Acta Acustica united with Acustica
Vol. 103 (2017) 506 – 522

DOI 10.3813/AAA.919079

Identifying Speech Quality Dimensions in a Telephone Conversation

Friedemann Köster, Dennis Guse, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin, Germany. friedemann.koester@tu-berlin.de

Summary
Speech telecommunication services are traditionally used for communication between two interlocutors interacting in a conversation. Thus, the quality of transmitted speech in a conversational situation, as perceived by the end-users, is the important indicator for service providers to evaluate their systems. In this context, it is not enough to only provide information about the overall quality but also to indicate reasons and sources for quality losses. In this article, we present an approach towards analyzing speech quality in a conversational situation by dividing a conversation into three separate phases and identifying corresponding quality-relevant perceptual dimensions, as perceived by the system users. The identified dimensions can be combined for the overall quality assessment and may separately be used to diagnose the technical reasons of quality degradations. For this, four separate subjective experiments to uncover the underlying dimensions in each conversational phase are conducted. The resulting quality-profile, consisting of seven perceptual dimensions, is then validated in an extensive conversational experiment triggering all three phases of a conversation using a new proposed test-paradigm. This allows deeply analyzing conversational speech quality for diagnosis and optimization of telecommunication systems and provides the fundamentals for instrumental diagnostic conversational speech quality measures.

PACS no. 43.71.An, 43.71.Gv

1. Introduction

Vocal human-to-human communication is the main purpose for using speech telephony services. Technological development within traditional and modern packed-based (Voice-over-IP) telephony networks can affect – and possibly also impair – the transmitted speech signal. The network and terminal device elements which are responsible for this (referred to as *quality elements* [1]), are codecs, bandwith limitation (narrowband (300-3400 Hz) and wideband (50-7000 Hz)), linear and non-linear filters, delay, packet loss, echo, and noise [2].

It is therefore of high priority for telecommunication providers to find out how end-users perceive and experience degradations. For this, assessing the quality of transmitted speech over telecommunication systems allows the providers to improve their services and counter possible issues. In this context, the quality of transmitted speech is also referred to the so-called *Quality of Experience* (QoE) that "is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state" [3].

Received 12 April 2016,
accepted 2 March 2017.

In telephony services passive subjective experiments with human participants in a laboratory context are common means to study and understand QoE (so-called *listening-only test*s, LOTs). In these experiments, overall (or integral) quality ratings on five-point *Absolute Category Rating* (ACR) scales are gathered [4]. The experiments yield a *Mean Opinion Score* (MOS) [5], representing the average integral quality rating of an average person.

Since subjective experiments are time and money consuming, the demand of telecommunication service providers for instrumental models to predict the overall quality of transmitted speech, as gathered in LOTs, raised. Research led to the development of multiple types and approaches (parametric and signal-based) for instrumental models [6]. Nevertheless, as described in [7], the aforementioned LOTs and the instrumental models hold two main limitations:

- **Integral quality**: Only the integral quality is taken into account, reasons for underlying sub-optimum quality are not uncovered.
- **Non-interactive settings**: The methods refer to the passive listening situation, but conversational and interactive aspects are not considered.

The first limitation (integral quality) points out, that two dissimilar speech samples impaired by different degradations, for example one by a bandwidth limitation and

one by background noise, can be rated with the same low MOS value. Having only the MOS value at hand, system providers cannot identify the reason for a possible quality loss, and therefore do not know how to improve their services. In LOTs, experimenters can of course directly ask for specific degradations, but in that case they have to be certain about the presence of these degradations beforehand. Thus, traditional methods do not provide diagnostic information. To counter this problem, new subjective [8] as well as new instrumental [9] diagnostic methods have been developed. They identify and assess quality-relevant *perceptual dimensions* to obtain diagnostic information.

The definition and the underlying idea of perceptual quality dimensions is the following: The output of a transmission system, a speech signal possibly degraded by the aforementioned quality elements, is perceived by the system user as a composition of explicit *features*, that are orthogonal (and thus independent) and represent recognizable and nameable characteristics of the speech sound [1, 2]. These features are *perceptual dimensions* in a *multidimensional perceptual space*. When the user judges quality, s/he makes use of these perceptual dimensions to determine a perceptual difference to an optimum, degradation-free situation. Overall quality can thus be determined on the basis of perceptual features. In turn, the features allow identifying reasons for quality losses. For example, two speech samples showing the same integral quality rating may exhibit different perceptual dimension judgments that are connected to specific quality elements.

In [10] Wältermann identified four perceptual quality dimensions for narrowband and wideband speech transmission in a listening-only situation. The benefit of quality dimensions is also pointed out by recent development within the *International Telecommunication Union* (ITU). The currently started work item *Perceptual Approaches for Multi-Dimensional Analysis* (P.AMD) [11] is aiming at developing a signal-based quality predictor that provides diagnostic information on the basis of the perceptual quality dimensions identified by Wältermann.

The second limitation (non-interactive settings) reveals that the aforementioned traditional methods only consider the unrealistic passive listening-only situation. Quality elements that affect the interaction or the speaking (for example echo or delay) cannot be determined in LOTs. To fill this gap, conversational tests [5, 12] and speaking test [13] have been designed.

Feasible solutions to both limitations have only been developed separately. This leads to the trade-off for an experimenter to either extract diagnostic information or to address different conversational phases in an experiment. This article presents one approach to address this trade-off by formulating and answering the following question:

What are the quality-relevant perceptual dimensions that an interactive conversational situation is composed of?

To answer this question, we follow the approach of combining the advantages of both solutions. More specifically, we identify quality-relevant perceptual dimensions in each conversational phase, namely in the *Listening*, the *Speaking*, and the *Interaction Phase*. Thus, this article has four main contributions:

- **Multidimensional analysis of a conversation**: The results of four experiments yielding the perceptual dimensions in the *Speaking* and the *Interaction Phase*.
- **A new quality-profile for conversational speech quality**: Together with the work conducted by Wältermann [10] the multidimensional analysis reveals seven perceptual dimensions underlying the conversational speech quality.
- **A new conversational test-paradigm**: For the direct quantization of the perceptual dimensions and for the validation of the quality-profile a new subjective conversational test-paradigm that separately addresses each phase of a conversation is established.
- **Validation of the proposed quality-profile**: Together with the new test-paradigm the quality-profile is validated in a final conversational experiment.

The new quality-profile and the work presented allows to assess and diagnose conversational speech quality in future work. In addition, it is the direct follow up of the studies conducted by Wältermann [10] and serves as a fundamental framework for developing diagnostic instrumental models to predict the quality of transmitted speech in a conversational situation as demanded in the current ITU-T work item *Objective Conversational Voice Quality Assessment Model* (P.CQO) [14].

The rest of the paper is organized as follows. In Section 2, a review of speech quality in a conversational situation is given. The perceptual dimensions are identified by auditory experiments with following multidimensional analyses. Section 3 gives an understanding of the paradigms used for the experiments. The actual experiments conducted to uncover the underlying perceptual dimensions in a conversation and their results are presented in Section 4. The identified dimensions are validated in a separate extensive conversation experiment using the new test-paradigm. The results and a discussion are illustrated in Section 5. Conclusions are drawn and an outlook towards future work is given in Section 6.

2. Speech Quality in a Conversational Situation

To provide information about a complete conversational situation (see limitation two (non-interactive setting) in section 1), a typical conversation has to be investigated with respect to all possible occurring situations. In a conversation, the interlocutors alternately adopt the roles of listener and talker which introduces interaction between the participants. In [15] and [16] a conversational process is described as a four-state model: while having a conversation, the participants either listen to what is said (01) or speak (10) while exchanging information. Additionally the participants can also both speak (11) or remain silent (00) at the same time.

Table I. Overview of the so far identified perceptual quality dimensions in a conversational situation (see [10]).

Conversational Phase	Perceptual Dim.	Description	Possible Source
Listening Phase	Noisiness	Background noise, circuit noise, coding noise	Coding, background noise
	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	*Unknown*	-	Sidetone or echo
Interaction Phase	*Unknown*	-	Delay

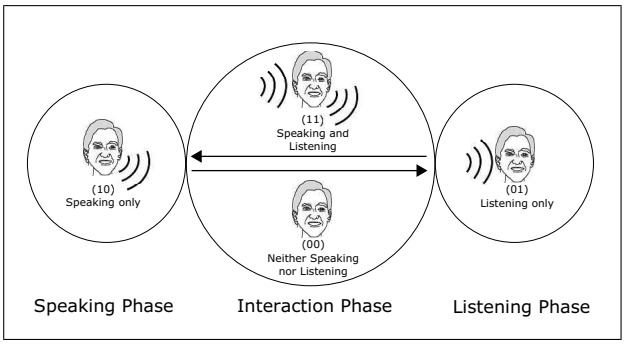

Figure 1. The three phases of a conversational process, as perceived by one participant [17].

