METHODS FOR ASSESSING THE QUALITY OF TRANSMITTED SPEECH AND OF SPEECH COMMUNICATION SERVICES

Friedemann Köster, Sebastian Möller, Jan-Niklas Antons, Sebastian Arndt, Dennis Guse, Benjamin Weiss
Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin, Berlin, Germany
{friedemann.koester, sebastian.moeller, jan-niklas.antons, sebastian.arndt, dennis.guse, benjamin.weiss}@tu-berlin.de

The quality of transmitted speech is the major indicator for telecommunication providers to classify their services. As a result, the assessment of quality is of high scientific and economic importance, and corresponding methods for assessing the quality of transmitted speech have been in the focus of multiple studies in the past. In this contribution, traditional methods are reviewed, their weaknesses are identified and improvements and extensions are proposed. The presented work covers subjective, diagnostic, instrumental, and physiological methods, as well as possibilities for evaluating long-term quality aspects.

INTRODUCTION

The use of telephony for vocal human-to-human communication is deeply integrated into our daily life. The technological development made telephony even more useful, e.g., with the emergence of mobile phones or Voice-over-IP. However, creating and maintaining a reliably working telephony system has become an even more complex task; e.g., applying heterogeneous end-user devices and network infrastructure as well as transcarrier-interconnectivity.

Within a telephony service, a speech signal can be degraded while recording, coding, transmission, decoding, and reproduction. Here, the perspective of the end-users, i.e., customers, is of major interest: Understanding how end-users perceive and experience degradations allows for improving telephony service and reacting efficiently to occurring issues like network overload.

The *Quality of Experience* (QoE) “is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment 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.” [1]. This definition of QoE and the development of methodologies to assess the subjectively perceived QoE is the objective of the European Network on quality of Experience in Multimedia Systems and Services (Qualinet COST IC1003). This is a collaboration of European QoE experts and some of the topics addressed in this article were part of this collaboration. Applied to telephony services, assessing and predicting QoE represents one of the major goals of current research. In order to understand QoE, experiments involving human participants are required. Typical methods for QoE assessment of transmitted speech are listening-only and conversational tests (with or without a task), where test subjects experience a stimulus and judge the quality. Such subjective studies enable to understand, if and how degradations are perceived under the presented condition and tasks of the study [2]. Quantitative feedback is often gathered using 5-point Absolute Category Rating (ACR) scale. The quantity evaluated from the scores is represented by the mean opinion score (MOS). A typical ACR listening-only scale resulting in a MOS (therefore also called MOS scale) can be seen in Figure 1 [3].

Quality of the speech:

excellent	good	fair	poor	bad
5	4	3	2	1

Figure 1: ACR Listening-quality scale according to [3].

For speech quality assessment usually short stimuli with a duration of up to a few seconds are used, which are sufficiently long to evaluate the quality related to codecs or network impairments. These types of subjective tests exhibit, however, certain inherent limitations:

- Stimuli must be carefully selected, so that all necessary conditions are covered while the amount of stimuli is still manageable.
- The effort of preparing and conducting the studies is significant.
- Only quantitative quality feedback is available.

The information collected during subjective studies forms the basis for the design of instrumental, so-called objective quality prediction models. A major complication is the fact that given subjective results are often only valid under the given conditions of the study in which they were collected (e.g. set of stimuli, choice of task etc.).

In the following, we present current work on the evaluation of speech quality that aims at overcoming the limitations outlined above. Such approaches include diagnostic methods, physiological measurements, and approaches to study temporal effects during the presentation of a single stimulus and over multiple distinct usage episodes. The ultimate goal is to include insights gathered by such recent approaches into traditional objective speech prediction models in order to extend their validity and therefore improve the accuracy of such models.

