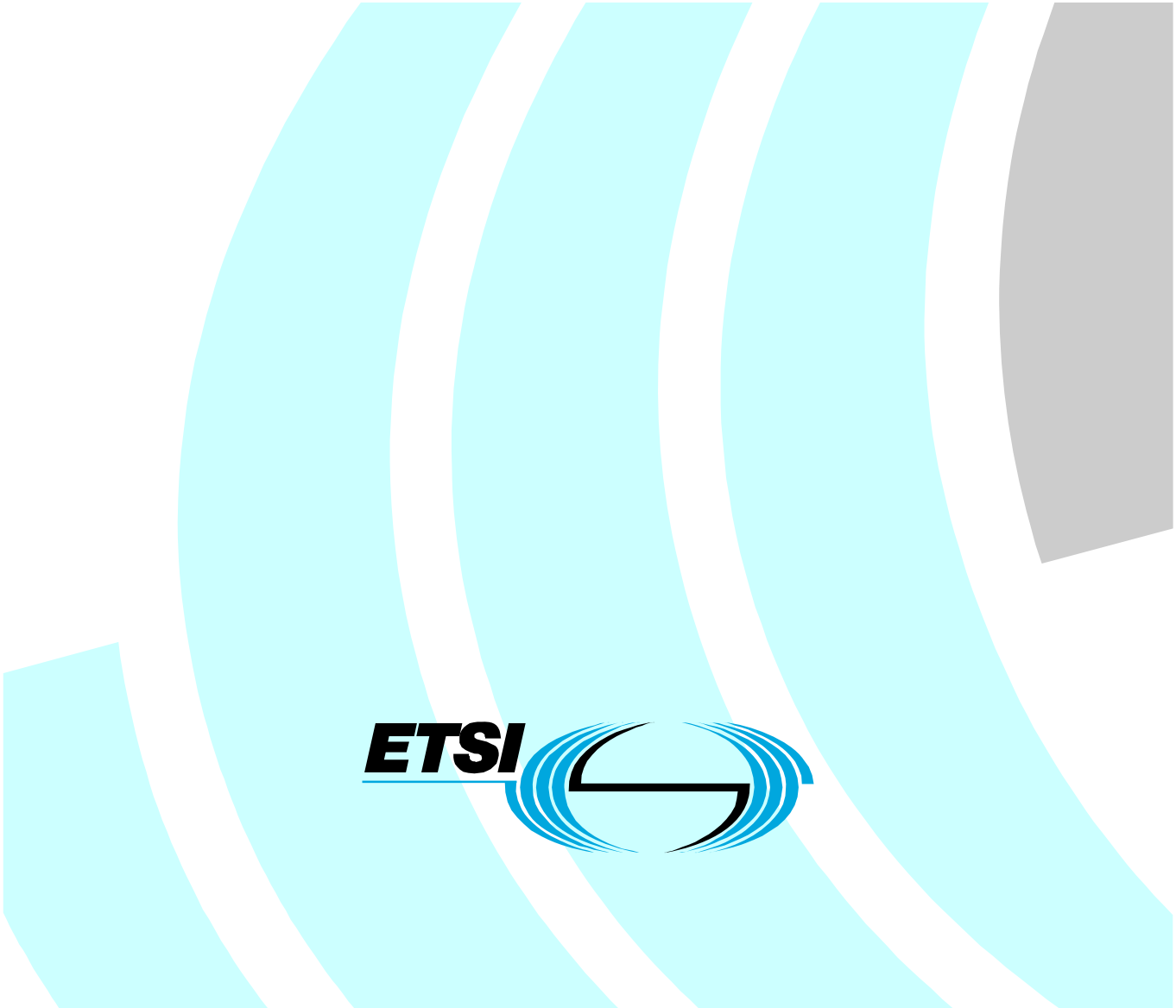
ETSI TR 102 948 V1.1.1 (2011-04)

Technical Report

**Speech and multimedia Transmission Quality (STQ);
Quality of Synthesized and Natural Speech
impaired by Packet Loss and Coding Algorithms**



Reference
DTR/STQ-00182

Keywords
quality, speech

ETSI

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

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

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

Contents

Intellectual Property Rights

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

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

Foreword

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

Introduction

In recent years, synthesized speech has reached a quality level which allows it to be integrated into many real-life applications, e.g. e-mail and SMS readers, etc. In particular, Text-to-Speech (TTS) can fruitfully be used in systems enabling the interaction with an information database or a transaction server, e.g. via the telephone network.

Modern telephone networks, however, introduce a number of degradations which have to be taken into account when services are planned and build up. The type of degradation depends on the specific network under consideration. In traditional, connection-based (analogue or digital) networks, loss, frequency distortion and noise are the most important degradations. In contrast, new types of networks (e.g. mobiles or IP-based ones) introduce impairments which are perceptively different from the traditional ones. Examples are non-linear distortions from low bit-rate coding-decoding processes (codecs), overall delay due to signal processing equipment, talker echoes resulting from the delay in conjunction with acoustic or electrical reflections, or time-variant degradations when packets or frames get lost on the digital channel. A combination of all these impairments will be encountered when different networks are interconnected to form a transmission path from the service provider to the user. Thus, the whole path has to be taken into account for determining the overall quality of the service operated over the transmission network.

The present document is addressed to network operators, service providers, users, manufactures and regulators. It considers the impact of some of the above mentioned impairments provided by IP-based networks on synthesized speech.

1 Scope

The present document provides information about the impact of specific types of packet loss and coding on speech quality predictions provided by ITU-T Recommendation P.862 [i.6] and P.563 [i.9] models, when both naturally-produced and synthesized speech are used. The variability of predictions with respect to the type of signal used (naturally-produced or synthesized) and loss conditions as well as their accuracy were assessed by comparing the predictions with subjective assessments.

The results indicate some implications for designers of speech communication systems.

It has to be emphasized that none of the instrumental algorithms investigated here (P.862 and P.563) were validated for synthesized speech. The presented analysis is a use case which is out-of-scope for these algorithms.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the reference document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

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

2.1 Normative references

The following referenced documents are necessary for the application of the present document.

Not applicable.

2.2 Informative references

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ITU-T Recommendation P.85 (1994): "A method for subjective performance assessment of the quality of speech voice output devices".

[i.2] D. Sityaev, K. Knill, T. Burrows: "Comparison of the ITU-T Recommendation P.85 standard to other methods for the evaluation of text-to-speech systems", in Proceedings of 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 - ICSLP), Pittsburgh (USA), pp. 1077-1080, 2006.

[i.3] M. Viswanathan, M. Viswanathan: "Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale", in Computer Speech and Language, vol. 19, pp. 55-83, 2005, ISSN 0885-2308.

[i.4] Y. Alvarez, M. Huckvale: "The reliability of the P.85 standard for the evaluation of text-to-speech systems", in Proceedings of 5th Int. Conf. on Spoken Language Processing (ICSLP 2002), Denver (USA), pp. 329-332, 2002.

[i.5] ITU-T Recommendation P.800 (1996): "Methods for subjective determination of transmission quality".

[i.6] ITU-T Recommendation P.862 (2001): "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs".

[i.7] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends: "Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality, Part I-Time-delay compensation", in J. Audio Eng. Soc., vol. 50, pp. 755-764, 2002, ISSN 1549-4950.

[i.8] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: "Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality, Part II-Psychoacoustic model, In J. Audio Eng. Soc., vol. 50, pp. 765-778, 2002, ISSN 1549-4950".

[i.9] ITU-T Recommendation P.563 (2004): "Single-ended method for objective speech quality assessment in narrow-band telephony applications".

[i.10] L. Malfait, J. Berger, M. Kastner: "P.563 - The ITU-T standard for single-ended speech quality assessment", in IEEE Transaction on Audio, Speech and Language Processing, vol. 14. No. 6, pp. 1924-1934, 2006, ISSN 1558-7916.

[i.11] S. Moeller: "Telephone transmission impact on synthesized speech: quality assessment and prediction", in Acta Acustica united with Acustica, vol. 90, pp. 121-136, 2004, ISSN 1610-1928.

[i.12] S. Moeller: "Quality of telephone-based spoken dialogue systems", Springer, New York (USA), Chapter 5, pp. 201-236, 2005, ISBN 0-387-23190-0.

[i.13] S. Moeller, D.-S. Kim, L. Malfait: "Estimating the quality of synthesized and natural speech transmitted through telephone networks using single-ended prediction models", in Acta Acustica united with Acustica, vol. 94, pp. 21-31, 2008, ISSN 1610-1928.

[i.14] D.-S. Kim: ANIQUE: "An auditory model for single-ended speech quality estimation", in IEEE Transaction on Speech and Audio Processing, vol. 13, No.5, pp. 821- 831, 2005, ISSN 1063-6676.

[i.15] D.-S. Kim, A. Tarraf: "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality", in Bell Labs Technical Journal, vol. 12, pp. 221-236, 2007, ISSN 1089-7089.

[i.16] ITU-T Recommendation G.729 (2007): "Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Exited Linear prediction (CS-ACELP)".

[i.17] ITU-T Recommendation P.862.1 (2003): "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[i.18] M. Chochlík, K. Grondžák, Š. Baboš: "Windows operating system core programming", (in Slovak), University of Žilina, Žilina (Slovakia), 2009, ISBN 978-80-8070-970-9.

[i.19] ITU-T Recommendation G.711 (1988): "Pulse code modulation (PCM) of voice frequencies".

[i.20] ETSI ETS 300 580-2 (2000): "Digital cellular telecommunications system (Phase 2) (GSM); Full rate speech; Part 2: Transcoding (GSM 06.10 version 4.2.1)".

[i.21] IETF RFC 3951 (2004): "Internet Low Bit Rate Codec (iLBC)".

[i.22] J.-M. Valin (2006): "Speex: A free codec for free speech", in Proceedings of Australian National Linux Conference (LCA 2006), Dunedin (New Zealand).

[i.23] 3GPP2 C.S0014-C (2007): "Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems".

[i.24] M. Yajnik, S. Moon, J. Kurose, D. Towsley: "Measurement and modelling of the temporal dependence in packet loss", in Proceedings of IEEE INFOCOM 1999 conference, New York (USA), vol. 1, pp. 345-352, 1999.

[i.25] W. Jiang, H. Schulzrinne: "QoS measurement of Internet real-time multimedia services", Technical Report (CUCS-015-99), Columbia University (USA), December 1999.

[i.26] H. Sanneck, N. T. L. Le: "Speech property-based FEC for Internet telephony applications", in Proceedings of the SPIE/ACM SIGMM Multimedia Computing and Networking Conference, San Jose (USA), pp. 38-51, 2000.

