INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2013-2016

COM 12 – C 267 – E

December 2012

English only

Original: English

Question(s): 15/12

STUDY GROUP 12 – CONTRIBUTION 267

Source: Deutsche Telekom AG

Title: Towards perceptual speech quality dimensions in a conversational situation

Abstract

In this contribution, we present first steps towards the diagnosis of speech quality in a conversational situation. We follow the approach that the conversational process consists of three phases, namely the Listening, the Speaking, and the Interaction phase. In earlier studies, for each of these phases perceptually relevant quality dimensions that provide diagnostic information have been extracted in separate tests. The presented work follows these results of the earlier studies and merges them in a sophisticated experimental paradigm addressing all three phases and their underlying dimensions in a conversational situation. The results show, that there is a need for a test paradigm that allows the participants to perceive each phase separately, and that each expected perceptual dimension is covered by a sufficient number of technical conditions. The presented work is targeted towards the current work item P.CQO of Q15 SG12 and is intended to give new impulses.

1. Introduction

Traditional methods for assessing the quality of transmitted speech (instrumental and subjective methods) provide reliable and valid results, but as described in (Köster et al. 2014), these methods hold certain limitations:

- Only the overall quality is predicted, but not the reasons underlying sub-optimum quality.
- The methods mostly refer to the passive listening situation.

The first point addresses the fact that the MOS value does not give insights into the source of a possible quality-drop, no diagnostic information is obtained. More precisely, two speech samples can be rated with the same MOS value while one is degraded by packet loss and the other one by background noise. To solve this limitation, ITU-T is currently working on two approaches to provide diagnostic information (P.TCA and P.AMD).

Contact:	Friedemann Köster Telekom Innovation Labs, TU Berlin Germany	Tel: +49 30 8353 58255 Fax: +49 30 8353 58409 Email: friedemann.koester@telekom.de
Contact:	Sebastian Möller Telekom Innovation Labs, TU Berlin Germany	Tel: +49 30 8353 58465 Fax: +49 30 8353 58409 Email: sebastian.moeller@telekom.de
Contacts:	Dennis Guse Telekom Innovation Labs, TU Berlin Germany	Tel: +49 30 8353 58874 Fax: +49 30 8353 58409 Email: dennis.guse@telekom.de

The second limitation describes the disadvantage that traditional methods address the passive listening-only situation. This only partly agrees with reality in telecommunications. Therefore, conversation tests to assess the quality in an interactive situation, and speaking tests to assess the quality of active speaking situations, have been designed.

Up to now both limitations and potential solutions have only been considered in isolation. This yields to a trade-off between considering either diagnostic information or the user's situation. This trade-off is tackled in (Köster & Möller 2014), where seven diagnostic quality dimensions for a conversational situation are proposed. However, these dimensions were identified in separate studies for each possible situation of a conversation (listening, speaking, and interacting) which have also been used by (Guéguin et al. 2008). In this contribution the proposed dimensions are picked up and evaluated in a joint experiment which addresses all three phases of a conversation.

After a short review of the proposed quality dimensions for a conversational situation, the conversational experiment is explained. Afterward, the results are presented and discussed. The contribution closes with a conclusion and an outlook on possible next steps.

2. Perceptual quality relevant dimensions of a conversational situation

Telephony systems are in general used for a conversation between two people. The two interlocutors take turns in speaking and listening. This leads to interaction between the two participants. A conversational process can be described as a four-state model: A participant can either listen or speak, and in addition both participants can speak or remain silent at the same time. This description is also recommended by the ITU-T (P.59 1993) and leads to a separation of a conversation into three phases as perceived by one participant (Guéguin et al. 2008): the Listening Phase, the Speaking Phase, and the Interacting Phase.

