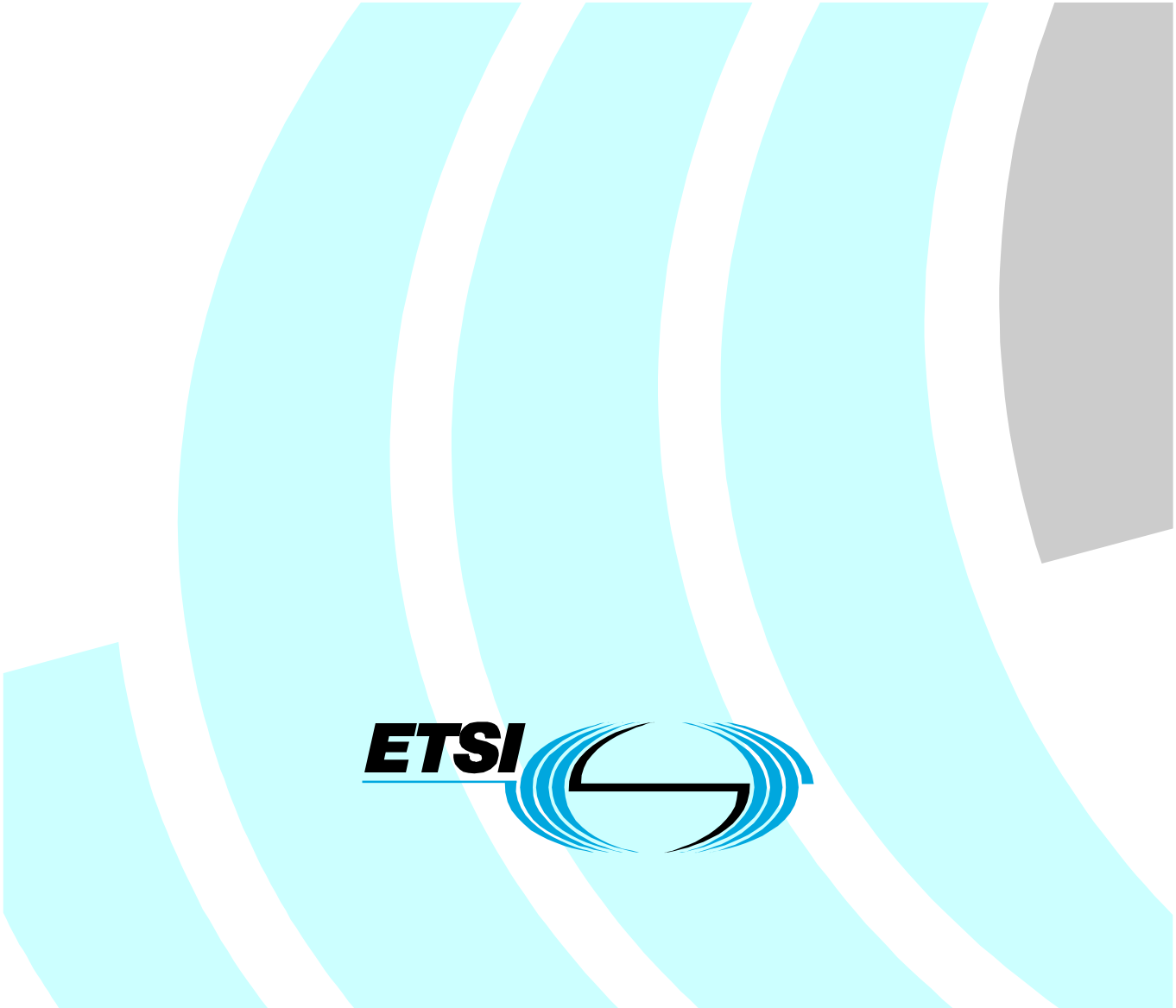
ETSI TR 102 648-2 V1.1.1 (2007-02)

Technical Report

**Speech Processing, Transmission and Quality Aspects (STQ);
Test Methodologies for ETSI Test Events and Results;
Part 2: 1st ETSI Plugtests Speech Quality Test Event Report**



Reference
DTR/STQ-00079-2

Keywords
interoperability, quality, speech, VoIP

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://webapp.etsi.org/IPR/home.asp).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Speech Processing, Transmission and Quality Aspects (STQ).

The present document is part 2 of a multi-part deliverable. Full details of the entire series can be found in part 1 [19].

1 Scope

The present document contains the anonymous Test Report from the 1st ETSI Plugtests Speech Quality Test Event.

2 References

For the purposes of this Technical Report (TR) the following references apply:

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

[1] ITU-T Recommendation P.800: "Methods for subjective determination of transmission quality".

[2] ETSI EG 201 377-1: "Speech Processing, Transmission and Quality Aspects (STQ); Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks".

[3] ITU-T Recommendation P.501: "Test signals for use in telephonometry".

[4] ITU-T Recommendation P.502: "Objective test methods for speech communication systems using complex test signals".

[5] ITU-T Recommendation P.58: "Head and torso simulator for telephonometry".

[6] ITU-T Recommendation P.57: "Artificial ears".

[7] ETSI TIPHON temporary document 17TD135: "Subjective and objective speech quality evaluation on speech data recorded at the SuperOp 99 event in Hawaii. Sophia Antipolis, March 2000".

[8] ITU-T Recommendation P.64: "Determination of sensitivity/frequency characteristics of local telephone systems".

[9] ITU-T Recommendation P.79: "Calculation of loudness ratings for telephone sets".

[10] ITU-T Recommendation G.122: "Influence of national systems on stability and talker echo in international connections".

[11] ITU-T Recommendation P.56: "Objective measurement of active speech level".

[12] ITU-T Recommendation P.830: "Subjective performance assessment of telephone-band and wideband digital codecs".

[13] ITU-T Recommendation P.810: "Modulated noise reference unit (MNRU)".

[14] 21TD68: "Proposal for 2nd Speech quality test event", Reinhard Scholl.

[15] 21TD95: "Preliminary test report", T-Nova Berkom & HEAD acoustics.

[16] 21TD101: "Test Spec", T-Nova Berkom & HEAD acoustics.

[17] 21TD116: "Preliminary test results: Explanation", T-Nova Berkom & HEAD acoustics.

[18] ETSI TS 101 329-5: "Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON) Release 3; End-to-end Quality of Service in TIPHON systems; Part 5: Quality of Service (QoS) measurement methodologies".

[19] ETSI TR 102 648-1: "Speech Processing, Transmission and Quality Aspects (STQ); Test Methodologies for ETSI Test Events and Results; Part 1: VoIP Speech Quality Testing".

[20] ITU-T Recommendation P.340: "Transmission characteristics and speech quality parameters of hands-free terminals".

[21] ETSI TBR 8: "Integrated Services Digital Network (ISDN); Telephony 3,1 kHz teleservice; Attachment requirements for handset terminals".

[22] ITU-T COM12-117E, March 2000.

3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AGC	Automatic Gain Control
ASL	Active Speech Level
CAS	Communication Analysis System
CSS	Composite Source Signal
ERL	Echo Return Loss
HATS	Head And Torso Simulator
IP	Internet Protocol
IRS	Intermediate Reference System
ISDN	Integrated Services Digital Network
JLR	Junction Loudness Rating
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
NOTE: Output of TOSQA.	
NIST	National Institute of Standards and Technology
OLR	Overall Loudness Rating
OVL	Over-Load Point
PBX	Public Branch Exchange
PLC	Packet Loss Concealment
PVS	PC Voice Switch
RLR	Receive Loudness Rating
RTP	Real time Transport Protocol
SLR	Send Loudness Rating
TMOS	TOSQA Mean Opinion Score
TOSQA	Telecommunications Objective Speech Quality Assessment
VAD	Voice Activity Detection

4 Summary

The European Telecommunications Standards Institute (ETSI) organized a special test event for VoIP (Voice over Internet Protocol) speech quality in Sophia Antipolis, France, from 23rd of October to 1st November, 2000. T-Nova Deutsche Telekom Innovationsgesellschaft mbH Berkom, in collaboration with HEAD acoustics GmbH, performed speech quality measurements on VoIP equipment of different manufacturers. Texas Instruments Incorporated and Alcatel co-sponsored the test event.

The aim of the test event was to determine the speech quality of various Voice over IP equipment under certain IP network conditions. During the test event, speech material as well as measurement data were collected by transferring voice samples and artificial signals across the Voice over IP setup. Speech quality was measured by both instrumental (objective) and auditory (subjective) methods. Both methods were used to measure the one-way speech quality (listening quality). The important transmission parameters determining conversational quality like double talk performance, background noise transmission and echo performance were accessed using sophisticated test signals and enhanced analysis methods as described in TS 101 329-5 [18] and recent ITU-T Recommendations.

The one-way speech transmission quality was evaluated by processing real speech samples and analysing it using the TOSQA algorithm. To validate the TOSQA algorithm, auditory reference tests according to ITU-T Recommendations of the P.800 series were carried out. Correlations of 91,6 % and 93,6 % for listening quality and connection quality, respectively, demonstrate the high accuracy of TOSQA for VoIP transmission scenarios tested here. A subset of speech recordings were carried out using the HATS (head and torso simulator) HMS II.3 of HEAD acoustics equipped with type 3.4 artificial ears. For these conditions a separate auditory test was conducted and the speech material was also assessed by the new version TOSQA2001 terminal extension. Here a correlation of 98 % was derived.

Instrumental measurements using sophisticated test signals and analysis methods according to recent ITU-T Recommendations of the P.500 series were conducted covering all conversational aspects like single talk and double talk periods or echoes. These tests are specially designed to analyse and optimize parameters determining conversational quality, quality of background noise transmission, the performance of echo cancellers and others.

These measurements were carried out at the acoustical interface using IP terminals or standard ISDN telephones mounted to the HATS and at the electrical interface for gateway testing. The results provide important information for the manufactures about conversational speech quality of their equipment. In particular, the tests determined parameters like:

- distortions, AGC (automatic gain control), VAD (voice activity detection) or PLC (packet loss concealment) implementations under single talk conditions;
- double talk performance influenced by level variations, clipping and echoes;
- echo canceller performance determined by convergence characteristics, spectral echo attenuation, NLP implementation;
- quality of background noise transmission, clipping, voice activity detection or the design of comfort noise injection.

Based on the results the following tests and test conditions for conversational speech quality are suggested for standardization.

Specific echo canceller tests for the VoIP equipment including low Echo Return Losses (ERL) of 6 dB (simulating worst case echo conditions in networks) and high ERL > 40 dB (simulating typical ISDN connections). On the one hand the implemented echo cancellers should guarantee a sufficient echo attenuation but on the hand the echo cancellers should not degrade the performance of the network for high ERL values in the echo path.

- These echo canceller tests should be carried out and analysed under single and double talk conditions.
- The occurrence of signal gaps (clipping) under double talk conditions should especially be tested under network condition including high ERL values. Again the implemented signal processing in VoIP equipment should not degrade the network performance if no packet loss and no delay jitter is introduced during the test.
- The quality of background noise transmission together with implemented comfort noise injection should be tested. The tests should determine the adaptation of injected comfort noise on the actual background noise level and spectrum.

The results of this test event are being published in the present document. Parts of it will be included in the document ETSI TIPHON 05013 TR 101 329-6 "Actual measurement test results". The report will also be presented at ETSI STQ and ITU-T Study Group 12. The data will provide input for new or enhanced standards and recommendations for enhanced VoIP communications. Furthermore, the results can be used for optimization of the manufacturers" VoIP equipment to improve the overall speech quality.