[i.27] W. Jiang, H. Schulzrinne: "Modelling of packet loss and delay and their effect on real-time multimedia service quality", in Proceedings of 10th International Workshop Network and Operations System Support for Digital Audio and Video (NOSSDAV 2000), Chapel Hill (USA), 2000.

[i.28] ITU-T Recommendation. P.830 (1996): "Subjective performance assessment of digital telephone-band and wideband digital codecs".

[i.29] J. Mullennix, S. Stern, S. Wilson, C. Dyson.: "Social perception of male and female computer synthesized speech", in Computers in Human Behavior, vol. 19, pp. 407-424, 2003, ISSN 0747-5632.

[i.30] S. Darjaa, M. Rusko, M. Trnka: "Three generations of speech synthesis systems in Slovakia", in Proceedings of XI International Conference Speech and Computer (SPECOM 2006), Sankt Peterburg (Russia), pp. 297-302, 2006, ISBN 5-7452-0074-X.

[i.31] L. Sun, G. Wade, B. M. Lines, E. C. Ifeachor: "Impact of packet loss location on perceived speech quality", in Proceedings of Internet Telephony Workshop (IPtel 2001), New York (USA), 2001.

[i.32] P. Arden: "Subjective assessment methods for text-to-speech systems", in Proceedings of Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology, Sheffield (UK), pp. 9-16, 1997.

[i.33] C. Delongu, A. Paoloni, P. Ridolfi, K. Vagges: "Intelligibility of speech produced by text-to-speech systems in Good and Telephonic Conditions", in Acta Acustica, vol. 3, pp. 89-96, 1995, ISSN 1610-1928.

[i.34] M. F. Spiegel, M. J. Altom, M. J. Macchi, K. L. Wallace: "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech", in Speech Communication, vol. 9, pp. 279-291, 1990, ISSN 0167-6393.

[i.35] A.W.Rix, J.G. Beerends, D.-S. Kim, P. Kroon, O. Ghitza: "Objective assessment of speech and audio quality - Technology and applications", in IEEE Transaction on Audio, Speech and Language Processing, vol.14, No.6,pp. 1890-1901, 2006, ISSN 1558-7916..

[i.36] ITU-T Del. Contr. D.123: "Proposed procedure for the evaluation of objective metrics", L.M. Ericsson (Author: Irina Cotanis), ITU-T Recommendation SG 12 Meeting, June 5-13, 2006, Geneva (Switzerland).

[i.37] ITU-T TD12rev1: "Statistical evaluation procedure for P.OLQA v.1.0", SwissQual AG (Author: Jens Berger), ITU-T Recommendation SG 12 Meeting, March 10-19, 2009, Geneva (Switzerland).

[i.38] P. Počta, J. Holub: "Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models", Submitted to Acta Acustica united with Acustica, July 2010 (under review).

[i.39] P.Počta, T. Terpák: "Packet Loss and Coding Impact on Quality of Synthesized Speech Predicted by PESQ and P.563 Models", in Proceedings of MESAQIN 2010 conference, Prague (Czech Republic), June 2010, pp. 26-36, ISBN 978-80-01-04569-5.

[i.40] P. Počta: "Impact of coding on quality of naturally-produced and synthesized speech", in Proceedings of 8th International workshop Digital Technologies 2010, Žilina (Slovakia), November 2010, ISBN 978-80-554-0304-5.

[i.41] QUALCOMM: "PESQ Limitations for EVRC Family of Narrowband and Wideband Speech Codecs", vol. 2008, 2008.

[i.42] ITU-T Recommendation P.56 (1993): "Objective measurement of active speech level".

[i.43] ITU-T Recommendation P.863 (2011): "Perceptual objective listening quality assessment".

3 Definitions, symbols and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

dependent losses: dependent packet loss is often referred to as 'bursty'

NOTE: It means that losses may extend over several packets, showing dependency between individual loss events. The burstiness is specified by conditional loss probability. This type of loss represents the loss distributions typically encountered in real networks. For example, losses are often related to periods of network congestion. Gilbert model is normally deployed to model this type of losses.

independent losses: each packet loss is independent (memoryless), regardless of whether the previous packet is lost or not

NOTE: This type of loss is normally modelled by Bernoulli model.

3.2 Symbols

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

ci_{95i}	the 95 % confidence interval
d	the number of degrees of freedom provided by the mapping function
dB	decibel
df	the number of degrees of freedom
δ_i	the standard deviation of subjective scores for stimulus i
F	*F-ratio (output parameter of ANOVA test)*
clp	conditional loss probability
M	the number of individual subjective scores
N	the number of stimuli considered in the comparison
p	probability that a packet will be dropped given that the previous packet was received
p^*	parameter characterizing the reliability of ANOVA test
q	probability that a packet will be received given that the previous packet was dropped
R	Pearson correlation coefficient
$rmse$	root mean square error
$rmse^*$	epsilon-insensitive root mean square error
ulp	unconditional loss probability
X_i	the subjective *MOS* value for stimulus i
$\overline{X}$	the corresponding arithmetic mean values of X
Y_i	the predicted *MOS* value for stimulus i,
$\overline{Y}$	the corresponding arithmetic mean values of Y

3.3 Abbreviations

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

ANIQUE+	Auditory Non-Intrusive Quality Estimation Plus
ANOVA	Analysis of Variance
CI	Confidence Interval
EVRC-B	Enhanced Variable Rate Codec version B
GSM-FR	Global System for Mobile Communications Full Rate codec
iLBC	Internet Low Bit Rate Codec
IP	Internet Protocol
ITU-T	ITU Telecommunication Standardization Sector
MAD	Mean Absolute Deviation
MOS	Mean Opinion Score
MOS-LQOn (P.563)	MOS-LQOn predicted by P.563

MOS-LQOn (P.862)	MOS-LQOn predicted by P.862
MOS-LQOn	MOS-Listening Quality Objective narrow-band
MOS-LQSn	MOS-Listening Quality Subjective narrow-band
MOSn	Mean Opinion Score narrowband
MS	Mean Square
PCM	Pulse Code Modulation
SPL	Sound Pressure Level
SS	Sum of Squares
TOSQA	Telekom Objective Speech Quality Assessment
TTS	Text-to-Speech
VoIP	Voice over Internet Protocol

4 Overview and related works

For determining the output quality of TTS systems (voice output devices), an application-oriented listening-only test described in ITU-T Recommendation P.85 [i.1] is recommended to be used. During such a test, participants have to solve a secondary task (e.g. to collect information which is contained in the sample) while listening to speech samples generated by TTS system. After the sample is finished, they have to judge different quality aspects on a set of 5-point category rating scales, such as overall impression, acceptance, listening effort, comprehension problems, articulation, pronunciation, speaking rate and voice pleasantness. By providing a secondary task, it is expected that the listeners' focus of attention is directed towards the contents of the speech signal and not towards its surface form alone. The arithmetic mean of all judgements collected on the "overall impression" scale is called a Mean Opinion Score (MOS). Although the method has been criticized for some deficiencies [i.2], [i.3] and [i.4], it is still the most commonly used method for the overall assessment of the speech output of TTS systems but when such output is impaired by transmission degradations, the modified versions of this test or classical test according to ITU-T Recommendation P.800 [i.5] are mainly deployed.

In order to quickly and economically optimize the speech output of automatic telephone services or to select between different TTS systems that are available in the market, network or service designers and system developers would like to have additional tools at hand. These tools should predict the quality perceived by the user - as it would be judged in an auditory test - on the basis of the speech signals generated by the system as well as degraded by network. Such tools are available for predicting the quality of natural speech transmitted over telephone channels, e.g. the standardized "P.862" model described in [i.6], [i.7] and [i.8] or the standardized "P.563" model defined in [i.9] and [i.10]. The former one is belonging to intrusive or comparison-based (full-reference) models, which are based on comparison between the degraded output signal and clean input signal of transmission channel. The clean speech signal is considered as the reference: the closer the transmitted signal is to this reference, the smaller the degradation and the higher quality. The difference is not calculated on the signal level but from an internal representation of the signals, consisting mainly of non-linear frequency analysis and loudness model. The latter is defined as a non-intrusive or single-ended (reference-free) model. The idea of such single-ended models is to generate an artificial reference (i.e. an "ideal" undistorted signal) from degraded speech signal and to use this reference in a signal-comparison approach. Once a reference is available, a signal comparison similar to the one of P.862 can be performed. The result of this comparison can further be modified by a parametric degradation analysis and integrated into an assessment of overall quality.

Some works have been carried out on study of quality of synthesized speech over the phone and performance of models for predicting and estimating the speech quality in case of synthesized speech usage. In [i.11], two questions were addressed whether: the overall amount of degradation is similar for synthesized compared to naturally-produced speech, and in how far can estimation models describing the quality impact on naturally-produced speech be used for estimating the effects on synthesized speech. Prototypical speech samples were first impaired by different degradations (e.g. circuit noise, low bit-rate coding, etc.) in controlled way, using a transmission simulation model. The samples were then judged upon by test subjects in an application-oriented listening-only scenario. It turns out that noise-type degradations exercise about the same quality impact on naturally-produced and synthesized speech. On the other hand, the impact of low bit-rate codecs is different for the two types of stimuli. In addition, the estimations of the transmission rating model which was investigated in this study (the E-model) seem to be in line with the auditory test results, both for naturally-produced as well as for synthesized speech, especially for uncorrelated noise. In [i.12], author extended the aforementioned work to new modelling examples with signal-based comparative measures, like P.862 and Telekom Objective Speech Quality Assessment (TOSQA). The results have shown that the both measures are capable of predicting quality of transmitted synthesized speech to a certain degree. All models (both mentioned signal-based models and E-model), however, do not adequately take into account the different perceptive dimensions caused by the source speech material and by the transmission channel. Moreover, they are only partly able to accurately predict the impact of signal-correlated noise. In [i.13], auditory MOS ratings for naturally-produced and synthesized speech samples transmitted over different telephone channels were estimated with three single-ended quality prediction models (Auditory Non-Intrusive Quality Estimation Plus (ANIQUE+), [i.14] and [i.15], Psytechnics model, and P.563). Mainly similar degradations to those introduced in [i.11] were used in this study. It was concluded that the investigated single-ended models mainly predict the effects of the transmission channel but not of the source speech material (naturally-produced or synthesized).

All previously mentioned works mostly focused on the impact of traditional network degradations (e.g. circuit noise, ambient noise, etc.) and coding on the quality of synthesized speech transmitted over phone. As mentioned before, new types of networks introduce new types of degradations, mainly time-variant degradations from packet loss or fading radio channels and non-linear distortions from newest low bit-rate coding-decoding processes (codecs).

