INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analyzing the Relation Between Overall Quality and the Quality of Individual Phases in a Telephone Conversation

Friedemann Köster, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Labs
Technische Universität Berlin, Germany
friedemann.koester@tu-berlin.de, moeller@tu-berlin.de

Abstract

Assessing and analyzing the quality of transmitted speech in a conversational situation is an important topic in current research. For this, a conversation has been separated into three individual conversational phases (listening, speaking, and interaction), and for each phase corresponding quality-relevant perceptual dimensions have been identified. The dimensions can be used to determine the quality of each phase, and the qualities of all phases, in turn, can be be combined for overall conversational quality estimation. In this article we present the work that has been conducted to identify the weights of the individual phases for the overall quality. For this, we conducted an experiment that allows the participants to perceive each phase separately and to gather the overall quality as well as the quality ratings for each individual phase. The results enable to create a linear model to predict the overall quality on the basis of the three phases. This allows to draw first conclusions regarding the relation between the individual phases and the overall quality and provides a major landmark towards a diagnostic assessment of conversational quality.

Index Terms: conversation, speech quality, conversational phases, quality estimation

1. Introduction

The quality of transmitted speech in a conversational situation as perceived by the system users – referred to as the *Quality of Experience* (QoE) [1] – is an important indicator for telephone service providers to improve and evaluate their services. Thus, assessing and understanding QoE is a fundamental task in current research. Usually, subjective listening-only experiments are conducted with naïve participants in a laboratory context to assess QoE. In these experiments – called *Listening-Only Tests* (LOT) – participants listen to a number of different stimuli and rate them on an *Absolute Category Rating* (ACR) scale (labeled from 1 – bad, to 5 – excellent). The ratings per stimulus are averaged to the *Mean Opinion Score* (MOS) [2, 3].

However, as discussed in [4], the mentioned method inherits two practical limitations: (i) The MOS value 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 LOTs disrespect conversational phases, like speaking or interaction.

The first limitation reveals that two differently impaired speech stimuli – one by e.g. temporal clipping and the other by circuit noise – can be rated with the same (low) MOS value, that alone does not give service providers information on how to improve their systems. The second limitation points out that the aforementioned method only addresses the passive listening-only situation. This only partly represents reality, as in a conversational situation speaking and interactive phases also occur.

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 [5]: 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 [6, 7]. They are connected to specific *quality elements* (e.g. codecs, filters, or packet-loss) [8]. Assessing these perceptual dimensions thus serves for diagnosing speech quality.

In separate listening, speaking and interaction experiments seven perceptual dimensions were identified for a conversational situation: four for the *Listening Phase* [9], two for the *Speaking Phase*, and one for the *Interaction Phase* [10]. The seven dimensions were validated in a complete conversation test [11]. The direct scaling of all dimensions requires a new test paradigm which allows the participants to perceive each conversational phase separately. The underlying idea – that has already been proven in [7] – 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. To follow this approach, the weights of the individual phases for the overall conversational quality have to be identified.

In this paper, we present the first results of a conversation experiment using the required test paradigm. The participants rated the overall conversational quality as well as the quality of each individual phase. The results allow to relate the overall conversational quality to the individual ratings for each phase, and thus to identify their weights for the overall quality.

After a short summary of the three conversational phases and the corresponding perceptual dimensions, we will present the conducted experiment. Then, the results are presented and the model to map the overall quality is introduced. We will close the paper with a conclusion and an outlook on future work.

2. Phases in a Conversational Situation

In a conversation, two interlocutors take turns in speaking and listening. This leads to an interaction between both participants that is described as a four-state model in [12]: One participant can either speak or listen, as in addition both participants can also speak or remain silent at the same time. From a speech-quality point-of-view, this leads to a separation of a conversation into three phases as perceived by one participant: The *Lis-*

http://dx.doi.org/10.21437/Interspeech.2016-255

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 1: Overview of the three phases and their seven perceptual quality dimensions for a conversational situation [9, 10, 11].

tening, the *Speaking*, and the *Interaction Phase*.

