STATE OF THE ART VOICE QUALITY TESTING

OPTICOM GmbH
Nägelsbachstr. 38
91052 Erlangen
GERMANY

Phone: +49 9131 / 530 200
Fax: +49 9131 / 5302020
EMail: info@opticom.de
Website: www.opticom.de

Further information:
www.psqm.org
www.pesq.org

White Paper by OPTICOM GmbH, Germany

CONTENTS

1 INTRODUCTION

QoS testing is one of the key issues in modern telecommunications. Whether it is during the development of VoIP equipment, setting up networks or while operating a mobile network, one will always be faced with the problem to determine and optimize the speech quality. New evaluation techniques based on modelling human perception have been devised to tackle this issue. This white paper intends to give the reader a comprehensive understanding why speech quality testing is required as well as an overview on the latest standards for state of the art voice quality testing.

2 FACTORS AFFECTING VOICE QUALITY

Traditional telephony networks are built to provide an optimal service for time-sensitive voice applications requiring low delay, low jitter and have constant but low bandwidth requirements. IP networks, on the other hand, were built to support non real-time applications such as file transfer or e-mail. These applications are characterized by their bursty traffic characteristics, with sometimes high bandwidth demand, but are not sensitive to delay.

Converging telephony and IP networks require that IP networks are enhanced with mechanisms that ensure the quality of service (QoS) required to carry voice over IP. This is especially important considering that users of the traditional telephony networks are used to quite high voice quality standards. Providing comparable service quality as in the traditional telephone networks will drive the initial acceptance and success of VoIP services.

2.1 Traditional Networks (POTS)

Figure 1 shows an example for the topology of a traditional public network (PSTN), such as those which were established until a few years ago. Due to the fact that in most countries a sole public organization was responsible for the network design, a homogeneous technology could be exploited.

On the PSTN, voice quality is influenced mainly by the quality of the handset, the loudness of the telephone, the acoustic echo generated between the speaker and microphone as well as interference on the line itself. The factors that can be used to characterize this behaviour are loudness, delay, echo, noise and cross talk. Most of these can be well assessed by traditional measurements like S/N, non-linear distortions etc.

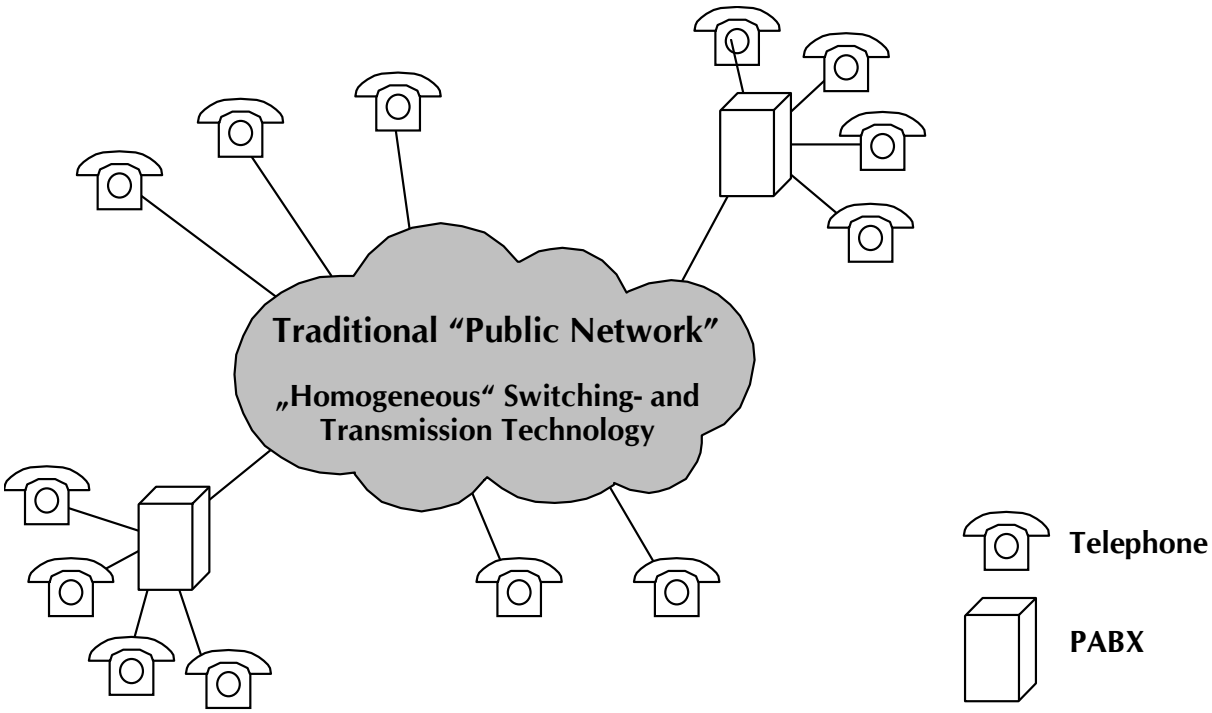


Figure 1: Topology of a traditional network

2.2 Digital Speech Compression

The speech quality of digital signals is primarily a function of the available bit rate. Modern techniques allow for bit rates of 8kbit/s and less, to transmit a speech conversation. This coding gain compared to wide band audio codecs can be achieved by focusing on the modelling of the human speech tract. As a consequence, the codec is highly adapted to transmit speech signals, and music signals or natural sounds will be significantly distorted. Figure 2 illustrates some common coding techniques in the context of bit rates, and the resulting speech quality.

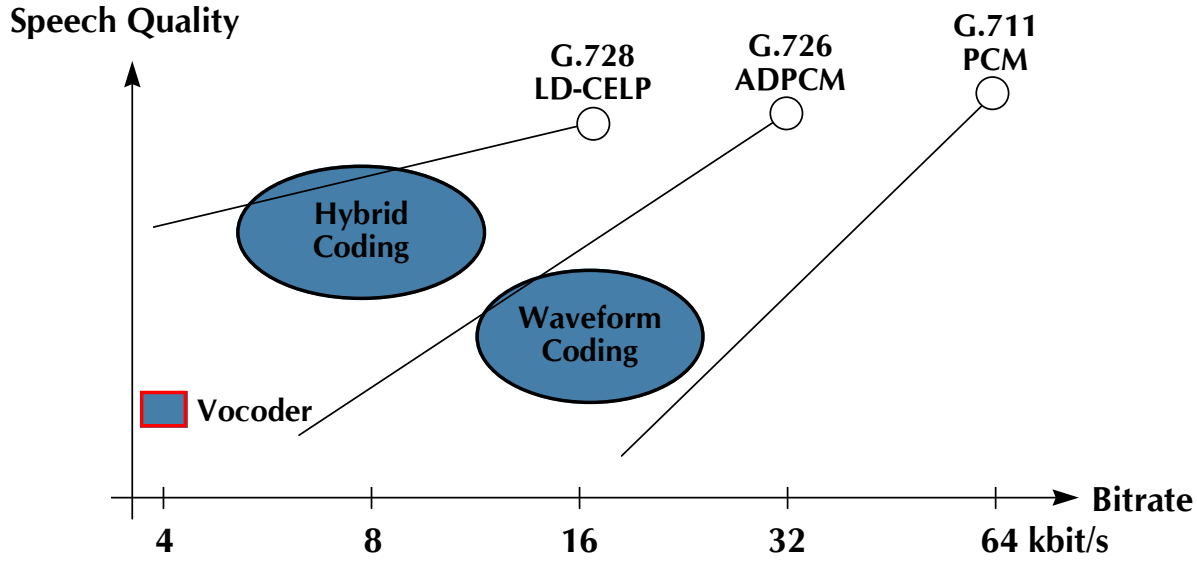


Figure 2: Speech Quality in the context of coding techniques [32]

In the case of VoIP typically the following codecs are being used: G.711 (64 kbit/s), G.723 (5.4 and 6.3 kbit/s), G.728 (16 kbit/s) and G.729 (8 kbit/s), as well as GSM Full-Rate.

While applying a series of evaluations, some general anchors can be found for the most common coding techniques, which are used today in the context of digital phone connections. Table 1 presents an overview on reasonable results that can be expected, when applying P.861 (PSQM) to telecommunication links. It should be noted that even with a very good ISDN connection, the best results will not even come close to 5 ("Excellent"). This compares with the experience of subjective tests, where the maximum MOS observed for an undistorted connection did not exceed 4.3. A reason for this might be the subjective impression of a telephone bandwidth limit to 4 kHz.

Coding Scheme	MOS	Subjective Interpretation
64kbit/s PCM A-law	**4.3**	good, almost excellent
32kbit/s ADPCM DECT	**3.8**	good
13kbit/s GSM Full-Rate	**3.4...3.7**	fair, almost good
...	**...**	

Table 1: Overview on reasonable results of PSQM and their subjective interpretation

2.3 Mobile Transmission

Speech coding is highly deployed in modern communication devices, although originally the evolution was driven specifically by the development of mobile phones. Due to the limited bandwidth, speech coding is a key element to "over air transmissions". However, the encoded data when sent through the air by radio frequencies will be exposed to a sensitive transmission link which is very likely to be affected by errors. Such errors may corrupt the transmitted data, and due to the lack of redundancy it may be difficult if not impossible to reconstruct the speech signal.

