Analyzing Perceptual Dimensions of Conversational Speech Quality

Friedemann Köster, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Labs
Technische Universität Berlin, Germany
friedemann.koester@telekom.de, sebastian.moeller@telekom.de

Abstract

Most telecommunication systems are used for communication between two people which interact during a conversation. In general, the quality of conversational speech is the major indicator for telecommunication-service providers to evaluate their systems. In this context, not only the assessment of the overall quality but also the analysis of the conversational speech quality is essential. We present an initial approach towards analyzing the conversational quality by separating a conversation in phases, and extracting individual corresponding perceptual dimensions of quality, as they are subjectively perceived by the system users. These dimensions can be combined for overall quality estimation and may separately be used to diagnose the technical reasons of quality degradation. For this reason, we review known and identify new dimensions of quality perception on the basis of subjective experiments. This enables to deeply analyze conversational speech quality for diagnosis and optimization of telecommunication systems.

Index Terms: conversation, speech quality, perceptual quality dimensions, multidimensional analysis

1. Introduction

For telecommunication providers it is important to evaluate the quality experienced by service users, the so-called Quality of Experience (QoE). For this, either subjective or instrumental methods can be used. In subjective experiments, naïve test participants judge the quality of transmitted speech signals on rating scales, such as the Mean Opinion Score (MOS) [1]. Since subjective tests are time and money consuming, instrumental methods of quality estimation have been developed. The most popular instrumental methods are the signal-based full-reference models such as Perceptual Evaluation of Speech Quality (PESQ) [2] and Perceptual Objective Listening Quality Assessment (POLQA) [3] that estimate the speech quality as experienced by a user in a listening-only situation. These models for assessing the quality of transmitted speech have two major disadvantages:

- If the resulting MOS is low, they provide little insight into the reason for the quality-loss.
- The resulting MOS refers only to the specific situation, for which the subjective quality judgments were obtained. Most subjective and instrumental methods are based on the passive listening situation, which hardly matches the reality in telecommunications.

To overcome the disadvantages outlined above, separate approaches have been followed. On the one hand, diagnostic methods have been developed to not only provide information about the overall quality but also about the quality-relevant perceptual dimensions related to certain signal impairments. Yet, these methods only exist for the passive listening situation. On the other hand, conversation tests assess the quality in an interactive situation, but these methods only provide an overall quality value without any diagnostic information. These approaches result in a trade-off between the amount of diagnostic information and the number of user's situation covered, which is of particular relevance in modern telecommunication systems [4]. In this paper we present an approach on how to combine the advantages of both the diagnostic information and considering the user's situation. After giving an overview of the basic idea and its related work, we present methods and results that have been analyzed in this work to extract perceptual dimensions in a conversational situation.

2. Conversational speech quality in related research

In every-day life, telecommunication systems are usually used for a conversation between two interlocutors. Therefore, to provide information about all possible user's situations, a typical conversation has to be investigated with respect to its situations. In a conversation, the interlocutors alternately adopt the roles of listener and talker which introduces interaction between the participants. Richards presents in [5] a description of the conversational process with 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. This classification is also recommended by the International Telecommunication Union (ITU) [6]. Following this classification, a conversation can also be interpreted as a statistical (e.g. Markov-Model) process based on the aforementioned states. With this classification, a conversation, as perceived by one participant, is composed of listening and talking periods, that alternate according to the interaction with the interlocutor. According to [7], this leads, again as perceived by one participant, to three phases of a conversation: the *Speaking Phase*, the *Listening Phase*, and the *Interacting Phase*. The *Speaking Phase* and the *Listening Phase* correspond to the states (10) and (01), the *Interacting Phase* describes the alternation of the states (10) and (01); the frequency of changes describes the degree of interaction. As a disturbing side-effect the states (00) and (11) can occur. Thus, from a speech-quality point-of-view, a conversation is affected by the degradations encountered in the *Listening Phase*, in the *Speaking Phase*, and those affecting the interactivity of the conversation, the *Interacting phase*. In the following, these three phases will also describe the possible user's situations during a conversation. To obtain diagnostic information on conversational speech quality (all user's situations), the three phases will be analyzed in detail in the following subsections.

