On the Impact of Combinations of Terminal and Network Degradations on the Conversational Quality of Wideband Telephony

Janto Skowronek[1], Eckhardt Schön[1], Alexander Raake[1], Michal Soloducha[1], Peter Voigt[2], Stefan Bleiholder[3], Frank Kettler[3]

[1] *Audiovisual Technology Group, TU Ilmenau, 98684 Ilmenau, Deutschland, Email: janto.skowronek@tu-ilmenau.de*

[2] *AVM GmbH, 10559 Berlin, Deutschland* [3] *Head Acoustics GmbH, 52134 Herzogenrath, Deutschland*

Motivation

In the field of quality assessment of telephony systems, one recent trend is to gain more knowledge about the impact of terminal equipment on the perceived quality. From a device manufacturers perspective, combinations of terminal and network degradations are quite well understood by the technical experts and also a comprehensive set of technical characterization methods for terminal equipment is available. However, the quality impact of such combinations on the naïve end user, especially in conversational situations, has hardly been studied yet.

Moreover, from a service operators perspective, it is not always possible to clarify, whether any occuring customer complaints are triggered by problems from the network, the terminal device or a combination of both. Hence, service improvements can be facilitated by extending quality monitoring and planning models (e.g. [1, 2, 3]) with a more sophisticated inclusion of terminal characteristics and their interaction with network impairments.

Study Goal and General Approach

To prepare the way for the above mentioned model extensions, the overall goal is to gain more insights on the impact of combinations of terminal and network degradations on the quality perceived by end users. For that purpose, we were primarily interested in obtaining an overview across a rather broad range of different *types* of impairments. Moreover, we addressed in this study also the impact of experimental contexts on results, an issue that is well known in the field (e.g. [4]). More precisely, we considered the following two context factors.

The first context factor is the *corpus effect* (e.g. [5]), a bias caused by the selection of stimuli in a particular experiment. That means, we opted for a single large-scale test instead of a series of smaller individual tests, as it is not trivial to combine the results of such individual tests. Note that *corpus effects* can be overcome by using a sufficient set of anchor conditions reused in individual tests (e.g. [5, 6]). However, at this point in time we did not have a clear picture about which conditions could serve as good anchor conditions; thus we kept the decision to continue with a single experiment.

The second context factor is the test paradigm in terms of listening-only test vs. conversation test. Listening-only tests have a number of advantages compared to conversation tests (see e.g. [7]), such as enabling a higher sensitivity of test participants, or allowing the testing of a larger number of conditions. Conversation tests on the other hand allow the testing of conditions that affect conversational quality, such as echo, delay and double-talk impairments. Those conditions can not be assessed in listening-only tests or they require very specific test protocols such as third-party listening tests. Moreover, conversation tests allow to collect quality ratings in a more natural test situation, an actual conversation. Further, after After considering the advantages and disadvantages, we opted for the conversation tests, as we thought that the more natural situation and inclusion of echo, delay and double-talk impairments better matched the primary study goal than a higher sensitivity of test participants and larger total number of test conditions.

Test Setup

Test Design: With the decision to run a conversation test, we needed to account for two actually contradicting objectives: (a) the need to cover a broad range of conditions to obtain the best possible picture and thus requiring a rather large number of test calls; and (b) sticking to a single conversation test in order to keep the experimental context as constant as possible, which in turn limits the number of possible test calls if the test should be practically feasible. As a compromise we applied a *Balanced Incomplete Block Design* (BIBD).

The essential idea of BIBDs is to distribute a set of conditions across test sessions such that three constraints are fulfilled: (1) each test session contains only a subset of the conditions; (2) the number of occurences across all sessions is the same for all conditions; (3) all possible combinations between any two conditions are equally often used. Since a BIBD is characterized by five parameters, a common notation is a 5-tuple of the form $(v - b - r - k - \lambda)$. Moreover, not all combinations of those five parameters are actually possible, and one can find literature on existing BIBDs. In [8] we found a (34-34-12-12-4) BIBD, which allows to have 12 calls (= 24 ratings) for 34 test conditions, while limiting the number of calls per test participant group to 12 and the overall number of groups to 34. Table 1 explains the meaning of the five parameters, their interpretation for the current test, and the actual values for the chosen design.

Test Conditions: While the chosen BIBD allowed to test 34 conditions, we still needed to make an informed selection of degradations to test. For that purpose we

Table 1: Meaning of the five parameters of a Balance Incomplete Block Design and the values for the current study.