Due to the benefit of such an event it is strongly recommended to continue the process of end-to-end speech quality testing. To support this idea a second ETSI VoIP test event is currently being prepared and planned.

5 Overview

The present document describes the test methodologies, the assessment methods and the results of the measurements which were carried out during the 1st ETSI VoIP speech quality test event. The aim of the test event was to determine the speech quality of various Voice over IP equipment under certain IP network conditions. During the test event, speech material as well as measurement data were collected by transferring voice samples and artificial signals across the Voice over IP setup. This material was analysed and the results are reported in the present document.

The analysis of the collected data can be split in two parts. In the first part the assessment of the one-way speech quality (listening quality) was performed by both, auditory and instrumental assessments. In the second part the analysis of various transmission parameters, double talk performance and background noise transmission was performed and different transmission parameters were indicated.

The one-way speech transmission quality was evaluated by processing real speech samples and analysing it using the TOSQA algorithm. TOSQA leads to MOS-comparable results. To validate the TOSQA results an auditory reference test was carried out.

A detailed description of the relationship of a reference MOS evaluation according ITU-T Recommendation P.800 series Recommendations and the relating TOSQA results is given.

In the second main part of the document, measurement results based on recent ITU-T Recommendations (P.500 series, P.340 [20]) are included. These measurement results provide information about various transmission parameters from which double talk performance and background noise transmission of the tested Voice over IP equipment can be derived.

All measurements performed for one-way transmission speech quality at the electrical interface using speech signals were performed by T-Nova. The data acquisition for the evaluation of the one-way speech transmission quality at the acoustical interfaces was performed by HEAD acoustics. The assessment of all TMOS values were carried out by T-Nova. Also the auditory tests were conducted in the speech quality test laboratory at T-Nova in Berlin.

The data acquisition and the evaluation of the various transmission parameters, double-talk and background noise performance using artificial test signals were performed by HEAD acoustics, at the electrical interface as well as at the acoustical interface.

6 Test Description

6.1 General Test Description

Measurements were conducted with two measurement scenarios:

- Connection between the electrical network-interfaces at the used PBX ("electrical-electrical").
- Measurement at the acoustical interface:
  - Acoustical transmitter interface - Acoustical receiver interface ("Acoustical-Acoustical").
  - Electrical transmitter interface - Acoustical receiver interface ("Electrical-Acoustical").

A detailed diagram of the measurement scenarios can be found in clause 6.2. Measurements were executed using two kinds of input signals:

- Human speech samples (German or English language or both).
- Artificial test signals (according to ITU-T Recommendation P.501 [3]).

The description of the signals is included in clause 6.4 speech quality assessment of the speech samples was performed using the method for instrumental speech quality estimation of Deutsche Telekom, TOSQA. For verifying the instrumental results auditory reference assessments were performed using speech material in German.

The artificial test signals were used for measuring of transmission parameters according the ITU-T Recommendation P.500 series. These methods are mentioned in clause 6.5.

6.2 Measurement Scenarios

6.2.1 Measurements using Electrical Interfaces

6.2.1.1 Measurement Setup

For the "electrical-electrical" measurements two kinds of input signals were used:

a) speech samples designed according to ITU-T Recommendation P.800 [1]; and

b) test signals according to ITU-T Recommendation P.501 [3].

The input signals were transmitted and recorded simultaneously, i.e. the sending and receiving process were started at the same time. Therefore exact delay measurements were possible.

For all kind of measurements a packet loss generator (NIST Net) and a packet loss monitor were used. Both entities were controlled by the PVS measurement equipment. The measurement setup is shown in figure 1.

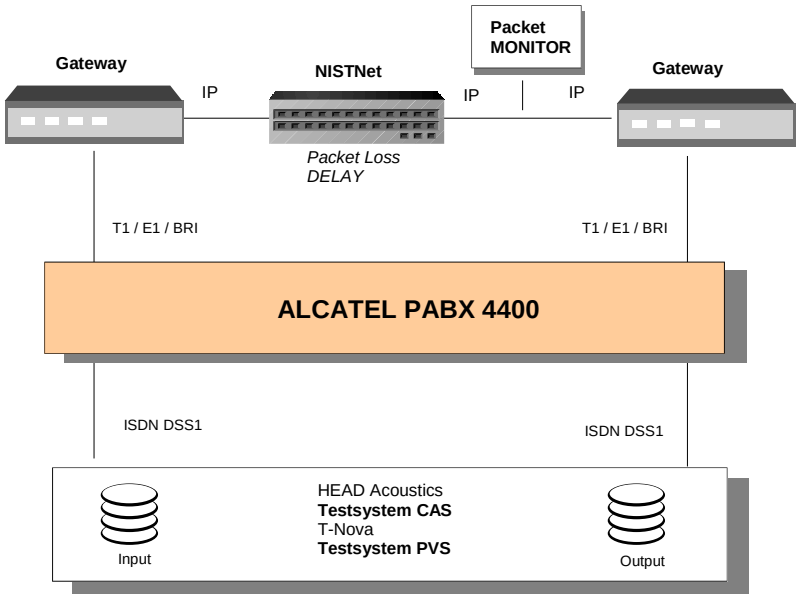


Figure 1: Electrical - Electrical Measurement Setup

6.2.1.2 Measurement Conditions

For all kinds of "electrical-electrical" measurements the following IP network conditions were used.

Table 1: Network Conditions for Electrical-Electrical Measurements

Condition	Packet Loss (Equal)	Additional Delay (see note 1)	Delay Variation
1	0	0	No
2	1 %	0	No
3	2 %	0	No
4	3 %	0	No
5	5 %	0	No
6	1 %	50 ms	20 ms
NOTE 1: Additional IP network delay was introduced by NIST Net. NOTE 2: Additional IP network delay was introduced by NIST Net.			

The additional delay in condition 6 was intended to ensure proper jitter (delay variation) generation by NistNet.

In such jitter condition the test network can cause situations where packets are reordered, if the packet size is very small. This effect can be avoided with high probability by using a certain packet length, which should be at least three times higher than the delay variation itself.

6.2.2 Measurements using Acoustical Interfaces

6.2.2.1 Measurement Setup

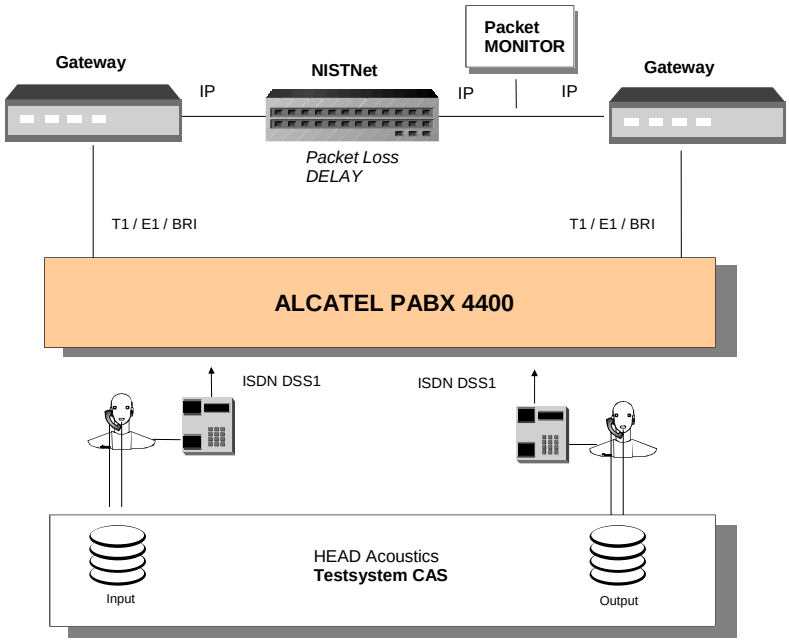


Figure 2: "Acoustical - Acoustical" Measurement Setup with Reference ISDN Terminals

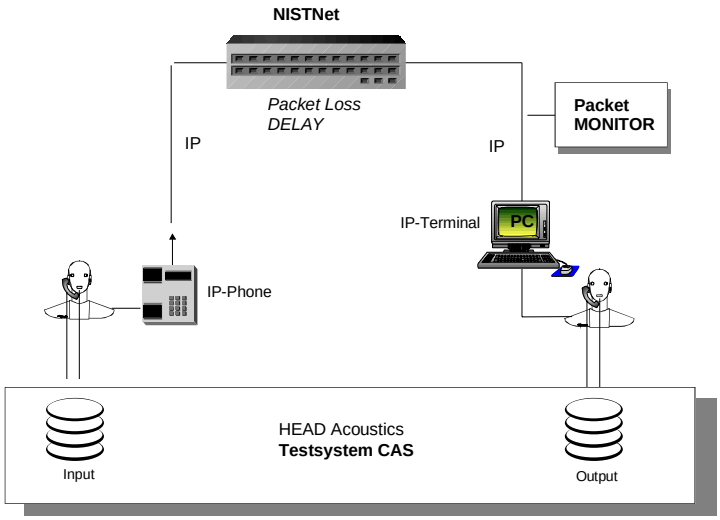


Figure 3: "Acoustical - Acoustical" Measurement setup with IP Terminals (handsets or headsets)

The reference terminals provided were standard digital handset terminals according to TBR 8 [21]. For all kind of measurements a packet loss generator and a packet loss monitor were included in the setup.

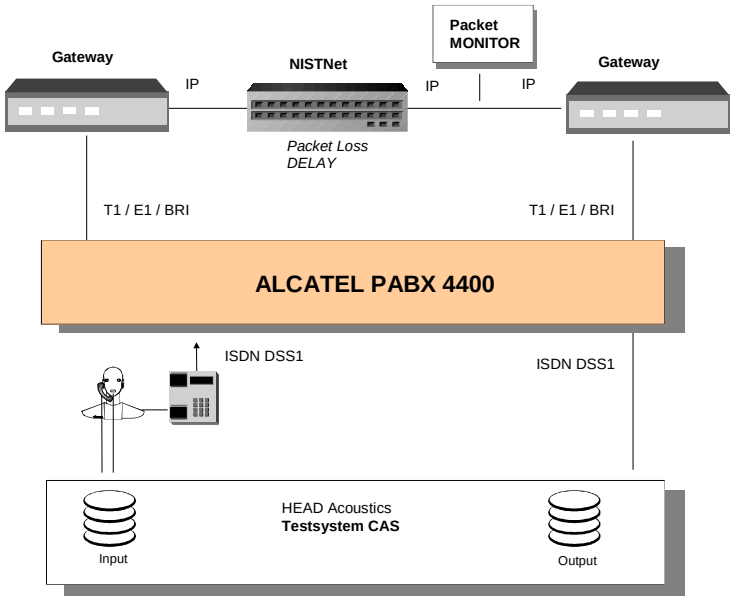


Figure 4: Measurement setup "Acoustical - Electrical" for Gateway to Gateway Configuration

For the tests the handsets of the terminals are applied to the HATS using the positioning as described in ITU-T Recommendation P.64 [8] with defined pressure force. The test sequences, natural speech as well as the artificial sequences were automatically introduced and recorded by the test system CAS and stored on hard disc.

Instead of the PABX telephone an IP telephone was used if provided in combination with a gateway which interfaces to the PABX. This is shown in figure 5.