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 [9], 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 [13]. In [10], 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 [10], 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. This quality profile was validated in [11], and an overview of the three phases and their perceptual dimensions can be seen in Table 1.

It is obvious that each of the three phases has a significant impact on the overall quality rating of conversational speech. However, their weights and the relation between phases and the overall quality have to be identified. In the next sections, we will present the experiment conducted to gather the necessary ratings and the results that lead to a model describing the relation between the individual phases and the overall quality.

3. Conversational Experiment

In [11] the demand for a test paradigm that allows the participants to perceive each phase of a conversation separately was discussed. The new test paradigm, its design according to [14], and the setup are introduced in the following. The test paradigm guides the participants through a structured conversation, a speaking, a listening, and an interactive scenario. This enables to gather quality ratings for the overall conversation and the three conversational phases in one test session.

3.1. Test design

The conversational experiment was carried out by 36 participants (18 female, 18 male) grouped into 18 pairs, aged between 18 and 51 years (Mean=30, Standard Deviation (Std) =7.69). The participants took an average of 1 hrs, 21 min to complete the test including instructions and rating tasks. According to [14] the experiment consisted of 3 sections:

In the first section the participants conducted a structured conversation. For this, a *Short Conversation Test* [15] was applied, in which the two participants had to solve tasks in role-plays (e.g. ordering pizza). After this section the participants were asked to rate the overall conversational quality MOS_{CO}.

Condition	Degradation	Phase(s) triggered
1	clean	none
2	Sidetone -5 dB	Speaking
3	Delay 1000 ms	Interaction
4	Echo 100 ms	Speaking
5	Packet-loss 10 % (no PLC)	Listening
6	White noise 30 dB attenuation	Listening
7	Attenuation 15 dB	Listening
8	Codec LPC-10	Listening
9	Noise(6) + Echo(4)	Listening and Speaking
10	Codec LPC-10(8) + Sidetone(2)	Listening and Speaking
11	Delay(3) + Packet-loss(5)	Interaction and Listening

Table 2: The eleven conditions used in the experiment and the phases they are intended to trigger.

In the second section, the two participants changed roles between listening or speaking. First, one participant read out two sentences and the other participant listened to what is said, second, the participants changed roles. After each part, the participants rated the overall speaking quality MOS_{SP} and the overall listening quality MOS_{LI}, respectively.

In the third section, the participants performed a *Random Number Verification Task* [15] to be sensitive for delay. The participants rated the overall interaction quality MOS_{IN}.

The MOS ratings are gathered on a continuous rating scale that showed to be more sensitive than the ACR scale [16]. The scale can be seen in Figure 1. The ratings $\in [0; 6]$ were transformed to ACR MOS ratings $\in [1; 5]$ according to [16].

For the three sections, the participants communicated over a transmission system that was distorted by eleven different degradations. The degradations and the phases they were supposed to trigger can be seen in Table 2. The order of degradations was randomized between participants.

Figure 1: Quality rating scale used in the conversational experiment (taken from [17]).

3.2. Technical setup

For the experiment a test system based on *Pure Data* [18], a graphical programming language for signal processing, was used. It enables to manipulate audio effects in real-time and thus to simulate degradations like echo or nonstationary degradations. The system was extended with multiple speech codecs (e.g. LPC-10), using open-source implementations. The same setup was used for the experiments discussed in [10] and [11].

The sound signal was captured and presented via a *Beyer Dynamic DT770* stereo headset. The participants were located in two sound-insulated test rooms which met the requirements according to [3].