To provide diagnostic information for the perceived quality in a conversational situation, each of the phases have been analysed in detail. More precisely, for each phase quality relevant perceptual dimensions have been identified with (I) Pairwise similarity and a following MDS or with (II) a Sematic Differential and a following PCA

Applying both methods in separate experiments in the three phases of a conversation, seven perceptual dimensions were identified and proposed (Köster & Möller 2014) (Wältermann, Raake & Möller 2010). An overview can be seen in Table 1. The table shows, that the seven perceptual dimensions split in the three phases of the a conversational situation: the Listening Phase is composed of four dimensions: *Noisiness, Discontinuity, Coloration* and *Loudness*; the Speaking Phase is composed of two dimensions: *Impact of one's own voice on speaking* and *Degradation of one's own voice*; the Interaction Phase is composed of only one dimension: *Interactivity*.

The proposed dimensions however, have been identified in separate listening, speaking, and conversation tests. It has not been validated and analysed, though, if the proposed dimensions can be identified and extracted altogether in a real conversational situation, consisting of all phases. Furthermore, it has to be evaluated how the perceptual dimensions behave with respect to dominance and occurrence if only one specific phase or a combination of multiple phases is degraded. Additionally, the interpretation of the proposed perceptual dimensions has to be investigated and analysed further. Especially the interpretation of the dimensions for the Speaking and the Interaction Phases are first approaches only and a more meaningful interpretation might be found in a comprehensive experiment.

Conversational Phase	**Perceptual dimension**	**Description**	**Possible Source**
Listening Phase	Noisiness	Background noise, circuit noise, coding noise	Coding
	Discontinuity	Isolated and non-stationary distortions	Packet loss

	Coloration	Frequency response distortions	Bandwidth limitations
	Loudness	Important for the overall quality and intelligibility	Attenuation
Speaking Phase	Impact of one's own voice	How is the back coupling of one's own voice perceived	Sidetone and echo
	Degradation of one's own voice	How is the back coupling of one's own voice degraded	Sidetone and echo
Interaction Phase	Interactivity	Delayed and disrupted interaction	Delay

Table 1: Overview of the seven identified and proposed perceptual quality dimensions for a conversational situation

For this aim, a sophisticated conversational experiment was conducted in which each of the conversational phases, as well as all of the proposed perceptual dimensions are addressed. In the following section the general procedure and structure of this experiment is presented.

3. Conversational experiment

To verify the proposed perceptual dimensions, we decided to conduct a SD experiment with a consecutive PCA. For this an experiment was conducted that addresses a regular conversation situation and additionally all phases of a conversational situation separately. The approach is based on the hypothesis that the results of the complete conversational experiment lead to similar results as the separate speaking, listening and interaction experiments. In the following the design and the setup of the experiment are explained in detail.

3.1. Test design

The conversational experiment was carried out by 40 participants (23 female, 17 male) grouped into 20 pairs, aged between 18 and 53 years. Since all of the possible phases of a conversation should be addressed the experimental task consisted of 3 sections:

I. In the first section, the task of the two participants was to conduct a Short Conversation Test (SCT) according to (P.805 2007). SCTs were used because their tasks represent everyday-life situations and provide a reasonable degree of interaction while being limited to an acceptable test duration. After each SCT, the participants had to judge the 28 APs representing all phases of a conversation.

II. The second section addresses the Listening and Speaking Phases. One of the participants is asked to read out 2 sentences while the other participant listens to what is read out. The sentences and procedures of the speaking part were similar to (Appel & Beerends 2002) and (Köster & Möller 2014). The listening part was analog to (Wältermann, Raake & Möller 2010). After the first sequence, the participants change roles so that each participant has to speak and listen. For each sequence, the participants were asked to rate 12 APs for the Speaking Phase and 14 APs for the Listening Phase.

III. The third section addresses the Interaction Phase. This task is supposed to be sensitive for possible delays in the transmission system. Therefore, so called random number verification tasks (RNVT) were used (P.805 2007). Accordingly, the participants are asked to alternately verify a set of numbers. The approach of the task is, that already a small amount of delay becomes perceivable due to the quick verification procedure. The participants were asked to rate 10 APs representing the Interaction Phase.

