This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2009 proceedings

Speech Quality while Roaming in Next Generation Networks

Sebastian Möller, Marcel Wältermann, Błażej Lewcio, Niklas Kirschnick, Pablo Vidales
Quality & Usability Lab, Deutsche Telekom Laboratories
Technische Universität Berlin
Berlin, Germany
{sebastian.moeller, marcel.waeltermann, blazej.lewcio, niklas.kirschnik, pablo.vidales}@telekom.de

Abstract— **In NGNs, handovers between different wireless access technologies provide seamless roaming during voice calls. The resulting speech quality depends on the audio bandwidth of the speech codecs used in the respective networks, as well as on degradations resulting from the handover, coding, and packet loss. We present the results of four listening experiments where speech quality is quantified as a function of network and codec characteristics, and compare them to estimations obtained from instrumental models. The results show when and under which circumstances a network handover and/or codec changeover should be scheduled in order to obtain better speech quality. This is important for the development of high-quality roaming strategies.**

Keywords- speech quality, Next Generation Network, handover, wideband speech transmission

I. INTRODUCTION

In Next Generation Networks (NGNs), the convergence of different wireless technologies will provide the user with transparent and ubiquitous access to speech and multimedia services. The independence of network and service layers enables users to move through geographical areas covered by different wireless network technologies, while the service is preserved. In order to guarantee seamless mobility also for time-critical services such as Voice-over-IP (VoIP), sophisticated mobility-enabling protocols [1][2][3] are necessary which ascertain a fast and robust roaming between different wireless networks (so-called vertical handovers).

Depending on the network technology and the available *network bandwidth*, speech signals need to be coded at different *audio bandwidths*. Whereas GSM currently mainly transmits the traditional telephone band (300-3400 Hz), HSDPA and WLAN may offer also wideband (50-7000 Hz) speech transmission. As a result of a handover, the codec may change as well, depending on the VoIP service provider. As a consequence, also the audio bandwidth provided by the codec may switch between narrowband (NB) and wideband (WB) within a single call. It is obvious that such audio bandwidth transitions affect the perceived quality of the connection. Additional degradations may be introduced by the handover itself (e.g. due to silence frames), the speech codec, as well as packet loss or discard. Despite the high relevance for NGN set-up, the perceptual effects linked to these new network scenarios have not yet been formally investigated to the authors' knowledge.

In order to design network handover and codec changeover strategies which provide an optimum quality to the user, the impact of different types of degradations on perceived overall quality (so-called Quality of Experience, QoE) needs to be quantified. Until now, subjective tests are the only valid and reliable means for this purpose. Once the perceptual effects have been quantified, they can be described by instrumental quality prediction models like the ones recommended by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T). This body currently recommends signal-based models which estimate the quality of transmitted speech in a listening-only situation by comparing the clean and degraded signals on a perceptual level [4], as well as parametric models which estimate conversational speech quality on the basis of transmission channel parameters (loudness ratings, noise levels, packet loss rates, etc.) [5]. Until now, none of these models has been validated to correctly predict the effects of NGN handovers and/or codec changeovers on user perception.

In this paper, we present the first part of an in-depth study on the effects of vertical handovers and codec changeovers on speech quality in NGNs. Our aim is to quantify the effects of different network and codec characteristics (handover point within a call, networks used before and after the handover, audio bandwidth available in these networks, and packet loss) on the quality perceived by the user. The following research questions guided our experiments, because they need to be answered in order to design optimum handover strategies:

1. *Which network and codec characteristics are the most relevant ones from a perceptual point-of-view?*
2. *Is it advantageous to switch from NB to WB whenever possible, or does audio bandwidth switching degrade perceived quality? If yes, under which circumstances?*
3. *Is it possible to predict the effects of network handovers and codec changeovers with existing quality prediction models?*

In order to get analytic insights into the perceptually relevant effects, we limited our first investigation to the listening-only situation. We additionally compared the subjective results to instrumental estimations using the cited models, in order to check whether we will be able to replace subjective listening-only tests in future studies. This way, we are able to provide initial answers to all three research questions. In the second part of the study, we will carry out conversation tests to counter-check the obtained results in a

