Perceived Listening Effort for ICC systems – a multi-language comparison

Jan Reimes[1], Jan Holub[2]

[1] HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail: telecom@head-acoustics.de
[2] Faculty of Electrical Engineering, CTU Prague, 16627 Prague, E-Mail: holubjan@fel.cvut.cz

Abstract

Communication inside a car cabin can be quite difficult depending on the driving situation, mainly due to high driving noise. Up to a certain degree, in-car communication (ICC) systems serve as a remedy for this situation. However, as for many speech-processing systems, often a trade-off between (improved) listening effort and (possibly decreased) speech quality has to be made.

In previous work, it was shown that the combined auditory assessment of speech quality and listening effort might be a suitable framework for the evaluation of ICC systems. Most of these experiments were conducted with a comprehensive test corpus, but so far only in German language. For auditory evaluations with speech samples in general, it is important that the language, which is used in the stimuli does match the one of the subjects. In particular, for the assessment of listening effort, this requirement becomes even more important.

This contribution presents new auditory results of an ICC system in three different languages. Identically processed speech samples in American English, Mandarin and German were obtained and tested in two different listening laboratories. Results of auditory tests and multi-language comparisons are presented and analysed.

1. Introduction

Listening effort (LE) is a widely used concept describing an impact of acoustic challenges in voice communication [1]. Listening effort is also recognized as one of the most significant aspects of telecommunication systems and services which affect user satisfaction [2]. Beside the well-known rating scale for Listening Quality (LQ, Annex B.4.5a), Recommendation ITU-T P.800 [3] also lists a rating scale for Listening Effort (LE, Annex B.4.5b).

When assessing LE subjectively, the question repeatability always arises and is therefore of great research interest [4]. To assure result relevancy, tests are sometimes held in different laboratories in parallel. Afterwards, the results are checked for their level of agreement. In those repeated experiments, usually one or more parameters differ among laboratories.

Typically, the deployed test subject nationality are varied - and in consequence, the language used [5]. The present study aims to find the level of inter-lab and test results repeatability, using test speech stimuli, which were processed identically in three different languages (American English, Mandarin and German).

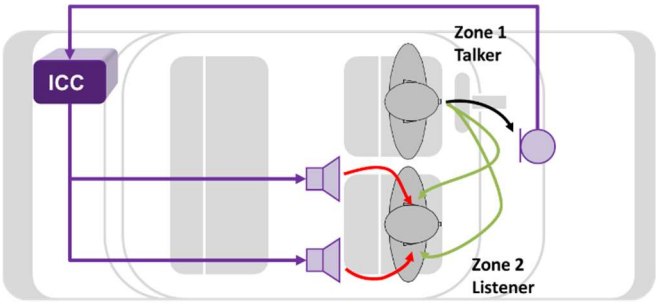

Figure 1: Test setup for ICC recordings

2. Experiment description

2.1 ICC speech samples design and acquisition

Figure 1 illustrates the test setup in a vehicle used for the acoustics recordings of in-car communication scenarios. The vehicle is a convertible (compact/sports car), with only two seats in the second row. The talker HATS (Head and Torso Simulator) is positioned at the driver's seat (zone 1), the listener is located at the co-driver's seat (zone 2).

For the generation of listening test samples, a similar approach as described in [7] was chosen. All test conditions were simulated by means of impulse responses, pre-recorded noise conditions and offline processing of the ICC system. Beside the condition *ICC off*, the offline processing includes a balanced base-line setting (gain = 0 dB), which is then extended by various gains and artificially increased delay. For gains 0, 5 and 8 dB, only a slight feedback compensation is conducted, while for gains 10 and 15 dB, a stronger compensation is applied.

Driving noise of three different conditions were recorded binaurally at the listener position (plus standstill, with running engine) as well as at the input microphones of the ICC system. In overall, 46 test conditions were generated. Table 1 provides an overview of the recordings generated for the listening test. Note that the delay of the ICC simulation was compensated in the reinforcement path, i.e. the additionally inserted delay represents the overall processing time.