DIAGNOSING THE QUALITY OF TRANSMITTED SPEECH

Traditional subjective tests only provide little insight into the reason of quality impairments. More precisely, two different speech stimuli can both be rated with the same MOS value, while, for example, one is degraded by background noise and the other one due to clipping. Thus, the MOS score does not provide any diagnostic information regarding the cause of a quality impairment. Two important observations have been exploited in the design of approaches that allow for revealing the cause of a quality reduction:

- Naïve listeners identify perceptual dimensions related to impairments, such as degraded sound-color, noisiness or continuity
- Experts identify technical causes of the transmission channel, which result in impairments like sub-optimum speech-level, speechspectrum or noise-level

Two methodologies have been used to identify perceptual dimensions of transmitted speech that base on the first observation above: (1) scaling perceptual differences of pairwise presented stimuli, and then mapping the perceptual distance to a *multidimensional space* (MDS) [4]; or (2) rating all stimuli independently on a set of bipolar scales (*Semantic Differential*, SD [5]) and reducing the space of judgments with the help of a factor analysis (*Principal Component Analysis*, PCA). Applying both methodologies to narrowband (300-3400 Hz) and wideband (50-7000 Hz) transmitted speech stimuli allowed for identifying three perceptual dimensions of the quality judgments: coloration, noisiness, and discontinuity [6]. The results are illustrated in Figure 2. A fourth dimension representing loudness was added in a later study. In [7], it was shown that the identified perceptual dimensions can be quantified directly in a subjective test.

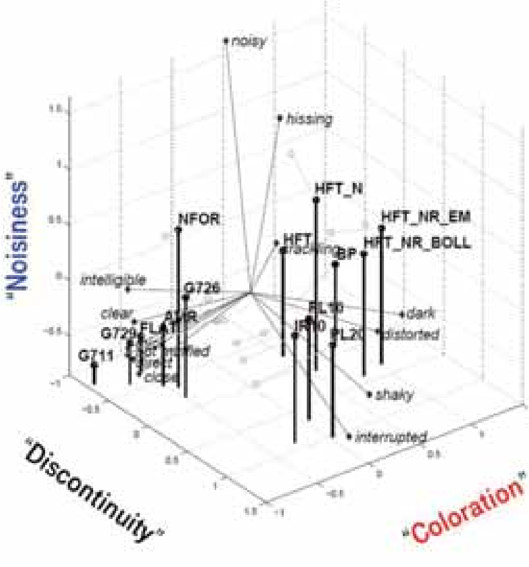

Figure 2: Results of the experiments for identifying the perceptual dimensions of transmitted speech [7].

The identification and estimation of perceptual dimensions is a current work item of the *International Telecommunication Union* (ITU-T) (working title P.AMD, *Assessment of Multiple Dimensions*). The instrumental estimation of perceptual dimensions will be discussed further in the next section.

The second observation mentioned above can be exploited as follows: First, expert-listeners identify the most dominant types of degradations ("Impairment type", Level 1, e.g. "Speech-level", "Speech-spectrum", etc., refer to Table 1) and rate them according to the three categories highly dominant, dominant, or less dominant. Afterwards, experts identify the detailed properties of each of the dominant degradations ("Degradation", Level 2, e.g. "Loud speech" or "Quiet speech" corresponding to "Speech-level", etc.) [8]. The experts can choose from the list of technical causes given in [9] where a total of 47 different impairments on Level 2, grouped into 9 categories on Level 1, are provided. Experts are asked to judge only those samples that received a MOS rating of 3.0 or lower in the subjective scores. The identification of technical causes is another work item of ITU-T with the working title P.TCA (*Technical Causes Analysis*).

Table 1: Extract of the P.TCA guidelines given in [9].