According to [17], this leads to three phases of a conversation: the *Speaking Phase* (10), the *Listening Phase* (01), and the *Interaction Phase* describing the alternation of the states (10) and (01). The frequency of changes describes the degree of interaction and as a side-effect the states (00) and (11) can occur. The three phases, as perceived by one participant, are illustrated in a state diagram in Figure 1.

Thus, from a speech-quality point-of-view, a conversation is affected by the quality elements encountered in the *Listening Phase* (codecs or filters), in the *Speaking Phase* (echo or sidetone), and those affecting the interactivity of the conversation in the *Interaction Phase* (delay, double talk and mutual silence impaired by signal processing in the devices or the network) [15, 17]. In the following, the three phases also describe the possible user's situations during a conversation. To obtain diagnostic information on conversational speech quality (addressing the described trade-off in Section 1), the three phases will be analyzed in detail in the following subsections.

2.1. Listening Phase

For the *Listening Phase*, subjective and instrumental methods to assess the listening quality are standardized and recommended by the ITU [5, 18]. To obtain diagnostic information (see the first limitation (integral quality) in Section 1) in the *Listening Phase*, the approach of identifying perceptual dimensions related to impairments, such as degraded colouration or noisiness, is used.

As mentioned in Section 1, perceptual dimensions are features of the multidimensional space formed by a perceptual event inside a listener. Typically, two methodologies that are described in Section 3 are used for the identification of perceptual dimensions. Using both methodologies, Wältermann uncovered four perceptual dimensions for the *Listening Phase* [10]: *colouration*, *noisiness*, *discontinuity*, and sub-optimal *loudness*. Each of these perceptual dimensions can directly be connected to the aforementioned quality elements (see Table I). Also, Wältermann showed that it is possible to directly quantify the identified perceptual dimensions in a subjective test [8].

The proposed subjective method is similar to what is recommended for noisy speech signals in [19] and is the orgin of the current ITU-T work item P.AMD.

Other proposals, e.g by Sen, who used the Diagnostic Acceptability Measure (DAM [20]), identified four to seven dimensions, which are subdimensions (and thus not orthogonal) of the dimensions identified by Wältermann [21].

In sum, the multidimensional space inside a listener in the *Listening Phase* is composed of four perceptual dimensions (*colouration*, *noisiness*, *discontinuity* and sub-optimal *loudness*) that lead to the development of diagnostic subjective and instrumental methods.

2.2. Speaking Phase

Technically, the *Speaking Phase* is usually distorted by degradations due to talker-echoes and sidetones (see [22] and [2]). Thus, separate *Talking and Listening Tests* [13] are conducted to assess the quality in the presence of those distortions. In these subjective tests, participants are asked to speak into a transmission system and rate afterwards, how the system affects one's own speaking [12]. But, simultaneously speaking and listening can cause considerable fatigue to test participants. Therefore, so called *3rd-Party-Listening-Tests* have been developed. In these tests the spoken and the back coupled of the own voice of a participant is recorded and afterwards both are rated by a third person [13]. However, these methods only determine an integral quality value, without diagnostic information.

2.3. Interaction Phase

The *Interaction Phase* covers not only the change from state (01) to (10) and the change from state (10) to (01), but also the states (00) and (11) (see Figure 1). Increased amounts of the states (00) and (11) technically occur especially due to transmission delay [23], which is particularly noticeable by a shift of the usual rhythm of conversation, leading to passive interruptions (occurs when a speaker

becomes interrupted by the delayed arrival of a counterpart's utterance) and active interruptions (occurs when an speaker starts to speak, while he still hears his counterpart talking) [24].

For the subjective assessment of the interaction quality, conversational tests have to be conducted that, unlike for the *Listening* or *Speaking Phase*, require two participants. To simulate a natural conversation, the participants usually follow particular scenarios. For example, the ITU-T recommends so called *short-conversations test*s (SCTs) in which the participants are asked to solve tasks in role-plays (ordering a plane-ticket or pizza), or so-called interactive tests in which the participants are asked to align numbers or addresses as fast as possible [25]. One of these interactive tests is the so-called *Random Number Verification Test* (RNVT) in which the participants are asked to alternately compare a predefined series of numbers (each participant has one series at hand while one number is different) [2].

Following these guidelines, the participants are generally asked to rate the overall quality, or the interruption effort. To deeper analyze the interactivity, values like the *Speaker Alternation Rate* or the *Conversational Temperature* have been introduced [26]. However, these values focus more on the alternation and turn-taking of the speakers, and less on the diagnoses or the perceptual space of the interlocutors.

2.4. Summary

The approach of the presented work is to combine the advantages of considering all possible user situations in a conversation, and of diagnosing the quality of transmitted speech on the basis of perceptual dimensions. Table I gives an overview of the currently known perceptual dimensions in a conversational situation. As it can be seen, except for the *Listening Phase* no perceptual dimensions have so far been identified. This leads to the formulation of the already stated research question in Section 1, what perceptual dimensions an interactive conversational situation is composed of.

To answer this question, perceptual dimensions in the *Speaking* and the *Interaction Phase* have to be uncovered. The identification of the perceptual dimensions and the underlying experiments are presented in the Section 3 and 4.

3. Experimental paradigms to uncover perceptual dimensions in a telephone conversation

For each of the two remaining phases of a conversation (*Speaking* and *Interaction Phase*) two experiments with two different experimental paradigms were conducted. Both paradigms follow different approaches to transform data into a low-dimensional space with particular advantages and drawbacks. In the field of audio research, the methods described have been part in numerous studies, see for example [27], [28], or [29], to name just a few. Because of that, we decided to use the same approaches for our studies.

Section 3.1 describes the method of *Multidimensional Scaling* (MDS) of dissimilarity or preference ratings gathered in a pairwise comparison experiment. The method of analyzing attribute ratings of a *Semantic Differential* (SD) experiment with a *Principal Component Analysis* (PCA) is introduced in section 3.2.

Using and comparing both methods leads a) to a more distinct interpretation of the resulting dimensions and b) helps to verify the validity of the results. Thus, the two paradigms in combination provide a solid statement about the actual nature of the underlying dimensions for the phase under investigation.

3.1. Multidimensional Scaling

In general, MDS is used as a multivariate technique and is mainly applied to find the number of dimensions required to represent perceptual attributes of stimulus objects in a low-dimensional multidimensional space [30].

The approach is to gather the dissimilarity between two pairwise presented stimuli. This results in a dissimilarity matrix for each participant. The MDS maps the (average) dissimilarities into distances. It is assumed, and it has been verified, that the psychological dissimilarities correspond to Euclidean distances (higher dissimilarities, higher distances) [8, 30, 31].

In the context of the presented work, we are interested in the quality of perceptual events, happening either during speaking or during interaction. Thus, our stimuli are obtained in an active or interactive instead of a passive situation, and instead of asking the participants for a dissimilarity rating, we gathered preferences values. The two different approaches of gathering dissimilarities and preferences have been analyzed and compared in different studies and experiments and revealed a high degree of correlation [32, 27].

Since we are not interested in individual preferences but in group tendencies, we are looking for a multidimensional solution for an average person, and the preference ratings are averaged over the individuals resulting into a single preference matrix.

However, the gathered preference data cannot be used in a classic MDS that uses dissimilarity data. Therefore, a so called non-metric MDS, also called ordinal MDS, is applied [33]. While a classic MDS is metric, that is, the model represent various properties of the data related to algebraic operations, non-metric MDS represent only the ordinal properties of the data [30].

The preference matrix serves as input for the non-metric MDS where the mapping is restricted to be a monotone function. ALSCAL is employed as a method for computing the non-metric MDS [34].

Following [30], to determine the resulting dimensionality, both, statistical fit parameters and the ability to interpret the resulting dimensions are considered. One important statistical fit parameter is the so-called *Stress*. It is actually a badness-of-fit parameter specifying how bad the resulting distances match with the given data. A reasonable dimensionality is found if the Stress value does

not decrease significantly with increasing the number of dimensions. Looking at a Scree plot (see for example Figure 6), ideally a sharp "elbow" marks the adequate dimensionality [30].

Using the MDS paradigm provides the advantage that the task for participants is practicable. No complex instructions are required and comparing two pairwise presented stimuli is straight forward. But, the interpretation of the resulting dimensionality and of the resulting dimensions is only possible on the basis of the known difference between the stimuli used. This may lead to intuitive and speculative interpretations. To express a valid interpretation it should be considered to compare the results of a MDS with other methods for minimizing dimensionality.