For all three sections, the participants were asked to communicate using a transmission system that was distorted by 11 different degradations (see Table 3) which were similar to previously conducted tests (Köster & Möller 2014) (Wältermann, Raake & Möller 2010). Each pair of participants first conducted one introduction session to get familiar with the test, and afterwards 11 sessions for each

degradation consisting of all 3 sections. The order of degradations was randomized between participants. Table 3 describes the experimental procedure and structure.

Test Section	Task P1	Task P2	Rating P1 (APs)	Rating P2 (APs)
I	SCT	SCT	28	28
II	Listening	Speaking	14	12
	Speaking	Listening	12	14
III	RNVT	RNVT	10	10

Table 2: Overview of the experimental procedure. (I) Conversation, (II) Listening and Speaking, (III) Interaction. P - Participant, APs - antonym-pairs, SCT - Short Conversation Scenario, RNVT - random number verification tasks.

Condition	Degradation	Condition	Degradation
1	Clean	7	Attenuation 15dB
2	Sidetone -5dB	8	Codec LPC-10
3	Dealy 1000ms	9	Noise+Echo
4	Echo 100ms	10	Codec LPC-10 +Sidetone
5	Packet loss 10%		
6	White noise 30db	11	Delay+Packet loss

Table 3: Conditions

3.2. Technical Setup

The test system (TheTelephone (Guse et al. 2015)) is based on Pure Data (PD), a graphical programming language for signal processing. It allows manipulating audio effects in real-time and thus enables to simulate acoustical degradations like echo, as well as non-stationary degradations. Additionally, the system was extended with multiple speech codecs including G.711 or LPC-10, using open-source implementations. The codec components also introduce effects like packet-loss or jitter on request. The system provides a mouth-to-ear delay of 70ms as baseline. The sound signal was presented via a Beyer Dynamic DT770 stereo headset. The participants were located in two sound-insulated test rooms which met the ITU-T requirements.

3.3. Determination of attributes for the SD

As described before, in the SD experiment a predefined set of attributes (APs) was given to the test participants. In the test the same APs as in the previous separate listening, speaking and, interaction tests were used. For the section I all APs were used. For section II and III the corresponding APs for each phase of a conversation have been rated. Table 6 shows all used APs (translated from German).

4. Results

The results of the conducted experiment are structured in four groups: first we analyse the results of the third section (Interaction Phase), second and third the results of the second section (Listening Phase as well as Speaking Phase), finally the results of the first section (Conversation Test) of the SD experiment.

4.1. Section III - Interaction Phase

The ratings of the 10 APs in the SD experiment result in a multidimensional space and are compared in terms of a correlations matrix of the SD attributes. This matrix shows, that the

addressed 10 attributes highly correlate with each other (average $r \approx 0.9$). The results of the following PCA indicate, that the 10 attributes can be described by one dimension, covering 85,429% of the variance of the 10 one-dimensional features. This result is similar to the one of the previously conducted separate interaction experiment and shows that the proposed dimension works as intended.

4.2. Section II - Listening Phase

The number of assigned dimensions as the result of the PCA for the Listening Phase in Section II can be seen in the Scree Plot in Figure 1a. Here only three dimensions are determined, covering 96,987% of the variance of the 14 APs. In separate LOTs, however, four dimensions were proposed. An explanation for this can be found by analysing the factor loadings for each feature to the determined three dimensions in Table 4. Dim 3 describes the dimension Loudness ('loud - quiet' (0,972)) and Dim 2 describes the dimension Noisiness (hissing (0,831), noisy (0,862), and crackling (0,866)), whereas Dim 1 seems to cover both dimensions Coloration and Discontinuity, correlating with the remaining 10 APs. Discontinuity has only been triggered by two conditions with 10% packet loss. Additionally one of the two conditions is combined with a low quality LPC-10 codec triggering Coloration. This leads to the result, that one dimension covers the APs for Discontinuity and Coloration. Thus, we think that the reduction of the dimensionality of the Listening Phase space from 4 (found in the previous experiments) to 3 (found in our experiment) is due to the limited number of conditions which could trigger these perceptual dimensions.