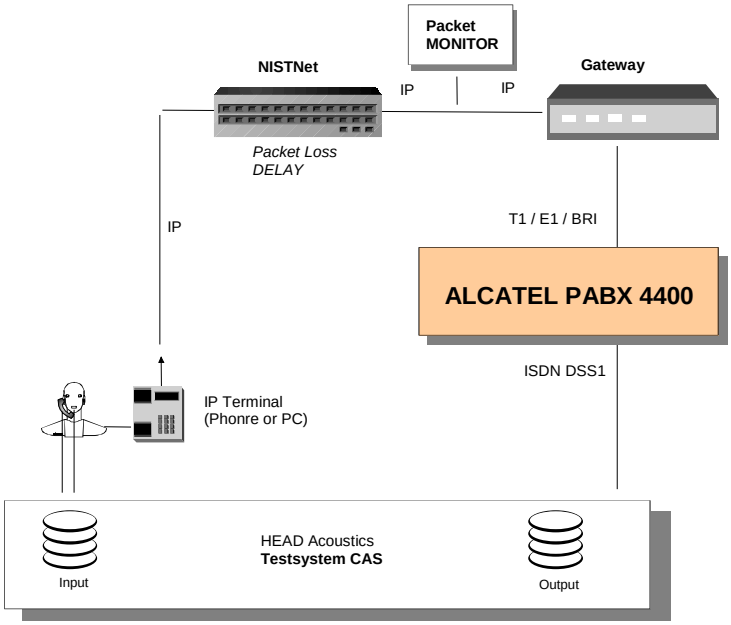


Figure 5: Measurement setup "Acoustical - Electrical" for IP-Terminal to Gateway Configuration

6.2.2.2 Measurement Conditions

The IP network conditions for all kinds of acoustical measurements ("electrical - acoustical" and "acoustical - acoustical") were:

Table 2: Network Conditions for All Kinds of Acoustical Measurements

Condition	Packet Loss (Equal)	Additional Delay (see note)	Delay Variation
1	0	No	No
2	0	100 ms	No
3	0	100 ms	20 ms
4	1 %	100 ms	No
5	1 %	100 ms	20 ms
6	3 %	100 ms	No
NOTE: Additional IP network delay was introduced using NIST Net.			

Also for all kinds of acoustical measurements it was recommended to use a packet-length of at least 60 ms for audio frames (cf. electrical-electrical measurements). This packet length had no influence on auditive or instrumental assessment methods.

6.3 Measurement Methodology

This clause describes the measurement procedure in detail, especially the verification procedure that specific adjustments on the behaviour of IP network were achieved.

As displayed in several figures of clause 6.2 a real time network simulator (NIST Net) was used to generate the IP network conditions. In terms of packet loss generation this network simulator uses a shaped random number generator to drop packets. To achieve the given percentage of packet loss, a fairly large number of packets need to cross the NIST Net device. If for example the packet loss rate is configured to 1 % the NIST Net device would need to drop one packet out of 100. But because of the underlying random number generator it may occur that out of 100 packets 0 or 2 packets will be dropped. The configured packet loss rate will actually be achieved just after a large number of packets (> 1 000) with a certain maturity.

To estimate the speech quality several voice samples were processed across the VoIP scenario. To assess the speech quality,4 voice samples of 8 seconds speech are used according to ITU-T Recommendation P.800 series. Using a certain voice codec, e.g. G.723.1 (one frame per packet) one second speech leads to 33 packets. If VAD is used and one takes the structure of the voice samples into account (50 % voice activity) an 8 seconds speech sample leads to about 130 packets. To achieve 1 % packet loss during this 8 second speech sample the NIST Net device would need to drop exactly 1 packet (0,76 % packet loss). Because of the random nature of the packet loss generation it will occur that 3 or more packets are dropped, even if the NIST Net is configured to 1 % packet loss. A drop of 3 packets out of 130 corresponds to a packet loss rate of 2,3 %. This would actually apply for condition C2 (2 % packet loss) to the current speech sample and would lead to wrong interpretation of speech quality results.

Therefore it is necessary to control the actual packet loss rate during the measurement process. Consequently, a packet monitor as shown in all figures of clause 6.2 was introduced. The packet monitor observes the network and calculates the actual packet loss rate based on RTP sequence numbers. This device is remotely controlled by the T-Nova measurement equipment PVS which itself transmits and receives simultaneously the voice samples using the electrical (ISDN) interface across the system under test. After transferring a voice sample, the PVS system checks the real packet loss rate and if necessary it repeats the same speech sample several times until the required packet loss rate is encountered.

Using 4 different voice samples (according to ITU-T Recommendation P.800 [1], 8 seconds each) for one particular condition leads to reasonably accurate results for time invariant voice transmission systems (e.g. PBX test, codec test). Because of the nature of VoIP technology the VoIP transmission system is time variant, especially in cases of packet loss. Even if it could be ensured that the number of packet losses matches the packet loss rate (by checking the packet losses using the packet monitor), the system is nevertheless time variant in terms of the position of packet losses in the speech signal. To avoid those influences, it is necessary to transmit more speech samples in order to compensate this effect on average. A number of 16 speech samples, still 8 seconds voice each, leads to reasonable results and increases the maturity of the results.

During this test event both approaches were used, the packet monitor and the increased number of speech samples.

6.4 Test Signals

6.4.1 Voice Signals

The speech samples were provided in two languages, German and English. Each sample contains two short concatenated sentences read by the same speaker. For both languages a set of four speech samples spoken by four different speakers (two male and two female each) were prepared. Each of these speech samples are 8 sec in length and contain about 50 % speech activity.

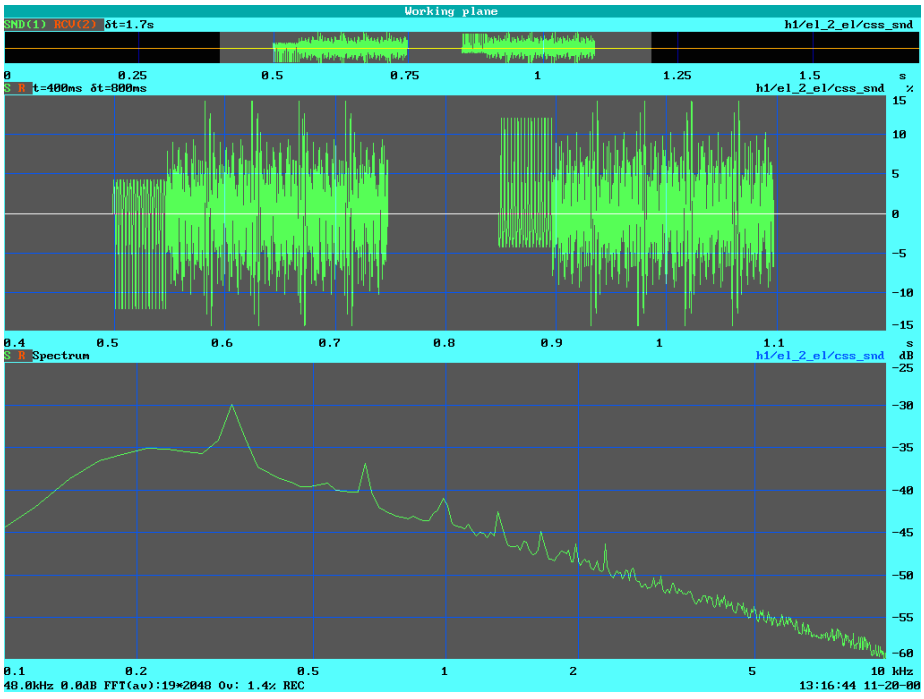
In case of the German speech material each sentence was used only once because of the usage in the auditory reference test. The four English speech samples are the same for all elaborated conditions in order to achieve easy comparisons with non-expert listeners between different test conditions. In addition one German speech sample was processed across all test conditions 8 times. In total 16 speech samples were transmitted in each condition.

All speech samples were pre-filtered with a modified IRS(send) filter [12]. This source speech material was provided with a sampling frequency of 8 kHz and an Active speech level (ASL, [11]) of -26 dB re. OVL.

6.4.2 Artificial Test Signals

The test signals which were used for the objective tests conducted by HEAD acoustics during the event are published and defined in ITU-T Recommendation P.501 [3]. These speech-like test signals represent important characteristics of real speech, e.g. voiced parts such as the vowels in real speech, unvoiced parts such as most of the consonants, power density spectrum or signal modulation vs. time. These signals have the advantage of not being limited to one language or one speaker. The corresponding analysis methods are described in ITU-T Recommendation P.502 [4]. In the following these test signals are briefly described.

Figure 6 shows the so-called "Composite Source Signal". It consists of 2 signal bursts with a duration of 250 ms each and a pause of approximately 100 ms (upper large window). Each burst consists of a voiced part, a shaped pseudo random noise sequence and a pause. The power density spectrum (lower window in figure 6), derived from the Fourier-Transform of this test signal, reproduces the spectral characteristics of real speech.



NOTE: Upper window: time signal, lower window: power density spectrum derived by Fourier transform.

Figure 6: Composite Source Signal

For tests under single talk conditions, this signal is fed into the measurement set-up, and the transmitted signal is recorded and analysed. Parameters such as one way transmission delay, frequency responses and others can be determined using this test signal.

A periodical repetition of this Composite Source Signal with variable signal level for each signal burst is shown in figure 7. This test signal is used to determine the sensitivity threshold and the switching characteristic of voice activity detection (VAD).

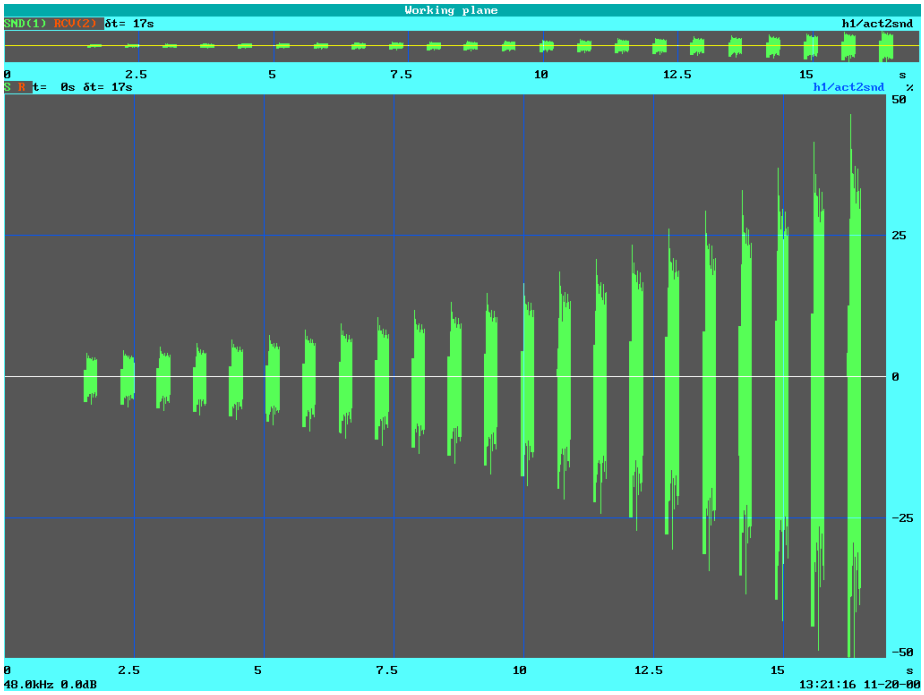
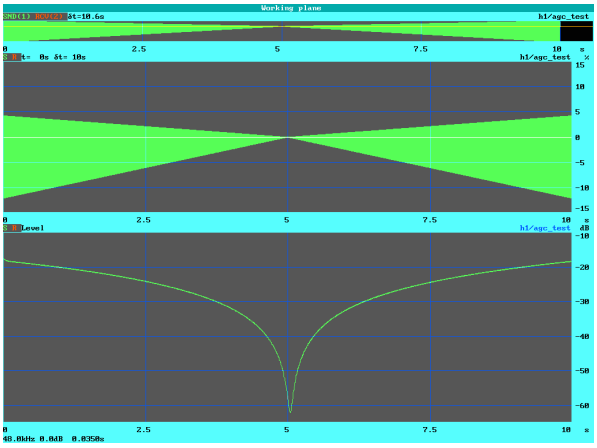


Figure 7: Test signal to determine the activation sensitivity under single talk conditions

A test signal developed to determine the behaviour of implemented automatic gain control (AGC) is shown in figure 8. It consists of a periodical repetition of a voice sound (the same which is used in the voiced part of the CSS in figure 6) with level variations vs. time.