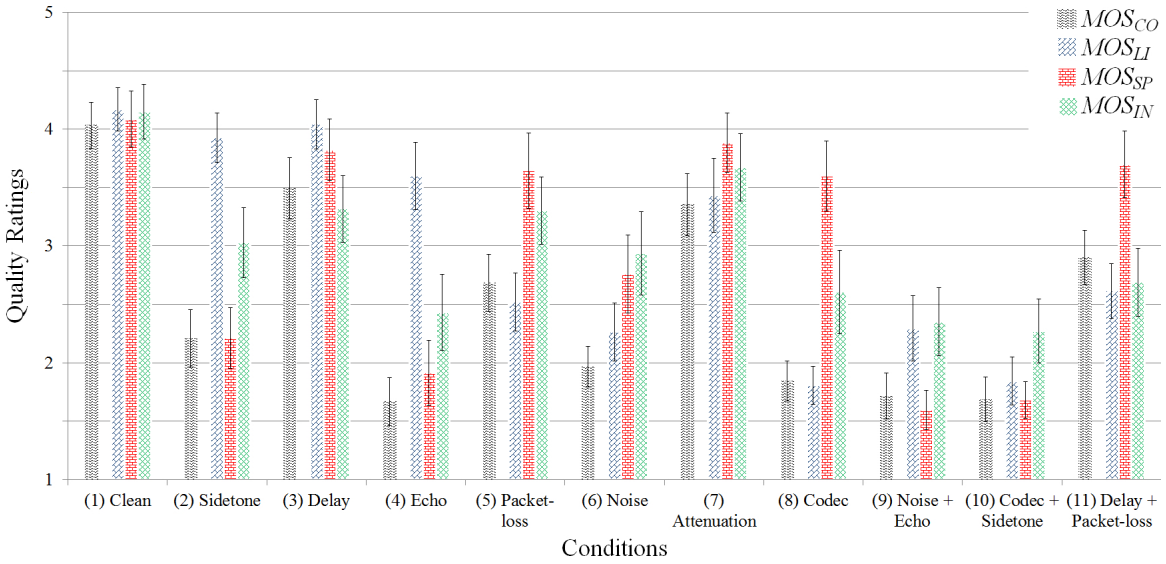

Figure 2: Subjective quality ratings resulting from the conversational experiment; the overall conversational quality (MOS_{CO}), and the quality of the three conversational phases (MOS_{LI},MOS_{SP}, and MOS_{IN}). The error-bars display the 95% confidence intervals.

4. Results

First, the key characteristics of the gathered ratings are discussed. The means of the standard deviation ($\varnothing Std$), calculated per condition, are given in Table 3. The values lie within the range of standard deviations as typically also obtained in standard ACR experiments [17].

In addition, a repeated measure ANOVA between the conditions and the four quality ratings as depended variables was carried out. The results are also given in Table 3. The terms df_n and df_d denote the degrees of freedom of the numerator and denominator of the F-test, respectively. The results show that the used conditions have a significant influence on all four quality ratings ($p < .01$).

The quality ratings can be seen in Figure 2. The resulting MOS_{CO}, MOS_{LI}, MOS_{SP}, and MOS_{IN} values for all eleven conditions are illustrated. It can be seen, that for the seven conditions (2 to 8) in which the occurring degradation is intended to trigger one phase of a conversation, the corresponding ratings of the triggered phases show a significant impact (cf. Table 2 and, e.g. condition 8, where the codec triggers the ratings for the *Speaking Phase*).

Also, the ratings indicate that the overall quality seems to be anchored to the quality ratings of the degraded phase. This can for example be seen in condition 2. Here, the transmission system was distorted by sidetone, effecting the ratings of the *Speaking Phase*. As it can be seen, the overall conversational quality MOS_{CO} and the quality of the *Speaking Phase* MOS_{SP} received almost the same rating. Conditions 4, 5, and 8 show a similar effect, with lower characteristics though.

The three remaining conditions 9, 10, and 11 (intended to trigger more than one phase) show analog ratings, except that here two phase quality ratings are anchored to the overall quality (cf. Condition 10 where MOS_{CO} is almost equal to MOS_{LI} and MOS_{SP}).

Furthermore, the results reveal that an attenuation and a transmission delay only have a slight impact on the overall conversational quality (a drop of about 0.5 MOS points). The other degradations, however, show to have a more crucial impact on the overall conversational quality (a drop of about 2.0 and more MOS points).