Fullband speech samples from [6] have been used for processing. They have been recorded by four talkers (two males, two females) and are available in American English (ENG), Mandarin (MAN) and German (GER) language. The sentences were simple meaningful sentences as described in Annex B.1.4 of [3]. Each source speech file contains up to 16 sentences.

Table 1: Driving conditions and ICC parameters

Speed [km/h]	Nbr. of Processings	Max. Gain [dB]	Min. Delay [ms]	Max. Delay [ms]
0	3 + ICC-off	10	15	45
50	13 + ICC-off	15	15	65
100	13 + ICC-off	15	15	65
120	13 + ICC-off	15	15	65

Each sentence is centred inside a time window of 4.0s. With active speech durations between 2 and 3 seconds, leading and trailing silence parts resulted in approximately 0.5-1.0 seconds of noise in the degraded sample.

In the following, the term *database* refers to a set of 46 test conditions, which are generated by the aforementioned offline processing in one of the three languages.

In addition to each database, 12 reference conditions according to Annex B of [8] were designed using the noise type *Full-size car 130 km/h*, as per clause 8 of [9].

2.2 Subjective test design

Subjective tests were designed according to [3], and for each sample, both LE and LQ following ACR methodology have been assessed. The order of LE and LQ questions has been balanced, i.e. the subjects were asked for their assessment of LE and consequently for LQ in half of the listening sessions, and for LQ and then LE in the other half. Within each session, identical question order was kept for all samples.

The sessions have been designed not to exceed 1.5 hour in duration. Test duration comprised 50% of actual listening time and 50% test overhead including administration, initial briefing, preliminaries, and breaks.

In both labs, speech files have been played back using diffuse-field equalized headphones (Sennheiser HD600), calibrated before and verified after the experiment. Listening environment conformed to requirements stated in [3], providing reverberation time less than 185 ms and noise not exceeding 30 dB_{SPL} (A) with no peaks in frequency spectra.

2.3 Subjective data acquisition

At least 24 subjects have been used for each test database, each panel with an independent randomization following "partially-balanced/randomized blocks" experimental design described in [10]. The subjects' gender and age structure is provided in Table 2.

Table 2: Gender and age distribution

Language	Ratio #f / #m	Avg. Age [Years]	Std. Dev. Age [Years]
ENG	1.00	33.4	9.9
MAN	1.00	30.5	10.2
GER	0.92	29.2	8.6

All subjects have been native speakers of the tested language with normal hearing assessed by subjects' introductory self-assessment. No additional hearing tests have been performed prior the subjective testing.

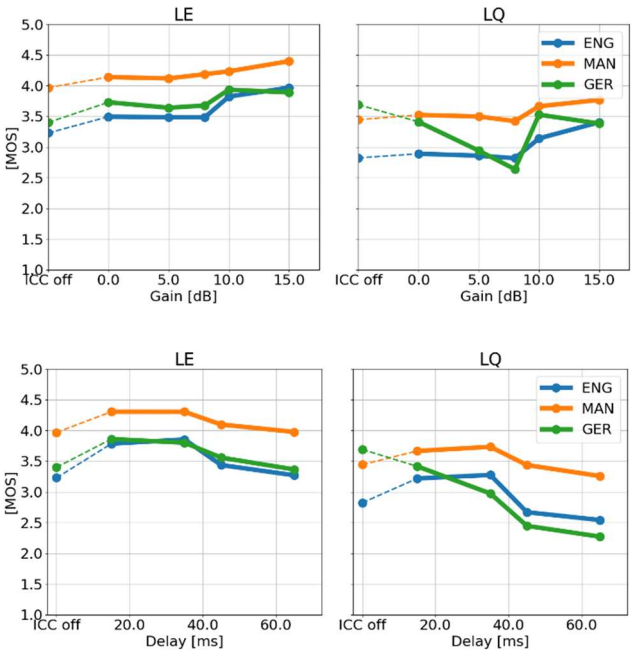

Figure 2: Test results for LE/LQ at 50 km/h versus gain (upper) and versus delay (lower)

3. Auditory Results

The results of the listening tests are illustrated for each driving condition in Figure 2 to Figure 4. Each figure shows the results for LE/LQ versus the ICC system parameters (additional) gain and processing delay. In case of more than one condition for a certain delay or gain, the per-condition results are averaged.