Impairment type (Level 1)	**Degradation (Level 2)**
Speech - level	*Loud speech*
	Quiet speech
	Loudness varies
	Speech level fluctuations
	Temporal speech clipping
	Choppy speech
	Self-clipping
	Speech cut-outs
Speech - spectrum	*Timbre varies*
	Muffled speech
	Sharp speech
	Coloured speech
Noise - level	*Line sounds dead*
	Loud noise
	Noise level fluctuations
	Temporal noise clipping
	Noise cut-outs

Obviously, there are links between the technical causes and the perceptual dimensions. In a first study, the results of a P.TCA annotation experiment have been analyzed with respect to the reliability of the annotations, as well as with respect to the relationships between technical causes, perceptual dimensions, and overall quality [10]. The results showed that there is a need for all - P.TCA cause analysis, P.AMD perceptual dimension analysis, and overall MOS scores - as these three metrics are only partly correlated and thus contain complementary information.

All of above mentioned approaches concentrate on a "passive" listening-only situation. Other phases appearing during the usage of a communication service, like speaking (impaired by e.g. side tone or echo) and interacting (impaired by e.g. delay) are not addressed. Concerning this, the following approach is currently under investigation [11]: The perceptual dimensions of the listening phase described above add up with the dimensions related to the speaking phase and with the dimensions related to the interacting phase. In initial studies, separate tests for the speaking and the interacting phases have been conducted, resulting in two dimensions for the speaking phase and in one dimension for the interacting phase [12]. These dimensions will have to be validated in subsequent studies, and until now, they have not proven to be orthogonal to each other. Subjective methods have to be developed for this target, which would allow for combining all phases of a conversation in one experimental paradigm.

INSTRUMENTAL ESTIMATION

Conducting subjective experiments to evaluate a telecommunication service is very time consuming, especially when large numbers of participants are required. Therefore, the need for instrumental (or so-called "objective") estimation models has grown over the past years. A variety of approaches are useful for estimating subjective ratings of the quality of transmitted speech. They can be divided into three groups based on the input information that they require:

1. Intrusive signal-based models; these models compare the input and output signal of a transmission channel and map the signal differences to a predicted rating, using a perceptual weighting

2. Non-intrusive signal-based models; these models rely only on the (degraded) output signal of a transmission channel and map signal characteristics to a predicted rating

3. Parametric models; these models use the information of a parametric description of the elements (e.g., Send Loudness Rating, Talker Echo Loudness Rating, or Roundtrip delay) of the transmission channel and map it to subjective ratings

A large amount of intrusive signal-based models have been developed in order to estimate the overall quality of the listening situation in a laboratory environment. The long-term standard of the ITU-T had been the so-called PESQ (*Perceptual Estimation of Speech Quality*) [13] model, which has been replaced by its successor POLQA (*Perceptual Objective Listening Quality Assessment*) [14]. These two models proved to be reliable for the estimation of the overall quality but do not address diagnostic features.

Estimators that represent the perceptual effects of certain system components (e.g. filters for coloration) were described in [15] in order to provide diagnostic information for the perceptual dimensions from [7]. These estimators have been further improved resulting in the DIAL (*Diagnostic Instrumental Assessment of Listening quality*) model [16]. This intrusive model combines the dimension estimators with a predictor for the overall quality in order to provide reliable instrumental assessment of both the overall quality and its corresponding perceptual dimensions. A block-diagram describing the DIAL model is depicted in Figure 3.

Furthermore, the DIAL model and the POLQA model were analyzed in terms of quality degradation indicators related to the perceptual dimensions [17]. More precisely, it was shown that indicators extracted from the two algorithms (e.g., *ERB* (Equivalent Rectangular Bandwidth), L_n (Noise Loudness), or *LTL* (Long-Term Loudness)) can be used to predict the subjective ratings of the perceptual dimensions. These results are intended to be used in the P.AMD project to develop a model to predict subjective ratings for the four perceptual dimensions coloration, noisiness, loudness and discontinuity.