Parameter		Interpration for current test	Value
v	# treatments	# technical conditions	34
b	# blocks	# sessions, i.e. # test participant groups	34
r	# replications	# observations = # calls per technical condition	12
k	block size	# test calls in a session	12
λ	# concurrences, i.e. # blocks containing any 2 points	# sessions containing any pair of technical conditions $\Rightarrow$ *overlap* of conditions	4

conducted a discussion round with ten experts from the three project partners TU Ilmenau, AVM and HEAD acoustics, in which we applied a three-step selection process: 1. selection of types of degradations, 2. selection of combinations, and 3. selection of actual values.

The terminal degradations were realized using differently modified devices provided by AVM; the background noise was realized with an eight-channel acoustic simulation [9] provided by HEAD acoustics, and the terminal degradations were realized using Netem. Some conditions could only be realized as asymmetric calls, i.e. the degradation was only present in one direction but not in the other. Therefore we needed to run two calls instead of one to get the planned number of ratings. As this was necessary in eight cases, the number of actual conditions grew from 34 to 42. Table 2 summarizes all chosen conditions.

Test Procedure: We invited 34+1 groups of naïve test participants. Thus we collected ratings from 70 participants (male/female ratio 44/56, age 19...61, mean age 25), who were affiliated with the TU Ilmenau (mainly undergraduate and PhD students). The test participants came to two sessions of about 45-65 minutes each, in which the participants had conversations over the test system, using the standardized short conversation test scenarios of [10].

Analysis Method

The analysis goal was on identifying any different behaviors when combining the degradations, but not on hypotheses testing in the sense of finding significant differences between conditions. For that reason, the analysis was mainly visual, while running statistical tests avoided that we *see more in the data than there actually is.*

Before outlining the visual analysis, a remark on the chosen statistical test method. The text book procedure for non-normal BIBD data is a Durbin test with non-parametric PostHoc tests. However, we refrained from this, since (a) the data is strictly speaking not a BIBD given the issue of asymmetric calls (see above) meaning that a Durbin test cannot be applied without data modification, and (b) these procedures are intended for testing specific hypotheses, which was not the goal here. Instead we chose as alternative non-parametric pairwise comparisons, i.e. Man-Whitney-U tests, to obtain an *indication* of the statistical significance of any observed effects.

Table 2: Technical characteristics of the test conditions.

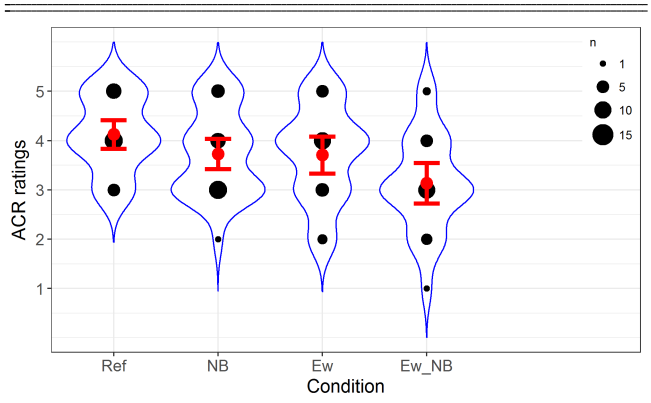
Reference condition	
Next-Generation DECT phones (AVM C5), local network, WB: 50 - 7000 Hz, G.722 codec, Overall Loudness Rating 10 dB, mouth-to-ear delay 150ms, echo attenuation 96dB, double-talk attenuation 96 dB	
Single degradations	
PLs	Packet loss, random, 8%
PLw	Packet loss, random, 2%
Es	Talker echo attenuation 12 dB at 300ms round-trip delay
Ew	Talker echo attenuation 36 dB at 300ms round-trip delay
L-6	Send loudness reduced by 6dB
L+6	Send loudness increased by 6dB
Fs	Bandpass filter 400-1000 Hz (asymmetric condition: realized as a cascade of two devices)
Ref|Fs	asymmetric condition: participant has *Ref* condition while conversation partner has *Fs* condition
Fw	Highpass filter 1000 Hz
NB	Narrowband: 300-3400 Hz, codec tandem G.711 & G.726
DTs	Double-talk attenuation 30dB
DTw	Double-talk attenuation 12dB
Ds	(echo-free) delay of 800ms one-way
Dw	(echo-free) delay of 400ms one-way
Ns	asymmetric condition: transmitted background noise (street noise) with 70 dB(A) noise level at far end, i.e. in environment of conversation partner
Nw	asymmetric condition: same noise at 55 dB(A)
Ref|Ns	asymmetric condition: participant has Ref condition but is in the background noise environment *Ns*
Ref|Nw	asymmetric condition: participant has *Ref* condition but is in the background noise environment *Nw*
Combined degradations	
symmetric conditions: *PLs_NB*, *PLs_NB*, *Ds_DTs*, *Dw_DTw*, *Fw_L-6*, *Es_Ds_DTs* (with Ds adjusted to 500ms), *Ew_Dw_DTw* (with Dw adjusted to 300ms), *Es_L-6*, *Ew_L+6*, *Es_NB*, *Ew_NB*, *Es_DTs*, *Ew_DTw*	
asymmetric noise conditions (notation analog to *Ns* & *Ref|Ns* above): *PLs_Ns*, *PLs|Ns*, *PLw_Nw*, *PLw|Nw*, *DTs_Ns*, *DTs|Ns*, *DTw_Nw*, *DTw|Nw*	
asymmetric conditions: *Fw_L+6* (cascade of two devices), *Ref|Fw_L+6* (participant has *Ref*, while conversation partner has *Fw_L+6*)	