Currently, these types of degradations are poorly investigated, especially with respect to their influence on synthesized speech [i.11]. That is the reason for exhaustive investigation of their impact on quality of synthesized speech. In particular, the present document provides information about an impact of specific types of packet loss and coding on speech quality predictions provided by P.862 and P.563 models, when both naturally-produced and synthesized speech are used. Two synthesized speech signals generated with two different Text-to-Speech systems and one naturally-produced signal were investigated. In addition, the variability of P.862's and P.563's predictions with respect to type of signal used (naturally-produced or synthesized) and loss conditions as well as their accuracy were assessed by comparing the predictions with subjective assessments. Finally, the aim of this study is three-fold: firstly, it would be beneficial to know whether the investigated models are able to provide valid predictions of perceived quality for the given application domain. Secondly, it would be worth to discover whether the impact of the packet loss and new coding approaches on the quality of synthesized speech is different from the impact on naturally-produced speech. Thirdly, it would be useful to find out which of the investigated modelling approaches is the most adequate one for the given task.

5 Experiment description

5.1 Experimental scenario

One-way VoIP session was established between two hosts (VoIP Sender and VoIP Receiver), via the loss simulator (Figure 1). In case of loss simulator, two currently most widely used models have been deployed for the purpose of packet loss modelling, namely Bernoulli and Gilbert loss model. More details about loss models can be found in clause 5.2. For this experiment the ITU-T Recommendation G.729AB encoding scheme [i.16] was chosen. In the measurements, two frames were encapsulated into a single packet; thus corresponding to a packet size of 20 milliseconds. Adaptive jitter buffer, G.729AB's native Packet Loss Concealment, and Voice Activity Detection/Discontinuous Transmission were implemented in the VoIP clients used. The jitter buffer does not play any role in case of this experiment because of small constant jitter inserted by the loss simulator during the measurement.

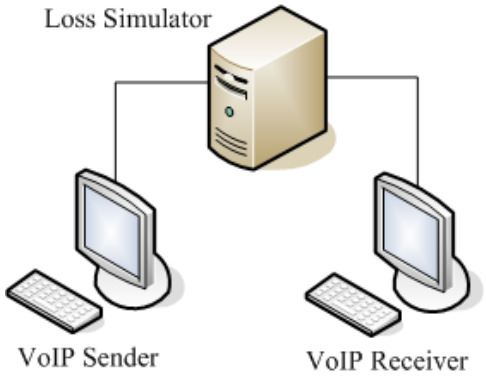


Figure 1: Experimental scenario

The reference signals described in clause 5.3 were utilized for transmission through the given *VoIP* connection. For coding experiment, the experimental scenario with loss simulator and VoIP clients (VoIP Sender and Receiver) was replaced just by coding algorithms, like ITU-T Recommendation G.729AB [i.16], ITU-T Recommendation G.711 [i.19], GSM-FR (ETS 300 580-2) [i.20], Internet Low Bit Rate Codec (iLBC) [i.21], Speex [i.22] and Enhanced Variable Rate Codec version B (EVRC-B) [i.23] but naturally speech quality assessment procedure was not changed and followed the description presented in clause 5.4. In case of EVRC-B codec, the noise suppression was only disabled in comparison to default settings. In other cases, default settings were applied.

5.2 Packet loss models

Packet loss is a major source of speech impairment in VoIP. Such a loss may be caused by discarding packets in the IP networks (network loss) or by dropping packets at the gateway/terminal due to late arrival (late loss). Several models [i.24] and [i.25] have been proposed for modelling network losses, the currently most widely used of them will be briefly discussed in the following clauses.

5.2.1 Bernoulli model

In the Bernoulli loss model, each packet loss is independent (memoryless), regardless of whether the previous packet is lost or not. In this case, there is only one parameter, namely the average packet loss rate (P_{pl}), which can be mathematically described by the following formula:

$$P_{pl} = \frac{n_l}{n} 100 \quad (1)$$

where n_l is the number of lost packets and n is the total number of transmitted packets in a trace.

5.2.2 Gilbert model

Most research in VoIP networks uses a Gilbert model to represent packet loss characteristics [i.24] and [i.26]. In 2-state Gilbert model as shown in Figure 2, State 0 is for a packet received (no loss) and State 1 is for a packet dropped (loss). p is the probability that a packet will be dropped given that the previous packet was received. 1-q is the probability that a packet will be dropped given that the previous packet was dropped. 1-q is also referred to as the conditional loss probability (*clp*). The probability of being in State 1 is referred to as unconditional loss probability (*ulp*). The ulp provides a measure of the average packet loss rate and is given by [i.27]:

$$ulp = \frac{p}{p+q} \quad (2)$$

The *clp* and *ulp* are used in the paper to characterize the loss behavior of the network.

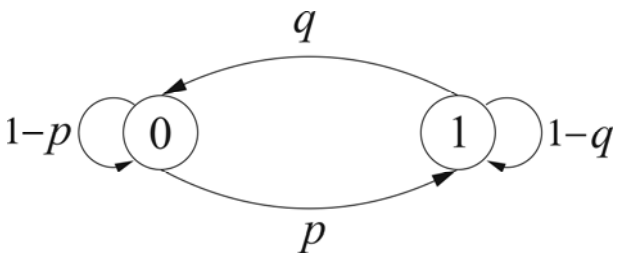


Figure 2: Gilbert model

Six independent loss and eleven dependent loss conditions were chosen to cover cases of interest. They consist of combinations of packet loss rate (from 0 % to 15 %) in case of independent losses and unconditional loss probability (ulp, 0 %, 1,5 %, 3 %, 5 %, 10 % and 15 %), conditional loss probability (clp, 70 % and 80 %) in case of dependent losses and 40 initial seeds to simulate different loss locations/patterns in both cases.

5.3 Reference signals

The reference signals selection should follow the criteria given by ITU-T Recommendations P.830 [i.28] and P.800 [i.5]. The reference signals should include talk spurts (sentences) separated by silence periods, and are normally of 1 s to 3 s long. They should also be active for 40 % to 80 % of their duration.

Following the criteria given by [i.5] and [i.28], three meaningful and non-technical sentences in Slovak with different length were defined for the purpose of this experiment. On basis of those sentences, speech files have been generated by two TTS systems (male voices) and recorded from one natural speaker (male). The natural speech sample was recorded in an anechoic environment; he was not professional speaker. The decision about using male voice came from the previous study published in [i.29]. The tests have proved that the message produced by the male synthetic voice was rated as more favourable (e.g. good and more positive) and was more persuasive, in terms of the persuasive appeal, than the female synthetic voice. These particular differences are perceptual in nature, and more likely due to differences in synthesis quality between male and female voices.

TTS system 1 is diphone synthesizer developed at the Institute of Informatics of the Slovak Academy of Sciences. That is the second version of Slovak TTS system (Kempelen 1.x), which is based on concatenation of small elements of a pre-recorded speech signals, mainly diphones. For the purpose of this experiment, the recent version of this synthesizer (Kempelen 1.6) has been used. More information about this type of synthesizer can be found in [i.30], clause 3. TTS system 2 is unit selection synthesizer also developed at same institute as TTS system 1. In case of this experiment, the recent version of this synthesizer (Kempelen 2.1) has been deployed. A new approach called pre-selection of element-candidates based on a phonetic analysis of the orthoepic transcription of text is deployed in recent version of this synthesizer. More information about this synthesizer can be found in [i.30], clause 4. It has to be noted that the speech material has not been specifically optimized after generation. In particular, very small pronunciation errors or inadequate prosody has not been corrected.

Finally, three reference signals (namely Natural, Diphone and Unit) in length of 12 seconds were applied. To avoid the differences in MOS values between the signals caused by different perceptual impact of same loss locations when the signals with unlike distributions of talk spurts are used [i.31], the same distributions and very similar durations of talk spurts (different talkers used) were deployed. Because the listening level has proven to be an important factor for the quality judgments of synthesized speech [i.32], all speech samples have been normalized to an active speech level of -26 dB below the overload point of the digital system, when measured according to ITU-T Recommendation P.56 [i.42] and stored in 16-bit, 8 000 Hz linear PCM. Background noise was not present.

5.4 Objective assessment

Finally, speech quality was objectively assessed by P.862 and P.563 algorithms. The quality was assessed on electrical interface. In case of P.862 algorithm, the scores were then converted to MOS-Listening Quality Objective narrow-band (MOS-LQOn) values by this equation.

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 * x + 4.6607}} \quad (3)$$

where *x* and *y* represent the raw P.862 score and the mapped MOS-LQOn, respectively. The equation mentioned is defined by ITU-T Recommendation P.862.1 [i.17]. In case of P.862 and P.563 scores calculation, some batch data processing techniques proposed in [i.18] were used.

5.5 Subjective assessment

The subjective listening tests were performed in MESAQIN.com laboratory in Prague according to ITU-T Recommendation P.800 [i.5]. Always up to 9 listeners were seated in listening chamber with reverberation time less than 190 ms and background noise well below 20 dB SPL (A). All together, 25 listeners (11 males, 14 females, 21 years to 30 years, mean 24,08 years) participated in the tests. 18 of them reported to have no experience with synthesized speech. The subjects were paid for their service.

The samples were played out using high quality studio equipment in random order and presented by two loudspeakers to the test subjects. Results in Opinion Score 1 to 5 were averaged to obtain MOS-Listening Quality Subjective narrowband (MOS-LQSn) values for each sample.

Because of big amount of very similar objective measurement data for dependent losses (clp = 70 % and 80 %), there was a need to make the decision which condition is better to test in order to limit the number of samples used in subjective tests. In other words, which condition provides us more data that can prove the behavior of models investigated? At the end, the decision was made to use second group of dependent losses, namely clp = 80 % due to some effects related to burstiness of losses reported in clause 6.1. Finally, the subjective tests were done for independent losses and dependent losses clp = 80 %. All together, 108 speech samples were selected for subjective testing of loss impact, 54 for each type of losses investigated here. Always 3 samples represented one network testing condition (the packet loss or the combination of ulp's and clp) and type of the signal used. In each sample collection, the best, average and worst cases were chosen from speech quality perspective. These were selected out of all recorded samples by expert listening. In addition to loss experiment, the subjective test for coding experiment was also realized, 6 current codecs were investigated (see clause 5.1) which results in 18 samples (6 codecs × 3 kinds of signal) involved in this part of subjective test. To having balanced sessions from impairment as well as size perspective, the samples from coding experiment were combined with samples from loss experiment, as follows: all samples from independent losses experiment (54 samples) and 9 samples of coding experiment, namely samples belonging to ITU-T Recommendation G.711 [i.19], iLBC and ITU-T Recommendation G.729 [i.16] codecs (all together 63 samples) belong to session No.1 and all samples from dependent losses experiment (54 samples) and the rest of samples of coding experiment (EVRC-B, GSM-FR and Speex) create session No.2 (containing 63 samples as well).

