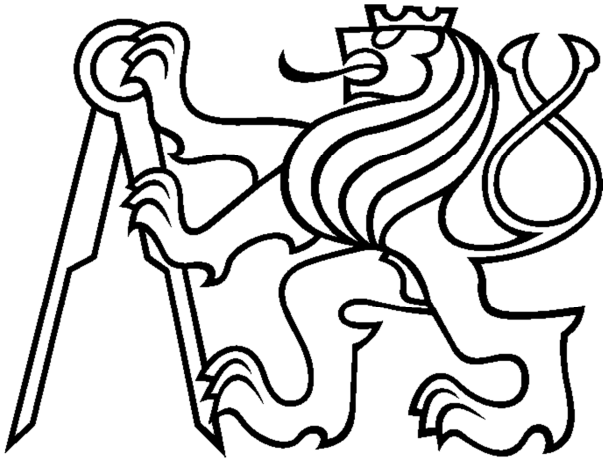
Czech Technical University in Prague

Faculty of Electrical Engineering



Doctoral Thesis

February 2017

Oldřich SLAVATA

Czech Technical University in Prague

Faculty of Electrical Engineering
Department of Measurement

Impact of IP Chanel Parameters on the Final Quality of the Transferred Voice

Doctoral Thesis

Oldřich SLAVATA

Prague, *February 2017*

Ph.D. Programme: *Electrical Engineering and Information Technology, P2612*

Branch of study: *Measurement and Instrumentation, 2601V006*

Supervisor: ***Prof. Ing. Jan Holub,*** **Ph.D.**

Poděkování

Na tomto místě bych předně rád poděkoval svému školiteli Prof. Ing. Janu Holubovi, Ph.D. za poskytnutý námět, konzultace a odborné vedení při tvorbě této disertační práce.

Další poděkování patří všem kolegům z katedry měření na FEL, kteří svými konzultacemi pomohli vyřešit spoustu dílčích problémů. Speciálně bych rád poděkoval pí. Kočové za její ochotu a vstřícnost při řešení administrativních záležitostí.

V neposlední řadě bych rád poděkoval své rodině za vytvoření výborného zázemí, za trpělivost a za plnou podporu při tvorbě této disertační práce.

Čestné prohlášení

Čestně prohlašuji, že disertační práci na téma „Impact of IP Chanel Parameters on the Final Quality of the Transferred Voice“ jsem vypracoval samostatně a použil jsem úplný výčet citací z literatury uvedené v seznamu na konci této práce. Nemám námitek proti půjčování nebo zveřejňování mé disertační práce, nebo jejích částí. Taktéž nemám námitek proti využití výsledků této práce Elektrotechnickou fakultou ČVUT, a to se souhlasem katedry měření

Abstrakt

Předmětem této disertační práce je měření kvality přenosu hlasu v IP síti. Vývoj v oblasti digitalizace, kódování a přenosu řeči probíhá velmi rychle, stejně jako vývoj síťových prvků a zvyšování přenosové kapacity a spolehlivosti sítě. To klade zvýšené nároky také na vývoj a implementaci metod pro měření kvality přenosu hlasu. Objektivní algoritmy i subjektivní metody testování musí reagovat na nové druhy poruch a rušení. Problémem je, že metodika subjektivních a částečně i objektivních testů je založena na doporučení ITU-T P.800 z roku 1993 a byla navržena pro použití v klasických telefonních sítích. Při současné nízké četnosti a nepravidelném výskytu chyb jsou používané vzorky příliš krátké.

V první části této práce jsou popsány protokoly a kodeky, které se nejčastěji využívají pro přenos řeči. Dále jsou popsány metody a algoritmy, které se v současnosti využívají pro měření kvality přenosu. V práci je využíván zejména algoritmus POLQA dále pak PESQ a 3SQM.

Ve druhé části práce je v sérii experimentů ověřen vliv různých parametrů IP sítě na kvalitu přenosu hlasu. Testované parametry jsou: zpoždění, variace zpoždění (jitter), ztrátovost paketů, QoS a různé kodeky. Také je vyhodnocen vliv délky vzorku na výsledky měření pomocí objektivního algoritmu. Pomocí subjektivních testů je zkoumán vliv zpoždění na kvalitu konverzace a vliv stresu hodnotitelů na výsledky těchto testů.

V další části práce je navržena metodika pro subjektivní testy s využitím dlouhých vzorků. Předmětem zkoumání je zde kromě délky vzorku také způsob sběru a vyhodnocení dat od hodnotitelů. Zkoumané metody jsou: klasická s jedním hodnocením na konci, ekvidistantní s hodnocením v pravidelných intervalech a náhodná s hodnocením kdykoliv v průběhu vzorku. Tyto metody se liší jak ve výsledném hodnocení vzorku, tak v jeho nejistotě.

Díky využití delších vzorků je možné zmenšit počet opakování a tím subjektivní testy zjednodušit při zachování požadované přesnosti. Experimentální výsledky prokázali vliv Jitteru, ztrátovosti paketů, použitého kodeku a metody QoS na kvalitu přenosu. Rovněž byl prokázán vliv statického zpoždění na kvalitu konverzace. Z porovnání různých objektivních algoritmů jsou patrné rozdíly ve výsledcích, způsobené tím jak se algoritmy se postupně vyvíjejí.

Abstract

The subject of this dissertation is the measurement of quality of voice transmission over IP networks. Development in digitization, encoding and transmission of the speech is very fast, as well as the development of network elements and increasing transmission capacity and network reliability. It also puts increased demands on the development and implementation of methods for measuring the quality of voice transmission. Objective algorithms and subjective test methods must respond to new kinds of faults and disturbances. The problem is that the methodology of subjective and partly objective tests is based on ITU-T Recommendation P.800 of 1993 and was designed for use in conventional telephone networks. With the current low rate and irregular occurrence of errors, the used samples are too short.

In the first part of this thesis, the protocols and codecs which are most often used for speech transmission are described. The following describes the methods and algorithms currently used to measure the transmission quality. In this thesis is mainly used algorithm POLQA then PESQ and 3SQM.

In the second part of this thesis, the effect of various IP network parameters on the voice quality is verified in the series of experiments. Tested parameters are delay, delay variation (jitter), packet loss, QoS and various codecs. The influence of the length of the sample on the measuring results using an objective algorithm is also evaluated. The effects of static delay on the quality of the conversation and the effects of stress evaluators on the results of these tests are investigated using subjective tests.

In the second part of this work, a methodology for subjective tests using long samples is suggested. The subject of the investigation is in addition to the length of the sample as well the method of collection and analysis of data from the evaluators. The investigated methods are: a classic one with evaluation at the end, equidistant with ratings at regular intervals and random with ratings anytime during the sample. These methods differ in the final evaluation of the sample and its uncertainty.

Thanks to the use of longer samples, it is possible to reduce the number of repetitions and the subjective tests are simplified while maintaining the required accuracy. Experimental results showed the influence of jitter, packet loss, the codec used and methods of QoS on the transmission quality. Influence of static delay on the quality of the conversation was also confirmed. Differences in results are noticeable in comparison of objective algorithms. It is caused by the the progressive development of algorithms.

List of Abbreviations

VoIP	Voice over Internet Protocol
UDP	User Datagram Protocol
TCP	Transmission Control Protocol
IP	Internet Protocol
SIP	Session Initiation Protocol
IETF	Internet Engineering Task Force
ACR	Absolute Category Rating
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure
MOS	Mean Opinion Score
POLQA	Perceptual Objective Listening Quality Assessment
GSM	Groupe Spécial Mobile
MOS-LQS	Mean Opinion Score - Listening Quality Subjective
ALS	Active Level of Speech
SPL	Sound Presure Level
ERP	Ear Reference Point
ISDN	Integrated Services Digital Network
xDSL	Digital Subscriber Line
QoS	Quality of Service
IHL	Internet Header Length
DSCP	Differentiated Services Code Point
ECN	Explicit Congestion Notification
RFC	Request for Comments
HTTP	Hypertext Transfer Protocol
SMTP	Simple Mail Transfer Protocol
SB-ADPCM	Adaptive Sub Band Differential Pulse-Code Modulation
AMR-WB	Adaptive Multi-rate Wideband
AMR	Adaptive Multi-Rate
UMTS	Universal Mobile Telecommunication System
DTX	Discontinuous Transmission
VAD	Voice Activity Detection
CNG	Comfort Noise Generation
CELP	Code-Excited Linear Prediction
ACELP	Algebraic Code-Excited Linear Prediction
MOS-LQO	Mean Opinion Score - Listening Quality Objective
FFT	Fast Fourier Transform
IRS	Impulse Response Samples

SNR	Signal to Noise Ratio
LPC	Linear Predictive Coding
PCM	Pulse-Code Modulation
WAV	Waveform Audio File Format
UTP	Unshielded Twisted Pair
FIFP	First In First Out
WFQ	Weighted Fair Queuing
CBWFQ	Class-based Weighted Fair Queuing
LLQ	Low-latency Queuing
CD-WRED	Class-based Weighted Random Early Detection
ECN	Explicit Congestion Notification
LFI	Link Fragmentation and Interleaving
iLBC	Internet Low Bitrate Codec
MOS-CQS	Mean Opinion Score - Conversation Quality Subjective
RMSE	Root-Mean-Square Error

Table of contents

1 Introduction

Throughout the history, the telephone and data networks always developed separately. At the end of the eighties, modems enabling transmission of modulated data in the telephone channel in the band 0.3 to 3.4 kHz started to appear. With the progress of digitization of transmission and switching systems it is no longer necessary to modulate computer data and it can be transmitted in the whole range of telephone networks. Due to the development of mobile telephony, data transmission and Internet access has become, for most home users, the main purpose of the fixed telephone network. Furthermore, increasing network bandwidth along with constantly perfected methods of digital processing of acoustic and visual signals allow users to use the Internet for voice and video communication in real time.

The trend of the twenty-first century is the gradual convergence of telephone and data networks. The goal is the most efficient use of existing lines and switching equipment to meet user demand for faster and more stable connections. With the XDSL technology, it is possible to use telephone lines for high-speed data transmission outside of the band 0.3 to 3.4 kHz, which is designed for call data. In the second decade of the twenty-first century, there is a rapid increase in the volume of transmitted data in mobile networks. Permanent access to the Internet, email, or social networks is for many users the main way of using the mobile phone. Telephony is becoming only an adjunct to data and multimedia services, and it begins to approach a full merger of the two networks and the provision of telephone services using only data network.

Internet telephony is becoming a lucrative line of business because it allows local providers to break the monopoly of big operators. In the resulting competitive environment, it is necessary to objectively measure and evaluate the quality of services. Generally, we are ensuring implementation of telephone services and the quality of the actual call connection. Transferring a phone call via the Internet network poses special problems. Network parameters such as delay, packet loss, used codec or bandwidth all affect the quality of the transmitted audio signal.

1.1 VoIP

Voice over Internet Protocol is a technology that allows voice transmission in digital form in packets of UDP / TCP / IP. It is used for phone calls over the Internet or other computer networks. Its development began in the 1990s and in the future it will probably completely replace traditional voice service.

As already indicated, for voice transmission on the network layer the IP protocol is used and on the transport layer UDP protocol is used, which unlike TCP prefers the speed of delivery before the integrity of transmitted data. For the IP

telephony is a delay caused by inspection of each packet (TCP), generally a bigger problem than sporadic outages packets (UDP). Currently, the most commonly used protocol for the connection and signalling of a telephone call on an IP network is SIP created in IETF.

2 State of the art

2.1 Objective testing of voice transmission quality

In recent decades, the most applied approach in the problems of measuring the quality of voice transmissions was based on imitation of human perception. These methods use the algorithms that imitate the human perception of sound (voice) and then predict the behavior of subjective evaluators during listening tests. When testing the quality of voice, so-called ACR tests (absolute category rating) are often used, during which listeners assess the quality of the degraded sample speech without knowing the undistorted original. ACR listening tests, conducted by ITU, most frequently use assessment on the five-point scale, which is also used in the objective evaluation methods standardized in ITU. These standardized methods are PSQM (perceptual speech quality measure, ITU-T P.861 recommendation, 1996), [1], and his successor PESQ (perceptual evaluation of speech quality, ITU-T P.862, 2000) [2].These measurement standards were aimed at assessing voice quality in narrowband telephony (bandwidth 100-3500 Hz), PESQ extension for wideband transmissions (50-7000 Hz) was completed in 2005 [3]. With this extension, PESQ provides a very good correlation of the estimates with the results of subjective tests for narrowband [4] (Table 2.1) and reasonable correlation for wideband tests. The measure of the quality in these methods (and also in subjective tests) is a five-point scale MOS (mean opinion score) defined in ITU-T P.800 [5].

Table 2.1 Correlation coefficient, eight unknown subjective tests (PESQ) [4]

Test	*Type*	*Correlation*
1	Mobile; real network measurements	0.979
2	Mobile; simulations	0.943
3	Mobile; real networks, per file only	0.927
4	Fixed network; simulations, 4–32 kbit/s codecs	0.992
5	Fixed network; simulations, 33-64 kbit/s codecs	0.974
6	VoIP; simulations	0.971
7	Multiple network types; simulations	0.881
8	VoIP frame erasure concealment; simulations	0.785

In conjunction with the rapidly increasing availability and speed of the Internet, new services began to emerge using VoIP communication and new ways of coding and speech compression. With that are coming the new kinds of interference and distortions such as the time-warping. PESQ was not originally designed for these problems, and the results thus began to be unreliable. Therefore, ITU-T initiated the development of a new standard for evaluating speech quality. The new algorithm was meant to be the technological update of PESQ. Bandwidth was increased from 7 kHz (wideband) to 14 kHz (super wideband). In addition high-quality reference speech samples were recorded in an environment with low noise and low reverb. It used a sampling frequency of 48 kHz and voice samples with a bandwidth of 14 kHz.

Six developers submitted algorithms for assessment. Of these, ITU-T has selected three that met the requirements and were selected for final standardization. These three algorithms were from developers OPTICOM, SwissQual, and TNO. The candidate algorithms were evaluated by comparing the results with data from subjective tests, which included a large number of newly created databases of samples, which were, for the tested algorithms, an unknown. The selection was based on the error in predicting the MOS value (Mean Opinion Score), offsetting a 95% confident interval for each value of MOS. Later, were all three algorithms combined into a so-called composite model. It turned out that the combined model has better results than the individual sub-algorithms, and it was selected for standardization. According to the original working title, this model was named POLQA (Perceptual Objective Listening Quality Assessment) and was approved as a new standard for measuring voice quality in narrowband, wideband, and super wideband applications by ITU-T in 2011. Relevant ITU-T Recommendation has a number P.863 [6].

It should be noted that although POLQA operates in a super wideband mode at 48 kHz, it should not be used to evaluate the transmission of music signals. POLQA algorithm is adapted for speech and in its development and debugging only speech signals were used. To evaluate the transmission quality of the music it is better to use the standard ITU-T BS.1387 (PEAQ).

2.2 Requirements for the successor of PESQ

The first requirement for the algorithm, which was to become the successor of the PESQ P.862, was that it has to be technically compatible with existing and previously standardized methods of measuring the quality of speech. Especially with methods that use so-called Full-reference approach. In these methods, the degraded signal is compared with a non-disturbed reference signal and

subsequently evaluated. The tested system is seen as a "black box, " and there is no other information about it except the input and output signal.

The new algorithm should assess the quality of speech transmission on a five-point MOS scale, which is used in ACR listening tests. Important is the best match of the results of objective and subjective tests. The new method should also be more accurate than the PESQ and should permit reliable measurement of the quality even in the case of interference and degradation which PESQ has been designed for, such as the following list (taken from [7]):

- *New and advanced coding technologies*
- *Voice quality enhancement devices*
- *Time stretching and compression techniques*
- *Influence of acoustic coupling devices and the room acoustics during insertion/recording*
- *Influence of the loudness of presentation of the speech signal*
- *Influence of linear distortions and spectral shaping („frequency response“)*
- *Influence of bandwidth (intermediate bandwidth to common telephony bands).*

This requires a large set of test data, wherein the degraded signal is evaluated by subjective listeners, with a wide variety of different disturbances. In addition to this new degradation and interference, the conditions for which PESQ provides accurate and correct results were also included in the file. The result is the following list (taken from [7]) of conditions/degradation/interference for which a new algorithm (POLQA) must provide accurate results:

- *Single and tandem speech codecs as used in telecommunication scenarios today*
- *Packet loss and concealment strategies (packet switched connections)*
- *Frame- and bit-errors (wireless connections)*
- *Interruptions (such as unconcealed packet loss or handover in GSM)*
- *Front-end-clipping (temporal clipping)*
- *Amplitude clipping (overload, saturation)*
- *Effects of speech processing system such as noise reduction systems and echo cancellers on clean speech*
- *Effects of speech processing systems such as noise reduction systems (adaptation phase and converged state) and echo cancellers on pre-noised speech*
- *Effects of speech coding systems on pre-noised speech*
- *Variable delay (Voice-over-IP, video-telephony) and time warping*
- *Gain variations*
- *Influence of linear distortions (spectral shaping), also time-variant*

- *Non-linear distortions produced by the microphone/transducer at acoustical interfaces*
- *Voice enhancement systems in networks and terminals and their effects on listening quality*
- *Reverberations caused by hands-free test setups in defined acoustical environments*

The new algorithm has to deal with all these problems and must provide reliable results across a wide variety of interference. According to the wide range of different kinds of interference, there also were special requirements for the organization of subjective tests and the creation of a database of test samples. It was necessary that all kinds of interference are equally represented in test samples played to subjective listeners.

2.3 Methodology of subjective tests

Development of new algorithms for objective testing of voice quality requires an extensive database of reliable subjective data. The data consists of the reference and the degraded signal and subjective quality assessment for each degraded signal. Development of algorithm POLQA was focused on super-wideband voice transmission. On the other hand, most subjective tests are realized in narrowband or wideband range. The new procedures for experiments were created. Experiments have been much more difficult to prepare and conduct due to the widespread transmission band, lower noise floor and stricter requirements for minimization of side effects.

Generally, the tests were very similar to methodology P.800 [5]. They have used at least four spokespeople for each test. The results were presented as Mean Opinion Scores for Subjective Listening Quality (MOS-LQS) at the five-point scale. The tests were conducted as ACR method. To reduce side effects, a number of common conditions and variables were included in each test.

To test the modern communication systems with high-quality voice transmission, it is necessary that the reference recordings used in the objective and subjective tests have adequate quality. Therefore, there are high demands for reference recordings. Recordings shall be made in a room with low reverberation (reverberation time of less than 300 ms over 200 Hz), ideally in an anechoic room. For recording, a directional microphone should be used and microphone distance from the speaker's mouth should be about 10 cm. Speech signals are sampled at 48 kHz and bandpass filtered at 50 Hz to 14 kHz.

Each reference set consists of two sentences separated by a gap of a minimum of 1 s and maximum of 2 s. The minimal length of active speech in each file is 3 s. Beginning active speech lies between 0.5 and 2 seconds from the recording

start and end of an active speech lies between 0.5 and 2 seconds from the end of the recording. A set of at least sixteen samples from four different speakers is used in test. As part of the test, no text can be repeated.

Digital active speech level (ALS according to ITU-T P.56 [8]) of the test signal should be at the level of -26 dBov (dB overload) for playback at a nominal level. The corresponding nominal SPL (Sound pressure level) in the acoustic domain is 73 dB above the ear reference point (ERP) using diotick headset (the same signal in both ears, it is used for super wideband), or 79 dB above the ERP using IRS headphones and monotonous play (narrowband, wideband). For all tests in the super wideband domain, the level is set from -20 dB to +6 dB of the nominal, in order to evaluate the influence of the volume on the final quality.

In listening tests, the ACR type of tests (absolute category rating), where subjective evaluators listen to degraded voice samples without the knowledge of the original are most used. Listeners evaluate each sample on the scale. In the telecommunications sector, most commonly used is five-point scale MOS [5].

Until recently, the most commonly used in listening tests was a narrowband speech (audio bandwidth of 100 Hz to 3500 Hz), less then wideband speech in the range of 50 Hz to 7 kHz. In contrast, the super-wideband tests are extending the band to the range of 20 Hz to 14 kHz.

It is necessary to take into account that the evaluators are often relating their evaluation to the best available sample in the test. This leads to an overestimation of the quality of narrowband samples. Such a sample is usually, in the pure narrowband test, rated higher grade than if it is included in the set with (super)wideband samples.

About 24 subjective evaluators aged 15-65 years, with evenly represented groups 15-30, 31-50, 51-65 should be used in each test. They should be equally represented by both men and women.

Another condition for listening tests are described in the requirements for Mean Opinion Score for Listening quality subjective (MOS-LQS) [5]. In order to compare the results of different listening tests, the following 12 key conditions (list is taken from [9]) is situated in each test:

- *Clean 0 dB, −10 dB, and –20 dB relative attenuation*
- *Multiplicative noise (MNRU conditions* [10]*) with signal-to-noise ratios of 10 dB and 25 dB (using P.50* [11] *shaped noise for modulation)*
- *Additive noise with a signal-to-noise ratio of 12 dB using Hoth noise* [12] *and 20 dB using babble noise*
- *Linear filtering with narrowband telephone characteristic (300–3400 Hz), bandpass filters 500–2500 Hz, and 100–5000 Hz*

- *Temporal clipping with 2% and 20% packet loss, packet size 20 ms without packet loss concealment*

The ITU-T Study Group 12 POLQA benchmark consisted of three phases: training phase, the validation and selection process, and finally, the integration phase. During the training phase, the algorithm POLQA was used with a total of 16 wideband and super-wideband databases of samples. Stage evaluation and selection was carried out using eight databases. Besides these tests, POLQA was also used to evaluate the 38 narrowband databases to verify the behavior of the algorithm in the narrowband mode. This allows for comparison of POLQA and older algorithm PESQ in narrowband telephony.

The database used in the training phase was made available to all applicants. In contrast, the database used in the later stage of the validation and selection was not accessible for applicants and was largely established after handing algorithms in ITU-T. The selection is therefore realized using the data that was "unknown" for the algorithm.

2.4 Specific problems with TCP/IP networks

At present, more and more phone calls are transmitted as data in TCP / IP networks. Compared to traditional telephone networks the IP networks are relatively reliable. Errors such as massive packet loss or significant delays are relatively rare. This can cause problems due to the fact that the methodology for subjective and partly objective tests is based on ITU-T Recommendation P.800 of 1993, which was designed for use in classic telephone networks. Recommended sample length for listening tests is 8-12 seconds, from this, the active speech is about a half. If we take into consideration, that for most coders one TCP/IP packet carries from one to three voice packets, each containing 20 ms of the speech signal, we get 400 TCP/IP packets for the whole sample. The error rate of 2 percent means that in the worst case, we will lose eight packets. This together makes 480 ms of the speech signal, but not necessarily continuous. With such a low error rate, there is the negligible probability that the lost packets contain pauses between speeches or words that do not alter the meaning. Such a cases can then distort the results of subjective and objective tests.

An approach which is currently used in objective tests is based on multiple repetitions of transmission of a single sample and statistical processing of results. This allows eliminating the influence of coincidence in error distribution and also reduces the uncertainty of type A in the final estimate of MOS. For subjective tests, this procedure cannot be used because of inappropriate extension of test and the effect of repetition, which leads to fatigue and loss of

attention of evaluators. One possible solution is to use a longer sample for example 120 s. Various methods of subjective testing using long samples will be discussed later in this thesis (chapter 7.2).

2.5 Summary

Voice transmission quality can be measured objectively using a computer algorithm, or subjectively using listening tests. A properly executed subjective tests are more accurate, but organizationally more difficult, more expensive and time-consuming. The development of objective methods is aiming at the best correlation with subjective tests. Recently, the availability and speed of Internet access are rapidly increasing. Transferring a call in the form of data over TCP / IP is gradually replacing traditional telephony in fixed network, and shortly it will probably replace mobile voice services GSM. Some problems and features of IP networks such as packet loss, delay and jitter affect the quality of voice transmission. Development of communications in an IP network brings higher demands for quality and speed of transmission and thus the rapid development of codecs and compression algorithms. The original testing methodology developed for conventional fixed networks or ISDN lines provides unreliable and distorted results in the measurement of the IP network. Simultaneously with the development of new codecs, new test algorithms and procedures as POLQA algorithm (ITU-T P.863) or subjective tests using long samples are also produced.

3 Objectives

Development of new algorithms for coding and compression is very fast, as well as the development of network elements and increasing the capacity of the transmission network. This brings new challenges in measuring the quality of voice transmission.

The first objective of this study is to verify the effect of the IP channel parameters on the quality of voice transmission. Using objective algorithms, we can test the rate of this effect and also check the behavior of these algorithms in modern VoIP environment.

Investigated parameters of VoIP transmission are:

- Packet loss
- Delay
- Jitter and jitter buffer
- The codec and its settings
- The length of the sample
- Methods of QoS and media access control

Above mentioned parameters will be evaluated in a series of experiments using both objective and subjective measurement methods. The aim is to determine the applicability of current subjective and objective measurement methodologies in the context of new coders, procedures and changing characteristics of the transmission network.

One of the most important problems that arise in subjective tests according to the current methodology is that the errors are relatively rare and irregular and used voice samples are too short to make the mistakes to occur. Possible solution is using longer samples.

The second objective of this work is to propose options for the development of new methods for measuring voice quality in IP telephony using long samples. Within this, it is necessary to propose and compare different methods of voting and obtaining data from users.

4 Technical Background

This chapter provides a brief overview of technologies that are commonly used in IP telephony.

4.1 Network Protocols

The protocol is a standard that controls communication and data transmission in computer network. Typical tasks of the protocol are to:

- Detect physical connection
- Establish the connection
- Configure the connection parameters
- Format the messages and data
- Deal with corrupted messages
- Detect connection loss
- Terminate the connection

Protocols operate at different layers of electronic communications by its purpose.

4.1.1 IP - Internet Protocol

Internet Protocol [13] is the fundamental protocol of the network layer and the entire Internet. It performs broadcasting of datagrams based on network IP addresses contained in its header. Each datagram is an independent unit of data that contains all the necessary information about the recipient, the sender and sequence number of the datagram in the message. Datagrams travel through networks independently, and their order of delivery may not match those of the message. Datagram delivery is not guaranteed, reliability, if required, must be provided on a higher layer (TCP, applications). Each datagram includes a header that has the following structure (Table 4.1).

Table 4.1 IPv4 Header Format

Byte	0		1	2	3
0 - 3	Version	IHL	DSCP,ECN	Total Length	
4 - 7	Identification			Flags	Fragment Offset
8 - 11	Time To Live		Protocol	Header Checksum	
12 - 15	Source IP Address				
16 - 19	Destination IP Address				
20 - 23	Options				
24 - …	Data				

Version - IP version (currently IPv4, IPv6 is ready).

IHL - **Internet Header Length.** The length of the header in the four-bit words (may vary because of the optional capabilities).

DSCP - Differentiated Services Code Point. Originally defined as the Type of service (ToS) field. This item was meant to enable to select the type of service and consequently to adapt the routing (prefer speed, integrity, bandwidth, etc.). In practice, this has not been implemented. At present, the item is used for similar purposes - bears the mark for mechanisms to ensure services with defined quality e.g. VoIP.

ECN - Explicit Congestion Notification. This field allows end-to-end notification of network congestion without dropping packets. ECN is an optional feature that is only used when both endpoints support it and are willing to use it. It is only effective when supported by the underlying network.

Total Length - Datagram length in bytes.

Identification - Each packet has a unique identifier. If the datagram is corrupted during the transmission, fragments that belong together can be identified by this field. (Have the same identifier).

Flags - serves to control fragmentation. Two are defined: *More Fragments* meaning "I am not the last" and *do not fragment* prohibiting to fragment this datagram.

Fragment Offset - indicates the position of the fragment in the original datagram. The unit is eight bytes.

Time to live - protection from loops, each router decrements this value when passing. When the timer reaches zero, the datagram is dropped.

Protocol - determines the destination protocol at a higher layer (e.g. TCP - 6, UDP - 17, ICMP - 1, ...).

Header Checksum - field is used for error-checking of the header. When a packet arrives at a router, the router calculates the checksum of the header and compares it to the checksum field. If the values do not match, the router discards the packet. The router must always calculate the new checksum due to the decrementation of the TTL field.

Source IP Address - IP address of the network interface that has sent the datagram.

Destination IP Address - IP address of the network interface to which the datagram is intended.

Options - Various supplementary information or requirements. For example, you can prescribe a series of addresses, which the datagram has to pass. It is underused in practice.

4.1.2 UDP - User Datagram Protocol

UDP [14] provides a simple interface between the network and the application layer. It makes no guarantee of delivery and once the message is sent, it does not maintain any state. It only adds checksums and the ability to distribute UDP packets between different applications running on the target computer (source and destination port) to the functions of the network layer. The header of UDP packet is seen on Table 4.2.

Table 4.2 Header of UDP packet

	Bit 0 - 15	16 - 31
0	Source port	Destination port
32	Length	Checksum
64	Data	

Source port - not required, by default it is set to 0.

Destination port - identify the target application (eg .: SIP - port 5060).

Length - the length of the UDP packet, including the data in bytes.

Checksum - 16-bit checksum is covering the header and data.

4.1.3 The difference between TCP and UDP

TCP is a connection-oriented protocol. The connection is opened by the client or the server, and then the transfer can go in both directions. Important features of the protocol are:

- Reliability - TCP uses acknowledgment of receipt, re-sending, timeouts and requests undelivered data. There is no data loss during transmission.
- Maintaining of the order - If a packet arrives out of order, cache on the receiving end will take care that they are sorted correctly.
- Higher demands - the transmission is time-consuming and also requires sending "service" packets, e.g., To establish a connection.

