New developments in VoIP testing

N. Rohrer, S. Poschen, F. Kettler

HEAD acoustics GmbH, 52134 Herzogenrath, Germany

Abstract

The migration of conventional telephone networks to packet based networks is an ongoing process. More and more telephone carriers migrate their networks to IP based technologies. This happens on the one hand in the backbones without knowledge or interaction of customers, but on the other hand also at the end users homes with VoIP gateways integrated in DSL routers or VoIP capable phones. Besides this process company in-house telephony networks also migrate from dedicated telephone lines to IP integrated solutions with VoIP capable phones at the user's side. In contrast to the benefit of lower costs and integration of additional data services like phone books new challenges arise. Therefore it is necessary to adapt measurement methods to cover all upcoming problems.

Packet Loss and Jitter

Packet based networks as they are used for VoIP transmission are subject to packet loss and jitter. These impairments affect listening quality of the transmitted speech and background noise. Methods to prevent or cover quality degradation are packet loss concealment and jitter buffers. As buffering the jitter is a less complex problem, the coverage of packet loss might lead to serious difficulties. Therefore in most cases packet loss concealment is an integral part of a speech codec or additional reference implementations are published. If the packet loss concealment is not part of the codec, it may sometimes be implemented in a non-uniform way or not at all. This happens due to hard- or software constrains in the telephone devices.

In order to get an impression of the robustness of the VoIP device against network impairments it is necessary to test the equipment under realistic and reproducible test conditions. These measurements result in a quality rating which allows a comparison to other devices or implementations.

Test scenarios

The test scenarios must cover a wide range from a best case to a worst case scenario. Also they must be reproducible to allow retesting or to document progress in packet loss concealment development.

A best case test scenario simulates a clean network without packet loss and jitter, which means a fully undisturbed and clean network with sufficient bandwidth. The worst case test scenario depends on the situation under which a device might be used. For example a phone which is intended to be used in a managed LAN will have fewer problems with jitter or packet loss than a customer premises device which is connected via a low dimensioned DSL connection.

Some representative test scenarios were chosen by ETSI for the ETSI SQTE [1] in order to test VoIP devices under widespread but defined conditions. The packet loss varies between 0% and 5% in these tests. In one test an additional jitter with an average of 20 ms is used.

Statistics vs. Pattern

During the 5th ETSI SQTE packet loss and jitter was inserted in the IP network by the statistical based NISTNet tool [2] and monitored in order to ensure the accuracy and distribution of impairments. Using this statistical approach, several repetitions of each measurement are needed. The statistical distribution of packet loss does not necessarily lead to an accurate number of lost packets during each test run. Additionally, these tests are not exactly reproducible, because it obviously makes a difference if the packet loss affects speech or silence parts. In order to balance this problem, the measurement duration must be long enough to compensate the statistical effect.

Another possibility is to run the same test with the same pattern of packet loss and jitter all the time. Appropriate tools like NetEM [3] might be used which allow the application of previously stored impairment patterns. In order to ensure that the same impairment occurs at the same part of speech or silence all the time, the start of the pattern must be synchronized to the beginning of the speech transmission in the measurement.

This leads to several problems as the measurement computer and the computer running the network impairment tool are very often two devices which must be synchronized. Also the test signal must be fed into a running RTP stream. If the reference gateway is not identical with the measurement computer this leads to the problem that the speech is not synchronized to the packets in the stream. The result is that the speech might start in the beginning or at the end of a packet. This changes the influence of a lost packet on speech quality.

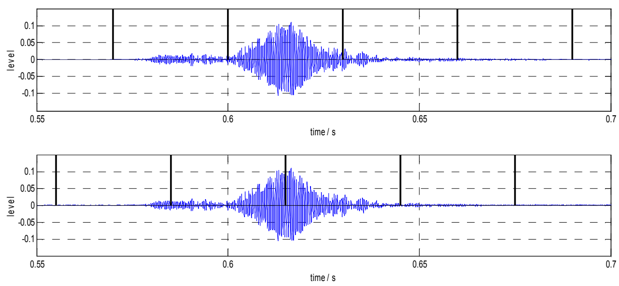

Figure 1: Packeting of a speech sample with different time offsets

Figure 1 illustrates this problem: A speech sequence is divided into packets by the vertical black lines each 30 ms. In one case, the main part of the speech, starting at 0.6 seconds, is packed into one packet. In the other case the main part of the speech is spread over two packets.