Due to interaction of speech coding and transmission, errors in a mobile transmission can cause heavy distortions that sound quite different than traditional "analog" distortions. These artefacts cannot be assessed by traditional measures.

2.4 Packet Based Transmission

The PSTN network uses digital voice transmission for greater efficiency in the backbone. This requires digitizing the analogue voice signal, which affects the voice quality. The VoIP Gateway interconnects the IP network with the PSTN network and adopts voice processing and signaling schemes. The gateway components affecting the voice quality are speech codec, silence suppression mechanism, and comfort noise generator.

In addition, the IP network, even without active voice components, affects the voice quality through its tendency to lose packets and to add extensive jitter to the signal. The H.323 PC terminal also affects the clarity through its speech codec, silence suppression mechanism, and microphone and loudspeaker quality.

Packet loss

Packet loss is not uncommon in IP networks. As the network, or even some of its links, becomes congested, router buffers get filled up and start to drop packets. Another cause can be route changes due

to network links going down. An effect similar to packet loss occurs when a packet experiences a large delay in the network and arrives too late to be used to reconstruct the voice signal.

For non-real-time applications, such as file transfers, packet loss is not critical since the protocol allows retransmission to recover dropped packages. However, in case of real-time, voice information has to arrive in a certain time window to be useful to reconstruct the voice signal. Retransmission would add extensive delay to the reconstruction and cause clipping or uttering, resulting in unintelligible speech.

To avoid packet loss for real-time applications, mechanisms are required in the IP network to assure minimum throughput for selected applications. These mechanisms will minimize packet loss, as well as delay, for the higher priority traffic such as voice. Different router mechanisms can be utilized to meet this objective. These include various prioritization schemes, such as Weighted Fair Queuing (WFQ) and router flow control mechanisms such as the Internet Engineering Task Force's (IETF) Multi-Protocol Label Switching (MPLS) tagging scheme or use of Type of Service (ToS) bits in the IP header. All these mechanisms require prior configuration by a network administrator who must decide what priority and resources to provide for each specific service class. A more dynamic alternative for assigning resources is the Resource Reservation Protocol (RSVP), which permits a voice terminal or voice gateway to request a specific IP quality of service.

Regardless of which is used, a more serious problem remains. QoS is defined on an end-to-end basis, and therefore requires that sufficient network resources be provided throughout the network path. This is not an overwhelming issue for an enterprise network or single ISP environment where all resources can be administered through one network manager. However, it is almost impossible to administer today when multiple ISPs or service providers are involved, as is the case in virtually every national or international long distance call.

In addition, this fulfillment of QoS assumes that all routers in the network are equally capable of identifying voice traffic and providing the network resources required. This is the exception rather than the rule in today's IP networks because standards for many of these mechanisms have not been finalized and implemented by equipment manufacturers.

2.5 Modern CTI Networks

From the above paragraphs we can conclude that various technologies will be exploited in modern CTI networks, thus leading to heterogeneous networks as indicated in Figure 3. And in contrary to the traditional public network, this means today we will find a multitude of public, and corporate business networks interacting in a modern telecommunication business.

One can easily conclude that due to widespread use of compression technology, typically several times in cascade, an objective quality metrics is a must to guarantee for an end-to-end quality of service. Measurements according to ITU-T P.861 can be effectively applied to monitor, and to assure for certain quality levels inside a public or corporate network. This is especially true if telephony services are based on leased lines which are under the responsibility of third parties.

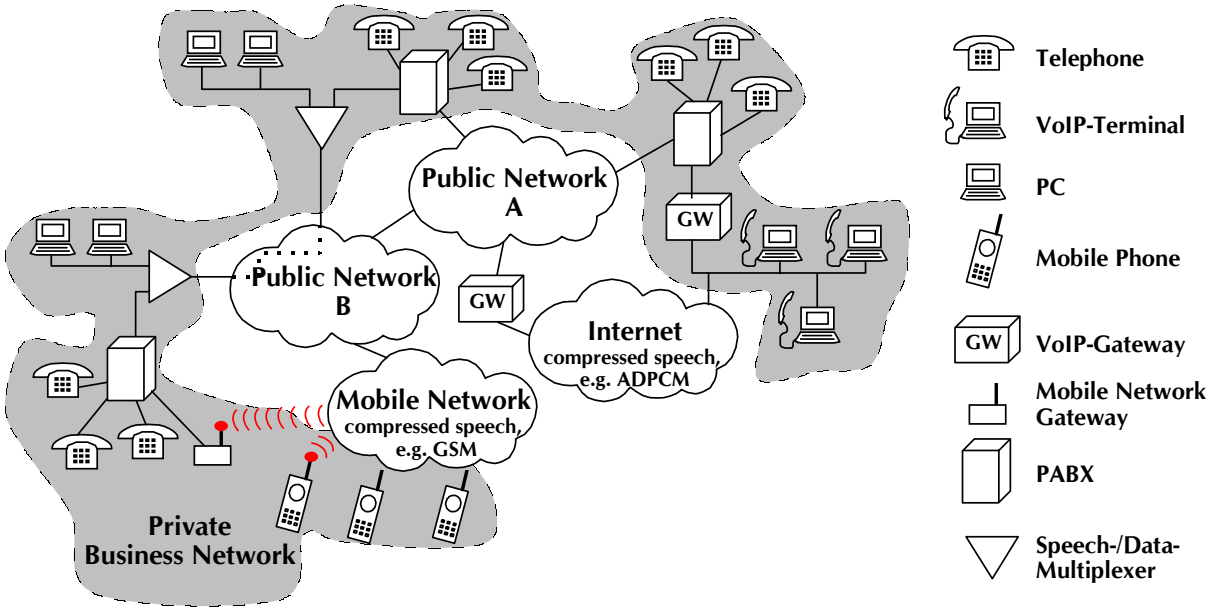


Figure 3: An example for the topology of today's telecommunication networks

3 WHAT IS QUALITY?

In the context of modern telecommunications, we often talk about **quality of service** ("QoS"). However, depending on the degree of perfection and the context, many aspects can be addressed as a matter of 'quality'. General considerations how to evaluate the quality standard of an international telephone service are addressed in ITU-T recommendation E.420. This recommendation outlines some aspects for quality of service, starting with principly being able to provide the customer with the ability to use the desired services. Following quality aspects are the level of service for reaching users in foreign countries (connection establishment, connection retention), the quality of the connection and also the billing integrity.

3.1 Call Connectivity

Call connectivity is characterized by the level of **connection establishment** and **connection retention**. Connection establishment can be characterized by the **answer seizure ratio** ("ASR"). This figure gives the percentage ratio of seizures, resulting in an answer, compared to the total number of seizures. For well developed networks, such as in the U.S. and in European countries, one could expect a ratio of 60...70%. Of course, such figures include user behaviour, as dialing the wrong number will also be counted.

3.2 Availability and Reliability

Once a connection can principally be established, the availability of the service around the hour (busy hours) as well as the reliability (e.g. dropped calls, wrong numbers) become an issue. Also the **post dial delay** ("PDD") can determine the degree of reliability, as perceived by customers. The PDD is defined as the time from the end of dialling until the start of ringing. Another parameter is also the **post gateway answer delay** ("PGAD").

3.3 Speech Intelligibility

An important prerequisite before we can start talking about speech quality is the question if at all we can understand the other party. If the speech quality is at the bottom end, it may be difficult to understand the context, especially in a foreign language. Speaking in terms of information theory, the question is then whether parts of the transmitted information are getting lost. In this case an averaged quality score is of course not of much help as it does not characterize the degree of missing information.

A common used term which is aimed to characterize speech intelligibility is "clarity", indicating how much information can be extracted from a conversation. Speech intelligibility depends on a large variety of factors, and only few are well understood. For example, certain frequency bands are more important for intelligibility than others: 250-800 Hz is less important for speech intelligibility than 1000-1200 Hz. Intelligibility also depends on the speech material. Complete sentences are usually much better understood than a sequence of unrelated words due to the logical word flow in a sentence. Subjective test procedures are defined based on spoken syllables, however these procedures cause too much effort to be applied in the daily operation.

3.4 Speech Quality

Once the call could be reliably established, and the voice at the other end can principally be understood, we talk of voice quality or speech quality. (We want to avoid the term audio quality, as this term is more related to wide band audio, such as 20 kHz audio bandwidth).