Results for the three different languages are shown as individual curves in each graph. Note that the differences across the three languages are analysed in section 4.

All results for standstill condition (0 km/h) for LE and LQ are all located in the upper MOS range (> 4.5), they are left out here for sake of simplicity.

The results for the driving condition 50 km/h are shown in Figure 2. As expected, results for LE only slightly improve for increasing gains, because even with deactivated ICC system, already an moderate LE is obtained (between 3.3 and 4.0). Results for LQ are almost constant versus additional gain.

With increasing processing delay, speech degradations are introduced, which is slightly noticeable for LE and more obvious for LQ.

A striking issue here is the drop of LQ in language GER at gain 8 dB (about 1.0 MOS compared to *ICC off*). As mentioned in the description of the ICC system, conditions up to this gain only include a slight feedback compensation and audible artefacts are more likely to occur here.

For higher gains, LQ improves again due to a more advanced feedback compensation. Surprisingly, these audible degradations at 8 dB (and partially also at 5 dB) were only assessed in this manner by the German panel. This effect is visible in the results of the next two driving conditions as well.

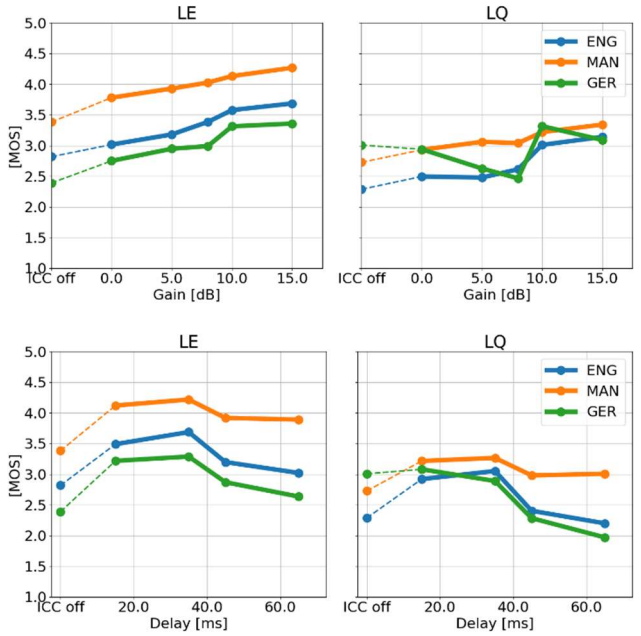

Figure 3: Test results for LE/LQ at 100 km/h versus gain (upper) and versus delay (lower)

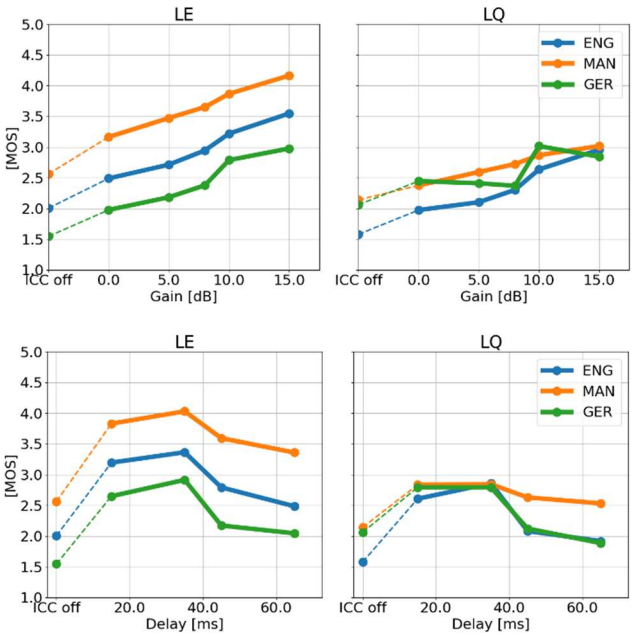

Figure 4: Test results for LE/LQ at 120 km/h versus gain (upper) and versus delay (lower)

The results for the driving condition 100 km/h are shown in Figure 3. Here the impact of increasing gain of the ICC system is more visible than in the previous condition. For all languages, the improvement in LE is about 1.0 MOS. Also for LQ, a slight improvement versus gain can be observed.