It is obvious that a loss of the packet containing the main part of the speech (upper signal) is more disturbing than losing a packet containing only half of the speech information. Furthermore a packet loss in the beginning of a speech phase is more difficult to interpolate or substitute by packet loss concealment algorithms than a packet loss in the middle of the speech phase, when the previous packet carries already important speech information. Interpolation is easier in this case. An ideal connection in a measurement system would synchronize the speech signal and the packet stream and thus guarantee the same speech content in the same packet. This is only possible in sending direction of the synchronized reference gateway.

Another criterion for realistic and efficient objective measurements is the choice of appropriate patterns. Figure 2 shows an example of two different patterns providing the same packet loss rate. Packet loss which is marked by the vertical red lines is distributed more across the speech parts in the upper plot and more across the silence parts in the lower plot. The first sentence in the lower plot is nearly not affected by packet loss, whereas it is significantly disturbed in the upper plot. This will obviously lead to different MOS-LQO results.

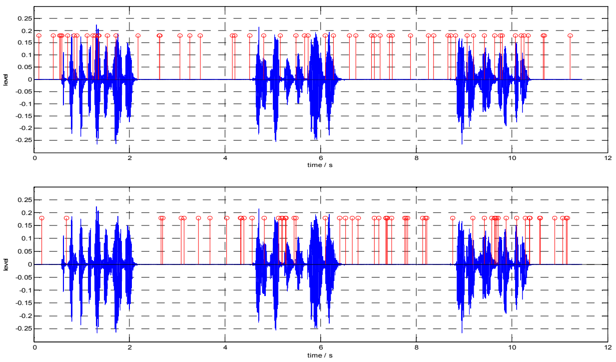

Figure 2: Examples for different packet loss distributions with same packet loss rate

The values in table 1 show the influence of different patterns on the results of a listening speech quality test. All values are averaged over 6 measurements applying speech sequences of 32 s each with a speech activity of 50% for each sequence; the total analysis time amounts to 192 s. The packet loss rate was 6% in all test cases and a G.722 codec was used on VoIP demo equipment.

Pattern	Statistic	A	B	C
MOS-LQO	3.3	3.1	3.3	3.5

Table 1: MOS-LQO results for different packet loss distributions

The statistic value is best reproduced by pattern B. Pattern A distributes the packet loss more across active speech parts. The resulting MOS-LQO values are therefore lower than for the statistical approach. Pattern C concentrates the network impairments more to the silence phases of the used test signal. Consequently the resulting listening speech quality expressed by the MOS-LQO value is higher.

Reproducing real networks

The behavior of a real network can be simulated in laboratory tests by monitoring packet loss and jitter in the target network under different conditions (e.g. traffic, load, different services running…). Effects like packet loss distribution or burst errors can be modeled in the network simulation tool using specific patterns. To prevent potential problems resulting from an unbalanced pattern which could for some reasons interfere with the strengths or weaknesses of the implemented packet loss concealment algorithms, different patterns should be created and used to calculate an average value. If the test is reproducible in a sense that packeting, triggering of impairment pattern and the pattern are weighted, tests must not be repeated because in each test the specified packet loss or jitter rate is strongly fulfilled.

Also the influence of network components in the test network must be considered. Figure 3 shows an example for a wrong dimensioned PPPoE server in a DSL test network. The upper graph shows the delay between the incoming IP packet and the corresponding PPPoE packet measured on the network ports of the server. The processing time of the PPPoE server, expressed by the delay on the y-axis, is not stable. Note that the delay is scaled logarithmically to get a better impression of the problem. The processing time is in a range between 200 ms and 10 s. This is an extreme example but shows the necessity to check all network components. The resulting jitter shown in the lower diagram is unacceptably high to be used in a test network. A simulation of specific network conditions is not possible if such network components are used. This network problem would dominate all other simulated network impairments.