UDP is a simpler protocol for sending independent messages. It is used in applications where 100% reliability is not needed, or even bringing unwanted delays (streaming audio or video, online games, VoIP). Important features are:

- Speed - nothing is controlled or validated, data are sent quickly and smoothly.
- Does not maintain the order - does not guarantee that the message reaches the recipient in the same order they were sent.
- Does not guarantee delivery - datagram may get lost on the way. The control, if necessary, must be realized on a higher layer.

4.1.4 SIP - Session Initiation Protocol

The Session Initiation Protocol is used for transmission of the signalization in the Internet telephony. Actually it is standardized in RFC 3261 [15] Typically it uses UDP port 5060, but can also work on TCP 5060. SIP was created as a reaction to the older and more complex H.323 protocol of ITU-T. The project creators were to develop a modern internet protocol based on well-proven principles of simplicity and decentralization. Therefore SIP is based on the proven HTTP and SMTP.

For creating and managing SIP call, the protocol must provide the following five actions:

- Localization of subscriber
- Determining the status of the participant (free, busy, forwarded ...)
- Determine of options participant (codec, bit rate ...)
- Establishing a connection
- Control of the connection (changes in progress, termination)

The most important methods (commands) of the SIP Protocol are:

- INVITE - request to establish a connection
- ACK – acknowledgment of the call connection
- BYE – ending the call
- CANCEL – termination of SIP sessions before the call is established
- REGISTER – registration of the subscriber station
- OPTIONS – query for the server options

Responses do not have identical designations, but they are organized into groups in the numerical code:

1XX – informational messages (ex. „100 trying“, „180 ringing“)

2XX – Successful execution of the request („200 OK“)

3XX – redirecting

4XX – client error, incorrect request („403 forbidden“)

5XX – server error („500 server internal error“, „501 not implemented“)

6XX – fatal error („606 not acceptable“)

SIP devices can establish a session directly between themselves, but it is more common they apply to one or more SIP proxy servers. These servers can work as so-called SIP registrator on which each participant is registered.

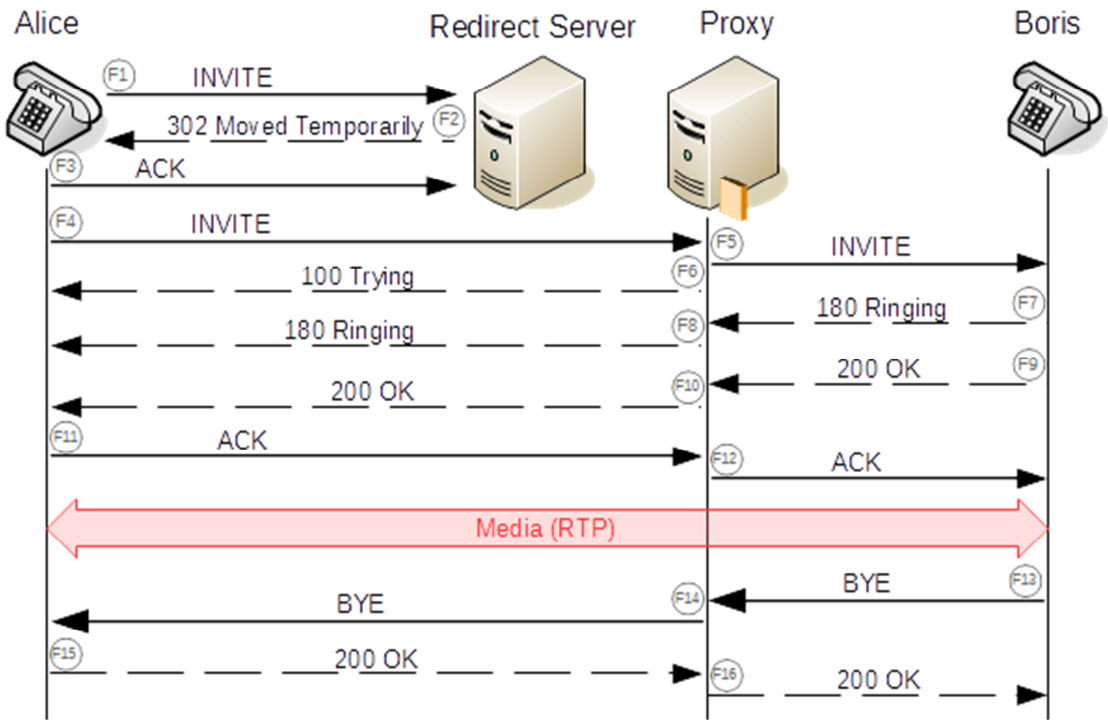


Fig. 4.1 SIP comunication [16]

4.2 Codecs

Codec (a composite of the first syllables of the words "coder and decoder" or "compression and decompression") is a device or a computer program that can transform data stream or signal. Coder stores data in encrypted form (mostly for the purpose of transmission, storage or encryption), by contrast, the decoder is used for restoring the original or near original form of data that is suitable for playback, viewing, or other manipulation. Codecs are essential components of the software for editing (cut) of multimedia files (voice, music, movies), and are often used for phone calls, video calls and distribution of the multimedia data in networks (streaming).

Encoding (compression) can be lossless or lossy.

Lossless compression means that after decoding the source signal can be restored to a state identical to the original. Examples of lossless audio codecs are:

- FLAC
- ALAC
- Monkey's Audio APE
- RealAudio Lossless
- Windows Media Lossless

Lossy compression means that some information in the process of compressing is intentionally dropped or suppressed. After decoding, the input signal cannot be reconstructed in its original quality, but the quality decline is compensated by a significant decline in the volume of data. The rate of reduction in the volume of data depends on the used coder and its settings. In practice, a decrease to 10% of the original volume is relatively common. The subjective quality of the resulting signal is not significantly different from the original. This can be achieved by understanding the mechanism of human perception and thus the possibility to exclude the part of the signal that is not critical to human ears. Examples of lossy codecs are:

- MP3
- G.711
- SPEEX
- G.729
- AAC
- G.722
- Ogg Vorbis

In IP telephony, it is important to maintain high connection stability, low packet loss, and short delays. One way this can be achieved is to low a bitrate required for the transmission, therefore, VoIP uses almost exclusively lossy codecs.

Audio Codecs used in this thesis are: G711, G722, AMR, G.722.2, G.729, Speex

4.2.1 G.711 (PCM)

G.711 [17] (standardized in 1988) is the most common, most widespread and simplest standard for the digitizing audio signal used in telecommunications. For simplicity, however, is also the least economical to bitrate. It uses pulse code modulation (PCM) and is often referred to by the abbreviation as well. Sampling frequency in this format is 8kHz, and a resolution is 8 bits, this implies bandwidth of 64 kbit / s, which is the basic capacity of the speech channel in ISDN telephony.

For encoding the signal a logarithmic compression, where the twelve or thirteen-bit signal is converted to an eight-bit signal is used. The logarithmic compression of the speech signal in Europe and Australia uses a formula known as A-law, in North America and Japan formula μ-law. In North America and Japan a higher compression is used, because in some of the local telephone network only seven bits for voice are available and eighth bit is used for signaling. In Europe all eight bits are used for voice transmission. Signaling is transmitted by a separate channel. Voice in a digital telephone network In Europe and Australia has the higher quality than the same transfer using the telephone network in North America or Japan. The signal must be transcoded at the interface of these networks. The used format and other information are transmitted via the signaling.

Other variants of G.711 are:

- G.711.0 (sometimes referred to as G.711 LLC), standardized in 2009, provides lossless compression.
- G.711.1 standardized in 2008, allows the use of a sampling frequency of 16kHz for higher transmission quality. Data flow is 64, 80 or 96 Kbit/s

4.2.2 G.722

G.722 [18] is a wideband (50 Hz - 7 kHz) audio codec standardized in 1988. The sampling frequency is 16 kHz with a resolution of 14 bits. The bitrate is 48, 56 and 64 kbit / s. Codec uses a technology SB-ADPCM (Adaptive Sub Band Differential Pulse-Code Modulation) allowing to change the size of the

quantization step, and thereby reducing the bitrate at the same subjective quality of transmission.

Codecs G.722.1 and G.722.2 (AMR-WB) are not the variants of G.722 but separate codecs that use a different compression technology for lower bitrate.

4.2.3 AMR (AMR-NB)

AMR (Adaptive Multi-Rate audio codec) [19] is narrowband audio codec standardized in 1999. It is mainly used for speech transmission. The sampling rate is 8 kHz and resolution is 13bit. For coding it uses the ACELP technology (Algebraic Code Excited Linear Prediction). Codec enables to set the level of compression depending on the required transmission quality or depending on the available network bandwidth.

Compression can be set in eight steps resulting in the following bit rates: 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 and 4.75 kbit / s.

AMR codec was established in 1999 as the standard codec of 3GPP for speech transmission. Since then it has been widely used in GSM and UMTS.

AMR uses Discontinuous Transmission (DTX), with Voice Activity Detection (VAD) and Comfort Noise Generation (CNG). The purpose is to reduce the required bitrate. Codec carries no or very little data during pauses in speech.

4.2.4 G.722.2 (AMR-WB)

G.722.2 (Adaptive Multi-rate Wideband) [20] is a wideband codec designed for speech transmission. It is based on AMR codec. It provides higher quality voice transmission thanks to an extended frequency range of 50 to 6400 Hz (7000 Hz for mode 23.85 kbit / s). The sampling frequency is 16 kHz. Codec enables to set the following bit rates (in kbit/s): **6.60**, **8.85** (do not provide wideband, used in the case of a bad connection), **12.65** (baseline, provides transmission quality comparable with the codec G.722 with bitrate of 48kbit / s), **14.25, 15.85, 18.25, 19.85, 23.05, 23.85** (providing quality comparable to G.722 with bit rate of 64 kbit / s).

4.2.5 G.729

G.729 [21] is a narrowband codec for speech transmission. The sampling frequency is 8 kHz, the resolution is 16 bits. In the basic variant, this codec has a bitrate of 8kbit / s. G.729 also uses the ACELP technology. Due to the effort to

the lowest possible bit rates while maintaining quality, the codec is quite complex and therefore computationally intensive. Due to the low bitrate, this codec is widely used in VoIP applications. Later, several additions were created that add additional functionality.

G.729a (Annex A) is a version compatible with the original G.729, which reduces the computational complexity for the price of a small reduction in transmission quality.

G.729b (Annex B) is a variant which adds the technology Voice Activity Detection (VAD), Discontinuous Transmission (DTX) and Comfort Noise Generation (CNG) to reduce bitrate. This variant is not compatible with the original G.729.

G.729ab combines the functions of variants A and B, is only compatible with G.729b.

Other variants of the codec add another level of bitrate, lower 6.4 kbit / s (Annex D, F, H, I, C +) and higher 11.8 kbit / s (Annex E, G, H, I, C +). Annex J adds support for wideband transmission.

The complete list of Appendices can be found in [22].

4.2.6 GSM (GSM-FR, GSM 06.10)

GSM was the first standard for digital speech coding used in the GSM mobile phone networks. Codec has been standardized by ETSI in 1994 [23]. The sampling frequency is 8kHz, resolution 13 bit and bitrate 13 kbit / s.

Codec is based on the technology RPE-LTP (Regular Pulse Excitation - Long Term Prediction). Order of the linear prediction is lower than in modern codecs. Currently, this codec is replaced by GSM 06.60 or AMR.

4.2.7 Speex

Speex [24] is an audio codec designed specifically for VoIP applications. It is based on the CELP technology (Code-Excited Linear Prediction). It is a free software that is the part of the project ogg. The first version was released in 2003. Codec works with a sampling frequency of 8, 16 and 32 kHz depending on the required transmission quality and bandwidth. Bitrate is widely adjustable from 2-44 kbit / s. Use of the VBR technology allows codec to change the bitrate dynamically any moment during the transmission, depending on the "intensity" of the transmitted signal. For example, some consonants require a higher bitrate than others for the same transmission quality. Codec also uses

technology to save data transmission during pauses in speech (VAD, DTX). Speex also supports stereo encoding.

5 Measuring Speech Transmission Quality

This chapter provides an overview of technologies and algorithms for measuring voice quality, which are used in this thesis. A detailed description of the algorithms can be found in the relevant recommendations.

5.1 MOS

For evaluating the quality of voice transmission the MOS scale (Mean Opinion Score) is used. Individual variants of the concept of MOS are defined in ITU-T P.10 [25], which regulates and unifies the terminology in the issue of measuring the quality of voice transmission. MOS value is an estimate of the subjective evaluation of users. The most commonly used is the five-point scale (Table 5.1) detail defined in ITU-T P.800 [5].

Table 5.1 MOS scale

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

There are several methods to obtain MOS values. The most accurate is a subjective test, where the MOS value is obtained directly from users. Subjective tests are divided into the conversation (ITU-T P.800 Annex A) and listening (ITU-T P.800 Annex B). Conversational tests are more demanding of time and organization than listening, they are used mainly for testing the parameters of the transmission, which cannot be tested via a simple listening test, such as delay.

Intrusive methods return results nearest to subjective tests. They are based on a comparison of the original and transferred sample. These algorithms use psychoacoustic models of human perception, seek to mathematically describe the human perception of sound, and find variables which have a direct impact on the perceived quality of voice signal. Intrusive methods include PESQ (Perceptual Speech Quality evaluation of) according to ITU-T P.862 (P.862.1) [26] and POLQA (Perceptual Objective Listening Quality Assessment) ITU-T P.863.

Another type of quality measurement are non-intrusive methods. These methods do not use the reference signal and final MOS is calculated only on the parameters of the transferred sample. The disadvantage of these methods is the lower accuracy and reliability, but it is useful if the reference signal is not available. An example of the non-intrusive method is 3SQM defined in Recommendation ITU-T P.563 [27].

5.2 PESQ

PESQ Perceptual Evaluation of Speech Quality is an ITU-T Standard (Rec. P.862, [2]) that has been used for objective measurement of speech transmission quality. The basic philosophy of the algorithm PESQ and its successor POLQA is shown in Fig. 5.1.

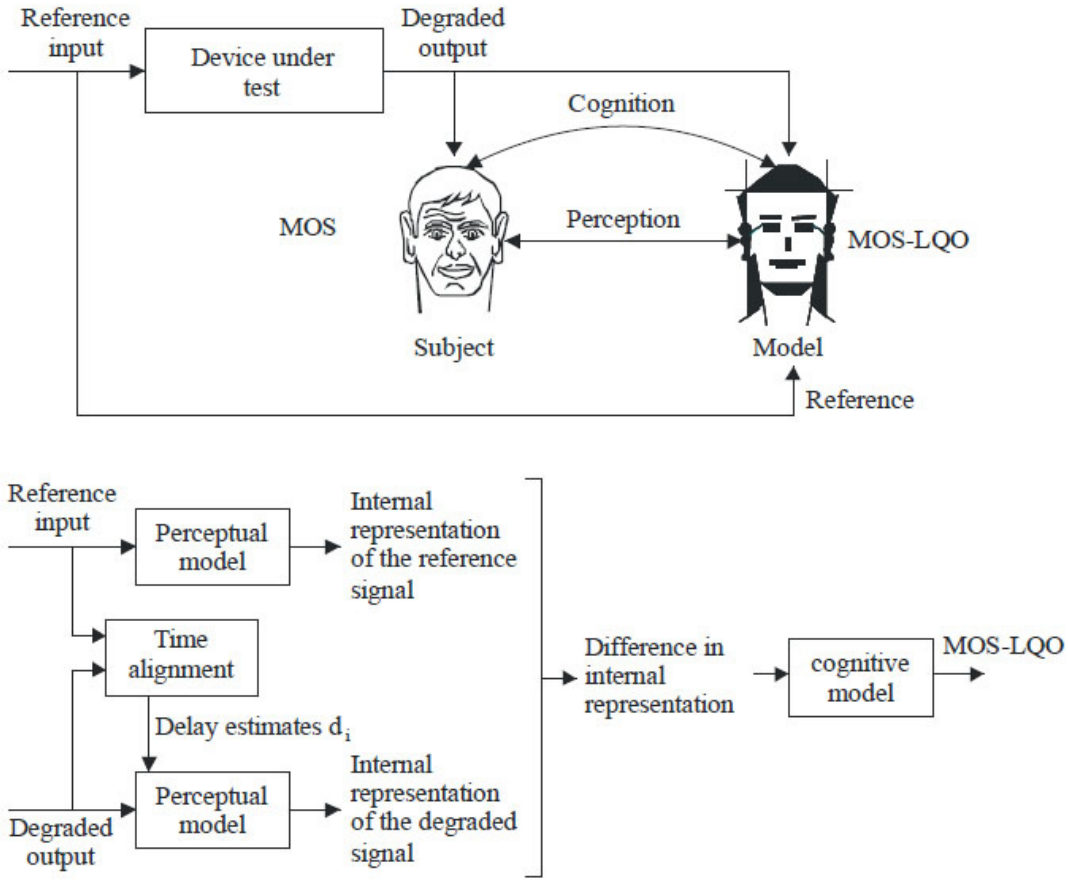


Fig. 5.1 Overview of the basic philosophy of the objective algorithms, taken from [6]

5.2.1 Algorithm Description

Entry of the algorithm are two audio recordings, represented by two data vectors. The first vector represents the original (reference signal). The second vector represents transferred (degraded) signal.

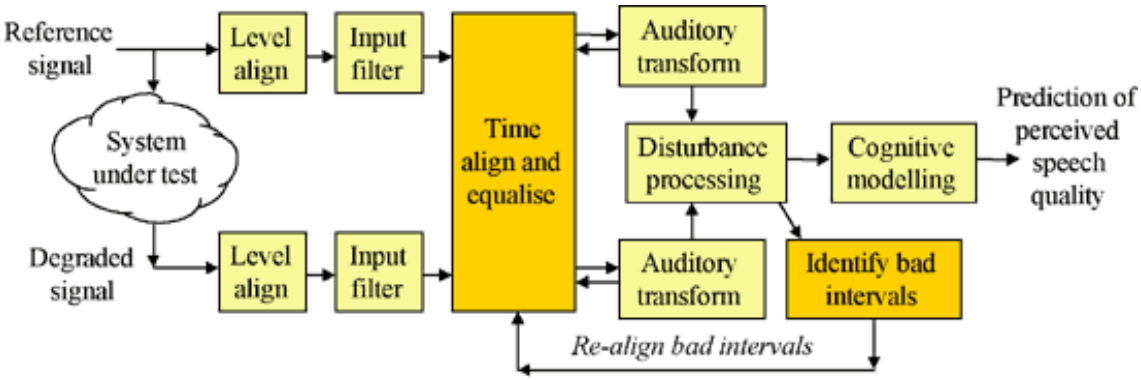


Fig. 5.2 Overview of the PESQ algorithm [28]

5.2.2 Level alignment

The purpose of this block is to align the original and degraded signal to the same predefined amplitude level 79 dB SPL. The alignment procedure is as follows:

- The filter cuts off all below 250 Hz, between 250 and 2000 Hz it is flat and further decreases linearly with variable slope through the following points: [2000 Hz, 0 dB]; [2500 Hz, -5 dB]; [3000 Hz, -10 dB]; [3150 Hz, -20 dB]; [3500 Hz, -50 dB]; [4000 Hz, -500 dB].
- Filtered waveforms are squared, and the average amplitude is calculated
- The gain needed for alignment to the desired level is separately calculated for the original and degraded signal and it is applied to the **unfiltered** version signals. The filtered version is used only for calculating the gain.

This step does not correct any errors in the recording of the sample. Only adjusts the volume to the level required for further processing. When an amplifier is overdriven (clipping of amplitude), the volume decreases but disturbing harmonics remain in the sample. Conversely, if the sample was recorded too quietly when the volume aligned, the noise is also amplified to an audible level.

5.2.3 Filtration

Since you cannot determine the type of handset used when recording the sample, the algorithm calculates with the receiving characteristics of the IRS handset. Calculation of IRS version of the original and degraded signal in the PESQ algorithm is implemented as follows:

- FFT over the length of the file
- filtering in the frequency domain according to the IRS receiving characteristics
- inverse FFT

This procedure makes the algorithm relatively insensitive to the filtering of the handset.

5.2.4 Time alignment

For the final result of the algorithm PESQ it is very important to compare the corresponding sections of the signals. Therefore, it is important to align any delays of the degraded signal against the original. This block operates on the basis of the correlation between the original and degraded signal.

- In the first step, the algorithm calculates the mutual correlation function of the two signals, the maximum of this function indicates a time shift (delay) at which both signals best overlap.
- In the second step the samples are divided in half and for each half separately, the correlation is calculated again. The delay may vary during transmission, and therefore the ideal shift for both halves of the sample may be different. If the correlation of the individual parts is better than the correlation of the whole sample, the shift calculated in the second step is added to the shift from the first step and the algorithm continues to the third step. The criterion for judging whether the correlation is better than in the first step is the width of the peak around the maximum of the correlation function (narrow sharp tip is a good correlation, broad poor). If the correlation in the second step is the same or worse than in the first, samples are shifted according to the delay calculated in the first step and the algorithm continues to psycho-acoustic transformation.
- In the third step, the halves are divided further into quarters and the correlations are calculated again. If they are better than in the second step, it proceeds by dividing into eighths, and so on recursively until correlation is worse, or the parts are too short or the maximum number of divisions is reached.

5.2.5 Auditory transformation

Auditory (or psycho-acoustic) transformation is the most important part of the algorithm. In this block, the parameters of the original and degraded signal are analyzed using a mathematical model of the human auditory system.

- The sample is divided into sections of length 16 ms with 50% overlap.
- From each section 256-point FFT is calculated.
- Series of the results of the FFT is divided into 17 frequency bands, called "bark bands" (width of the bands slightly increases towards the end of the series).
- For each of the seventeen bands energy therein is summed.
- The energy is converted back to volume levels.
- Results are weighted in view of the varying sensitivity of the human ear to different frequencies.
- Thresholding - all under the defined minimum level is set to zero.

The result is a transformation vector of seventeen values for each 16 ms interval. These vectors are then sorted into a matrix according to the time sequence of the respective sections in the signal.

5.2.6 Disturbance processing and Cognitive Modeling

Matrices of the original and degraded signals are placed side by side, and the appropriate elements are deducted. If there is a section with significantly larger errors in the resulting matrix, it is sent again (once only) for time alignment to eliminate the error in the initial time alignment. Separately, positive and negative differences between matrices are summed. The sum of the positive differences and the sum of negative differences are multiplied by different coefficients because the human ear is more sensitive to the added interference than to the missing signal. The weighted sum of the differences is then subtracted from the maximum value of five, and the result is the value of the MOS for the sample.

5.2.7 ITU–T Recommendation P.862.1

The P.862 algorithm provides rough results in a range from -0.5 to 4.5. The results need to be further converted to MOS-LQO (P.800.1) by the mapping function (1) defined in ITU-T P.862.1 [26]. Thus remapped values can be directly compared with the results of other measurement methods.

The mapping function is defined as follows:

$$y = 0{,}999 + \frac{4{,}999 - 0{,}999}{1 + e^{-1{,}4945 * x + 4{,}6607}} \tag{1}$$

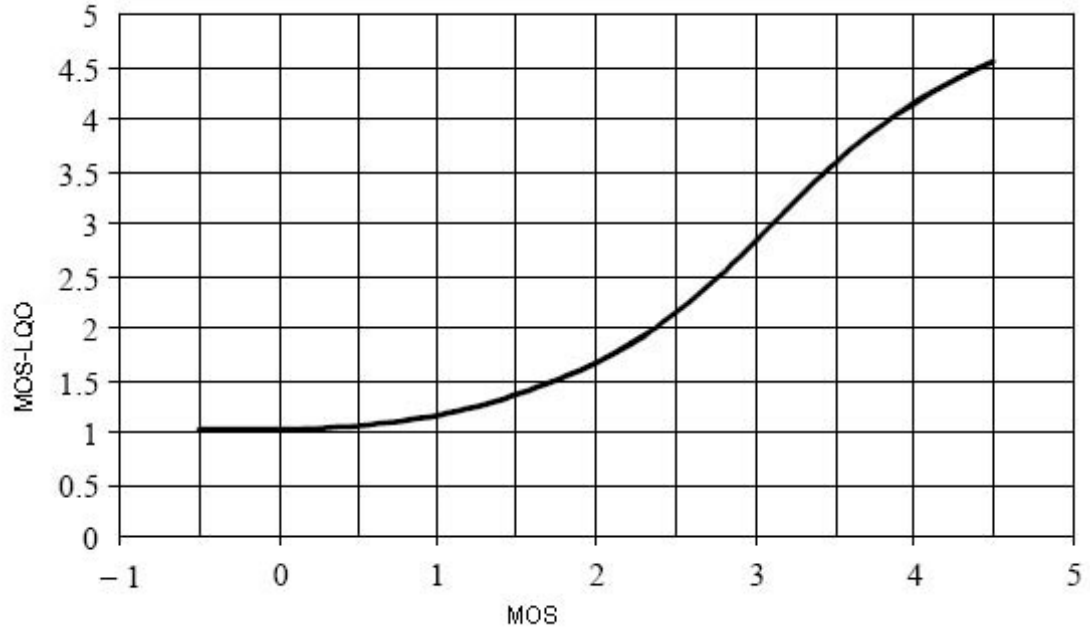


Fig. 5.3 Mapping function from $MOS_{P.862}$ to MOS-LQO, taken from [26]

5.3 3SQM

3SQMis non-intrusive method for measuring listening quality of the voice signal. The algorithm consists of three separate parts which have different methods of calculating the MOS.

Part 1 - In the sample, parameters typical for computer signal processing such as: signal-to-noise ratio (SNR), the length of suspension and damping, time cropping, etc., are calculated. Range of values of these parameters is then used to estimate the value of MOS.

Part 2 - A complex "cleaning" function is applied at the degraded sample. Missing parts are recalculated, the sample is filtered and further regulated. This purified sample, together with the original, is used as input signal for the simplified PESQ (without time alignment) and its output is an estimate of MOS.

Part 3 - The main part of this block is a precision LPC model of the human vocal tract. This 'synthesizer' attempts to pronounce the degraded sample. The result is compared with the original sample. Everything different in the original sample is considered as unnatural to the human vocal tract and marked as damage

caused during sample transfer. The sum of this added disturbance is used to calculate the MOS estimation.

The most distant of these three estimates of MOS is dropped and the arithmetic mean of the remaining two is the resulting estimate of MOS for the entire algorithm.

5.4 POLQA

POLQA Perceptual Objective Listening Quality Assessment is an ITU-T Standard (Rec. P.863, [6]) that covers a model to predict speech quality using digital analysis of the speech signal.

5.4.1 Algorithm Description

Entry of the algorithm is two audio recordings represented by two data vectors composed of 16-bit PCM samples. The first vector represents the original (reference signal). The second vector represents transferred (degraded) signal. The algorithm itself consists of a block of time alignment, evaluation, and modification of the sampling frequency and the actual psychoacoustic model that calculates the resulting values of MOS. A general overview of the ITU-T P.863 algorithm is shown in Fig. 5.4.

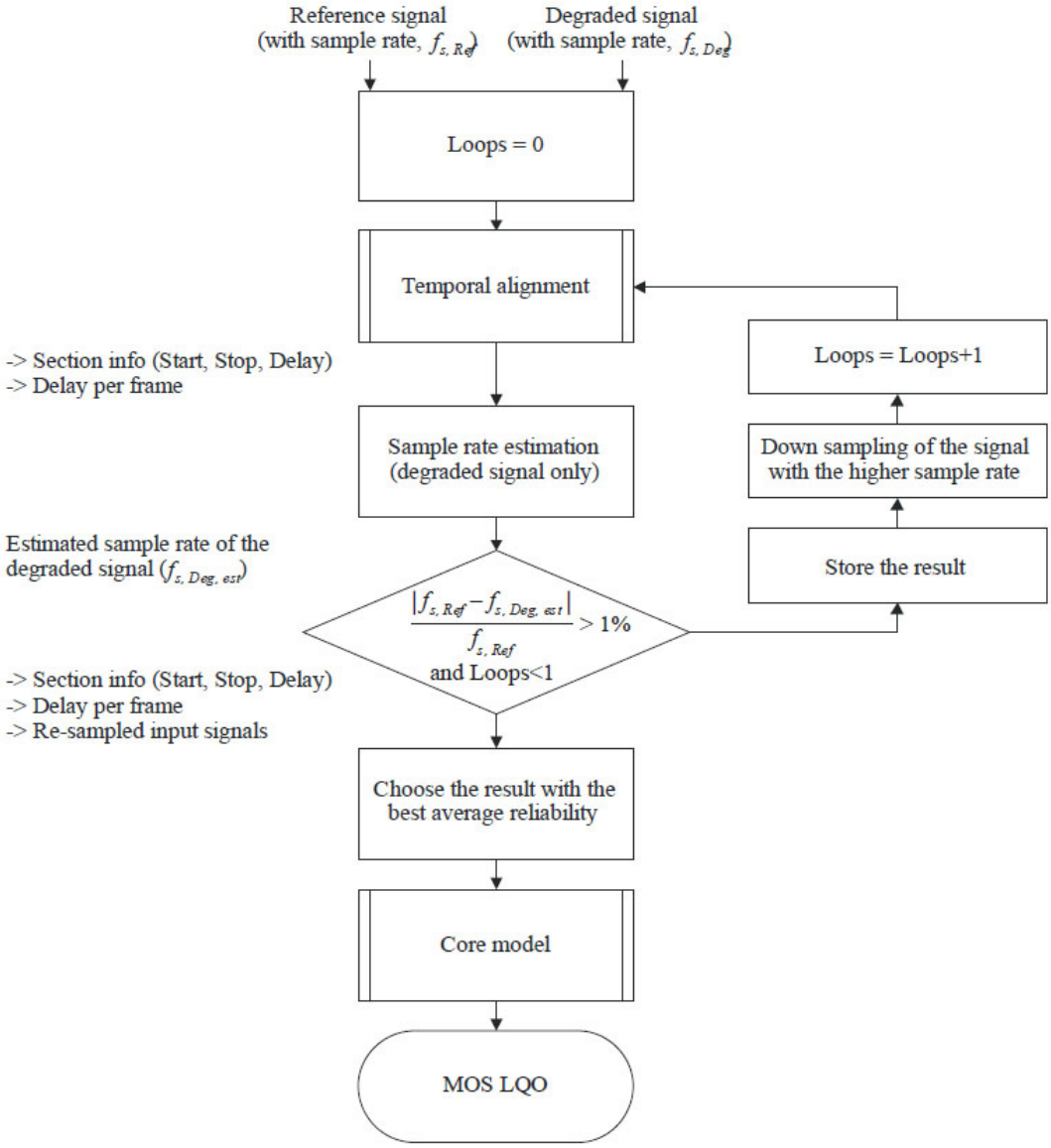


Fig. 5.4 General overview of the ITU-T P.863 algorithm, taken from [6]

Signal delay can vary during the transmission, so it is important to evaluate the delay on sufficiently short sections of the signal. The delay is always expressed as a difference of the degraded signal compared to the original because the algorithm, if possible, searches the corresponding segment of the original signal for each segment of the degraded signal. In the algorithm POLQA, this is implemented in three steps. First, the signal is divided into the so-called macro frame of the same length. The length depends on the sampling frequency of the signal. Evaluation of sampling rate is based on the delay information calculated in the block of time alignment. If the sampling frequency of the two signals differs by more than 1%, the higher frequency signal is resampled.