6 Experimental results

In this clause, the experimental results for objective assessment and comparison with subjective scores for both investigated impacts (loss, coding) are described and explained in more details, respectively.

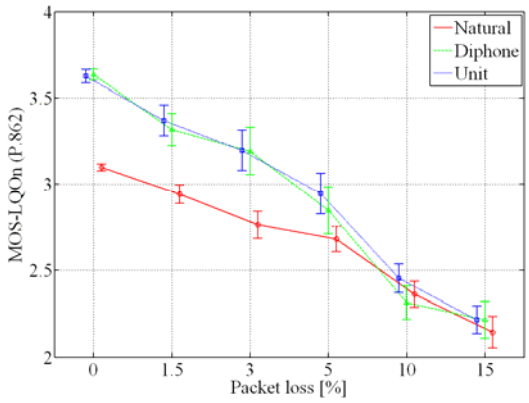
6.1 Impact of packet loss

6.1.1 Experimental results for objective assessment

The measurements were independently performed 40 times (40 different loss patterns) under the same packet loss (independent losses) and the same pair of *ulp* and *clp* (dependent losses) and the same signal. The average MOS-LQOn score, 95 % Confidence Interval (CI) and Mean Absolute Deviation (MAD) were calculated. The next clauses describe experimental results for the both examined types of losses in more details.

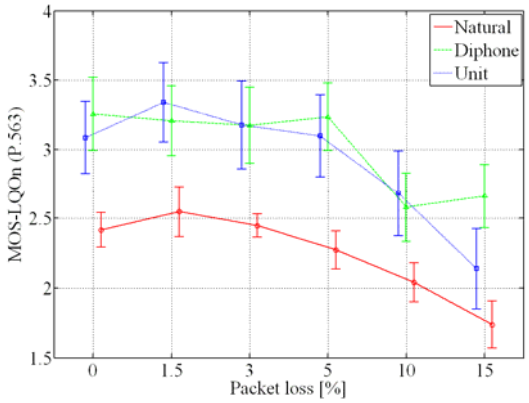
6.1.1.1 Independent losses

Using a Bernoulli model gives us the possibility to analyze P.862's and P.563's behavior only from two perspectives, namely packet loss and type of the signal used (Natural, Diphone, and Unit). Figures 3 and 4 depict differences between investigated signal types in speech quality evaluation, provided by P.862 and P.563 respectively. It can be seen from above-mentioned figures that a sort of the reference signal used (naturally-produced (Natural) or synthesized (Diphone and Unit)) has a significant impact on overall speech quality. In particular, the synthesized speech seems to be more prone to packet loss impairments than the natural speech. This might be due to higher poorness of the synthesized speech from a phonetic point of view, i.e. that there are fewer variations and fewer redundancies as also reported in [i.33] and [i.34], which results in speech quality differences when both types of speech are transmitted over telephone channel in the presence of packet losses. In addition, it can also be seen that both models got much higher MOSn values for synthesized signals, especially for 0 % packet loss. The similar effect has been obtained in [i.12]; see Figures 5.15 and 5.16. Unfortunately, the author did not specify the reason for this effect. Probably, that is due to some differences in 'artificiality' dimension between the naturally-produced and the synthesized signals coded by ITU-T Recommendation G.729 [i.16] codec, which can be perceived as degradations by the models. In case of the synthesized signal, small differences have been seen by the models and the models decreased the score according to that. On the other hand, the models saw higher differences in 'artificiality' dimension for natural signal and naturally considered that as higher degradation. The reported behavior was also motivation for us to investigate the impact of other (mainly newest) codecs on final MOSn score in such a case (see clause 6.2) from the point of view of objective as well as subjective assessments. Moreover, there is no difference between synthesized signals used from this perspective because of similar 'artificiality' dimension introduced by both synthesizers.



NOTE: The vertical bars show 95 % CI (derived from 40 measurements) for each loss and signal type.

Figure 3: MOS-LQOn predicted by P.862 [i.6] (MOS-LQOn (P.862)) as a function of packet loss for different types of signal used in case of independent losses



NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 4: MOS-LQOn predicted by P.563 [i.9] (MOS-LQOn (P.563)) as a function of packet loss for different types of signal used in case of independent losses

However, it can be seen from Figure 4 that non-monotonic results have been obtained in case of P.563 model. Such behavior could be explained by poorer performance of speech reconstruction and temporal clipping detection modules deployed in P.563 algorithm, in reported case. As described in [i.10], packet loss is one of the problematic conditions for P.563, and for any no-reference model, because it is not possible to guess what information has been lost. On the other hand, this problem can be alleviated by packet loss concealment algorithm implemented in codec but this algorithm usage can also negatively influence the performance of vocal tract model (part of the speech reconstruction module) as well as temporal clipping detection module by avoiding the abrupt loss of energy [i.10]. If the loss is concealed properly, the modules are not able to recognize the loss but in opposite case, the modules can detect it when the abrupt loss of energy is higher than the thresholds used in both cases. Those situations can affect the process of the quasi-clean reference speech signal creating and distortion-specific parameters detection and finally the behavior of P.563 in such a case. Moreover, as can be also seen from the aforementioned figure, the non-monotonic results occur more frequently for synthesized speech signals. This type of speech is more vulnerable to packet loss impairment (as also stated above). Due to that, the problematic situations related to packet loss concealment algorithm pointed out before, obtain more frequently.

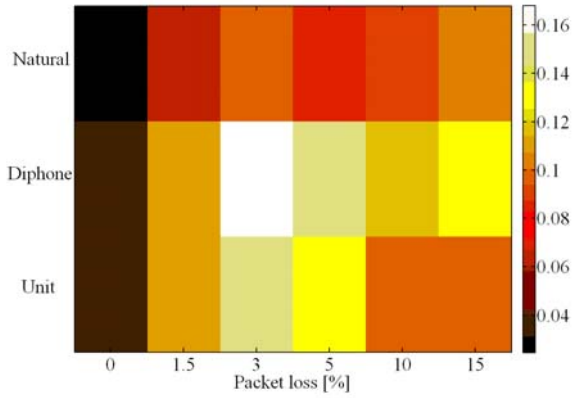


Figure 5: MAD of MOS-LQOn's predicted by P.862 [i.6] at each point of loss space and type of the signal used in case of independent losses

Figures 5 and 6 show MAD's of MOS-LQOn's (P.862) and MOS-LQOn's (P.563), which have been obtained for this experiment. It can be seen from Figures 5 and 6 that the deviations of predictions for naturally-produced speech are smaller than those for synthesized speech, especially for P.563 model. First fact is related to smaller sensitivity of the naturally-produced speech to packet loss impairments than the synthesized speech (as also pointed out above). Second one might be caused by earlier reported poorer performance of some P.563's modules when the synthesized speech is deployed. Naturally, the deviations of predictions rise for higher packet losses in both cases because of higher probability of losses obtained at active speech intervals in such a case.

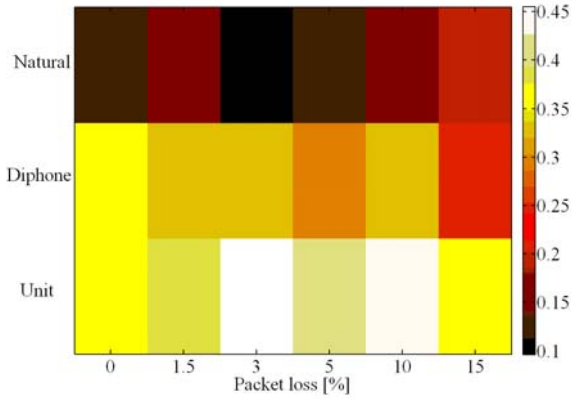
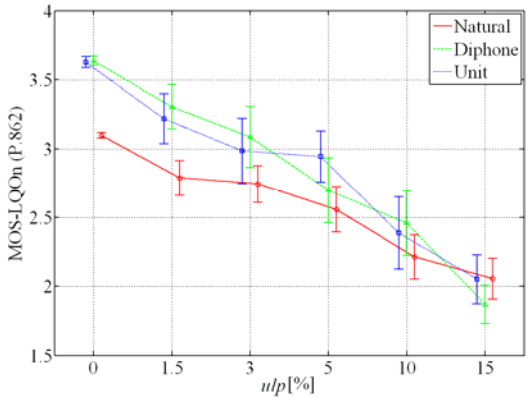


Figure 6: MAD of MOS-LQOn's predicted by P.563 [i.9] at each point of loss space and type of the signal used in case of independent losses

Two two-way analyses of variance (ANOVA) were conducted on MOS-LQOn's (P.862) and MOS-LQOn's (P.563) using packet loss and type of the signal as fixed factors (Table A.1 and A.6). The highest *F*-ratios were clearly found for the packet loss (F = 1 493,55, p^* < 0,01) in case of P.862 usage and for the type of signal used (F = 273,06, p^* < 0,01) in case of P.563 usage. Moreover, the signal factor (MOS-LQOn's (P.862)) and packet loss (MOS-LQOn (P.563)) showed a little bit weaker effect on quality than firstly mentioned factors for P.862 as well as P.563 based predictions, with F = 290,96, p^* < 0,01 and F = 87,73, p^* < 0,01, respectively. The realized *ANOVA* tests revealed that different factor has affected the average MOS-LQOn values for each model investigated. In particular, P.563 model seems to be more sensitive to type of the signal used than P.862. It has to be emphasizes that P.563 model has been built for monitoring the quality degradation produced by transmission channel on naturally-produced speech and thus has been trained to disregard the effect of the specific voice, and has not been trained on synthesized speech. Probably, those facts are responsible for such big impact of signal factor on P.563's predictions, reported in this experiment.

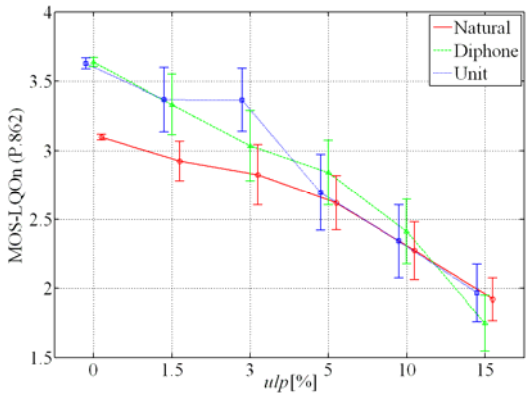
6.1.1.2 Dependent losses

Using a Gilbert model extends the possibilities to investigate P.862's and P.563's behavior to three perspectives, namely ulp, clp and naturally type of the signal used. The experimental results for all investigated clp's are depicted in Figures 7 to 10. It can be observed how speech quality drops, as expected, with both clp and ulp. Also, it is clear that a kind of the reference signal used could seriously influence the quality in case of dependent losses. Obviously, same effect as in first case (independent losses) was obtained. It means that the synthesized speech is more vulnerable to packet loss impairments than the natural speech, also in case of dependent losses. Moreover, the higher burstiness of losses (expressed by clp parameter) leads to higher non- monotonicity of predictions provided by P.563 model (see Figures 9 and 10) than for independent losses. It can be pronounced that is due to poorer performance of packet loss concealment algorithm under bursty losses, as widely reported in scientific papers (for instance in [i.31]).



