Overall End-to-End Conversation Quality in Complex Scenarios

Stefan Bleiholder, Nils Rohrer, Frank Kettler

HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail:telecom-consulting@head-acoustics.de

Abstract

Conversational quality may be diversely disturbed, e.g. by distorted speech sound, audible echo, sluggish interaction due to long delay, insufficient double talk capability and more. The complexity arises from the perceptual weighting of these single disturbances in the conversation context, the combination of multiple disturbances and the mutual influence of disturbances from the phones used on both ends. Thus, the determination of overall end-to-end communication quality is a very complex and challenging task, both, for test subjects in auditory conversational tests, and -even more- for instrumental models. A conversational test was conducted using commercial wireless phones on both ends, which have been specifically manipulated in order to cover a wide range of possible impairments. The results reflect overall end-to-end quality for single disturbances on one side, on both sides, multiple disturbances and also the mutual influence and interaction between the phones. The results of this study are discussed in this contribution and future work is outlined.

Introduction

In order to systematically evaluate the complexity of conversational quality rating by humans, a conversational test was designed and conducted using a number of diversely modified terminals. Thus, one-dimensional quality impairments (such as impaired sound quality or echo) could be assessed as well as multi-dimensional quality impairments (such as impaired sound quality and echo), all in conversational context. The demands on the conversation test procedure are high: it needs to guarantee a realistic, natural conversation between two naïve subjects, needs to be reproducible and well-balanced for both subscribers, short on the one hand, but covering all aspects of a natural conversation on the other hand (single talk, interaction and double talk).

ITU-T P.805 [1] suggests two methods to test overall conversational quality, the so-called “Kandinsky” Tests and Short Conversation Tests:

- Kandinsky Test, advantage: natural discussion between subjects about given neutral task, designed to initiate a natural conversation including single and double talk periods. Disadvantage: time consuming
- Short Conversational tests, advantage: short duration using guided predefined tasks for subjects, high reproducibility. Disadvantage: unnatural conversation, low probability of double talk occurrence.

It was decided to use the Kandinsky tests for this purpose due to its higher naturalness of conversations. Pictures with colored geometrical figures, e.g. by Kandinsky, are overlaid with scattered numbers in the picture, differently for each subject to initiate the conversation. The task for both subjects is to identify identical numbers at identical positions. Typically, a natural conversation will develop. When there is a single number on only one picture, this will cause a single talk situation between both subjects. Vice versa, identical numbers at similar, but different positions cause interaction and double talk situations between both users. For each conversation test, test participants used new pictures with new numbers. The quantity and the distribution of numbers in both pictures was tested in several test runs before, in order to balance the duration and the naturalness of each conversation.

Test persons were invited for the test sessions, the duration was limited to approximately 90 minutes. Within this time frame 24 conditions should be tested, thus the discussion time needed to be limited to approximately 3 minutes per condition. For expert test subjects, this is a common and feasible task. For naïve test subjects, judging overall conversational quality in a complex scenario with multiple possible impairments in about 3 minutes per condition is a challenging task. In order to ease the test conduction for subjects and ensure a relaxed atmosphere,

- tests subjects were invited pairwise, they needed to know each other before,
- they were briefed very systematically and thoroughly with a detailed oral explanation of the task and the operation of the web-based user interface.
- An additional hands-on video presentation of two expert test subjects was given to the naïve ones showing typical elements of a telephone conversation, such as single talk, interaction and natural double talk.
- The test rooms were decorated with objects from everyday life in any given living room in order to create a relaxed environmental atmosphere for the subjects.
- The first two conversations were for training purpose to get familiar with the procedures, typical impairments and the user interface.
- A timer for both subjects ensured a minimum 3 min duration of each conversation.

The overall quality rating was given on a 5-point ACR-scale as per [2] with intermediate steps similar to the ones in [3]. The plausibility of the scores chosen by the user was checked as the conversational test went on. Test persons received an expense allowance for participation.