978-1-4244-3435-0/09/$25.00 ©2009 IEEE

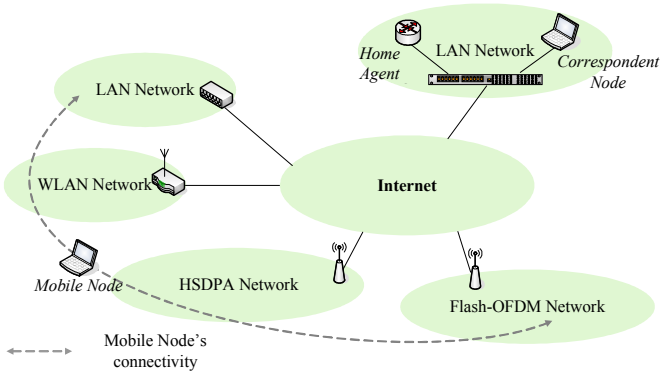

Figure 1. Overview of the NGN testbed.

more realistic situation, and use them for improving the mentioned quality prediction models for NGN handovers.

A detailed analysis of the perceptual effects requires controllable network conditions. For this purpose, an NGN testbed has been developed which allows to manipulate the mentioned network characteristics. This testbed is described in Section II. With the help of the testbed, long and short speech samples have been degraded in a controlled way, and auditory tests have been carried out in order to assess the associated quality. The organization of these tests is described in Section III, and Section IV summarizes the main results. The subjective scores are finally compared to the estimations obtained with both signal-based and parametric quality prediction models in order to test the validity of these approaches, see Section V. Section VI draws conclusions for developing handover strategies which are optimal from a perceptual quality point-of-view.

II. NGN TESTBED

The NGN testbed uses Mobile IPv4 to enable seamless handovers between different radio access technologies, i.e. continued service (potentially with packet loss) when network handovers are performed. It consists of a Home Agent (HA), a Mobile Node (MN), and a Correspondent Node (CN) which communicate with each other to establish a VoIP connection. With this set-up, an NGN incorporating LAN, WLAN, Flash-OFDM (Orthogonal Frequency Division Multiplex) and HSDPA (High Speed Downlink Packet Access) network technologies is emulated. CN and MN are implemented on laptops with Linux 2.6.18.2, and the HA is a Cisco 2620XM router with CISCO IOS 12.2(8r). Details on the hardware and software components as well as on their communication are described in [6], and an overview is depicted in Fig. 1.

The VoIP framework is implemented with the help of the PJSIP software [7] to which extensive modifications have been made. It makes use of the SIP/SDP parameter negotiation, parallel media stream establishment, and RTP packet filtering to enable codec changeover during a call. Different changeover techniques have been developed and compared in [3][6], showing that a soft codec changeover with a shared network port seems to be the best solution.

Different speech codecs and packet loss concealment techniques can be handled by the extended VoIP client. For the experiments described in the following, we used ITU-T Rec. G.722.2 at 23.05 (Tests 1) and 12.65 kbit/s (Tests 2) for WB speech transmission, and ITU-T Rec. G.711 at 64 kbit/s for NB transmission. The first choice was guided by auditory test results showing that this codec provides close-to-perfect transmission quality, whereas the G.711 logarithmic PCM is the most widespread NB codec. Both codecs use the recommended packet loss concealment (PLC). As parallel media streams are allocated during the transition phase between networks, a codec changeover can be scheduled before or after network handover. The testbed can additionally introduce random packet loss of a defined percentage *Ppl* in order to degrade network conditions in a controlled way.

III. DESCRIPTION OF THE EXPERIMENTS

In NGN scenarios, speech quality is expected to vary during a call, as a result of network handover and/or codec changeover. Thus, in order to quantify quality, the entire length of a call has to be considered. Standard listening-only tests which make use of speech samples of 4-8 s length [8] are not suitable for this purpose. On the other hand, conversational tests – despite being comparable to normal telephone usage and thus being ecologically valid – place a content-related focus on the user's attention; in such a situation, users are generally less analytic in their judgments, and it might happen that subtle perceptual differences get blurred.

As a compromise, we opted for a two-fold test protocol: (a) We simulated conversations of 60 s length by concatenating 5 meaningful speech segments alternating with pauses, playing them back to the test participants, asking them to answer content-related questions during the pauses, and asking for an overall quality judgment at the end of the simulated conversation; this approach has been developed in [9] and is now recommended for call-quality measurement in [10]; (b) The composing segments of the simulated conversations and some additional segments of approx. 6 s length were presented to the participants in a standard listening-only context, asking for an overall quality rating after each sample. We carried out two tests of the first type (Tests 1a and 2a) and two corresponding tests of the second type (Tests 1b and 2b). The following sections describe the test conditions, set-up and participant group in more detail.