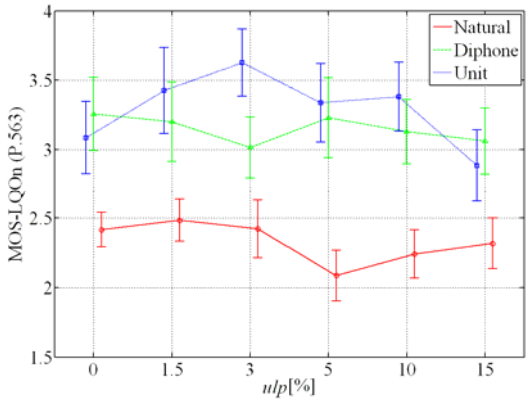
NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 7: MOS-LQOn predicted by P.862 [i.6] as a function of unconditional loss probability for different types of signal used in case of dependent losses (clp = 70 %)



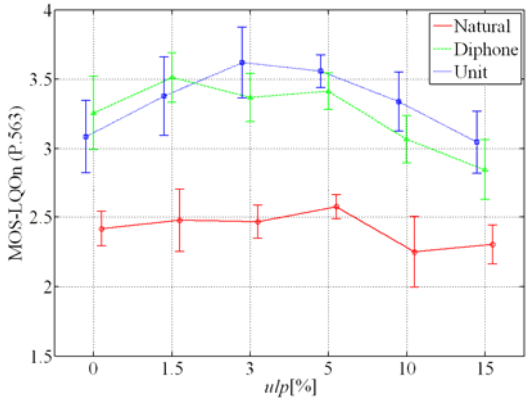
NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 8: MOS-LQOn predicted by P.862 [i.6] as a function of unconditional loss probability for different types of signal used in case of dependent losses (clp = 80 %)



NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 9: MOS-LQOn predicted by P.563 [i.9] as a function of unconditional loss probability for different types of signal used in case of dependent losses (clp = 70 %)



NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 10: MOS-LQOn predicted by P.563 [i.9] as a function of unconditional loss probability for different types of signal used in case of dependent losses (clp = 80 %)

In Figures 10 and 11, the MAD of MOS-LQOn's (P.862) and MOS-LQOn's (P.563) for 70 % clp can be seen. Unsurprisingly, P.862's and P.563 predictions deviation behavior is also similar as obtained in previous case. Interestingly, the MAD was increased for dependent losses and all signal types deployed but only in case of P.862 predictions. It was clearly expected that MAD will be negatively affected by burstiness in both cases (P.862 and P.563) but probably this effect was masked by more influencing effects, like poorer performance of speech reconstruction and temporal clipping detection modules in presence of packet loss.

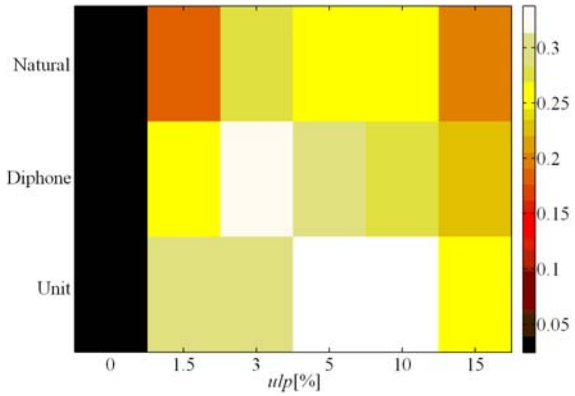


Figure 11: MAD of MOS-LQOn's predicted by P.862 [i.6] at each point of loss space and type of the signal used in case of dependent losses (clp = 70 %)

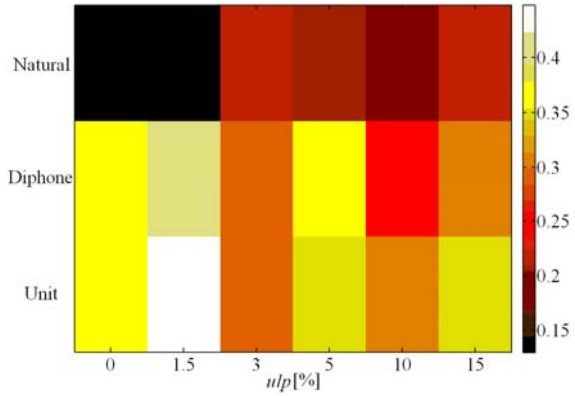


Figure 12: MAD of MOS-LQOn's predicted by P.563 [i.9] at each point of loss space and type of the signal used in case of dependent losses (clp = 70 %)

Similarly as for independent losses, four two-way ANOVA's were carried out on MOS-LQOn's (P.862) and MOS-LQOn's (P.563) for all investigated *clp*'s, using *ulp* and signal type as fixed factors (Tables A.3 to A.6). In principle, similar results as for independent losses were obtained. However, the higher impact of signal type (expressed by *F*-ratio; $F = 494{,}78$, $p^* < 0{,}01$ for *clp* = 70 % and $F = 709{,}56$, $p^* < 0{,}01$ for *clp* = 80 %) was obtained for dependent losses (increased by higher burstiness) in P.563 case, see Tables A.2, A.5 and A.6. Contrariwise, the loss impact (expressed by packet loss (independent losses) or *ulp* (dependent losses)) was decreased by higher burstiness in P.862 case (see Tables A.1, A.3 and A.4) but still remains the most influencing factor.

6.1.2 Comparison between subjective and predicted quality scores

In the following clauses, auditory MOS values (MOS-LQSn) will be compared to the predictions of the two investigated models, namely intrusive P.862 and non-intrusive P.563. The comparison will be performed for all experimental conditions (independent and dependent losses), i.e. all combinations of type of the signal and network conditions (packet loss or combinations of ulp and clp), respectively. It has to be noted that the experimental conditions for dependent losses were restricted to clp = 80 % conditions in this case due to similarities in the results obtained for both types of dependent loss conditions, as described in clause 5.4. However, the MOS-LQSn values will have been influenced by the choice of conditions in the actual experiment. In order to account for such influences, model predictions are commonly transformed to range of conditions that are part of the respective test [i.35]. This may be done e.g. using a monotonic 3rd order mapping function. Such functions have been determined for each model, each signal and each experiment individually, maximizing the correlation, minimizing the root mean square error and epsilon-insensitive root mean square error, see below.

The performance of models will be quantified in terms of Pearson correlation coefficient *R*, the respective root mean square error (*rmse*) and epsilon-insensitive root mean square error (*rmse**) as follows [i.36] and [i.37]:

$$R = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}} \tag{4}$$

and

$$rmse = \sqrt{\left(\frac{1}{N-d}\sum_{i=1}^{N}(X_i - Y_i)^2\right)} \tag{5}$$

with X_i the subjective MOS value for stimulus *i*, Y_i the predicted MOS value for stimulus *i*, $\overline{X}$ and $\overline{Y}$ the corresponding arithmetic mean values, *N* the number of stimuli considered in the comparison, and *d* the number of degrees of freedom provided by the mapping function (*d* = 4 in case of 3-order mapping function, *d* = 1 in case of no regression). On the other hand, the epsilon-insensitive root mean square error can be described as follows:

$$Perror_i = \max\left(0, \left|X_i - Y_i\right| - ci_{95_i}\right) \tag{6}$$

where the ci_{95i} represents the 95 % confidence interval and it is defined by [i.37]:

$$ci_{95_i} = t(0.05, M)\frac{\delta_i}{\sqrt{M}} \tag{7}$$

where *M* denotes the number of individual subjective scores and δ_i is the standard deviation of subjective scores for stimulus *i*. The final epsilon-insensitive root mean square error is calculated as usual but based on *Perror* with the formula (6):

$$rmse^* = \sqrt{\left(\frac{1}{N-d}\sum_{i=1}^{N}Perror_i^2\right)}. \tag{8}$$

The correlation indicates the strength and the direction of a linear relationship between the auditory and the predicted MOS values; it is largely influenced by the existence of data points at the extremities of the scales. The root mean square error (*rmse*) describes the spread of the data points around the linear relationship. The epsilon-insensitive root mean square error (*rmse**) is similar measure like classical *rmse* but *rmse** considers only differences related to epsilon-wide band around the target value. The 'epsilon' is defined as the 95 % confidence interval of subjective MOS value. By definition, the uncertainty of MOS is taken into account in this evaluation. For an ideal model, the correlation would be *R* = 1,0 and the *rmse* and *rmse** = 0,0.

All *R*, *rmse* and *rmse** will be calculated for the raw (not regressed) MOSn predictions and for the regressed MOS-LQOn values, obtained with the help of the monotonic mapping functions and both (the regressed and the not regressed MOSn predictions) will also be separated according to the type of signal used, in order to get an indication of the characteristics of the individual models on different types of source data.

6.1.2.1 Independent losses

At the beginning, it should be noted that 95 % confidence intervals for MOS-LQSn values presented in this comparison computed according to equation (7) were on average 0,2955 MOS (for Natural signal), 0,2625 MOS (for Diphone signal), 0,2847 MOS (for Unit signal). Figures 13 and 15 compare the MOS-LQSn values and the raw model predictions, namely MOS-LQOn (P.862) and MOS-LQOn (P.563). The corresponding correlations R and root mean square errors (rmse) and epsilon-insensitive root mean square errors (rmse*) are given in Table 1. The correlation calculated over all test conditions varies between 0,8915 and 0,9471 for P.862 and 0,5197 and 0,7356 for P.563 model (see Table 1). For P.862, the correlation coefficient is higher for 'unit' signal (synthesized speech generated by unit selection synthesizer) than for naturally-produced signal and diphone type of synthesized speech. Moreover, the smallest rmse and rmse* have been also obtained for synthesized speech generated by unit selection synthesizer. Contrariwise in P.563 case, the correlation is higher for naturally-produced speech but interestingly the smallest rmse and rmse* have been again attained for 'unit' signal.

