Introducing a new Test-Method for Diagnostic Speech Quality Assessment in a Conversational Situation

Friedemann Köster, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin, Deutschland,
Email: friedemann.koester@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

Traditional methods for assessing the quality of transmitted speech are usually limited to specific restrictions. On the one hand, the methods only ask test participants to judge how the overall quality is perceived, and on the other hand users are placed in simulated situations, like listening- or speaking-only, that do not reflect reality. To overcome this limitations the speech quality in a conversational situation was analyzed by dividing a conversation into three separate phases and identifying seven corresponding quality-relevant perceptual dimensions. The extracted dimensions can be combined for the overall quality assessment and may separately be used to diagnose the technical reasons of quality degradation. In this article, we present a new method for directly assessing the identified quality dimensions in a conversational situation. The benefit of this new analytic assessment method is the reduced experimental effort due to the reduced number of necessary scales and thus the number of judgments per condition. Additionally, the method enables to analyze conversational speech quality for diagnosis and optimization.

Introduction

In traditional and modern packed-based (Voice-over-IP) networks the transmitted speech can be affected, and also impaired, during recording, coding, transmission, decoding, and playback. The network and terminal devices responsible for this, also called *quality elements* [1], are codec, bandwidth limitation (narrowband (300 - 3400 Hz) and wideband (50 - 7000 Hz)), linear and non-linear filters, delay packet loss, echo and noise [2]. Therefore, it is of high importance for telecommunication providers to evaluate how system end-users perceive and experience possible degradations, allowing to improve their services. Traditionally, this is done by assessing the quality of transmitted speech over telecommunication services, the so-called *Quality of Experience* [3]. For this, passive listening-only subjective experiments with naïve participants in a laboratory context are conducted. Participants are asked to judge the perceived overall quality on five-point *Absolute Category Rating* (ACR) scales yielding in the *Mean Opinion Score* (MOS) representing the average quality rating [4, 5].

However, the traditional methods only gather information about the overall quality and do not give insights into reasons for possible low quality ratings - no diagnostic information are provided. Furthermore, the methods are limited to specific simulated listening situations that do not represent the reality. To overcome both limitation the hereafter approach has been followed:
In [6] a conversational situation is 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 leads to a separation of a conversation into three phases as perceived by one participant [7]: 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 analyzed in detail. More precisely, for each phase quality relevant perceptual dimensions have been identified. Perceptual dimensions are defined as orthogonal and thus independent features of the multidimensional space formed by a perceptual event provoked by an acoustical speech wave inside a listener [8]. Perceptual dimensions are directly connected to aforementioned quality elements and thus provide diagnostic information. Two methodologies are used for the identification of the perceptual dimensions: (I) *Pairwise Similarity* (PS) with a *Multidimensional Scaling* (MDS) [9]; or (II) *Semantic Differential* SD [10] with a *Principal Component Analysis* (PCA).

Applying both methods in separate experiments in the three phases of a conversation, seven perceptual dimensions were identified, proposed and validated [11, 12, 13]. An overview can be seen in Table 1 . The table shows the seven perceptual dimensions and its connected quality elements 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*.

Apart from the fact that the SD and MDS methodologies are necessary to extract the perceptual dimensions, they inherent one major drawback: Due to the relatively large number of attributes (SD) and pairwise comparisons (MDS) both methods are time-consuming. Thus, the number of conditions to be assessed is limited due to the enormous experimental effort. In this article we present a **new methodology that allows to directly quantify the proposed seven dimensions**. This enables to increase the number of conditions to be assessed. Additionally the method gives the possibility to actually create conversational databases with recorded conversations and subjective overall quality and dimension rat-

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 on speaking	How is the backcoupling of one's own voice perceived	Sidetone and echo
	Degradation of one's own voice	How is the backcoupling 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.

ings. Moreover, having the databases available, diagnostic instrumental conversational speech quality estimation models based on perceptual dimensions can be trained.