2.1. Listening phase

Regarding the user's situation, the *Listening Phase* is the most respected phase of a conversation and most subjective and objective methods that evaluate the quality of transmitted speech are based on this phase. For the subjective assessment of the listening quality, methods standardized by the International Telecommunication Union (ITU), like ITU-T Rec. P.800 [8], are used. One approach for obtain diagnostic information is using algorithms that identify perceptual dimensions related to signal impairments, such as noisiness or coloration of the timbre. Two methodologies have been used for the identification of the perceptual dimensions: (1) scaling perceptual differences of pairwise presented stimuli, and then mapping the perceptual distance to a multidimensional space (MDS [9]); or (2) rating all stimuli independently on a set of bipolar scales (Semantic Differential, SD [10]) and reducing the space of judgments with the help of a factor analysis (Principal Component Analysis (PCA)). Using both methodologies for narrowband and wideband speech stimuli, Wältermann [11] identified three perceptual dimensions: *coloration*, *noisiness*, *discontinuity*. Later, a fourth dimension representing *loudness* was added. Other proposals, e.g by Sen [12] who used the Diagnostic Acceptability Measure (DAM [13], similar to SD), identified four to seven dimensions, which are subdivisions of the dimensions identified by Wältermann. Also, Wältermann [14] showed that it is possible to directly quantify the identified perceptual dimensions in a subjective test. In sum, the *Listening Phase* is the most researched phase of a conversation: quality-relevant dimensions have been identified and subjective methodologies have been validated.

2.2. Speaking Phase

Technically, the *Speaking Phase* is usually distorted by degradations due to talker-echoes and side-tones. Thus, separate *Talking and Listing Tests* [15] 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 [16]. However, simultaneously speaking and listening can cause considerable fatigue to test participants. Therefore, so called *3rd-Party-Listening-Tests* have been developed, in which the spoken and the heard of a participant is recorded and afterwards both are rated by a third person [15]. The importance of the *Speaking Phase* is also shown in the work by Appel and Beerends [17] who introduced the instrumental model PESQM (Perceptual Echo and Sidetone Quality Measure). However, these methods only determine an integral quality value (MOS) without diagnostic information. To investigate the user's situation *Speaking Phase* with respect to diagnostic information a multidimensional analysis has to be conducted.

2.3. Interacting phase

The *Interacting Phase* covers not only the change from state (01) to (10) and the change from state (01) to (10) but also the states (00) and (11). The latter may only occur in adequate amounts, otherwise the interaction is perceived as unusual. Increased amounts of the states (00) and (11) occur especially due to transmission delay, which is particularly noticeable by a shift of the usual rhythm of conversation, leading to passive interruptions (occur when a speaker becomes interrupted by the delayed arrival of a counterpart's utterance) and active interruptions (occur when an speaker starts to speak, while he still hears his counterpart talking) [18]. For the subjective assessment of the interacting quality conversational tests have to be conducted which, unlike for the *Listening* or *Speaking Phase*, require two participants and therefore raise the level of realism through a structured conversation. To simulate a natural conversation, the participants usually follow particular scenarios. For example, ITU-T Rec. P.805 [19] recommends so called short-conversations test (SCT) [16] in which the participants are asked to solve tasks in role-plays (e.g. ordering a plane-ticket) or so-called interactive tests [20] in which the participants are asked to align numbers or addresses as fast as possible. Following these guidelines, the participants are generally asked to rate the overall quality or the interruption effort, but no further diagnostic information is provided. To the authors knowledge no former studies have fully addressed a conversation with respect to more diagnostic information.

3. Identifying the perceptual dimensions of the *Speaking Phase*

An initial study to identify the perceptual dimensions of the *Speaking Phase* has been carried out to obtain diagnostic information in a Talking and Listing Test [21]. In this study, 16 participants were asked to speak short sentences with durations of 10 to 12 s into a transmission system that was distorted by talker-echo and side-tone. For 16 conditions the participants were asked to rate the overall quality and 12 attributes in a SD experiment. The results of the overall quality evaluation, which was done for a sanity check, were similar to the studies that were conducted by Appel and Beerends [17]. The results of the PCA on the ratings stemming from the SD experiment identified two perceptual dimensions of the *Speaking Phase*. One dimensions reflects the impact of the own heard voice on the speaker while speaking, it covers attributes like helpful, irritating, exhausting, distracting or fluent. The negative and positive ends of this dimension would thus be labeled with "high impact on speaking" and "no impact on speaking", respectively. The second dimension covers attributes like reverberant, clear, thin and distorted, which suggests that this dimension reflects the degradation of the own heard voice. The ends of this dimension would be labeled with "own voice not degraded" and "own voice degraded". However, these two dimensions still have to be validated with a subsequent MDS study, which will be done in the following.