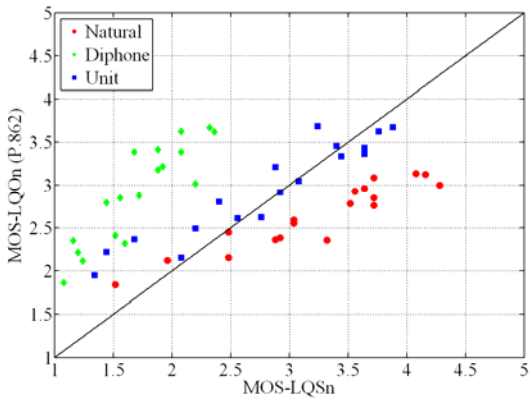


Figure 13: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (not regressed)

On the other hand, Figures 14 and 16 depict the subjective MOSn values (MOS-LQSn) and the regressed model predictions (MOS-LQOn (P.862), MOS-LQOn (P.563)). As attempted to use 3rd order regression (as mentioned above) has occasionally lead to non-monotonic results (only in case of P.563 model), the 2nd order regression was used instead. The orders of monotonic mapping functions are reported in the respective table, namely in Table 2. Table 2 also shows that the correlation coefficients slightly increase in all cases, and that the root mean square errors and epsilon-insensitive root mean square errors are considerably reduced, after applying mapping functions.

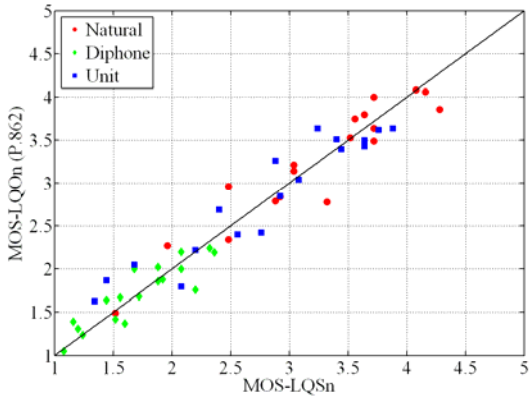


Figure 14: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (regressed)

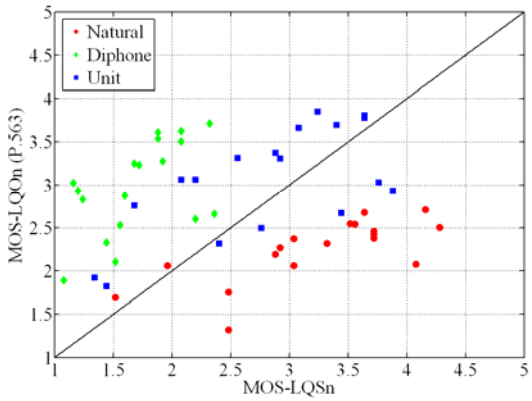


Figure 15: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (not regressed)

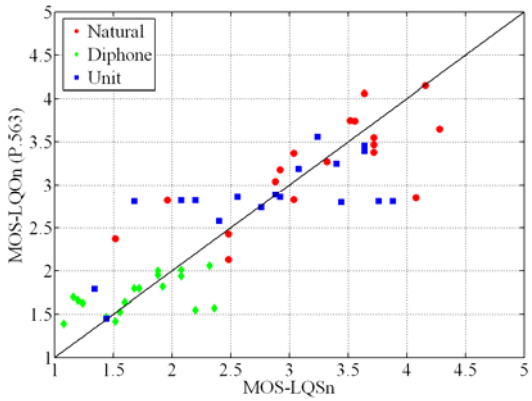


Figure 16: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (regressed)

Table 1: Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]) as well as MOS-LQOn (P.563 [i.9]) before regression for independent losses

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9366	0,1740	0,1148
	Diphone	0,8915	0,2959	0,2320
	Unit	0,9471	0,0712	0,0476
P.563	Natural	0,7356	0,2708	0,2064
	Diphone	0,5197	0,3251	0,2679
	Unit	0,6474	0,1480	0,0934

One two-way *ANOVA* was conducted on *MOS-LQSn*'s using packet loss and type of the signal as fixed factors (clause A.2.1, Table A.7). The clearly highest F-ratio was found for the signal factor (F = 350,72, p^* < 0,01). Moreover, the packet loss factor showed a little bit weaker effect on quality than former factor, with F = 99,31, p^* < 0,01. The realized *ANOVA* test revealed that subjects were more sensitive to the type of the signal used than to the independent losses. Probably, that was due to differences between the types of the investigated signals, especially from phonetic point of view (i.e. that the synthesized speech contains fewer variations and fewer redundancies and sounds sometimes less natural (mainly older approaches of speech synthesis)). Those differences were higher than impairments caused by independent losses and forced the listeners to change their opinions according to type of the signal used and not according to amount of impairments heard from speech sample assessed. A diagnostic analysis of the test data exposed that this effect mainly happened in case of listeners without previous experience with synthesized speech (in our case, 72 % of subjects reported no previous experience with synthesized speech, see clause 5.4). In addition, it was also found that one of the synthesized signals, namely 'diphone' signal (sounds less natural than 'unit' and 'natural' signals) was particularly disliked (on average over all conditions 'diphone' samples were rated by approx. 1,11 MOS-LQSn worse than the samples generated by unit selection synthesizer and by approx. 1,5 MOS-LQSn worse than naturally-produced samples). Excluding this signal from the analysis, the influence of type of the signal was decreased and packet loss became dominant factor (F (packet loss) = 87,99, p^* < 0,01; F (type of signal) = 42,02, p^* < 0,01), more details in Table A.8. This behavior is in line with the behavior of P.862, as can be clearly seen from Table A.1. Naturally, the impact of packet loss (F = 1493,55, p^* < 0,01) is much higher than in subjective test case because of a bit different approach employed in the comparison-based models. This type of the models only focuses on the impairments and not on type of the speech used, as humans intermittently do. In contrast to P.862 (Table A.1) as well as the modified human's behavior (Table A.8), the ANOVA results obtained for P.563 model (see Table A.2) are inverse. It has to be again emphasized that P.563 model has been built for monitoring the quality degradation produced by transmission channel on naturally-produced speech. Thus, the model has been trained to disregard the effect of the specific voice, and he does not been trained on synthesized speech. Probably, those facts are responsible for such behavior. On the other hand, the ANOVA results obtained for P.563 (see Table A.2) are surprisingly close to the first results obtained for auditory test (Table A.7). This fact supports our previous statement that P.563 is more sensitive to type of the signal than P.862 and subjects and on the other hand it looks like that he dislikes some kinds of signals, as the humans in our subjective test.

Table 2: Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]) as well as MOS-LQOn (P.563 [i.9]) after regression and order of monotonic mapping function for independent losses

	Type of the signal	R	rmse	rmse*	Order of monotonic mapping function
P.862	Natural	0,9429	0,0654	0,0519	3
	Diphone	0,8934	0,0468	0,0481	3
	Unit	0,9517	0,0594	0,0453	3
P.563	Natural	0,7529	0,1249	0,0858	2
	Diphone	0,5282	0,0854	0,0653	2
	Unit	0,6999	0,1382	0,0939	3

6.1.2.2 Dependent losses

Firstly, it should be mentioned that 95 % confidence intervals of MOS-LQSn values for Natural signal, Diphone signal and Unit signal computed according to equation (7) were on average 0,22827 MOS, 0,2532 MOS and 0,2299, respectively. Figures 17 and 19 show the MOS-LQSn values and the raw model predictions for dependent losses, and Table 3 lists the respective correlations, root mean square errors and epsilon-insensitive root mean square errors. As observed for independent loss test, the correlation between auditory judgements and instrumental predictions varies considerably between voices and models (see Table 3). For P.862, the correlation coefficient is highest for naturally-produced speech. Moreover, the smallest rmse and rmse* have been attained for synthesized speech generated by unit selection synthesizer, likewise as for independent losses. On the contrary in P.563 case, the correlation is higher for 'diphone' signal but interestingly the smallest rmse and rmse* have been obtained for 'natural' signal.

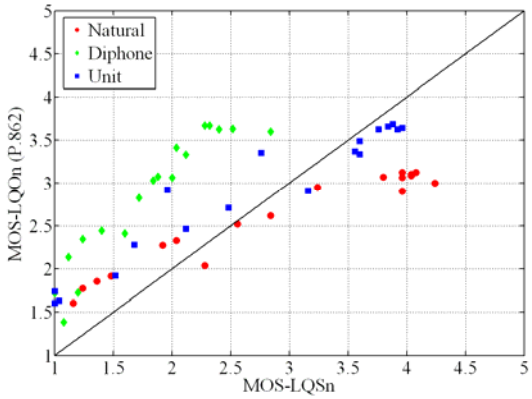


Figure 17: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for dependent losses (not regressed)

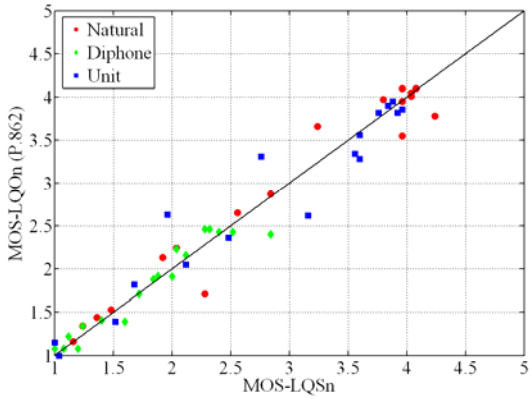


Figure 18: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for dependent losses (regressed)

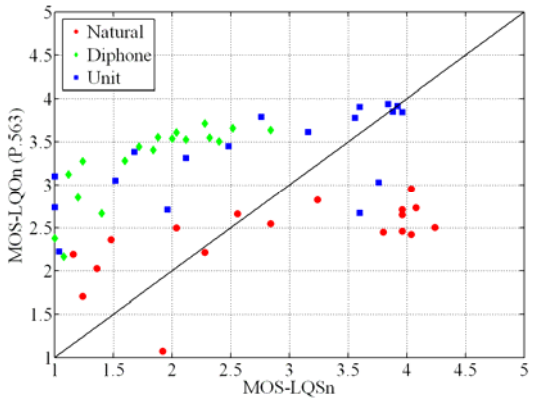


Figure 19: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for dependent losses (not regressed)

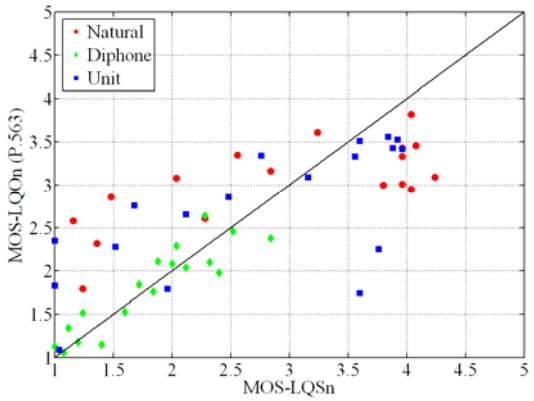


Figure 20: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for dependent losses (regressed)

Comparing the performance of the two investigated models, alike as in previous case, P.862 model once more attains the highest correlations and smallest root mean square errors and epsilon-insensitive root mean square errors for all types of signals used in this study, as clearly expected.