A. Test Conditions

Test 1a concentrates on WB/NB transitions and the effects of packet loss on perceived quality. It contains two conditions with pure NB and WB calls, 4 conditions where packet loss continuously increases until the middle (3rd segment) of the call, and then switching occurs to a loss-free network with a different codec (or not), and 4 conditions where NB→WB or WB→NB transitions occur at the beginning (2nd segment), or at the end (4th segment) of a call. We consider packet loss rates of 10-20% to be realistic constraints when a handover should be executed at latest. Table I summarizes the conditions. The corresponding Test 1b contains all segments of the simulated conversations, plus additional samples with similar degradations, addressing also Flash-OFDM networks. The resulting list of 25 segments for this test is not reproduced here, but a rough description of the conditions is provided in Fig. 3.

TABLE I. TEST 1A CONDITIONS. H: HSDPA; W: WLAN; *PPL*: PACKET LOSS IN %; SWITCHING AT THE BEGINNING (BEG.), MIDDLE (MID.) OR END OF A SIMULATED CALL.

No.	Network(s)	Codec(s)	*Ppl* per segment
1	H	G.711	0
2	W	G.722.2	0
3	H	G.711	0,10,20,10,10
4	W	G.722.2	0,10,20,10,10
5	H→W mid.	G.711→G.722.2	0,10,20,0,0
6	W→H mid.	G.722.2→G.711	0,10,20,0,0
7	H→W beg.	G.711→G.722.2	0
8	H→W end	G.711→G.722.2	0
9	W→H beg.	G.722.2→G.711	0
10	W→H end	G.722.2→G.711	0

TABLE II. TEST 2A CONDITIONS. SEE TABLE I FOR EXPLANATIONS.

No.	Network(s)	Codec(s)	*Ppl* per segment
1	W	G.722.2	0
2	H	G.711	0
3	H→W beg.	G.711→G.722.2	0
4	H→W mid.	G.711→G.722.2	0
5	W→H mid.	G.722.2→G.711	0,3,3,0,0
6	W→H mid.	G.722.2→G.711	0,5,5,0,0
7	W→H mid.	G.722.2→G.711	0,10,10,0,0
8	W→H→W	G.722.2↔G.711	0
9	H→W→H→W	G.711↔G.722.2	0
10	W	G.722.2	0,0/5,5,5/0,0

Test 2a was designed to put a magnifier on the most interesting findings of the first test. It focuses on the switching position within a simulated conversation, as well as on additional packet loss rates, see Table II. The corresponding Test 2b with short samples also included different network load and high packet-loss-rate scenarios for limited WLAN networks (overall 27 conditions).

B. Test Set-up

Tests 1a/b and 2a/b were carried out at distinct points in time, with different participant groups. Test participants were invited to a sound-insulated laboratory, were instructed about the purpose of the test, and listened to the samples in three sessions of approx. 25 min. each (2 sessions for parts a, 1 session for parts b). Speech samples were presented over a Sennheiser HMD 410 headset at a comfortable listening level, with a background level below 35 dB(A) [8]. At the end of each simulated conversation of part a, as well as after each sample of part b, participants had to rate the overall quality on a 5-point absolute category scale, with 5 corresponding to "excellent" and 1 to "bad" quality. The test set-up and scale followed mainly the requirements given in [8] and [10]. 13 participants took part in Test 1a, 24 in Test 1b, 14 in Test 2a, and 17 in Test 2b. They were recruited from the normal telephone-user population, did not report any hearing impairment, and received a voucher in return for their effort.

IV. ANALYSIS OF EXPERIMENTAL RESULTS

Fig. 2 shows the auditory judgments of the simulated conversations of Test 1a, averaged over all test participants and samples used in each condition (Mean Opinion Scores, MOS, and standard deviations, std.). As expected, the pure WB condition (#2) is rated best, and better than the pure NB condition (#1). The packet loss of conditions #3 and #4 impacts quality significantly (analysis of variance with $F(9,27)=10.247$; Tukey HSD post-hoc test with $p<0.001$); these conditions have the lowest quality of the entire test, showing that packet loss seems to be the most dominant factor for quality degradation. Conditions #5 and #6 differ from #3 and #4 in that a network handover and a codec changeover occur, leading to a different audio bandwidth and no packet loss at the end of the call. In both cases, the switching results in a significant improvement of perceived quality (Tukey HSD, $p<0.05$). Apparently, network handover can be an efficient tool for quality improvement in case that packet loss impairs quality significantly. This also holds when a transition from WB to NB is necessary, as the comparison between conditions #4 and #6 shows.