Figure 4 illustrates how speech intelligibility and speech quality are depending on the data rate, respectively the **bit error rate** (BER). First, it can be seen that the higher the bit rate, the more likely a good speech quality (not only intelligibility) will be obtained. However, the effect of bit errors increases with a lower data rate due to the increased lack of redundancy. As a summary, speech quality is first interfered by artefacts but with lower bit rates and thus more sensitivity to errors, speech intelligibility is interfered by information loss.

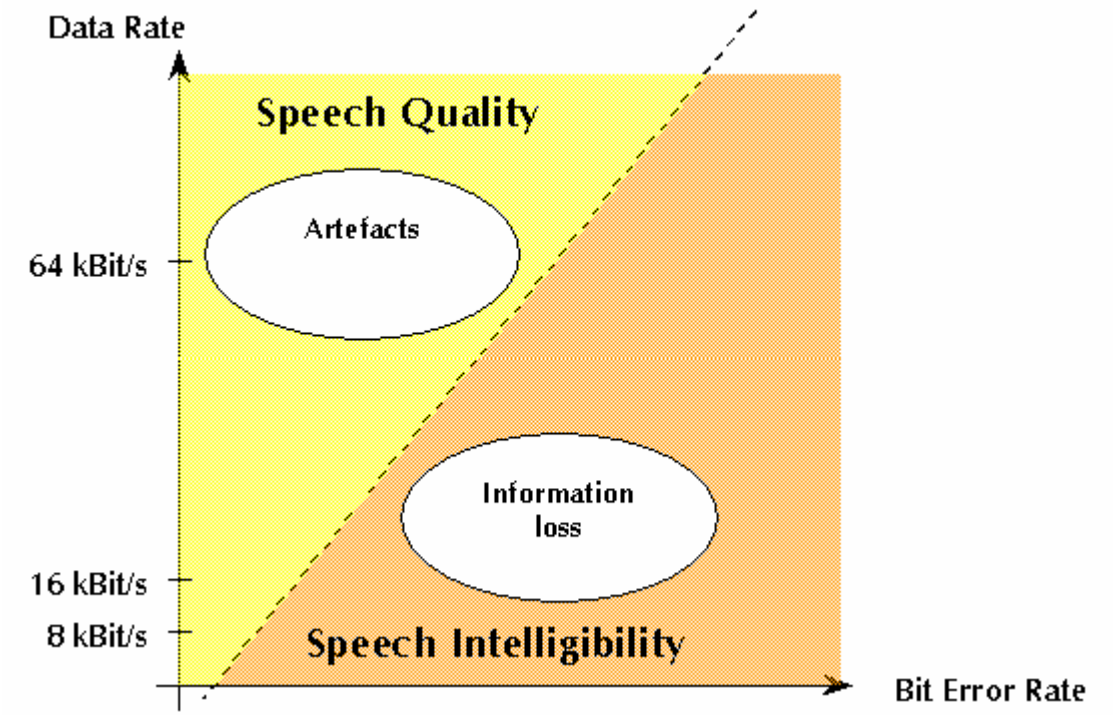


Figure 4: Speech intelligibility and speech quality in the context of Bit Error Rate and Data Rate

Voice quality can be measured by two ways, either by introducing test calls (an "intrusive" method), or by monitoring ("listening" to) the normal traffic ("non-intrusive"). In order to end up with reliable results, it is however mandatory to insert a well defined speech signal into the device under test. For this reason, state-of-the-art voice quality testing is always intrusive. This may, however, generate additional traffic, which is the reason why non-intrusive models like CCI (P.562) are under development. The speech quality is usually scored on an opinion scale ("MOS").

4 TESTING VOICE QUALITY

4.1 Subjective Testing

Due to the lack of international standards for measuring the perceived voice quality, until a few years ago, the only widely accepted assessment procedures for voice quality were listening tests.

Useful methods for testing telephone band speech signals were first standardized within the ITU-T[1]. Recommendation P.800 [16] defines the absolute category rating test method (ACR) which has been used for the assessment of speech codecs since 1993. Within the ACR test method, the ITU five grade impairment scale is applied (see Table 2). It should be noted, that because of the telecommunication environment, testing is done without a comparison to an undistorted reference. This compares with a

[1] International Telecommunication Union, Geneva, (former CCITT), see also http://www.itu.org

typical situation of a phone call, where the listener has no access to a comparison with a reference, for example the original voice of the other party. However, a listening test according to P.800 could be regarded as a comparison between a test signal and a reference "in the mind" of the listener. This is because of the fact that the listener is very familiar with the natural sound of a human voice.

For comparison reasons, and in order to be able to merge the results of different individuals, it is necessary to adjust the listeners' opinions to an absolute scale. For this purpose, predefined examples with well defined noise insertions of fixed modulated noise reference units (MNRU, [17]) are presented at the beginning of a test. Each sample represents an example distortion corresponding to the ITU-T version of the five grade impairment scale.

Impairment	Grade
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 2: The ITU-T five-grade impairment scale

Based on these test conditions a population of typically 20 to 50 test subjects will be presented with an identical series of speech fragments. Every test subject will be asked to score each sample by applying the impairment scale. After statistical processing of the individual results, a mean opinion score (MOS) can be calculated. With thorough setups, such test results can be reproduced quite well, even at different locations. Of course, the effort needed in terms of subjects and time is tremendous. And of course, such test methods can not be applied within a practical or field environment in the daily life.

4.2 Evolution of Objective Testing

The design of objective measurement methods based on human perception goes back to the eighties, and is based on the research work of Zwicker, Schröder, Brandenburg et al. The first algorithm that was implemented into a real measurement system was NMR [4] in 1989. The best known algorithms in the past were PAQM [2], PSQM [3], NMR [4], PERCEVAL [25], DIX [31], OASE [29], POM [6]. Except for PSQM, all of these were developed to assess the quality of wideband audio codecs. This is due to the fact that the widespread use of perceptual codecs started earlier in the broadcast environment, than it did in telecommunications. In 1996 PSQM was standardized as ITU-T Rec. P.861 for speech quality measurement. It showed superior correlation with subjective tests compared to all the other proposals that were not based on human perception. Out of PAQM, PSQM, NMR, PERCEVAL, DIX, OASE and POM, PEAQ was developed as a joint collaboration. PEAQ was standardized in 1998 as ITU-R Rec. BS.1387 for wideband audio testing. With the ongoing development of speech coding, especially for packet transmission, also newer algorithms for speech quality measurement were developed, like PSQM+, PSQM99, MNB, PAMS, TOSQA, PACE and VQI. Verification tests performed by the ITU showed that far the best of these was PSQM99. The second best was PAMS, but none of these proposals was good enough for a revision of the P.861 standard. Consequently PESQ was developed, which is PSQM99 with an improved delay compensation. PESQ was standardized in 2000 as ITU-T Draft Rec. P.862.

When comparing all of the relevant measurement algorithms they can be broken down to a block diagram as shown in figure 5. Although they significantly differ in the way they try to model human perception, they also show a very high degree of similarity in their basic structure.

This structure consists of two inputs, one for the (unprocessed) reference signal and one for the signal under test. The latter may for example be the output signal of a codec that is stimulated by the reference signal.

In a first signal processing step the peripheral ear is modelled ("perceptual model", or "ear model"). Of course, the implementations of the peripheral ear model differ widely between the various algorithms. In general one can say that for wideband audio signals this part of the algorithm is more important than for speech quality measures, and therefore it is modelled more accurately in e.g. PEAQ. Also one can observe significant improvements here between the first algorithms like PAQM or NMR and the latest developments like PEAQ. PEAQ probably uses the most accurate and most detailed perceptual model that has ever been implemented until today.

In a consecutive step, the algorithm models the audible distortion present in the signal under test by comparing the outputs of the ear models. The information obtained by this process is called MOVs ("Model Output Variables"), and may be useful for a detailed analysis of the signal.

The final goal instead is deriving a quality measure, consisting of a single number that indicates the audibility of the distortions present in the signal under test. To achieve this, some further processing of the MOVs is required, which simulates the cognitive part of the human auditory system. Various proposals exist for this step. They range from algorithmic descriptions (e.g. PESQ) to artificial neural networks (e.g. PEAQ). While most algorithms require time aligned input signals, the process how to achieve this is usually not part of the model description. Just with the newer speech quality measures like PESQ, the delay compensation is an integral part of the model.

