EVALUATION OF AMR-NB AND AMR-WB IN PACKET SWITCHED CONVERSATIONAL COMMUNICATIONS

Hervé Taddei[1], *Imre Varga*[1], *Laëtitia Gros*[2], *Catherine Quinquis*[2]
Jean Yves Monfort[2], *Frank Mertz*[3], *Thorsten Clevorn*[3]

[1]Siemens AG ICM MP, Haidenauplatz 1, D-81675 Munich, Germany
Email: firstname.name@siemens.com
[2]FTR&D, 2 Avenue Pierre Marzin, F-22307 Lannion, France
Email: firstname.name@rd.francetelecom.com
[3]RWTH Aachen University, Institute of Communication Systems and Data Processing, Germany
Email: name@ind.rwth-aachen.de

ABSTRACT

The introduction of Packet Switched (PS) networks (e.g. IMS) creates a need to evaluate the speech transmission quality when using the 3GPP default speech codecs. 3GPP SA4 have created a work item in 3GPP Release 6 on "Performance characterization of default codecs for PS conversational application". France Telecom R&D and Siemens proposed a test framework consisting of an UMTS simulator for the air interface and an IP network simulator. ITU-T recommendation P.800 [1] is used for the quality estimation of the transmission. The real-time conversational test results show that the AMR-NB and AMR-WB speech codecs are well suited for PS conversational applications. Furthermore, the results clearly show a higher understanding when using AMR-WB instead of AMR-NB.

1. INTRODUCTION

3GPP SA4 standardized the AMR-NB (Adaptive Multi Rate Narrow Band) and the AMR-WB (Adaptive Multi Rate Wide Band) speech codecs at 8 and 16 kHz sampling frequency respectively for speech conversational service. In the first part of the characterization phase of testing, the quality of these codecs was evaluated in circuit-switched transmission at various channel conditions (3GPP TR 26.975 & 26.976 [2]).

With the introduction of PS networks in mobile telephony, e.g. IMS (IP Multimedia Subsystem), there is a need to evaluate the speech transmission quality when using these codecs with IP/UDP/RTP protocols. 3GPP SA4 have created a work item in 3GPP Release 6 on "Performance characterization of default codecs for PS conversational application". France Telecom R&D and Siemens proposed a test framework consisting of an UMTS simulator for the air interface and an IP network simulator for the transmission of the IP packets on the Core Network (3GPP-SA4 S4-030564/65 [2]). ITU-T P.800 [1] is used for the estimation of the quality of the transmission.

In this paper we present and analyze the listening test results conducted in French language for clean speech conditions. Section 2 is dedicated to the presentation of the conversational test method. The overall simulator is described in Section 3. The test results are discussed in Section 4.

2. CONVERSATIONAL TEST PROTOCOL

The protocol described below evaluates the effect of degradation such as delay and dropped packets on the quality of communication. It corresponds to the conversation-opinion tests recommended by the ITU-T P.800 [1]. Contrary to listening tests, conversation-opinion tests are suited to assess the effects of impairments that can cause difficulty while conversing (such as delay).

Two acquainted subjects, seated in separate sound-proof rooms, hold a conversation for around 3 minutes with the support of a pretext given by the coordinator. The pretexts set for this protocol are those developed by the Ruhr University (Bochum, Germany) within the context of ITU-T SG12 [3]. These scenarios have been developed to allow a balanced conversation between both participants and to stimulate discussion. They are derived from typical situations of every day life: railways inquiries, rent a car or an apartment, etc. 32 non-expert subjects (16 couples) were used in order to conduct a relevant statistical analysis of the data. Eighteen test conditions were evaluated as described in Sec. 3.1.

After each conversation (corresponding to one specific condition), the subjects have to judge the quality by ticking the appropriate box on 5-category scales corresponding to five different questions. These questions, extracted from the ITU-T recommendations P.800, P.830, P.831, P.832 [1, 4, 5, 6], enable to cover all aspects of speech quality

0-7803-8603-5/04/$20.00 ©2004 IEEE

in a conversational context (interaction, intelligibility, voice quality of the conversational partner):

Q1: How do you judge the global quality of the communication?

Q2: How do you judge the quality of the voice of your partner?

Q3: Did you have difficulties to understand some words?

Q4: How do you judge the conversation when you interacted with your partner?

Q5: Did you perceive any impairment (noises, cuts,...)?

Two statistical analyses are conducted on the five dependent variables (i.e. the measured criteria/questions). The first analysis consists in a Multiple ANalysis Of VAriance (MANOVA), which globally indicates the possible effect of the experimental factors (i.e. the various conditions) on the perceived quality, all the dependent variables considered. Then, a specific ANOVA is run on each dependent variable to test these effects on each subjective quality aspect separately.