5.4.2 Time Alignment

The basic concept of time alignment is:

- Distribution of signals to each other corresponding segments of equal length, and calculation of the delay for each section separately
- If possible for each section of the degraded signal, the corresponding part of the original signal is searched. Therefore, the delay is expressed as an offset of the original signal from the degraded and not vice versa.
- The delay of each frame is updated successively, so as to avoid scanning of too long segment of the signal, which would be time and processing power consuming.

A block of time alignment consists of blocks of **filtration, presets, coarse adjustment, fine adjustment,** and **merging sections**. The input signal is divided into frames of equal length whose length depends on the sampling frequency of the input signal.

The principle of finding delay (time shift of two signals) is to calculate the maximum of the correlation function of the two signals. The calculated correlation is stored in the histogram, the signals are shifted by one step, and the correlation is calculated again. The position of the highest value found in the histogram corresponds to the delay.

At the beginning of the algorithm original and degraded signal are bandpass **filtered**. The shape of the filter depends on the operating mode. In narrowband, mode signals are filtered from 290 Hz to 3300 Hz, then in the super wideband mode from 320 Hz to 3400 Hz. These filtered signals are used only for time alignment, the functional model itself uses a different filter.

Presets searches sections of active speech in the signal, calculates a first estimate of delay of each frame and determines the scan range which is necessary for determining the delay of each frame (uncertainty of the initial estimate of delay).

Coarse adjustment gradually refines the delay values for each frame using a reverse search (backtracking). Resolution increases gradually in order to maintain the correlation function and keep searched ranges as short as possible.

Fine adjustment determines the final delay value of each frame. The accuracy of this step is determined by resolution of the final step of the coarse adjustment.

Merging section merges segments with the same or very similar delays under the so-called sectional information.

Time alignment process has the following features:

- There is no limit of the static delay.
- The algorithm is designed to handle the changes in the length of the delay up to 300ms, but there is no hard limit.
- The delay may vary from frame to frame.
- Small difference in the sampling rate (up to 2 %) is compensated within the block of time alignment. The larger difference is detected and resolved in a different part of the algorithm.
- Signals stretched or compressed in the timeline, with or without pitch correction, are evaluated correctly (unlike PESQ).
- The evaluation also works with very noisy signals (SNR less than 0 dB).
- Signals with variable level are evaluated correctly.

Overview of the first part of the algorithm POLQA is shown in Fig. 5.5.

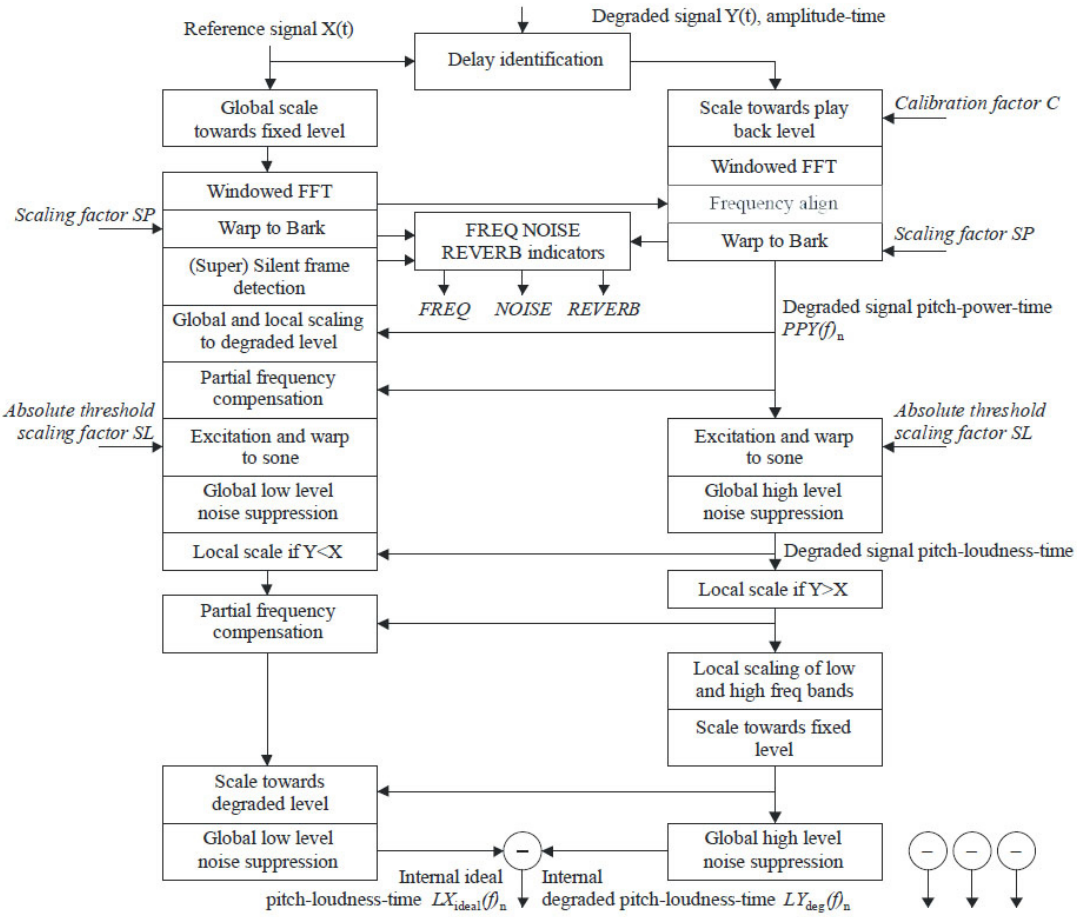


Fig. 5.5 Overview of the first part of the algorithm POLQA, taken from [6]

5.4.3 The Core Model

The main element of this part of the algorithm is a perceptual model which is calculated for four different settings to cover all major types of interference. Interference, in this case, is divided to added interference and lack of signal. For both types, it takes into account the strong and weak effect. Entries of perceptual model are waveforms and delay information. The outcome is a "disturbance density," which is a measure of the interference perceptibility in the signal.

The perceptual model also provides indicators of disturbances in the frequency spectrum, noise, disturbance that is associated with reverb and echo, spectral flatness and changes in volume level. If very high interference is detected the number of " disturbance densityes" is reduced from four to two, one for the added interference and one for the missing signal.

The disturbance density is only a measure of perceptibility of interference in the signal. It determines how easy or difficult it is to perceive this interference when listening to the test sample. It does not take into account any cognitive effects that are important if the real listeners in the subjective test are asked to assess the quality of the presented sample.

Subjective evaluators basically convert the "disturbance density" on their own degree of discomfort and inconvenience when listening. POLQA algorithm performs this conversion by correcting values of "disturbance density" for the following situations:

- Significant level variations
- Many frame repetitions
- Strong timbre
- Spectral flatness
- Noise switching during speech pauses
- Many delay variations
- Strong variations of the disturbance density over time
- Strong variations of the loudness of the signal

All these operations are performed on the frames of length 32 or 43ms (depending on the sampling frequency) using a 50 % overlap between the adjacent frame and for each Bark band separately. In the final step, all indicators in both time and frequency domains are grouped together in order to calculate the final value of MOS LQO.

Overview of the second part of the algorithm POLQA is shown in Fig. 5.6.

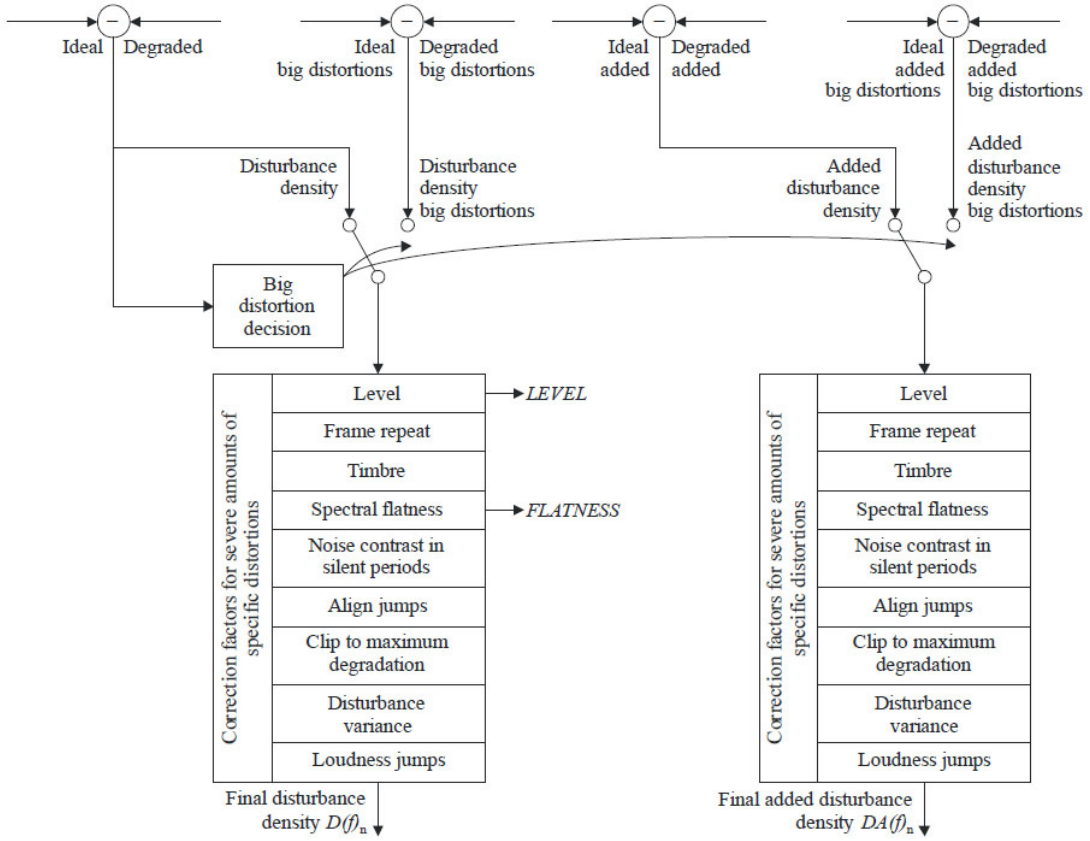


Fig. 5.6 Overview of the second part of the algorithm POLQA, taken from [6]

5.4.4 POLQA perceptual model

One of the biggest changes in POLQA compared to PESQ is the application of the reference signal idealization concept. The idea is that POLQA simulates ARC tests where the audience (assessors) do not have a reference signal, and the degraded sample is evaluated only according to their own ideas about how the perfect signal should look. It is also expected that subjects have an idea of how the perfect sample should sound. This implies that if listeners were ot rate the reference sample which is not ideal (for example it has a wrong volume level, noticeable noise or reverberation, etc.), they would rate it as worse. In block idealization, therefore, POLQA corrects minor errors of the reference signal so that in a further process the degraded signal is compared with the signal which is closest to the ideal reference. As with the idealization of the reference signal, it is processed in the preparation of the degraded sample. Some errors (such as small changes in pitch or linear frequency distortion), which are hardly noticeable for the subjective listener (unlike objective algorithm) and in the ACR test it will unset, are corrected.

The perceptual model starts with adjusting the volume of the reference signal. The average volume of active speech is set to -26 dBov[1]. The degraded signal is not corrected in a similar way. It is presumed that any variation of the volume of the degraded signal from the reference will be scored as an error.

In the following step, the two signals are divided into blocks of 32 or 43 ms (depending on the sampling frequency) with 50 % overlap and the spectrum is calculated for each block using the FFT. Consequently, small changes in the pitch of the degraded signal are corrected (frequency dewarping). Subsequently, the spectra are transformed into psychoacoustic scale by merging the various spectral lines in the so-called critical bands. The used scale is similar to the bark bands. The result is a "power spectral density." In this step the first three indicators of interference are calculated for the following disorders: frequency distortion, added noise and room reverberation.

Then, the "sensory effect" (excitation) of each band is derived. This includes thresholding and scaling of individual bands in both time and frequency domains. The result is a psychoacoustic representation that shows how strongly and loudly will be the given component of the signal perceived by listeners.

Then it is following another idealization of the reference signal where inappropriate tone and stationary noise is filtered. At the same time, the algorithm partially removes noise and linear frequency distortion of the degraded signal.

Final "density disturbance", which is a measure of audibility and perceptibility of interference in the degraded signal, is calculated using the comparison of the corresponding excitations Overview of the third part of the algorithm POLQA is in Fig. 5.7.

[1] dB(overload) – the amplitude of a signal (usually audio) compared with the maximum which a device can handle before clipping occurs.

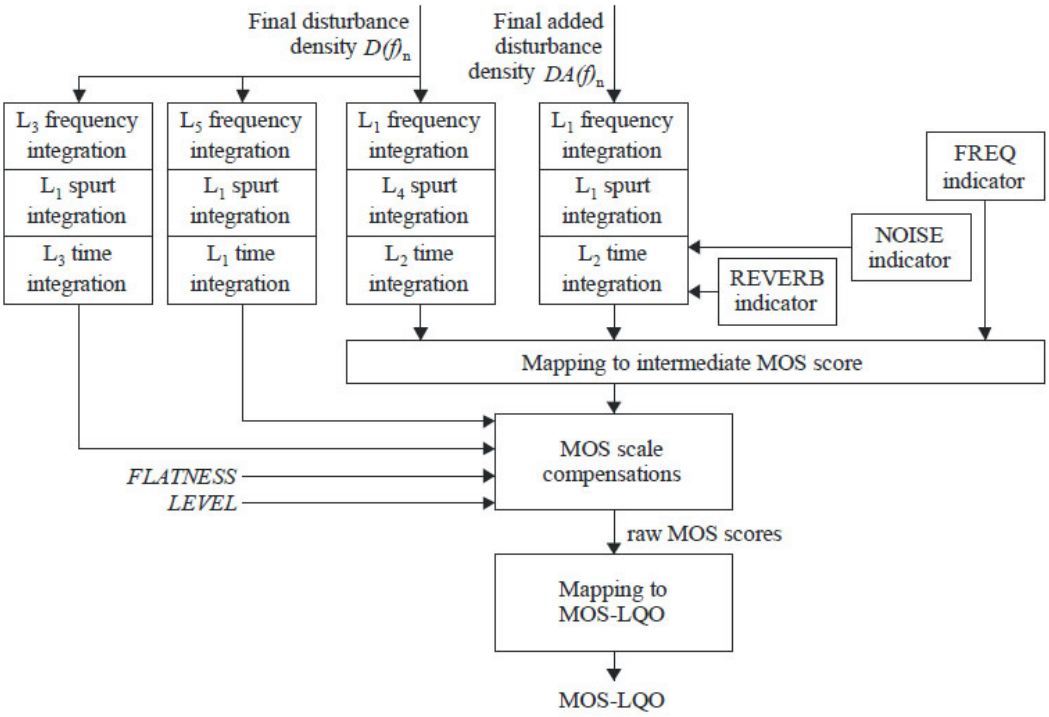


Fig. 5.7 Overview of the second part of the algorithm POLQA, taken from [6]

5.4.5 Operating modes of POLQA

POLQA algorithm allows you to switch two operating modes that are used to evaluate the wideband and narrowband signals.

In the wideband mode, other input filter and other mapping function to compute the final MOS LQO is used than in the narrowband setup. For wideband measurement sampling frequency of 48kHz is required. For narrowband, the allowable values are 8kHz, 16kHz, and 48kHz. Implementation of the algorithm POLQA from OPTICOM contains a library allowing to resample the signal to the desired value.

In the wideband mode, the possible bandwidth limitations of the test sample are evaluated as degradation and the final MOS is lowered. Also in this setting, the appropriate wideband reference signal must be used. The signal quality is modeled as if the listener in subjective tests evaluated it using headphones (diffuse field equalized headphone) in diotic mode (mono signal is played in both ears).

The narrowband mode, the test signal is compared with a narrowband reference. Bandwidth limitations typical for conventional phone calls are not classified as an error or interference, and the grade is reduced much less than in

the wideband mode. Modeling sound quality is equivalent to the subjective tests using the IRS type handset and mono signal played in one ear.

When interpreting the results of the algorithm POLQA it is always necessary to specify what settings was used. Although both modes produce results on the same five-point scale, it is strictly prohibited combining wideband and narrowband results. If it is necessary to compare the features of wideband and narrowband networks, the wideband algorithm must be used for both cases.

5.4.6 Perceptual results

This chapter provides an overview of the measured parameters.

MOS-LQO

The most important result of the algorithm POLQA is MOS-LQO. This number directly describes the quality of voice/speech on the MOS scale. Values are defined in UTU-T Recommendation P.863 that use the assessment on a scale similar to the MOS, ranging from 0 (worst) to 4.5 (best) for the narrowband mode and from 0 to 4.75 for wideband. The results of the algorithm POLQA itself are therefore numerically slightly different from the results obtained in subjective listening tests using real listeners.

By a large comparison with the results of subjective tests a mapping function (2,3) that converts the results of the P.863 algorithm to MOS-LQO was defined. It is directly comparable with the results of subjective tests or older algorithms. All available implementations of POLQA algorithm contain this feature.

Mapping functions are defined as follows:

Narrowband mode:

$$MOS_{LQO} = 0{,}79 + 0{,}0036 * MOS_{P.863} + 0{,}2117 * MOS_{P.863}^2 - 0{,}0065 * MOS_{P.863}^3 \quad (2)$$

Wideband mode:

$$MOS_{LQO} = 0{,}276 + 0{,}7203 * MOS_{P.863} - 0{,}00756 * MOS_{P.863}^2 + 0{,}01141 * MOS_{P.863}^3 \quad (3)$$

G.107 R-factor

Libraries algorithm POLQA also contain a mapping function for the conversion of MOS-LQO to the scale used in measurement according to G.107 (E-model) [29], wherein the resulting parameter is equivalent to Ie-Value (also often referred to as an R-factor). The scale ranges from 0 (worst) to 100 (best). Mapping function between R-scale and MOS-LQO is defined in G.107.

5.4.7 Reporting and averaging of results

If possible, the results of the algorithm POLQA should be published after averaging a number of measurements. At minimum, the measurement by two male and two female voices should occur during the same test condition. Averaging can then be carried out in the MOS domain. The second option is to use the concatenated voice sample which contains short segments from different speakers.

When publishing the results, it should always be stated, whether the averaging took place in MOS domain or signal domain.

It is likely that the results obtained using the concatenated sample will be different from the results obtained by averaging the results of each of its parts. This is due mainly to the different length of the samples and the non-linear averaging during algorithm POLQA. Influence of length of the sample to a final value of the MOS is handled in more detail in another part of this work.

Of course, there are applications where the averaging is not possible. This concerns, for example, drive tests or tests that take place in real networks with non-deterministic behavior. In this case, it is possible to publish the results without averaging, but it is useful to highlight this fact.

If averaging is possible, the maximum and minimum values measured during testing and the number of samples used for averaging should be included in the results , in addition to the actual average. Additionally, it is also advisable to calculate 95% confidence intervals (measurement uncertainty). The number of samples used for averaging depends on the specific application, in particular on the frequency and regularity of occurrence of the test conditions and the required measurement accuracy.

Results are ideally displayed in graphical form.

5.4.8 Accuracy of POLQA results

A common mistake when interpreting the results of the algorithm POLQA is overestimating the accuracy of the calculated results.

Listeners In subjective listening tests only have the five-point scale without decimals available. When attending thirty evaluators the theoretical accuracy of the result is 0.03 MOS. However, this applies only in the ideal case where there is no other source of inaccuracy. In fact, the reliability of the listening test is considerably lower.

The accuracy of results of POLQA algorithm depends on the application, the sample used and tested conditions. For individual measure, the uncertainty is

typically 0.3 MOS. Measurement accuracy can be further improved by averaging and specifying confidence intervals (measurement uncertainties).

5.4.9 Limits of algorithm POLQA

Static delay cannot be evaluated in listening tests. Although the delay is measured in POLQA algorithm, it is not considered as an error of the degraded signal. Delay variations that occur in sections of active speech are evaluated accordingly.

5.5 Network Simulation and Emulation

Simulator (emulator) of IP environment is a software tool that allows modelling of the behaviour of networks and network elements. It has many applications in the design and construction of complex network projects (businesses, schools, telecommunication networks, etc.). Another area of use is the development and production of hardware network elements themselves and finally, the development and implementation of software such as transmission protocols, codecs, and communication tools (instant messengers, IP phones). Depending on the application, these programs are divided into simulators, emulators, and traffic generators.

Simulator - It is a purely virtual instrument. It allows the developer to create, using one computer, a model of the entire network and its behavior under various loads and various modes of operation. The range of simulated network depends only on the capabilities of a particular simulator and power of the used computer. The advantage is the ability to quickly and cheaply test the proposed network and detect possible errors in the project before investing in construction. Problems of this method are mainly that the simulated structure may differ from the one later realized, and real network elements may behave differently than their virtual images.

Emulator - Unlike the simulator, the emulator does not generate the whole network, but only its outward manifestations. This allows one to test the response of real elements to be involved in any large network. A computer with implemented emulator can, in laboratory conditions, represent hundreds of the network elements and thousands of kilometres of the lines. Emulated transfer properties are in particular: delay, jitter, packet loss, duplicate packets, and bandwidth. The advantage of this method is in testing real elements against the real load. The problem may be that the emulated load differs from the real network in the detailed parameters (delay distribution, selecting of the "lost" packets, etc.).

Traffic Generator - It is the type of emulator that does not produce the transmission errors, but introduces to the network a traffic load (data transfer, requests to servers, etc.). Essentially it emulates the terminal network elements.

6 Objective Experiments

Objective experiments are a suitable substitute for subjective testing. They are less time consuming and easier to organize.

The purpose of the following series of experiment is to verify the effect of the IP channel parameters on the quality of voice transmission. The second goal is to check the behavior of the testing algorithms in modern VoIP environment.

6.1 Evaluation of objective speech transmission quality measurements in packet-based networks

This experiment presents an analysis of the relation between IP channel characteristics and final voice transmission quality. The NISTNet emulator is used for adjusting the IP channel network. The transmission quality criterion is an MOS parameter investigated using the ITU-T P.862 PESQ, future P.863 POLQA and P.563 3SQM algorithms. Jitter and packet loss influence are investigated for the PCM codec and the Speex codec.

6.1.1 Experiment description

Test-bed

The test-bed (Fig. 6.1) consisted of three computers, an Opera audio analyzer, and interconnecting cables. A concatenated speech file in WAV format (8kSa/S, 16bit), 16.75 s in length, was used. The file contained four short sentences spoken by four different speakers (two men, two women) and adequately covered the entire human speech spectra. Due to this fact, the concatenated file was used as an effective replacement for testing using multiple speech samples.

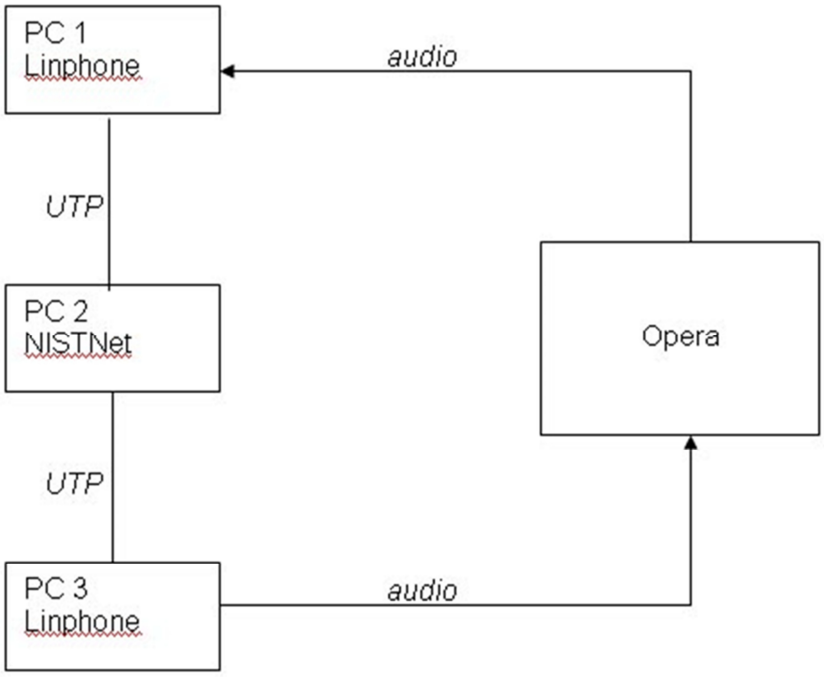


Fig. 6.1 Test-bed

The signal was transferred from an audio output “line 1 out” of the Opera analyzer to an audio input (microphone) of PC 1. PC 1 and PC 2 were connected by a UTP network cable (subnet 192.168.0. X), as were PC 2 and PC 3 (subnet 192.168.1. X). PC 2 was therefore fitted with a two-port network interface card. The test signal was transferred from PC 1 to PC 3 using a VoIP call in the Linphone program, using PCM (G.711) and Speex codecs. From PC 3, the audio output (headphone) signal was led back into the audio input “line 2 in” of the Opera analyzer. The NISTNet emulator [30] was running on PC2, which (according to the specific settings) introduced transmission errors between PC 1 and PC 3. The results depend on the accuracy and repeatability of the network simulation. We proved by several experiments [31], [32] that NISTNet suits these requirements satisfactorily. It was also used in other experiments [33]. The measured samples were adjusted in Adobe Audition 3.0 (converting stereo → mono) and then tested using the POLQA (ITU-T P.863) PESQ (ITU-T P.862) and 3SQM (ITU-T P.563) algorithms [34]. The PESQ algorithm output was recalculated to the value of MOS-LQO (Listening Quality Objective) according to a mathematical prescription defined in ITU-T P.862.1. According to the official wording of P.862, the effect of packet loss on CELP coded transmission can be tested in this way. It was not tested for PCM transmissions affected by packet loss, but the recommendation itself does not prevent any user making such tests [2], [35].

6.1.2 Tested transfer parameters

Jitter

The limited speed of signal transmission in the network and signal processing components on the route, e.g. routers and converters, cause signal delay. The speed of the signal transmission is a particular problem when a call is made over a long distance or is tranferred via satellite for part of the route. The delay alone does not affect the quality of the transferred signal. However, the delay is usually not constant. The sender generates packets at the same time intervals, but the network parameters may be changed during a call. Consequently, the transmission delay varies during the call. This phenomenon is called jitter and may cause problems with the delivery of packets. It may change their order and thus impair the signal quality. It can be buffered to some extent to compensate on the recipient site. In this experiment, the jitter buffer of Linphone was set to default 60 ms. The NISTNet emulator allows the mean value of the delay to be set (parameter delsigma). The delay of each packet is randomly generated, with a normal distribution around this value. In this experiment the following values (in ms) were adjusted: 0, 10, 20, 40, 60, 80, 100, 125.

Packet Loss

The root cause of packet loss during transmission may be a route failure (drop-out of the satellite or microwave links), or saturation of the router buffer. Sometimes the packet is not used in the reconstruction of the signal, due to its excessive delay. Losses may be dependent (the probability of packet loss depends on whether the previous packet was lost) or independent. A suitably long speech sample with a high speech activity factor should be used in the case of independent losses, in order to assure uniform distribution of impairments in different measured samples. However, the sample length is limited by the requirements of the recommendation (20s as given by P.862.3). Independent losses were used in this experiment, and the following values were adjusted (in %) 0, 1, 2, 4, 6, 8, 10, 12.

6.1.3 Results

For each IP channel parameter setting, five samples were measured and processed. Using statistical processing of the results, a confidence interval of CI95 was calculated. It is displayed in the graph as error bars. The speech sample that was used is long enough even for the low packet loss values that were tested. Five repetitions are enough to achieve satisfactorily low result dispersion and uncertainties. The results for PESQ [2], [26] and 3SQM [27] are in agreement with previous experiments [36]–[39]

Jitter

Changes in the delay in the transmission have a major impact on the quality of the transferred voice. The Linphone jitter buffer was set to its default value of 60 ms. When setting the parameter delsigma to 40 ms and higher, the jitter buffer on the receiver side can no longer fully compensate for errors caused by the disorderly packet delivery. When the jitter is higher than 100 ms, the transferred signal is almost unintelligible. The results for the PCM codec are depicted in Fig. 6.2, and for Speex codec in Fig. 6.3. In both cases, the new POLQA algorithm predicts a higher MOS value than PESQ. Non-monotonicity of the graph in the range of 0–20 ms, particularly evident in the POLQA algorithm, is probably caused by an outlying result from one sample. It can be assumed that the measurement of significantly larger numbers of samples would cause extermination of the graphs. This may be a subject of further experiments. The graphs also show that the 20–40 ms samples have significantly greater variance than in the rest of the chart. This is a critical area for the jitter buffer on the receiver side, where one sample is buffered and the other, which is slightly different, is not buffered.