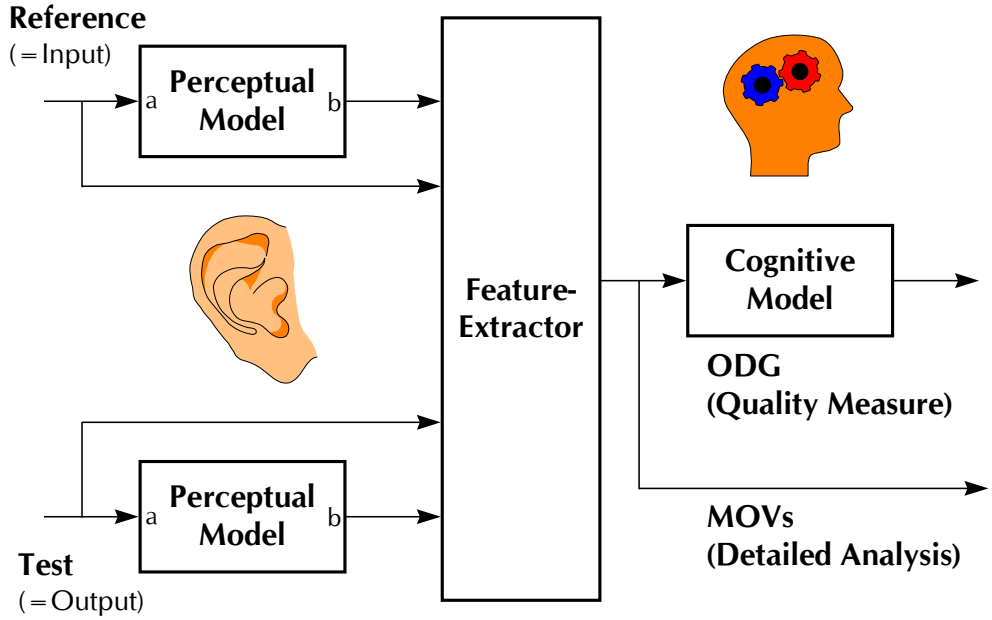


Figure 5: The structure of the generic perceptual measurement algorithm

Summary:

As a summary, we can note that objective testing of voice quality based on perceptual techniques works, because it models both, **the human ear** (perceptual modelling) and **the judgement behaviour** of a test subject (modelling the brain).

4.3 International Standardization

International standardization of perceptual audio measurement techniques was mainly driven by two expert groups within the International Telecommunications Union (ITU).

Within the telecommunication sector of the ITU, in 1996 study group 12 finalized recommendation P.861 [19] for the objective analysis of speech codecs. After a wide-ranging comparison of proposed methods, the group opted for the PSQM algorithm. PSQM correlated up to 98 percent with the scores of subjective listening tests.

In 1999/2000 P.861 was revised. This resulted in the draft recommendation P.862 ("PESQ") which is more suitable for measurements on real networks than P.861 was.

Within the study period 1994 – 1998 the ITU-R had established task group 10/4 with the scope to recommend an objective, perceptual based model to evaluate the quality of wide band audio codecs. After collecting a set of proposals, the group of model proponents opted for a joint collaboration to derive an improved model. The model was recommended as a measure for the perceived audio quality ("PEAQ") under recommendation BS.1387 in late 1998 after thorough verification.

All three standards, ITU-T P.861, ITU-T P.862 and ITU-R BS.1387, today represent the state-of-the-art technique for the objective evaluation of the perceived speech/audio quality. It should be noted, however, that all of these techniques were derived from modelling the corresponding subjective experiment by an algorithm based approach. Thus it is essential to understand the scope of the modelled subjective experiment when trying to interpret the calculated results.

4.4 PSQM, PSQM+

PSQM

The algorithm to calculate the perceptual speech quality measure (PSQM) was introduced by Beerends in 1993 [3]. This development by KPN Research represents an adapted version of the more general perceptual audio quality measure (PAQM) [2], optimized for telephony speech signals. This is due to the observation that the psychoacoustic effects known from masking experiments seem to differ in significance, when comparing the perception of speech and music signals. One reason might be that the human brain possibly recalls the reference sound of familiar voices more accurately from the daily life experience, compared to music sounds. Up to now, no single homogeneous approach has been presented that would allow for high correlation with both, speech, and music signals without adapting algorithm parameters [1].

Figure 6 shows a block diagram of a basic model of the PSQM algorithm. Within PSQM, the physical signals constituting the source and coded speech are mapped onto psychophysical representations that match the internal representations of the speech signals (the representations inside our heads) as closely as possible.

As depicted in Figure 6, the quality of the coded speech is judged on the basis of differences in the internal representation. This difference is used for the calculation of the noise disturbance as a function of time and frequency. In PSQM, the average noise disturbance is directly related to the quality of coded speech.

Besides perceptual modelling, the PSQM method also uses cognitive modelling in order to get high correlations between subjective and objective measurements [19]. The result is the estimated quality of the received signal.

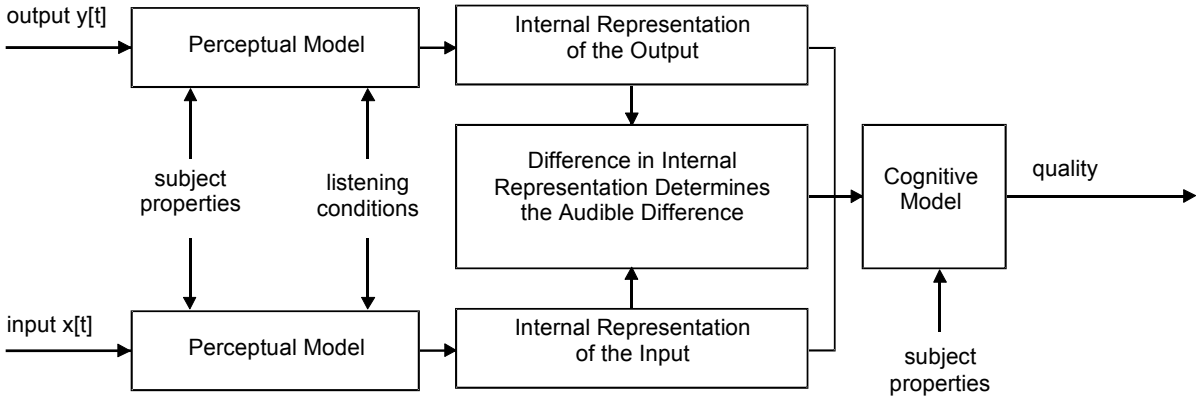


Figure 6: Block diagram of the basic model of the PSQM algorithm [19]

The PSQM algorithm was one out of several proposals that had been input to study group 12 of ITU-T in 1995 for the purpose of international verification. Further proposals were the EPR Algorithm ("Expert Pattern Recognition"), which consisted of measures of the "LPC Cepstrum Function", "Information Index", and the "Coherence Function" (CHF). In a test series conducted by the Japanese phone corporation NTT, including listening tests in Japan and Italy, the highest correlation was achieved with PSQM results, when compared to the subjective tests. Consequently, PSQM was recommended by the ITU-T in 1996 for the objective quality measurement of telephone band speech codecs. Since then, it has been used intensively for R&D applications, and is now more and more applied to field applications in networks.

PSQM+

The standard version of PSQM as defined by P.861 has three major drawbacks:

- The time alignment as defined by P.861 is very crude and not suitable for practical measurements on noisy lines. The standard time alignment is based on the detection of the first 0.5ms window where the reference and the test signal exceed a certain minimum energy. This point is taken as the starting point of both files. Any noise captured during the measurement may also trigger this detection and thus lead to entirely wrong results. In OPERA™ this problem does not exist anymore. OPERA™ adds a sophisticated time alignment algorithm to PSQM and PSQM+. This OPERA™ time alignment works perfectly for files with a constant delay and may be used for signals with varying delays too. However, as this advanced time alignment is not part of the standard, other PSQM implementations may not handle delay properly and may fail totally on varying delays.

- The asymmetry processing of PSQM weights loud distortions much stronger than a human listener would do. PSQM+ uses some special means to overcome this.

- On time clipped passages (e.g. caused by dropouts or packet loss) the opposite effect shows up. These distortions are not enough taken into account. PSQM+ uses a special scaling algorithm to eliminate this problem.

A verification of these modifications of the standard showed a significant improvement of the correlation between the measured PSQM+ scores and the subjective listening test results, when compared to PSQM. These improvements make PSQM+ much more suitable for measurements on VoIP networks than the standard PSQM.

4.5 PESQ

When PSQM was standardized as P.861, the scope of the standard were at that time state of the art codecs as they were mainly used for mobile transmission, like GSM. VoIP was not yet a topic at this time. The requirements for measurement equipment have changed dramatically since then. As a consequence, the ITU setup a working group for revision of the P.861 standard to cope with the new demands arising from modern networks like VoIP. With these networks the measurement algorithm has to deal with much higher distortions as with GSM codecs, but maybe the most eminent factor is that the delay between the reference and the test signal is not constant anymore.

A first approach to overcome these problems was the development of PSQM+. It could well handle the larger distortions as they are caused by e.g. packet loss, but still had significant problems with the compensation of the varying delay. OPTICOM added in it's OPERA™ system a Delay Tracking feature that implemented an easy way to solve that issue in most cases, without loosing the option of realtime operation. Although this feature failed for some signals, it was the only available method up to date to achieve reliable results for the speech quality of VoIP networks.