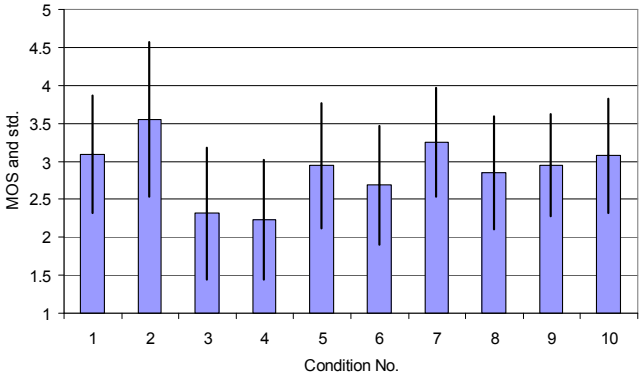

Figure 2. Results of Test 1a. See Table I for conditions.

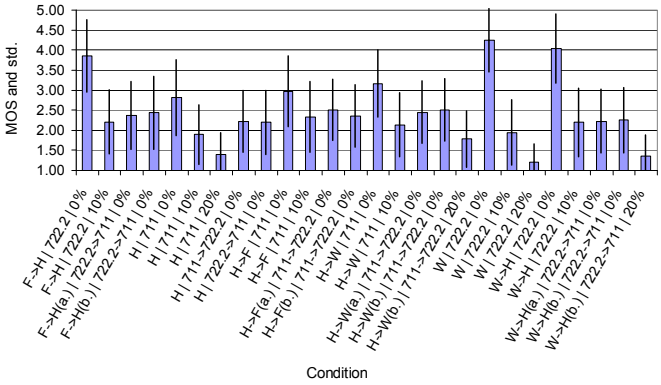

Figure 3. Results of Test 1b. First field: network conditions: W: WLAN; H: HSDPA; F: Flash-OFDM; →: network handover; (b.): codec changeover before network handover; (a.): codec changeover after network handover. Second field: codec conditions: 711: G.711 at 64 kbit/s; 722.2: G.722.2 at 23.05 kbit/s; →: codec changeover. Third field: packet loss conditions (*Ppl* in %).

Even when no packets are lost, switching from NB to WB may be advantageous in case that a significant time period remains in order to take profit of the improved quality. A comparison between conditions #7, #8 and #1 shows that quality improves when switching from NB to WB at the beginning of a call; however, when switching occurs at the end of a call, the quality degrades compared to the pure NB case. For the opposite (WB→NB) direction, switching definitely degrades quality; the longer the WB connection remains established, the better the perceived quality.

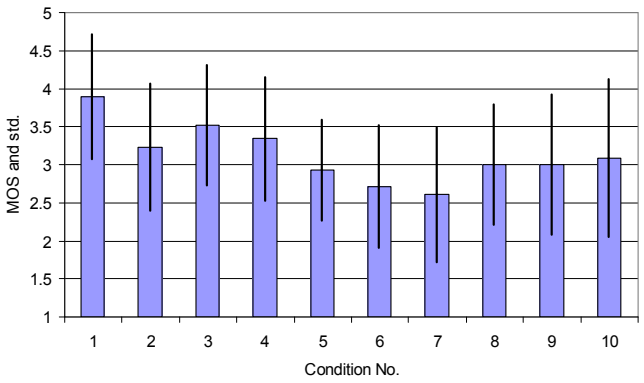

Figure 4. Results of Test 2a.

The results of Test 1b help to quantify the trade-off between audio bandwidth and packet-loss degradations, see Fig. 3. As in Test 1a, packet loss is the most important quality degradation. In case of zero packet loss, a network handover without changing the codec degrades quality only slightly. However, if the codec has to be changed as well, this has a significant effect on perceived quality. For WB→NB transitions, the impact of bandwidth switching is roughly equivalent to a 5-10% packet loss degradation. For the NB→WB transition, quality improves in all cases, and the improvement is again equivalent to a 5-10% packet loss in the improved condition. These findings are valid both for the transitions between WLAN/HSDPA and Flash-OFDM/HSDPA. No significant difference was found between codec switching before and after the handover.