3.2. Semantic Differential

In a SD experiment, a previously determined set of attributes is given to the participants in terms of bipolar scales. The extremities of each scale are labeled with a pair of opposite attributes, so called *antonym-pairs* (APs) (for example *loud* vs. *quiet*), each describing a one-dimensional feature. The intensity of each feature within a given condition has to be judged by the test participants.

Using the Principal Component Analysis (PCA) on the average ratings of the participants, only the components with eigenvalues above one are kept. The columns of the resulting matrix are the principal components (PCs) and correspond to the coordinates of the points representing the APs in the dimension-reduced space. Finally, the result is transformed into a rotation matrix satisfying the VARIMAX criterion [35]. The rotation causes that correlating scales are summarized by one axis, which leads to a more simple structure. Detailed information about the SD and the PCA can be found in [8] or [36], for example.

Compared to the MDS paradigm, the interpretation of the resulting dimensions is supposed to be easier because it is assumed that each dimension is represented by a cluster of APs giving the experimenter direct hints on which aspects are covered. Nevertheless, to get a valid interpretation of the dimensions it is recommended to conduct both, a MDS and a SD experiment. The disadvantage of the SD paradigm is that significant effort has to be conducted to determine the APs beforehand (see Section 4.1.3 and Section 4.2.3).

4. Uncovering perceptual dimensions in the *Speaking* and *Interaction Phase*

As described in Section 2, the *Listening Phase* has been the subject of research towards understanding and uncovering the perceptual space of a listener. In this section we present studies that have been conducted to investigate the perceptual space of participants in a conversational task. To do this, a conversation is split into three phases according to Section 2, and experiments analogue to the *Listening Phase* (following the paradigms presented in Section 3) are conducted for the *Speaking* and the *Interaction Phase.*

Table II. Conditions and Overall Quality Results for the SD Experiment in the *Speaking Phase*. α: Attenuation [dB], β: Roundtrip Delay [ms], σ: Standard Deviation.

Condition	Values		Overall	
	α	β	Rating	σ
1	−25	0	1.6	0.75
2	−20	0	1.9	0.82
3	−15	0	2.0	0.94
4	−10	0	2.4	0.82
5	0	0	4.4	0.53
6	10	0	3.2	0.81
7	15	0	3.5	0.78
8	20	0	4.1	0.71
9	0	50	2.3	0.72
10	0	100	2.5	0.61
11	0	150	2.1	0.33
12	0	200	2.3	0.64
13	−10	100	1.9	0.71
14	−20	100	1.8	0.58
15	−20	150	1.5	0.47
16	−10	300	1.5	0.46

Part of the work illustrated in this section is based on the data presented in a former publication [37].

4.1. Speaking Phase

To uncover the perceptual dimensions of the *Speaking Phase* both methodologies (MDS and SD) are applied. Since the speaking is usually impaired by sidetone and talker-echo (see Section 2.2), for both experiments a passive speaking-only test with these two degradations was carried out with the goal to investigate how hearing one's own voice while speaking influences the speaking, and how the participant perceives their own voice.

4.1.1. Technical setup

The test system for the two tests conducted for the *Speaking Phase* is implemented with the help of the graphical programming language tool for modeling and simulating dynamic systems [38]. The system was developed to simulate sidetone and talker-echo. For the sidetone distortion the direct back coupling of the spoken voice with different levels of attenuation is used. For the talker-echo the delayed back-coupled and attenuated spoken voice with varying delay values is used. The conditions used can be seen in Tables II and IV. The direct back coupling had a delay of < 10 ms and is recorded as 0 ms delay. The attenuation level is simulated in association to the input speech level. Note that some conditions simulate degradations with strong characteristics to guarantee that all naïve participants perceive the effects of sidetone and echo. The conditions were presented in randomized order. An EDIROL USB AudioCapture UA-25EX soundcard was used, together with a Sennheiser HMD 46 ATC 300 Headset. The back coupling was presented diotic. The participants were set in a test room which meets the requirements according to [5].

4.1.2. Test design

For both test-paradigms (SD and MDS) basically the same task had to be conducted by the participants. For each presented condition or comparison the participants were asked to read out aloud a text that appeared to them on the test screen. Each piece of text consisted of two to three sentences, and all together 27 randomly presented text-pieces were used. One text-piece could for example look like this (translated from German):

> "Can you please give me the best connection between Munich and Duisburg. I have to arrive on Saturday at 12.30pm latest."

To avoid the participants pay too much attention on reading the text, they were asked to learn the text by rereading it at minimum three times. Thus, it was ensured that the participants could speak the text as freely as possible, simulating a real *Speaking Phase*.

4.1.3. SD Experiment

As mentioned in Section 3.2, in an SD experiment a predefined set of attributes (APs) is given to the test participants in terms of bipolar scales. In order to find proper attributes, two pre-tests were conducted.

In a first test, participants were asked to freely use the degraded test setup. Their task was to gather as many descriptions of the degraded test setup as possible. In sum, a list of 25 APs were collected by 3 experts. Experts were chosen because it was assumed that they can describe the system adequately. However, since they are very experienced with telecommunication degradations, they might also be biased. Thus, in the second test, 10 naïve participants, according to the definition in [39], were asked to use the degraded test setup and select 5 of the 25 APs they think describe the system best.

Based on the overall frequency of selection, a set of 11 APs were finally selected: *exhausting - not exhausting; requires concentration - requires no concentration; distracting - not distracting; not fluent - fluent; loud - quiet; not helpful - helpful; thin - thick; distorted - undistorted; unclear - clear; reverberant - anechoic; irritating - not irritating*.

The actual test was carried out by 16 naïve participants (4 female, 12 male) aged between 21 and 36 years (mean age 26.4). For each condition (see Table II) the participants were asked to fulfill the task described in Section 4.1.2. After each task for each condition, the participants are asked for their subjective rating of the overall quality (MOS) for a validity check, and of the APs introduced before.

The scale shown in Figure 2 was used for the overall quality ratings (taken from [12]). This scale is, in particular, useful because it avoids scale-end effects and is more sensitive in comparison to the classical ACR scale [40]. A similar scale (see Figure 3) with only two labels was used for the AP ratings. We used a within-subject-test-design.

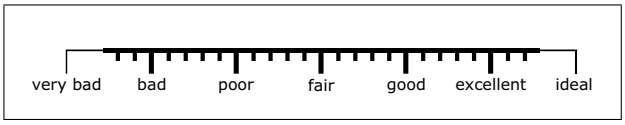

Figure 2. Overall quality rating scale (taken from [12]).

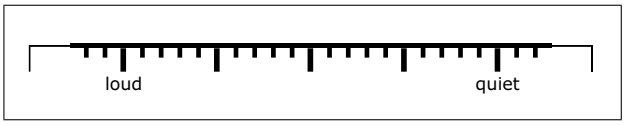

Figure 3. AP rating scale used for the SD experiments.

4.1.4. Results of the SD Experiment

The results of the conducted SD experiment for the *Speaking Phase* are structured in two groups: first, we analyze the results of the overall quality, second, the results of the PCA on the AP ratings stemming from SD experiment are presented.

The results of the overall quality ratings are presented in Table II. The ratings are similar to the studies made in [41]. The standard deviations lie within the range of standard deviations as typically also obtained in standard ACR experiments [12]. Additionally, a repeated measure *ANalysis Of VAriance* (ANOVA) [42] between the conditions and the overall quality ratings as depended variables was carried out. The results show that the conditions have a significant impact on the overall quality judgments of the test subjects ($F(4.04, 60.64) = 33.86, p < .01$). With this data it is proofed that the different degradation levels worked as intended (falling quality - lower rating / rising quality - higher ratings).

To analyze the results of the PCA, first, the number of resulting perceptual dimensions has to be identified. As described in Section 3.2, the number of the dimensions is found by keeping only components with Eigenvalues above one. To visualize the results a *Scree Plot* can be seen in Figure 4.

The figure shows that two components have eigenvalues above one, resulting in two dimensions. The determined two dimensions cover 95.3 % of the variance of the eleven APs. Table III shows the factor loadings for each of the eleven features to the determined two dimensions.

It can be seen, that the first dimension (Dim 1) covers seven of the eleven features with loadings above 0.8. These seven features (concentration, loud, fluent, distracting, exhausting, irritating, helpful) describe how the hearing of the own voice is perceived by the speaker and what impact or effect hearing the own voice could trigger inside the listener. More precisely, the results for the first dimension show that hearing one's own voice can, for example, be very irritating and can handicap the fluency of the speaking.

