Modeling the Overall Conversational Quality Based on the Identified Underlying Perceptual Quality Dimensions

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

Abstract

Speech telecommunication systems are usually used by two interlocutors engaging in a conversation. In this context, evaluating the quality of conversational speech is important for system developers to assess their services. In addition, the quality of conversational speech respects all quality elements of telecommunication systems and thus considers listening, speaking, and interactive impairments. In contrast, it is not enough to provide information about the overall conversational quality, but also to provide diagnostic information in terms of pointing out sources for possible low quality ratings. For this, a conversation is separated into three individual conversational phases (listening, speaking, and interaction), and for each phase corresponding perceptual quality dimensions were identified. We present the linear combination of subjective dimension scores to determine the quality of each conversational phase, and the qualities of the three phases, in turn, are combined for overall conversational quality modeling. The developed model shows to provide reliable results on the available data and gives insights into the relation between perceptual quality dimensions, conversational phases, and the overall conversational quality. The presented model forms the basis for developing new instrumental diagnostic conversational quality models and allows deeply analyzing conversational speech quality for diagnosis and optimization of telecommunication systems.

Introduction

As discussed in [1], traditional method to asses the quality of transmitted speech inherent two major limitations: (i) The overall quality value – *Mean Opinion Score* (MOS) – does not provide any insights into the reason for sub-optimum quality – no diagnostic information can be extracted – and (ii) the MOS values gathered in listening-only tests disrespect conversational phases, like speaking or interaction.

To overcome both limitations with one novel method, the approach of analyzing the different phases of a conversations was followed. For this, a conversation was split into three phases according to [2]: The *Listening*, the *Speaking*, and the *Interaction Phase*. Quality-relevant perceptual dimensions were identified to analyze the three phases. Perceptual dimensions are defined as orthogonal and thus independent features of a multidimensional space formed by a perceptual event inside the listener [3, 4]. They are connected to specific *quality elements* (e.g. codecs, filters, or packet-loss) [5]. Assessing these perceptual dimensions thus serves for diagnosing speech quality.

In separate listening, speaking and interaction experiments seven perceptual dimensions were identified and validated for a conversational situation [6].

The *Listening Phase* corresponds to the situation in which the participant is listening to a vocal message. The phase can be impaired by *quality elements* like codec, noise, non-optimal signal level, or packet-loss. In [7], four perceptual dimensions were extracted for the *Listening Phase*, namely: *Noisiness*, *Coloration*, *Loudness*, and *Discontinuity*.

The *Speaking Phase* corresponds to the situation in which the participant is actively speaking. This phase can be impaired by *quality elements* like sidetone or echo. Both impairments lead to the effect that the speaker is confronted with a backcoupling of the own voice which makes the production of speech more difficult for the speaker [8]. In [2], the two perceptual dimensions *Impact of one's own voice on speaking* and *Degradation of one's own voice* were extracted for this phase.

The *Interaction Phase* describes the alternation of speaking and listening; the frequency of changes describes the degree of interaction. As a disturbing side-effect *mutual silence* (both participants remain silent) and *double talk* (both participants speak) could occur. The phase is mainly impaired by the *quality element* delay. In [2], the single perceptual dimension *Interactivity* was extracted for the *Interaction Phase*.

This sums up to seven perceptual dimensions grouped into three phases of a conversation. An overview of the three phases and their perceptual dimensions can be seen in Table 1. In order to provide diagnostic information, a test method for a direct scaling of the seven perceptual quality dimensions is required, [9]. This method was applied in a first pilot test for eleven conditions under test in [10]. The gathered ratings now give the possibility to deeply analyze the quality of transmitted speech in a conversational situation. The underlying idea – that has already been proven in [4] – is that the dimensions, as they are orthogonal, can be combined to a quality rating for each conversation phase, and that the quality ratings for each phase, in turn, can be used to determine the overall conversational quality, defined as the *quality profile*.

To follow this approach, the weights of the individual

Table 1: Overview of the three phases and their seven perceptual quality dimensions for a conversational situation [7, 2, 6].

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

Table 2: Multiple linear regression analysis for predicting the overall listening quality (MOS_{LI}) on the basis of its four underlying perceptual quality dimensions; Noisiness (MOS_{Noi}), Discontinuity (MOS_{Dis}), Coloration (MOS_{Col}), and Loudness (MOS_{Lou}).