Test 2a confirms most of the findings of the first test, cf. Fig. 4. Once again, pure WB is rated best, and packet loss has the highest impact on speech quality. A comparison of conditions #2, #3 and #4 shows that switching from NB to WB is advantageous if it occurs early in the call. Switching around the middle of the call results in approximately the same quality as keeping NB for the entire call. Switching to loss-free NB because of packet loss in the WB network (#5, #6 and #7) does not improve quality: already with 3% packet loss in WB switching is worse than the pure NB case (#2). A two-segment-long period of 5% packet loss in the middle of a WB call (#10) is roughly equivalent in quality to the same period of NB in that call (#8). Multiple switching (#8 and #9) is always worse than single switching (#3 and #4).

The results of Test 2b (not reproduced here to save space) confirm the importance of packet loss, and show that the differences between WLAN and Flash-OFDM are negligible when the same codec is used. The impact of bandwidth switching depends in this case on the basic quality level provided by the WLAN: If the quality is high, then the impact of bandwidth switching is remarkably high as well; if the basic quality level is already low due to high packet loss in the network, then switching codecs does not have a strong effect. In other words: packet-loss degradations are able to mask the positive effect of switching to a higher audio bandwidth.

V. Estimation of Quality Judgments

Auditory tests like the ones described in the previous sections are expensive and time-consuming; thus, we checked the available quality prediction models as to whether they provide valid estimations also for NGN handover situations. Two different types of models have been checked: A parametric model which estimates quality on the basis of the parameters describing the network conditions, and a signal-based model which estimates quality as a perceptually-weighted distance between the input and the output signals of the network under test.

We used an extended version of the E-model as a parametric model. It is based on the original algorithm described in [5] for NB networks and has been modified to take into account WB transmission (by linearly extending the underlying transmission rating scale from 100 to 129), WB speech codecs (by defining codec-specific equipment impairment factors), and packet loss. The necessary modifications are all described in [11] and are recommended in the most recent update of [5]. As the E-model does not yet consider NB↔WB transitions, we decided to calculate two separate scores for the samples in which such NB↔WB transitions occur, then calculate an average of the underlying transmission ratings, and transform this back to the MOS scale. As a signal-based type of model, we used the WB extension of the PESQ model which is described in [12]. It mainly consists of the corresponding NB version [4], but applies a WB input filter and a different mapping function.

It has to be emphasized that both types of prediction models have not yet been validated for the scenarios investigated here. In addition, they estimate instantaneous speech quality of short samples, and not of entire conversations. Thus, we applied the predictions only to Tests 1b and 2b. The parametric E-model uses packet loss as an input parameter, which has not been manipulated in a controlled way in Test 2b; thus, for this test, only the signal-based model can be used. We compare the model estimations to the auditory test results in terms of MOS, and calculated the Pearson correlation r and the root mean squared error σ for each test.

The results are given in Table III. The parametric E-model does only use general information on the network condition (average packet-loss percentage *Ppl*, codec type) as an input; consequently, the predictions are not accurate. In contrast to this, WB-PESQ is able to recognize speech signal degradations caused by the network handover, codec changeover and packet loss. For Test 1b, this leads to very good prediction accuracy, even better than the value of $r = 0.93$ which is obtained for in-scope data [13]. The prediction accuracy is slightly lower for Test 2b which contains conditions with G.722.2 coding at 12.65 kbit/s; WB-PESQ has been shown to have bigger problems in predicting the effects of this codec compared to the 23.05 kbit/s bit-rate used in Test 1a/b, see e.g. [14].

TABLE III. Pearson Correlation and Root Mean Squared Error between Auditory and Estimated MOS.

Test	WB E-model		WB-PESQ	
	r	σ	*r*	σ
1b	0.58	1.44	0.96	0.44
2b	n.a.	n.a.	0.89	0.70

VI. Consequences for Handover Strategies

The auditory results help to answer the research questions raised in the introduction:

1. The most important network characteristic is the packet loss rate. The second most important characteristic is the audio bandwidth. Switching the audio bandwidth is roughly equivalent to the quality degradation of 5-10% packet loss, in both NB and WB conditions. The degradations due to network handover (make-before-break) alone – without codec switching – are negligible in comparison to packet loss and bandwidth switching.
2. Switching codecs is advantageous if the packet loss rate is high, and if the changeover helps to reduce the packet loss impact. Switching codecs in order to take profit of a larger audio bandwidth is advantageous only if a sufficiently long period of WB speech transmission remains. From the limited results of our tests, it seems that this minimal length is around 30 s. Unfortunately, it is not possible to know the remaining length of a call. As a workaround, handover strategies could consider the perceived quality and use this value to estimate multiple scenarios (using different remaining lengths), and then select the most beneficial for the user, assuming past and current conditions.