The non-degraded input signal of a transmission channel is usually not available outside of the laboratory environment. Therefore, it is demanded by the industry being able to estimate the quality of transmitted speech on the basis of the degraded output signal alone. The ITU-T recommends its standard P.563 [18] for this purpose, which estimates an overall MOS of the transmitted speech quality. It provides reliable but not as highly correlated results as the intrusive models do. This is comprehensible, as the information carried by the input signal is not available. A model that estimates the perceptual dimension from [7] is currently under study so as to provide diagnostic information with a non-intrusive method. The approach is similar to that of the DIAL model, except that only the output signal is used. A non-intrusive estimator for each perceptual dimension is therefore required. While a first estimator for the perceptual dimension nosiness has already been part of a study [19] and showed good results, estimators for the other three dimensions coloration, discontinuity, and loudness are still under development.

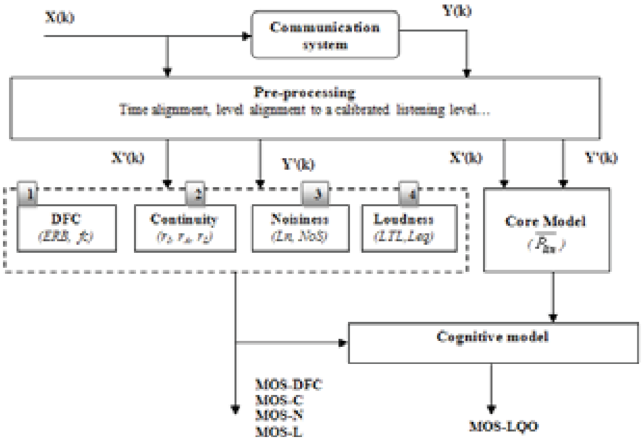

Figure 3: Overview of the DIAL model [17].

Before a transmission channel is installed, it is interesting to know during the planning phase what level of quality of the transmitted speech can be expected. For this purpose, so-called parametric models are used, which do not depend on the speech signals. These parametric models use a set of parameters that define each element of a transmission channel from the talker to the listener (e.g., loudness ratings, echo, and codecs). A prominent example of a parametric model is the E-Model, a network planning tool for the prediction of conversational and listening speech quality. The E-model is recommended by the ITU-T for narrowband and wideband network planning [20,21]. For the diagnostic information, the aforementioned three dimensions (discontinuity, noisiness, coloration) can also be estimated with a parametric approach similar to the E-model. Using parametric estimations of the perceptual dimensions, a dimension-based version of the E-model was developed, called the DNC (*Discontinuity, Nosiness, Coloration*) model [7].

The presented models all facilitate the assessment of the quality of transmitted speech and some of them also provide diagnostic information of the estimated quality. However, all of the presented models refer only to the listening situation, except for the E-Model, so that certain relevant properties of a communication channel are not covered. In [22] an intrusive model which combines listening, talking, and conversational features was developed for the estimation of the conversational speech quality. It estimates a quality value for each phase, and then maps the three values to asses an overall conversational quality. It does not provide any diagnostic information to its user, however.

PHYSIOLOGICAL QUALITY ASSESSMENT

Standard subjective tests lack information on the cognitive state of the test participant, and any physiological responses due to the presented stimuli. However, these tests currently build the basis for quality estimation algorithms. When physiological responses due to quality variations in the presented signal are better understood, they could enrich and improve current models significantly.