Conversational Test Conduction

In the following, two conversational tests are described that were conducted with expert test subjects (CT1) and a number of naïve test subjects (CT2). The physical setup was according to **Figure 1**. Two test subjects are sitting in two different office rooms, in front of a table with a computer and four different cordless telephone devices (DECT, AVM FRITZ!FON C5). The phones are paired with a FRITZ!BOX DSL router. The DSL router is registered at an Asterisk private branch exchange software acting as the VoIP registrar. In between the two DSL routers a Netem based impairment generator introduces delay, jitter and packet loss into the transmission path.

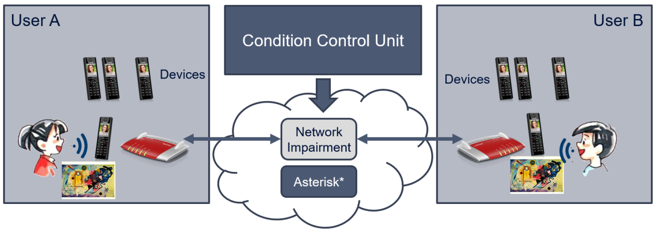

Figure 1: Test setup

Six out of eight cordless devices have been intentionally modified (impaired) in terms of signal processing and acoustic properties (see **table 1**). The DSL router was equipped with a special firmware allowing artificial echo generation. The computer provides a GUI where the users can start the next test run and a means to rate on the connection, once the conversation is over. Depending on the test run chosen by the users, the network and routers are automatically configured according to the impairment properties of the test.

Table 1: Used Devices

Device	Impairment	Direction
#1	high-quality, off-the-shelf device	-
#2	frequency response bandpass	Tx
#3	frequency response highpass	Rx
#4	double talk attenuation 12 dB	Tx
#5	double talk attenuation 30 dB	Tx
#6	high-quality, of-the-shelf device	-
#8	level variation 6 dB quieter	Tx
#10	level variation 9 dB louder	Tx

The GUI also informs both test subjects, which phone to use for the next call. Hence, it was possible to create and test an end-to-end telephone conversation including all major impairment factors important for a conversation. Background noise simulation was intentionally omitted, to limit the total number of test conditions.

Test Conditions

Table 2 shows the labels for all conditions that were tested in CT1 and CT2 including the cordless devices used for the respective call. Since not all conditions can be tested symmetrically, each side (User A, B) of the condition is labeled independently. Training conditions are omitted. For the reference condition 16, the connection was established between device #1 and device #6, both are the original high-quality cordless terminals with balanced acoustic properties (individually labeled as “nominal” in **Table 2**). The devices have nominal loudness ratings and balanced frequency response characteristics in Tx and Rx direction, very good double talk (DT Type 1) and echo performance values (Echo Loss > 46 dB) and a low round-trip delay (70 ms). For the other conditions, different impairments are combined, some implemented in the terminals, some in the network. In total, six different kinds of impairments were used, coloration via filter characteristics, lowering the transmitted frequency bandwidth (see **Table 3**), level variations (see **Table 4**), changed double talk properties (see **Table 5**), introduced packet loss, delay (see **Table 6** and **Table 7**) and echo (see **Table 8**).

Table 2: Conditions and Devices

Nbr.	Label 1	Label 2	Device 1	Device 2
1	EA12	EA12	#1	#6
2	Pl8	Pl8	#1	#6
3	DT12, EA(36-9)	L+9, EA36	#10	#4
4	DT30, D500	D500	#1	#5
5	EA(36-9)	L+9, EA36	#10	#6
6	DT30	nominal	#1	#5
7	D500	D500	#1	#6
8	DT12, EA(36+6)	L-6, EA36	#8	#4
9	D200	D200	#1	#6
10	DT12	L-6	#8	#4
11	nominal	L+9	#10	#6
12	nominal	Rx HighPass	#1	#3
13	DT12, EA36	EA36	#1	#4
14	Nominal	Rx HighPass, BandPass	#2	#3
15	DT12	L+9	#10	#4
16	Ref	Ref	#1	#6
17	DT12, D500	D500	#1	#4
18	EA36	EA36	#1	#6
19	nominal	BandPass	#2	#6
20	DT12	nominal	#1	#4
21	EA(36+6)	L-6, EA36	#8	#6
22	nominal	L-6	#8	#6