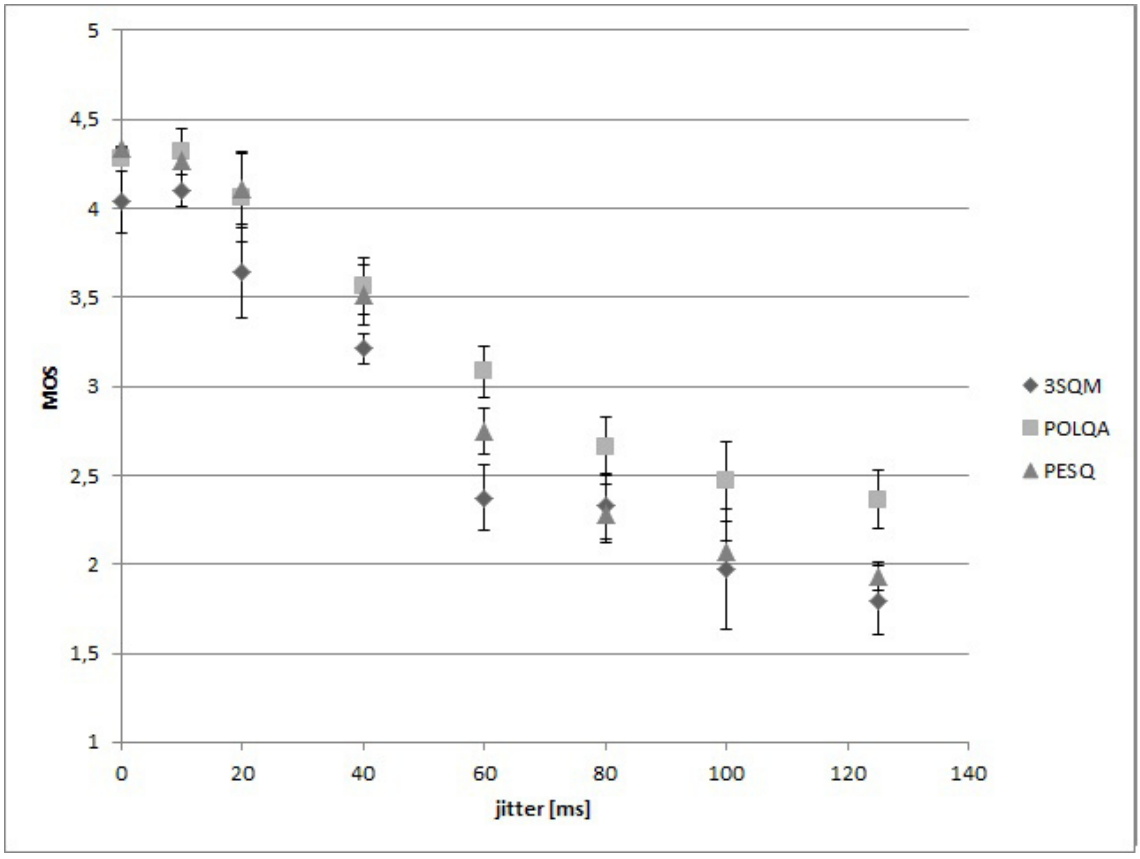


Fig. 6.2 MOS as a function of jitter (PCM codec).

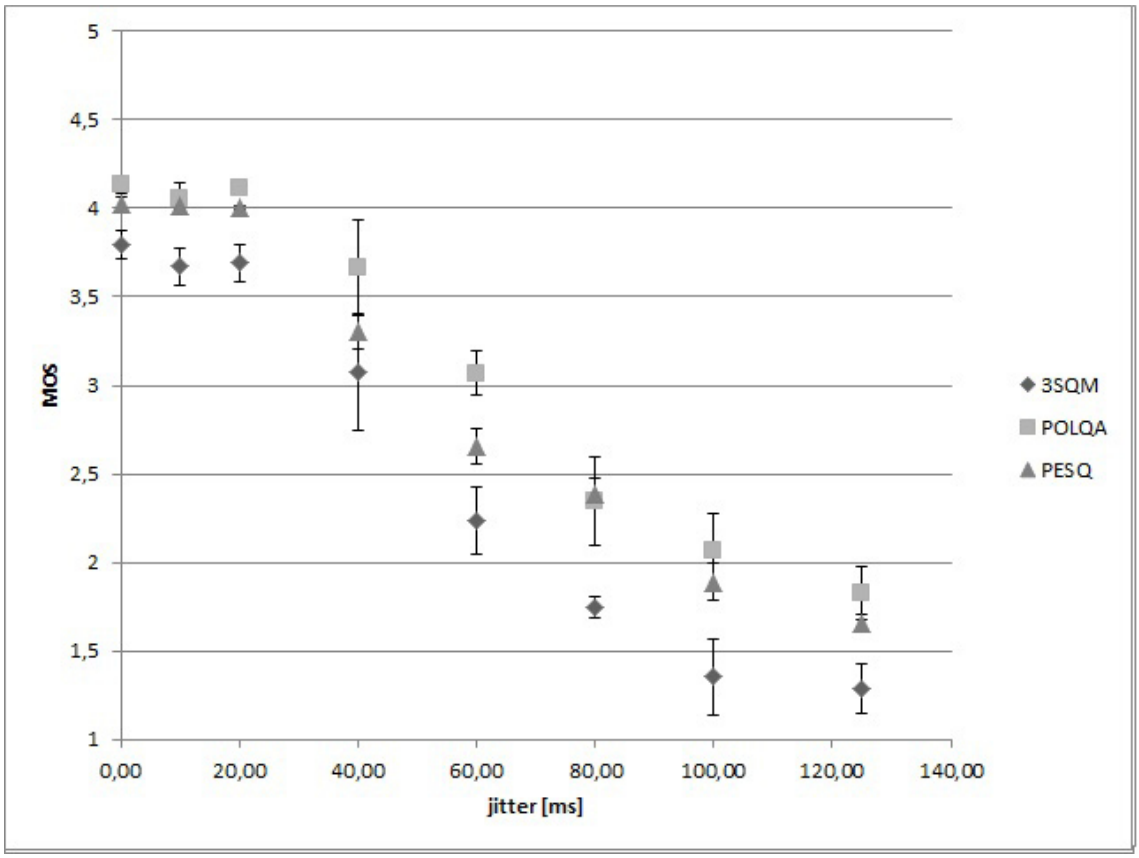


Fig. 6.3 MOS as a function of jitter (Speex codec).

Packet loss

Packet loss also affected the quality of the transferred voice. The results can also be affected by Packet Loss Concealment (PLC), implemented by Linphone. Comparing our results with [36], it is obvious that our Linphone had no PLC implemented. The MOS value drops below 4 already for 1% packet loss and for losses greater than 10% the transferred signal was unintelligible. The results for the PCM codec are shown in Fig. 6.4, and for the Speex codec in Fig. 6.5. Both graphs show an evident decrease already for 1% packet loss. Again, it can be seen that the POLQA algorithm predicts higher MOS values than PESQ. The POLQA result dispersion is probably due to the measurement procedure (number of samples). It is possible that POLQA is more sensitive to disturbances, and therefore provides larger variance of the results for the same samples than PESQ. Using more samples would probably cause extermination of the graphs.

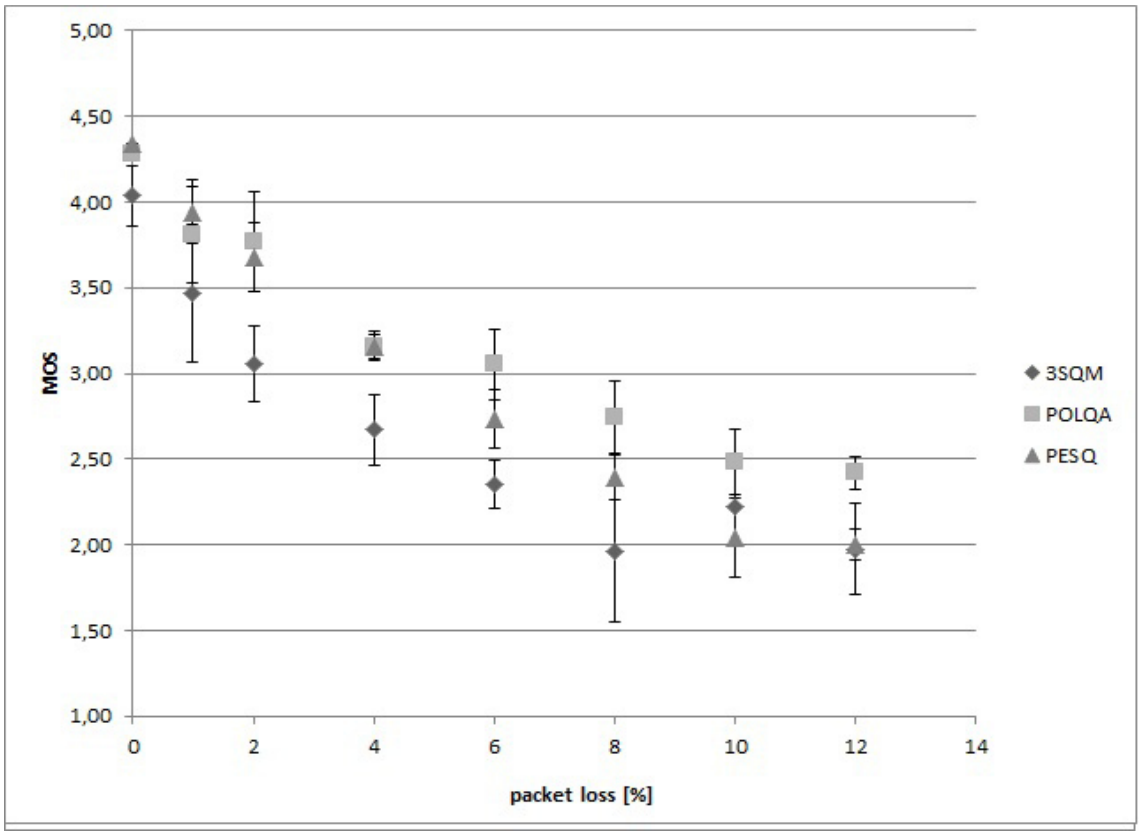


Fig. 6.4 MOS as a function of packet loss (PCM codec).

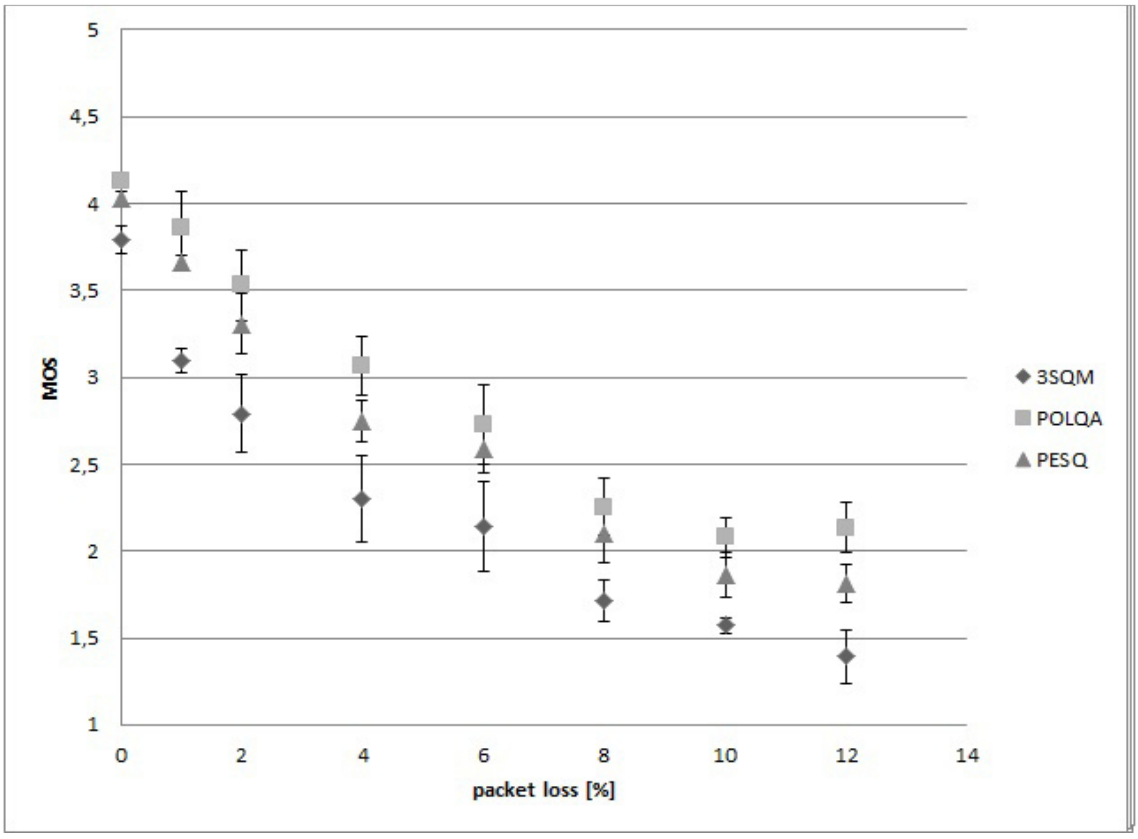


Fig. 6.5 MOS as a function of packet loss (Speex codec).

Combination of packet loss and jitter

In a real traffic, all kinds of defects occur simultaneously. The following charts show the dependence of MOS on a combination of jitter and packet loss. When the receiver decodes the transmitted signal, an excessively delayed packet has the same effect as a lost packet, because neither can any longer be used for reconstructing the transmitted signal. As a result, exposure to both disorders simultaneously causes a faster decline in quality than the separate effects of only one of them. The results for the PESQ algorithm and the PCM codec are shown in Fig. 6.6, while the results for the PESQ algorithm and the Speex codec are depicted in Fig. 6.7. The Speex codec, designed specifically for VoIP, provides only a slightly lower quality of the transmitted voice signal than the common and widely quoted PCMcodec. However, Speex needs only half the bit rate (32 Kbps compared to 64Kbps PCM). The results for the POLQA algorithm and the PCM codec are depicted in Fig. 6.8, and the results for the POLQA algorithm and the Speex codec are shown in Fig. 6.9. A slight difference can also be seen between the PCM and Speex codecs. The dispersion of the results of the POLQA algorithm is also clearly visible. Due to its greater sensitivity than PESQ, the POLQA algorithm probably requires a larger number of samples for statistical analysis.

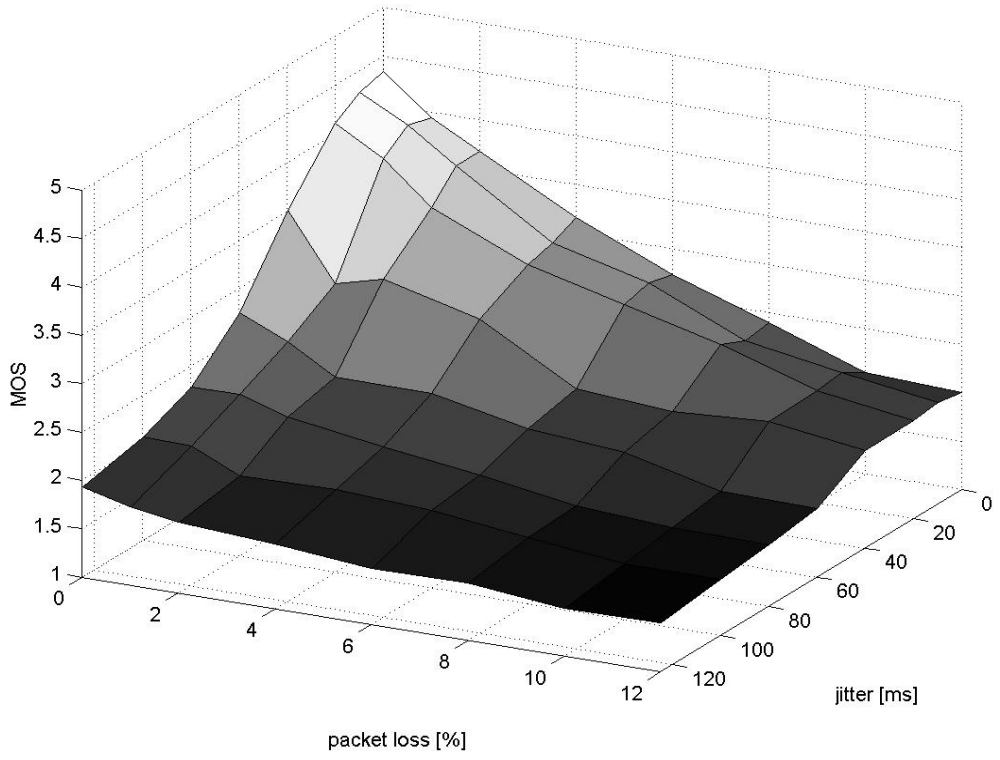


Fig. 6.6 MOS as a function of packet loss and jitter (PESQ algorithm, PCM codec).

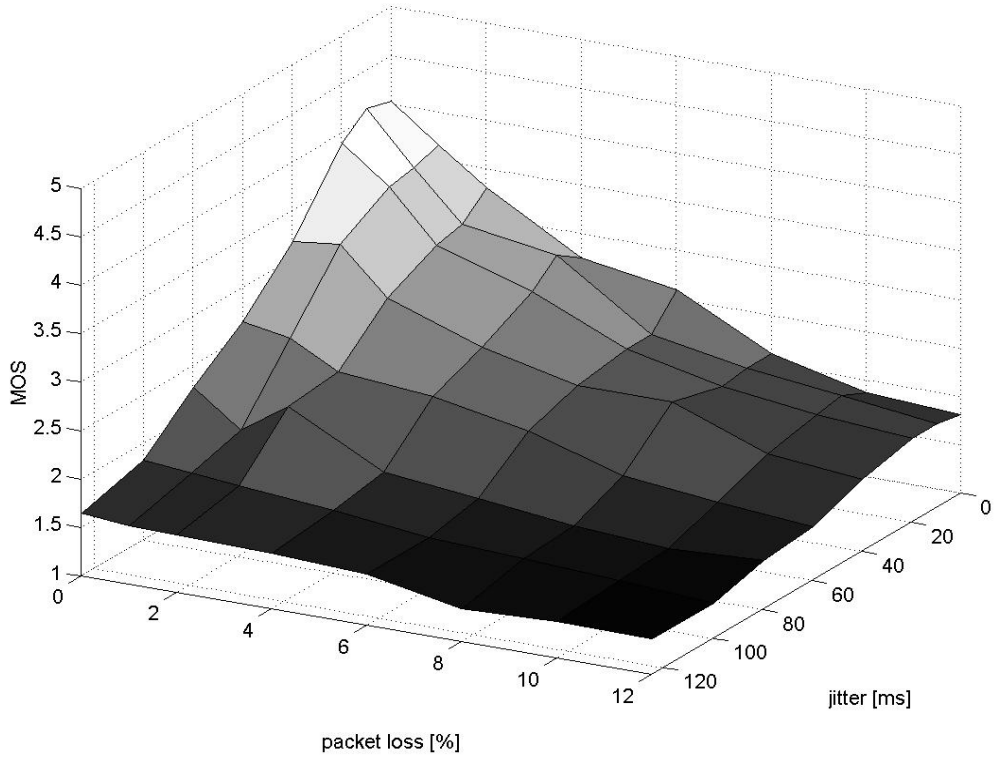


Fig. 6.7 MOS as a function of packet loss and jitter (PESQ algorithm, Speex codec).

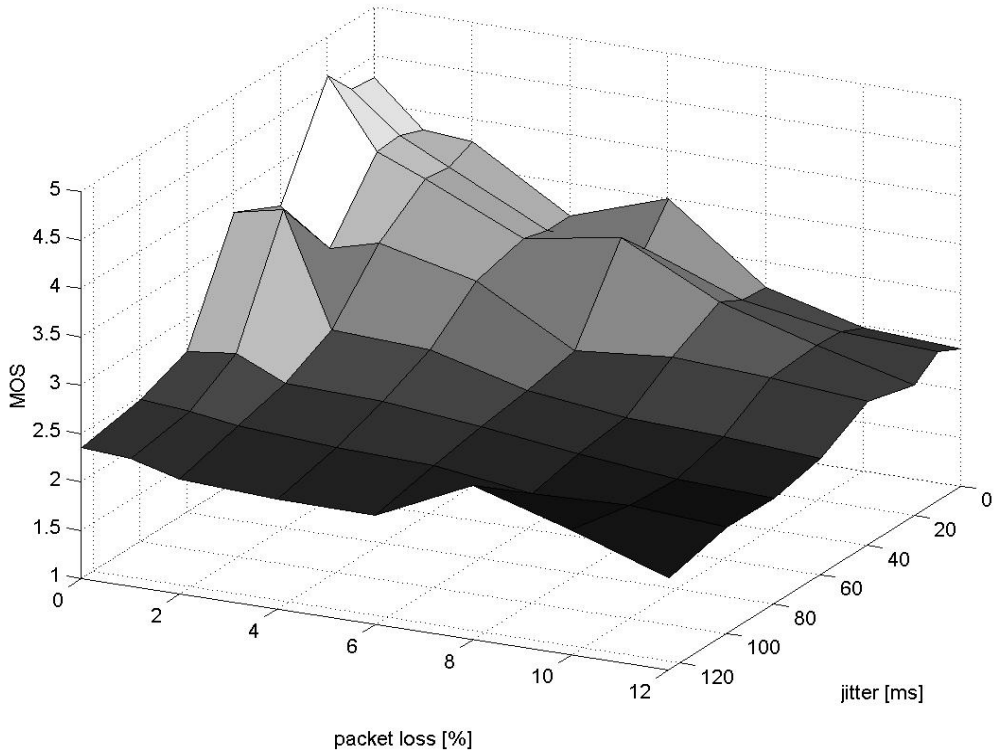


Fig. 6.8 MOS as a function of packet loss and jitter (POLQA algorithm, PCM codec).

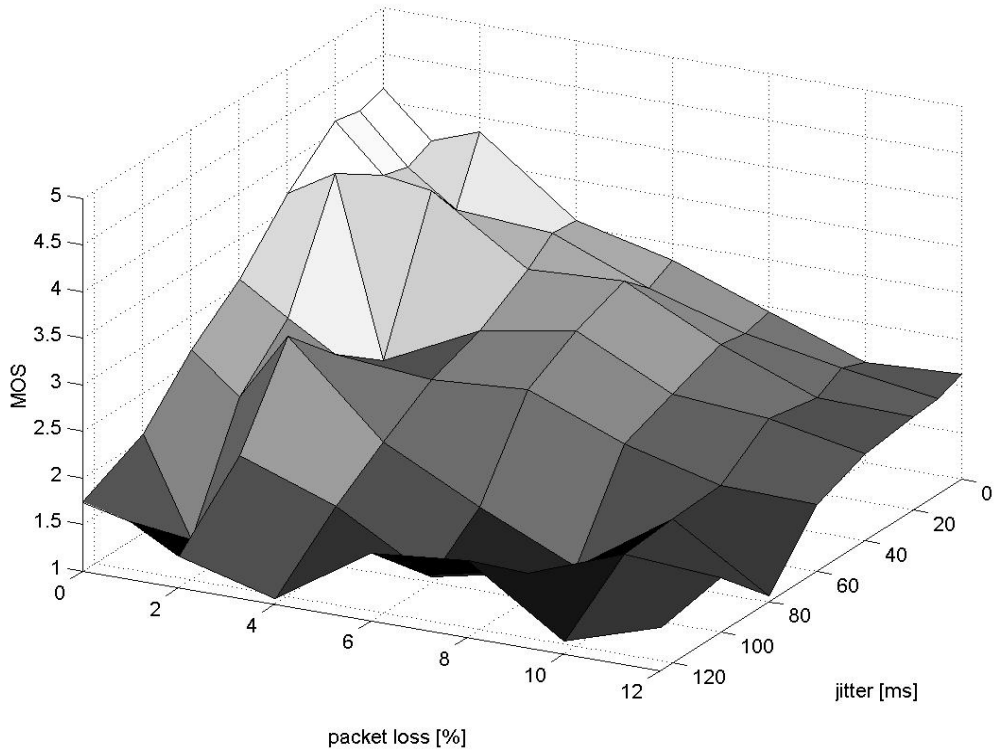


Fig. 6.9 MOS as a function of packet loss and jitter (POLQA algorithm, Speex codec).

6.1.4 Conclusion

This experiment has verified the impact of changes in delay and packet loss on the quality of voice transmission in IP networks and compares the results of the Speex codec with PCM. It has also compared the results of the new POLQA testing standard with older algorithms. The main objective was to identify the relation between the network parameters and the MOS values as delivered by different objective algorithms. A second objective was to compare the Speex codec with the PCM reference codec. This codec, designed specifically for VoIP, provides almost the same transmission quality at half the required transmission rate. The third objective was to explore the MOS predictions of the new POLQA testing standard. This algorithm predicts a higher MOS value for most samples, but its results are highly scattered. It appears that more samples are required for proper function when testing packet loss or jitter.

6.2 Impact of Jitter and Jitter Buffer on the Final Quality of the Transferred Voice

This experiment addresses the relation between packet delay variations (jitter), the length of the jitter buffer and final voice transmission quality. Network emulator NISTNet is used for adjusting the IP channel. VoIP client Linphone is used for adjusting the buffer length. The criterion of transmission quality is a MOS parameter investigated with algorithms ITU-T P.862 PESQ and P.863 POLQA.

6.2.1 Experiment description

Test-bed
The test bed was very similar to the previous experiment (Fig. 6.1). PC2 introduces a disturbance into the signal transmission between PC1 and PC3. Recording and playback of samples are carried out using the OPERA audio analyzer.

A concatenated speech file in WAV format (8kSa/S, 16bit) of the length 16.75 s was used. The file contains four short sentences spoken by four different speakers (two men, two women) and covers the entire human speech spectra adequately.

6.2.2 Tested transfer parameters

Jitter
Emulator NISTNet allows setting the mean value of delay (parameter delsigma). Delay of each packet is randomly generated with a normal distribution around this value. As an example Fig. 6.10 shows a histogram of delay when delsigma was set to 10ms. In this experiment the following values (in ms) have been adjusted: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120. The range of measured values was derived from the results of diploma thesis of Oldrich Slavata [31].

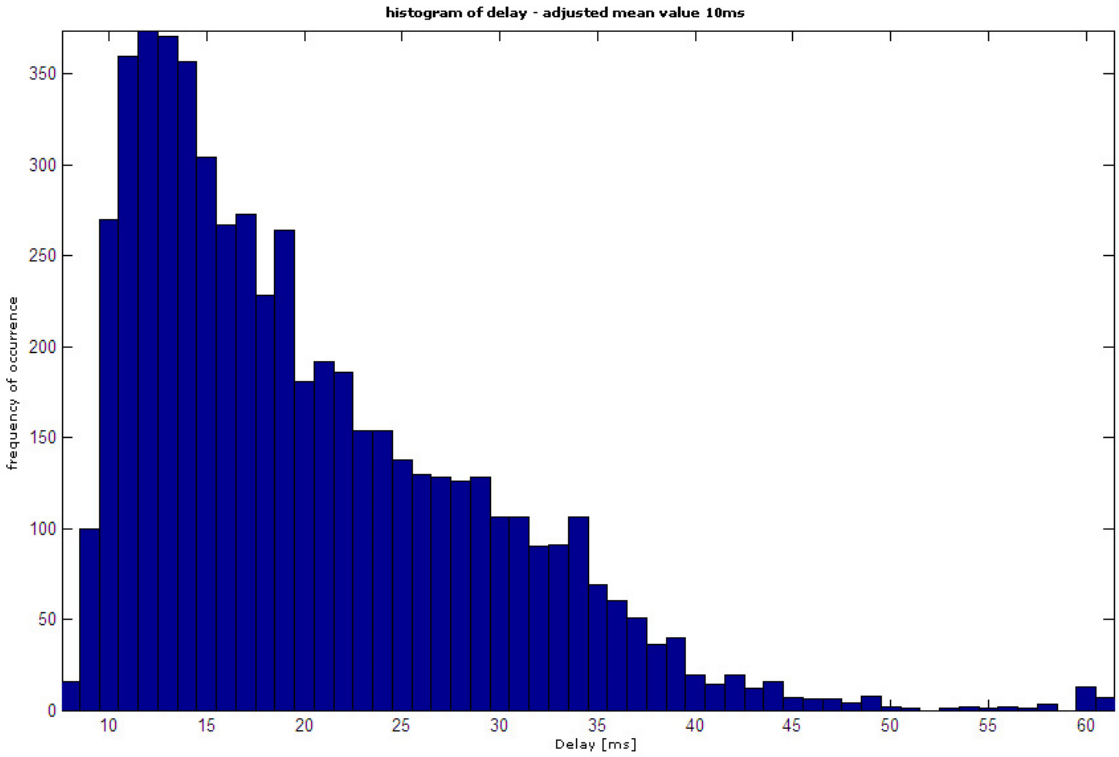


Fig. 6.10 Histogram of delay - adjusted mean value 10ms

Jitter Buffer

Delay variations can be partially buffered to compensate on the recipient side. The longer the buffer is, the better it works, but its length is limited because it extends overall delay and may affect the conversational quality. In Linphone, a simple buffer with a predetermined length is implemented. Algorithm with adaptive length is not implemented. Most commonly used values are around 60ms. In this experiment, the following values (in ms) of Linphone jitter buffer have been adjusted: 0, 30, 60, 90, 120, 180, 240. Comparison of the results of [31], [36], [39] shows that there is not any packet loss concealment algorithm implemented in Linphone.

6.2.3 Results

For each setting of IP channel parameter, ten samples were measured and processed. The used speech sample is long enough even for low values of jitter tested. Ten repetitions are enough to achieve satisfactorily low results dispersion and uncertainties. The confidence intervals (CI95) were computed.