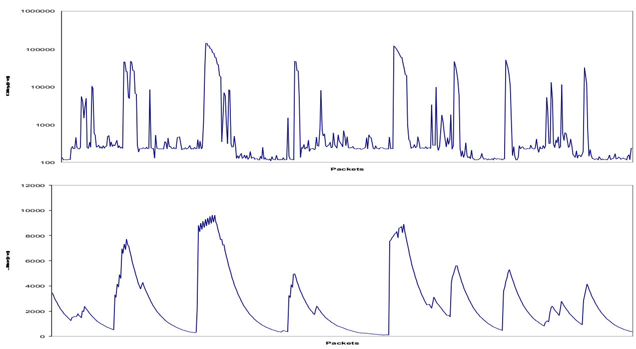

Figure 3: Influence on delay and jitter of an overloaded PPPoE server in a test network

Actual gateways

As an example for listening speech quality measurements under different network conditions figure 4 shows the measured MOS-LQO values for VoIP gateways from different manufacturers measured during the 5th ETSI SQTE. The codec used for this test was G.711 [4] with a packet length of 20 ms. Besides the average values the minimum and maximum values are also given. It can be seen that the resulting MOS-LQO values are widely spread. This means that the packet loss concealment is implemented in different ways and G.711 Appendix I [5] is not used in all gateways. The jitter buffer test (last test condition on the x-axis) shows, that some gateways cannot handle jitter appropriately. Other implementations handle jitter in a way that the same MOS-LQO result as without jitter can be measured. Only the one-way delay is higher under the jitter condition resulting from the increased jitter buffer size.

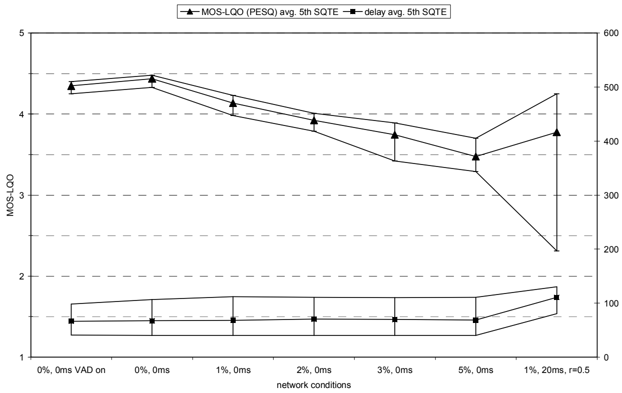

Figure 4: Minimum, maximum and average MOS-LQO results (upper curves) and one-way delay (lower curves) for several VoIP gateways under different packet loss and jitter rates.

Wideband

VoIP phones today often provide wideband speech transmission capability. On the network provider side the networks are upgraded to wideband capability. This is realized by several wideband codecs like G.722 [6], G.711.1 [7] for VoIP and AMR-Wideband (G.722.2 [8]) for mobile phones. Besides the software, the hardware such as speakers, microphones and housings must also fulfill new requirements to transmit wideband speech signals. As wideband capable hardware becomes more and more available, new testing methods are required. Besides the extension of the frequency range from 300 Hz to 3400 Hz in the narrowband case to 50 Hz to 7 kHz in wideband scenarios other problems arise.

Hardware requirements

In a wideband capable phone speaker and microphone must be designed to transmit frequencies between 50 Hz to 7 kHz. This leads to new challenges for the acoustical design especially if the housing of the speaker is limited in size due to design constraints. Standardization and test specifications need to apply new requirements for testing. New tolerance schemes for wideband phones must be defined. Figure 5 shows a tolerance scheme for measurements in receiving direction for narrowband and wideband terminals together with exemplary measurement results. The tolerance scheme for narrowband is defined in [9]. It can be seen that the wideband tolerance scheme as suggested in [11] is not only an extension of the tolerance scheme that is used for narrowband phones, but shows completely new characteristics. Contrary to the narrowband measurement which uses linear equalization and DRP/ERP correction the measurements for wideband phones in receiving direction are conducted with diffuse field equalization. This leads to more reproducible results if different artificial head measurement systems are used (Head and Torso Simulators acc. to ITU-T P.58 [10]).