The second dimension (Dim 2) covers three features (distorted, clear, reverberant) with loadings above 0.7 and two features (helpful, thick) with loadings slightly below 0.4. The dimension seems to be descriptive in terms of representing the degree of degradation and impairment of the

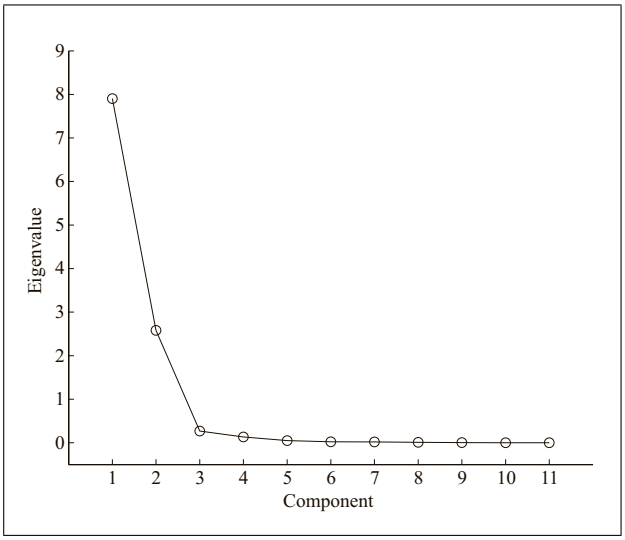

Figure 4. Scree Plot for the PCA on the SD experiment in the *Speaking Phase*.

own voice the speaker perceives hearing one's own voice. In other words, the resulting dimensions describes possible frequency distortions of the sidetone and the echo path. This is mostly determined by results of the loadings for the features "distorted", "clear", "reverberant", and "thick".

The low number of features and the inconsistent loading values show that the second dimension seems to be weak. However, a final reflection (see Section 4.1.7) of the resulting dimensions is only possible when having also the results from the MDS experiment at hand.

4.1.5. MDS Experiment

As mentioned in Section 3.1, in a MDS experiment the preferences of two pairwise presented stimuli is judged by the participants. Having N conditions this leads to $N(N-1)$ comparisons. Assuming that the preference between stimulus A and stimulus B is the same as the preference between stimulus B and stimulus A, this leads to $(N(N-1))/2$ comparisons [43]. Using the 16 conditions of the SD experiment this would lead to 120 comparisons.

For a feasible experiment conducted in approximately one hour this would take too long. Therefore, only 9 randomized conditions (see Table IV) were used for the test leading to 36 comparisons. Condition eight and one are alike and serve as anchor-conditions for a validity check.

To create the complete distance matrix for the ordinal MDS, one half of the participants judged the preference between stimulus A and stimulus B and the other half the preference between stimulus B and stimulus A. For each comparison, the participants were asked to speak the textpiece (see Section 4.1.2) once for condition A and once for condition B. They could redo the comparison as often as desired.

Afterwards, the participants had to judge whether they prefer stimulus A over stimulus B (and vice-versa) on the scale presented in Figure 5. The conditions were presented in randomized order and the MDS experiment was carried out by 22 naïve participants (14 female, 8 male) aged between 18 and 36 years (mean age 25.9) (different from the SD experiment).

Table III. Factor loadings (> 0.3) of the PCA on the SD experiment in the *Speaking Phase* - VARIMAX rotated (Dim - Dimension).

Antonym-pair	Dim 1	Dim 2
exhausting - not exhausting	.993	
concentration - no concentration	.991	
distracting - not distracting	.980	
not fluent - fluent	.991	
loud - quiet	.988	
not helpful - helpful	.893	.378
distorted - undistorted		.939
unclear - clear	-.612	.761
reverberant - anechoic	.351	.875
irritating - not irritating	.937	
thin - thick	-.874	.378

Table IV. Conditions for the MDS Experiment in the *Speaking Phase*. α: Attenuation [dB], β: Roundtrip Delay [ms].

Condition	α	β
1 (S0)	0	0
2 (E50)	0	50
3 (Sminus25)	25	0
4 (S20)	−20	0
5 (E250)	0	250
6 (Sminus10E150)	10	150
7 (S10E150)	−10	150
8 (S02)	0	0
9 (Sminus10)	10	0

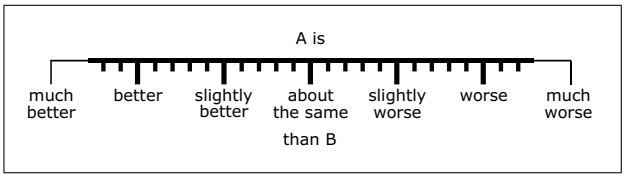

Figure 5. Preference comparison rating scale used in the MDS experiments.

4.1.6. Results of the MDS Experiment

The adequate dimensionality is found if the badness-of-fit parameter *Stress* does not decrease significantly with a further increase of the number of dimensions. To visualize the results a Scree Plot can be seen in Figure 6.

The figure shows that the sharp "elbow" is located at the second dimension, thus, two dimensions are extracted for the MDS experiment. This result is similar to the result of the SD experiment.

To analyze and compare the dimensions the resulting space of the MDS (see Figure 7) has to be inspected. Looking at the two anchor-conditions (S0 and S02) the resulting space of the MDS shows that these two conditions are positioned with a short distance, indicating, that the different quality levels worked as intended.

Dimension one shows that from left to right the conditions start with strong characteristics (strong echo or loud sidetone - S10E150, E250, S20) and end with rather weaker characteristics (quiet sidetone, e.g., Sminus10, Sminus25). The anchor-conditions are located in the middle of the scale. A strong echo or a loud sidetone results in a high impact on the speaking abilities of the speaker. In turn, a quiet sidetone does not have a impact on the speaking. A popular example for this effect is when a speaker is confronted with a loud background noise. In this case, the speaker automatically raises the voice to mask the noise. This effect is called the *Lombard Effect* [44, 45]. The same effect, but in the opposite direction, can be observed when a speaker is confronted with a loud copy of his or her own voice over a headset, a loud sidetone. In this case, the speaker automatically lowers the voice [46]. In [41], the term *self-listening comfort* is introduced to describe this influence. These introduced effects of the used conditions are reflected in the results for the first identified dimension. Looking again at this result, the scale of dimension one (from right-low, strong echo or loud sidetone, to left-high, weak echo or quiet sidetone) describes the impact on the speaker of hearing one's own voice while speaking.

For dimension two the scale starts with the anchor-condition S0 and then covers stepwise the conditions with stronger degradations (the higher, the stronger the degradation). In general, if the sidetone is delayed, the speaker starts to feel uncomfortable. For delays below 30 ms (considered as sidetone) and high levels, the direct signal and the delayed version will be interfered at the speakers ears which leads to a *comb-filtered* version of the signal [47]. The user will perceive this as a colouration in the sound of his or her own voice [41]. If the delay exceeds 30 ms (considered as echo) and the sound level is high, the speaker will experiences difficulties in talking. This is expressed in a slower speaking in terms of the speaking rate and pauses between words [48]. On the other hand, if the level is low, even high delayed echo hardly gives any degradation. Thus, a back coupled and delayed version of the own voice is perceived as a coloured and thus degraded version of the own voice by the speaker. Transferring this to the results of the MDS experiment, the identified dimension shows that stronger degradations lead to a more degraded perception of the own voice than weaker degradations. Hence, the scale of dimension two (from bottom-low to top-high) thus seems to describe the degree of degradation of the own voice the speaker perceives hearing one's own voice.

4.1.7. Conclusion

The results of the SD (see Section 4.1.4) and the MDS (see Section 4.1.6) experiment reveal a high degree of similarity.

In the SD experiment, the first resulting dimension covers APs that describe the impact of the own heard voice on the speaker while speaking. The same properties can be seen in the results of the MDS experiment where the first dimension describes from low to high the characteristics

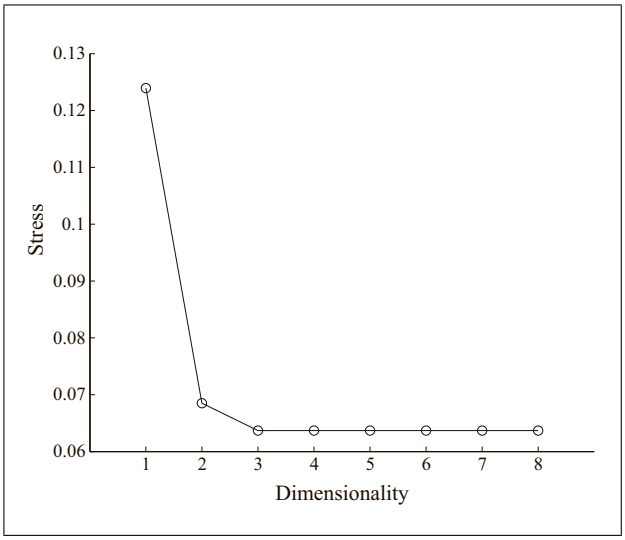

Figure 6. Scree Plot for the MDS on the comparison judgments in the *Speaking Phase*.