NOTE: Upper window: time signal, 10 s, lower window: signal level vs. time.

Figure 8: Test signal to measure AGC behaviour

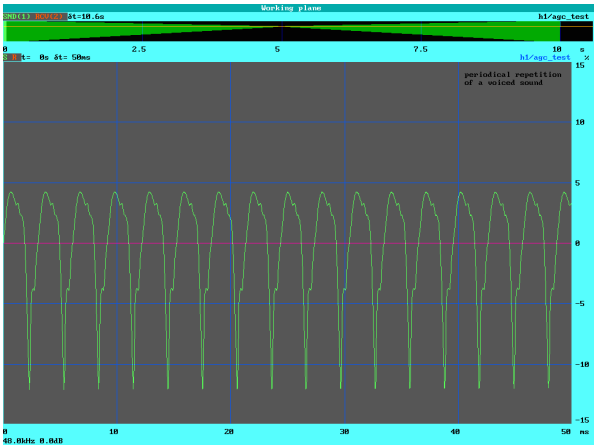
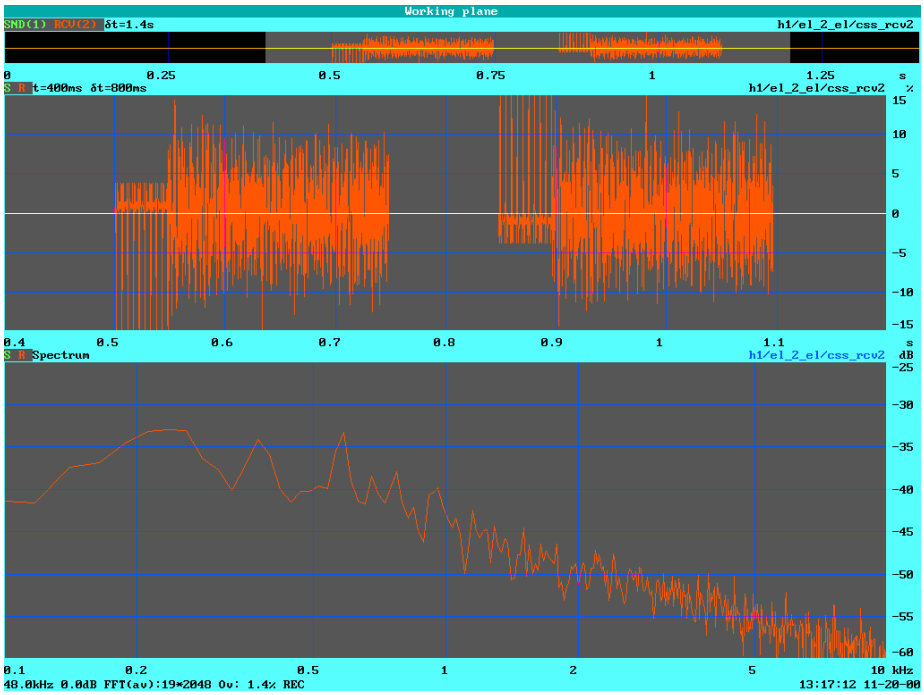


Figure 9: Enlarged time sequence from figure 8 (50 ms) showing in detail the periodical repetition of the voiced sound

The evaluation of double talk performance - both subscriber talk simultaneously - requires a second test signal to be applied simultaneously at the opposite transmission path. The two signals that simulate double talk need to be decorrelated. Figure 10 shows a second decorrelated Composite Source Signal.



NOTE: Upper window: time signal, lower window: power density spectrum derived by Fourier transform.

Figure 10: Second Composite Source Signal to simulated double talk

If both signals (the one from figure 6 and the one from figure 10) are applied simultaneously to the measurement setup, this simulates a double talk period, specific parameters determining transmission quality under double talk conditions can be analysed. These two Composite Source Signals can be combined in various ways to a two channel signal (including for example level variation or other signal characteristics) to simulate specific double talk situations during the tests. From the analysis of two Composite Source Signals those quality parameters that occur during periods of double talk can be determined (such as signal level variations or echo during double talk).

Figure 11 gives an example for a double talk signal. The Composite Source Signal in both channels is periodically repeated with a level variation of 20 dB in each transmission direction. The green test signal is the one shown in figure 6 and is fed into the measurement set-up in the sending direction, the red signal is the one shown in figure 10 and simultaneously fed in the receiving direction. The yellow part shows the periods of "double talk", i.e. where both green and red signals are present. Note that the entire signal sequence has a duration of 32 s.

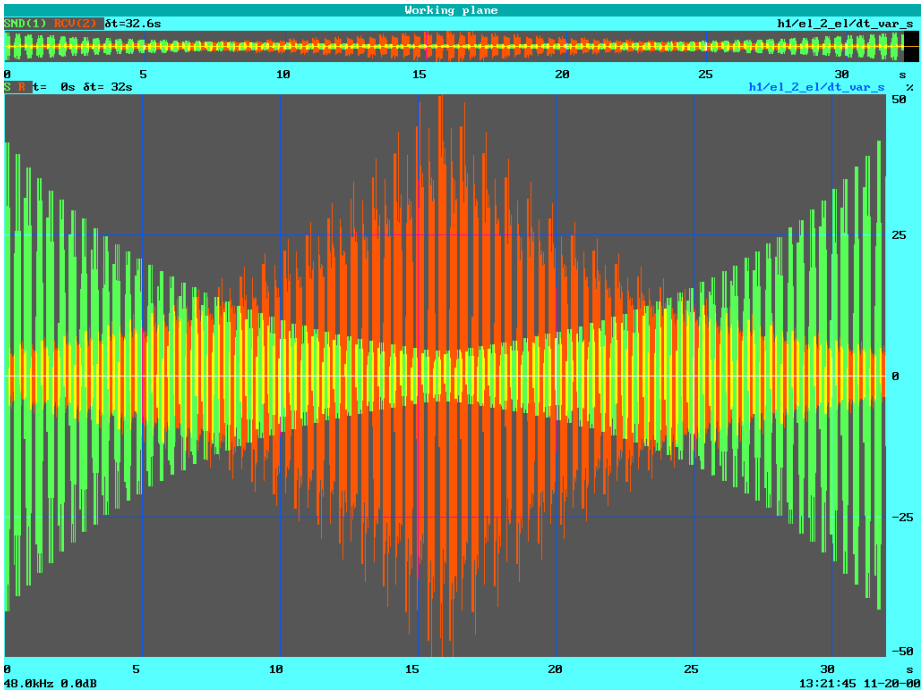


Figure 11: Test signal to simulated double talk based on the periodical repetition of the two Composite Source Signals (sequence length 32 s)

A sequence of 2 seconds from this double talk signal is shown in figure 12 in order to demonstrate the signal composition in detail. The two test signals reproduce short single talk periods in both transmission directions (only one signal is active, either green or red) and real double talk periods (both signals are active simultaneously as indicated in yellow). Double talk sequences using real speech look just like the test signals for the double talk situation is one example shown in figure 13 where the green part is a male voice and the red part a female voice.

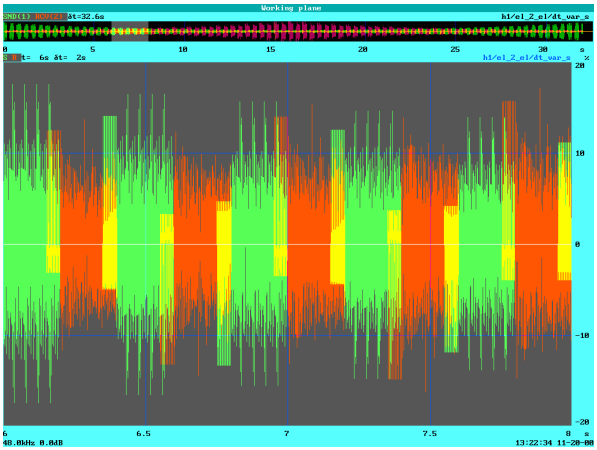
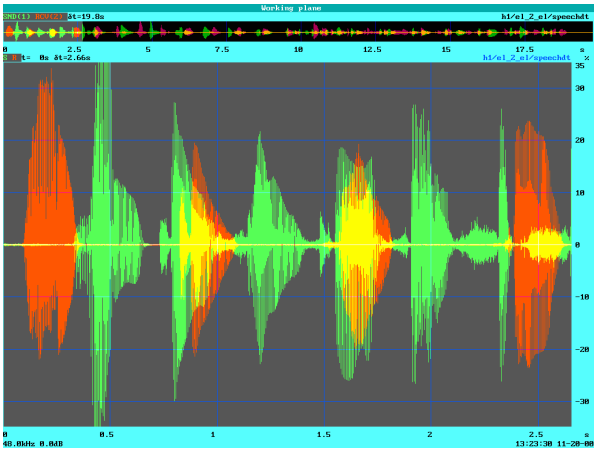


Figure 12: Enlarged time sequence from figure 11 showing in detail the periodical repetition of the voiced sound



NOTE: Green: male voice, red: female voice.

Figure 13: Typical double talk sequence using real speech

The transmission characteristic for a background noise signal can be determined using the test signal shown in figure 14. The sequence is a random noise signal with Hoth spectrum and is applied with increasing level to the measurement object. The level varies from infinite to -25 dBV. If the transmitted signal is recorded and analysed the systems reaction on this background noise can be determined for different input signal levels.