The conditions are labeled according to the effective impairment, which is audible for the test persons: In condition 6, named “DT30” (user A) and “nominal” (user B),

- User A experiences double talk attenuation of 30 dB (introduced by User B device) and otherwise an unimpaired connection compared to the ref. condition 16.
- User B experiences an unimpaired connection.

For more complex conditions, e.g. condition 3, labeled “DT12, EA(36-9)” for User A, three impairments are combined: User A perceives

- a double talk attenuation of 12 dB,
- an echo from the far end side (User B) with an echo attenuation of 36 dB
- the echo amplified by the sending level elevation of 9 dB, implemented in the device of User A (resulting in an echo attenuation of 27 dB).

Table 3: Filter Conditions

Condition Id	Label	Filter
19	BandPass	Tx BandPass Dev.#2
12	Rx HighPass	Rx HighPass Dev. #3
14	Rx HighPass, BandPass	Combination of Dev. #2,#3

Table 4: Level Variation Conditions

Condition Id	Label	Level
3,5,11,15	L+9	+9 dB mic gain (louder)
10,21,22	L-6	-6 dB mic gain (quieter)

Table 5: Double Talk Conditions

Condition Id	Label	DT attenuation
3,8,10,13,15,17,20	DT12	12 dB
4,6	DT30	30 dB

Table 6: Packet loss Conditions

Condition Id	Label	Packet loss
2	Pl8	8%

Table 7: Delay Conditions

Condition Id	Label	Additional delay
9	D200	200 ms
4,7,17	D500	500 ms

Table 8: Echo Conditions

Condition Id	Label	Echo attenuation
3	EA12	12 dB
5,7,10,13,21	EA36	36 dB

For CT1, 12 expert test subjects voted on all conditions, six on each side. For CT2, 24 naïve test subjects voted on all conditions, 12 for each side. The results are averaged to end-to-end Mean Opinion Scores, designated as $\mathrm{MOS}_{\mathrm{E2E}}$ on both sides A and B ($\mathrm{MOS}_{\mathrm{E2E,A}}$, $\mathrm{MOS}_{\mathrm{E2E,B}}$)

Discussion of Results

Figure 2 shows the $\mathrm{MOS}_{\mathrm{E2E}}$ averaged for all expert test subjects, separately for User A (blue) and User B (orange), including the 95 % confidence intervals. These intervals are given informatively only, as 6 votes are statistically insufficient.

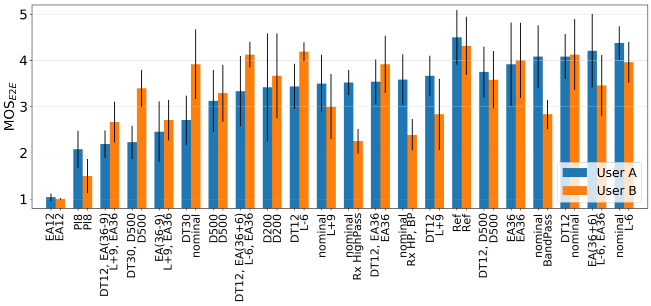

Figure 2: CT1 Test Results, Expert Test Subjects

The corresponding $\mathrm{MOS}_{\mathrm{E2E}}$ results averaged for all naïve test subjects for user A (blue) and user B (orange) are given in **Figure 3.** For all test subject groups, results of unsymmetrical conditions might lead to different MOS values for user A and B, because the auditory experience for each user differs.