Interestingly, a bit higher correlations as well as a bit smaller root mean square errors have been obtained for dependent losses mostly in P.862 case, see Tables 1 to 4. On the contrary, the a bit smaller epsilon-insensitive root mean square errors have been mostly attained for independent losses.

Table 3: Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]) as well as MOS-LQOn (P.563 [i.9]) before regression for dependent losses

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9723	0,1690	0,1130
	Diphone	0,9430	0,2590	0,1972
	Unit	0,9660	0,1099	0,0831
P.563	Natural	0,6260	0,2535	0,1953
	Diphone	0,8114	0,3625	0,3060
	Unit	0,6751	0,2549	0,2255

Table 4: Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]) as well as MOS-LQOn (P.563 [i.9]) after regression and order of monotonic mapping function for dependent losses

	Type of the signal	R	rmse	rmse*	Order of monotonic mapping function
P.862	Natural	0,9766	0,0642	0,0559	3
	Diphone	0,9613	0,0394	0,0507	3
	Unit	0,9686	0,0732	0,0538	3
P.563	Natural	0,6260	0,2178	0,1604	1
	Diphone	0,9049	0,0609	0,0528	3
	Unit	0,6751	0,2009	0,1629	1

Likewise as in previous case, one two-way ANOVA was conducted on MOS-LQSn's using *ulp* and type of the signal as fixed factors (Table A.9). In practice, similar results as for independent losses were obtained. However, the smaller impact of signal type (expressed by F-ratio; $F = 145,46$, $p^* < 0,01$) was obtained for dependent losses than for independent losses ($F = 350,72$, $p^* < 0,01$) in case of all signals involved in analysis, see Tables 11 and 13. On the other hand, the loss factor is currently more influential than before but still does not overcome the signal factor. As in previous case, the 'diphone' signal was again excluded from the analysis, see the reasons above in clause 6.1.2.1. The loss impact (expressed by packet loss (independent losses) or *ulp* (dependent losses)) again came to be dominant factor, when excluding 'diphone' signal from the analysis, see Table A.10. Moreover, the impact of the signal factor was considerably decreased in comparison to previous case.

On the basis of this comparison, it can be pronounced that the subjective tests also confirmed the objective experimental results (presented in clause 6.1.1), namely vulnerability effect (by the behavior of the P.862 predictions and auditory ratings in scatter plots as well as by lower subject ratings especially for higher losses, see Tables 1 to 4 and Figures 13, 14, 17 and 18, respectively). On the other hand, the effect is not as dominant as shown in clause 6.1.1.1. The reason for that is discussed below (see clause 6.2).

6.2 Impact of different codecs on subjective and objective scores

The codecs investigated here cover a wide range of different types of degradations. In particular, the ITU-T Recommendation G.729 [i.16] AB, Speex, iLBC, GSM-FR and EVRC-B introduce 'artificiality' dimension, unnatural sounding whereas the ITU-T Recommendation G.711 [i.19] produce no perceptual degradation (natural sounding), (informal expert judgements).

Figures 21 to 23 show a fundamental difference in the quality judgements for natural speech and synthesized speeches provided by auditory test, P.862 and P.563, when processed by those codecs. The 95 % confidence intervals of MOS-LQSn values for the investigated signals (Natural, Diphone, Unit) computed according to equation (7) were on average 0,3494 MOS, 0,2668 MOS and 0,2754 MOS, respectively. In particular, a comparison of P.862 and P.563 predictions to the auditory MOSn values is shown in Figure 21 for naturally-produced speech. It is possible to see from the mentioned figure that 'artificially sounding' codecs are rated significantly worse in both models' predictions compared to the auditory test. Whereas for the ITU-T Recommendation G.711 [i.19] codec (natural sounding codec) the predicted quality especially provided by P.862 is in better agreement with the auditory results, as in previous case. Furthermore, P.563 model under-predicts the quality much more than P.862 in all cases.

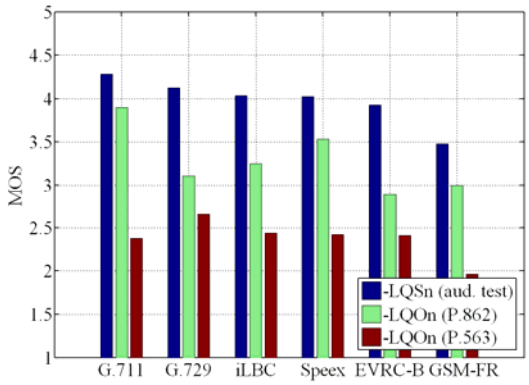


Figure 21: Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by P.862 [i.6] as well as by P.563 [i.9] for naturally-produced speech

The picture is quite different for synthesized signals, see Figures 22 and 23. In Figure 22, the comparison of the auditory ratings with the predictions provided by the two investigated models for 'diphone' signal can be seen. As discussed above (see clause 6.1.2.1), 'diphone' signal (sounds less natural than 'unit' and 'natural' signals) was particularly disliked by test subjects. This is probably the reason for such small ratings provided by subjects. In general, it would be expected that his behavior will be in line with the behavior of the second type of synthesized signal, namely 'unit' signal because of similar behavior attained for objective results (see Figures 3 and 4, 7 to 10, clauses 6.1.1.1 and 6.1.1.2). On the basis of the presented fact, it was decided to omit the 'diphone' signal from the further analysis of the behavior of synthesized speech under coding impairments. On the other hand, the behavior of the 'diphone' signal can be used as an example how higher unnaturalness of the signal can affect the opinions of the test users. Figure 23 depicts the effect of the investigated codecs on MOS-LQSn and MOS-LQOn predicted by P.862 as well as P.563 models for 'unit' signal. In contrast to naturally-produced speech (see Figure 21), the predictions of both models are in good agreement - with the exception of some predictions provided by P.563 model, like for ITU-T Recommendation G.711 [i.19] codec, etc.- with the auditory ratings. Regarding the behavior of P.563 for ITU-T Recommendation G.711 [i.19] codec, probably, the degraded 'unit' signal (coded by ITU-T Recommendation G.711 [i.19]) contains similar degree of 'artificiality' as the internal reference (quasi-clean reference speech) provided by P.563's speech reconstruction module (based on linear prediction). The P.563 model saw it as small degradation and the final score was partly decreased according to the dimension of the discussed degradation. Same also holds true for other codecs with similar behavior, namely Speex, GSM-FR, etc. On the other hand, such effect is not a case for P.862 because of the different kind of objective model (the internal reference is not used in this case). As can be seen from Figure 21, opposite effect has been detected for naturally-produced speech. The degraded natural speech (coded by ITU-T Recommendation G.711 [i.19]) contains much less 'artificiality' than the internal reference (quasi-clean reference speech) provided by P.563's speech reconstruction module (based on linear prediction). Finally, the psychoacoustic model considered the difference in 'naturalness' as degradation and consequently decreased the final score according to the dimension of this difference.

Moreover, when comparing the behavior of the synthesized speech with the behavior of naturally-produced speech from auditory ratings perspective see Figure 24 (excluding 'diphone' signal from this comparison because he was disliked by subjects in the test), there are some differences between subject ratings for the 'unit' signal and 'natural' signal. The observed differences may be due to differences in quality dimensions perceived as degradations by the test subjects. Whereas the 'artificiality' dimension introduced by the investigated 'unnatural sounding' codecs is additional degradation for the naturally-produced speech, this is not a case for the synthesized speech, which already carries a certain degree of artificiality. Furthermore, it looks like that the synthesized speech is insensitive to degradations introduced by the investigated codecs – except for GSM-FR codec - because of high degree of 'artificiality' dimension introduced by synthesizer. Regarding the GSM-FR codec behavior, probably this codec introduce some additional degradation to artificiality (for instance noisiness), which is a reason for lower scores for synthesized as well as naturally-produced speech. Our results are well in line with the results described in [i.12]. The synthesized speech is assessed a little more pessimistically than natural speech for ITU-T Recommendation G.729 codec, which is shown in Figure 5.12 in [i.16], p.225. On the other hand, the synthesized speech is rated a bit more optimistically by subjects than naturally-produced speech for IS-54 codec and its combinations. The effect is much more dominant for its combinations. Unfortunately, this codec as well as its combinations were not investigated in this study but then the GSM-FR codec was involved in this study which belongs to similar family of codecs. The same behavior as for IS-54 in [i.12] was also reported here for GSM-FR, probably because of very similar special techniques deployed in both codec-families. Regarding the predictions of P.862 (see Figures 5.15 and 5.16 in [i.12]), which were also investigated in the discussed study, they are more or less in line with our results, particularly for ITU-T Recommendation G.729 codec (see Figures 21 and 23).

Unfortunately, the study published in [i.16] is mainly focused on the different types of codecs and its combinations. This study can serve as an extension of the study published in [i.12].

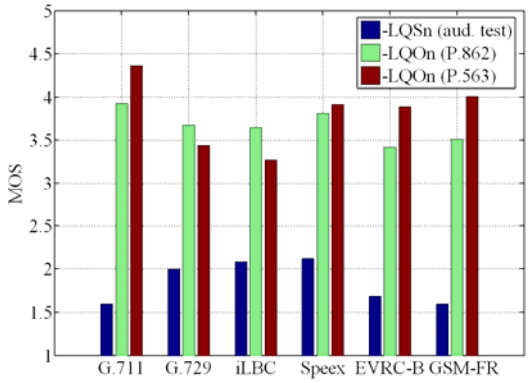


Figure 22: Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by P.862 [i.6] as well as by P.563 [i.9] for synthesized speech generated by Diphone synthesizer

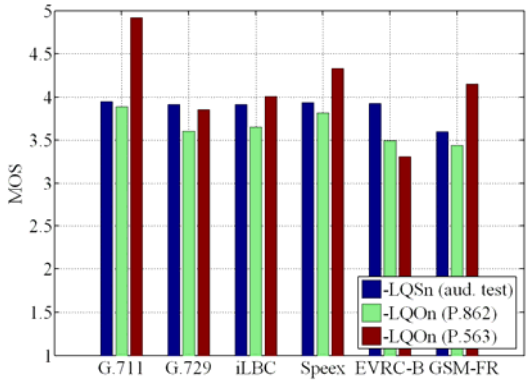


Figure 23: Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by P.862 [i.6] as well as by P.563 [i.9] for synthesized speech generated by Unit selection synthesizer

In addition, it looks like the both models have serious problem to correctly predict the quality of the natural speech impaired by present 'unnatural sounding' codecs like ITU-T Recommendation G.729 [i.16] (they predict the quality a bit pessimistically than was judged in the test), see Figure 21. The P.563 model is even much more pessimistic than P.862 in this case. On the other hand, this fact partially masks the vulnerability effect (see clause 6.1.1) but the effect is still here as can be seen from Figures 13, 14, 17 and 18, especially for lower MOSn values (higher packet losses).