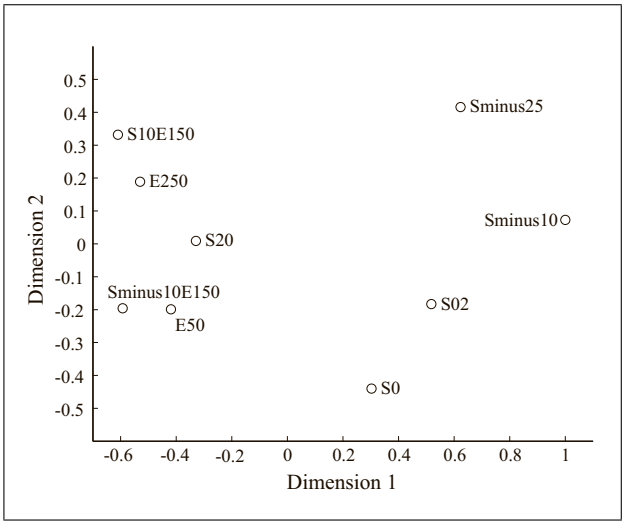

Figure 7. Results of the MDS experiment in the *Speaking Phase*.

(weak to strong echo/sidetone) of the conditions. In both cases the resulting dimensions seem to represent the impact of the degraded transmission system on the speaker while speaking.

The second resulting dimension in the SD experiment covers attributes that describe the amount of degradation of the conditions ("distorted - undistorted", "unclear - clear", "reverberant - anechoic"). In the MDS experiment the second identified dimension is also describing the same effects starting with the reference conditions ending with highly degraded conditions (strong echo/sidetone). Following from this, in both experiments the two identified dimensions seem to portray the degradation of one's own voice perceived by the speaker.

In sum, the result of the multidimensional analysis in terms of two subjective tests identified two perceptual dimensions. It was mentioned that a loud sidetone might decrease the voice of a speaker and that a back coupled and delayed version of one's own voice is perceived as

a colouration in the sound of the own voice by the user. These two effects match the two dimensions identified in the multidimensional analysis. One dimensions describes the impact on the speaker a back coupling might have (for example decreasing the voice) and the other dimensions describes the degraded perception of the own voice (for example a coloured sound).

However, it has to be mentioned again that the two identified dimensions might be dependent from each other in terms of their presence. While a degradation of one's own voice is only perceived when the own voice has also an impact on the speaking, a back coupling of the own voice might only have an impact on the speaking without perceiving a degradation of the own voice. Until now, this is just an assumption and has to be verified in an additional experiment.

Following from these results we like to propose to call the two perceptual dimensions of the *Speaking Phase* a) the ***impact of one's own voice on speaking*** (scaled from "no impact on speaking" (−1) to "high impact on speaking" (1)) and b) the ***degradation of one's own voice*** (scaled from "own voice not degraded" (−1) to "own voice degraded" (1)).

4.2. Interaction Phase

To uncover the perceptual dimensions of the *Interaction Phase*, again both methodologies (MDS and SD) are applied. Especially interactive experiments are sensitive for the quality element delay (see Section 2.3) that impairs the interaction of two interlocutors. So, for both experiments a conversation test was carried out to investigate how the user's interaction in a call is affected by varying amounts of transmission delay.

4.2.1. Technical setup

For the experiments a test system based on *Pure Data* (PD [49]), a graphical programming language for signal processing, was used. 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 on request, and were used in the validation experiment of Section 5.

The sound signal was presented via a *Beyer Dynamic DT770* stereo headset. In both setups the participants were located in two sound-insulated test rooms which met the requirements according to [5].

4.2.2. Test design

For the conversational tasks, SCTs (see Section 2.3) were used and modified by updating dates and currencies. The SCTs were selected because their tasks represent everyday-life situations and provide a reasonable degree of interaction while being limited to acceptable test duration.

In both experiments, each pair of participants first conducted one introduction SCT scenario to get familiar with the test design. In the SD experiment the participants were asked to give their rating on the APs for each condition and each SCT.

In the MDS experiment only one of the two participants was able to switch between two conditions. The one participant was asked to rate the comparison of two conditions with regard to the interaction between both interlocutors.

4.2.3. SD Experiments

Again, to conduct the SD experiment a predefined set of APs has to be found. To find suitable attributes, two pretests were conducted (similar to the SD experiment of the *Speaking Phase*).

In the first test, as many descriptions as possible were collected by 6 experts, resulting in a list of 42 different antonyms. In the second test, 15 naïve participants were asked to select 5 of the 42 attributes they think describe the system best. Based on the overall frequency of selection, a set of 10 antonym-pairs were finally selected: *not exhausting - exhausting; easy - hard; unpleasant - pleasant; not frustrating - frustrating; effective - ineffective; does not require concentration - requires concentration; lazy - agile; clear - confusing; relaxing - annoying; distracting - not distracting*.

The actual experiment was carried out by 32 naïve participants (8 female, 24 male) aged between 19 and 31 years (mean age 25.2) paired in 16 groups of two interlocutors. The test-system was distorted by eight randomized different values of one-way transmission delay (0, 300, 600, 900, 1300, 1700, 2100, and 2500 ms) resulting in eight conditions.

For each condition the participants were asked to play through one SCT scenario, rate the overall quality for a validity check, and then score on the APs introduced before.

Again, the same scales as in the SD experiment for the *Speaking Phase* were used (compare Figure 2 and Figure 3). We used a within-subject-test-design.

4.2.4. Results of the SD Experiments

The results of the conducted SD experiment are structured in two groups: first we analyze the results of the overall quality as a validity check, then the results of the SD experiment.

After averaging the ratings of the overall interaction quality over the conditions, a repeated measure ANOVA between the conditions and the overall quality ratings as depended variables was carried out. The result shows that the amount of delay has a significant impact on the judgment of the test subjects ($F(4.93, 152.75) = 17.19, p < .01$). This data indicates that the different degradation levels worked as intended (low delay - high overall quality / high delay - low overall quality).

The judgments show that the addressed 10 attributes highly correlate with each other (average r≈0.9). The results of the following PCA indicate, that the 10 features can be described by one dimension, covering 96.12 % of

the variances of the 10 one-dimensional features. The resulting factor loadings for each of the 10 features can be seen in Table V.

The outcome shows that all features are covered by one dimension with high loadings above 0.9. Regarding the ten features the resulting dimension seems to describe the convenience or the challenge of interacting. But, a final interpretation (see Section 4.2.7) of the dimension is again only possible after analyzing the MDS experiment.

4.2.5. MDS Experiments

In the case of the *Interaction Phase* the task in the MDS experiment is to judge the preference of two pairwise presented amounts of transmission delay. The eight conditions used in the SD experiment would lead to 28 comparisons and thus SCTs. Again, this would be too much for a feasible experiment. Therefore, only five randomized conditions (0, 500, 1000, 1500, and 2000 ms) were used leading to 10 comparisons.

As done for the MDS experiment in the *Speaking Phase*, one half of the participants judged the preference between condition ***A*** and condition ***B*** and the other half the preference between condition ***B*** and condition ***A*** to create the complete distance matrix for the ordinal MDS. As an exception for this experiment, only one of the two participants was asked to judge whether they prefer condition ***A*** over ***B***, the other participant acted as dummy. This procedure was followed because only one of the participants was able to change the condition and thus was able to judge their preference. The rating was again done on the scale shown in Figure 5.

The conditions were presented in randomized order and the MDS experiment was carried out by 52 naïve participants grouped in 26 pairs. Thus, the results are based on the ratings of 26 participants (10 female, 16 male) aged between 20 and 32 years (mean age 24.6) (different from the SD experiment).

4.2.6. Results of the MDS Experiment

The MDS reveals a Stress below 0.5 showing that the resulting space is one-dimensional. The space can be seen in Figure 8.

Looking at the figure, it can be seen that the resulting dimension starts with the highest delay (2000 ms) and then covers stepwise the conditions with lower delay until reaching the lowest value (0 ms).