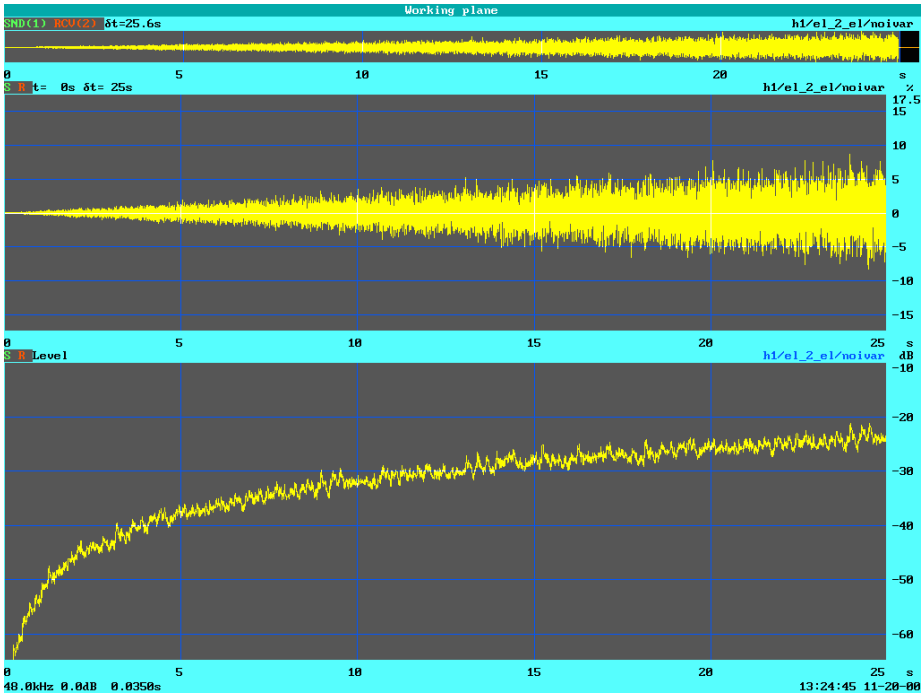


Figure 14: Noise signal with increasing level to determine quality of background noise transmission

Moreover, a realistic background noise signal - recorded in a cafeteria - was used for the tests. This realistic signal has higher signal level variations vs. time due to voice babble in the background, laughing or other sounds from the cafeteria. Recordings were carried out using this test signal from a realistic background noise scenario to generate listening examples in order to compare it to the results obtained with the test signal is shown in figure 14. In addition CS signals and real speech were applied in sending direction (the same test direction as for the background noise signal) or in receiving direction (the opposite transmission path). The test signal is shown in figure 15. The background noise signal and the test signals transmitted in the same direction (sending direction) are displayed in green colour. The signal which is applied to the opposite transmission path is represented in red colour.

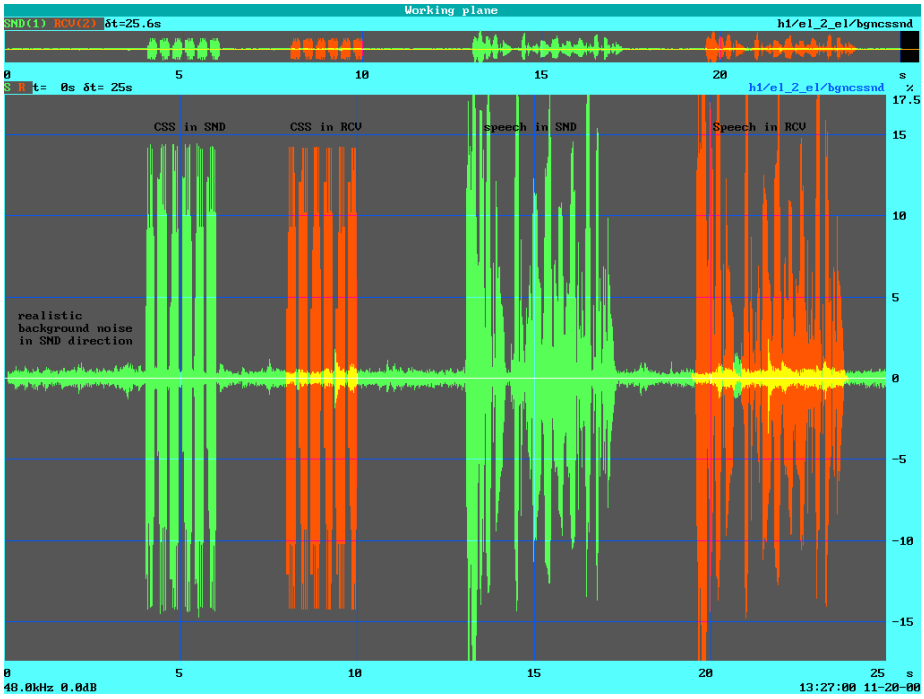


Figure 15: Realistic background noise (students cafeteria) with additional CSS and speech sequences applied in the same direction (green) or in the opposite transmission path (red)

The test signal for measuring echoes in the connections consists of the periodical repetition of the composite source signal as shown in figure 16. The composite source signal is repeated to achieve an appropriate signal length (12 s in figure 16).

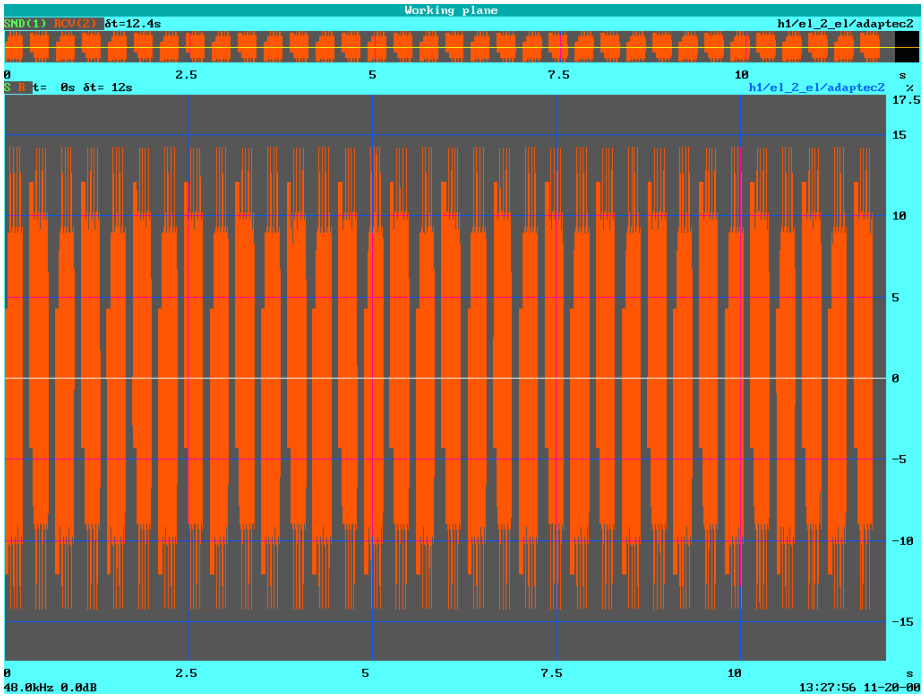


Figure 16: Periodical repetition of the composite source signal to measure echo during single talk

An additional test signal to determine echo during double talk is represented in figure 17. The red signal is applied in one direction of the measurement set-up and the green signal simulates the double talk. The sequence shown in figure 17 in the upper window represents a single talk situation in receiving direction for about 2 s, then the double talk sequence (yellow colour) is applied for again 2 s and the sequence ends with another short single talk period. The power density spectrum calculated by Fourier transform is given in the lower window. The 2 signals show comb-filter spectra, which is necessary to distinguish between the double signal (coming from the near end) and the echo signal (coming from the echo path as a reaction on the receive signal) by appropriate filtering.

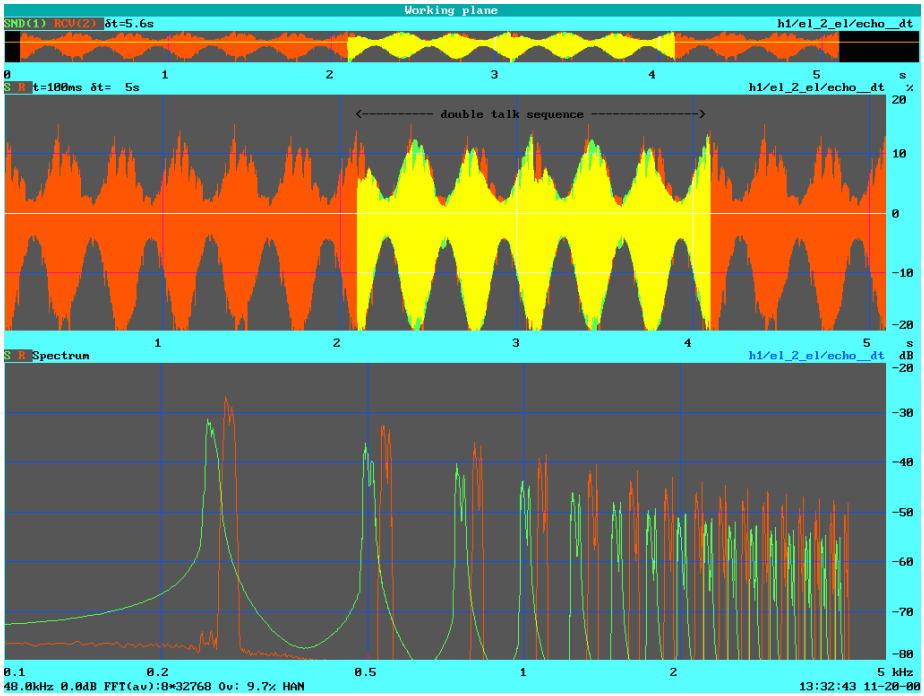


Figure 17: Two channel test signal with comb-filter structures to determine echo during double talk

6.5 Assessment Methods

6.5.1 Auditory Assessment

The auditory test results lead to speech quality ratings expressed through mean opinion scores MOS [1]. These MOS values represent the average test result derived from all individual ratings of a group of untrained test persons, who assess the auditory perceived speech sound quality. It should be noticed that for the whole measurements auditory assessments took place with a subset of the recorded speech material in order to verify the instrumental quality measures against the auditory test results.

However, the speech processing and the speech samples were designed in this way, that T-Nova Berkom is able to carry out separate auditory tests according to the ITU-T Recommendation P.800 series with the processed German speech material. This test would be carried out on request outside this ETSI VoIP test event on speech quality. It would provide auditory assessments for all test conditions in the "electrical-electrical" measurement part.

6.5.2 Instrumental Assessment

Latest psychoacoustic instrumental analyses using the Telecommunications Objective Speech Quality Assessment method TOSQA [2] lead to an one dimensional test result TMOS with a high correlation to quality scores gained by auditory listening only tests. These speech quality values describe listening quality and contain effects by one-way speech transmission which are perceived by a listener. The method is validated for VoIP transmission scenarios [7] and therefore applicable for the scenarios to be tested during the event. However, a subset of speech material was used for a cross-check analysis in order to demonstrate the performance of TOSQA for the entire transmission scenarios of this test event.

6.5.3 Instrumental Computational Assessment Using Speech-like (P.501) Test Signals

The auditory perceived quality for speech controlled, non-linear or time-variant systems is influenced by additional parameters like echo disturbances, double talk performance, switching characteristics, background noise transmission and others (see [2], [3] and [16]). These parameters like talking-related impairments (e.g. echo) or conversational aspects (e.g. double talk performance) determine the overall quality of the complete system. Tests based on sophisticated test signals and analysis methods were developed to determine the corresponding instrumental parameters. Depending on the interfaces used during the tests (acoustical, electrical) parameters according to the following list was measured [8], [9] and [10]:

- One-way delay in send and receive direction.
- Send loudness rating SLR, receive loudness rating RLR, junction loudness rating JLR, overall loudness rating OLR ("mouth to ear").
- Frequency responses and distortion, switching characteristics like minimum activation level, sensitivity of double talk detection.
- Double talk performance.
- Background noise transmission at idle mode, with near end signal, with far end signal.
- Echo delay, single talk echo, double talk echo.

These tests are performed in order to determine instrumental quality parameters for the given connections. The instrumental tests for the determination of implemented parameters is meant to check common requirements in telephonometry and to identify parameters which may lead to auditory perceived conversational quality degradation.