The aforementioned drop in LQ is also visible in this driving condition, but less pronounced (about 0.5 MOS compared to *ICC off*). Due to the increased amount of background noise, the additional artefacts seem to be less audible than before.

The results for the driving condition 120 km/h are shown in Figure 4. Since this scenario provides the highest background noise level, it also contains the lowest results for LE and LQ with *ICC off*. Especially LE is significantly improved (up to 1.5 MOS) with increasing gains.

With increasing delay, a decrease in LE and LQ can be observed for all languages. However, for both attributes, even the worst delay conditions provide at least same (in most cases better) results than for *ICC off*.

4. Analysis of Test Language

As seen in the results of Figure 2 to Figure 4, results of the three different languages differ considerably across identically processed conditions. In order to globally investigate these differences, a correlation analysis in form of a scatter plot between the three listening tests is conducted and shown in Figure 5 (only for LE). Subjects of the MAN listening test evaluated much higher scores compared to ENG and even more to GER. For moderate conditions and better (> 3.5 MOS), ENG and GER provide highly comparable results. For conditions of worse LE, the results for GER are more pessimistic than for ENG.

As a second analysis, pairwise comparisons with consideration of 95% confidence intervals (CI95) were evaluated for each couple of tested languages. CI95 values per condition were calculated according to ITU-T P.1401 [11] methodology.

First, pairs for each combination of LE/LQ values per condition within the same test language were created. Then the absolute differences were calculated for all these value pairs. The magnitudes of these differences were then corrected by the sum of the two corresponding CI95 values, which takes the uncertainty of the auditory data into account.

In case the sign of the difference was switched by this correction, such a comparison is assumed to be equal, i.e. the difference is set to zero. According to equation (1), for each test language with $N = 46$ test conditions, $M = 1035$ comparison pairs and differences are formed.

$$M = \frac{(N-1) \cdot N}{2} \tag{1}$$

These M corrected differences are then compared between all couples of languages. In case the sign between two corresponding comparisons is different, this is considered as a single ranking error between these languages.

The ranking errors are counted for all possible comparisons and is then divided by M. If this value is 0.0, this would indicate that all conditions are evaluated with exactly the same order in both languages. Vice versa, a value of 1.0 would indicate that all conditions between two test languages were evaluated exactly in the reversed order.

The paired comparison results are shown in Table 3 (for LE) and Table 4 (for LQ). The analysis provides a quite high consistency in ranking of all three languages - independent of their absolute shifts.