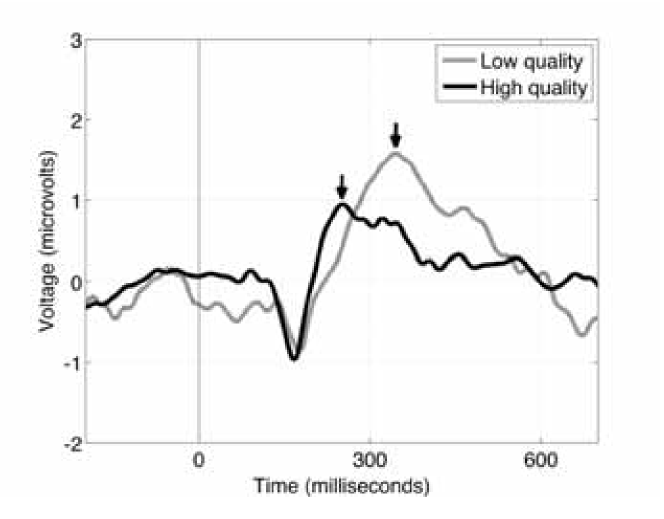

Figure 4: Exemplary grand average across all subjects of ERP plots for one stimulus in high quality and one stimulus in low quality at channel CPz. Reverberation was implemented as quality impairment of the low quality stimulus (reverberation time = 1500 ms). Arrows denote maximum amplitude (P300 peak for the low quality stimulus).

Brain activity measures are one methodology to better understand the cognitive processes underlying quality perception. Therefore, *electroencephalography* (EEG) has been introduced to the research area of speech QoE. EEG measures voltages on the participants' scalp. Using EEG, two basic analysing techniques can be considered – *eventrelated potentials* (ERPs) and *analysis of frequency bands*. ERPs are neural responses due to external events, such as presented audio signals. Measures of ERPs were used to confirm subjective results, as the amplitude of a positive deflection of the ERP after about 300ms of stimulus onset [23], P300, was shown to be gradually bigger for stronger degraded stimuli (see Figure 4). Furthermore, the P300 was shown in some cases to be even more sensitive than the behavioural answer, as in some trials subjects did not report a distortion on the behavioural level, but a similar EEG pattern as in cases where subjects did report a distortion was detected [24].

Frequency analysis is another possibility to analyse EEG signals. The obtained frequencies can be divided into several sub-bands, which provide information on the cognitive state of the listener. In several studies, it could be shown that participants tended to be more fatigued when listening to low-quality audio signals. This was detected when analysing the alpha and theta frequency power band, which are indicative for fatigue and drowsiness [25].

Another brain-based method is *functional near-infrared spectroscopy* (fNIRS), which is based on (de-)oxygenated blood flow and builds on neuroimaging techniques [26]. fNIRS was used in a study investigating the quality of synthetic speech. Significant correlations between subjective ratings and obtained fNIRS features were found [27]. So far, no study recording fNIRS using natural transmitted speech has been performed.

Similarly to brain-based measures, peripheral measures such as galvanic skin response, eye movement features, or heart-rate variability are of interest. To the authors' knowledge, only little research has been performed on peripheral physiology in the area of speech QoE, but would definitely be desirable to better understand the physiological processes underlying the speech quality perception and judgment.

LONG-TERM-EVALUATION

For longer usage periods of speech services, spanning for example over minutes or hours, or for repeated usage over weeks and months, subjective assessment methods can be used similarly to the ones for shorter samples. However, there are two major challenges with such longer time frames. Firstly, information on the variability in service quality is lost. Secondly, such retrospective judgements have proven to be far more difficult to estimate by instrumental means.

Continuous assessment methods have been used in (quasi-) laboratory conditions in order to address the first challenge mentioned above. Sliders were used to collect quality ratings continuously, applying, e.g., the *Continuous Evaluation of Time Varying Speech Qualit*y (CETVSQ) [28] similarly like in the experiment on noise and loudness perception [29,30]. Here, a MOS scale [3] is recommended to assist the usage of the sliders. At the end of each sequence, a retrospective overall quality rating is asked for on the same scale. An alternative is segmenting longer stimuli into smaller units and applying established assessment methods for short samples as discussed above. This was done, for example, to simulate conversational structures with alternating listening-only and talking phases in order to develop an estimation model for long-term speech quality [31,32].

Regarding the second challenge, a rich body of results indicating some kind of weighted average relationship between instantaneous or short-term ratings with retrospective, episodic judgments has been collected. The unweighted average tends to be too optimistic for estimating retrospective overall quality. Instead, extreme and longer degradations, as well as degradations with temporal proximity to the retrospective rating have to be weighted stronger. Refer to [33] for more information and estimation models.