7 Results

7.1 Auditory Reference Test

7.1.1 Performance of the Auditory Test

The auditory test was performed in the Berlin speech quality labs at T-Nova and was carried out according to ITU-T Recommendation P.800 [1]. Each circuit condition was tested with four different speech samples (short, concatenated sentences, 2 male and 2 female speakers each). Speech samples were played back to the subjects via a conventionally shaped standard telephone handset in a low-noise test cabinet (room noise floor < 30 dB(A)), at a listening level of about 79 dB(A) SPL. Test subjects were required to judge the listening speech quality on a 5-point ACR overall speech quality scale with German scale labels, as recommended in ITU-T Recommendation P.800 [5]. Before starting the first test session, 8 example speech samples were played back to the test subjects, which were expected to range in quality from good to poor. In addition to the tested VoIP scenarios also 11 reference conditions were integrated in the test. These are single speech codecs (G.723.1and G.729) from a simulation process as well as different MNRU conditions [13]. To avoid the influence of presentation order on the test result, groups of test subjects listen to the samples in different randomized orders. All in all 25 test subjects took part on the auditory test. Four votes per each condition and test subject yield the Mean Opinion Sore (MOS) for the condition under test. So the MOS per condition is an average of 100 single votes.

7.1.1.1 TOSQA Results

In the similar manner as subjects assess the speech quality in the listening test, TOSQA rates the four speech samples for each condition. Here exactly the same speech material from the auditory test was used. TOSQA compares this transmitted and possibly distorted speech material with the clean input speech material as reference.

The four TOSQA results (TMOS) of the four speech samples were averaged and are the basis for comparison with the MOS values gained by the auditory test.

Figure 18 plot shows mean opinion scores gained by auditory experiments on the X axis. The Y axis shows the speech quality value TMOS calculated by TOSQA. The TMOS values were transformed in a common way by a third order monotonous mapping.

In an ideal prediction of quality, the instrumental results would be equal to the MOS-values. In this case all symbols would be on the 45° line in the diagram. All in all TOSQA shows a good prediction of speech quality, the correlation coefficient after third order mapping is 0,916.

This result is based on the four German samples which are used in the auditory reference test. In addition to the four German samples, also four English samples were transmitted via each network condition, followed by another 8 German samples All of them were assessed by TOSQA. The resulting mean TMOS averaged over all 16 samples was also compared with MOS values gained by the auditory test. In this case the correlation coefficient between MOS and "mean TMOS" increases slightly to 0,921.

This values should be also compared with the requirement which was defined in the ITU-T Q13/12 for such kind of conditions (cf. ITU-T document COM12-117E, March 2000). During the competition for the new standard for objective speech quality measurement a minimum correlation coefficient of 0,80 was required for real VoIP recordings if the test corpus was not used for training the objective model. This is exactly the same situation we had in our VoIP test here.

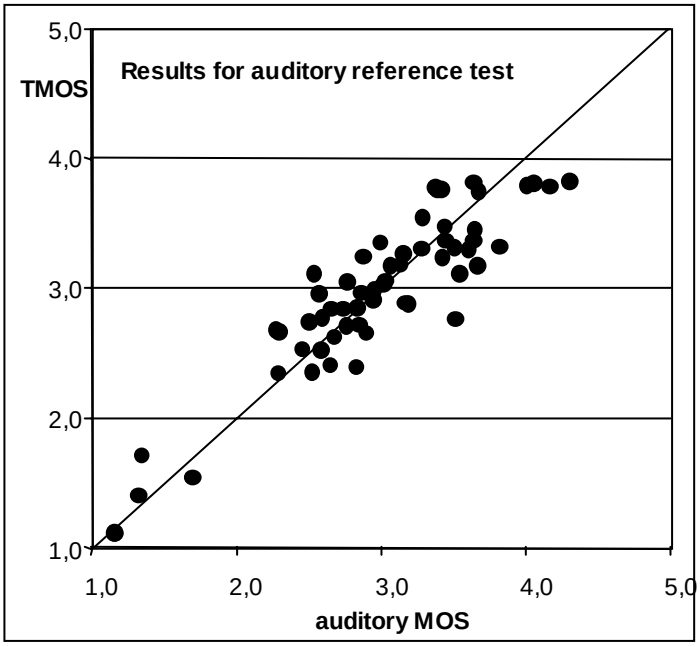


Figure 18: MOS versus TMOS for VoIP conditions and references

Figure 19 shows a more detailed scatter plot with the G.711 conditions as unfilled circles, all other VoIP conditions as filled circles and the reference conditions as grey diamonds.

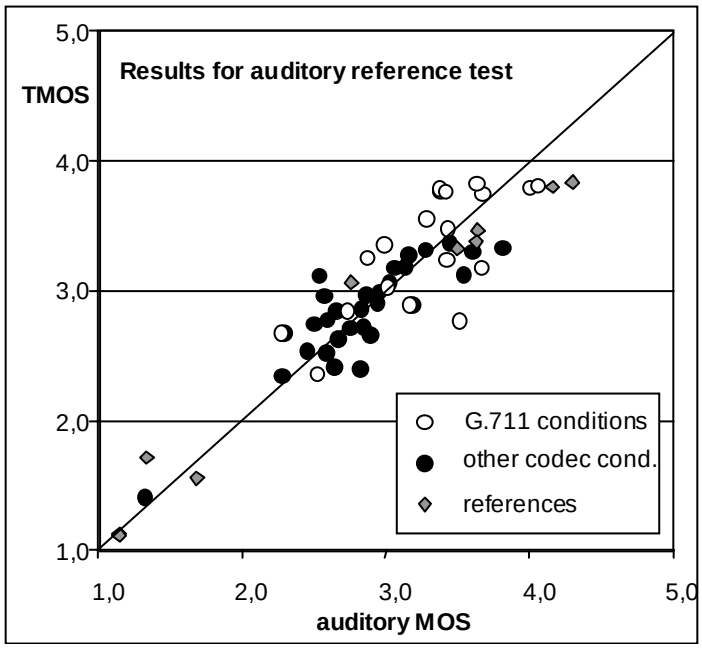


Figure 19: MOS versus TMOS for VoIP conditions and references, detailed plot

TOSQA calculates not only a speech quality value which is focused on quality during speech activity, but also an additional value called "connection quality: CQ". This algorithm is still under test and therefore not referenced in ITU-T. Here besides other values, a special weighting of intervals without speech activity is taken into account for the connection quality result. If the current version of this algorithm of TOSQA "connection quality" will be used for assessment, the correlation increases further to 0,935.

During the competition for a new standard for objective speech quality measurement within ITU-T Q13/12, also a VoIP database were assessed. This database was recorded during the ETSI Tiphon SuperOP, Hawaii in September 1999. Here a correlation of 0,949 between TOSQA and auditory MOS could be reached. In case of TOSQA "connection quality" this result is increasing to 0,965.

These higher correlations are caused by the smaller range of included conditions in this test. This so called "SuperOP-test" contains only G.711 and G.723.1 conditions under different rates of packet loss and delay jitter. There only two coding algorithm were tested and only one manufacturer was involved in the test.

7.2 Speech Quality Estimation Using Voice Signals

Voice samples were processed across the VoIP scenario for the estimation of the speech quality. Because of the random behaviour of the IP network emulation (NIST Net) in total 16 voice samples (8 seconds voice each, German and English) were processed in each transmission scenario. For minimizing the random influence of the NIST Net all 16 TMOS values obtained for each condition were averaged.

Tables 3, 4 and 5 show the speech quality estimation from the codec types which were mostly used in the participants VoIP equipment. For each condition, an averaged TMOS as well as the minimum and the maximum value are provided. In some cases the differences between TMOS min and TMOS max are quite large, even if the same codec under similar network conditions is used. These deviations are mainly caused by the fact that the implementations provided by the participants are slightly different, for example for packet length, different packet loss concealment as well as the entire voice activity detection algorithms.

7.2.1 G.711 Codec

Table 3 shows the average test results from C1 to C6 for the G.711 codec.

Table 3: G.711, VAD on, PLC on/off, PL = 20ms/30 ms

Condition	TMOS average	TMOS min	TMOS max
C1: Drop = 0 %, Delay = 0 ms, Jitter = 0 ms	4,2	4,2	4,2
C2: Drop = 1 %, Delay = 0 ms, Jitter = 0 ms	3,7	3,3	4,0
C3: Drop = 2 %, Delay = 0 ms, Jitter = 0 ms	3,4	3,1	3,8
C4: Drop = 3 %, Delay = 0 ms, Jitter = 0 ms	3,4	2,8	3,7
C5: Drop = 5 %, Delay = 0 ms, Jitter = 0 ms	3,1	2,7	3,5
C6: Drop = 1 %, Delay = 50 ms, Jitter = 20 ms	3	2,7	3,3

As expected the TMOS values are decreasing with higher packet loss values. In general the results show that for the G.711 codec the reached TMOS values, especially for C1, are fully comparable with the simulated reference values.

The averaged TMOS values were obtained by assessment of G.711 codec implementations with PLC (packet loss concealment) as well G.711 as implementations without PLC algorithm. Furthermore the summary assessment includes also implementations with both packet lengths, 20 ms and 30 ms.

7.2.2 G.723 Codec

Table 4 shows the test results from C1 to C6 for the G.723.1 (6.3) codec.

Table 4: G.723.1 (6.3), VAD on, PLC on/off, PL = 30 ms

Condition	TMOS average	TMOS min	TMOS max
C1: Drop = 0 %, Delay = 0 ms, Jitter = 0 ms	3,3	3,2	3,5
C2: Drop = 1 %, Delay = 0 ms, Jitter = 0 ms	3,2	3,1	3,4
C3: Drop = 2 %, Delay = 0 ms, Jitter = 0 ms	3,1	3,0	3,2
C4: Drop = 3 %, Delay = 0 ms, Jitter = 0 ms	3,0	2,9	3,1
C5: Drop = 5 %, Delay = 0 ms, Jitter = 0 ms	2,8	2,7	2,9
C6: Drop = 1 %, Delay = 50 ms, Jitter = 20 ms	2,7	2,0	3,1

Also the average values obtained by the assessment of G.723.1 implementations are decreasing with higher packet loss values. In general the results show that for the G.723 codec the reached TMOS values, especially for C1, are almost comparable with the simulated reference values.

The averaged TMOS values were obtained by assessment of G.723.1 codec implementations with PLC (packet loss concealment) as well as G.723.1 implementations without PLC algorithm. In all test cases the packet length was 30 ms.

7.2.3 G.729 Codec

Table 5 shows the average test results from C1 to C6 for all vendors for the G.729 codec.

Table 5: G.729, VAD on, PLC on/off, PL= 10ms/30 ms

Condition	TMOS average	TMOS min	TMOS max
C1: Drop = 0 %, Delay = 0 ms, Jitter = 0 ms	3,3	3,1	3,5
C2: Drop = 1 %, Delay = 0 ms, Jitter = 0 ms	3,1	2,9	3,3
C3: Drop = 2 %, Delay = 0 ms, Jitter = 0 ms	3,0	2,9	3,2
C4: Drop = 3 %, Delay = 0 ms, Jitter = 0 ms	2,9	2,9	3,0
C5: Drop = 5 %, Delay = 0 ms, Jitter = 0 ms	2,7	2,6	2,9
C6: Drop = 1 %, Delay = 50 ms, Jitter = 20 ms	2,3	1,8	2,7

Also the average values obtained by the assessment of G.729 implementations are decreasing with higher packet loss values. In general the results show that for the G.729 codec the reached TMOS values, especially for C1, are almost comparable with the simulated reference values.

The averaged TMOS values were obtained by assessment of G.729 codec implementations with PLC (packet loss concealment) as well as G.729 implementations without PLC algorithm. Furthermore the summary assessment includes also implementations with packet lengths of 10 ms to 30 ms.