3.1. Test paradigm

In the MDS paradigm, the dissimilarity between pairwise presented stimuli is judged by each participant, resulting in a dissimilarity matrix for each participant. To generate this, a continuous scale labeled with "not similar at all" and "very similar" at its extremities was employed. The analyses discussed here are based on an average data matrix that is created computing the means over the participants.
The idea of MDS is the mapping of the dissimilarities into distances of a space smaller dimension by a mapping function. It is assumed that the psychological dissimilarity corresponds to an Euclidean distance. PORXSCAL was employed as a method for computation [22]. The new dimensionality is determined by the so-called parameter *Stress*, a badness of-fit measure indicating how bad the distances match the given dissimilarities [14].

3.2. Test design and technical setup

The Talking and Listening Test was carried out by 22 participants (14 female, 8 male) aged between 18 and 36 years. The

Table 1: *Conditions for the Talking and Listening Test*

condition	Attenuation [dB]	Roundtrip-Delay [ms]
1 (S0)	0	-
2 (E50)	-	50
3 (Sminus25)	25	-
4 (S20)	-20	-
5 (E250)	-	250
6 (Sminus10E150)	10	150
7 (S10E150)	-10	150
8 (S02)	0	-
9 (Sminus10)	10	-

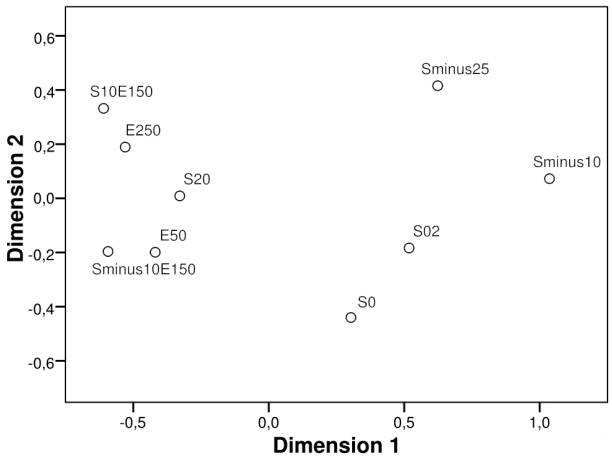

Figure 1: Results of the MDS study; normalized $[-1; 1]$

task was to read out loud a text consisting of two to three sentences (27 randomly presented texts). The test-system was distorted by side-tone and echo. For the side-tone distortion the direct back coupling of the spoken voice with different levels of attenuation and for the echo the delayed back coupled and attenuated spoken voice with varying values was generated. We used the same system as in [21]. The conditions are presented in Table 1, condition 8 is the same as condition 1; they serve as anchor-conditions. After a short training, the participants had to judge the dissimilarity between pairwise presented stimuli; in each passage of the test, the participants had to read out the same text for two conditions. All conditions were compared to each other resulting in 36 tests runs. All participants were students from the local university and were rewarded for their participation.

3.3. Results

A MDS was carried out with the generated mean dissimilarity matrix. In general, the adequate dimensionality is found if the Stress does not decrease significantly with a further increase of the number of dimensions. In this study the MDS resulted in two dimensions (see figure 2). This result is equivalent to the result of the previous conducted SD experiment. In the SD experiment the two resulting dimensions were describing I) the impact of the own heard voice on the speaker while speaking and II) the degradation of the own heard voice. These two dimensions can also be seen in the resulting space of the MDS experiment in figure 1. In this figure dimension 1 shows the following: From left to right the conditions start with strong characteristics (strong echo or loud side tone - e.g. S10E150, E250, S20) and end with rather weaker characteristics (quiet side tone, e.g. Sminus10, Sminus25). The two anchor-conditions S0 and S02 are positioned in a moderate amount. This represents the impact of the conditions on the speaker while speaking, validating and corresponding to condition I of the SD experiment.