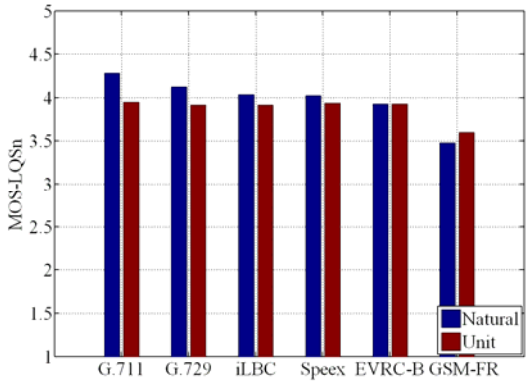


Figure 24: Comparison of the subjective ratings for naturally-produced speech with the ratings for synthesized speech generated by Unit selection synthesizer

Comparing the performance of the two investigated models from coding impairments perspective, P.862 model again over-performs P.563 model, mainly for naturally-produced speech.

7 Conclusions

In the present document, auditory MOSn values for the naturally-produced and synthesized speech samples transmitted over different simulated telephone channels were predicted with one comparison-based (P.862) and one single-ended (P.563) quality prediction models. The main goal of this study was to gain a better understanding of behavior of both models predictions under different types of losses, coding schemes and signals as well as to assess their accuracy by comparing the predictions with subjective assessments. Additional information with regard to this study can be found in [i.38], [i.39] and [i.40]. It has to be again emphasized that none of the instrumental models investigated here (P.862 and P.563) was verified for synthesized speech, the presented analysis is an out-of-domain use case for these models.

Three specific questions were addressed in this investigation (see clause 4). The first question can be answered in a positive way. All in all, the predictions provided by P.862 model seem to be in line with the auditory ratings, especially for loss conditions. Unfortunately, P.862 has some problems with the coding impairments, particularly when the naturally-produced speech is used. On the other hand, P.563 is less accurate as P.862 in both cases (loss conditions and coding impairments). Finally, it is possible to pronounce that both models are capable of predicting the quality of the transmitted synthesized speech under the investigated conditions to a certain degree, which is also confirmed by the reasonable correlations to the results of the auditory tests. Addressing the second question, the packet loss and coding impairments have a different impact on the quality of naturally-produced speech and synthesized speech. More precisely, the synthesized speech seems to be slightly more affected by loss degradations (especially for higher losses) and this kind of speech is also insensitive to the coding impairments provided by current codecs, like ITU-T Recommendation G.729 [i.16], iLBC, etc. Comparing the both models, the P.862 model seems to cope best with the both degradations investigated here (question 3). Finally, it can be concluded on the basis of the results obtained in this study that P.862 model can be used for assessing the quality of synthesized speech impaired by packet loss and coding algorithms with very good accuracy (see Tables 1 to 4, for more details). On the other hand, it is not possible to deploy P.563 model in such case (poor performance obtained).

8 Implications and future work

The results have some implications for the designers of telecommunications networks (speech communication systems) and of speech synthesis providers. As mentioned before, the synthesized speech is a bit more prone to packet loss impairments than naturally-produced speech. This suggests that concealment method with better performance has to be used for the networks with a significant percentage of services based on synthesized speech to obtain a similar degree of quality as for natural speech. On the other hand, the results of the second experiment (coding impact) show that the synthesized speech is not sensitive to the coding impairments provided by present codecs, like ITU-T Recommendation G.729 [i.16], Speex, etc. This allows designers to select an arbitrary codec from the current codecs available in the market without any impact on the ultimate speech quality. Moreover, the designers have to be careful to use objective models like P.862, etc. for predicting the quality of natural speech coded by present 'unnatural sounding' codecs because of the misleading predictions pointed out above. Similar statement related to EVRC family codecs has been reported in [i.41].

Future work will focus on the following issues. Firstly, it would be beneficial to investigate the performance of new intrusive model for predicting speech quality, contained in ITU-T Recommendation P.863 [i.43] under conditions investigated here (for instance as a part of characterization phase of this model). Secondly, on the basis of the results obtained for P.563 model, it would be useful for network operators and service providers, etc. to design new non-intrusive model for such conditions (synthesized speech and IP impairments). Thirdly, the extension of the E-model towards the synthesized speech impaired by the time-varying and coding impairments would be also desirable.

Annex A:
ANOVA results

A.1 ANOVA for objective results

In the next clauses, the detailed results of the analysis of variance (ANOVA) conducted on MOS-LQOn for independent and dependent losses can be found.

A.1.1 Independent losses

Tables 5 and 6 provide the results of ANOVA carried out on the independent losses test results (Dependent variable: MOS-LQOn (P.862) and MOS-LQOn (P.563)) described in more detail in clause 6.1.1.1.

Table A.1: Summary of ANOVA conducted on MOS-LQOn's (P.862 [i.6]) in case of independent losses

Effect	SS	df	MS	F	p*
Packet loss (1)	141,477	5	28,2954	1493,55	0,0000
Type of the signal (2)	11,024	2	5,5122	290,96	0,0000
(1)*(2)	6,619	10	0,6619	34,94	0,0000
Error	13,299	702	0,0189		
Total	172,42	719			

Table A.2: Summary of ANOVA conducted on MOS-LQOn's (P.563 [i.9]) in case of independent losses

Effect	SS	df	MS	F	p*
Packet loss (1)	57,65	5	11,5301	87,73	0,0000
Type of the signal (2)	71,775	2	35,8874	273,06	0,0000
(1)*(2)	5,925	10	0,5925	4,51	0,0000
Error	92,263	702	0,1314		
Total	227,613	719			

A.1.2 Dependent losses

In Tables 7 to 10, the results of ANOVA for the dependent losses test results and the all investigated *clp*'s (Dependent variable: MOS-LQOn (P.862) and MOS-LQOn (P.563)) are shown. More details about this can be found in clause 6.1.1.2.

Table A.3: Summary of ANOVA conducted on the MOS-LQOn's (P.862 [i.6]) in case of dependent losses (clp = 70 %)

Effect	SS	df	MS	F	p*
ulp (1)	175,701	5	35,1402	503,71	0,0000
Type of the signal (2)	9,712	2	4,8558	74,48	0,0000
(1)*(2)	9,89	10	0,989	14,18	0,0000
Error	48,974	702	0,0698		
Total	244,277	719			

Table A.4: Summary of ANOVA conducted on the MOS-LQOn's (P.862 [i.6]) in case of dependent losses (clp = 80 %)

Effect	SS	df	MS	F	p*
ulp (1)	174,971	5	34,9942	358,24	0,0000
Type of the signal (2)	14,551	2	7,2753	69,6	0,0000
(1)*(2)	13,649	10	1,3649	13,97	0,0000
Error	68,575	702	0,0977		
Total	271,745	719			

Table A.5: Summary of ANOVA conducted on the MOS-LQOn's (P.563 [i.9]) in case of dependent losses (clp = 70 %)

Effect	SS	df	MS	F	p*
ulp (1)	13,105	5	2,6211	23	0,0000
Type of the signal (2)	129,218	2	64,6089	494,78	0,0000
(1)*(2)	4,707	10	0,4707	3.6	0,0001
Error	91,667	702	0,1306		
Total	238,697	719			

Table A.6: Summary of ANOVA conducted on the MOS-LQOn's (P.563 [i.9]) in case of dependent losses (clp = 80 %)

Effect	SS	df	MS	*F*	*p**
ulp (1)	11,02	5	2,204	20,07	0,0000
Type of the signal (2)	135,982	2	67,9908	709,56	0,0000
(1)*(2)	2,832	10	0,2832	2,96	0,0012
Error	67,266	702	0,0958		
Total	217,1	719			

A.2 ANOVA for subjective results

In the next clauses, the detailed results of ANOVA conducted on MOS-LQSn for independent and dependent losses can be found.

A.2.1 Independent losses

Tables 11 and 12 provide the results of ANOVA carried out on the independent loss test results (Dependent variable: MOS-LQSn) described in more detail in clause 6.1.2.1.

Table A.7: Summary of ANOVA conducted on the MOS-LQSn's in case of independent losses

Effect	SS	df	MS	*F*	*p**
Packet loss (1)	388,73	5	77,747	99,31	0,0000
Type of the signal (2)	549,16	2	274,581	350,72	0,0000
(1)*(2)	35,75	10	3,575	4,57	0,0000
Error	1042,83	1332	0,783		
Total	2016,47	1349			

Table A.8: Summary of ANOVA conducted on the MOS-LQSn's in case of independent losses excluding diphone signal

Effect	SS	df	MS	*F*	*p**
Packet loss (1)	368,57	5	73,715	87,99	0,0000
Type of the signal (2)	35,2	1	35,204	42,02	0,0000
(1)*(2)	5,61	5	1,122	1,34	0,0455
Error	743,97	888	0,838		
Total	1153,36	899			

A.2.2 Dependent losses

Tables 13 and 14 show the results of ANOVA carried out on the dependent loss test results (Dependent variable: MOS-LQSn) described in more detail in clause 6.1.2.2.

Table A.9: Summary of ANOVA conducted on the MOS-LQSn's in case of dependent losses (clp = 80 %)

Effect	SS	df	MS	*F*	*p**
ulp (1)	427,06	5	85,413	74,77	0,0000
Type of the signal (2)	332,32	2	166,16	145,46	0,0000
(1)*(2)	76,38	10	7,638	6,69	0,0000
Error	1521,57	1332	1,142		
Total	2357,34	1349			

Table A.10: Summary of ANOVA conducted on the MOS-LQSn's in case of dependent losses (clp = 80 %) excluding diphone signal

Effect	SS	df	MS	*F*	*p**
ulp (1)	455,93	5	91,187	68,88	0,0000
Type of the signal (2)	7,84	1	7,840	5,92	0,0151
(1)*(2)	13,07	5	2,613	1,97	0,0801
Error	1175,52	888	1,324		
Total	1652,36	899			

History

Document history		
V1.1.1	April 2011	Publication