The graphs show a close correlation between the ratio of jitter/jitter buffer and the final quality. If the buffer is significantly shorter than jitter range, it cannot compensate, and the final score is low. If the buffer is significantly longer, jitter effects are almost unnoticeable. In areas where both values are close, it leads to more variance values (Fig 6.11). This is a critical area, where one sample is buffered and another, slightly different, isn't.

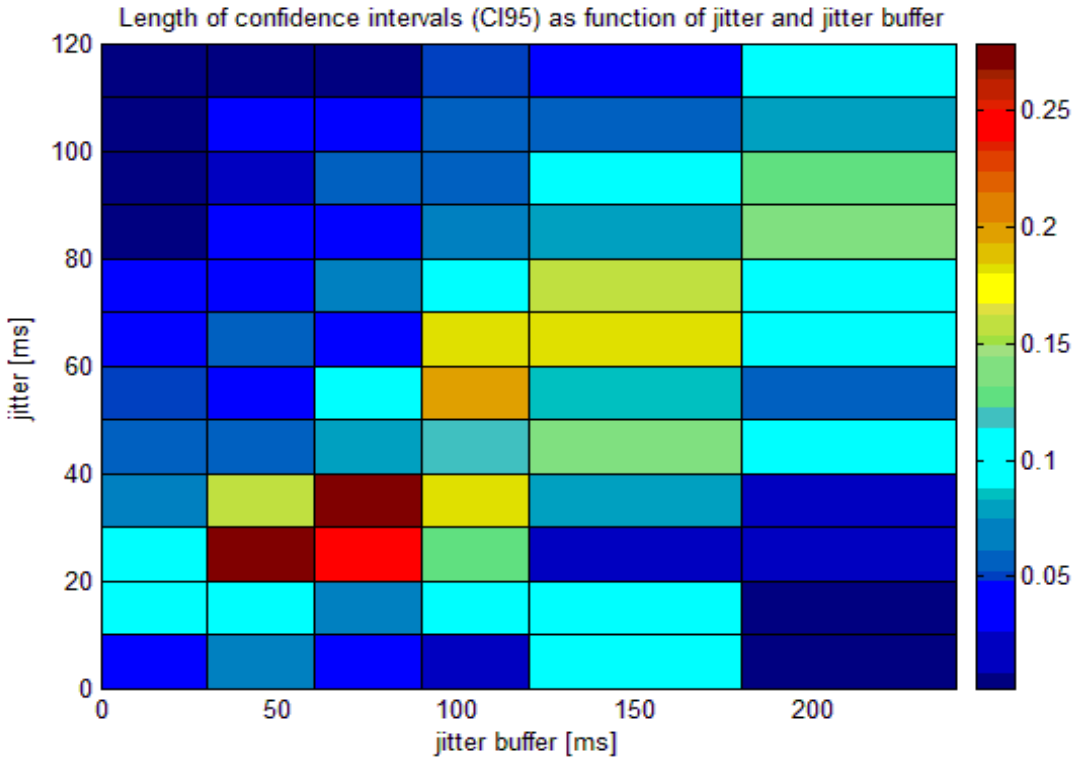


Fig. 6.11 Length of confidence intervals (CI95) as function of jitter and jitter buffer (PESQ)

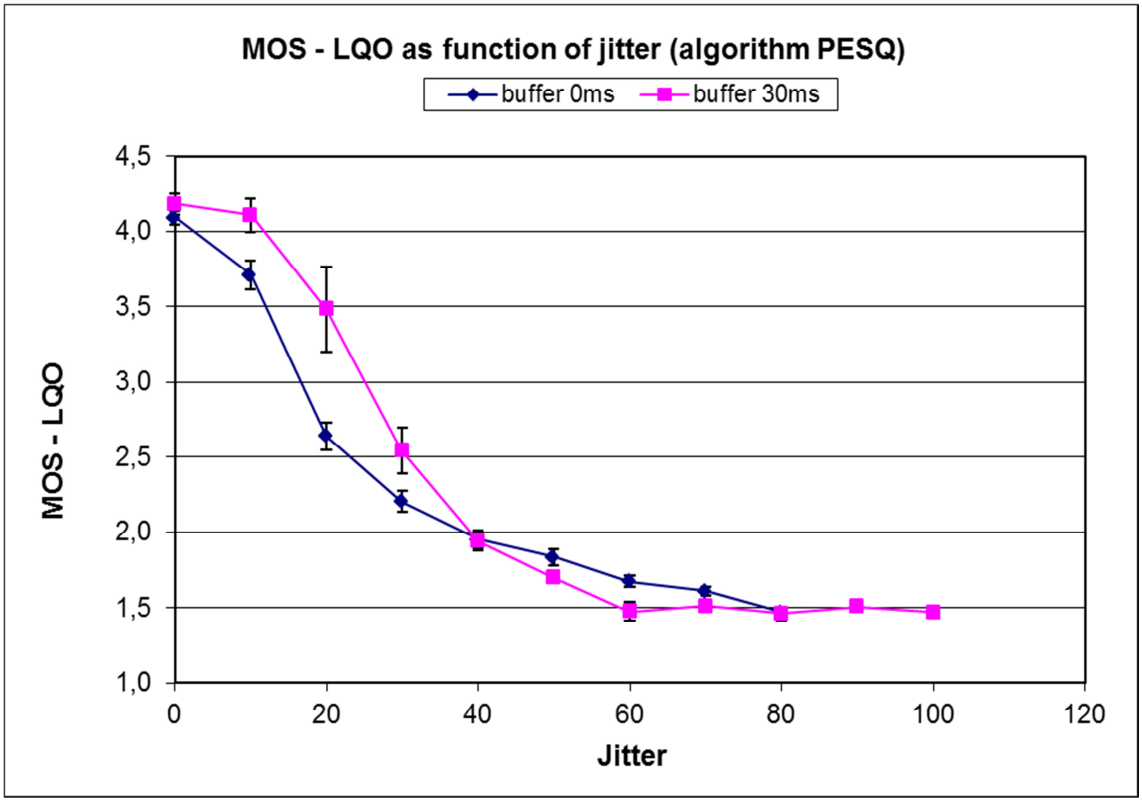


Fig. 6.12 MOS - LQO as function of jitter, comparison of buffer length 0ms and 30ms (PESQ)

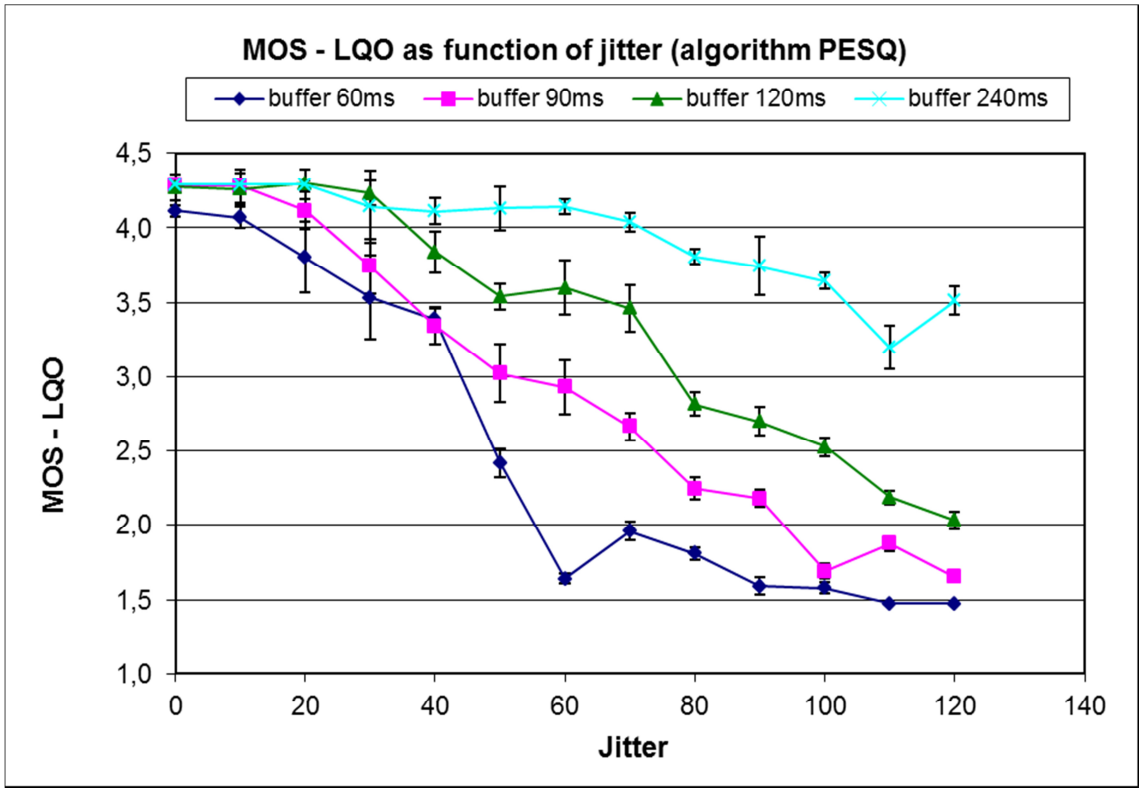


Fig. 6.13 MOS - LQO as function of jitter, comparison of buffer length 60ms, 90ms, 120ms and 240ms (PESQ)

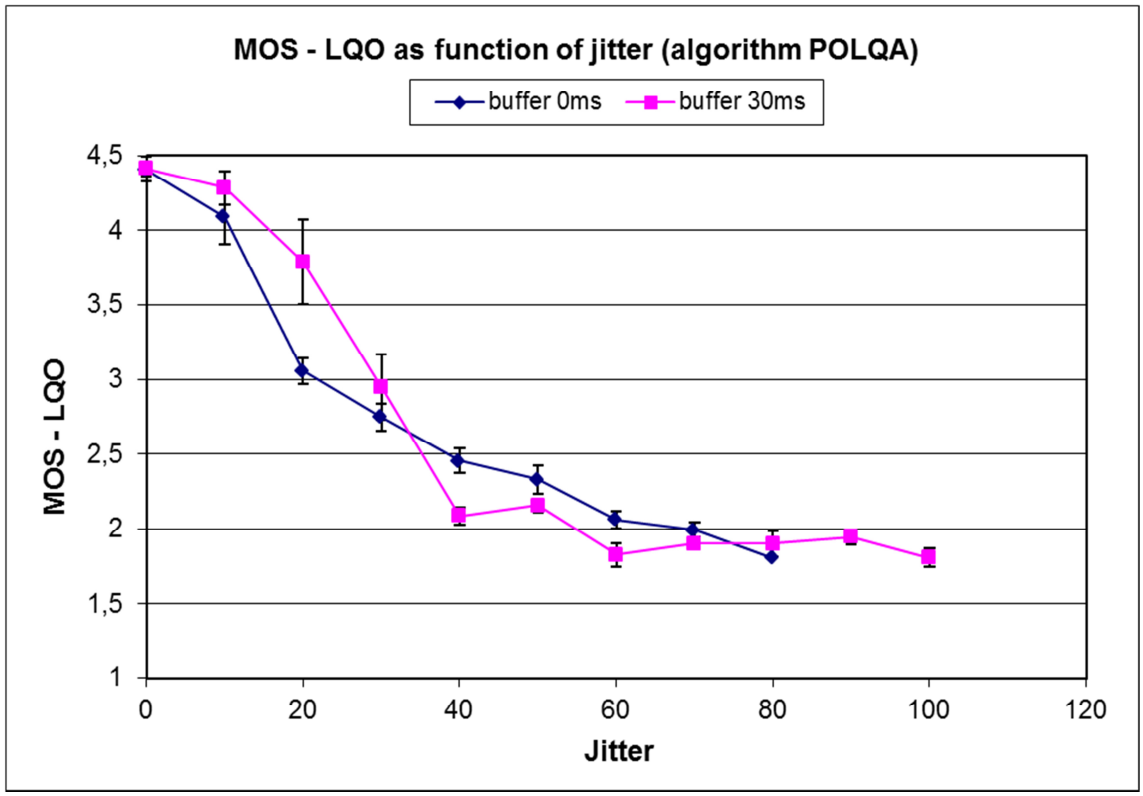


Fig. 6.14 MOS - LQO as function of jitter, comparison of buffer length 0ms and 30ms (POLQA)

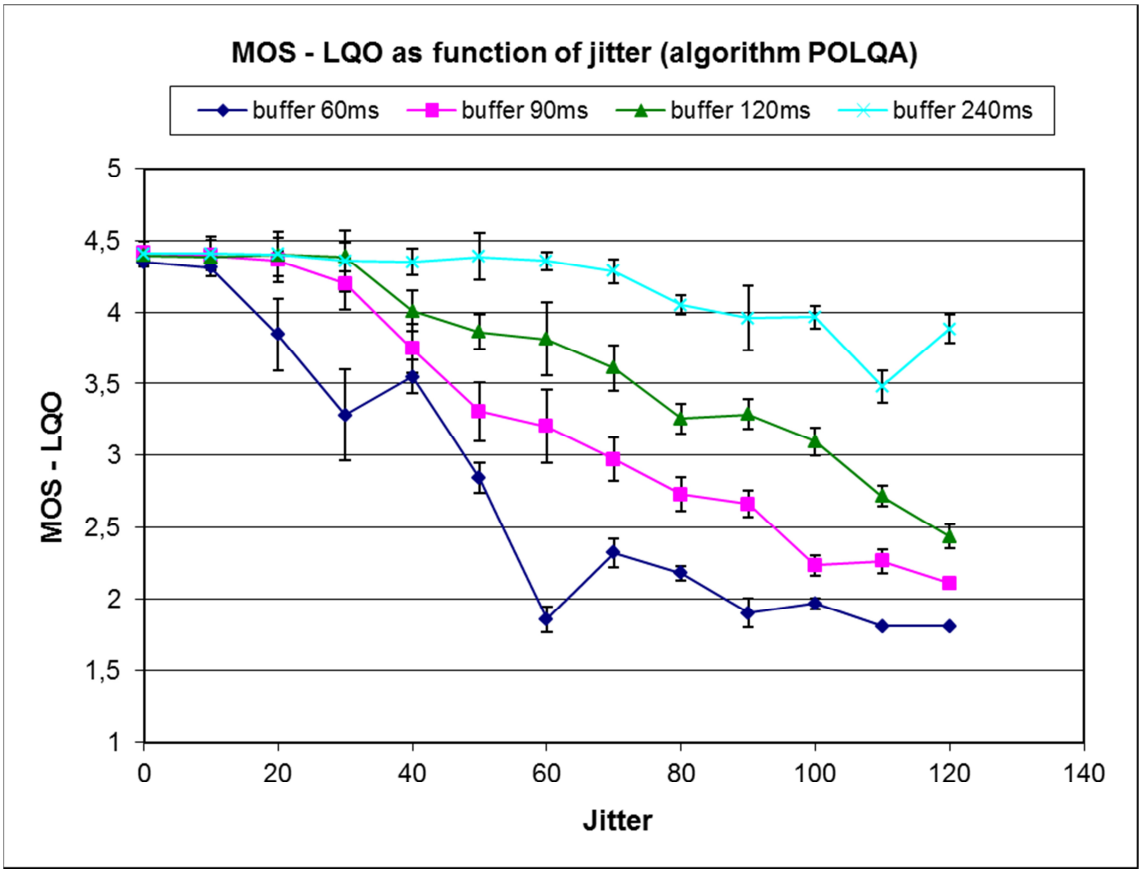


Fig. 6.15 MOS - LQO as function of jitter, comparison of buffer length 60ms, 90ms, 120ms and 240ms (POLQA)

The results confirm the assumption that the longer the buffer is, the better it compensates the differences in packet delay. It is important that the growth of effect is not linear. Buffer shorter than 60ms is unreliable and cannot compensate the relatively small variation in delay (Fig. 6.12, 6.14). Increasing the length of the buffer to 90 and 120ms provides the fastest improvment of effect throughout the adjusted range of jitter.

Probably the biggest difference is between the buffer lengths 60 and 90ms. The increase of 30ms here provides significantly improved efficiency (Fig. 21, 23).

For future experiments, it is important that the buffer of 90ms has significantly better results than Linphone's default 60ms (Fig. 6.13, 6.15) while the increase of the delay is still imperceptible to the listener. Another extension of buffer (180ms, 240ms) does not have such a dramatic effect, and the delays have a negative impact on the quality of conversation.

Compared to PESQ algorithm (Fig. 6.16) POLQA predicts (Fig. 6.17) usually a few tenths higher MOS.

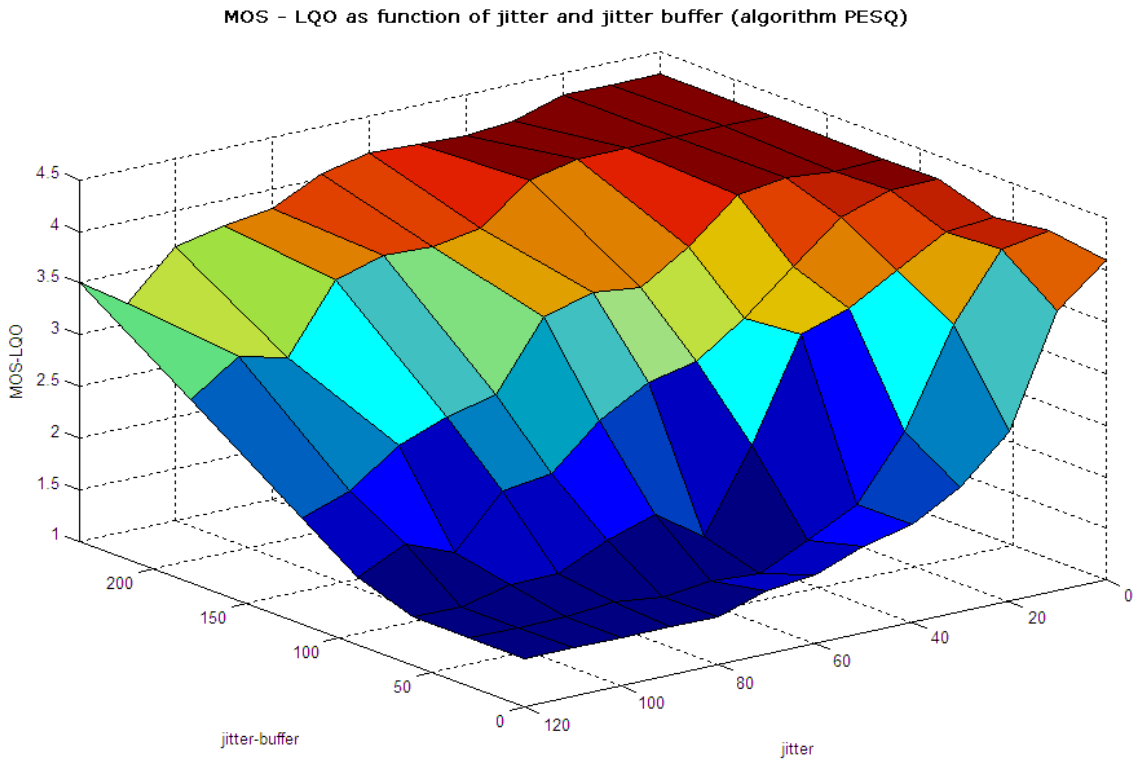


Fig. 6.16 MOS - LQO as function of jitter and jitter buffer (algorithm PESQ)

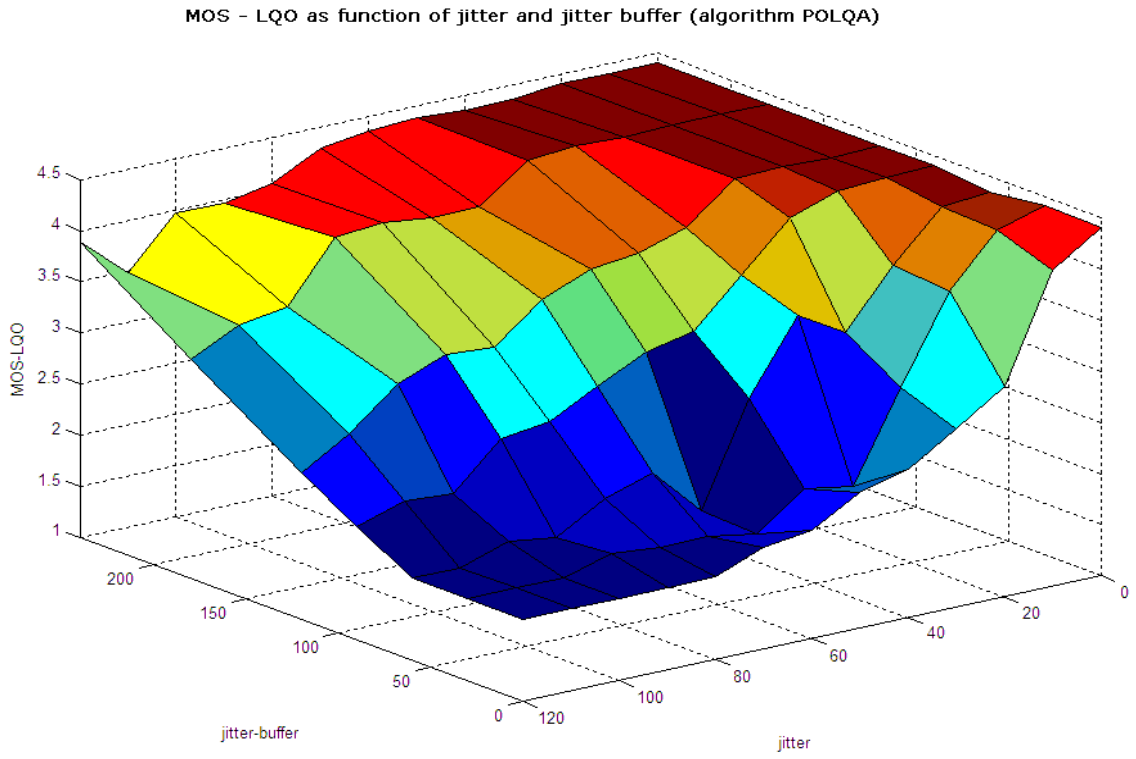


Fig. 6.17 MOS - LQO as function of jitter and jitter buffer (algorithm POLQA)

6.2.4 Conclusion

This experiment verifies the impact of jitter and jitter buffer on the quality of voice transmission in IP networks. It also compares the results of new testing standard POLQA with the older algorithm.

The main objective was to identify the relation between the ratio of jitter/jitter buffer and MOS values as delivered by different objective algorithms. In the case of the same jitter, longer buffer means better transmission, but the dependence is not linear.

The second objective was to explore MOS predictions of new testing standard POLQA. This algorithm predicts higher MOS value for most samples.

6.3 Effect of sample length on the objective quality evaluation of the transferred speech

This experiment investigates the effect of sample length on the outcome of the POLQA algorithm in evaluating the quality of transmission affected by packet loss. The NISTNet emulator is used for adjusting the IP channel network. The transmission quality criterion is an MOS parameter investigated using the ITU-T P.863 POLQA

6.3.1 Introduction

One of the parameters which most affect the quality of transmission is packet loss. The root cause of packet loss during transmission may be a route failure (drop-out of the satellite or microwave links) or saturation of the router buffer. Sometimes, the packet is not used in the reconstruction of the signal, due to its excessive delay. Losses may be dependent (the probability of packet loss depends on whether the previous packet was lost) or independent. A suitably long speech sample with a high speech activity factor should be used in the case of independent losses, in order to assure uniform distribution of impairments in different measured samples. However, the sample length is limited by the requirements of the recommendation.

6.3.2 Time structure and the length of test signals for POLQA

According to the recommendations set out in P.863 [6] and POLQA application guide [40] the test signal should contain speech sections separated by intervals of silence lasting 1 - 2 seconds. The length of a simple sentence is usually from 1 s to 3 s and varies by language. For the purposes of the POLQA algorithm each sample should contain at least 3 seconds of active speech. Most of the experiments used in calibrating and validating POLQA contained two sentences separated by silence interval, totaling 8 s in duration. In some cases, there were three or four sentences in a longer sample (up to 12 s). The maximum sample length recommended for use with POLQA is 20 s. In the application guide it is also noted that due to the nonlinear averaging the average of the results of short samples usually does not coincide with the result of a concatenated sample folded of them.

6.3.3 Experiment description and test signals

Test-bed

The test bed was very similar to the previous experiment (Fig. 6.1). It consisted of three computers and audio Analyser Opera, which was used for playback and recording of samples. Computers were arranged in two Ethernet networks with a contact point in PC2 which was equipped with two network cards. The NISTNet emulator [30] was running on PC2, which (according to the specific

settings) introduced transmission errors between PC 1 and PC 3. The results depend on the accuracy and repeatability of the network simulation. We proved by several experiments that NISTNet suits these requirements satisfactorily. The test signal was led by audio cable from the Opera to the audio input (line in) of PC1. A VoIP call was realized between PC1 and PC3 a using the softphone Faramphone and PCMA codec. From the output of PC3, the signal was led with the audio cable back into the Opera Analyser.

Independent losses were used in this experiment, and the following values were adjusted (in %) 0, 2, 4, 6, 8, 10, 12.

Test signals used in this experiment

In this experiment, samples in length of 4, 8, 12, 16 and 20s composed of six different sentences listed in Table 6.1. were used. The structure of test samples is listed in Table 6.2.

Table 6.1 List of used sentences

Sentence number	Speaker	Language
s1	Female	Chinese
s2	Male	Chinese
s3	Female	Czech
s4	Male	Czech
s5	Female	German
s6	Male	German

Table 6.2 Structure of test samples

Sample	Structure
4.1	S1
4.2	S2
4.3	S3
4.4	S4
4.5	S5
4.6	S6
8.1	S4 + S5
8.2	S6 + S2
8.3	S3 + S1
12.1	S4 + S6 + S3
12.2	S5 + S2 + S1
16.1	S2 + S3 + S4 + S5
20.1	S1 + S3 + S5 + S4 + S6

Each 4s sample was measured twice, 8s four times, 12s six times, 16 and 20s twelve times. This resulted in an equal number of twelve repetitions for each sample length and condition.

Processing of results

Recorded samples were converted from stereo to mono in Adobe Audition and then evaluated by algorithm POLQA. The MOS value for a given condition and the length of the sample is the arithmetic mean of twelve repetitions. Besides the MOS values, the uncertainties type A was evaluated for each sample length and test condition. It is displayed in graphs using error bars.

6.3.4 Results

Effect of packet loss

As expected, higher packet loss means a worse transmission quality. The results in terms of the effect of packet loss correspond to previous experiments (Fig. 6.18). The experiment of 2015 (yet unpublished) was measured in identical test environment and using the same software, using the concatenated sample in the German language in the length of 15s. A previous experiment in 2011 (chapter 6.1) was measured in a similar test environment, using the softphone Linphone and using the concatenated sample in the Czech language, in the length of 17s. Due to the length of the samples used in the previous experiments, they are compared in Fig. 26 with new measured results for the length 12s and 16s.

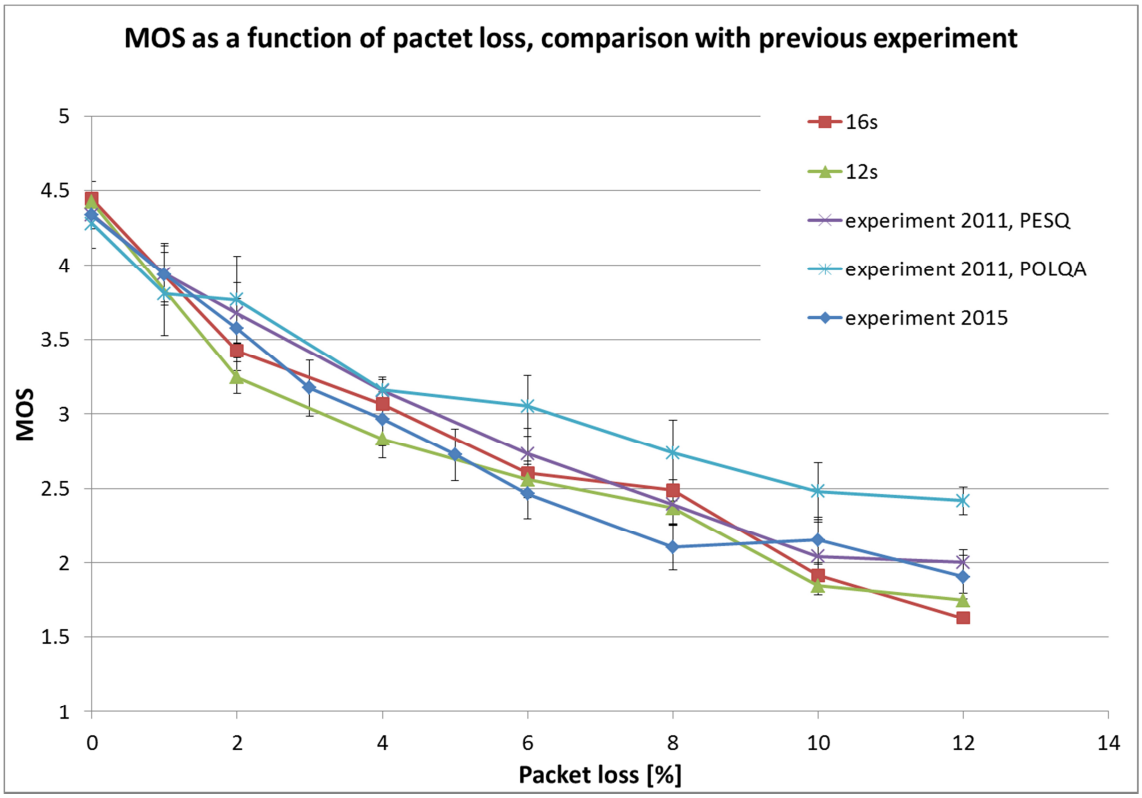


Fig. 6.18 MOS as a function of packet loss, comparison with the previous experiment

Effect of sample length

Authors of POLQA application guide state that the largest part of the calibrations was performed on samples of length 8s. For this reason, the result of this length will be taken as a reference for other measurements.