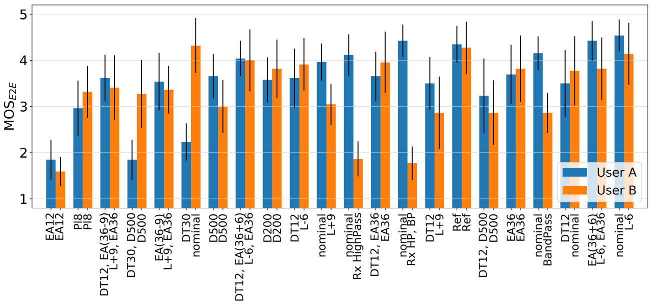

Figure 3: CT2 Test Results, Naïve Test Subjects

As the test structure and the task for test subjects is rather complex, the votes are analyzed with regard to plausibility:

- The echo conditions for User A (EA12, EA(36-9), EA36, EA(36+6)) for both groups of test subjects show increasing $\mathrm{MOS}_{\mathrm{E2E}}$ with increasing echo attenuation (see **Table 9**).
- The delay conditions (Ref, D200, D500) lead to decreasing $\mathrm{MOS}_{\mathrm{E2E}}$ scores with increasing delay (see **Table 10**).
- The double talk conditions show the same tendency: the stronger the double talk attenuation impairs, the lower the $\mathrm{MOS}_{\mathrm{E2E}}$ scores (see **Table 11**).
- For the filter condition with the Bandpass in Tx direction (Tx BP), both expert and naïve test subjects penalize the lower frequency range with an MOS of 2.8 and 2.9 respectively. For the Rx direction highpass, both groups rate the coloration with 2.3 and 1.9 MOS respectively. A combination of both highpass and bandpass filters does not show an additional significant decrease of the MOS scores, the scores seems to saturate already (2.4 (expert) and 1.8 (naïve), see **Table 12**).

Table 9: Echo Condition Scores

	EA12	**EA(36-9)**	**EA36**	**EA(36+6)**
Effective EA	12 dB	27 dB	36 dB	42 dB
$\mathrm{MOS}_{\mathrm{E2E}}$ Expert (A)	1.0	2.5	3.9	4.2
$\mathrm{MOS}_{\mathrm{E2E}}$ Naïve (A)	1.8	3.5	3.7	4.4

Table 10: Delay Condition Scores

	Ref	**D200**	**D500**
Effective delay	70 ms	270 ms	570 ms
$\mathrm{MOS}_{\mathrm{E2E}}$ Expert (A/B)	4.5/4.3	3.4/3.7	3.1/3.3
$\mathrm{MOS}_{\mathrm{E2E}}$ Naïve (A/B)	4.3/4.3	3.6/3.8	3.7/3.0

Table 11: Double Talk Condition Scores

	Ref	**DT12**	**DT30**
Effective DT attenuation	0 dB	12 dB	30 dB
$\mathrm{MOS}_{\mathrm{E2E}}$ Expert (A)	4.5	4.1	2.7
$\mathrm{MOS}_{\mathrm{E2E}}$ Naïve (A)	4.3	3.5	2.2

Table 12: Filter Condition Scores

	Ref	**Tx BP**	**Rx HP**	**Tx Bp & Rx HP**
$\mathrm{MOS}_{\mathrm{E2E}}$ Expert (B)	4.3	2.8	2.3	2.4
$\mathrm{MOS}_{\mathrm{E2E}}$ Naïve (B)	4.3	2.9	1.9	1.8

Figure 4 and **5** show the $\mathrm{MOS}_{\mathrm{E2E}}$ comparison between expert (light blue) and naïve (dark blue) test subjects for user A and user B respectively. For 62.5 % of all conditions, the deviation between $\mathrm{MOS}_{\mathrm{E2E}}$ scores for naïve and expert test subjects is lower than 0.5 MOS. Only for 20.8 % of all conditions the deviation is ≥ 0.7 MOS. On average, the expert test subjects rate the impairments equal or more critical than the naïve test subjects.