3. TEST BED DESCRIPTION

The VoIP Terminal simulator software consists of an AMR-NB/WB codec operating in floating point version running in real-time. The sound pick-up and reproduction are done via sound cards included in the PCs. The conditions of usage are taken from the existing 3GPP specifications. The RTP payload is constituted following the IETF RFC 3267 [7] and 3GPP TS 26.235 [2]. To reflect the 3GPP decisions in TS 26.236 [2], the following options have been chosen:

- The bandwidth efficient mode is used.
- Only one speech frame is encapsulated per RTP packet.
- The multi-channel session is not used.
- Interleaving of codec frames is not used.
- The internal CRC is not used.

The overall simulation is performed using 5 PCs.

- PC 1 and 5: VOIP Terminal Simulator Software
- PC 2 and 4: UMTS Air Interface Simulator
- PC 3: Network Simulator Software (NetDisturb)

3.1. Basic Principles

The platform simulates a PS interactive communication between two users using PC1 and PC5 as their relative VoIP terminals. PC1 sends AMR-NB/WB encoded packets that are encapsulated using IP/UDP/RTP headers to PC5. In fact, the packets created in PC1 are sent to PC2. PC2 simulates the air interface uplink (UL) transmission and then forwards the transmitted packets to PC4 through PC3 that simulates the IP core network via an Ethernet interface. PC4 simulates the air interface downlink (DL) transmission and then forwards the packets to PC5. PC5 decodes and plays the speech back to the listener.

The test conditions are built as a combination of 3 parameters. Two of them, called "IP Packet Loss Ratio" and "Radio", are used for AMR-NB/WB tests. The parameter "IP Packet Loss Ratio" has 2 values: 0% and 3%. The parameter "Radio" has 3 values: 10^{-2}, 10^{-3} and 5.10^{-4} Block Error Rate (BLER) of radio frames. The third parameter depends on the speech coder and on specific implementations. For AMR-NB, the third parameter is called "Mode + Delay" and defines the combination of the mode (the codec bitrate) and the end-to-end delay. This parameter has 3 different characteristics: 6.7 $kbit/s$ and 300 ms delay; 12.2 $kbit/s$ and 300 ms delay; 12.2 $kbit/s$ and 500 ms delay. For AMR-WB, the third parameter is called "Mode"; it combines the speech coder and the ROHC (RObust Header Compression) implementation. It has 3 different values: 12.65 $kbit/s$; 12.65 $kbit/s$ and ROHC; 15.85 $kbit/s$ and ROHC.

3.2. Air interface simulator

Siemens and RWTH Aachen developed a real-time system for simulating the UL/DL transmission of IPv6/UDP/RTP packets containing AMR-NB/WB speech frames over the UMTS air interface on a Linux platform. An algorithm for header compression according to IETF RFC 3095 [7] (ROHC) may be used optionally. Main part of the air interface simulator is an implementation of the Radio Link Control (RLC) protocol for assigning IP packets to radio frames. The underlying *physical layer* has previously been simulated offline for different radio channel qualities, and the resulting block error patterns are inserted within the real-time simulation.

RLC and RAB (Radio Access Bearer) settings

The following PS RAB (from 3GPP TS 34.108 [2]) has been used for air interface simulations: *Conversational/ Speech/UL:46 DL:46 kbps/PS RAB*. The RAB specifies the use of a PDCP header, RLC in unacknowledged mode, MAC in transparent mode, transport block size of 928 bits for each RLC PDU, Turbo Coding, and 16 bit CRC.

In RLC *unacknowledged mode* any residual bit errors in an RLC PDU lead to the discarding of this PDU. There are no retransmissions. Each TTI (transmission time interval) of 20ms, available IP packets are segmented/concatenated and placed into RLC PDUs of a fixed size, defined by the specified RAB. Appropriate RLC headers are added, containing a sequence number for detecting discarded PDUs and length indicators defining the packet boundaries. In the regular case, one IP packet is placed into an RLC PDU that is filled up with padding bits. Due to delayed packets there may also be more than one IP packet in the RLC transmission buffer to transmit in the current TTI.

After forming, the RLC PDU is passed to the error insertion block that decides if the transmission over the air

interface has been successful or if the PDU has to be discarded due to residual errors after channel decoding. The error insertion is based on error patterns consisting of binary decisions for each transmitted RLC PDU, resulting in a certain BLER (block error rate). Successfully transmitted PDUs are evaluated to reassemble the IP packets. A discarded PDU will result in none, one, or more lost IP packets, resulting in a certain packet loss rate of the IP packets and thereby in a certain FER of AMR-NB/WB frames. The reassembled IP/UDP/RTP/AMR-NB (or AMR-WB) packets are transmitted to the next system block.