Concerning the visual analysis, we opted for a combination of three data representations. First, we plotted the Mean Opinion Score (MOS) and its 95% confidence interval around it as (red) errorbar. Second, we generated (blue) violin plots, which represent the probability densitity function of the ratings, and can be seen as vertical and continous representations of the histograms. Third, we coded the number of ratings for each value on the Absolute Category Rating scale [10] by (black) dots with different sizes. Thus, these dots are a discrete representation of the histograms "seen from above".

Results

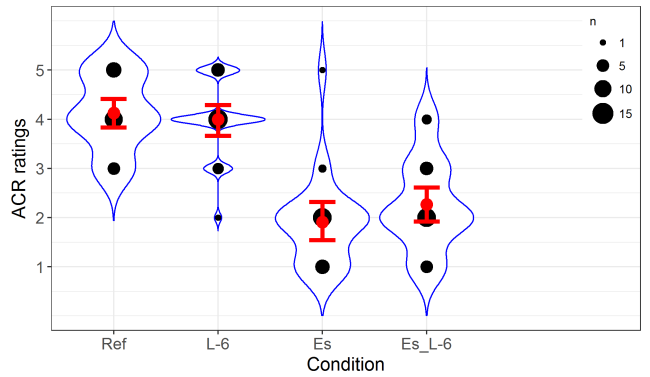
Since the data comprised 42 different conditions, we present in this paper first the detailed analysis for a few examples, followed by a brief summary of all results.

Example results in detail: Figure 1 presents the visual analysis for three example combinations: echo and narrowband (Panel a), echo and loudness (Panel b), and double-talk and noise (Panel c). In addition, we added the plots for the corresponding individual degradations, and we added for all condition pairs the p-values of the Man-Whitney-U tests. In the following discussion, the term *impact of a degradation* refers to the difference in MOS between a degradation and the reference *Ref*.



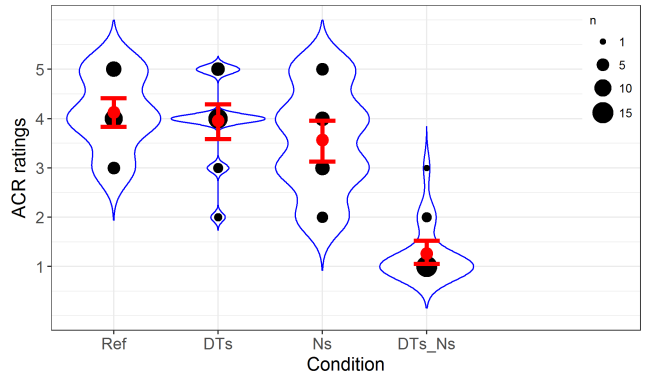
p-values:	*Ref* vs. *NB*:	$= 0.100$	*Ref* vs. *Ew*:	$= 0.139$
	Ref vs. *Ew_NB*:	$= 0.001$	*NB* vs. *Ew*:	$= 0.951$
	NB vs. *Ew_NB*:	$= 0.043$	*Ew* vs. *Ew_NB*:	$= 0.053$

Panel (a)



p-values:	*Ref* vs. *L-6*:	$= 0.637$	*Ref* vs. *Es*:	< 0.001
	Ref vs. *Es_L-6*:	< 0.001	*L-6* vs. *Es*:	< 0.001
	L-6 vs. *Es_L-6*:	< 0.001	*Es* vs. *Es_L-6*:	$= 0.117$

Panel (b)



p-values:	*Ref* vs. *DTs*:	$= 0.599$	*Ref* vs. *Ns*:	$= 0.059$
	Ref vs. *DTs_Ns*:	< 0.001	*DTs* vs. *Ns*:	$= 0.167$
	DTs vs. *DTs_Ns*:	< 0.001	*Ns* vs. *DTs_Ns*:	< 0.001

Panel (c)

Figure 1: Visualization of the perceptual ratings for three *representative* examples of the 42 tested conditions. For explanation of figure elements see text.