With the new draft ITU standard P.862 (PESQ) this problem is now finally eliminated. It combines the excellent psychoacoustic and cognitive model of PSQM+ with a time alignment algorithm that perfectly handles varying delays. The only drawback of PESQ is that it is absolutely not designed for realtime applications. This is in turn why it cannot fully replace PSQM+. With PSQM and PESQ there will soon be two standards that cover the entire problem of measuring speech quality. Figure 7 gives an overview of the structure of the PESQ algorithm and shows the new blocks which have been added to the PSQM algorithm.

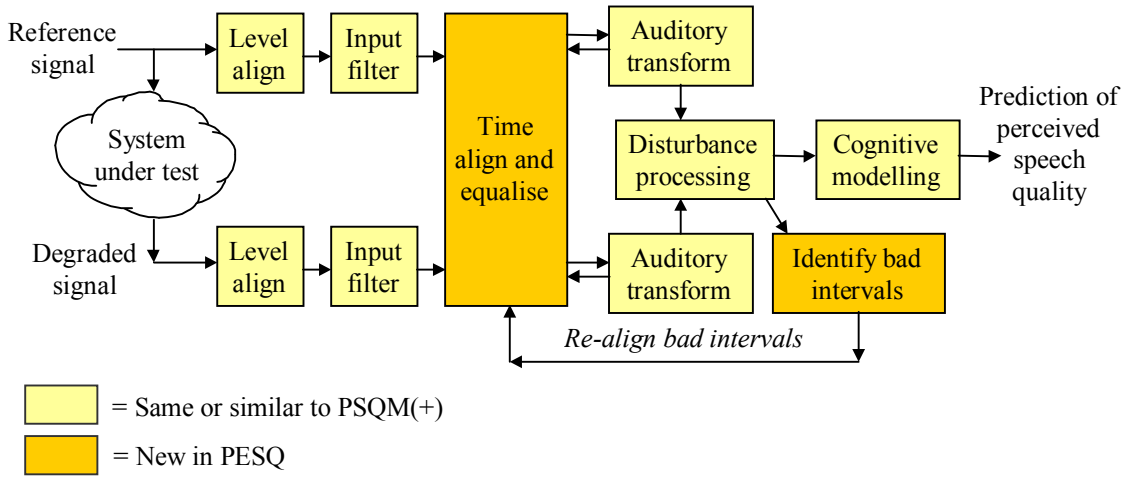


Figure 7: The structure of the PESQ algorithm

5 OPERA™ – A COMPREHENSIVE TEST TOOL

OPTICOM's new generation of quality testers, called OPERA™ – short for „Objective Perceptual Analyzer" – represent the latest development to objectively evaluate and assure the quality of compressed speech and wide-band audio signals by modelling the human ear: OPERA™ is not only suitable for assessing a single processing device. With OPERA™ you can achieve a comprehensive analysis of the end-to-end quality, from the studio source to the receiver, or from the caller to the called. OPERA's flexible scalability may range from a single stand alone tester up to powerful network-wide setups with distributed systems sharing information over IP.

The open framework concept of OPERA™ allows to add advanced measurement algorithms as plug-ins in the future as soon as they will become available. In addition, user defined measurement algorithms may be integrated upon request. OPERA™ is available as a software suite, a completely pre-installed workstation system and a completely pre-installed portable system. The operating system used is Microsoft Windows NT 4.0.

In the following we want to give you a short overview of the functionality and the operation of OPERA™. Basically, a measurement of end-to-end quality comprises two steps, first signal acquisition and second the analysis of the test signal. OPERA™ allows to perform measurements interactively or fully automated according to a predefined schedule.

5.1 Signal Acquisition

There are several kinds of signal sources available with the OPERA™ measurement system:

- WAVE files
- Analog loop start interfaces
- External VoIP gateways / terminal adapters
- Sound board interface

WAVE Files

In this case OPERA™ compares audio files. Supported file formats are WAVE files containing either plain PCM and a-law or μ-law. The PSQM algorithm is defined for sampling rates of 8 kHz and 16 kHz.

Analog Loop Start Interfaces

When assessing POTS (plain old telephone system), one can connect the OPERA™ system to the network using the analog loop start interfaces of a voice interface board. The available signal acquisition software **OptiCall™** is used to perform test calls. OptiCall™ sends a reference file containing speech through the

connected network and records the received signal. The reference signal and the recorded test signal are stored in WAVE file format which can then be analyzed with the OPERA™ analyzer.

External VoIP gateways / terminal adapters

Currently external VoIP gateways (or VoIP terminal adapters, respectively) are used to connect the analog loop start interfaces to an IP network. The signal acquisition is handled by the OptiCall™ software as described above. OPTICOM is working on a solution that captures the data directly from the IP network without the need of external gateways and additional D/A and A/D conversion.

Sound Board Interface

Online measurements with phone lines are currently not supported. However, by using a sound board interface, OPERA™ can perform online analysis of audio sources. In addition, the **OptiRec™** application will soon be available for recording WAVE files from audio sources.

5.2 Interactive Operation for the Developer

Once the user has chosen the signal sources, he is ready to analyze his test signal. Therefore OPERA™ includes an analyzer that offers a detailed view on the measurement results. With this tool the user may measure the quality according to PSQM, PSQM+, PESQ, as well as delay and echo. Figure 8 shows a screenshot of a PSQM measurement.

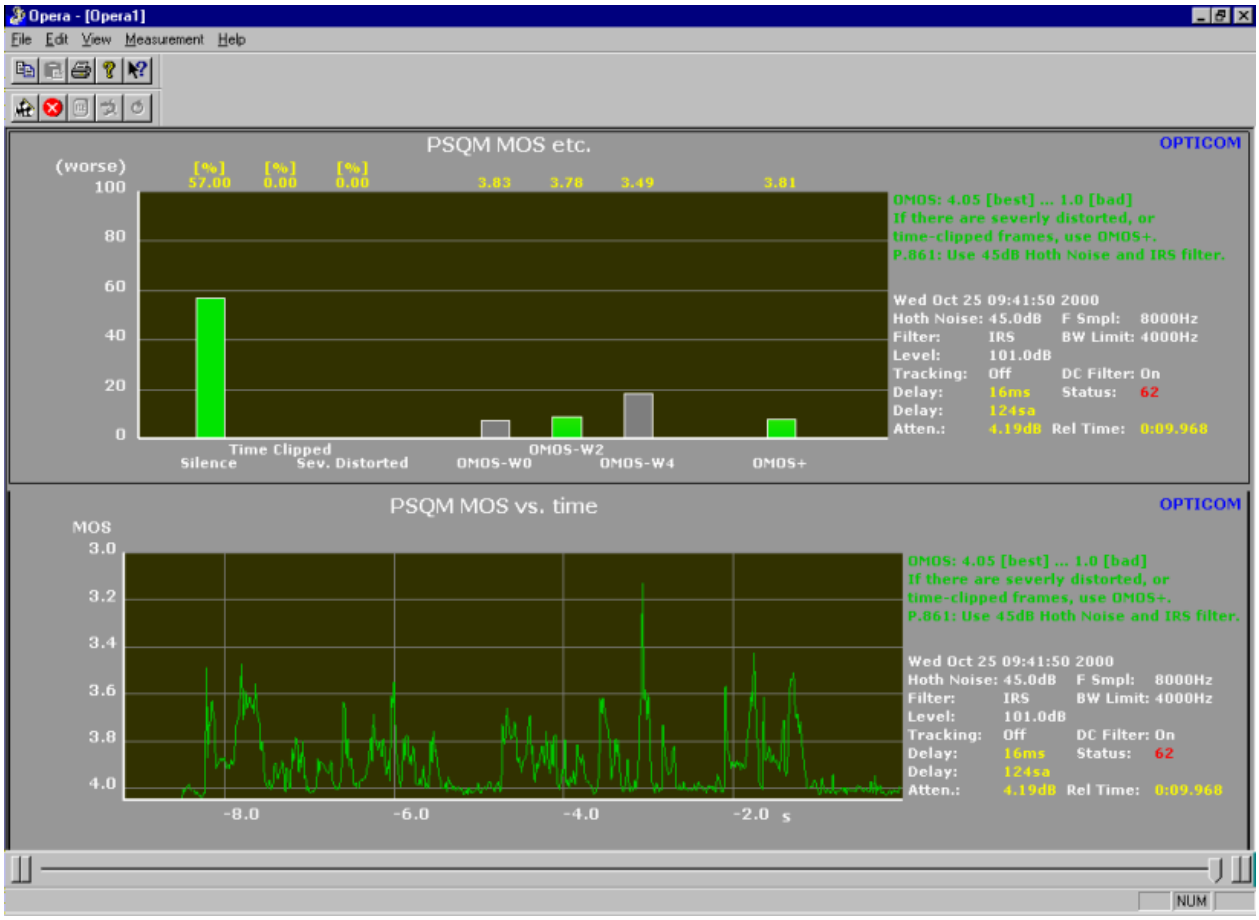
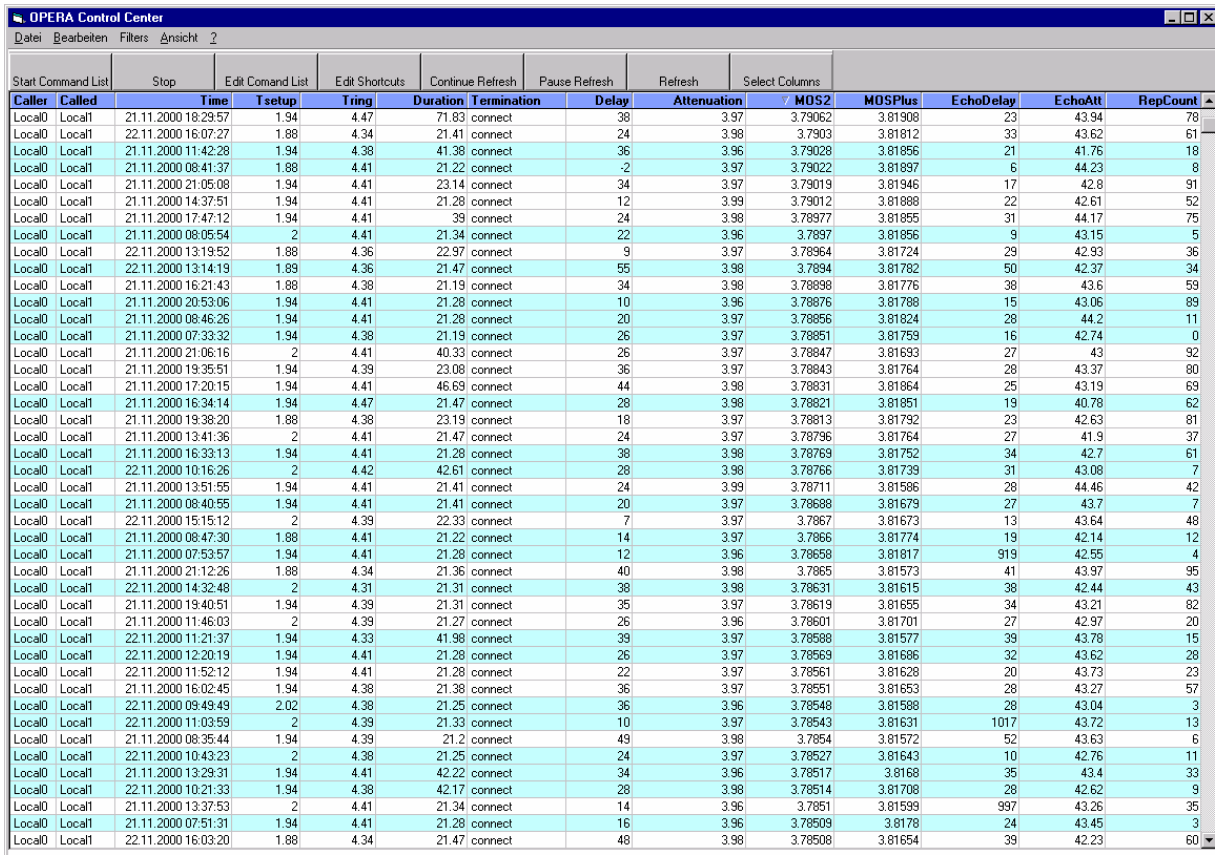