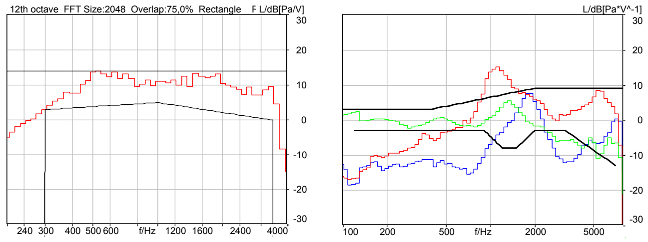

Figure 5: Tolerance scheme with measurement result for narrowband (left) and wideband phones (right) in receiving direction

Hardware of early-to-market wideband phones is often derived from narrowband phones. The phone software is extended with wideband codecs. Typical problems of the hardware that was not intended to be used for wideband can be seen in figure 6. Four different frequency responses of existing wideband phones show the different problems. The left frequency responses are measured on phones which are equipped with wideband capable hardware. The frequency responses on the right show typical results of narrowband hardware used in wideband environment. The upper frequency response shows a strong attenuation towards high and low frequencies whereas the plot of the lower frequency response shows attenuation to low frequencies but no attenuation to higher frequencies.

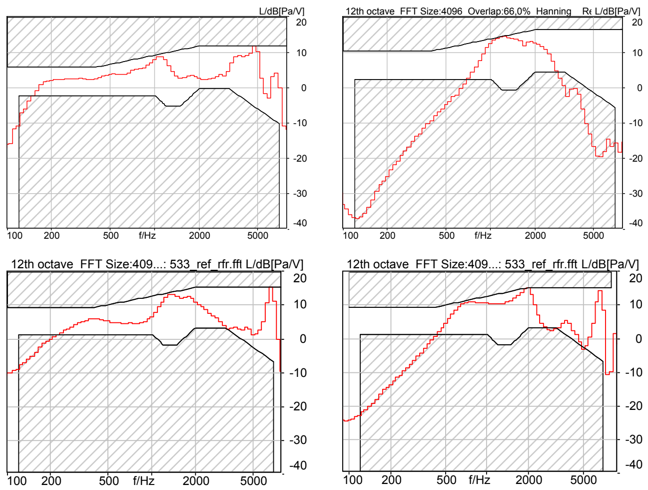

Figure 6: Different frequency response results for actual wideband phones in receiving direction

Echo Perception

Echo disturbances are annoying in telephone conversations. The sensitivity for echoes increases with higher round trip delays. In VoIP connections one-way delays in the range of 150 ms can be measured. Besides the delay, the frequency content of the echo signal also influences perception and annoyance [11]. The tolerances from narrowband scenarios cannot be applied or linearly extended for wideband communication. A suggestion for a new tolerance scheme for the spectral attenuation of wideband echoes considering the echo delay was made in [11]. Figure 7 shows the tolerance scheme for four different round trip delays.

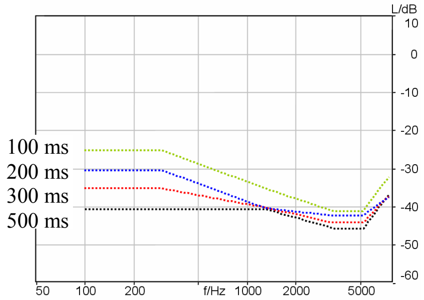

Figure 7: Tolerance scheme for spectral echo attenuation under consideration of echo round trip delay

Measurements are conducted with actual wideband capable VoIP phones in handset mode. The results are verified with the tolerance schemes for different round trip delays. The curves for two phones are shown exemplarily in figure 8 and 9.

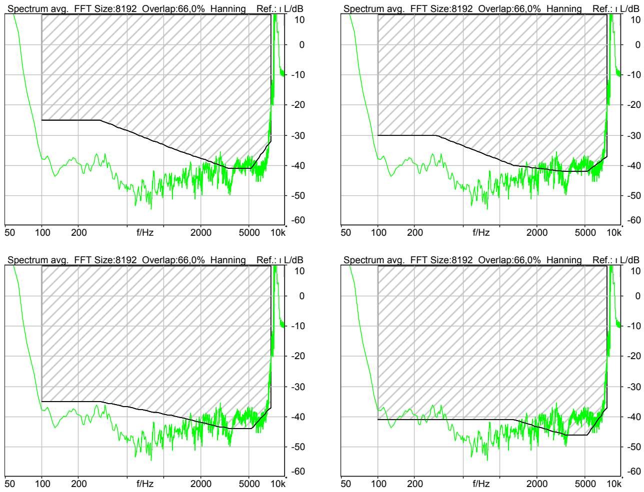

Figure 8: Tolerance schemes for different round trip delays with echo measurement result (100 ms: top left, 200 ms: top right, 300 ms: bottom left, 500 ms: bottom right)

The result in figure 8 indicates constant echo attenuation over the complete frequency range.