Panel (a): The combination of echo and narrowband is an example for a summation of individual impacts. In terms of MOS (red dots), the impact of *Ew_NB* is approximately the sum of the impacts of *Ew* and *NB*. Comparing the distributions (blue violin plots and black dots), one sees a clear shift downwards for the combined degradation, which confirms the observation in terms of MOS.

Panel (b): The combination of echo and loudness is an example for an inhibition effect. Since the MOS of *Es_L-6* are slightly (though not significantly) higher than the MOS of *Es*, we see the trend that the combination of *Es* and *L-6* inhibits the impact of *Es* alone. A slight upwards shift of the distribution from *Es* to *Es_L-6* confirms this behavior. While this *inhibition effect* is a result in terms of the perceptual ratings, this particular case is also an example for an underlying technical interaction: *L-6* means a reduced speech level of 6dB in send-direction and the level of the talker echo signal is also reduced, as the echo is generated in the transmission chain behind the sending device.

Panel (c): The combination of noise and double-talk is an example for a complex combination of different effects. In terms of MOS and the distributions, the individual degradations support each other in the sense that the impact of the combined degradation is much stronger than the sum of the individual impacts. Moreover, the strong background noise in the far end (*Ns*) caused the double-talk attenuation (*DTs*) in the near-end device always when the near-end speaker was talking, even when the speaker in the far-end was not speaking. Since this behavior is strictly speaking not a real double-talk impairment, this *DTs_Ns* condition is an example for a technical interaction that generates a new *type* of impairment. Furthermore, this condition shows that test participants take also conversational cues and not only signal-related cues into account: the conversation partner at the far-end could hardly understand the near-end speaker, who in turn gave also low ratings even though he could understand the far-end speaker fairly well (except in case of a *true* double-talk situation).

Summary on found effects: After repeating the analysis above for all conditions, we compiled an overview of the effects found, which is shown in Table 3. Concerning the perceptual impact, the table shows that the perception of such combined degradations is more complex than a simple addition. Even though the combination of two degradations is sometimes a summation of the two individual degradations, we have also observed additional effects, i.e. dominance, mutual support, and inhibition (for details see Table 3). Moreover, this picture is even more complex. First, some of the effects are caused by a technical interaction in the sense that one degradation directly changed the strength or even the character of the other degradation. Second, another influencing aspect is the conversational structure, e.g. how often there were double-talk periods or how interactive was the conversation; aspects that are particularly relevant for the impact of the double-talk and delay degradations. Third, degradations can be either attributed to the system or to other aspects of the situation. For instance, the impact of delay is sometimes attributed to the conversation partner [11] instead of the system. And in the current study, the data show that the environmental noise is seen by many participants as part of the test condition and not as some aspect of the environment. Fourth and finally, in

some conditions the test participants used different *types* of quality features for their judgement: quality features that they directly perceive from the signal and quality features that they perceive from the conversation flow.

Table 3: Summary of the found effects and additional influencing factors when combining individual degradations.

Effects in terms of ratings	
Description	Found for conditions
Undetermined: Hardly any effect of combination and of individual impairments	*Dw_DTw, DTw_Nw, DTw|Nw, (Ew_Dw_DTw)*
Ambiguous: No clear direction of effect, i.e. distribution spreads in two directions	*L+6_Fw,L-6_Fw*
Summation: Combined impairment is (approximately) a sum of individual impairments	*Ew_DTw, Ew_L+6, Ew_NB, PLw_NB, PLs|Ns, Ew_Dw_DTw*
Dominance: Stronger impairment hardly / slightly changed by weaker impairment	*Ds_DTs, Es_DTs, Es_NB, PLs_NB, PLw_Nw, PLw|Nw, (Ew_Dw_DTw)*
Weighted sum: Effect of combined impairment is between sumation and dominance	*Es_DTs, PLs_Ns, Es_Ds_DTs*
Mutual support: Combined impairment is stronger than sum of individual impairments	*DTs_Ns, DTs|Ns*
Inhibition: Combined impairment weaker than any individual impairment	*Es_L-6*
Boundary: Impairment close to scale boundary hardly changed by other impairment	*(Es_DTs), (Es_NB), (Ew_Dw_DTw)*
Technical influences	
Description	Found for conditions
One degradation increases/decreases strength of the other degradation	*Ew_L+6, Es_L-6*
Combination generates impairment with a new character	*DTs_Ns, DTs|Ns*
External influences	
Description	Relevant for conditions
Conversational structure: amount of double-talk probably too low (to be checked)	*Dw_DTw, Ds_DTs, DTw_Nw*
Conversational structure: impact of degree of interactivity to be checked	*Dw_DTw, Ew_Dw_DTw, Ds_DTs, Es_Ds_DTs*
Attribution of impairment to system or to conversation partner to be checked	*Dw_DTw, Ew_Dw_DTw, Ds_DTs, Es_Ds_DTs*
Attribution of impairment to test condition, i.e. seen as a quality problem	PLs_Ns, PLs|Ns
Attribution of impairment to environment, i.e. not seen as a quality problem	PLw_Nw, PLw|Nw
Quality features: conversational features appear to be used/important	*Ref|L+6_Fw, Ref|Fs*