Figure 8: Screenshot of the OPERA™ analysis tool

5.3 Automated Operation for the Operator

The OPERA™ Control Center (see Figure 9) provides the opportunity of performing automated test calls. The Control Center takes care of both, the signal acquisition process and the analysis. With this tool the user may arrange a schedule where he defines the test calls to be performed, e.g. during a certain period of the day. Running automated measurements repeatedly, he gets a reliable overview of the quality of service on the examined network. Operating in the automated mode, something more than 3000 test calls a day can be analyzed by a single OPERA™ system.



Caller	Called	Time	Tsetup	Tring	Duration	Termination	Delay	Attenuation	MOS2	MOSPlus	EchoDelay	EchoAtt	RepCount
Local0	Local1	21.11.2000 18:29:57	1.94	4.47	71.83	connect	38	3.97	3.79062	3.81908	23	43.94	78
Local0	Local1	22.11.2000 16:07:27	1.88	4.34	21.41	connect	24	3.98	3.7903	3.81812	33	43.62	61
Local0	Local1	21.11.2000 11:42:28	1.94	4.38	41.38	connect	36	3.96	3.79028	3.81856	21	41.76	18
Local0	Local1	21.11.2000 08:41:37	1.88	4.41	21.22	connect	-2	3.97	3.79022	3.81897	6	44.23	8
Local0	Local1	21.11.2000 21:05:08	1.94	4.41	23.14	connect	34	3.97	3.79019	3.81946	17	42.8	91
Local0	Local1	21.11.2000 14:37:51	1.94	4.41	21.28	connect	12	3.99	3.79012	3.81888	22	42.61	52
Local0	Local1	21.11.2000 17:47:12	1.94	4.41	39	connect	24	3.98	3.78977	3.81855	31	44.17	75
Local0	Local1	21.11.2000 08:05:54	2	4.41	21.34	connect	22	3.96	3.7897	3.81856	9	43.15	5
Local0	Local1	22.11.2000 13:19:52	1.88	4.36	22.97	connect	9	3.97	3.78964	3.81724	29	42.93	36
Local0	Local1	22.11.2000 13:14:19	1.89	4.36	21.47	connect	55	3.98	3.7894	3.81782	50	42.37	34
Local0	Local1	21.11.2000 16:21:43	1.88	4.38	21.19	connect	34	3.98	3.78898	3.81776	38	43.6	59
Local0	Local1	21.11.2000 20:53:06	1.94	4.41	21.28	connect	10	3.96	3.78876	3.81788	15	43.06	89
Local0	Local1	21.11.2000 08:46:26	1.94	4.41	21.28	connect	20	3.97	3.78856	3.81824	28	44.2	11
Local0	Local1	21.11.2000 07:33:32	1.94	4.38	21.19	connect	26	3.97	3.78851	3.81759	16	42.74	0
Local0	Local1	21.11.2000 21:06:16	2	4.41	40.33	connect	26	3.97	3.78847	3.81693	27	43	92
Local0	Local1	21.11.2000 19:35:51	1.94	4.39	23.08	connect	36	3.97	3.78843	3.81764	28	43.37	80
Local0	Local1	21.11.2000 17:20:15	1.94	4.41	46.69	connect	44	3.98	3.78831	3.81864	25	43.19	69
Local0	Local1	21.11.2000 16:34:14	1.94	4.47	21.47	connect	28	3.98	3.78821	3.81851	19	40.78	62
Local0	Local1	21.11.2000 19:38:20	1.88	4.38	23.19	connect	18	3.97	3.78813	3.81792	23	42.63	81
Local0	Local1	21.11.2000 13:41:36	2	4.41	21.47	connect	24	3.97	3.78796	3.81764	27	41.9	37
Local0	Local1	21.11.2000 16:33:13	1.94	4.41	21.28	connect	38	3.98	3.78769	3.81752	34	42.7	61
Local0	Local1	22.11.2000 10:16:26	2	4.42	42.61	connect	28	3.98	3.78766	3.81739	31	43.08	7
Local0	Local1	21.11.2000 13:51:55	1.94	4.41	21.41	connect	24	3.99	3.78711	3.81586	28	44.46	42
Local0	Local1	21.11.2000 08:40:55	1.94	4.41	21.41	connect	20	3.97	3.78688	3.81679	27	43.7	7
Local0	Local1	22.11.2000 15:15:12	2	4.39	22.33	connect	7	3.97	3.7867	3.81673	13	43.64	48
Local0	Local1	21.11.2000 08:47:30	1.88	4.41	21.22	connect	14	3.97	3.7866	3.81774	19	42.14	12
Local0	Local1	21.11.2000 07:53:57	1.94	4.41	21.28	connect	12	3.96	3.78658	3.81817	919	42.55	4
Local0	Local1	21.11.2000 21:12:26	1.88	4.34	21.36	connect	40	3.98	3.7865	3.81573	41	43.97	95
Local0	Local1	22.11.2000 14:32:48	2	4.31	21.31	connect	38	3.98	3.78631	3.81615	38	42.44	43
Local0	Local1	21.11.2000 19:40:51	1.94	4.39	21.31	connect	35	3.97	3.78619	3.81655	34	43.21	82
Local0	Local1	21.11.2000 11:46:03	2	4.39	21.27	connect	26	3.96	3.78601	3.81701	27	42.97	20
Local0	Local1	22.11.2000 11:21:37	1.94	4.33	41.98	connect	39	3.97	3.78588	3.81577	39	43.78	15
Local0	Local1	22.11.2000 12:20:19	1.94	4.41	21.28	connect	26	3.97	3.78569	3.81686	32	43.62	28
Local0	Local1	22.11.2000 11:52:12	1.94	4.41	21.28	connect	22	3.97	3.78561	3.81628	20	43.73	23
Local0	Local1	21.11.2000 16:02:45	1.94	4.38	21.38	connect	36	3.97	3.78551	3.81653	28	43.27	57
Local0	Local1	22.11.2000 09:49:49	2.02	4.38	21.25	connect	36	3.96	3.78548	3.81588	28	43.04	3
Local0	Local1	22.11.2000 11:03:59	2	4.39	21.33	connect	10	3.97	3.78543	3.81631	1017	43.72	13
Local0	Local1	21.11.2000 08:35:44	1.94	4.39	21.2	connect	49	3.98	3.7854	3.81572	52	43.63	6
Local0	Local1	22.11.2000 10:43:23	2	4.38	21.25	connect	24	3.97	3.78527	3.81643	10	42.76	11
Local0	Local1	21.11.2000 13:29:31	1.94	4.41	42.22	connect	34	3.96	3.78517	3.8168	35	43.4	33
Local0	Local1	22.11.2000 10:21:33	1.94	4.38	42.17	connect	28	3.98	3.78514	3.81708	28	42.62	9
Local0	Local1	21.11.2000 13:37:53	2	4.41	21.34	connect	14	3.96	3.7851	3.81599	997	43.26	35
Local0	Local1	21.11.2000 07:51:31	1.94	4.41	21.28	connect	16	3.96	3.78509	3.8178	24	43.45	3
Local0	Local1	22.11.2000 16:03:20	1.88	4.34	21.47	connect	48	3.98	3.78508	3.81654	39	42.23	60