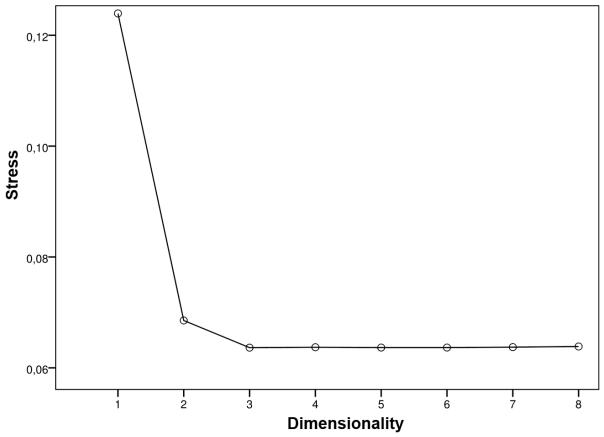

Figure 2: Scatterplot

Dimension 2 in figure 1 is representing the amount of degradation of the conditions. The scale starts with the anchor-condition S0 and then covers stepwise the conditions with higher degradations. Thus, the second dimension of the MDS experiment is describing the same effects as the dimension II of the SD experiment. Therefore, the results of the SD experiment are validated with the presented MDS experiment.

Following from these results we like to call the two perceptual dimensions of the *Speaking Phase* a) the *impact of one's own voice on speaking* (scaled from "high impact on speaking" (-1) to "no 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)).

For a sanity check two Anker-conditions (S0 and S02) were used. The resulting space of the MDS study in figure 1 shows that these two conditions have a short distance, indicating, that the different quality levels worked as intended and that the ratings are reliable.

4. Identifying the perceptual dimensions of the *Interacting Phase*

To provide diagnostic information for the *Interaction Phase* of a conversation the approach of identifying perceptual dimensions is used, and a PCA on attribute ratings stemming from a SD experiment is applied. Since the interaction is distorted by delay [23], a conversation test was carried out to investigate how the user interaction in a call is affected by varying amounts of transmission delay.

4.1. Test design

The conversation test was carried out by 32 participants (8 female, 24 male) divided into 16 pairs, aged between 19 and 31 years. Conversational tasks from the SCTs in ITU-T Rec. P.805 [19] 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 an acceptable test duration. The participants were asked to communicate over a test system that was distorted by 8 different values of delay (see table 2) which were

determined in internal pre-tests. In each test, each pair of participants first conducted one introduction SCT scenario to get familiar with the test degradations, and afterwards 8 SCT scenarios with randomly presented delays. The participants first had to judge the overall quality for a sanity check and second the SD attributes. All participants were students from the local university and were rewarded for their participation. Table 2 describes the experimental conditions.

4.2. Technical setup

The test system was implemented with three *DELL Inspirion mini 1012* notebooks. One of these notebooks served as a server with a *Debian 6 'squeeze'* operating system and *Astrisk* installed as a PBX (private branch exchange) plus *netem* as network emulator. The other two notebooks were used as clients. They had a *Windows XP* operating system and a *X-Lite VoIP Client* installed as a softphone. The sound signal was presented via a *EDIROL USB AudioCapture UA-25EX* external soundcard and a *Sennheiser HMD 46 ATC 300* stereo headset. The participants were located in two sound-insulated test rooms which met the requirements according to [8].

Subjects	32
Mean age	25.18 (σ = 6.43)
Female / Male	F:8 / M:24
Network + Codec	VoIP/netem + G.711
Delay [ms]	0,300,600,900,1300,1700,2100,2500

Table 2: Experimental Conditions

4.3. Determination of attributes for the SD

In the SD experiment, a predefined set of attributes was given to the test participants in terms of bipolar scales. The extremities of each scale are labeled with pairs of opposite attributes, so-called antonyms, each describing a one-dimensional feature. In order to find proper attributes, two pre-tests were conducted. 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*.

4.4. Results

The results of the conducted evaluation are structured in two groups: first we analyze the results of the overall quality as a sanity check and second the results of the SD experiment.
After averaging the ratings of the overall interaction quality over the conditions, one-way ANOVA was carried out, showing that the amount of delay has a significant impact on the judgment of the participants ($F=13.405$, $p<0.01$). With this data we had a proof that the different quality levels worked as intended and that the ratings are reliable.
The ratings of the SD-Experiment result in a multidimensional space that can be reduced with a PCA. For this purpose the ratings of the SD attributes are compared in terms of a correlations matrix of the SD attributes. This matrix shows, that the addressed 10 attributes highly correlate with each other (average $r\approx0.9$). The results of the following PCA indicate, that the 10 attributes can be described by one dimension, covering 96.12 % of the variances of the 10 one-dimensional features. The authors propose to call the identified perceptual dimension the *interactivity*.