Recent studies on repeated usage, such as regular calls every day, have provided controversial results. Applying weighting procedures as for episodic quality has failed so far at improving modelling performance [34]. A limitation that is inherent to all long-term methods is the required time effort. Only a limited number of conditions and stimuli can therefore be included. Hence, conditions must be very well selected.

CONCLUSIONS

In this contribution, traditional as well as new subjective and objective methods for assessing the quality of transmitted speech have been presented and discussed. Although speech quality is a well researched field, there still remains a vast number of open questions. This is partly also due to the fact that the technological infrastructure changes rapidly.

On the subjective side, special experiments can provide diagnostic information by extending them with the assessment of the perceptual dimensions noisiness, coloration, discontinuity, and loudness. Further research towards new assessment paradigms and the identification of additional conversational dimensions has to be conducted for being able to diagnose a whole conversational process that involves listening and talking.

Multiple intrusive and parametric approaches are available and provide reliable estimations regarding the objective estimation of the quality of transmitted speech. However, the need for nonintrusive models is still strong for monitoring purposes. This is also the case for models covering all conversational aspects. These models have to rely on new subjective paradigms, though.

Physiological measures are a valid test method for the assessment of speech based on neuronal states, and it could be a good complement to standard subjective tests in the future. This will particularly be of importance for high-quality media assessment in which the cognitive state of consumers is of interest.

Future work on long-term evaluation methods will provide insights into how quality evolves over longer episodes, as well as over multiple episodes. This will complement well-known short-term effects and thus enable telecommunication service providers to optimize their service for longer usage periods.

Studies combining long-term evaluation and physiological measures can build a foundation for better understanding the cognitive consequences of low-quality speech. This has already been successfully introduced in [25]. Finally, these consequences could be incorporated into objective metrics.

BIBLIOGRAPHY

[1] S. Möller, P. Le Callet, and A. Perkis, "Qualinet White Paper on Definitions of Quality of Experience," COST Action IC 1003, Lausanne, 1.1 edn., 2012.

[2] A. Raake and S. Egger, "Quality and Quality of Experience," in *Quality of Experience: Advanced Concepts, Applications, Methods*. Heidelberg: Springer, 2014, pp. 11 – 34.

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

[4] I. Borg and P. Groenen, *Modern Multidimensional Scaling - Theory and Applications*, 2nd, Ed. New York, NY: Springer Series in Statistics, 2005.

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

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

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

[8] ITU-T Temporary Document TD 686 (GEN/12), *Expert Listening for P.TCA.*: International Telecommunication Union; Rapporteur Q.16/12 (L. Malfait), 2011.

[9] ITU-T Temporary Document TD 650rev1 (GEN/12), *Requirement Specifications for P.TCA (Technical Cause Analysis).*: International Telecommunication Union; Rapporteur Q.16/12 (L. Malfait), 2011.

[10] S. Möller, F. Köster, J. Skowronek, and F. Schiffner, "Analyzing Technical Causes and Perceptual Dimensions for Diagnosing the Quality of Transmitted Speech.," in *Proc. 4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, Vienna, AT, 2013, pp. 30 – 35.

[11] F. Köster and S. Möller, "Towards a New Test Paradigm for the Subjective Quality Assessment of Conversational Speech," in DAGA, Meran, IT, 2013, pp. 440 – 443.

[12] F. Köster and S. Möller, "Analyzing Perceptual Dimensions of Conversational Speech Quality," in *Interspeech 2014*, Singapore, 2014, pp. 2041 – 2045.

[13] ITU-T Recommendation P.862.2, *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. Geneva: International Telecommunication Union, 2007.

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

[15] K. Scholz, *Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen*. Kiel: Shaker Verlag, 2008.

[16] N. *Côté, Integral and Diagnostic Intrusive Prediction of Speech Quality*. Berlin: Springer, 2011.