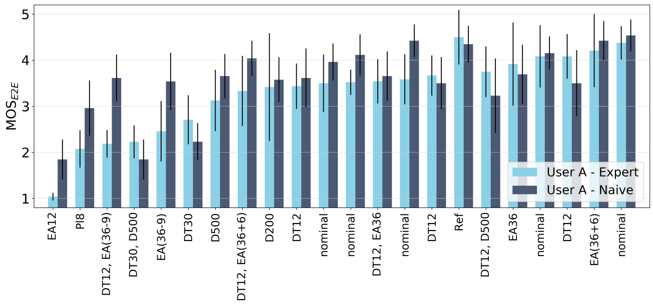

Figure 4: $\mathrm{MOS}_{\mathrm{E2E}}$ User A - Expert vs. Naïve Test Subjects

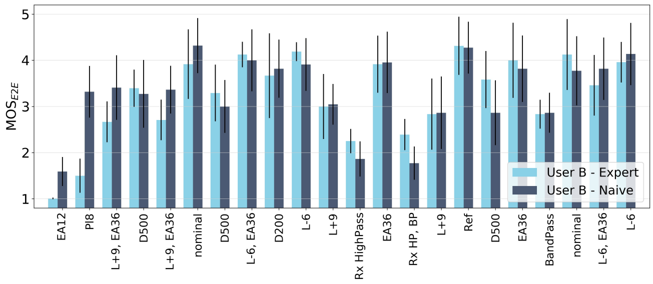

Figure 5: $\mathrm{MOS}_{\mathrm{E2E}}$ User B - Expert vs. Naïve Test Subjects

There are mainly two aspects, where expert and naïve subjects seem to have different requirements:

- For the very strong and annoying echo condition EA12, the experts give an unambiguous 1.0 MOS vote on both sides (both Users A and B) with CI95 of 0.1 (User A) and 0.016 (User B). Vice versa, the naïve subjects still give a 1.8 MOS (User A) and 1.6 MOS (User B) score, even though they complained about the strong echo in their notes for this condition. Potentially, naïve subjects are accustomed to echoes in communication scenarios and therefore do not penalize echoes as strongly as experts do.

- A similar tendency can be seen in the complex 27 dB echo condition "EA(36-9)", and in the same condition with an additional 12 dB double talk attenuation "DT12, EA(36-9)" for User A.
- The second aspect is the annoyance caused by packet loss (condition "PL8" on both sides). Experts judge significantly more critically.

Since under the majority of conditions there are only minor differences between expert and naïve test results and the overall root-mean-squared deviation is 0.6 MOS (User A) and 0.5 MOS (User B), the results of CT1 and CT2 were combined in the following. These results are shown in **Figure 6**. All observations derived from the CT2 and CT1 tests hold true also for the combined results. In the following, further analyses and conclusions refer to these results.

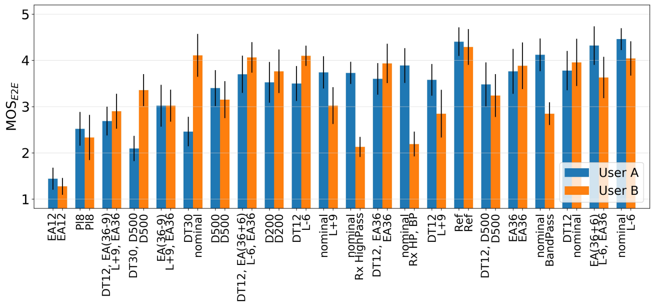

Figure 6: Joined Results for Expert and Naïve Subjects

From Single-Side MOS to MOS_{E2E}

An end-to-end MOS score, as it is perceived by the subscriber, is always influenced by the acoustic properties of both devices used in the connection, the influence of the network and the reaction of the terminals on these impairments. There are well-known instrumental methods available to measure the performance of terminals, e.g. E-MOS analysis according to [4] for echo performance or MOS-LQO according to ITU-T P.863 [5] for listening quality. The echo test conditions in **Table 13** for example aim on this interactive influence between the terminals, all in the context of conversational quality.