	Mean Std	ANOVA			
	$\varnothing Std$	df_n	df_d	F	p
MOS_{CO}	0.64	6.2	218.3	73.76	$< .01$
MOS_{LI}	0.69	5.8	202.9	74.22	$< .01$
MOS_{SP}	0.76	4.9	173.8	70.76	$< .01$
MOS_{IN}	0.88	6.9	242.3	21.49	$< .01$

Table 3: Statistical analysis of the ratings gathered in the conversational experiment; the overall conversational quality (MOS_{CO}), and the quality of the three conversational phases (MOS_{LI},MOS_{SP}, and MOS_{IN}).

	MOS_{CO}	MOS_{LI}	MOS_{SP}	MOS_{IN}
MOS_{CO}	1	.647	.818	.897
MOS_{LI}	.647	1	.289	.648
MOS_{SP}	.818	.289	1	.753
MOS_{IN}	.897	.648	.753	1

Table 4: Correlations between the overall conversational quality (MOS_{CO}) and the quality of the three conversational phases (MOS_{LI},MOS_{SP}, and MOS_{IN}). The correlations are significant at a $p < 0.01$ level.

Table 4 gives the correlations between the four quality ratings. It can be seen that the ratings for the individual phases show a significant correlation with the overall conversational quality. Also, the speaking and listening ratings significantly correlate with the *Interaction Phase*. This was expected, since the *Interaction Phase* describes the frequent change from speaking to listening and thus is connected to both phases. In turn, the ratings for the *Speaking* and *Listening Phase* have a low (but significant) correlation. In sum, while the *Speaking* and the *Listening Phase* seem to be mostly independent from each other, there is a significant correlation between the overall conversational quality and the three individual phase qualities as well as between the *Interaction Phase* and the *Speaking* and the *Listening Phase*.

Predictor	Standardized β Coefficient	T-stat	$P_r > \|t\|$
MOS_{LI}	.269	1.47	.183
MOS_{SP}	.454	2.15	.069
MOS_{IN}	.381	1.43	.194

Table 5: Multiple linear regression analysis.

5. Model

The data obtained in the conversational experiment aims at identifying the relationship between the overall conversational quality MOS_{CO} and the quality of the *Listening Phase* MOS_{LI}, the *Speaking Phase* MOS_{SP}, and the *Interaction Phase* MOS_{IN}. Thus, based on the preceding subjective results, $\widehat{MOS}_{CO}$ is estimated from subjective MOS_{LI}, MOS_{SP}, and MOS_{IN} values.

For this, we decided to apply a multiple linear regression. Linear regression was chosen (i) for its simplicity and (ii) by its similarity to the estimation of audiovisual quality, where two dimensions (audio and video) are estimated with linear models [19]. Hence, the overall conversational quality is estimated according to the following regression equation:

$$\begin{aligned}\widehat{MOS}_{CO} = \alpha + \beta \times MOS_{LI} \\ + \gamma \times MOS_{SP} + \delta \times MOS_{IN}\end{aligned} \quad (1)$$

The analysis of the linear regression is given in Table 5. There, the standardized β coefficients and the significance test for each predictor (T-stat and $P_r > |t|$) are given. The regression reaches a R^2 value of .89 and a $RMSE$ of 0.34. The significance test reveals that the three predictor coefficients are not statistically significantly different from zero ($p > .05$). This can be explained with a high collinearity (*Variance Inflation Factor*$(VIF) > 2$) of the three predictors and their shared variances. However, the ANOVA of the regression model shows that it is significant ($F(3,7) = 18.21$, $p < .01$).

The regression model allows to replace the coefficients from Equation 1 (α, β, γ, and δ) with values that enable to estimate the overall conversational quality. This leads to the following equation:

$$\begin{aligned}\widehat{MOS}_{CO} = -1.02 + 0.25 \times MOS_{LI} \\ + 0.39 \times MOS_{SP} + 0.54 \times MOS_{IN}\end{aligned} \quad (2)$$

Applied on the subjective ratings for the three conversational phases, the regression model estimates the subjective MOS_{CO} values with a correlation of $\rho = .94$ and an $RMSE$ of 0.27. Figure 3 displays the regression between the estimated $\widehat{MOS}_{CO}$ values and the subjective MOS_{CO} values.