In literature, it is described that a transmission delay may lead to three effects [50]. First, the delay leads to an *interruption*. Interruptions are distinguished between *active* and *passive* interruptions. Active interruptions occur when one interlocutor starts to speak, while he or she still hears the other interlocutor speaking. Passive interruptions occur when one interlocutor gets interrupted by the delayed arrival of a statement of the other interlocutor. Second, due to the transmission delay, the perception of a conversation, in terms of structure and pattern, may considerably be different from one interlocutor to the other, while both are participating in the same conversation. Third, if the test subjects perceive an unnatural rhythm of the conversational flow, they adapt their behavior.

Table V. Factor loadings (> 0.3) of the PCA on the SD experiment in the *Interaction Phase* - VARIMAX rotated (Dim - Dimension).

Antonym-pair	Dim 1
distracting - not distracting	.971
exhausting - not exhausting	.988
concentration - no concentration	.979
unpleasant - pleasant	.981
clear - confusing	.960
lazy - agile	.995
easy - hard	.993
relaxing - annoying	.979
not frustrating - frustrating	.982
effective - ineffective	.977

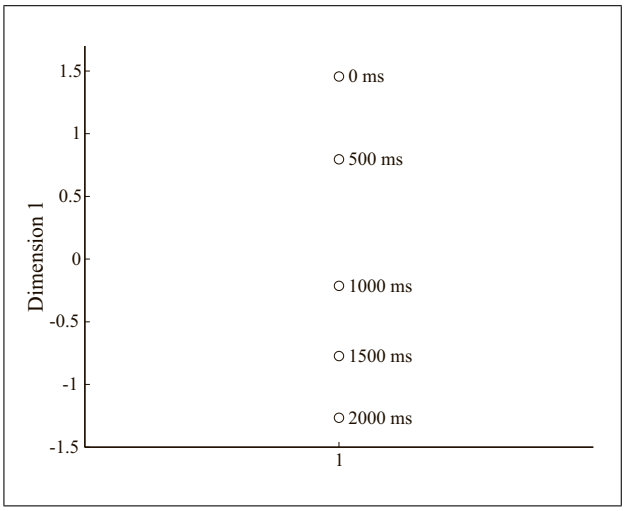

Figure 8. Results of the MDS experiment in the *Interaction Phase*.

The result of the MDS experiment thus seem to merge these three effects into one dimension. The resulting scale of the dimension (from bottom-high to top-low) seems to describe the effort or difficulty to interact with the interlocutor as described in [50].

4.2.7. Conclusion

Again, the results of the SD (see Section 4.2.4) and the MDS (see Section 4.2.6) experiment reveal a high degree of similarity.

In the SD experiment, the resulting dimension covers APs that describe the convenience or the difficulty of interacting. The same characteristics can be seen in the results of the MDS experiment where the resulting dimension describes from low to high the effort or difficulty to interact (high to no delay). Thus, in both cases the resulting dimension seems to represent the degree of facility/difficulty to interact.

It was mentioned earlier that a transmission delay may lead to passive and active interruptions that shift the natural interactive rhythm in a conversation. These interrup-

Table VI. Overview of the seven identified and proposed perceptual quality dimensions for a conversational situation.

Conversational Phase	Perceptual Dimension	Description	Possible Source
Listening Phase	Noisiness	Background noise, circuit noise, coding noise	Coding, circuit or background noise
	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	Frequency distortions of the sidetone and echo path
Interaction Phase	Interactivity	Delayed and disrupted interaction	Delay

tions also lead to a different perception (in terms of the two interlocutors) of the conversational structure. In addition, too high amounts of delay are related to a rising user dissatisfaction [51]. The results of the two conducted multidimensional analysis combine these findings of the user perception as the identified dimension seems to cover the effects of a delayed speech transmission. The resulting dimension can be described with used APs (see Table V) and the characteristics of the dimension is depended on the amount of transmission delay.

Following from these results we would like to propose to call the resulting perceptual dimensions of the *Interaction Phase* the ***interactivity*** (scaled from "easy to interact" (−1) to "hard to interact" (1)).

4.3. Resulting Quality Dimensions in a Conversational Situation

In memory of the aforementioned research question (see Section 2.4) and the two limitations (see Section 1), we now have a set of seven proposed dimensions for a entire conversation.

While the *Listening Phase* was already part of different studies and revealed four perceptual dimensions, two additional perceptual dimensions for the *Speaking Phase* and one perceptual dimension for the *Interaction Phase* were identified. An overview of the perceptual quality spaces resulting from the multidimensional analysis can be seen in Table VI. The seven perceptual dimensions are proposed to be called:

- Coloration,
- Noisiness,
- Discontinuity,
- Loudness,
- Impact of one's own voice on speaking
- Degradation of one's own voice,
- Interactivity.

The two identified dimensions for the *Speaking Phase*, the *impact of one's own voice on speaking* and *degradation of one's own voice* seem to cover the space spanned by the degradations sidetone and echo. However, also other degradations (e.g. loud background noise) might not only affect the *Listening Phase*, but also the *Speaking phase*.

For the *Interaction Phase*, the perceptual dimension *interactivity* was identified. We see mainly two explanations for this result: firstly, we identified the perceptual dimension with the help of an SD experiment that is based on prior determination of antonyms. In our case we conducted two pre-tests with naïve participants and with experts, separately. However, the high correlation of the attributes suggests that the attributes only cover a certain limited space. This is due to the fact that the stimuli that we presented varied only with respect to delay. This brings us to our second explanation: the only quality element we varied was the delay. We did not consider quality elements of the *Listening Phase* or the *Speaking Phase*, which might have provoked other dimensions.

So far, the three phases were treated mostly independent. It is not known and has to be analyzed if the results of the multidimensional analysis for the *Speaking Phase* and *Interaction Phase* would be different when quality elements of all phases are considered in one single tests. In particular, it has to be verified if the separately identified dimensions can still be uncovered in a real conversational situation. Also, it is not known yet how the presences of multiple degradations affect the characteristics of the seven perceptual dimensions. For example, in [17] or [52] it was investigated that the conversational quality is rated more critically for echo than for transmission delay. If this could be adapted for the identified dimensions is not known yet. For this, additional studies to investigate and identify the conversational quality profile are necessary.

The multidimensional analysis revealed the perceptual quality spaces for each phase of a conversation that in sum is composed of seven perceptual dimensions. This set of perpetual dimensions allows diagnosing conversational speech quality in future work. However, this set of perceptual dimensions still has to be validated and their characteristics in a conversational test (and not in separate SOTs or LOTs) have to be investigated. For this, at first a new subjective test-paradigm that allows considering all three conversational phases and their perceptual dimensions has to be developed. Using the developed test-paradigm then enables to verify the proposed perceptual spaces. The pro-

posed paradigm and the subsequent study is presented in the next section.

5. Validation Experiment

To verify the new set of quality dimensions, we created a new test-paradigm that separately addresses each phase of a conversation as well as a short structured conversation scenario. The proposed test-paradigm is presented in Section 5.1. The approach of the validation experiment is based on the hypothesis that the resulting dimensions of the separate conducted listening, speaking, and interaction experiments can also be identified using the new paradigm. We decided to conduct an additional SD experiment (see Section 3.2) to analyze the identification of the dimensions, however, in future the paradigm will be used to directly quantify the seven dimensions. In the following the paradigm and the results of the experiment are explained in detail. Parts of the work presented in this section is based on the data presented in the publication [53].

5.1. Test design and new test-paradigm

Since all of the possible phases of a conversation should be addressed the new test-paradigm consists of 3 sections:
(I) In the first section, the task of the two participants is to conduct a SCT. This section represents a regular everyday-life conversational scenario of about two to four minutes length. After each SCT, the participants first have to judge the overall quality and second the 28 APs representing (and used in) all phases of a conversation.
(II) The second section addresses the *Listening* and *Speaking Phases*. One of the participants is asked to read out a text while the other participant listens to what is read out. The sentences and procedures of the speaking part are similar to the previous conducted studies in Section 4.1. The listening part is analog to [10]. 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 rate the 11 APs for the *Speaking Phase* and 14 APs for the *Listening Phase* [10].
(III) The third section addresses the *Interaction Phase*. This task is supposed to be sensitive for possible delay in the transmission system. Therefore, RNVTs are used. Accordingly, the participants are asked to alternately verify a set of numbers. The participants are asked to rate the 10 APs representing the *Interaction Phase*.

The experiment was carried out by 40 participants naïve (23 female, 17 male) grouped into 20 pairs, aged between 18 and 53 years (mean age 28.7). For all three sections, the participants were asked to communicate using a transmission system (see Section 4.2.1) that was distorted by 11 randomized different degradations (see Table VII) which were analogue to the previously conducted tests.