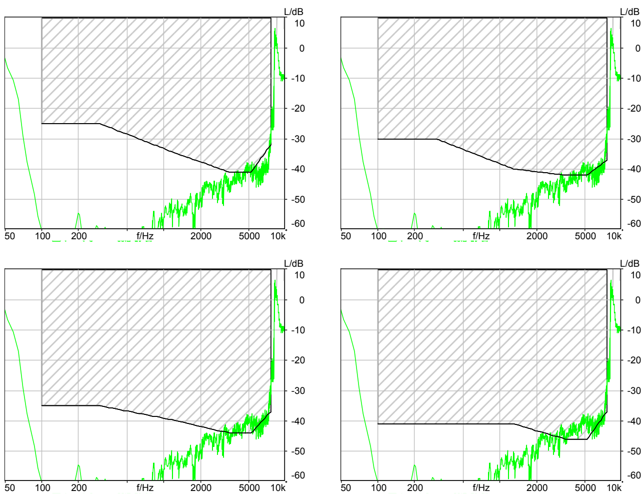

Figure 9: Tolerance schemes for different round trip delays with echo measurement result (100 ms: top left, 200 ms: top right, 300 ms: bottom left, 500 ms: bottom right)

The tolerance scheme for a 100 ms round trip delay is only slightly violated at high frequencies. Stronger violations occur for higher round trip delays. The results from the second phone in figure 9 show different characteristics and indicate a different AEC performance. Attenuation in the low frequency range is high. The upper frequency range does not seem to be sufficiently attenuated.

Conclusion

Testing of VoIP equipment under realistic conditions requires measurement methods which allow a reliable reproduction of different network conditions. If network impairment patterns are used, the synchronization of these patterns to the audio stream as well as a reproducible packeting of audio data are essential requirements.

The tendency to wideband speech transmission requires a new tolerance scheme for a wideband frequency response. Together with extending transmission delays in IP networks the need of new echo measurements increases. A new tolerance scheme for wideband echo considering different round trip delays addresses this problem. It also takes into account the higher sensitivity to echoes in frequency ranges around 3 kHz.

References

[1] Reference to the ETSI SQTE homepage
URL: http://www.etsi.org/plugtests/SQTE/SQTE.htm

[2] Reference to the NISTNet homepage
URL: http://snad.ncsl.nist.gov/nistnet/

[3] Reference to a NetEM homepage
URL: http://www.linuxfoundation.org/en/Net:Netem

[4] ITU-T Recommendation G.711 : Pulse code modulation (PCM) of voice frequencies; 11/1988

[5] ITU-T Recommendation G.711 : A high quality low-complexity algorithm for packet loss concealment with G.711; 09/1999

[6] ITU-T Recommendation G.722 : 7 kHz audio-coding within 64 kbit/s; 11/1988

[7] ITU-T Recommendation G.711.1 : Wideband embedded extension for G.711 pulse code modulation; 03/2008

[8] ITU-T Recommendation G.722.2 : Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB); 07/2003

[9] TBR 38: Public Switched Telephone Network (PSTN); Attachment requirements for a terminal equipment incorporating an analogue handset function capable of supporting the justified case service when connected to the analogue interface of the PSTN in Europe; 05/1998

[10] ITU-T Recommendation P.58: Head and torso simulator for telephonometry; 08/1996

[11] Wideband Echo Perception, S. Poschen, F. Kettler, HEAD acoustics GmbH, Herzogenrath, Germany
A. Raake, S. Spors, Deutsche Telekom Laboratories, Berlin, Germany, IWAENC, 09/2008, Seattle, WA, USA