Predictor	Standardized β Coefficient	T-stat	$P_r > \|t\|$
MOS_{Noi}	.559	7.06	.00
MOS_{Dis}	.472	4.05	.01
MOS_{Col}	.110	.69	.51
MOS_{Lou}	.130	1.38	.21

phases for the overall conversational quality, and the weights of the perceptual quality dimensions for the quality of each individual phase have to be identified. In this paper, the ratings of the pilot test, see [10], from the basis for relating the gathered ratings and thus for *modeling the overall conversational quality based on the identified underlying perceptual quality dimensions*. For this, a multiple linear regression is applied for each conversational phase and its underlying perceptual quality dimensions. In addition, the resulting three models provide input for a final linear regression model to map the overall conversational quality on the basis the three estiamted quality values for each individual phase. The paper will close with a short summary and an outlook.

Listening Phase

The analysis of the relation between the dimension ratings and the overall quality of the *Listening Phase* is based on the ratings gathered in [10]. Thus, the relation between the overall listening quality MOS_{LI} and the ratings of its four underlying perpetual dimensions (*Noisiness* (MOS_{Noi}), *Discontinuity* (MOS_{Dis}), *Coloration* (MOS_{Col}), and *Loudness* (MOS_{Lou})) is analyzed with a linear regression model.

The analysis of the linear regression is given in Table 2. The regression reaches a R^2 value of .97 and a $RMSE$ of .17. The significance test reveals that two of the four predictor coefficients are not statistically significantly different from zero ($p > .05$). This can be explained with a high collinearity ($VIF > 2$) of the two predictors and their shared variances since only eleven data points are available, see [10]. In addition, *Loudness* shows to have no significant impact on the overall listening quality. Again, this result can be explained with the low number of data points (two) for the *Loudness*. Testing more conditions triggering the *Loudness* in future experiments should make the *Loudness* predictor significant.

In similar studies presented in [4], the relation between the perceptual quality dimensions of the *Listening Phase* and the overall listening quality was analyzed in a listening-only test. It was shown that *Discontinuity* and *Noisiness* have the highest impact on the overall listening quality. Thus, the results presented in [4] are analogue to the results present in Table 2 (see the standardized β coefficient). In addition, the ANOVA of the regression model shows that it is significant ($F(4,6) = 56.791, p < .01$).

Table 3: Multiple linear regression analysis for predicting the overall speaking quality (MOS_{SP}) on the basis of its two underlying perceptual quality dimensions; Impact of one's own voice (MOS_{Ios}) and Degradation of one's own voice (MOS_{Dos}).

Predictor	Standardized β Coefficient	T-stat	$P_r > \|t\|$
MOS_{Ios}	.033	.05	.95
MOS_{Dos}	.903	1.47	.18

The regression analysis allows replacing the regression coefficients of a regression model with values that enable to estimate the overall listening quality as follows:

$$\begin{aligned}\widehat{MOS}_{LI} = &- 1.955 + .436 \cdot MOS_{Noi}\\ &+ .516 \cdot MOS_{Dis}\\ &+ .117 \cdot MOS_{Col}\\ &+ .305 \cdot MOS_{Lou}\end{aligned} \tag{1}$$

Speaking Phase

Again, the analysis of the relation between the dimension ratings and the overall quality of the *Speaking Phase* is based on the ratings gathered in [10]. Thus, the relation between the overall speaking quality MOS_{SP} and the ratings of its two underlying perpetual dimensions (*Impact of one's own voice* (Ios) and *Degradation of one's own voice* (Dos)) is analyzed with a linear regression model.

The analysis of the linear regression is given in Table 3. The regression reaches a R^2 value of .87 and a $RMSE$ of .38. The significance test reveals that the two predictor coefficients are not statistically significantly different from zero ($p > .05$). Again, this can be explained with a high collinearity ($VIF > 2$) of the two predictors and their high correlation of $> .80$. The two dimensions seem to be depended from each other in terms of their presence and thus do not provide significant predictors. In addition, again, the conducted experiment only provided four data points for the two perceptual dimensions. Thus, the dependency of both perceptual

Table 4: Multiple linear regression analysis for predicting the overall interaction quality (MOS_{IN}) on the basis of its one underlying perceptual quality dimension; Interactivity (MOS_{Int}).