Table 13: Echo Condition Scores

	EA12	EA(36-9)	EA36	EA(36+6)
MOS_{E2E} Joined (A)	1.4	3.0	3.8	4.3

Table 14: Filter Condition Scores

	Ref	Tx BP	Rx HP	Tx Bp & Rx HP
MOS_{E2E} Joined (B)	4.3	2.8	2.1	2.2

The E-MOS score of a terminal with 12 dB or 36 dB echo loss can be measured in instrumental tests, but the sensitivities of the other terminal or network elements need to be taken into account to derive the effective end-to-end $MOS_{E2E,echo}$ for the echo perception. In the same way an end-to-end $MOS_{E2E,LQO}$ score for User A is influenced by both the MOS-LQO performance of the device A in Rx direction and the MOS-LQO performance of the device B in Tx direction. The instrumental methods available today measure the performance of terminals. However, relevant is the end-to-end MOS_{E2E} to reproduce users' perception, again, in the context of a conversation (see **Table 14**).

From single quality dimensions to overall conversational quality

The main factors influencing the overall end-to-end quality of a conversation are known in principle, such as listening quality (level, coloration, distortion), absence of echo, interactivity between subscribers, double talk performance and interaction between terminals and network. These different quality dimensions have all the same effect on the overall quality, the lower the scores for each dimension, the lower the overall perceived quality. But the absolute influence of each dimension on overall quality is different for each dimension. **Table 15** compares conditions EA12 and DT30 for User A, one impaired by strong echo, the other by strong double talk attenuation. Condition EA12 shows an overall $MOS_{E2E,A}$ score of 1.4 whereas condition DT30 results in an $MOS_{E2E,A}$ of 2.5 in the conversation test, a difference of more than 1.0 MOS. Interestingly, the individual contribution of each dimension (E-MOS, DT-MOS) is virtually identical. The E-MOS score for a 12 dB echo attenuation calculated as per [4] is 1.8 MOS. According earlier investigations [3], the DT-MOS score for a 30 dB DT attenuation also results in 1.8.

Table 15: Contribution of different Dimensions

	EA12 (A)	DT30 (A)
$MOS_{E2E,A}$ Overall Joined	1.4	2.5
MOS Dimensional	1.8 (E-MOS)	1.8 (DT-MOS)

Conclusion

Two conversational tests with expert and naïve test subjects were conducted in a complex VoIP scenario with multiple impairments of different conversational. The impairments were selected and judged as single disturbances and combined disturbances, all in conversational context. The MOS_{E2E} results for the single disturbances can be used to estimate the individual contribution of each quality aspect in the conversational context. In the same way, the MOS_{E2E} results derived from the multiple disturbances can be used to estimate and combine the individual contributions on overall quality in the conversational context. The first analysis motivates the assumption that individual MOS_{E2E} scores for each quality dimension contribute differently to the overall quality. Further work is planned to justify the assumptions and extend the validity to other communication scenarios.

References

[1] ITU-T Recommendation P.805, "Subjective evaluation of conversational quality," 04/2007.

[2] ITU-T Recommendation P.800, "Methods for subjective Determination of Transmission Quality," 08-1996.

[3] S. Bleiholder, J. Reimes and F. Kettler, "Auditory Assessment of Echo during Double Talk and Double Talk Distortions," in *DAGA*, Rostock, 2019.

[4] S. Bleiholder, J. Reimes and F. Kettler, "Super-Wideband Extension of a Perceptual Based Echo Assessment," in *ITG Fachtagung*, Oldenburg, 2018.

[5] ITU-T Recommendation P.863, "Perceptual objective listening quality prediction," 03/2018.