First we present the rating scales used for the new test-method. Second we will explain the actual test procedure in detail. The article closes with a conclusion and an outlook on future work

Dimension Rating Scales

The new subjective assessment method provides a means for quantifying the seven quality relevant perceptual dimensions in a conversational situation (noisiness, discontinuity, coloration, loudness, impact of one's own voice on speaking, degradation of one's own voice, and interactivity) directly by means of seven descriptive scales. Thus, each scale is dedicated to one particular dimension. The extremities of each scale are labeled with the antonym-pairs describing the corresponding dimension. This enables to directly quantify separate scores for each perceptual dimension present in a conversational situation.
Figure 1 shows the graphical scale layout. The continuous scales were chosen over traditional ACR scales because they showed to be more sensitive [14]. While the labels on the left of the scales describe no impairment in the relating dimension, the labels on the right describe the maximum impairment. Thus the scales are considered to be unipolar.

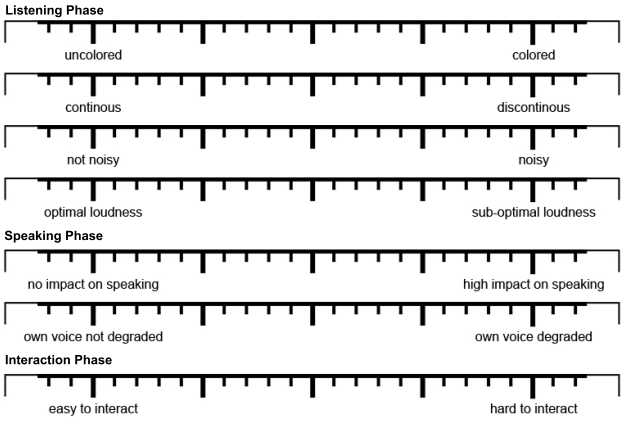

Figure 1: Dimension scale design.

Test Procedure

The new test-method is supposed to provide diagnostic information for a conversational situation. Therefore, the method follows common paradigms for subjective conversational tests as described in [15]. For each condition, or transmission system properties under test, two participants in tow separate rooms according to [4] are required. The basic test set-up can be seen in Figure 2.

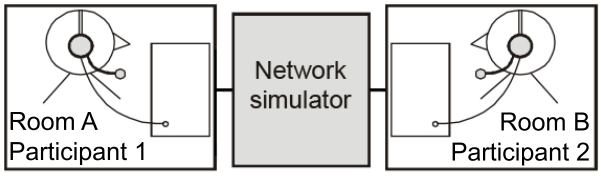

Figure 2: Test-method set-up.

In [13] it was shown that with general conversation scenarios like the *Short Conversation Test* (SCT) or the the *Random Number Verification Task* (RNVT) alone participants are not capable of identifying all of the seven perceptual dimensions. It seemed that too many cognitive resources are bound by this tasks due to the fact that the attention of the test participants is rather on the content of the conversation, and on the dialogue flow. Thus, it is important to establish a test-method that specifically allows the participants to perceive each phase separately, in addition to a natural conversation paradigm. Therefore, the new test method to assess one condition is composed of three sections:

(I) In the first section, the task of the two participants is to conduct a SCT scenario according to [15]. The 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. Thus, this section represents a regular everyday-life conversational scenario of about 2-4 minutes length. After each SCT, the participants are asked to judge the overall quality (according to [4]), and then the seven perceptual dimensions 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 two sentences while the other participant listens to what is read out. The sentences and procedures of the speaking part are similar to [16] and [12]. The listening part is analog to [11]. After the first sequence, the participants change roles so that each participant has to speak and listen. For each sequence, the participants are asked to judge the overall quality of the speaking as well as the two dimensions for the *Speaking Phase* and the overall quality of the listening as well as the four dimensions 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, RNVTs are used [15]. The participants are asked to judge the overall quality of the interaction and the *Interactivity* representing the *Interaction Phase*.

An overview of the test procedure for both participants can be seen in Figure 4.