7.2.4 Summary of Results

Figure 20 provides a summarized overview of all TMOS results. It displays also the speech quality assessments of the references that are used in this test to compare the results of the real processed voice samples against codec simulations as well as MNRU conditions (numbers indicate the adjusted Q values).

The assessment has shown that in case of the G.711 codec under error free conditions the highest TMOS values of the systems under test were reached. The obtained TMOS value is nearly identical to the simulated G.711 reference. In case of packet loss conditions a degradation of speech quality could be detected. Furthermore there is a certain range of deviation of the resulted TMOS maximum and TMOS minimum values. This could be caused by different settings like PLC on / PLC off and different packet lengths. This effect is especially observed in case of the tested G.711 conditions.

Due to the high basic quality of the G.711 codec distortions decrease the speech quality value more in an absolute value than it could be achieved in case of low bit rate codecs.

The TMOS minimum values under packet loss conditions of 2 % and more are close together for G.723.1 and G.729 codec type. They seems independent on the used codec type.

The reference values in case of the low bit rate codecs G.723.1 and G.729 were nearly reached only by some implementations under error free conditions.

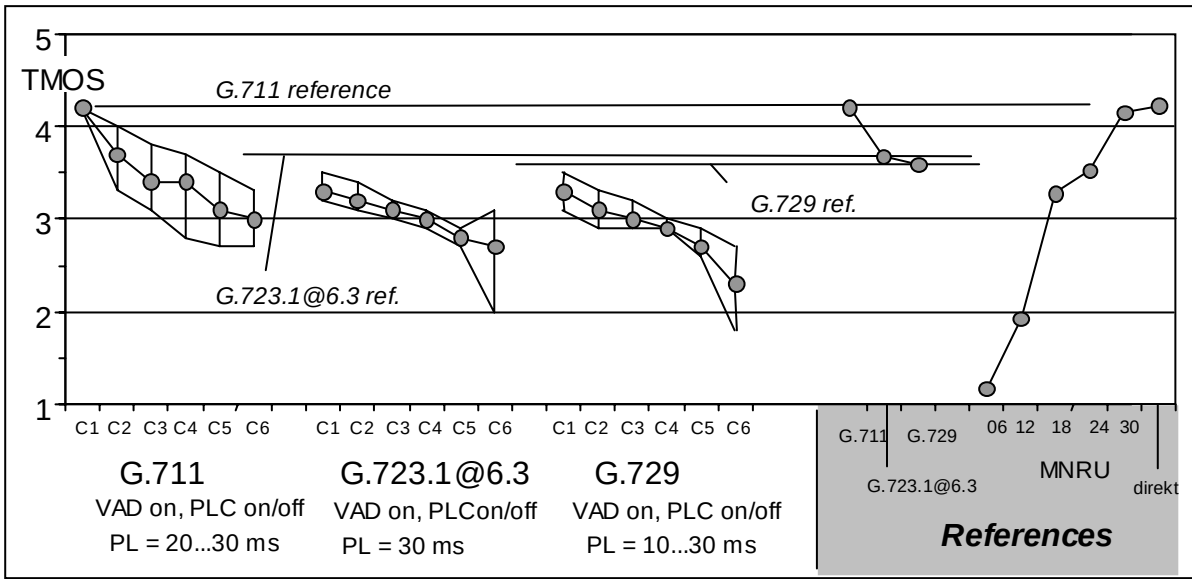


Figure 20: Summary of all conditions (including reference conditions)

7.3 Advanced Measurements on Communicational Quality

7.3.1 Parameters determining speech sound quality under single talk conditions

Figures 21 and 22 demonstrate two analysis examples obtained from recordings under network conditions with simulated packet loss. The figures show an enlarged time sequence from the recorded signal using the periodical repetition of a voiced sound as test signal.

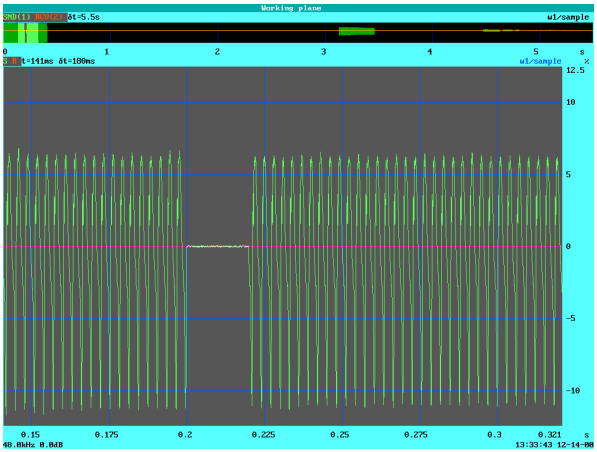


Figure 21: Occurrence of packet loss

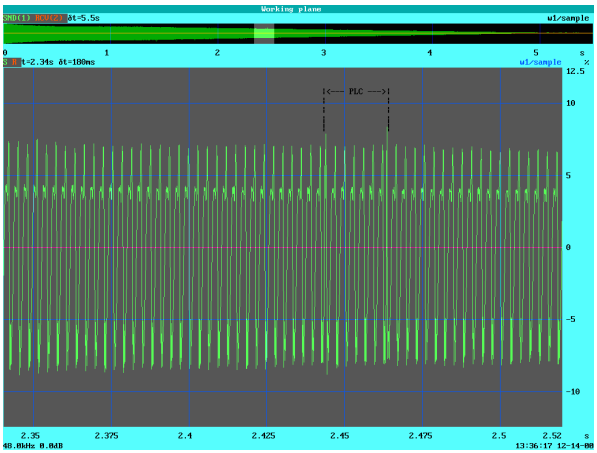


Figure 22: Packet loss concealment

A resulting signal gap due to one lost IP packet can be analysed in figure 21. Packet loss concealment (PLC) is obviously not implemented in the equipment which was under test. A typical PLC implementation can be seen in figure 22. The lost IP packet obviously is substituted by a previous frame, but signal irregularities at the beginning and the end of this substituted frame can be analysed. These resulting peaks are audible and annoying.

This result reflects one kind of PLC implementation which can be used to cover the influence of lost packets in the network.

7.3.2 Transmission Characteristics for Background Noise

The test signal consists of the spectral shaped noise signal (Hoth spectrum) with increasing test signal level vs. time. The measurement result derived from the test via one ISDN line and the PBX is shown in figure 23 for comparison. The red signal in the upper windows represents the test signal, the green signal the measured signal (transmitted signal). Note, that the overlapped red and green colour results in yellow. The analysis curves in the lower window show the calculated level vs. time (calculated in time domain using a 35 ms time constant). The transmission is linear, no level variations can be observed, the two curves (red for the original test signal level, green for the transmitted signal) are in parallel. Typical results derived from tests during the event are shown in figure 24.

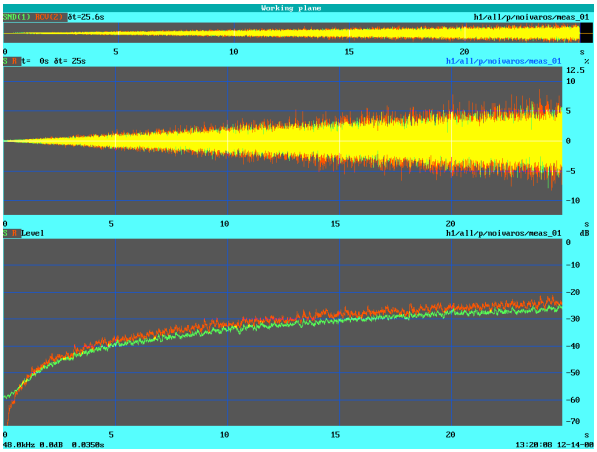


Figure 23: Transmission characteristic for background noise (PBX reference connection)

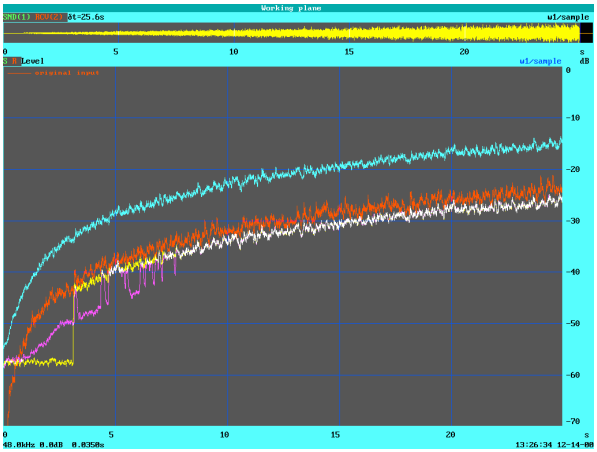


Figure 24: Typical analysis results derived from tests with 3 manufacturers (red: original test signal)

The results displayed in yellow, magenta and cyan represent different kinds of implementations. An activation level threshold (yellow), the influence of automatic gain control (AGC, cyan) and an implementation with adaptive comfort noise injection can be analysed clearly.

7.3.3 Transmission Performance under Double Talk Conditions

The transmission characteristics under double talk conditions were determined using the test signal consisting of two decorrelated composite source signals which are periodically repeated and applied with different signal levels to both sides of the connections. The recordings were carried out in one direction and consequently the signal which is fed at the other end of the connection should be transmitted and no additional signal components should occur. The following four figures demonstrate typical measurement results which were obtained during the event. In each figure the transmitted and recorded signal is shown in green and the original test signal is displayed in red. In order not to confuse the reader the double talk signal which is fed in the opposite transmission path is not given in these figures.