Figure 9: : Screenshot of the OPERA™ Control Center

Such automated test call performances are not only restricted to one single OPERA™ system, the Control Center also manages measurements with distributed OPERA™ measurement systems that intercommunicate via an IP-connection. Figure 10 shows an example of monitoring the quality of a telecommunication network with distributed OPERA™ measurement units.

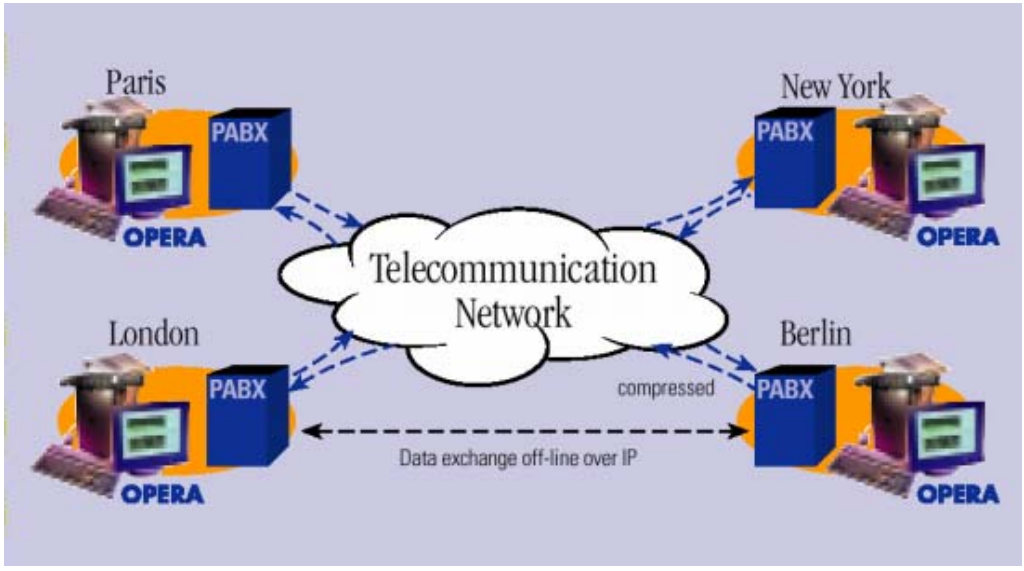


Figure 10: Example of a distributed system of OPERA™ measurement units

5.4 Example Results

Table 3 shows some measurement results obtained from applying PESQ, PSQM and PSQM+ to a Voice over IP network. OPERA™ was attached to this network through two VoIP Gateways. The codecs used in this example were either G.711 (64 kBit) or G.723 (6.3 kBit) codecs. First we performed the measurement on the network without any load and then on a heavily loaded network in order to monitor the influence of the network congestion on the speech quality.

	PESQ	PSQM	PSQM+	PSQM(v)	PSQM+(v)	Delay(PESQ)
G.711 No Load	3.68	3.23	3.62	3.77	3.83	70..100ms
G.711 Load	1.45	*1.00*	*1.00*	*1.00*	*1.00*	200..800ms
G.723 No Load	3.32	2.93	3.54	2.91	3.53	206ms
G.723 Load	1.35	*1.00*	*1.00*	*1.00*	*1.00*	250..500ms

Table 3: Measurement results obtained from applying PESQ, PSQM and PSQM+ to a Voice over IP network

Since PESQ is the measure recommended for this situation, we take these results as a reference to evaluate the suitability of PSQM and PSQM+ for this application, too. By looking at the PESQ values only, it becomes obvious that the network load has a significant impact on the speech quality. For the G.711 codec the quality drops from 3.68 to only 1.45 for the busy network. This quality drop is about the same for the G.723 codec, but due to the higher compression rate one could transfer ten times as many calls on the same network compared to G.711 without any significant loss of speech quality.

The effect of the network load is not only visible by looking at the speech quality, which is at the lower limit of the scale for all three measures. There is also a significant influence on the delay of the line, especially when combined with G.711. Although G.723 shows a much higher minimum delay, the delay variation is much less than with G.711. The latter shows a delay of 435ms ±84%(!), whereas G.723 varies just ±33% around and average of 350ms.

If these results are now compared to those obtained from measurements with PSQM or PSQM+, it appears that both measures may surprisingly well be used for the assessment of the unloaded network, too. Of course this is mainly due to the time alignment algorithm implemented in the OPERA™ system. If the delay tracking feature of OPERA™ is switched on (see columns PSQM(v), PSQM+(v) in Table 3) the

MOS becomes slightly better. This is obvious since inefficient signal preprocessing may contribute to measure distortions which are not present in the assessed signal.

This example demonstrates not only the use of PESQ for VoIP applications, furthermore it also shows that you may well use PSQM or better PSQM+ as long as you combine it with an advanced and robust time alignment algorithm. Of course the preference should always be to use PESQ, but if it is not possible because of e.g. realtime requirements, you are still fine with PSQM+, even for VoIP applications.

6 ABOUT OPTICOM

OPTICOM, the world leader in perceptual voice and audio quality testing solutions, and the technologies provider of techniques such as PSQM, PSQM+, PEAQ and PESQ addresses the testing advantages of utilizing ITU's current and proposed standards for today's and future networks.

Under the mission statement "quality is our business ", OPTICOM focuses on top notch developments to gain for its customers improved quality in audio and video communications. With the new OPERA™ family of perceptual analyzers, the company proves it's worldwide reputation for state-of-the-art solutions to improve the audio quality of new media.

OPTICOM was founded by its President Michael Keyhl in 1995 as a "spin-off" company of the Fraunhofer-Institute, Germany's leading organization for applied research. OPTICOM's developers benefit from their broad experience in the research and development of perceptual based coding and evaluation techniques, such as MP3 and NMR, lasting back to the late 1980's.

Through many international contacts and cooperations with leading research organizations, OPTICOM has today gained an active role in the international standardization business, e.g. of the new ITU-R standard "PEAQ". OPTICOM is also continuously active in, or observing the work of the AES, EBU, ITU-T, ETSI, ISO/MPEG and others.

After being successfully in business for more than four years, the company is fast growing and seeking to expand the number of their employees. OPTICOM is located in Erlangen, Northern-Bavaria, GERMANY, and has just recently opened offices and distributionship channels in the USA and Asia. For more information, please feel free to visit **www.opticom.de**.