Remark: For some conditions not only a main effect but also alternative effects are reasonable. For alternatives that we considered as less likely, the corresponding conditions are put into parentheses.

Discussion and Conclusions

In terms of lessons learned, we can report that it is practically feasible to test combinations of degradations in a single conversation test by a combination of two design aspects: (a) a stepwise selection of a representative set of degradations, and (b) the application of a balanced incomplete block design BIBD allowing for 34 conditions. However, we can also report that it is not trivial to obtain data that truely fulfill the characteristics of a BIBD when running a conversation test. In our case, first technical constraints required to realize a few asymmetric conditions, meaning that the number of actual calls needed to be increased. Second, running a conversation test with real devices required a rather complex setup with partially manual operations, which in turn is to a certain extent errorprone and thus leads to missing data points. As a consequence, the data analysis is not trivial either, as the practical problems mentioned can prevent from running statistical *text book procedures*.

Nevertheless, in this study we identified a number of different effects and influencing factors, which show that the combination of degradations is more than a simple addition. This has some implications on the final goal, which is a better inclusion of terminal device characteristics into the E-Model [3] or alternative approaches such as the Voice over IP monitoring model ITU Rec. P.564 [1]. More specifically, having observed different effects for different combinations of degredations suggests that such model extensions require interaction terms, which comprise different functions for different types of degradations. Further model improvements can be expected when the conversational structure is included, e.g. by means of a conversation structure surface analysis (e.g. [12, 13]) or by means of allowing the model to switch between different modes (such as the interactivity terms in the current version of the E-Model [2]). Investigations in both directions, modeling with specialized interaction terms and inclusion of the conversational structure, are currently planned as the next steps.

Acknowledgments

We thank Tatiana Surdu and Frank Hofmeyer for their support in running and analyzing the conversation test. This work was financed under the grant number ZF4087102MS5 by the German Federal Ministry of Economics and Technology *BMWi*.

References

[1] ITU-T, "Recommendation P.564 – Conformance testing for voice over IP transmission quality assessment models", 2007.

[2] ITU-T, "Recommendation G.107 – The E-model: a computational model for use in transmission planning", 2014.

[3] ITU-T, "Recommendation G.107.1 – Wideband E-model", 2015.

[4] U. Reiter et al., "Factors Influencing Quality of Experience", in: Quality of Experience - Advanced Concepts, Applications, Methods (Eds. S. Möller % A. Raake), Springer, 2014.

[5] M.-N. Garcia, "Parametric Packet-based Audiovisual Quality model for IPTV services", Springer, 2014.

[6] ITU-T, "Rec. P.833 – Methodology for derivation of equipment impairment factors from subjective listening-only tests", 2001.

[7] S. Möller, "Assessment and Prediction of Speech Quality in Telecommunications", Kluwer Academic Publishers, 2000.

[8] J.W. Di Paola et al., "A List of (v,b,r,k,λ) Designs for r $\leq$ 30", in: Proc. 4th Conf. Combinatorics, Graph Theory and Computing, 1973.

[9] ETSI, "Technical Specification TS 103 224 – A sound field reproduction method for terminal testing including a background noise database", 2014.

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

[11] K. Schoenenberg et al., "Why are you so slow? ? Misattribution of transmission delay to attributes of the conversation partner at the far-end", Intern. J. of Human-Computer Studies 72 (5), 2014.

[12] P.T. Brady, "A statistical analysis of on-off patterns in 16 conversations", Bell Syst. Tech. J. 47 (1), pp 73-99, 1968.

[13] K. Schoenenberg et al., "On interaction behaviour in telephone conversations under transmission delay", Speech Communication 63-64, pp. 1-14, 2014.