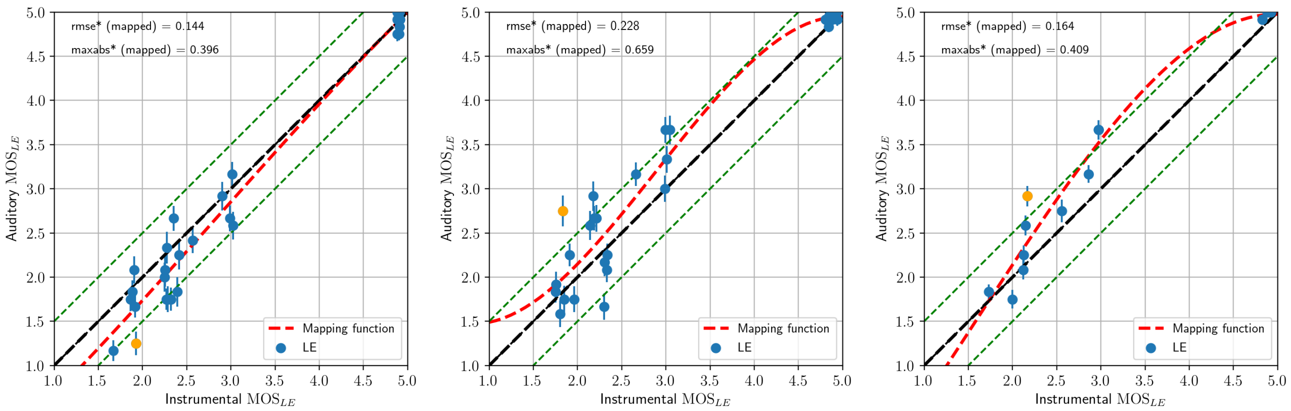
With 16 single speech sentences available per condition (four male & female talkers, two sentences each), the whole corpus can be divided into several orthogonal groups. The following sections provide three examples for training and test of the prediction model. For each example, the material is deliberately divided in half, i.e. 240 samples are used for training and test.

Figure 4 illustrates the results in form of a scatter plot. Each point represents the averaged auditory and instrumental results per condition. Vertical markers indicate the 95% confidence interval (CI95), aggregated according to [13]. As performance metrics, *rmse** and *maxabs** [13] are provided, which take the uncertainty of the auditory data (CI95) into account.

Split per sentence: For the first example, all talkers are used to train the model, but only sentences *s1* are employed. For testing, the same talkers are evaluated - but with sentence *s2* (unknown during training). A slight bias / over-prediction can be observed, but in general, performance as shown in Figure 4a is quite accurate.

Split per talker: For the second example, all *male* talkers with both sentences *s1/s2* are used to train the model. For testing, all *female* talkers are employed, which were unknown during the training. As shown in Figure 4b, a decrease in prediction performance can be observed. This is not unexpected as the impact of unbalanced training material is also known from e.g., speech/audio quality assessment methods or in the domain of speech recognition.

Split per condition: For the third example, the 30 conditions are randomly split. In contrast to the previous two examples, each point in the scatter plot now repre-



(a) Only sentences *s2*

(b) Only *female* talkers

(c) Random conditions

Figure 4: Prediction results (test) for specific partitions

sents 16 samples (instead of eight before), which is also indicated by the smaller CI95 markers. Figure 4c shows a highly accurate prediction performance.

Conclusions

A comprehensive study on perceived listening effort and speech quality for ICC systems was conducted in previous work [2], which is based on binaural simulations and noise recordings in a realistic manner. The intentionally introduced degraded modes were rated according to the expectation, namely much lower than the balanced/optimum setting.

For the instrumental assessment of perceived listening effort, a prediction model is currently being developed. The current status of the preliminary algorithm was briefly introduced. Due to the considerable amount of test samples and conditions of the auditory data, several cross-validation experiments were successfully conducted. For a set of balanced training material, the model is able to robustly predict $\mathrm{MOS_{LE}}$ for samples unknown during the training procedure.

For further development, auditory databases related (but not limited) to ICC must be available for an elaborated training procedure. Especially the intermediate range of LE (3.0-4.0 $\mathrm{MOS_{LE}}$) should be investigated in more detail.

References

[1] ITU-T Recommendation P.ICC. *TD-GEN-0654r2 Draft of In-Car Communication Audio Specification*, November 2018.

[2] Jan Reimes and Christian Lüke. Perceived listening effort for in-car communication systems. In *ITG-Fachtagung Sprachkommunikation*. VDE Verlag, Oldenburg, Germany, September 2018.

[3] Arne Pusch, Jan Rennies, Henning F. Schepker, and Simon Doclo. Höranstrengung als Messverfahren zur Evaluation von Near-end listening enhancement Algorithmen. In *Fortschritte der Akustik - DAGA 2018*, volume 44, pages 543–546, Berlin, Germany, 2018.

[4] Jan Rennies and Gerald Kidd. Binaural listening effort in noise and reverberation. In *Fortschritte der Akustik - DAGA 2018*, volume 44, pages 615–616, Berlin, Germany, 2018.

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

[6] ETSI TS 103 558 V0.0.4. *Methods for objective assessment of listening effort (draft; expected to be finalized 03/2020)*, Feb 2019.

[7] ITU-T Recommendation P.835. *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, Nov. 2003.

[8] ITU-T. *Practical Procedures for Subjective Testing*. ITU, 2011.

[9] ITU-T Recommendation G.160 Amendment 2. *Voice enhancement devices - Appendix II*, March 2011.

[10] Roland Sottek. A hearing model approach to time-varying loudness. *Acta Acustica united with Acustica*, 102(4):725–744, Jul / Aug 2016.

[11] Rui Wan, Nathaniel I. Durlach, and H. Steven Colburn. Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. *The Journal of the Acoustical Society of America*, 136(2):768–776, 2014.

[12] Nathaniel I. Durlach. Binaural signal detection: Equalization and cancellation theory. *Foundations of Modern Auditory Theory*, Vol. 2, 02 1972.

[13] ITU-T Recommendation P.1401. *Statistical analysis, evaluation and reporting guidelines of quality measurements*, Jul. 2012.