   An interaction could be observed between packet-loss and audio-bandwidth degradations:
   - If packet loss is high (low basic quality), the impact of audio bandwidth on perceived quality is low.
   - If packet loss is low (high basic quality), the impact of audio bandwidth on overall call quality is high.
3. Parametric quality prediction models like the E-model are not yet able to estimate the speech quality resulting from codec changeovers. Signal-based models like WB-PESQ do a better job, but do not always perform as well as on in-scope data.

The results may help to design efficient network handover and codec changeover strategies. In bad network conditions (high packet loss rate), a handover should be made if the packet loss rate can be reduced by this step. In this situation, it is not important whether the audio bandwidth can be maintained or not; the reduction of packet loss should be the ultimate goal. In contrast to this, a handover can also be fruitful in good network conditions (low packet loss rate). In this case, switching to a higher audio bandwidth can help to significantly improve quality. The improvement is most effective if it occurs early in the call; the remaining call duration should be more than 30 s. Switching from WB to NB is always linked to a loss in quality; the pure network handover without codec switching, however, does not significantly impact the perceived quality.

VII. Future Work

We presented the first part of an in-depth study of the effects of network and codec characteristics in NGNs on perceived speech quality. In order to obtain valid and analytic results, we decided to use an NGN testbed to generate typical stimuli, and to judge them in a simulated conversation, as well as in a standard listening context. With the testbed working in real time, additional conversation tests will be carried out in order to validate the results in a more realistic setting. On the basis of these tests, instrumental quality prediction models like PESQ [4] and the E-model [5] will be extended in order to better take into account handover effects and time-varying quality during a call, e.g. in the way described in [9]. In this way, extended models can be used to improve handover strategies in real time, depending on measured network characteristics. In addition, the obtained results need to be mapped to actually observed network behavior.

Acknowledgment

The experiments have been carried out in the "Mobisense" project funded by Deutsche Telekom AG. Contributions came from the project "Attribute-based Speech Quality Measures" funded by the DFG (MO 1038/5-2). This funding and the fruitful discussions with team members of T-Labs and DAI-Labor are gratefully acknowledged.

References

[1] IETF RFC 3344, IP Mobility Support, (Author: C. Perkins) Internet Engineering Task Force (IETF), 2002.

[2] H. Schulzrinne E. Wedlund, "Application-Layer Mobility Using SIP", SIGMOBILE Mob. Comput. Commun. Rev., 4(3), 2000, pp. 47–57.

[3] M. Wältermann, B. Lewcio, P. Vidales, and S. Möller, "A Technique for Seamless VoIP-Codec Switching in Next Generation Networks", in: IEEE Int. Conf. on Communications (ICC 2008), Beijing, 2008.

[4] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs", Int. Telecom. Union, Geneva, 2001.

[5] ITU-T Rec. G.107, "The E-Model, and Computational Model for Use in Transmission Planning", Int. Telecom. Union, Geneva, 2005.

[6] P. Vidales, N. Kirschnick, B. Lewcio, F. Steuer, M. Wältermann, and S. Möller, "Mobisense Testbed: Merging User Perception and Network Performance", in: Proc. 4th Int. Conf. on Testbeds and Research Infrastructures for the Development of Networks & Communities, Innsbruck, March 18-20, 2008.

[7] PJSIP – Open Source SIP Stack and Media Stack for Presence, Im/instant Messaging, and Multimedia Communication, 2008, http://www.pjsip.org.

[8] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality", Int. Telecom. Union, Geneva, 1996.

[9] J. Berger, A. Hellenbart, R. Ullmann, B. Weiss, S. Möller, J. Gustafsson, and G. Heikkilä, "Estimation of 'Quality per Call' in Modelled Telephone Conversations", in: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, 2008.

[10] ETSI TR 102 506, "Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call", European Telecommunications Standards Institute, Sophia Antipolis, 2007.

[11] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment Factor Framework for Wideband Speech Codecs", IEEE Trans. Audio, Speech and Language Process. 14(6), 2006, 1969- 1976.

[12] ITU-T Rec. P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs", Int. Telecom. Union, Geneva, 2005.

[13] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality – technology and applications," IEEE Trans. Audio, Speech and Language Process., vol. 14, no. 6, 2006, pp. 1890–1901.

[14] N. Côté, S. Möller, V. Gautier-Turbin, A. (2006). "Analysis of a Quality Prediction Model for Wideband Speech Quality, the WB-PESQ, " in: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 115-122.