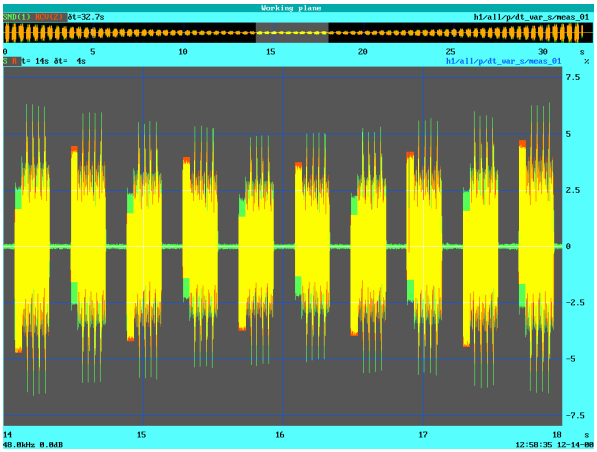


Figure 25: Result for the PBX connection

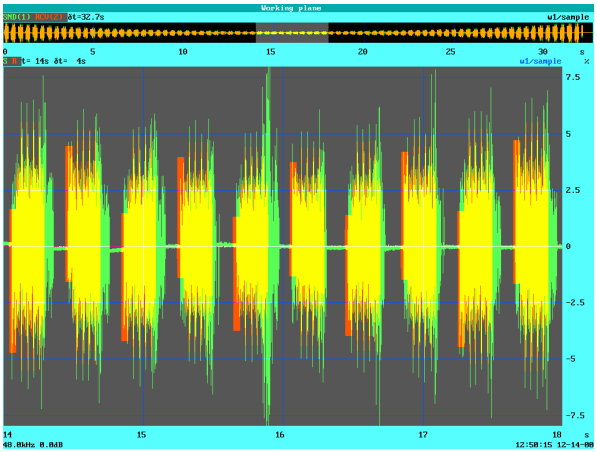


Figure 26: Typical result demonstrating echo components and short term clipping

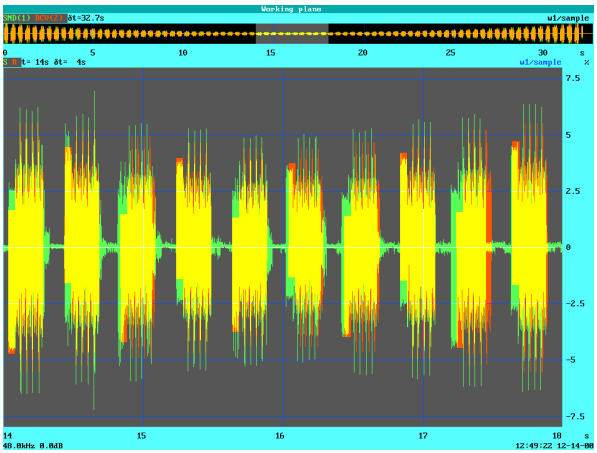


Figure 27: Typical result demonstrating echo components

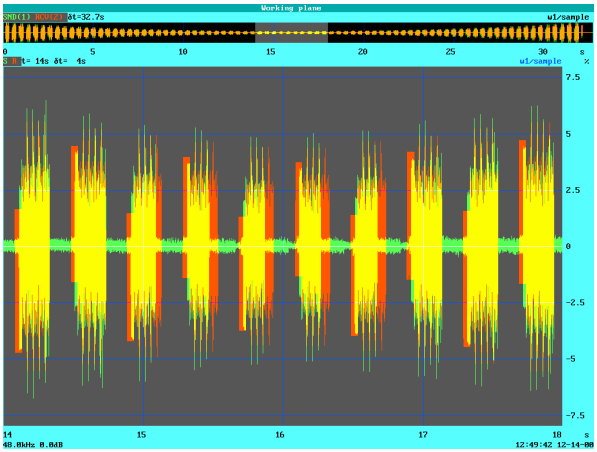


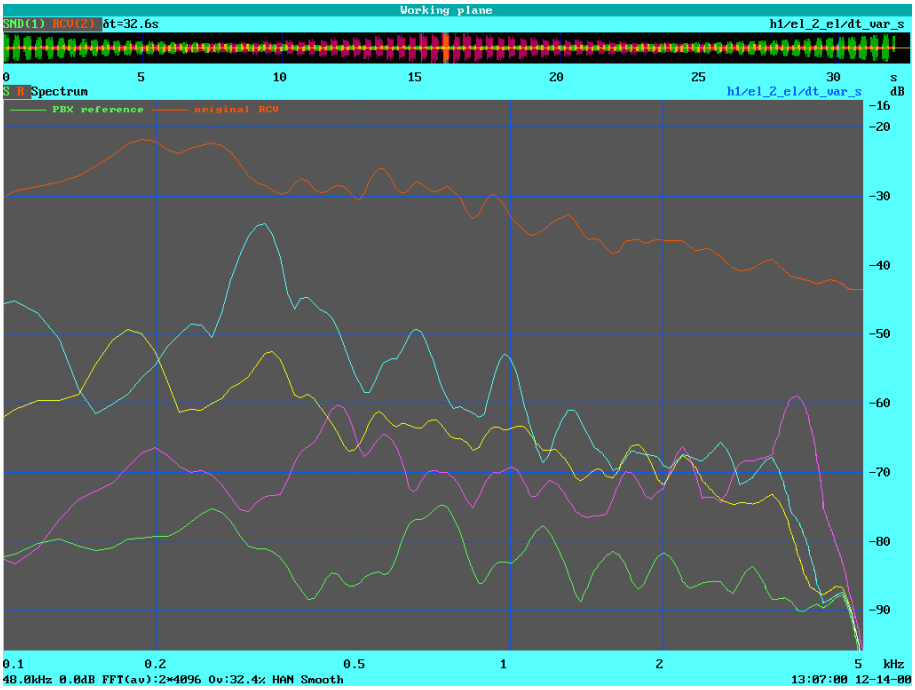
Figure 28: Typical result demonstrating echo components and clipping

Figure 25 shows the example for the analysis obtained for the PBX reference connection and points out that all signal bursts are completely transmitted. Clipping does not occur. During the pauses between 2 signal bursts the channel noise floor is recorded. No additional signal components which could be identified as disturbances (like echoes or others) occur.

The complete transmission of all signal bursts can also be seen in figure 26 except a short frontend clipping. Additional high level echo components (green) can be determined during the pauses between the signal bursts. Basically the same result can be analysed in figure 27 for another tested connection. One example for the occurrence of signal clipping (parts of the green signals are missing) is demonstrated in figure 28. An additional high level noise floor is recorded between the signal bursts in this example.

These three measurement results demonstrate some disturbances which are introduced by the implemented echo cancellers and the non-linear processors.

Figure 29 compares the power density spectra analysed during the signal pauses under double talk conditions. The red curve represents the curve for the original test signal which was applied in the opposite transmission path. The green curve was derived from the measurement of the PBX connection. The other curves in cyan, yellow and magenta represent the results for 3 different test connections. Obviously significant differences compared to the PBX connection occurred for the connections under test during the event.



NOTE: Red: original test signal, green: PBX connection, others: typical results measured during the event.

Figure 29: Comparison of power density spectra analysed in the pauses between two signals bursts during the double talk period

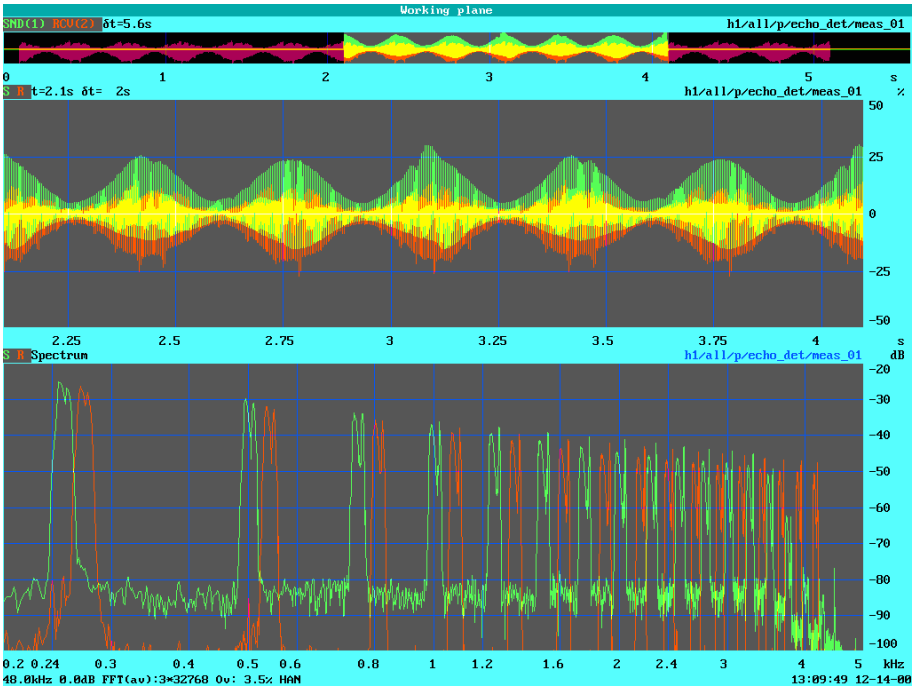
Further analysis demonstrates that these signal components are caused by echoes which only occur during double talk periods. It should be noted that the test connection was completely 4-wire and therefore physically did not produce any echo.

7.3.4 Detailed Analysis of Echo during Double Talk

A specific test signal according to ITU-T Recommendation P.501 [3] consisting of a two channel signal with a comb filter structure was used for a more detailed analysis of the echo disturbances during double talk. This signal is suited to determine and measure the echo more precisely.

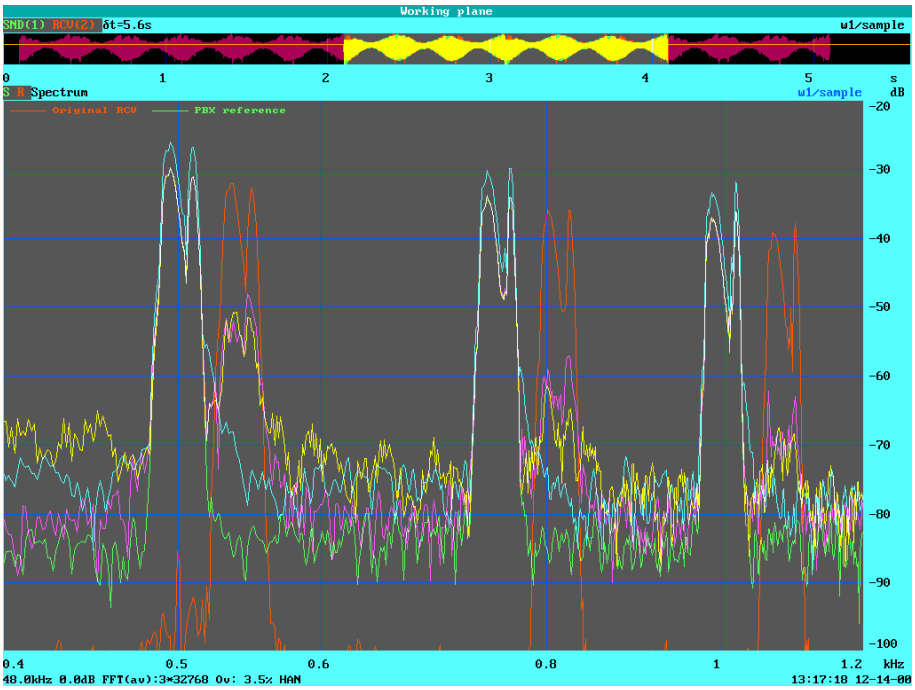
The test result which was obtained for the PBX connection is analysed in the example in figure 30. The time signals are shown in the upper window. The analysis window (lower window) shows the power density spectrum calculated by Fourier transform for the measured signal (green) and the original test signal in receiving direction (red). In principle signal components in the measured (green) signal which correlate to the excitation frequencies of the original signal (red, note that this was the original one which may cause echoes) can be analysed as echoes.

The analysis in figure 30 shows that the PBX connection was echo-free under double talk conditions. A more detailed analysis example for the enlarged frequency range between 400 Hz and 1,2 kHz is given in the following figure 31. Again typical measurement results obtained for 3 different manufacturers are analysed together with the PBX connection (green). The figure shows significant echo components for some of the tested connections. These components can be detected in the frequency range which was excited by the original test signal (red).



NOTE: Upper window: time sequences, lower window: power density spectra green: measured signal containing double talk signal and echo, red: original test signal.

Figure 30: Determination of echo during double talk test result for the PBX reference connection



NOTE: Green: PBX reference connection, red: original test signal, others: typical results measured during the event.

Figure 31: Enlarged frequency range with a comparison of typical test results

8 Conclusion

The present document gives an overview about the kind of measurements, the measurement methodologies and the major results from the 1st ETSI speech quality test event.

The test event was very successful and - as indicated by the feedback of all participating companies - very useful for the manufactures. The results of the listening speech quality demonstrate the state of the art performance of the equipment under single talk conditions. The test scenarios and the measurements as they are described above determine additional parameters which highly influence the conversational quality. The tests and the results obtained during the speech quality test event clearly pointed out the importance to consider the end-to-end scenario.

It can be assumed that for all participating manufactures the results can be used for optimization of the VoIP equipment to improve the overall speech quality.

It is strongly recommended to continue the process of end-to-end speech quality testing.

History

Document history		
V1.1.1	February 2007	Publication