Predictor	Standardized β Coefficient	T-stat	$P_r > \|t\|$
MOS_{Int}	.911	6.62	.00

dimensions should be analyzed further. At this point, a definite interpretation of their impact on the overall speaking quality is difficult as their predictors are interchangeable. The ANOVA of the regression model shows that it is significant ($F(2, 8) = 28.104, p < .01$).

To estimate the overall speaking quality MOS_{SP}, the regression coefficients from a linear regression model are replaced with the regression predictors resulting from the regression model analysis:

$$\begin{aligned}\widehat{MOS}_{SP} =& .144 + .026 \cdot MOS_{Ios}\\ &+ .819 \cdot MOS_{Dos}\end{aligned} \tag{2}$$

Interaction Phase

Again, the analysis of the relation between the dimension rating and the overall quality of the *Interaction Phase* is based on the ratings gathered in [10]. Thus, the relation between the overall interaction quality MOS_{IN} and the ratings of its underlying perpetual dimensions *Interactivity* is analyzed with a linear regression model.

The analysis of the linear regression is given in Table 4. The regression reaches a R^2 value of .83 and a $RMSE$ of .32. The significance test reveals that the predictor coefficient is statistically significantly different from zero ($p < .05$). The ANOVA of the regression model shows that it is significant (F(1,9) = 43.867, $p < .01$).

Again, the regression allows replacing the regression coefficients from a linear regression model with the regression predictors as follows:

$$\widehat{MOS}_{IN} = - .299 + .942 \cdot MOS_{Int} \tag{3}$$

Overall Conversational Quality

Knowing the relations between the perceptual quality dimensions and the three conversational phases allows modeling the ratings for each conversational phase. Now, to estimate the overall quality, the weights of the three individual conversational phases for the overall conversational quality have to be identified.

The analysis of the relation between the three conversational phases ratings and the overall conversational quality is again based on the ratings gathered in [10]. Thus, the relation between the overall conversational quality MOS_{CO} and the ratings of its three conversational phases (the *Listening Phase* (LI), the *Speaking Phase* (SP), and the *Interaction Phase* (IN)) is analyzed with a linear regression model.

The analysis of the linear regression is given in Table 5. The regression reaches a R^2 value of .97 and a $RMSE$ of .15. The significance test reveals that one of the three predictor coefficients is not statistically significantly different from zero ($p > .05$). Again, an explanation for this is the low number of available data points. However, the p value is not below but equal .05 ($p = .05$). Thus, all predictors of the applied regression model have a significant impact for the model. In turn, this means that all three conversational phases make a significant contribution for estimating the overall conversational quality. The ANOVA of the regression model shows to be significant ($F(3, 8) = 81.588, p < .01$). Again, the multiple regression model reveals the regression coefficients to estimate the overall conversational quality based on its three conversational phases:

Table 5: Multiple linear regression analysis for predicting the overall conversational quality (MOS_{CO}) on the basis of its three conversational phases; the *Listening Phase* (MOS_{LI}), the *Speaking Phase* (MOS_{SP}), and the *Interaction Phase* (MOS_{IN}).

Predictor	Standardized β Coefficient	T-stat	$P_r > \|t\|$
MOS_{LI}	.199	2.26	.05
MOS_{SP}	.442	4.75	.00
MOS_{IN}	.457	4.38	.00

$$\begin{aligned}\widehat{MOS}_{CO} =& - .393 + .188 \cdot MOS_{LI}\\ &+ .354 \cdot MOS_{SP} + .477 \cdot MOS_{IN}\end{aligned} \tag{4}$$

For the overall conversational quality, the *Speaking Phase* has a higher impact than the *Listening Phase* (see Table 5). From this it follows, that degradations that affect the *Listening Phase* (e.g. attenuation) only partly affect the overall conversational quality. Degradation concerning the *Speaking Phase* (e.g. echo), however, show to have a high impact regarding the overall conversational quality. However, as described before, it has to be further investigated how the impact of the two speaking dimension can be seen in relation to each other, and what conclusions could be drawn from these investigations in the context of the conversational phases.

The models presented for the *Listening*, the *Speaking*, and the *Interaction Phase* and the regression model to estimate the overall conversational quality based on the three conversational phases can be combined. This results in a quality profile and a model to estimate the overall conversational quality based on the seven perceptual quality dimensions of a conversational situation. The resulting quality profile is formed in two consecutive steps:

1. The quality of the three conversational phases (MOS_{LI}, MOS_{SP}, and MOS_{IN}) is estimated based on their underlying perceptual dimensions according to (1), (2), and (3).
2. The overall conversational quality MOS_{CO} is estimated based on the estimations of the conversational phases in step 1 according to (4).