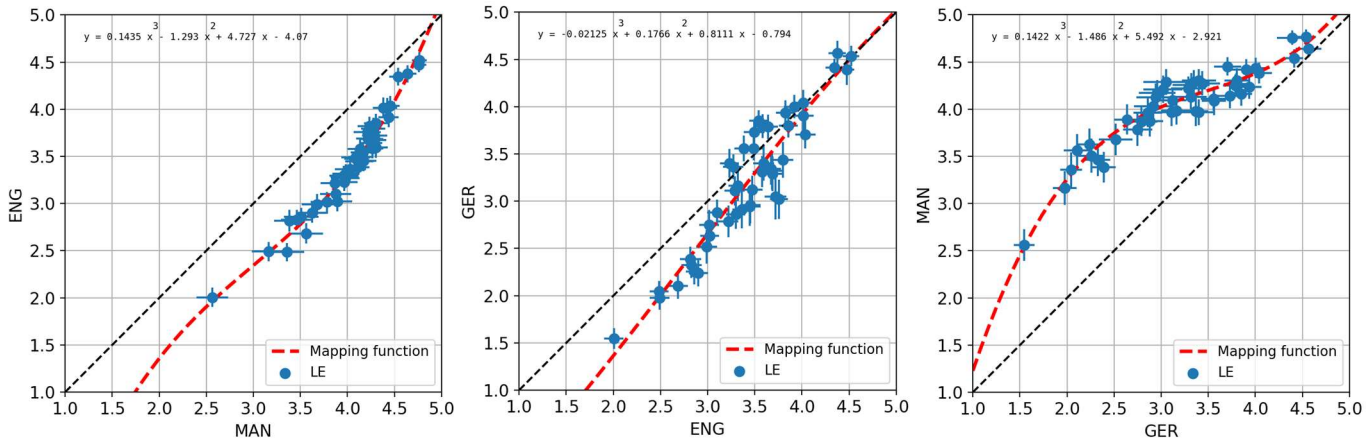

Figure 5: Correlation analysis for LE between ENG, MAN and GER language

While the slightly increased ratings for LQ (up to 2.51%) may be explained by the drop in LQ (see section 3), there are either no or only negligible number of rank errors for LE.

Table 3: Paired comparison ratings for LE

	ENG	MAN	GER
ENG	-	0.00%	0.68%
MAN	0.00%	-	0.00%
GER	0.68%	0.00%	-

Table 4: Paired comparison ratings for LQ

LQ	ENG	MAN	GER
ENG	-	0.48%	2.03%
MAN	0.48%	-	2.51%
GER	2.03%	0.51%	-

5. Conclusions

The results of data analysis of three auditory tests made in two different laboratories show the level of inter-lab repeatability in case of identically processed test speech samples are used. While results for and ENG and GER are quite close regarding LE and LQ, larger differences for MAN could be observed. The panel of Chinese evaluated the speech material consistently higher than the others. It may be discussed if this observation of this study is due to a different perception of speech and noise ("cultural bias") or if it can be explained by the tonal phonology of the Mandarin language (may have benefits on the rather stationary driving noise).

6. Acknowledgements

The listening tests conducted at Mesaqin.com in English and Mandarin language were funded by the ETSI Project STF 575, "Methods for Objective assessment of Listening Effort based on subjective test data bases".

The listening test conducted at HEAD acoustics GmbH in German language were part of the research project ZF4707701SS9, which is funded by the German Federal Ministry of Economics and Technology (BMWi).

References

[1] Peelle, J.E. (2018) Listening Effort: How the Cognitive Consequences of Acoustic Challenge are Reflected in Brain and Behavior, Ear and Hearing, 39(2): 204-214.

[2] Strand, Julia & Brown, Violet & Merchant, Madeleine & Brown, Hunter & Smith, Julia. (2018). Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures. Journal of Speech Language and Hearing Research.

[3] Rec. ITU-T P.800: "Methods for subjective determination of transmission quality", 08/1996.

[4] Holub, J., Avetisyan, H., Isabelle, S. Subjective speech quality measurement repeatability: comparison of laboratory test results. Int. Journal Speech Technology 20, 69–74 (2017).

[5] Goodman, D. J., & Nash, R. D. (1982). Subjective quality of the same speech transmission conditions in seven different countries. IEEE Transactions on Communications, 30(4), 642–654.

[6] ETSI TS 103 281: "Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals", V1.3.1, 2019-05.

[7] Jan Reimes and Christian Lüke, "Perceived listening effort for in-car communication systems", 13th ITG conference on Speech Communication, Oldenburg, Germany, September 2018.

[8] ETSI TS 103 558: "Methods for objective assessment of listening effort", v1.1.1, 2019-11.

[9] ETSI ES 202 396-1: "Background noise simulation technique and background noise database", V1.7.1, 2017-10.

[10] ITU-T Handbook of Subjective Testing Practical Procedures.

[11] ITU-T Rec. P.1401 "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models", 01/2020.