Physical layer simulations

The UMTS physical layer simulations for the offline generation of the UL and DL block error patterns were carried out following the parameters of the conformance tests for base stations (3GPP TS 25.141 [2]) and user equipment (3GPP TS 34.121 [2]). Channel coding was performed by the rate-1/3 Turbo code with the decoder using the Log-MAP algorithm and 4 iterations.

As test scenario for the physical layer the "outdoor to indoor and pedestrian test environment" defined in [8] was selected. This test environment is characterized by a channel impulse response model based on a tapped-delay line with four taps. The root mean square of the delays is 45 ns. Each tap exhibits a classic Doppler spectrum. A walking speed of 3 km/h is assumed.

Block error patterns of length 15000, corresponding to five minutes of transmission, were generated separately by two different simulations for UL and DL. Transport blocks were declared erroneous, i.e., RLC PDUs were declared missing, if the CRC at the receiver failed. At a BLER of 10^{-2}, burst errors of transport blocks up to a length of five blocks occurred due to the fading characteristic of the channel. The generated block error patterns were used in a wrap-around mode in the real-time simulator.

3.3. IP network simulator

The IP network simulator is designed to apply parameters which are inherent to IP transmission such as packet loss and transmission delay. Packet loss can be associated to a law such as uniform law or any other computed law. In the same way the delay parameter can be a fixed value or follow a given law.

The IP network simulator is a PC including two LAN 10/100 Mbps cards. The software generating the IP transmission impairments is able to generate independently the parameters for each direction path. On both links, one can choose delay and loss laws.

However, for these conversation tests, only the parameter packet loss is variable. The delay parameter is a fixed value for each condition and only two values are used. For each test condition, the settings are similar for the two transmission paths.

4. LISTENING TEST RESULTS

The MOS results obtained by the listening tests (3GPP-SA4 S4-030806 [2]) were studied by two statistical analyses as described in Sec. 2. Correlations were computed between all subjective criteria to analyze their importance and dependencies. These correlation coefficients are all significant and rather well correlated (correlation coefficients for AMR-WB are lower than for AMR-NB). Therefore, the effects obtained with the four other criteria are rather similar to those obtained with the *Global quality* criterion. In particular, the *Voice quality* criterion confirms the observations made for the *Global quality*.

For AMR-NB the effect of the "Radio" parameter (BLER) is statistically significant for all criteria. However, with AMR-WB this statistical significance can only be observed for the *Understanding* and *Impairment perception* criteria. For the *Global quality* and the other criteria the effect is not (or weakly) statistically significant. The analysis of the "Mode+Delay"/"Mode" parameter reveals no (or a weak for AMR-NB) statistical significance for the criteria. The effect of the "IP Packet loss ratio" is statistically significant for AMR-NB and AMR-WB. The Mean Opinion Scores (MOS) obtained for 3 % of IP packet loss ratio are systematically less than those obtained for 0 % of IP packet loss ratio, independent of the other parameters.

Thus, the subjects were mainly sensitive to the packet loss ratio. No quality differences were perceived between the different modes, and the quality differences perceived between the different radio conditions are weak.

4.1. AMR Narrowband

Figures 1 and 2 depict the MOS results for the *Global quality* and *Understanding* criteria for AMR-NB. The poorest radio condition significantly influences the *Global quality*, independent of the speech coder and of the delay. The *Global quality* at 12.2 $kbit/s$ with 300 ms delay is scored slightly less for radio condition 5.10^{-4} than for radio condition 10^{-3}. Nevertheless, the results are in confidence interval and moreover the radio conditions are close to each other. The same effect is visible for all criteria.

Note, due to the relatively high delay ($\geq$ 300 ms), the scores for the *Interactivity* criterion were always below 4, independent of the test condition.