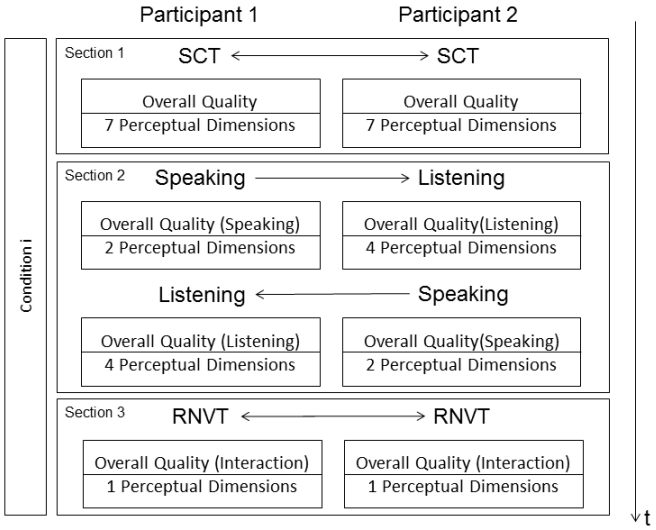

Figure 4: Overview of the test procedure. SCT - Short Conversation Scenario, RNVT - Random Number Verification Task.

Dimension Rating Schema

The dimension rating schema of the new test-method is comparable with the schema for the pure listening phase [8] or for analyzing noisy signals [17]. Each of the three separate sections of the new test-method includes an assignment (speaking, listening, SCT, RNVT) as well as a overall quality an a dimension assessment task. As these assessment tasks are analogue we will representative explain the rating task for Section I in detail.
After the overall quality assessment according to [4] the dimension scales (see Figure 1) are presented separately and consecutively. This is to reduce the bias due the presentation order [18]. Before the participants are asked for they ratings, they are asked to conduct the given task once. Afterwards, the participants first give their judgments on the overall quality and second on the seven perceptual dimensions. The detailed rating schema for Section I can be seen in Figure 3.
The conditions to be assessed are presented in randomized order. Additionally, the order of the dimension scales is permuted for each participant. The schema can be seen in Table 2. For each participant the order of the scales is held constant to avoid confusion of the scales.

Instructions and Training

A detailed written description of the test-method is given to the participants to ensure an equal level of knowledge. The instructions fist gives an overview over the scales and how they should be used. It is explained, that in the experiment the *characteristics* of a conversation are supposed to be judged and that this judgment is done seven scales. Each scales is labeled with an attribute at each end that describes the characteristic to be judged. The scales are described in detail using the high correlated attributes according to the SD experiment conducted to extract the perceptual dimensions. Second the test procedure is explained. The introduction introduced each section and its relating assignments and task to the participant.
In the training the two participants run through one test sequence as described in Figure 4. This is done to ensure that the participants get to know the test procedure as well as get used to the usage of the scales and the test method.

Test Duration

Section	Task	Duration [s]
Section 1	SCT	180
	Rating	70
Section 2	Speaking	30
	Rating	20
	Listening	30
	Rating	40
Section 3	RNVT	70
	Rating	10
Overall duration		450

Table 3: Average test duration.

Table 3 shows the estimations for the duration of one sequence of the new test-method. The average 450 seconds (or 7.5 minutes) can vary depending on the delay the system uses resulting in longer or shorter durations (duration one Sequence S = 7.5).
The training and the introduction together take up to 30 minutes until the procedure and the scales are understood (training and introduction T = 30).

Assuming the experimenter plans to test 15 different telephone system network settings (conditions C = 15). The total duration of the experiment using the new test method would than be 142.5 minutes (compare Equation 1).

$$C \cdot S + T = 15 \cdot 7.5 + 30 = 142.5 \tag{1}$$

To avoid participants fatigue the experiment should than be divided into two 70 minutes sessions.

Conclusions

In this article we present a new test-method for diagnosing speech quality in a conversational situation. The test-method is based on the perceptual dimensions identified in three phases of a conversation. For each dimension a separate rating scale with describing antonym-pairs is introduced. The test procedure and the dimension rating schema as well as the training and introduction point out the advantages of the new method. It is now possible to reduce the experimental effort and thus analyze more telephone system network settings in shorter time.