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 VIII describes the experimental procedure and structure. Keep in mind that the rating of all APs take up to 10 minutes per condition. Therefore, the experiment was split into two sessions *á* 60 minutes to avoid participants fatigue. In addition, the participants were allowed to have extra pauses when required. Again, we used a within-subject-test-design.

Table VII. Conditions used for the validation experiment.

Con.	Degradation
1	clean
2	Sidetone -5 dB attenuation
3	Delay 1000 ms
4	Echo 100 ms
5	Packet loss 10 % (no PLC)
6	White noise 30 dB attenuation
7	Attenuation 15 dB
8	Codec LPC-10
9	Noise(6) + Echo(4)
10	Codec LPC-10(8) + Sidetone(2)
11	Delay(3) + Packet loss(5)

Table VIII. 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.

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

5.2. Attributes for the SD

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.

5.3. Results

The results of the conducted experiment are structured in five groups: first we analyze the results of the overall quality, second the results of the third section (*Interaction Phase*), third and fourth 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.

5.3.1. Overall quality

After averaging the ratings of the overall conversational quality over the conditions, a repeated measure ANOVA between the conditions as independent and the overall conversational quality ratings as depended variable was carried out, showing that the conditions have a significant impact on the judgment of the test subjects ($F(7.01, 224.14)$

Table IX. PCA results - VARIMAX rotated; Factor loadings (> 0.6; except Speaking Phase > 0.2). Boldface printed values are used for identifying the *Dimension* (Dim).

	Section I: Conversation			Section II: Listening			Section II: Speaking		Section III: Interaction
Antonym-pair	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 1
interrupted - continuous	**.760**			**.939**					
distant - close	**.892**			**.876**					
crackling - not crackling			**.901**		**.866**				
not noisy - noisy			**.901**		**.862**				
muffled - not muffled	**.738**			**.913**					
shaky - steady	**.746**			**.866**					
indirect - direct	**.827**			**.904**					
dark - bright	**.821**			**.928**					
unintelligible - intelligible	**.792**			**.929**					
not hissing - hissing			**.913**		**.831**				
clear - unclear	**.717**			**.863**			.869	**.401**	
thin - thick	**.827**			**.832**				**.994**	
distorted - undistorted	**.720**			**.884**			.942	**.280**	
loud - quiet		**.975**				**.972**	.932	**-.256**	
not fluent - fluent	**.736**						**.983**		
not helpful - helpful	.632	**.665**					**.990**		
reverberant - anechoic		**.826**					**.971**		
irritating - not irritating		**.767**					**.985**		
distracting - not distracting		**.760**					**.987**		**.834**
exhausting - not exhausting	.662	**.664**					**.991**		**.984**
concentration - no concentration	.645	**.712**					**.990**		**.970**
unpleasant - pleasant	.621	**.708**							
clear - confusing	.619	**.706**							**.980**
lazy - agile	**.772**								**.878**
easy - hard	.651	**.669**							**.983**
relaxing - annoying		**.685**							**.986**
not frustrating - frustrating	.649	**.688**							**.994**
effective - ineffective	.710	**.617**							**.976**

= 45.88, $p < .01$). With this it is proofed that the different degradation levels worked as intended (falling quality - lower rating / rising quality - higher ratings).

5.3.2. Section III - *Interaction Phase*

The results of the following PCA indicate, that the 10 attributes can be described by one dimension, covering 85.4 % of the variance of the 10 one-dimensional features. The resulting factor loadings can be seen in Table IX. This result is similar to the one of the previously conducted separate interaction experiment, and shows that the proposed dimension works as intended.

5.3.3. Section II - *Listening Phase*

The Scree Plot (see Figure 9a) of the PCA shows that only three potential dimensions result for the *Listening Phase* in Section II. The three dimensions are determined, covering 96.9 % of the variance of the 14 APs. In separate LOTs, however, four dimensions were proposed.

An explanation for this can be found by analyzing the factor loadings for each feature to the determined three dimensions in Table IX. 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

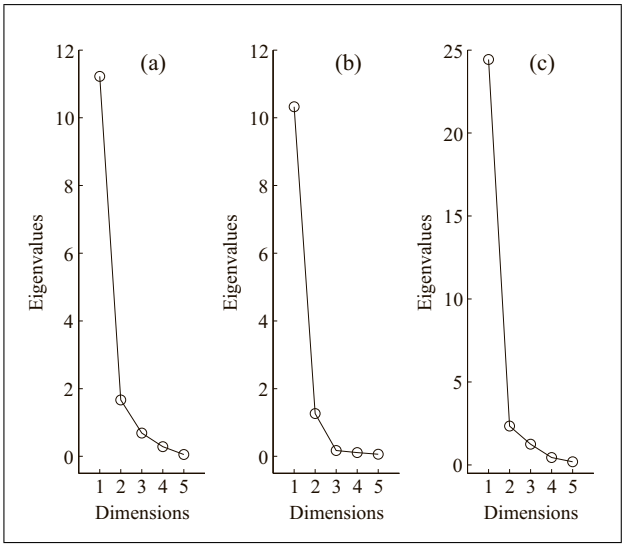

Figure 9. Scree Plots for the PCA (Validation Experiment); a - Listening Phase (II); b - Speaking Phase (II); c - Conversation Test (I).

the remaining 10 APs. *Discontinuity* and *Coloration* have only been triggered by two conditions.

Additionally, for each dimensions (*Discontinuity* and *Coloration*) one of the two conditions is combined with a different degradation that might mask the *Discontinuity* and *Coloration* degradation. Also, in [54] it was observed that in a diagnostic listening experiment subjects reflect in the colouration scale distortions that are not clearly to classify to any of the other three dimensions. These facts could be the reason of 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 identification experiments) to 3 (found in the validation experiment) is due to the limited number of conditions which could trigger these perceptual dimensions.

5.3.4. Section II - *Speaking Phase*

The Scree Plot (see Figure 9b) of the PCA shows that two potential dimensions result for the *Speaking Phase* in Section II. These two dimensions are determined, covering 96.5 % of the variance of the 11 one-dimensional features.

Two dimensions have also been discovered in the separate speaking test, termed *Impact of one's own voice on speaking* (covering features like helpful, irritating, exhausting, distracting or fluent) and *Degradation of one's own voice* (covering features like reverberant, clear, thin and distorted). Looking at the factor loadings for the *Speaking Phase* (see Table IX), it can be seen, that Dim 1 covers the same features as in the previous tests. Dim 2 explicitly only covers the features "thin", and with lower values "clear" (0,401) and "distorted" (0,280). These two features are also covered by Dim1.

Additionally, the feature "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.

5.3.5. Section I - Conversation Test

The Scree Plot (see Figure 9c) of the PCA shows that three potential dimensions result for the Conversation Test in Section I. These three dimensions are determined, covering 96.6 % 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, it seems 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 IX) 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), 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 (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 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 dimensions are still valid for a proper diagnosis of the quality of transmitted speech in a conversational situation.

5.4. Discussion

The results of the validation experiment show 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, reducing the number of separately perceivable dimensions in this phase. Thus, in subsequent experiments the presented test-paradigm (see Section 5) that specifically allows the participants to perceive each phase separately, in addition to a natural conversation paradigm, should be used.

Additionally, 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. In two condition the degradations triggering both dimensions might be masked, 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.

6. Conclusions and Outlook

The work presented in this contribution analyzed a conversational situation for the purpose of diagnostic quality assessment. The target of the work was to identify underlying quality-relevant perceptual dimensions of a conversational situation.

For this, we analyzed a conversation based on a separation into three phases: the *Listening Phase*, *Speaking Phase* and *Interaction Phase*. While the *Listening Phase* had been already object of multidimensional analyzes in related research work, perceptual dimensions characterizing the *Speaking Phase* and the *Interaction Phase* are still not well explored. Thus, we presented four initial experiments that enabled us to identify three new perceptual dimensions; *interactivity* (*Interaction Phase*), the *impact of one's own voice on speaking*, and the *degradation of one's own voice* (*Speaking Phase*). To analyze a conversation with respect to both, the user's situation and the diagnostic information, we now have a set of seven perceptual dimensions:

- Coloration,
- Noisiness,
- Discontinuity,
- Loudness,
- Impact of one's own voice on speaking,
- Degradation of one's own voice,
- Interactivity.

However, these dimensions have only been analyzed in separate studies for each phase. Therefore, a global conversation test using a new test-paradigm 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 their 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 results revealed that too many cognitive resources are bound in a conversational task, and thus the proposed dimensions are difficult to identify. Therefore, the test-paradigm used in the experiment, which specifically allows the participants to perceive each phase separately (and thus their underlying perceptual dimensions), is proposed for diagnostic quality assessments in a conversational situation.