4.2. AMR Wideband

Figures 3 and 4 show that the IP packet loss ratio seems to affect the quality mainly for the radio conditions 10^{-2} and 5.10^{-4} (especially for 12.65 $kbit/s$). Considering the *Global quality* criterion, the MOS is always equal or better than 4 for 0% IP packet loss ratio. For 3% IP Packet loss

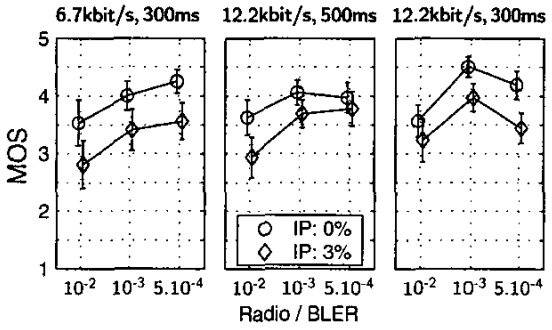

Fig. 1. AMR-NB: *Global quality* criterion MOS for various Radio conditions, Mode+Delay and IP Packet loss

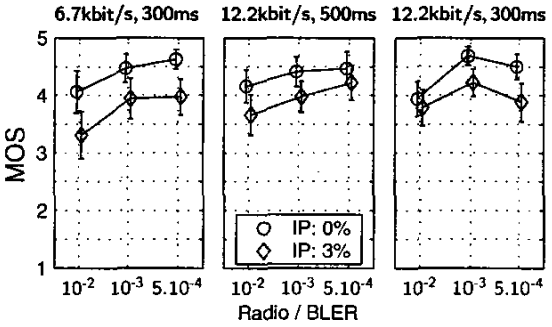

Fig. 2. AMR-NB: *Understanding* criterion MOS for various Radio conditions, Mode+Delay and IP Packet loss

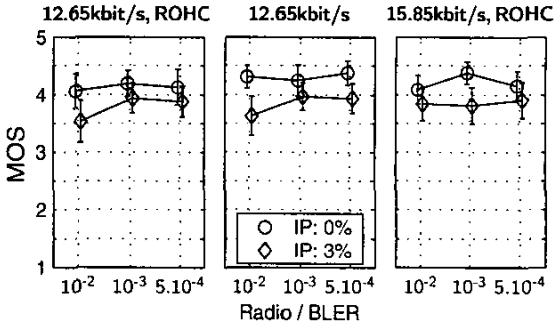

Fig. 3. AMR-WB: *Global quality* criterion MOS for various Radio conditions, Mode and ROHC

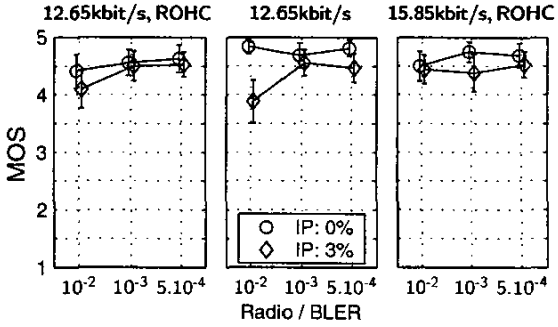

Fig. 4. AMR-WB: *Understanding* criterion MOS for various Radio conditions, AMR Mode and ROHC

ratio, the MOS is between 3.5 and 4, whatever the mode and radio conditions.

The mean scores observed for the *Understanding* criterion in Fig. 4 are very high for most test conditions, showing that the use of AMR-WB instead of AMR-NB increases significantly the understanding, even with impairments.

From the test results it seems that the end-to-end delay does not significantly impact the *Interactivity* criterion in the tested conditions. The mean MOS values for this criterion are mostly a little above 4.

5. CONCLUSION

The real-time conversational test results show that the AMR-NB and AMR-WB speech codecs are well suited for PS conversational applications.

For the narrowband and wideband conditions under test, the subjects were mainly sensitive to the IP packet loss ratio, whatever the mode, the delay and the radio conditions considered. No significant quality differences were perceived between the different modes (coder and ROHC), and the perceived quality is not significantly influenced by the different radio conditions.

Even if the AMR-NB and AMR-WB have not been tested in the same test sequence, it can be noted that the MOS scores collected in wideband are higher than the ones in narrowband. In particular, it can be noted that the understanding criterion is scored very high in wideband tests. The defaults perception criterion shows a similar result, even if it is not so evident.

6. REFERENCES

[1] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, August 1996.

[2] 3rd Generation Partnership Project (3GPP), *www.3gpp.org*.

[3] ITU-T SG 12 COM12-35, *Development of scenarios for short conversation test*, 1997.

[4] ITU-T Recommendation P.830, *Subjective performance assessment of telephone-band and wideband digital codecs*, 1996.

[5] ITU-T Recommendation P.831, *Subjective performance evaluation of network echo cancellers*, December 1998.

[6] ITU-T Recommendation P.832, *Subjective performance evaluation of hands-free terminals*, May 2000.

[7] Internet Engineering Task Force (IETF), *www.ietf.org*.

[8] ETSI TR 101 112, *Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03)*, April 1998.