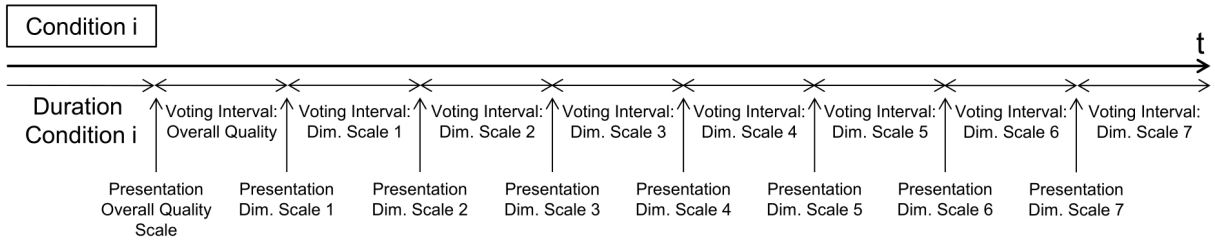

Figure 3: Condition, scale presentation, and rating for section 1.

Participant	Dim Scale 1	Dim Scale 2	Dim Scale 3	Dim Scale 4	Dim Scale 5	Dim Scale 6	Dim Scale 7
1	dis	col	noi	lou	ios	dos	int
2	col	noi	lou	ios	dos	int	dis
3	noi	lou	ios	dos	int	dis	col
4	lou	ios	dos	int	dis	col	noi
5	ios	dos	int	dis	col	noi	lou
6	dos	int	dis	col	noi	lou	ios
7	int	dis	col	noi	lou	ios	dos
...	...	...	...	...	...	...	...

Table 2: Presentation order of the dimensions scales. noi - Noisiness, dis - Discontinuity, col - Coloration, lou - Loudness, ios - Impact of one's own voice on speaking, dos - Degradation of one's own voice, and int - Interactivity.

In addition, the new methods serves as a baseline for creating conversational databases that allow to create and optimize conversational speech quality estimators.

References

[1] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*, Springer Science & Business Media, Berlin, 2005.

[2] A. Raake, *Speech Quality of VoIP Assessment and Prediction*, John Wiley & Sons, Chichister, West Sussex, 2006.

[3] Qualinet, "Qualinet White Paper on Definitions of Quality of Experience," 2013, (Version 1.2, eds. P. Le Callet, S. Möller, A. Perkins), European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland.

[4] ITU-T Recommandation P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva, 1996.

[5] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung*, Teubner Verlag, 1998.

[6] D.L. Richards, *Telecommuncation by Speech: The Transmission Performance of Telephone Networks*, Butterworths, London, UK, 1973.

[7] M. Guéguin and et al., *On the Evaluation of the Conversational Speech Quality in Telecommunications*, EURASIP J.Adv. Signal Process, 2008.

[8] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*, Springer, Berlin, 2012.

[9] I. Borg and P. Groenen, *Modern Multidimensional Scaling - Theory and Applications*, Springer Series in Statistics, New York, NY, 2005.

[10] C. Osgood, *The Measurement of Meaning*, University of Illinois Press, Urbana, IL, 1957.

[11] M. Wältermann, A. Raake, and S. Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission," Acta Acustica united with Acustica, 2010, pp. 1090–1103.

[12] F. Köster and S. Möller, "Analyzing Perceptual Dimensions of Conversational Speech Quality," in *Proc. 15th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2014)*, Singapore, Singapore, 2014, pp. 2041–2045, ISCA Interspeech 2014 Prceedings.

[13] F. Köster and S. Möller, "Perceptual Speech Quality Dimensions in a Conversational Situation," in *Proc. 16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015)*, Dresden, Germany, 2015, ISCA Interspeech 2015 Prceedings.

[14] F. Köster, D. Guse, M. Wältermann, and S. Möller, "Comparison Between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech," in *Fortschritte der Akustik, DAGA 2015: Plenarvortr. u. Fachbeitr. d. 41. Dtsch. Jahrestg. f. Akust.* 2015, DEGA.

[15] ITU-T Recommendation P.805, *Subjective Evaluation of Conversational Quality*, International Telecommunication Union, Geneva, 2007.

[16] R. Appel and J.G. Beerends, "On the quality of hearing one's own voice," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, 2002.

[17] ITU-T Recommendation P.835, *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Supression Algorithm*, International Telecommunication Union, Geneva, 2003.

[18] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," *J. Audio Eng. Soc*, vol. 56, no. 6, pp. 427–451, 2008.