Executing the two combination steps leads to the following equation to estimate the overall conversational qual-

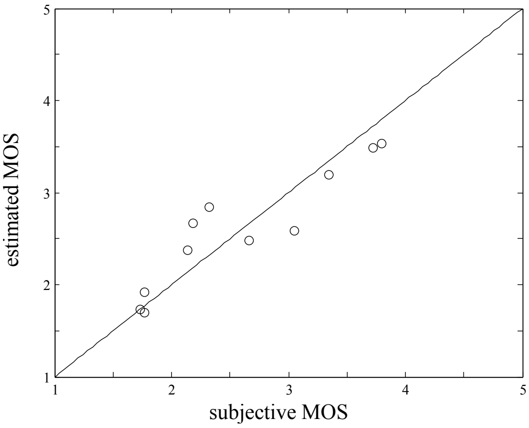

Figure 1: Estimated $\widehat{MOS}_{CO}$ vs. subjective MOS_{CO} based on the new quality profile.

ity based on its seven underlying perceptual dimensions:

$$\begin{aligned}\widehat{MOS}_{CO} = &-.393 + .188 \cdot \widehat{MOS}_{LI} \\ &+ .354 \cdot \widehat{MOS}_{SP} + .477 \cdot \widehat{MOS}_{IN}\end{aligned} \tag{5}$$

Applying (5) on the available ratings for the seven perceptual dimensions results in an estimation with a correlation of $\rho = .91$ and an error of $RMSE = .30$. Figure 1 shows an overview of the estimated $\widehat{MOS}_{CO}$ values and the subjective MOS_{CO} values for the eleven data points. It can be seen that the developed model provides reliable results with a high correlation and a small error.

Conclusions and Outlook

In this paper, the data from a already conducted experiment was used for modeling the overall conversational quality based on the identified underlying perceptual quality dimensions. For this, the relations and weights between the seven identified perceptual quality dimensions, the three conversational phases, and the overall conversational quality have been analyzed and identified. In sum, four regression models were developed. The resulting quality profile uses three of the developed models to estimate the quality of each of the three conversational phases based on their underlying perceptual quality dimensions. To estimate the overall quality, a fourth developed model that is based on the quality values of the three conversational phases, is used. The new quality profile thus gives information about the overall conversational quality and the quality of the three conversational phases based on the seven perceptual quality dimensions. The developed model provides reliable results with a high correlation and a small error.

However, the analyses showed that more data for more conditions under test are needed to poof the robustness of the new quality profile. The presented models are based on one experiment that tested eleven conditions [10]. Thus, more varying conditions with different characteristics should be tested in the future to provide more data points to train and validate the proposed quality profile.

In sum, analyses allow analyzing, diagnosing, and estimating the quality in a telephone conversion. In addition, the quality profile forms the fundamental base for future instrumental quality models. Based on the needs and the requirements the model should fulfill, the quality profile gives the possibilities to estimate the overall conversational quality either on the base of its seven perceptual quality dimensions or on its three conversational phases.

References

[1] F. Köster, S. Möller, J.-N. Antons, S. Arndt, D. Guse, and B. Weiss, "Methods for Assessing the Quality of Transmitted Speech and of Speech Communication Services," *Acoustics Australia*, vol. 42, no. 3, pp. 179 – 184, December 2014.

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

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

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

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

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

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

[8] ITU-T, *Handbook of Telephonometry*, International Telecommunication Union, Geneva, 1992.

[9] F. Köster and S. Möller, "Introducing a new Test-Method for Diagnostic Speech Quality Assessment in a Conversational Situation," in *Fortschritte der Akustik – DAGA 2016: Plenarvortr. u. Fachbeitr. d. 42. Dtsch. Jahrestg. f. Akust.*, Berlin, 2016, DEGA.

[10] F. Köster and S. Möller, "Analyzing the Relation Between Overall Quality and the Quality of Individual Phases in a Telephone Conversation," in *Proc. 17th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2016)*, San Francisco, USA, 2016, pp. 2493–2497, ISCA Interspeech 2016 Proceedings.