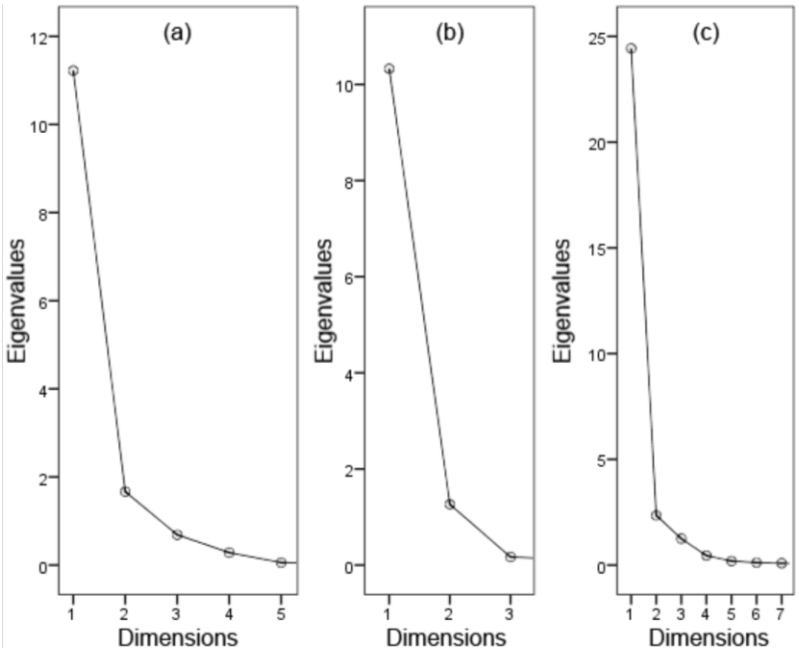

Figure 1: Scree Plots for the PCA; a - Listening Phase II; b - Speaking Phase II; c - Conversation Test I.

AP	Dim 1	Dim 2	Dim 3
interrupted - continuous	.939	.297	.095
distant - close	.876	.255	-.392
crackling - not crackling	.300	.866	.086
not noisy - noisy	.424	.862	.064
muffled - not muffled	.913	.368	-.112

shaky - steady	.866	.481	.072
indirect - direct	.904	.403	-.041
dark - bright	.928	.313	-.180
unintelligible - intelligible	.929	.349	-.048
not hissing - hissing	.518	.831	.103
clear - unclear	.863	.495	-.047
distorted - undistorted	.884	.449	.082
thin - thick	.832	.412	-.334
loud - quiet	-.100	.179	.972

Table 4: Factor loadings of the PCA results - VARIMAX rotated; Listening Phase.

4.3. Section II - Speaking Phase

The number of assigned dimensions as the result of the PCA for the Speaking Phase in Section II can be seen in the Scree Plot in Figure 1b. Two dimensions are determined covering 96,569% of the variance of the 11 one-dimensional features. Two dimensions have also been discovered in the separate speaking test, then termed *Impact of one's own voice on speaking* (covering APs like helpful, irritating, exhausting, distracting or fluent) and *Degradation of one's own voice* (covering APs like reverberant, clear, thin and distorted). Looking at the factor loadings for the Speaking Phase (see Table 5), it can be seen, that Dim 1 covers the same features as in the previous tests. Dim 2 explicitly only covers the AP thin, and with lower values clear (0,401) and distorted (0,280). These two features are also covered by Dim1. Additionally, AP reverberant, intended for Dim 2, is only respected by Dim 1. We explain this result with condition 9, where the echo is mixed with noise. In the perception of the participants, the noise seems to mask the echo degradation. Thus, only condition 4 covers pure reverberation, which potentially led to the presented outcome. We think that the limited coverage of the 2 dimensions (this experiment) in comparison to the interpretation of the two proposed dimensions (previous experiment) is again due to the number of conditions triggering the dimensions.