7 REFERENCES

[1] BEERENDS J. G., *Measuring the Quality of Speech and Music Codecs, an Integrated Psychoacoustic Approach*, 98th AES Convention, Paris 1995, Preprint #3945

[2] BEERENDS J. G., STEMERDINK J. A., *A perceptual audio quality measure based on a psychoacoustic sound representation*, J. Audio Eng. Soc., Vol. 40, No. 12, pp. 963-987, 1992

[3] BEERENDS J. G., STEMERDINK J. A., *A perceptual speech quality measure based on a psychoacoustic sound representation*, J. Audio Eng. Soc., Vol. 42, No. 3, pp. 115-123, 1994

[4] BRANDENBURG K., *Evaluation of Quality for Audio Encoding at low Bit Rates*, 82nd AES Convention, London 1987, Preprint #2433

[5] BRANDENBURG K., SPORER Th.*: 'NMR' and 'masking flag': Evaluation of Quality using Perceptual Criteria*, Proc. of the 11th International AES Conference on Audio Test and Measurement, Portland 1992, pp. 169-179

[6] COLOMES C., LEVER M., RAULT J.B., DEHERY Y.F., *A perceptual model applied to audio bit-rate reduction*, J. Audio Eng. Soc., Vol. 43, pp. 233-240, 1995

[7] ETSI Technical Report ETR 250, *Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks*, ETSI 1996

[8] FLANAGEN J. L., *Speech Analysis, Synthesis and Perception,* Springer, Berlin - Heidelberg - New York, 1972

[9] HERRE J., EBERLEIN E., SCHOTT H., SCHMIDMER Ch., *Analysis Tool for Realtime Measurements using Perceptual Criteria*", Proc. of the 11th International AES Conference on Audio Test and Measurement, Portland 1992, pp.180-190

[10] ISO/IEC/JTC1/SC29/WG11 Draft Document N1557, *Evaluation Methods and procedures for MPEG-4 tests*, 1997

[11] ITU-R Recommendation BS.562-3, *Subjective assessment of sound quality*

[12] ITU-R Recommendation BS.1116-1, *Methods for the Subjective Assessment of small Impairments in Audio Systems including Multichannel Sound Systems*, 1997

[13] ITU-R Recommendation BS.1387, *Method for Objective Measurements of Perceived Audio Quality (PEAQ)*, 1998

[14] ITU-T Contribution COM12-74-E, *Review of Validation Tests for Objective Speech Quality Measures*, March 1996

[15] ITU-T Recommendation E.420, *Checking the Quality of the International Telephone Service – General Considerations,* 1988

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

[17] ITU-T Recommendation P.810, *Modulated Noise Reference Unit (MNRU),* 1996

[18] ITU-T Recommendation P.830, *Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs*, 1996

[19] ITU-T Recommendation P.861, *Objective Quality measurement of telephone-band (300 - 3400 Hz) speech codecs*, 1996

[20] KARJALEINEN M., *A New Auditory Model for the Evaluation of Sound Quality of Audio Systems*, Proc. of the ICASSP 1985, pp. 608-611

[21] KEYHL M., HERRE J., SCHMIDMER Ch., *NMR Measurements of Consumer Recording Devices Which Use Low Bit-Rate Audio Coding*, 94th AES Convention, Berlin 1993, Preprint #3616

[22] KEYHL M., HERRE J., SCHMIDMER Ch., *NMR Measurements on Multiple Generations Audio Coding*, 96th AES Convention, Amsterdam, 1994, Preprint #3803

[23] KEYHL M., SCHMIDMER Ch., HERRE J., HILPERT J., *Maintaining Sound Quality - Experiences and Constraints of Perceptual Measurements in Today´s and Future Networks*", 98th AES Convention, Paris, 1995, Preprint #3946

[24] KEYHL M., SCHMIDMER Ch., WACHTER H., *A Combined Measurement Tool for the Objective, Perceptual Based Evaluation of Compressed Speech and Audio Signals,* 106th AES Convention, Munich, 1999, Preprint #4931

[25] PAILLARD B., MABILLEAU P., MORISETTE S., SOUMAGNE J., *PERCEVAL: Perceptual evaluation of the quality of audio signals*, J. Audio Eng. Soc., Vol. 40, 21-31, 1992

[26] PA&SQM PC-Software, Version 6.0, *User Manual*, OPTICOM 1997

[27] PRACHT St., *Voice Quality*, COMMUNICATE, November 1998, p. 43-46

[28] SPORER Th., *Evaluating Small Impairments with the Mean Opinion Scale - Reliable or Just a Guess?*, 101st AES Convention 1996, Preprint #4396 (E-1)

[29] SPORER Th., *Objective Audio Signal Evaluation - Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio*, 103rd AES Convention, New York, 1997 Preprint #4512

[30] TERHARDT E., *Calculating Virtual Pitch*, Hearing Research, Vol. 1, 1979, p. 155-182

[31] THIEDE Th., KABOT E., *A New Perceptual Quality Measure for Bit Rate Reduced Audio*, 100th AES Convention, Copenhagen, 1996, Preprint #4280

[32] VARY P., HEUTE U., HESS W., *Digitale Sprachsignalverarbeitung, Teubner-Verlag, Stuttgart, 1998*

[33] ZWICKER E., FELDTKELLER R., *Das Ohr als Nachrichtenempfänger*, Hirzel-Verlag, Stuttgart, 1967

[34] ZWICKER E., *Psychoakustik*, Springer-Verlag, Berlin - Heidelberg - New York, 1982

8 GLOSSARY OF TERMS

ACR
Absolute category rating test method according to the ITU-T recommendation P.800 used for the assessment of speech codecs. Within the ACR test method, a five grade impairment scale is applied. Because of the telecommunication environment the testing is done without a comparison to an undistorted reference.

ADPCM
Adaptive Difference-Pulse-Code-Modulation. According to standard ITU-T G.726. Bit rate of 32 kbit/s (also possible with 16, 24 and 40 kbit/s).ADPCM

Artefact
Spurious effects or imperfections introduced into a signal as a result of digital signal processing.

ASR
Answer Seizure Ratio. Defines the ratio between sucessful call attempts and the total number of calls.

BER
Bit Error Rate.

CCI
Call Clarity Index.

CELP
Code Excited Linear Prediction.

CTI
Computer Telephony Integration.

ETSI
European Telecommunications Standardization Institute.

IETF
Internet Engineering Task Force.

ISDN
Integrated Services Digital Network.

ITU-R
International Telecommunication Union, Geneva, Radiocommunciation sector (former CCIR), see also http://www.itu.org.

ITU-T
International Telecommunication Union, Geneva, Telecommunication sector, (former CCITT), see also http://www.itu.org.

LD-CELP
Low-Delay CELP Speech Coder. According to standard ITU-T G.726. Bit rate of 16 kbit/s.

MNRU
Modulated noise reference units.

MOS
Mean listening-quality Opinion Score, or simply Mean Opinion Score. The MOS is the mean of the given scores for a device under test of all test subjects in a subjective listening test.

MOV
The Model Output Variables are intermediate output values of the perceptual measurement method. These variables are based on basic psycho-acoustical findings and may therefore be used to characterize the coding artefacts further.

MPLS
Multi-Protocol Label Switching.

NMR
The measurement scheme NMR (Noise-to-Masked-Ratio) [Brandenburg, 1987] evaluates the level-difference between the masked threshold and the noise signal. A DFT with a Hann window of about 20 ms is used to analyse the frequency content of the signal. The transform coefficients are combined to

bands according to the Bark scale. The masked threshold is estimated for each band. The slope of the masked threshold is derived using a worst case approach taking into account the fact that the slopes are steeper for weak signals but run into the absolute threshold at higher levels. The absolute threshold is adapted to the resolution of the input signal (usually 16 bits), but not to psycho-acoustic demands. Due to these facts NMR is robust to changes of the reproduction level. The pitch scale resolution is about 1 Bark. Since the required computational power is low it was possible to implement NMR as a real time system at an early stage of its development.

The model has been in use since 1987 and has proven its basic reliability.

The most important output values of NMR are the masking flag rate, giving the percentage of frames with audible distortions, as well as the total and mean NMR which are different ways of averaging the distance between the error energy and the masked threshold.

Off-Line measurements
Measurement procedure which does not interact with the ongoing programme transmission.

On-Line measurements
Measurement procedure which relies on the ongoing programme transmission, or parts thereof.

PBX
Private Branch Exchange.

PCM
Pulse Code Modulation. According to standard ITU-T G.711. Bit rate of 64 kbit/s.

PDD
Abbrevation for Post Dial Delay. Refers to the time elapsed between the last dial tone and the first response of the network.

PEAQ
Stands for "Perceptual Evaluation of Audio Quality", the perceptual measurement technique recommended for wide band (music) audio signals as ITU-R BS.1387 in 1999. See also www.peaq.org.

PESQ
Stands for "Perceptual Evaluation of Speech Quality", an update to ITU-T P.861. Expected to be released in 2001 as ITU-T P.862. See also www.pesq.org.

PGAD
Post Gateway Answer Delay.

POTS
Plain Old Telephony Service (often used to characterize the traditional analog telephone service).

PSQM
Stands for "Perceptual Speech Quality Measure", the perceptual measurement technique recommended in ITU-T P.861. See also www.psqm.org.

PSTN
Public Switched Telephone Network.

QoS
Quality of Service.

Reference Signal
Test excerpt, reproduced without the processing by a test object, used as a comparison basis for an impairment test.

RSVP
Resource Reservation Protocol.

Subject
A test person evaluating the stimuli in a listening test.

TAPI
Telephony API (An application protocol interface defined by Microsoft, Intel).

ToS
Type of Service.

VoIP
Voice over Internet Protocol, a series of techniques permitting transmission of telephony over the

Internet. Often makes use of ITU-T G.7xx audio compression recommendations.

WFQ
Weighted Fair Queuing.