However, the results of the experiments also showed that particular issues should be analyzed in future experiments. The dependency of the two listening dimensions *Coloration* and *Discontinuity* as well as of the two speaking dimensions *Degradation of one's own voice* and *Impact of one's own voice* should be addressed in future studies.

The named issues are not considered to be a problem of the identified perceptual quality space or the test-paradigm, but rather of the particularities of the conducted experiments. In future test, care should be taken that each expected perceptual dimension is separately covered by a sufficient number of technical conditions. For example, for the perceptual dimensions that should be further analyzed, three distinct conditions with three different characteristics of a technical degradation should be applied. In addition, the training should be applied for future experiments using the new test-paradigm.

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 *Interaction* can be analyzed separately, without putting too much cognitive load on the test participants. Otherwise, perceptual dimensions which might be important for overall quality may remain unidentified.

In addition, independent laboratories should conduct experiments to further validate the identified perpetual quality space and the presented test-paradigm. Having additional ratings at hand could allow further analyzing and comparing the results. These results might give the possibility to push a standardization process for a new subjective conversational test-paradigm at ITU-T.

The ultimate aim of this work is that the conducted studies as well as the proposed test-paradigm and dimensions form a fundamental framework for the development of an instrumental conversational speech quality measure that is based on perceptual quality dimensions.

Acknowledgments

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under Grants MO 1038/20-1. We would like to thank Maxim Spur, Maxim Szepansky, Frank Haase and Marcel Wältermann for their help and comments.

References

[1] U. Jekosch: Voice and speech quality perception: Assessment and evaluation. Springer Science & Business Media, 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. http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf, 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] P. Vary, U. Heute, W. Hess: Digitale sprachsignalverarbeitung. Teubner-Verlag, Stuttgart, 1998.

[5] ITU-T Recommendation. P.800: Methods for subjective determination of transmission quality. 1996.

[6] S. Möller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, M. Wältermann: Speech Quality Estimation: Models and Trends. Signal Processing Magazine, IEEE **28** (2011) 18–28.

[7] F. Köster, S. Möller, J.-N. Antons, S. Arndt, D. Guse, B. Weiss: Methods for Assessing the Quality of Transmitted Speech and of Speech Communication Services. Acoustics Australia **42** (December 2014) 179 – 184.

[8] M. Wältermann: Dimension-based quality modeling of transmitted speech. Springer, Berlin, 2012.

[9] N. Côté: Integral and diagnostic intrusive prediction of speech quality. Springer, Berlin, 2011.

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

[11] ITU-T Temporary Document TD 438rev1 (GEN/12): Requirement Specifications for P.AMD (Perceptual Approaches for Multi-Dimensional Analysis). 2014.

[12] S. Möller: Assessment and prediction of speech quality in telecommunications. Kluwer, Boston, 2000.

[13] ITU-T Recommendation. P.831: Subjective Performance Evaluation of Network Echo Cancellers. 1998.

[14] ITU-T Recommendation. P.CQO: Objective Conversational Voice Quality Assessment Model (under study).

[15] D. Richards: Telecommuncation by speech: The transmission performance of telephone networks. Butterworths, London, UK, 1973.

[16] ITU-T Recommendation. P.59: Artificial Conversational Speech. 2001.

[17] M. Guéguin, et al.: On the evaluation of the conversational speech quality in telecommunications. EURASIP J.Adv. Signal Process, 2008.

[18] ITU-T Recommendation. P.863: Perceptual Objective Listening Quality Assessment. 2011.

[19] ITU-T Recommendation. P.835: Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Supression Algorithm. 2003.

[20] W. Voiers: Diagnostic Acceptability Measure for Speech Communication Systems. Hartford, USA, 1977, International Confeence on Acoustics, Speech and Signal Processing (ICASSP).

[21] D. Sen: Determining the Dimensions of Speech Quality from PCA and MDS Analysis of the Diagnostic Acceptability Measure. CZ-Prague, 2001, MESAQUIN.

[22] ITU-T: Handbook of telephonometry. International Telecommunication Union, Geneva, 1992.

[23] F. Hammer, P. Reichl, A. Raake: The Well-Tempered Conversation: Interactivity, Delay and Perceptual VoIP Quality. Proc. IEEE Int. Conf. on Communications (ICC), Seoul, Korea, 2005.

[24] S. Egger, R. Schatz, S. Scherer: It Takes Two to Tango - Assessing the Impact of Delay on Conversational Interactivity on Perceived Speech Quality. Proc. 11th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2010), Makuhari, Japan, 2010, ISCA Interspeech 2010 Prceedings, 1321 – 1324.

[25] ITU-T Recommendation. P.805: Subjective Evaluation of Conversational Quality. 2007.

[26] F. Hammer, G. Kubin: Quality Aspects of Packet-based Interactive Speech Communication. Dissertation. Graz University of Technology, Graz, Austria, Juli 2006.

[27] B. J. McDermott: Multidimensional Analyses of Circuit Quality Judgements. The Journal of the Acoustical Society of America **3** (April 1969) 774–781.

[28] V. V. Mattila: Ideal Point Modelling of Speech Quality in Mobile Communications Based on Multidimensional Scaling (MDS). Proc. 112th AES Convention, DE–Munich, May 10–13 2002, Audio Engineering Society.

[29] J. L. Hall: Application of multidimensional scaling to subjective evaluation of coded speech. The Journal of the Acoustical Society of America **110** (2001) 2167.

[30] I. Borg, P. Groenen: Modern multidimensional scaling - theory and applications. Springer Series in Statistics, New York, NY, 2005.

[31] J. Kruskal, M. Wish: Multidimensional scaling, quantitative applications in the social sciences. SAGE Publications, 1978.

[32] B. Choe: Nonmetric multidimensional scaling of complex sounds: Dimensions of preference ratings and perceived similarity of vehicle noises. Shaker, 2001.

[33] J. Kruskal: Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika **29** (1964) 115–129.

[34] Y. Takane, F. W. Young, J. De Leeuw: Nonmetric Individual Differences Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features. Psychometrika **42** (1977) 7–67.

[35] H. F. Kaiser: The Varimax Criterion for Analytic Rotation in Factor Analysis. Psychometrika **23** (1958) 187 – 200.

[36] C. Osgood: The measurement of meaning. University of Illinois Press, Urbana, IL, 1957.

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

[38] MATLAB / Simulink: version 8.3.0.532 (r2014a). The MathWorks Inc., Natick, Massachusetts, 2014.

[39] ITU-T Recommandation E.800: Definitions of terms related to quality of service. International Telecommunication Union, Geneva, 2008.

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

[41] R. Appel, J. Beerends: On the Quality of Hearing One's Own Voice. Journal of the Audio Engineering Society **50** (2002) 237–248.

[42] J. Bortz: Statistik. Springer-Verlag, 2005.

[43] F. Hinterleitner, C. Norrenbrock, S. Möller, U. Heute: What Makes this Voice Sound so Bad? A Multidimensional Analysis of State-of-the-Art Text-to-Speech Systems. Miami, USA, 2012, Proc. of the 2012 IEEE Workshop on Spoken Language Technology (SLT), 240–245.

[44] S. A. Zollinger, H. Brumm: The Lombard effect. Current Biology **21** (aug 2011) R614–R615.

[45] E. Lombard: Le signe de l'élévation de la voix. Annales des Maladies de L'Oreille et du Larynx **37** (1911) 101 – 119.

[46] H. Lane, B. Tranel: The lombard sign and the role of hearing in speech. Journal of Speech Language and Hearing Research **14** (dec 1971) 677.

[47] J. Proakis: Digital signal processing: Principles, algorithms, and applications, 4/e. Pearson Education, 2007.

[48] D. Vlaj, Z. Kacic: The influence of lombard effect on speech recognition. – In: Speech Technologies. InTech, jun 2011.

[49] M. Puckette: Puredata. http://puredata.info/, 2015. Accessed: 2015-08-26.

[50] F. Hammer: Quality aspects of packet-based interactive speech communication. 2006. PhD thesis, Graz University of Technology.

[51] ITU-T Recommandation G.114: One-way transmission time. International Telecommunication Union, Geneva, 2003.

[52] G. K. Helder: Customer evaluation of telephone circuits with delay. Bell System Technical Journal **45** (1966) 1157–1191.

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

[54] J. Berger, A. Llagostera: Multidimensional Evaluation and Predicting Overall Speech Quality. Proc. 16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015), Dresden, Germany, 2015, ISCA Interspeech 2015 Prceedings, 2549–2552.