AP	Dim 1	Dim 2
exhausting - not exhausting	.991	.090
requires concentration – requires no concentration	.990	.048
distracting - not distracting	.987	.076
not fluent - fluent	.983	.125
loud - quiet	.932	-.256
not helpful - helpful	.990	-.028
distorted - undistorted	.942	.280
unclear - clear	.869	.401
reverberant - anechoic	.971	-.029
irritating - not irritating	.985	.052
thin - thick	.026	.994

Table 5: Factor loadings of the PCA results - VARIMAX rotated; Speaking Phase.

4.4. Section I – Conversation Test

The number of potential dimensions as the result of the PCA for the Conversation Test in Section I can be seen in the Scree Plot in Figure 1c. Three dimensions are determined covering 96,637% of the variance of the 28 one-dimensional AP space. It was intended that the results of the PCA show that all seven dimensions are perceived in the Conversation Test. However, the results show that it seems as that only a limited number of dimensions can be perceived in a test paradigm like the SCTs that require the full attention of the test participants on the flow of the conversation, and not on the rating task. The factor loadings (Table 6) point out, that only the proposed dimensions Noisiness is distinct enough to be perceived separately in Dim 3 (hissing (0,913), noisy (0,901), crackling (0,901)). The other two assigned dimensions Dim 1 and Dim 2 represent a mix of the remaining 6 dimensions of the individual phases. Dim 1 represents the proposed Dimensions Coloration (muffled (0,738), 'dark-bright' (0,821), direct (0,827), clear (0,717), distant (0,892)) and Discontinuity (interrupted (0,760), shaky (0,746), distorted (0,720)) and could be related to the intelligibility. Dim 2 describes the cognitive load of the participant representing the dimensions Loudness (loud- quiet (0,975)) and the Impact of one's own voice on speaking (helpful (0,665), reverberant (0,826), distracting (0,760)). The remaining two dimensions Interactivity and Degradation of one's own voice are fused in Dim 1 and Dim 2. As mentioned before, we assume that this result is due to the limited cognitive resources test participants could dedicate to the detailed rating task, as these resources were bound by the conversation task of the STC. However, we argue that the results of the sections II and III of the experiment show that the seven proposed dimension are still valid for a proper diagnosis of the quality of transmitted speech in a conversational situation.

AP	Dim 1	Dim 2	Dim 3
indirect - direct	.827	.495	.173
exhausting - not exhausting	.662	.664	.325
unpleasant - pleasant	.621	.708	.327
distracting - not distracting	.533	.760	.366
irritating - not irritating	.564	.767	.291
not hissing - hissing	.292	.244	.913
thin - thick	.827	-.040	.472
distorted - undistorted	.720	.536	.390
clear - confusing	.619	.706	.335
unintelligible - intelligible	.792	.419	.425
lazy - agile	.772	.548	.289
shaky - steady	.746	.441	.494
easy - hard	.651	.669	.343
not noisy - noisy	.296	.234	.901
relaxing - annoying	.566	.685	.427
loud - quiet	-.056	.975	.084
requires concentration – requires no concentration	.645	.712	.251
interrupted - continuous	.760	.387	.376
crackling - not crackling	.239	.057	.901
clear - unclear	.717	.445	.529
not helpful - helpful	.632	.665	.372
dark - bright	.821	.406	.308

not frustrating - frustrating	.649	.688	.313
not fluent - fluent	.736	.550	.347
distant - close	.892	.354	.177
effective - ineffective	.710	.617	.328
reverberant - anechoic	.532	.826	.067
muffled - not muffled	.738	.509	.380

Table 6: Factor loadings of the PCA results - VARIMAX rotated; Conversation Test, Section I.