Results for all the length of the samples are shown in the graph (Fig. 6.19). The figure shows that the results for the lengths 12 s and 16 s are very close to a reference (average deviation of 0.15 (5.7%) and 0.11 (4.8%)). For most conditions, it lies within the confidence intervals.

In contrast, the results for the samples of length 4 s, and 20 s are significantly different from the reference (average deviation of 0.31 (11%) and 0.37 (17%)), and for most of the conditions lie outside the confidence intervals. These differences are clearly visible in Fig. 6.20 and Fig. 6.21.

The POLQA algorithm first divides the measured sample into short periods of time, and after that for each section separately, it evaluates the difference of the transferred sample from the original sample. These partial differences are added up and averaged, but not linearly.

For a sample of length 4s, there may be too little data (adds up to a few differences) and also measured condition may not occur in such a short sample, which especially applies to lower loss rates. Both of these lead to the fact that POLQA evaluates transmission as better when the error is objectively the same. The issue of an appropriate sample length for measuring the impact of packet loss is further discussed in chapter 7.2.

Conversely, at the upper limit of the recommended length, it is possible that it adds up too many sub-differences and despite the averaging, the resulting rating is lower than the result of reference length for the same error settings.

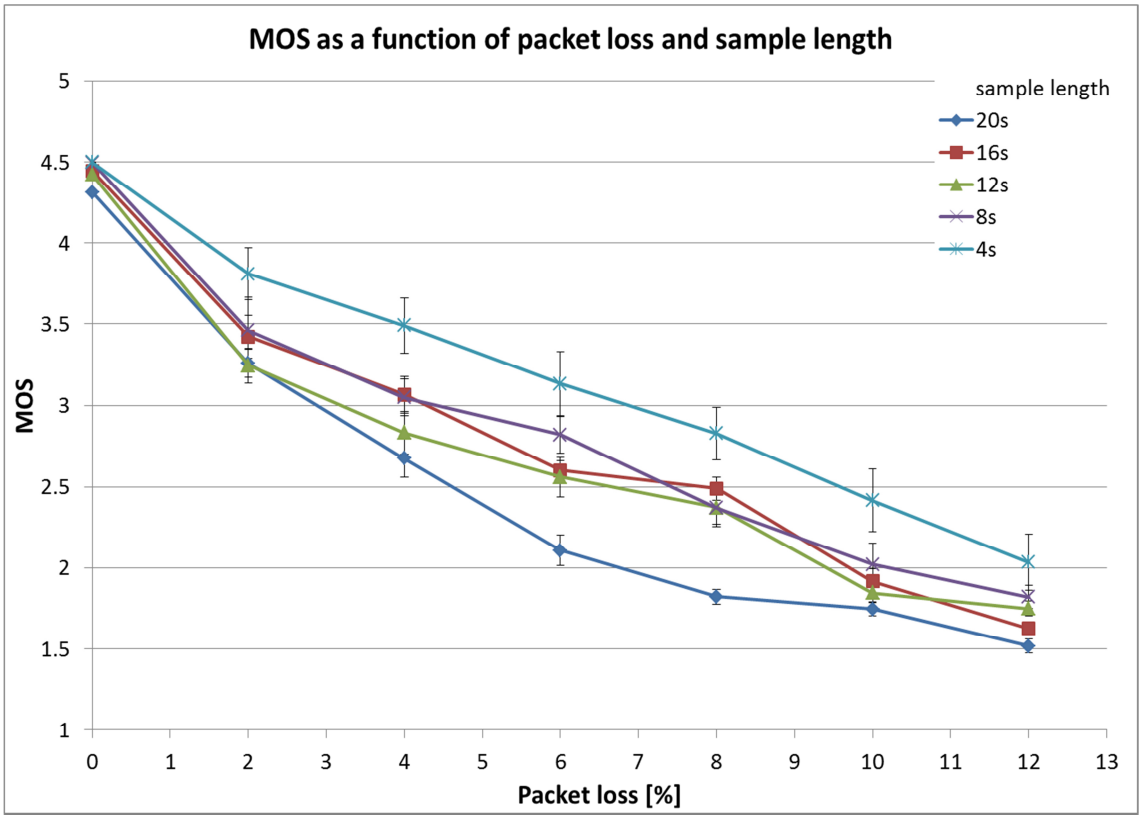


Fig. 6.19 MOS as a function of packet loss and sample length

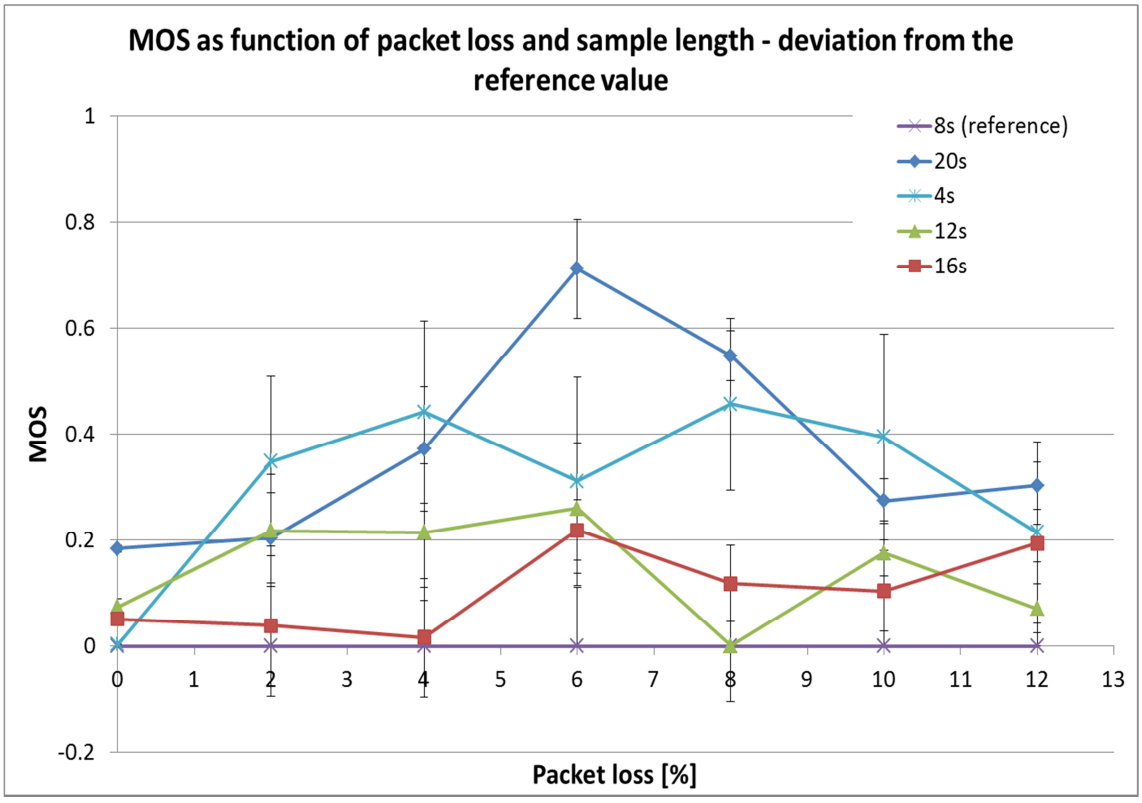


Fig. 6.20 MOS as function of packet loss and sample length - deviation from reference value

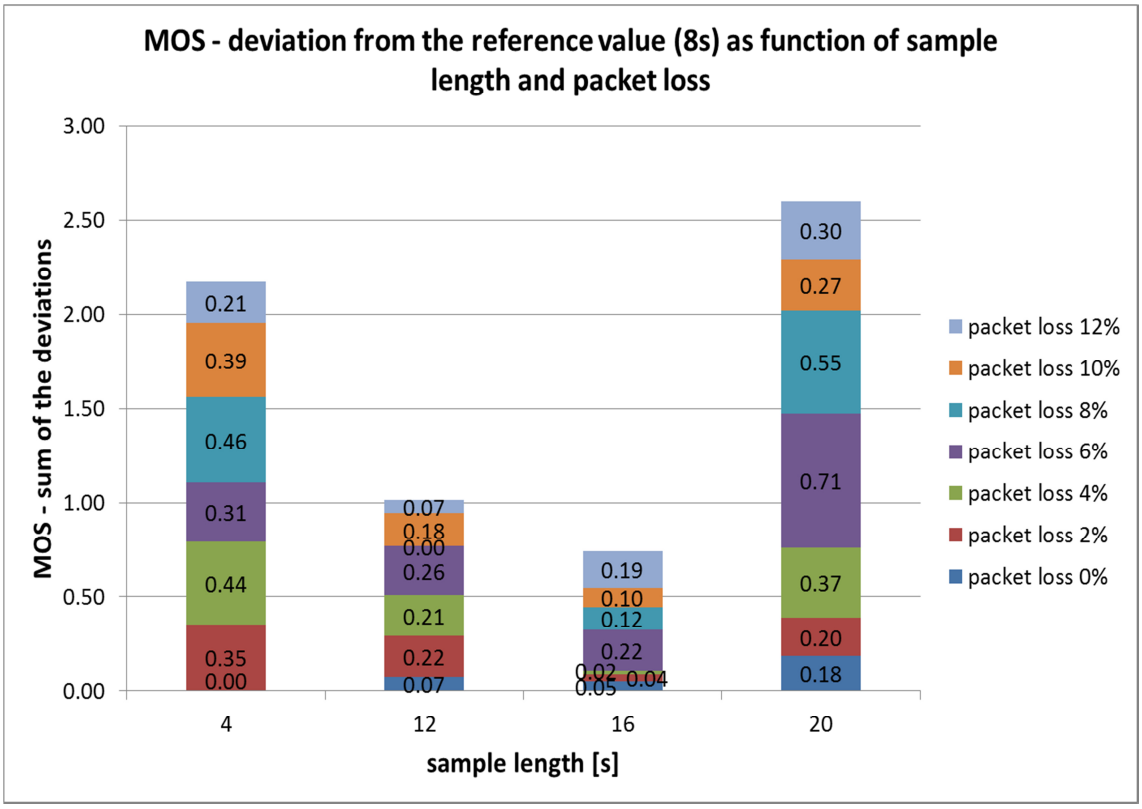


Fig. 6.21 MOS - deviation from reference value (8s) as function of sample length and packet loss

6.3.5 Conclusion

This experiment has verified the impact of packet loss on the quality of voice transmission in IP networks and compared the results of the POLQA algorithm depending on the length of used sample.

The first objective was to verify the effect of packet loss and compare it with the results of previous experiments. The differences are within the bounds of statistical error and can be caused by different hardware and software equipment.

The second and main objective was to identify the relation between the MOS value delivered by POLQA algorithm and length of a voice sample, used in the experiment. Within the interval permitted by the recommendation, algorithm results are very stable, only at the very borders, they exhibit statistically significant deviations.

6.4 Impact of codec and different methods of QoS on the final quality of the transferred voice in an IP network

This experiment deals with the analysis of the relation between used codec, QoS method, and final voice transmission quality. For adjusting of QoS, the Cisco 2811 router is used. For adjusting of codec, VoIP client Linphone is used. Criterion of transmission quality is a MOS parameter investigated with an algorithm ITU-T P.862 PESQ and P.863 POLQA

6.4.1 Introduction

For voice transmission, it is important that voice packets are delivered with minimum delay, with the variability of delay, and in the correct order. To some extent, we can accept the loss or dropping of some packets at the cost of low delay and increase in fluency. Different requirements on transmission are bringing the need to solve the simultaneous transmission of voice and data (or video) over a single network. The simplest way to solve problems is to increase the network bandwidth so that all protocols have sufficient capacity. But this method is expensive and not always realizable (because of various technological limits). Another possibility is the introduction of QoS classes and prioritization of network traffic according to their importance. Typical classes are:

Voice - Class for voice transmission.

Business-Critical –Class for important data (business applications, access to databases, etc.)

Best-Effort - Class for normal traffic (e-mail, web access, etc.)

Scavenger - Class of unwanted traffic that can also be inhibited

Voice and business-critical traffic should have a sufficient bandwidth and voice packets should additionally be sent as preferred. To ensure such capacity the following mechanisms are used:

Classification - manually configured or automatic classification of packets/frames

Marking - write specific values in the header of the packet/frame

Congestion Management - uses header tags in the decision in which class/queue the packet will be included

Congestion Avoidance - uses preventive dropping of packets to prevent congestion of the line

Policing and shaping - uses delaying or dropping of packets in traffic which reached the defined limit

Link Efficiency Mechanisms - increasing the efficiency of line with header compression and packet fragmentation

6.4.2 Experiment Description

Test-bed

The test bed (Figure 6.22) consisted of two small LAN networks connected with a serial line simulating the WAN connection with bandwidth limited to 384 kbit/s. The following network elements were used: IP phones Cisco 7942G, Switches: Cisco Catalyst 3560, Routers: Cisco 2811, Cisco 1841, PC and servers: Dell.

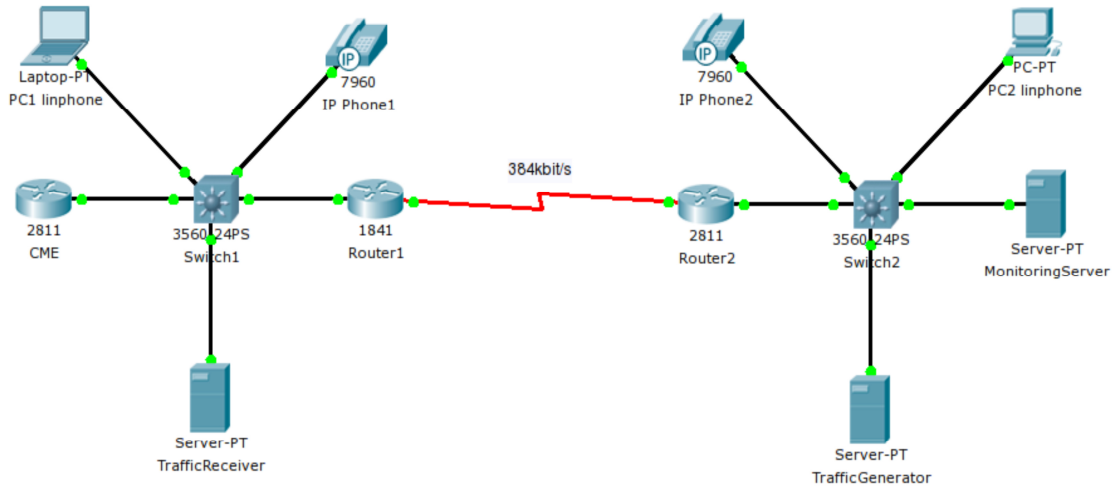


Fig. 6.22 Test-bed

A concatenated speech file in WAV format (8kSa/S, 16bit) of the length 16.75 s has been used. The file contains four short sentences spoken by four different speakers (two men, two women) and covers the entire human speech spectra adequately. Due to this fact, the concatenated file effectively replaces testing using multiple speech samples.

VoIP telephone call was realized between PC1 and PC2 using Linphone. On the side of the caller, the sample was directly played using Linphone. On the receiver side, it was recorded using Audacity. Various methods of QoS have been adjusted on the Router2. One of the purposes of the experiment was to verify the reliability of MOS estimates that are provided in Cisco IP phones. The test calls for each QoS settings were also made using these phones. To accelerate the experiment the dialing and recording was controlled by a bash

script. Softphone Linphone was chosen because it allows playback of a voice sample and script control. It has also been evaluated in previous experiments as the best in terms of quality of transmission [32]. The following codecs were used in Linphone: G.711, G.722, G729, iLBC. The iperf traffic generator was used to simulate the real traffic in the network.

To evaluate the quality of voice transmission, algorithms PESQ and POLQA were used. PESQ algorithm output was recalculated to the value of MOS-LQO (Listening Quality Objective) according to a mathematical prescription defined in ITU-T P.862.1. As follows from the official wording of P.862, the effect of packet loss on CELP coded transmission can be tested by it. It has not been tested for PCM transmissions affected by packet loss but the recommendation itself does not prevent any user from doing that and it was successfully used in previous experiments [31], [38].

Methods of congestion control and prevention

FIFO – First In First Out
The simplest type of queue. Packets are handled in the same order as they come. For larger loads, the queue fills up quickly, and it causes a delay and packet losses.

WFQ – Weighted Fair Queuing
Creates 16-256 queues according to the bandwidth. For the classification of packets to queues using automatic classification.

CBWFQ – Class-based Weighted Fair Queuing
Extending WFQ to support traffic classes. For each class, the queue is created in which all the traffic of this class is routed. The bandwidth allocated to classes by priority.

LLQ – Low-latency Queuing
CBWFQ with the added priority queue for real-time traffic (voice, video). Real-time data is cleared first.

CB-WRED – Class-based Weighted Random Early Detection
RED is a mechanism that randomly drops TCP packets before a queue is full, TCP traffic slows down, and there is no unsteady flooding of lines. Each class can be assigned a different RED profile.

ECN – Explicit Congestion Notification
Extension of WRED which adds information about the network congestion to the packet header.

LFI – Link Fragmentation and Interleaving
The method uses a fragmentation of long data packet, to reduce the delay of voice packets.

6.4.3 Results

For each setting of codec and QoS method, ten samples were measured and processed. Ten repetitions are enough to achieve satisfactorily low results of dispersion and uncertainties. The confidence intervals (CI95) were computed.

The graphs show a significant correlation between the used QoS method and the final voice quality. Typically the more advanced the method is, the better the quality of the transmission. Some methods have the same results in the quality of transmission, but they differ in the length of the delay represented by ping response (Fig. 6.23.). The length of the delay is important for the quality of the conversation.

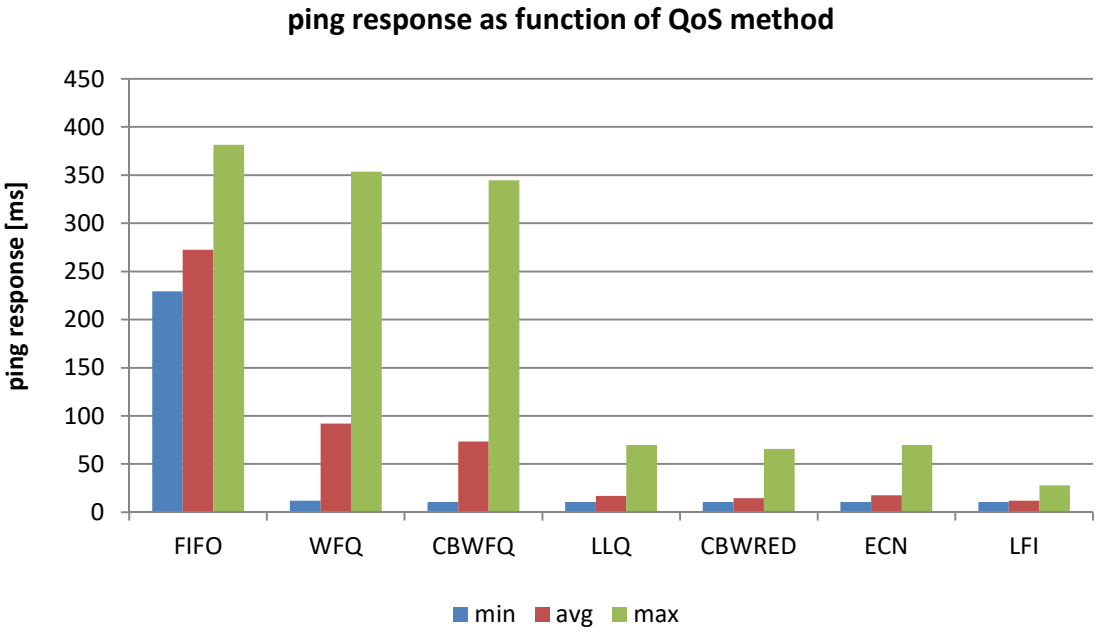


Fig. 6.23 ping response as function of QoS method

The following set of graphs (Fig. 6.24 – 6.27) shows the results for various codecs. Algorithms PESQ and POLQA have very similar results. Estimate of MOS using IP phone is very inaccurate, but it can be used for basic information e.g. if call is possible or not.

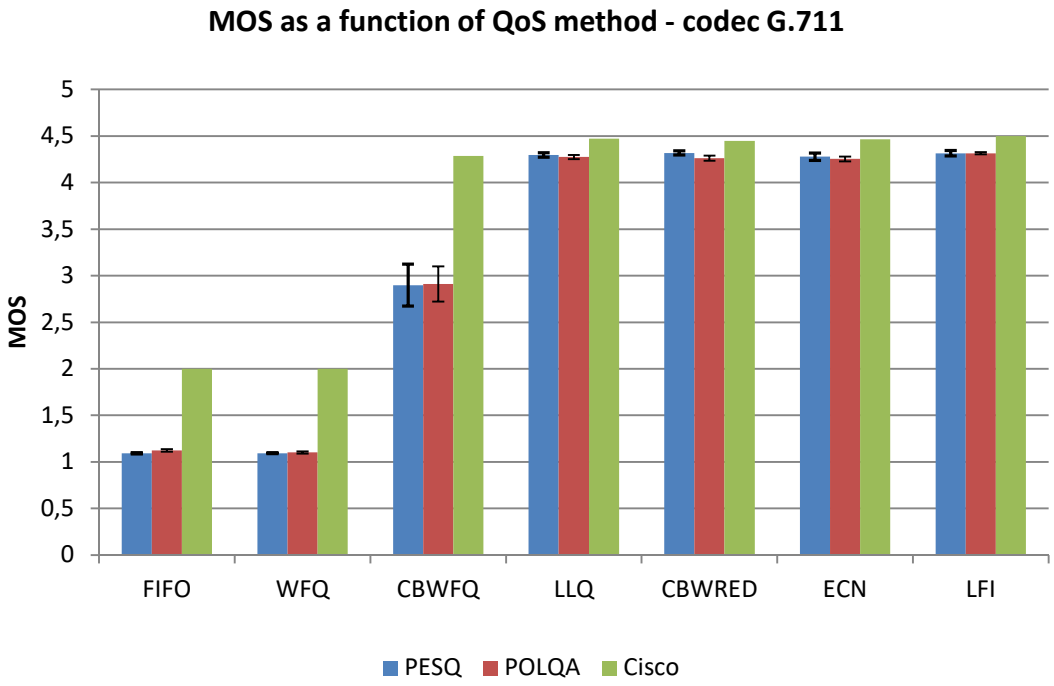


Fig. 6.24 MOS as function of QoS method - codec G.711

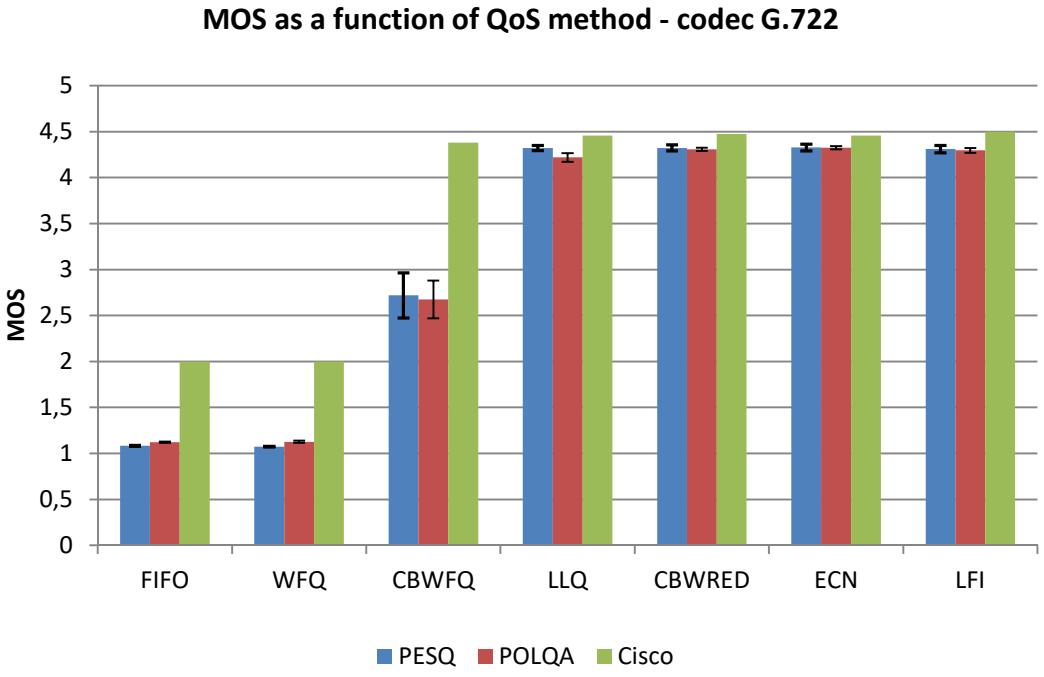


Fig. 6.25 MOS as function of QoS method - codec G.722

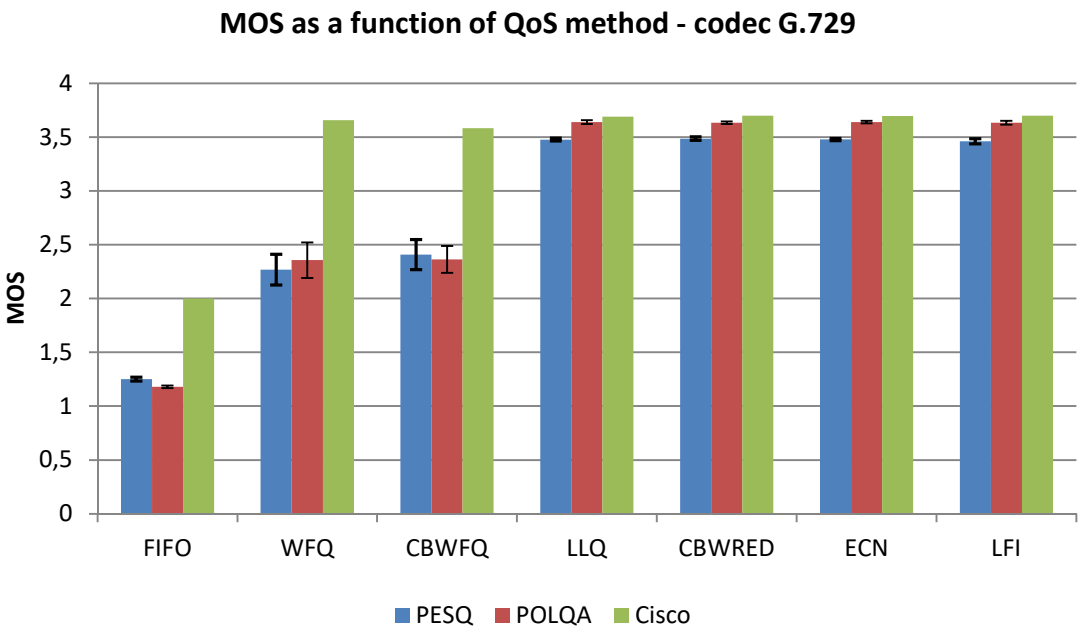


Fig. 6.26 MOS as function of QoS method - codec G.729

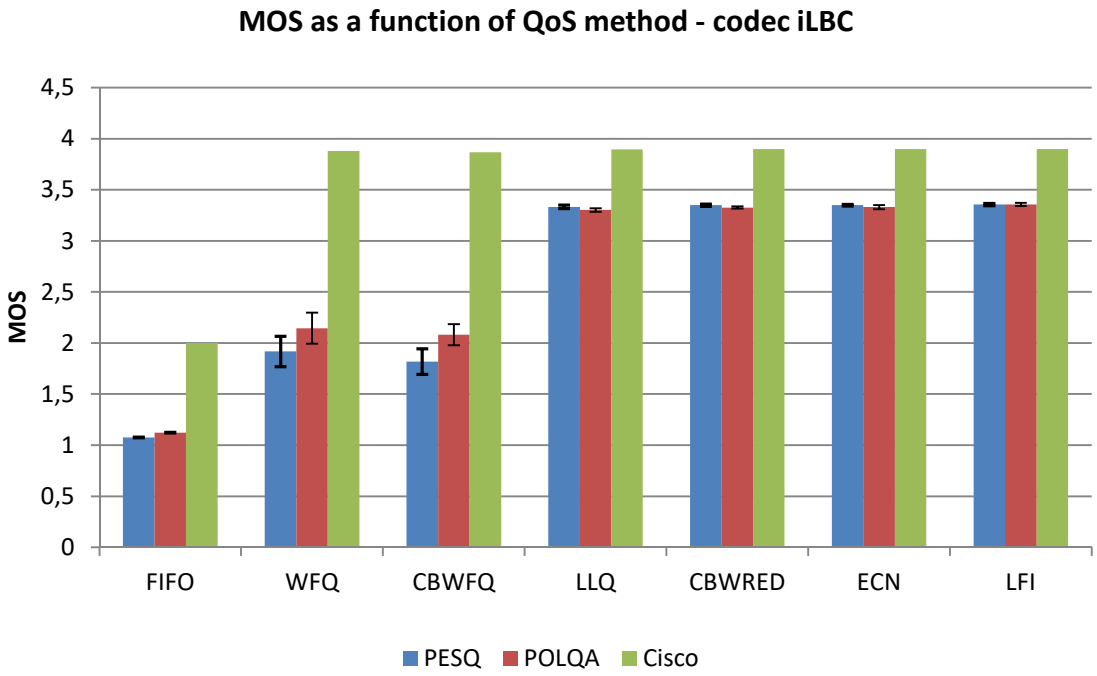


Fig. 6.27 MOS as function of QoS method - codec iLBC

Next set of graphs (Fig. 6.28 – 6.30) shows the results sorted by the algorithm used. The results show that high bitrate codecs G.711 and G.722 usually allow better transmission quality than a low bitrate G.729 and iLBC. But in the case of low bandwidth or worse QoS method, low bitrate codec can provide better results.