[17] S. Tiémounou, R.Le Bouquin Jeannès, and V. Barriac, "On the identification of relevant degradation indicators in super wiedeband listening quality assessment models," *Speech Communication*, vol. 55, no. 10, pp. 1047 – 1063, November – December 2013.

[18] ITU-T Recommendation P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*. Geneva: International Telecommunication Union, 2004.

[19] F. Köster, S. Möller, and G. Mittag, "Referenzfreie Schätzung der perzeptuellen Dimension Rauschhaftigkeit von übertragener Sprache," in DAGA, Oldenburg, 2014, pp. 501 - 502.

[20] ITU-T Reccomandation G.107, The E-model: a *Computational Model for Use in Transmission Planning*. Geneva: International Telecommunication Union, 2011.

[21] ITU-T Recommendation G.107.1, *Wideband E-mode*l. Geneva: International Telecommunication Union, 2011.

[22] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, *On the Evaluation of the Conversational Speech Quality in Telecommunications*.: EURASIP J.Adv. Signal Process, 2008.

[23] C. Duncan, R. Barry, J. Connolly, C. Fischer, P. Michie, R. Näätänen, J. Polich, I. Reinvang and C. Petten, "Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400," in *Clinical Neurophysiol.*, vol. 120, 2009, pp. 1883 – 1903.

[24] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, A.K. Porbadnigk and G. Curio, "Analyzing Speech Quality Perception using Electro-Encephalography," in *Journal of Selected Topics in Signal Processing*. IEEE, 2012, pp. 721 – 731.

[25] J.-N. Antons, F. Köster, S. Arndt, S. Möller, and R. Schleicher, "Changes of vigilance caused by varying bit rate conditions," in *IEEE Int. Workshop on Quality of Multimedia Experience (QoMEX*). IEEE, 2013, pp. 148 – 151.

[26] A. Villringer and U. Dirnafl, "Coupling of brain activity and cerebral blood flow: basis of functional neuroimaging," in *Cerebrovasc. Brain Metab. Rev. 7*, 1995, pp. 240 – 276.

[27] R. Gupta, K. Laghari, S. Arndt, R. Schleicher, S. Möller, D. O'Shaughnessy and T.H. Falk, "Using fNIRS to Characterize Human Perception of TTS System Quality, Comprehension, Fluency: Preliminary Findings.," in *4th International Workshop on Perceptual Quality of Systems (PQS)*, 2013, pp. 73 – 78.

[28] ITU-T Recommendation P.880, *Continuous evaluation of time varying speech quality*. Geneva: International Telecommunication Union, 2004.

[29] S. Kuwano, S. Namba, and Y. Nakajima, "On the noisiness of steady state and intermittent noises," *Journal of Sound and Vibration*, vol. 72, no. 1, pp. 87 – 96, 1980.

[30] H. Fastl, "Evaluation and measurement of perceived average loudness," *5th Oldenburg Symposium on Psychological Acoustics*, 1991.

[31] B. Weiss, S. Möller, A. Raake, J. Berger, and R. Ullmann, "Modeling Call Quality for Time-Varying Transmission Characteristics Using Simulated Conversational Structures," *Acta Acustica united with Acustica*, vol. 95, no. 6, pp. 1140 – 1151, 2009.

[32] B. Lewcio, *Management of Speech and Video Telephony Quality in Heterogeneous Wireless Networks*, Dissertation ed. Berlin: Fakultät für Elektrotechnik und Informatik, Technische Universität Berlin, 2013.

[33] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak and U. Reiter, "Temporal Development of Quality of Experience," in *Quality of Experience*.: Springer International Publishing, 2014, pp. 135 – 147.

[34] D. Guse and S. Möller, "Macro-Temporal Development of QoE: Impact of Varying Performance on QoE over Multiple Interactions," in *DAGA 2013*, Meran, IT, pp. 452 – 455.