5. Conclusion and Outlook

In earlier studies, seven perceptual quality dimensions for different phases of a conversational situation have been identified. However, these dimensions have only been analysed in separate studies for each phase. Therefore, a global conversation test addressing all three phases of a conversation and potentially triggering all seven dimensions was conducted. The experiment was divided into three sections I, II and III. While section II and III are addressing the three phases and its underlying dimensions, section I was supposed to simulate a conversation approaching all phases and dimensions in a realistic way, and with the test participants' attention on the conversation task.

The first key-finding of the experiment is that the proposed dimensions are difficult to identify in a realistic conversation situation, where the attention of the test participants is rather on the content of the conversation, and on the dialogue flow. It seems that too many cognitive resources are bound by this task, thus reducing the number of separately perceivable dimensions in this phase. Thus, it is important to establish a **test paradigm** that specifically allows the participants to perceive each phase separately, in addition to a natural conversation paradigm.

The second key-finding of the conducted experiment is that the results of section II Listening Phase and section I show that the two dimensions **Coloration and Discontinuity seem to merge**. We explain this finding with the peculiarities of the conducted experiment. One condition mixed up degradations triggering both dimensions, and the size of the experiment did not allow for more than one additional condition for each dimension. This finding has to be investigated in follow-up studies. More precisely, when designing test conditions care should be taken that each expected perceptual dimension is separately covered by a sufficient number of technical conditions.

In future experiments, it would be interesting to identify weights of the individual perceptual dimensions for the overall quality rating. We expect that the weighting of the dimensions will depend on the conversation task, and on the conversation structure induced by this task. For example, in a highly interactive setting emphasis might be given to the dimensions in the Speaking and Interaction Phase, whereas in less interactive settings the perceptual dimensions of the Listening Phase might dominate.

In sum, the contribution shows a test paradigm for a diagnosis of conversational quality should cover both - phases of realistic task-driven conversation structures as well as phases where the Listening, Speaking and Interacting Phases can be analysed separately, without putting too much cognitive load on the test participants. Otherwise, perceptual dimensions which might be important for overall quality may remain unidentified. The far end goal is to implement the demanded subjective test paradigm, to verify it and to give a base for the quality estimation in a conversational situation.

6. References

Appel, R & Beerends, JG 2002, 'On the Quality of Hearing One's Own Voice', *Journal of the Audio Engineering Society*, vol 50, no. 4, pp. 237-248.

Guéguin, M, Le Bouquin-Jeannès, R, Gautier-Turbin, V, Faucon, G & Barriac, V 2008, *On the Evaluation of the Conversational Speech Quality in Telecommunications*, EURASIP J.Adv. Signal Process.

Guse, D, Haase, F, Köster, F, Schiffner, F, Skowronek, J, Raake, A & Möller, S 2015, 'TheTelephone: A software-based flexible speech telephony system for conversational user studies', *TridentCOM*, submitted.

Köster, F & Möller, S 2014, 'Analyzing perceptual dimensions of conversational speech', *INTERSPEECH*, Singapore, Singapore.

Köster, F, Möller, S, Antons, J-N, Arndt, S, Guse, D & Weiss, B 2014, 'Methods for assessing the quality of transmitted speeh and of speech communication services', *Acoustics Australia* , vol 42, no. 3, pp. 179 - 184.

ITU-T Rec. P.59, 1993, *Artificial Conversational Speech*, International Telecommunication Union, Geneva.

ITU Rec. P.805, 2007, *Subjective Evaluation of Conversational Quality*, International Telecommunication Union, Geneva.

Wältermann, M, Raake, A & Möller, S 2010, 'Quality Dimensions of Narrowband and Wideband Speech Transmission'.

_____