5. Conclusion and Outlook

Two dimensions were already identified for the *Speaking Phase* in an initial SD experiment. With the subsequent MDS study these dimensions were confirmed. For both studies the same system with the same degradations was used, which could be an explanation for this. The two identified dimensions, 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 side-tone and echo. However, also other degradations (e.g. loud background noise) might not only effect the *Listening Phase* but also the *Speaking phase*. Therefore, in future work, it has to be analyzed, if these *Speaking Phase* dimensions can also be identified in a global conversation test with degradations concerning all three phases.
The key-finding for the *Interaction Phase* is that the perceptual dimension *interactive communicability* 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 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 degradation we varied was the delay. We did not consider degradations of the *Listening Phase* or the *Speaking Phase*, which might also make the given space smaller. In subsequent studies it has to be analyzed, if the results especially for the *Speaking Phase* and *Interacting Phase* would be different when degradations of all phases are considered in multiple similar tests. In particular, it has to be verified if the identified dimensions or the employed attributes of the phases correlate with each other, when e.g. taking echo as degradation into account, which might also effect the *Interaction Phase*. However, the identified dimension *interactivity* covers 96.12% of the variances of the attributes. This is a high value and since we used both, experts and naïve participants for the determination of the attributes, we are confident that the identified dimension is an adequate outcome for the *Interaction Phase*.

We analyzed a conversation based on a separation into three phases: the *Listening Phase*, *Speaking Phase* and *Interacting Phase*. While the *Listening Phase* had been already object of multidimensional analyses in related research work, perceptual dimensions characterizing the *Speaking Phase* and the *Interaction Phase* are still not well explored. Thus, we presented two inital experiments that enabled us to identify a new perceptual dimension *interactivity* (*Interaction Phase*) and to verify the dimensions *impact of one's own voice on speaking* and *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* and *interactive communicability*. In future work this set has to be validated (MDS and complete conversational experiment) with further research and experiments.

6. References

[1] ITU-T Recommendation P.10/G.100, *Vocabulary for performance and quality of service*, International Telecommunication Union, Geneva, 2006.

[2] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, International Telecommunication Union, Geneva, 2001.

[3] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*, International Telecommunication Union, Geneva, 2011.

[4] B. Lewcio, *Management of Speech and Video Telephony Quality in Heterogeneous Wireless Networks*, Fakultät für Elektrotechnik und Informatik, Technische Universität Berlin, Berlin, dissertation edition, 2013.

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

[6] ITU-T Rcomandation P.59, *Artificial Conversational Speech*, International Telecommunication Union, Geneva, 1993.

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

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

[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] D. Sen, "Determining the dimensions of speech quality from pca and mds analysis of the diagnostic acceptability measure," CZ-Prague, 2001, MESAQUIN.

[13] W.D. Voiers, "Diagnostic acceptability measure for speech communication systems," Hartford, USA, 1977, International Confeence on Acoustics, Speech and Signal Processing (ICASSP).

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

[15] ITU-T Recommendation P.831, *Subjective Performance Evaluation of Network Echo Cancellers*, International Telecommunication Union, Geneva, 1998.

[16] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer, Boston, 2000.

[17] 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.

[18] R. Schatz S. Egger and S. Scherer, "It takes two to tango - assessing the impact of delay on conversational interactivity on perceived speech quality," Makuhari, Japan, 2010, INTERSPEECH, pp. 1321 – 1324.

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

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

[21] F. Köster and S. Möller, "Towards a new test paradigm for the subjective quality assessment of conversational speech," Meran, IT, 2013, AIA-DAGA 2013 Conference on Acoustics.

[22] Patrick Mair Ingwer Borg, Patrick J.F.Groenen, *Multidimensionale Skalierung*, Rainer Hampp Verlag, 2010.

[23] Florian Hammer, Peter Reichl, and Alexander Raake, "The well-tempered conversation: Interactivity, delay and perceptual voip quality," in *Proc. IEEE Int. Conf. on Communications (ICC), Seoul, Korea*, 2005.