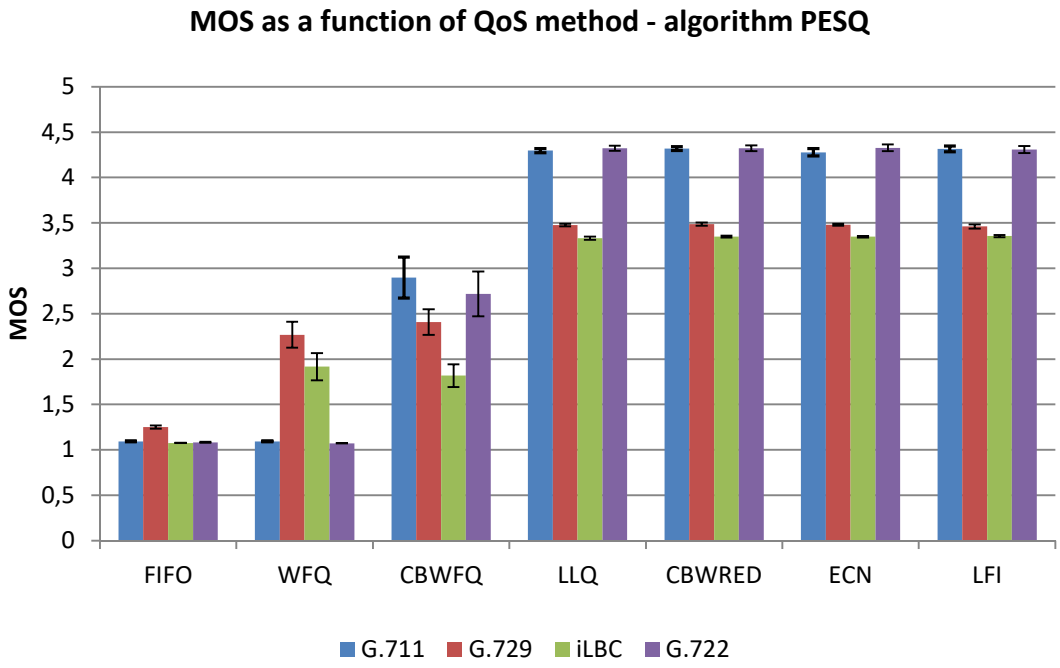


Fig. 6.28 MOS as function of QoS method - algorithm PESQ

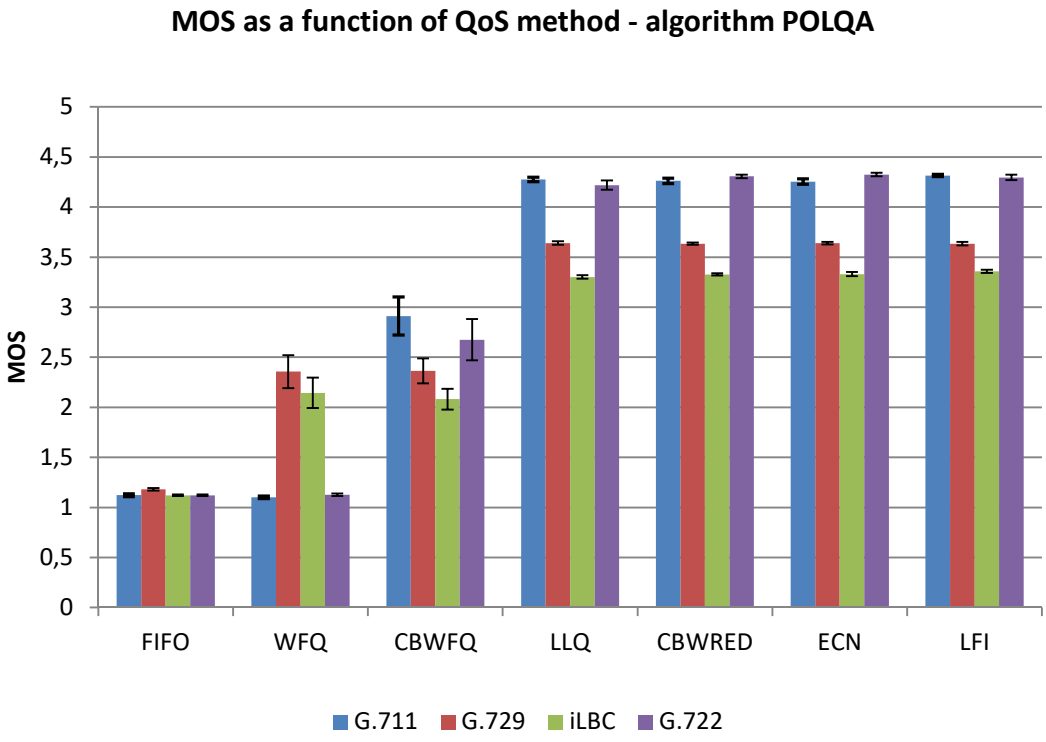


Fig. 6.29 MOS as function of QoS method - algorithm POLQA

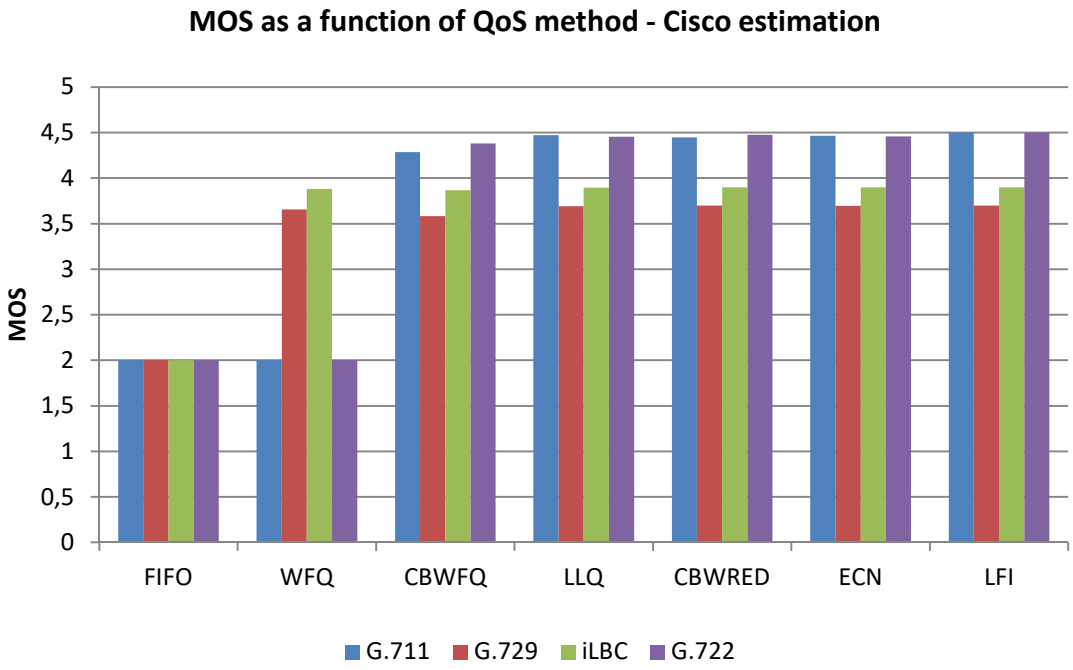


Fig. 6.30 MOS as a function of QoS method – Cisco estimation

The results confirm the assumption that the more advanced QoS method is, the better the quality of the transmission is. Simple FIFO queue is not suitable for the purpose of traffic management in converged networks. WFQ method brings a slight improvement for low bitrate codecs, but the resulting quality is still very bad. In the method, CBWFQ is noticeable improvement also in high bitrate codecs. LLQ method and others already provide transmission quality at the maximum limit for the current codec, but differ in the length of the response, therefore, delay and conversational quality.

6.4.4 Conclusion

This experiment verifies the impact of used codec and QoS method on the quality of voice transmission in IP networks.

The main objective was to identify the relation between the used QoS method and MOS values as delivered by different objective algorithms. It generally applies that the more advanced method is, the better the quality of the transmission is. Some advanced methods differ only in delay.

The second objective was to explore differences between commonly used codecs. Low bitrate codecs have worse results than high bitrate but in the case of low bandwidth, they may be useful.

7 Subjective Tests

The first purpose of these subjective experiments was to test the effect of static delay on the quality of the conversation. Effect of static delay cannot be evaluated using objective tests. The second task was to implement subjective tests using long samples and compare different data collection methods.

7.1 Conversational quality of a conference call in case of a stressed listener

One of the many requirements of the standard for subjective testing in laboratory conditions is that the respondents are at rest with no signs of physical or mental stress. This may cause differences between laboratory tests and the requirements of real users of the telephone network. A certain amount of stress for participants in a telephone conversation can never be ruled out in practice.

7.1.1 Experiment description

The purpose of this experiment was to test the effect of delay of voice information between the participants in a conference telephone call on the subjective assessment of the quality of conversation. Further to test and model the changes in the assessment of individual respondents based on their mental state.

Test-bed

The experiment was conducted in four separate rooms (Fig. 7.1). Due to the need to avoid direct contact between the respondents. One of the rooms fully meets the requirements of the standard (reverberation 182 ms, -60 dB). The second room is particularly suited for the listening tests, meets the requirements of reverberation time <500 ms, the other parameters have not been measured in the room. The third room is undefined in terms of acoustic parameters We can assume that even this room meets the reverberation time <500 ms. The technical background of the experiment, the network simulator, and posts of experiment supervisor are In the fourth room.

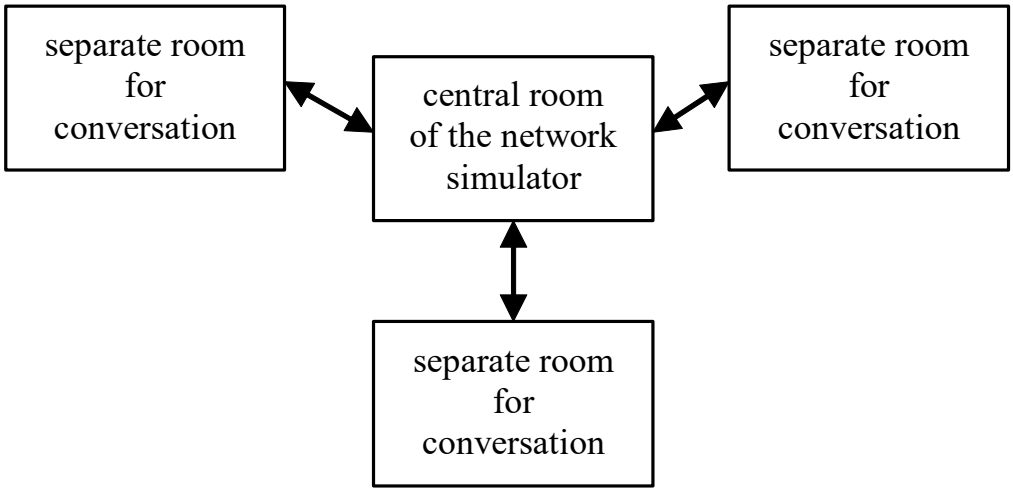


Fig. 7.1 Test-bed

Network simulator

Respondent's posts include telephone chassis with the standard handset. The signal from a handset microphone is preprocessed and routed into the central part of the simulator (Fig. 7.2). In the opposite direction signal is carried to the loudspeaker. The signal from the microphone is preprocessed in SHARK DSP 110 (impedance matching, volume treatment). The central part of the simulator consists of two digital signal processors Behringer Ultra-Drive Pro DCX 2496. The first processor (DCX A) made a filtration with a Butterworth highpass filter 48 th level, 303 Hz and low-pass filter Bessel 24 th level, 3031 Hz. Furthermore, the DCX A set the first part of the variable delay in the range of 1-582 ms.

The processors are connected so that each of the three inputs of DCX B is the sum of the two different analog outputs of DCX A. The second part of the delay is implemented in the DCX B, and output signals are carried into the handsets of the respondents.

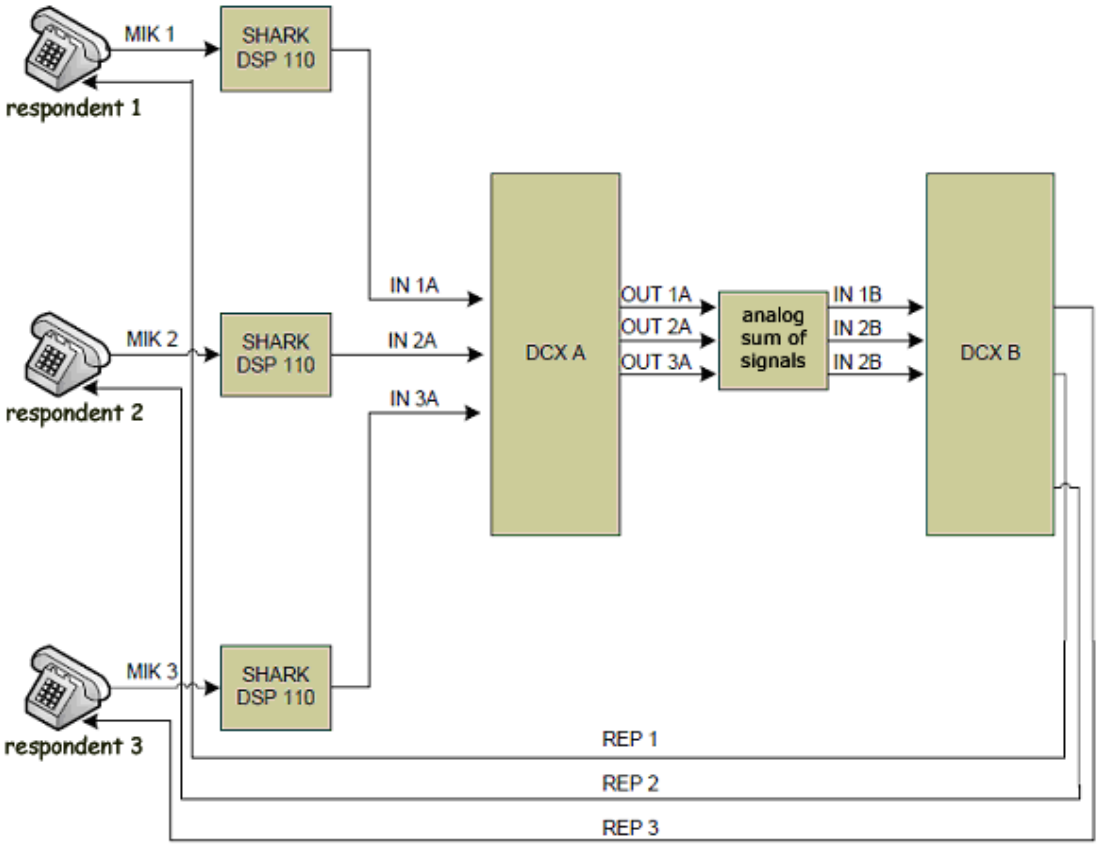


Fig. 7.2 Block diagram of a network simulator

Delay

Delay of a telephone call in an IP network has several different causes. On the side of the speaker, it is particularly encoding, packetization, and controller interface. On the side of listener, it is buffer, depacketization, and decoding. Causes of delays in the IP network itself are the particularly limited speed of signal transmission in the network and signal processing components on the route such as routers and converters. The speed of the signal transmission is a particular problem when the call is made over long distance or part of the route is led via satellite. In this experiment, the following values of delay were adjusted: 62, 337, 612, 887, 1176 ms.

Conversation scripts

For the purposes of the experiment, it was necessary to create the content of the conversation between the three respondents. The proposal is based on theoretical assumptions and normative recommendations for the scenarios of conversation between two participants. The plan assumed the implementation of several independent conversations in the Czech language, where each conversation is realized with the specific setup of delay. For this reason, it is necessary to ensure repeatability of scenarios. Respondents must be still

interested in the content of the conversation and must not show signs of tiredness or disinterest. High dynamics of the conversation (frequent changes of speakers) is necessary for the best assessment of the impact of the delay.

Variantion of freely defined scenarios were used in the experiment. The script does prescribe the exact content of the conversation, it only specifies the story and casts the respondents into the roles. Scenarios are chosen from ordinary conversation topics (planning of the weekend, buying gifts etc.). Each participant in the scenario has defined capabilities and requirements. The purpose of the conversation is to reach a compromise. Conversations are repeatable with changing the scenarios or changing roles of the respondents. The dynamics of the conversation is very high due to the open structure of the scenario, that resembles a common call.

Stress generating game

The stress generating game was created by a computer application (game) (Fig. 7.3). The screen displayed 10x10 matrix with numbers from 0 to 99, which is not repeated. The searched number was displayed besides the matrix. The task of the respondent was to quickly locate the number and identify it with the mouse click. The initial limit was 20 s from the 4th number then the average of the last three results. Unless the number was found, the result was counted 20 s. For each found number, respondents received a score inversely proportional to the length of the search. A total of 75 numbers was searched without repetition. The test was accompanied by sound effects (ticking clock, signaling of success/ failure etc.). The respondent was motivated to win prizes for outperforming the others. The average test time was 13 min.

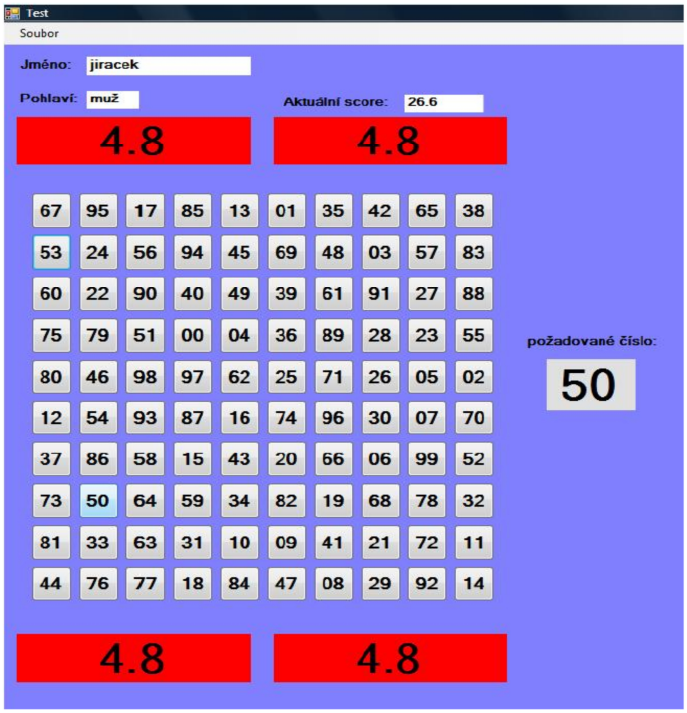


Fig. 7.3 Test application

Course of the experiment

Due to the capacity of the workplace and the design of the experiment, it was not possible to invite a large group of respondents. One session usually consisted of 6 respondents in two groups of 3. The group of respondents performed the prescribed scenario of the conversation. After two minutes, they were given permission to end the conversation and the respondents evaluated the influence of delays on the quality of the conversation in a prepared form. Delay values were adjusted in random order. After series of five conversations, respondents were asked to play a stress generating game on the computer. Immediately after the game the second series of five conversations followed.

A total of 56 respondents participated in the experiment. For technical reasons, it was possible to use data from 52 respondents.

7.1.2 Results

A complete set of ratings from 52 respondents was statistically processed and values of MOS - CQS including confidence intervals (95 %) were calculated. Furthermore, the results were approximated by estimation function and the degree of correlation of this function with the subjective data was calculated. The approximation was carried out using least squares in Microsoft Excel 2010. The real third-order polynomial was used as an estimation function.

Table 7.1 Summary of all results

	summary of all results	
delay value [ms]	MOS_{CQS}	confidence interval
62	4.47	0.14
337	4.31	0.14
612	3.85	0.18
887	3.71	0.21
1176	2.92	0.22

Table 7.2 Results without stress

	results without stress	
delay value [ms]	MOS_{CQS}	confidence interval
62	4,46	0,21
337	4,38	0,20
612	3,79	0,26
887	3,65	0,31
1176	2,62	0,28

Table 7.3 Results with stress

	results with stress	
delay value [ms]	MOS_{CQS}	confidence interval
62	4,48	0,19
337	4,23	0,20
612	3,90	0,26
887	3,77	0,29
1176	3,23	0,32

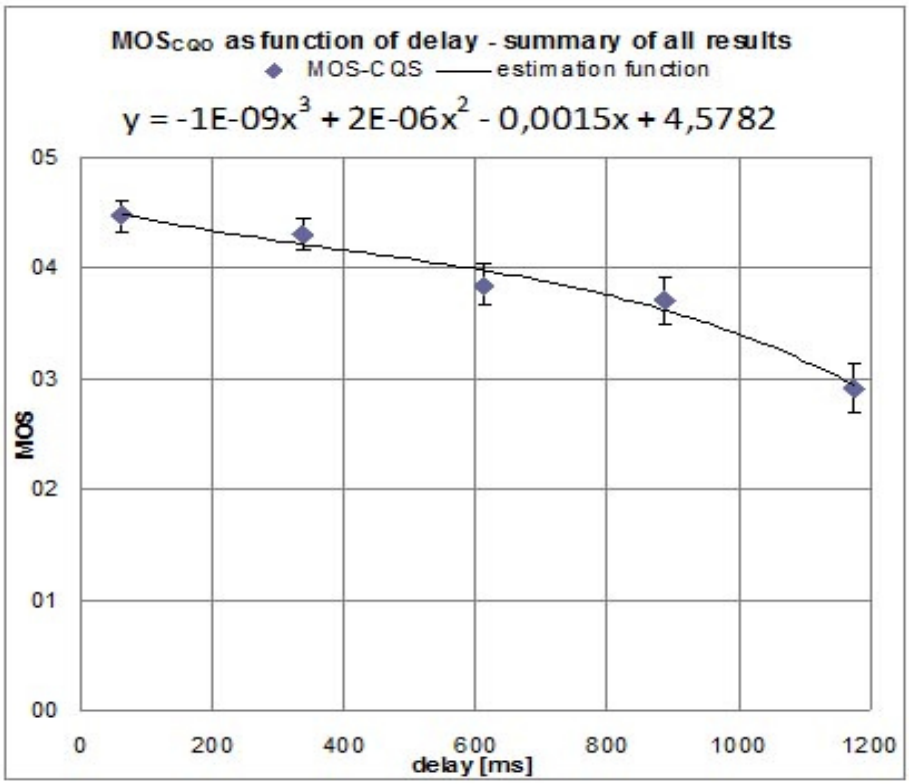


Fig. 7.4 Summary of all results

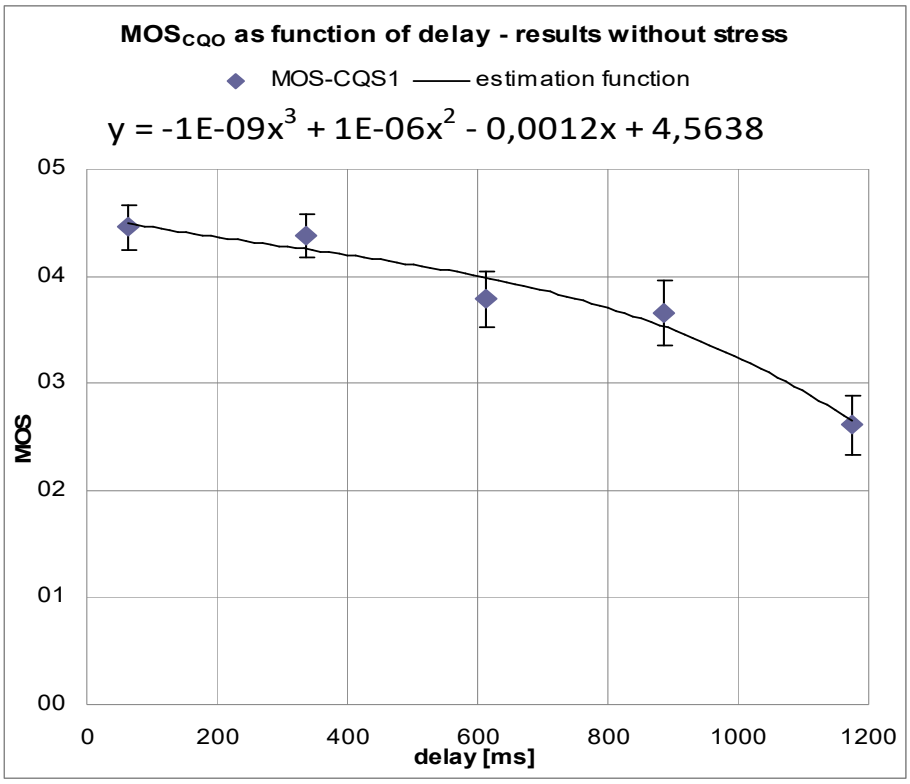


Fig. 7.5 Results without stress

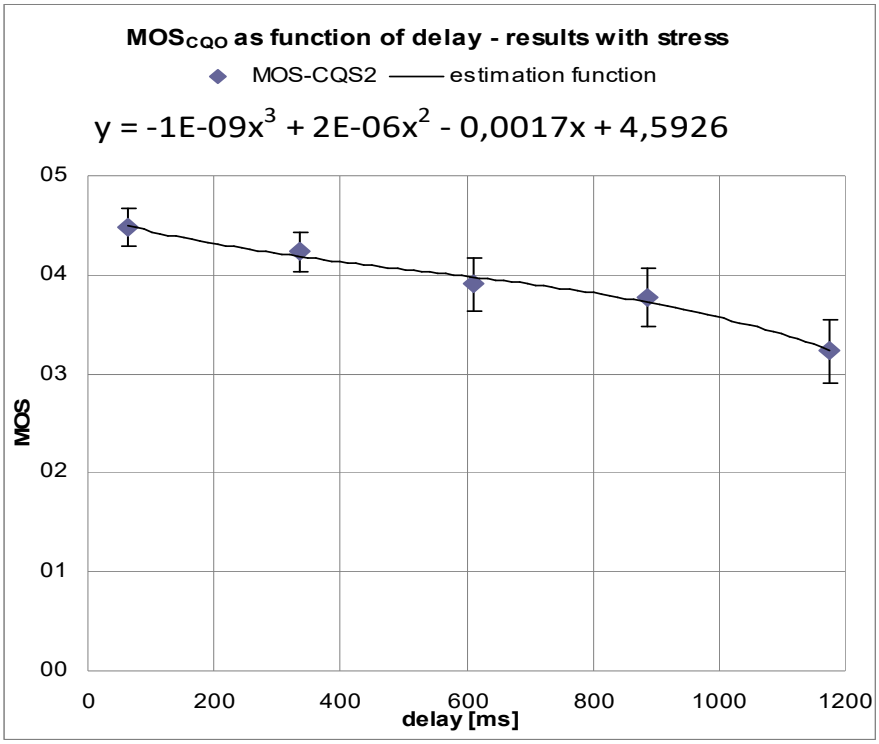


Fig. 7.6 Results with stress

Commonly used methods were selected for comparison of estimation function and subjective data. Correlation method and the RMSE are the most frequent. Their disadvantage is that do not take into consideration the reliability of subjective data, the confidence interval. This defect eliminates the method RMSE *.

Table 7.4 Correlation of estimating function with subjective results

	Camparison method			
Compared data	correlation	RMSE	RMSE*	Method MAX
No stres subjective results Universal estimating function	0.980	0.191	0.029	0.341
No stres subjective results No stres estimating function	0.983	0.123	0.000	0.197
Stres subjective results Universal estimating function	0.990	0.143	0.000	0.275
Stres subjective results Stres estimating function	0.994	0.045	0.000	0.072

Table 7.4 shows that the modified "stress" estimation function correlates with data after the stress better than a general purpose estimation function.

Interpretation of results

For each value of delay, difference rating was calculated before and after exercise. Resulting differences are shown in Graf (Fig. 7.7) including confidence intervals.

The results show that the impact load increases with increasing delay. For the 1176 ms it is after netting confidence interval is proven non-zero. The experiment was realized as three participants' conference call. In order to assess the impact of this, it was decided to compare these results to previously published test [41] in the Czech language with similar scenarios, however, between pairs of participants. The results of both experiments are in Fig. 7.8.