6. Discussion and Outlook

In preceding studies the quality of transmitted speech in a conversational situation has been analyzed by dividing a conversation in three phases. For each phase quality-relevant perceptual dimensions were identified and validated. The far-end goal is to know the weights of each perceptual dimension to map the quality of each of the three individual phases that sum up for the overall conversational quality.

We presented the results of a first pilot test using a new subjective conversational test paradigm. The results provide the information needed to create a linear model that predicts the overall conversational quality on the basis of its three phases.

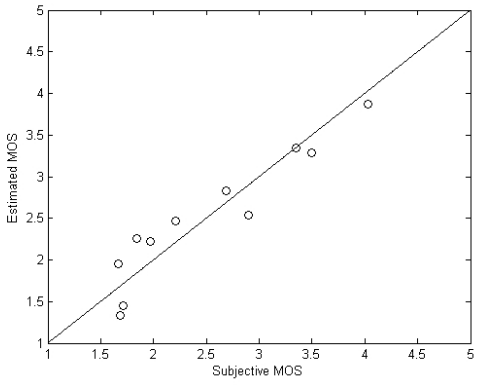

Figure 3: Estimated $\widehat{MOS}_{CO}$ vs. subjective MOS_{CO}.

The work presented in this article uncovers the relation between the overall conversational quality and the three corresponding conversational phases. It can be seen, that the *Interaction Phase* seems to have the biggest impact on the overall conversational quality. This is proven by the high correlation between the ratings of the phase and the overall quality (cf. Table 4) as well as by the high weighting in the regression model (cf. Table 5). A similar observation has also been made in [5] where the *quality element* delay (triggering the *Interaction Phase*) showed to have significant impact on the overall quality.

We explain this finding with the high importance of the *Interaction Phase* in a conversation: On the one hand, a conversation is always connected with a certain degree of interaction that typically affects the overall impression of a conversation, on the other hand, the *Interaction Phase* is connected to the *Speaking* and *Listening Phase* (cf. high correlation between the three phases in Table 4) and thus, interaction affects speaking and listening, and indirectly the overall conversational quality.

Regarding the *Speaking* and the *Listening Phase*, it can be seen that both phases are independent from each other (cf. low correlation in Table 4). For the overall conversational quality, the *Speaking Phase* has a higher impact than the *Listening Phase* (cf. regression model 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 quality. This could be explained with the high correlation between interaction and speaking (cf. Table 4), indicating that echo degradation might also have an impact on the interaction.

As a next step, the weightings of the perceptual dimensions for the three conversational phases have to be identified. Having these weightings at hand allows to deeply analyze a conversation and leads to possible models that estimate the conversational quality on basis of perceptual dimensions. In addition, more different conditions should be tested to verify the proposed conversational model and to create a wider picture of the proposed quality profile. This has to be validated with further research and experiments.

7. Acknowledgements

The presented work was supported by the Federal Ministry of Education and Research, Germany (01IS12056) and the Software Campus, as well as by the German Research Foundation (DFG), grant MO 1038/20-1. We would like to thank Frank Haase and Dennis Guse for their help and comments.

8. References

[1] Qualinet, "Qualinet White Paper on Definitions of Quality of Experience," 2013, (Version 1.2, eds. P. Le Callet, S. Möller, A. Perkins), European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland. [Online]. Available: http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf

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

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

[4] 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.

[5] M. Guéguin, R. L. Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the Evaluation of the Conversational Speech Quality in Telecommunications," *EURASIP J. Adv. Sig. Proc.*, vol. 2008, 2008.

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

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

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

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

[10] 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: ISCA Interspeech 2014 Prceedings, 2014, pp. 2041–2045.

[11] 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: ISCA Interspeech 2015 Prceedings, 2015.

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

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

[14] 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: DEGA, 2016.

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

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

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

[18] M. Puckette, "Puredata," http://puredata.info/, 2015, accessed: 2015-08-26.

[19] B. Belmudez, B. Lewcio, and S. Möller, "Call Quality Prediction for Audiovisual Time-Varying Impairments Using Simulated Conversational Structures," *Acta Acustica united with Acustica*, vol. 99, no. 5, pp. 792–805, 2013.