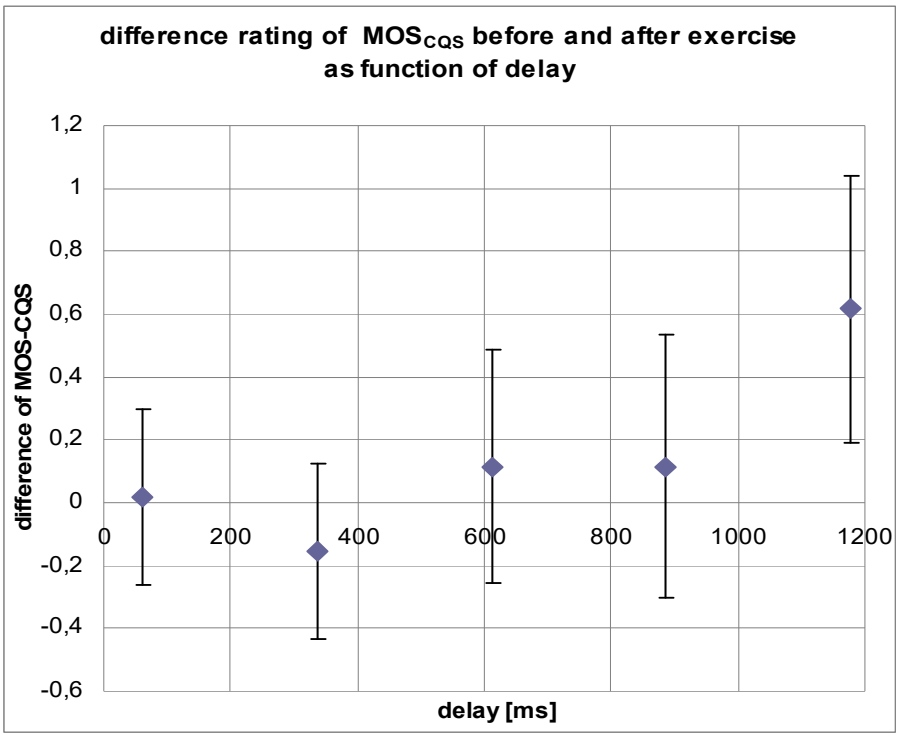


Fig. 7.7 Difference rating of MOS CQS before and after exercise as function of delay

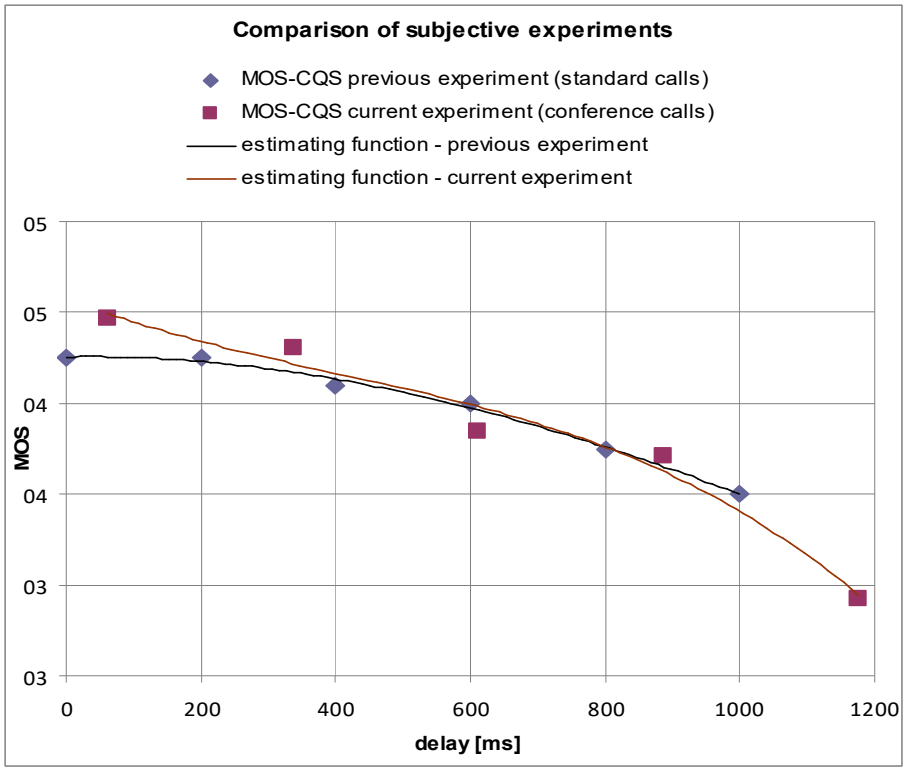


Fig. 7.8 Comparison of subjective experiments

The graph shows that in the range of delays 400 - 1000 ms, the results of both experiments are nearly identical. In the range of 0 - 400 ms have the results of an experiment with conference call have decreasing character as opposed to a constant in a second experiment. This may be due to greater dynamics of the conversation in a conference call where participants recognize the lower delays.

7.1.3 Conclusion

This experiment verifies the effect of the delay of voice information between the participants in a conference telephone call on the subjective assessment of the quality of the conversation. It also evaluates changes in the assessment of individual respondents based on their mental state.

Based on the results of the experiment, it has been shown that psychological stress has an impact on the respondent's rating of the quality of the conversation.

Influence of the fact that the test was organized as a conference with three parties has not been established. The difference between experiments with two and three participants in the conversation are very small and can be caused by large confidence intervals, time distance, or unspecified differences between the experiments.

7.2 Using long samples in subjective testing of voice transmission quality in IP network

This experiment deals with problems of speech transmission quality measurement in TCP/IP networks. It focuses on problems caused by the inaccuracy of current methods for measuring the quality degradation caused by packet loss and jitter. The proposed solution is to use longer samples for listening tests. This experiment compares three evaluation methods for listening tests.

7.2.1 Introduction

In recent years, there has been a rapid development of technologies in telecommunications. Users require faster and more comfortable transfer of voice, data and multimedia content, in larger volume and higher quality. Along with the development of transmission technologies, methods for measurement and evaluation of quality are also developed. The Very actual topic is subjective and objective (algorithmic) measurement of transmitted speech, video or multimedia quality in (common) case of transmission channel errors including packet loss and jitter delay for packet-based transmissions. The results of the measurement are influenced by many factors such as language, type of samples used, or encoders used in transmission. The main problem of present approaches in speech transmission quality measurement is that they have a basis in analog telephony and they have difficulty in reflecting modern trends which lead to new type of distortions in signal and even more dramatic change in listeners perception. Additional the technical aspect of communication is also changing perceptions and expectations of users. Expectations grow with experience of listeners and mass expansion of technology in the population. Freeware programs like Skype, Linphone, Viber, and many others are very popular among the internet users and, using a cell phones data connection, they often replace standard services even in mobile networks. Expansion of these programs allows their creators to develop and implement new modern coders and compression formats much more quickly than common telephone network or mobile operators. It is inadvertently changing expectations of listeners and their idea of quality scale. One of the consequences is that respondents in research made in [42] often evaluated mp3 as the higher quality record then lossless formats.

7.2.2 Conventional speech transmission quality measurement

Subjective speech transmission quality measurement is standardized in ITU-T Recommendation P.800 [5]. A measure of voice transmission quality is parameter MOS (Mean Opinion Score), it is a five point scale where 1 is the worst and 5 the best. The recommendation also defines testing chambers, its

reverberation, the length of session and used samples. There are two types of subjective tests, listening and conversational. In the listening test, the subjects are listening to prepared sound samples and rate the quality of it. In conversational test, subjects are performing a phone call thru the channel with defined parameters and rate the quality of transferred voice and comfort of conversation.

According to this recommendation, the speech material should consist of simple, meaningful, short sentences, chosen at random as being easy to understand (from current non-technical literature or newspapers, for example). Every sample then should then contain between two and five sentences, each from two to three seconds in length. The recommendation P.800 also defines that at least 100 votes per condition are required to obtain high precision and statistically significant results. It is impractical and difficult to have 100 listeners in one test. More frequented is the method with multiple samples for the same condition.

Another way to obtain MOS values is objective testing using computer algorithm.

7.2.3 Specific problems with TCP/IP networks

Nowadays more and more calls are transferred using data and TCP/IP networks. Compared to the original analog telephone network, these networks are relatively reliable, and in normal conditions errors are rare. Typical problems of IP network are delay, packet loss and jitter of delay. For example, in the case of packet loss, the error rate should not exceed few percent. For a typical value of 2% packet loss, it represents 160ms from 8s sample, but not necessarily continuous. Most codecs carry from one to three voice packets (20ms each) in one TCP/IP packet. The loss may involve three or four parts of the sample. The lost packets may contain silent breaks between words or even a word that will not change the meaning of the sentence. Resulting MOS of such a sample, obtained with the objective algorithm, will be below the actual quality, compared to subjective tests.

In the case of 2% packet loss, the objective MOS varies from 3.1 to 4.0, it causes that the A type uncertainty of this test is very high. The way to eliminate this effect, used in current research, is to transfer every sample multiple times and use the average of these acquired values as the output speech transmission quality. In [43] researchers transfer every sample ten times, and for every condition, they use same speech sample. Unfortunately, this method cannot be used in subjective tests. If we played the same sample multiple times, listeners will be affected by the repetitiveness of the speech material and wouldn’t be able to evaluate quality properly.

7.2.4 New methodology using long samples

We assumed that samples long approx. 80 - 120 seconds would allow us to decrease the time needed for subjective tests. In comparison with multiple samples, we can run only one test session, and it will better reflect the conditions of real life. Samples are long enough that even lower frequency errors are always evident and perceptible to listeners.

The length of samples recommended by P.800 has its meaning. Short samples are not affected by so-called recency effect, which causes that listener rates only the end of the sample and neglects the previous parts [44], [45]. In order to eliminate this effect we are trying two approaches in order to gain reliable results:

- Continuous evaluation based on beep marks – samples are enhanced with beep marks placed equidistantly in sample every 10 seconds
- Continuous evaluation with custom measurement application – evaluation is realized with the help of custom application and the time of change in score is fully in the hands of subject. Similar method is used in [46] for video quality assessment.

The advantage of these methods is that we gain multiple rates for one sample. This allows advanced processing of results like, for example, weighting the values according to their position in the sample.

7.2.5 Realization of tests

Tests were carried out in an anechoic room. Three sessions were held, each was attended by eight listeners. Following conditions were examined: packet loss in values of 1, 2, 3, 4, 5, 6, 8 and 10 % and jitter of delay in values of 10, 20, 30, 40, 50, 60, 80 and 100 ms plus one original and one clearly transferred sample. This means a total of 18 conditions. For every condition, a different, 90 seconds long sample from one of four speakers (two male and two female) was used. Table 7.5 shows the list of used samples and conditions. Samples were transferred with softphone Linphone using PCMa codec, played and recorded on the Opera audio analyser by OPTICOM. Required distortion was added into the samples using network emulator NISTnet. Besides the two methods mentioned above, the third method was used with one rate on the end of the sample, similar to the original P.800 recommendation, but using the same long sample.

Table 7.5 List of used samples

Sample number	condition	speaker
1	Clear	Male 1
2	Packet loss 1%	Female 1
3	Packet loss 2%	Male 2
4	Packet loss 3%	Female 2
5	Packet loss 4%	Male 1
6	Packet loss 5%	Female 1
7	Packet loss 6%	Male 2
8	Packet loss 8%	Female 2
9	Packet loss 10%	Male 1
10	Jitter 10ms	Female 2
11	Jitter 20ms	Male 2
12	Jitter 30ms	Female 1
13	Jitter 40ms	Male 1
14	Jitter 50ms	Female 2
15	Jitter 60ms	Male 2
16	Jitter 80ms	Female 2
17	Jitter 100ms	Male 1
18	Original	Female 2

7.2.6 Results

Figure 7.9 shows the results of all methods.

Classic method
In tests, it turned out that classic method with evaluation on the end of the sample maintains the same trends as other methods, but unfortunately, it has a very high uncertainty (CI95). These uncertainties can be reduced by re-assessment, but of course, it negates the advantage of using long samples. Surprisingly, this method does not show significant deviations from the other methods. The distribution of errors in the samples is quite regular so that the recency effect does not show much.

Equidistant method
The second method (equidistant evaluation during the playback of the sample) seems more appropriate. It achieves significantly lower uncertainties and basically simulates subjective tests using a larger number of shorter samples. This method is easy to implement and also processing of the results is simple. The only difference is the briefing before the test session. As a result, this method seems to be best suited for further use.

Free method

The third method (free evaluation during playback) should combine the advantages of both methods. However, it turned out that the uncertainty of this method is significantly higher than with the second method. The big problem was smaller quantities of votes than we expected. This method is also much more difficult for data processing.

In general, the results surprisingly show that voice samples spoken by a female have significantly lower ratings than male despite the lower distortion (for example samples number 4 and 5). This is probably due to the predominance of women in the panel of assessors. Overall, the gender of the speaker as well as the pitch, position, and pleasantness of voice has an impact on ratings, and it will be necessary to use a concatenated sample even for measuring of long samples.

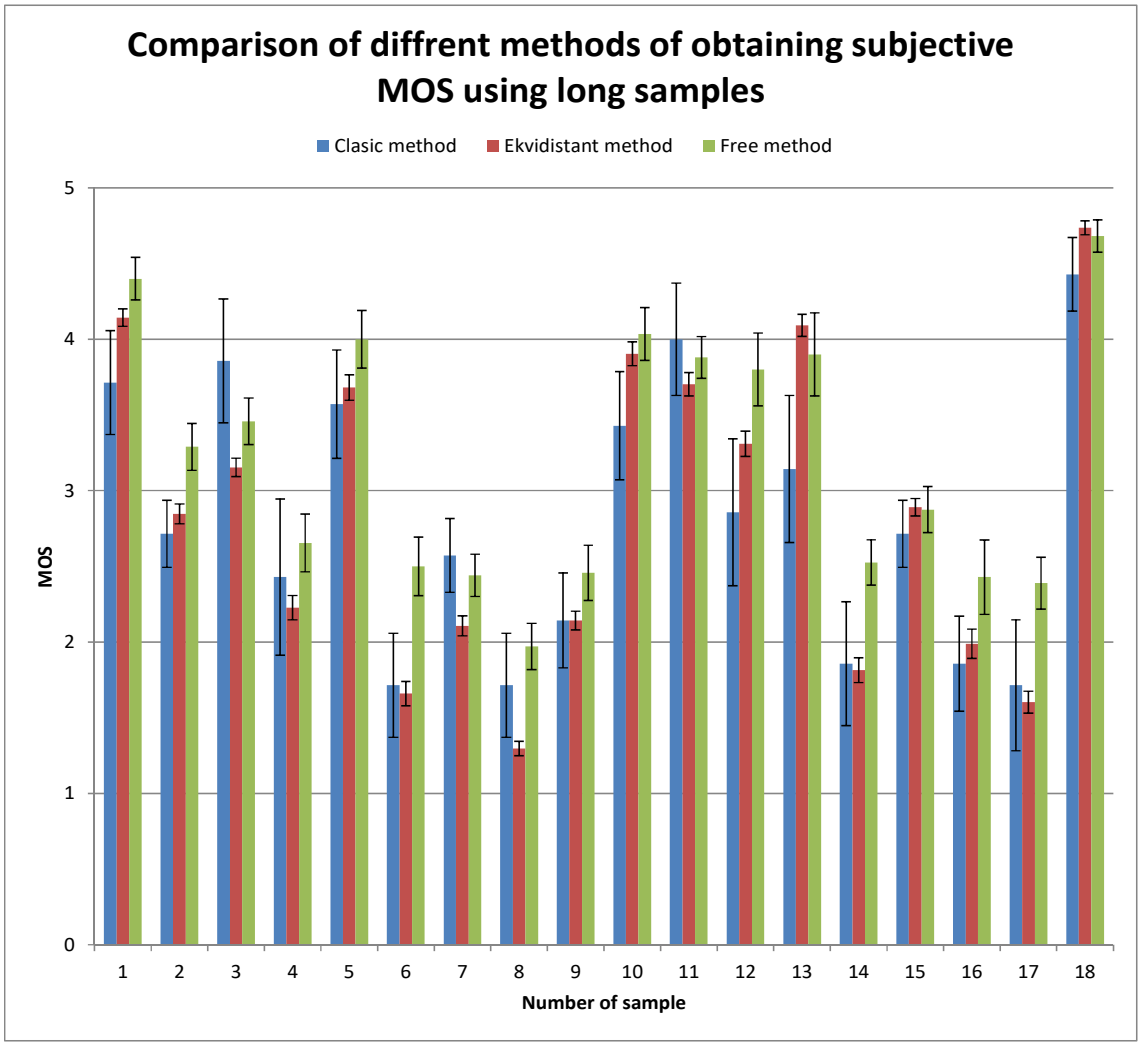


Fig. 7.9 Comparison of different methods of obtaining subjective MOS using long samples

7.2.7 Conclusion

Methodology currently used for speech transmission quality measurement is not entirely appropriate for use with conditions that include TCP/IP networks or any similar transfer method using packetization and buffers. It is based on repeated testing of samples with a length of 8 - 12s. This length is not sufficient for certain types of failures and a high number of repetitions causes problems in subjective tests. One of the ways to eliminate the disadvantages is the use of longer samples. Three different methods of obtaining MOS were used in this experiment. The equidistant method achieves lowest uncertainties, and it also has simple processing of results. For accurate measurement, it is also important to use concatenated samples from different speakers.

8 Conclusion

Compared to traditional telephone networks, the IP networks are relatively reliable. Errors such as massive packet loss or significant delays are relatively rare. This can cause problems because the methodology for subjective and partly objective tests is based on ITU-T Recommendation P.800 of 1993, which was designed for use in classic telephone networks.

Measurement according to this methodology in the modern network brings the following problems:

- Previously common disorders such as crosstalk or echo are nearly absent. There are other problems related to packet transmission (see above) and encoding (time warping).
- The occurrence of errors is relatively rare and irregular.
- The required sample length is too short to make mistakes to occur.

The purpose of this thesis was:

- In a series of experiments verify the effect of parameters of the IP network on the voice quality.
- Design and test the methodology for subjective tests using longer samples.

The following experiments were realized:

- Impact of the packet loss and jitter (chapter 6.1)
- Impact of the jitter buffer (6.2)
- Impact of the sample length (6.3)
- Impact of the codek and the QoS method (6.4)
- Impact of the delay and the stress of the listener (7.1)
- Impact of the packet loss, using long samples (7.2)

In an experiment using long samples, three different methods of collecting data from the evaluators were tested.

Following solutions were designed based on the results of experiments:

- In the framework of subjective tests, samples of 100 - 120s can be used. Thus a long sample minimizes problems with irregularities of failures and reduces the number of samples and repetition needed for conclusive tests.
- Equidistant method based on voting at regular intervals throughout the sample appears to be the best of the tested methods of data collection.

9 List of references

[1] ITU-T, “Rec. P.861, Objective quality measurement of telephone-band (300-3400 Hz) speech codecs,” *Int. Telecommun. Union, Geneva*, vol. 861, 1996.

[2] ITU-T, “Rec. P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” *Int. Telecommun. Union, Geneva*, vol. 862, p. 862, 2001.

[3] ITU-T, “Rec. P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs,” *Int. Telecommun. Union, Geneva*, 2007.

[4] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” *IEEE Int. Conf. Acoust. Speech, Signal Process. Proc. (Cat. No.01CH37221)*, vol. 2, pp. 2–5, 2001.

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

[6] ITU-T, “Rec. P.863, Perceptual Objective Listening Quality Assessment,” *Int. Telecommun. Union, Geneva*, vol. 863, 2011.

[7] J. G. Beerends *et al.*, “Perceptual Objective Listening Quality Assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement Part I-temporal alignment,” *AES J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

[8] ITU-T, “Rec. P.56, Objective measurement of active speech level,” *Int. Telecommun. Union, Helsinki*, vol. 56, 2011.

[9] J. G. Beerends *et al.*, “Perceptual Objective Listening Quality Assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement Part II–Perceptual Model,” *AES J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

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

[11] ITU-T, *Rec. P.50, Artificial Voices*. 1999.

[12] D. F. Hoth, “Room Noise Spectra at Subscribers’ Telephone Locations,” *J. Acoust. Soc. Am.*, vol. 12, pp. 499–504, 1941.

[13] IETF, *RFC 791 – Internet Protocol*. 1981.

[14] IETF, *RFC 768 - User Datagram Protocol*. 1980.

[15] IETF, *RFC 3261 - SIP: Session Initiation Protocol*. 2002, pp. 1–269.

[16] A. Korolev, “SIP call flow between UA, Redirect Server, Proxy and UA.” [Online]. Available: https://commons.wikimedia.org/wiki/File:SIP_call_flow_between_UA,_Redirect_Server,_Proxy_and_UA.png.

[17] ITU-T, “Rec G.711, Pulse code modulation of (PCM) of Voice frequencies,” *Int. Telecommun. Union*, 1993.

[18] ITU-T, *Rec. G.722, 7 kHz audio coding within 64 kbit/s*. 1988.

[19] 3GPP, *Technical Specification 26.071, Mandatory speech Codec speech processing function, AMR speech Codec, General description*. 1999.

[20] ITU-T, “Rec. G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-rate Wideband (AMR-WB),” *Int. Telecommun. Union*, 2003.

[21] ITU-T, “Rec. G.729, Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP),” *Int. Telecommun. Union*, 2012.

[22] “http://www.itu.int/rec/T-REC-G.729/e.” .

[23] ETSI, *ETS 300 961, Digital cellular telecommunications system (Phase 2+); Full rate speech Transcoding (GSM 06.10 version 5.1.1)*, no. ES 300 580-2. 1998.

[24] IETF, *RFC 5574 - RTP Payload Format for the Speex Codec*. 2009.

[25] ITU-T, “Rec. P.10, Vocabulary of terms on telephone transmission quality and telephone sets,” *Int. Telecommun. Union, Geneva*, pp. 1–42, 2006.

[26] ITU-T, “Rec. P.862.1, Mapping function for transforming P.862 raw result scores to MOS-LQO,” *Int. Telecommun. Union*, 2003.

[27] ITU-T, “P.563, Single-ended method for objective speech quality assessment in narrow-band telephony applications,” *Int. Telecommun. Union*, 2004.

[28] OPTICOM, “PESQ block diagram In: www.opticom.de [online]. [Cit. 13.12.2016]. Available at: http://www.opticom.de/technology/pesq.php.” .

[29] ITU-T, “Rec. G.107, The E-model: a computational model for use in transmission planning,” *Int. Telecommun. Union*, 2015.

[30] “http://www-x.antd.nist.gov/nistnet/.” .

[31] O. Slavata, “Neintruzivní měření kvality přenosu hlasu v telekomunikačních sítích,” CTU FEE, 2010.

[32] O. Slavata, “Měření kvality přenosu hlasu pro sítě typu VoIP,” CTU FEE, 2007.

[33] G. Sankaranarayanan and B. Hannaford, “Comparison of Performance of Virtual Coupling Schemes for Haptic Collaboration using Real and Emulated Internet Connections,” University of Washington, Seattle., 2007.

[34] I. Vondrka, “Implementation of the P.563 (3SQM) standard in PC’s using Lab/Windows CVI,” CTU FEE, 2005.

[35] Ditech Networks, “Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks,” 2007.

[36] E. Gündüzhan and K. Momtahan, “A Linear Prediction Based Packet Loss Concealment. Algorithm for PCM Coded Speech,” *IEEE Trans. Speech Audio Process*, vol. 9, no. 8, pp. 778–785, 2001.

[37] P. Počta and J. Holub, “Effect of speech activity parameter on PESQ’s predictions in presence of independent and dependent losses,” *Comput. Stand. Interfaces*, vol. 36, no. 1, pp. 143–153, 2013.

[38] P. Počta and J. Holub, “Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models,” *Acta Acust. United Acust.*, vol. 97, no. 5, pp. 852–868, 2011.

[39] Y. Stein and I. Druker, “The Effect of Packet Loss on Voice Quality for TDM over Pseudowires,” 2003.

[40] Head, “POLQA Application Guide,” 2012.

[41] J. Holub and O. Tomíška, “Delay Effect on Conversational Quality in Telecommunication Networks: Do We Mind?,” *Wirel. Technol.*, vol. 44, pp. 91–98, 2009.

[42] J. Sterne, “MP3: The meaning of format,” *Duke Univ. Press*, 2012.

[43] O. Slavata and J. Holub, “Impact of IP Channel Parameters on the Final Quality of the Transferred Voice,” in *Wireless Telecommunications Symposium*, 2012.

[44] L. Sun, I. Mkwawa, E. Jammeh, and E. Ifeachor, *Guide to Voice and Video over IP: For Fixed and Mobile Networks*. Springer, 2013.

[45] Q. Liu, Y. Liu, and D. Yang, “QoE Evaluation for Adaptive Streaming Based on Psychological Recency Effect,” 2013.

[46] ITU-T, *Rec. P.910, Subjective video quality assessment methods for multimedia applications*. 2008.

10 List of Publications

10.1 Related to this Thesis

10.1.1 Publications in Journals with Impact Factor

[1] SLAVATA, O. and HOLUB, J. Evaluation of Objective Speech Transmission Quality Measurements in Packet-based Networks. *Computer Standards & Interfaces.* 2013, **36**(3), pp. 626-630. ISSN 0920-5489. Available from: http://authors.elsevier.com/sd/article/S0920548913001244

10.1.2 Publications in ISI

[1] SLAVATA, O. and HOLUB, J. Impact of the Codec and Various QoS Methods on the Final Quality of the Transferred Voice in an IP Network. In: *2014 joint IMEKOTC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security.* 2014 Joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security. Madeira, 03.09.2014 - 05.09.2014. Madeira: IOPscience. 2015, ISSN 1742-6588. Available from: http://iopscience.iop.org/1742-6596/588/1/012011/

[2] HOLUB, J. and SLAVATA, O. Impact of IP Channel Parameters on the Final Quality of the Transferred Voice. In: STEWART, R. and BARTOLACCI, M., eds. *Wireless Telecommunications Symposium 2012 Papers and Presentation.* Wireless Telecommunications Symposium 2012. London, 18.04.2012 - 20.04.2012. Pomona (CA): California State Polytechnic University. 2012, pp. 1-5. ISBN 978-1-4577-0580-9. Available from: http://www.csupomona.edu/~wtsi/wts/Previous%20Conferences/WTS2012/index.htm

[3] SLAVATA, O. and HOLUB, J. Subjective Measurement and Objective Modeling of Voice Quality in a Conference Call for a Stressed Listener. In: HOLUB, J. and ŠMÍD, R., eds. *Measurement of Speech, Audio and Video Quality in Networks 2011.* Measurement of Speech, Audio and Video Quality in Networks 2011. Prague, 16.06.2011 - 17.06.2011. Praha: CTU Publishing House. 2011, pp. 14-19. ISBN 978-80-01-04848-1.

[4] HOLUB, J. and SLAVATA, O. Impact of IP Channel Parameters on the Final Quality of the Transferred Voice. In: ŠMÍD, R. and HOLUB, J., eds. *MESAQIN 2010.* Measurement of Speech, Audio and Video Quality in Networks. Praha, 03.06.2010 - 04.06.2010. Praha: České vysoké učení technické v Praze. 2010, pp. 18-22. ISBN 978-80-01-04569-5.

[5] SOUČEK, P., SLAVATA, O., and HOLUB, J. New Approach in Subjective and Objective Speech Transmission Quality Measurement in TCP/IP Networks.

In: 2014 joint IMEKOTC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security. 2014 Joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security. Madeira, 03.09.2014 - 05.09.2014. Madeira: IOPscience. 2015, pp. 1-4. ISSN 1742-6588 (34%, 33%, 33%)

[6] SLAVATA, O., SOUČEK, P., and HOLUB, J. Using long samples in subjective testing of voice transmission quality in IP network. In: POWELL, S, KETSEOGLOU, T, and SHIM, JP, eds. 2015 Wireless Telecommunications Symposium (WTS). Wireless Telecommunications Symposium (WTS). New York City, NY, 15.04.2015 - 17.04.2015. New York: IEEE. 2015, ISSN 1934-5070. ISBN 978-1-4799-677 (33.33%, 33.33%, 33.33%)

10.1.3Other Publications

[1] SLAVATA, O., HOLUB, J., and HÜBNER, P. Impact of Jitter and Jitter Buffer on the Final Quality of the Transferred Voice. In: *IDAACS-SWS\'2012.* Conferences on Intelligent Data Acquisition and Advanced Computing Systems. Offenburg, 20.09.2012 - 21.09.2012. Piscataway: IEEE. 2012, pp. 120-123. 1. ISBN 978-1-4673-4677-1. Available from: http://www.idaacs.net/sws2012/

[2] SLAVATA, O., "Neintruzivní měření kvality přenosu hlasu v telekomunikačních sítích", Diploma thesis, CTU FEE, 2010.

[3] SLAVATA, O., "Měření kvality přenosu hlasu pro sítě typu VoIP", Bachelor thesis, CTU FEE, 2007.

10.2Non-related to this Thesis

10.2.1Publications in ISI

[1] SLAVATA, O., SOUČEK, P., and HOLUB, J. New Concept of Laboratory Exercise on Temperature Measurements Using Thermocouple. In: *2013 Joint IMEKO (International Measurement Confederation) TC1-TC7-TC13 Symposium: Measurement Across Physical and Behavioural Sciences.* 2013 Joint IMEKO TC1-TC7-TC13 Symposium. Genoa, 04.09.2013 - 06.09.2013. Bristol: IOP Publishing Ltd. 2013, pp. 1-6. Journal of Physics Conference Series. ISSN 1742-6588. Available from: http://iopscience.iop.org/1742-6596/459/1/012059/pdf/1742-6596_459_1_012059.pdf

[2] SOUČEK, P., SLAVATA, O., and HOLUB, J. Innovation of Laboratory Exercises in Course "Distributed Systems and Computer Networks". In: *2013 Joint IMEKO (International Measurement Confederation) TC1-TC7-*

TC13 Symposium: Measurement Across Physical and Behavioural Sciences. 2013 Joint IMEKO TC1-TC7-TC13 Symposium. Genoa, 04.09.2013 - 06.09.2013. Bristol: IOP Publishing Ltd. 2013, pp. 1-4. Journal of Physics Conference Series. ISSN 1742-6588. Available from: http://iopscience.iop.org/1742-6596/459/1

10.2.2Other Publications

[1] HOLUB, J., et al. Towards Layer Adaptation for Audio Transmission. *International Journal of Interdisciplinary Telecommunications and Networking.* 2014, **6**(4), pp. 35-41. ISSN 1941-8663.

[2] SLAVATA, O. and HOLUB, J. Vliv použitého kodeku a QoS metody na výslednou kvalitu přeneseného hlasu v IP sítích [online]. *Journal of Physics: Conference Series.* ISSN 1742-6596.

[3] HOLUB, J., et al. Přenosné zařízení pro měření akustické odrazivosti a pohltivosti povrchů. *Sdělovací technika.* 2014, **61**(9), pp. 43-45. ISSN 0036-9942.

[4] HOLUB, J., et al. Towards Layer Adaptation for Audio Transmission [online]. In: *2014 Wireless Telecommunications Symposium (WTS).* 2014 Wireless Telecommunications Symposium (WTS). Pomona, 09.04.2014 - 11.04.2014. Piscataway: IEEE. 2014, pp. 1-4. ISBN 978-1-4799-1297-1.