ETSI TR 102 948 V1.2.1 (2012-12)

Technical Report

Speech and multimedia Transmission Quality (STQ); Quality Assessment of Synthesized Speech

Reference

RTR/STQ-194

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

Important notice

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, please send your comment to one of the following services:
http://portal.etsi.org/chaircor/ETSI_support.asp

Copyright Notification

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://ipr.etsi.org).

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 Recommendations P.862 [i.6], P.863 [i.39] and P.563 [i.9] models, when both naturally-produced and synthesized speech are used. The variability of the predictions with respect to the type of signal (naturally-produced or synthesized) and the loss conditions was evaluated. Moreover, the accuracy of predictions provided by the investigated models was assessed by comparing these predictions with subjective assessments.

The results indicate one implication for designers of speech communication systems.

It has to be emphasized that none of the instrumental algorithms investigated here (P.862, P.863 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] 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.24] W. Jiang, H. Schulzrinne: "QoS measurement of Internet real-time multimedia services", Technical Report (CUCS-015-99), Columbia University (USA), December 1999.

[i.25] 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.26] 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.27] ITU-T Recommendation. P.830 (1996): "Subjective performance assessment of digital telephone-band and wideband digital codecs".

[i.28] 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.29] 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.30] 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.31] 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.32] 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.33] 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.34] 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.35] P. Počta, J. Holub: "Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models", In Acta Acustica united with Acustica, vol.97, pp.852-868, 2011, ISSN 1610-1928.

[i.36] 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.37] 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.38] ITU-T Recommendation P.56 (1993): "Objective measurement of active speech level".

[i.39] ITU-T Recommendation P.863 (2011): "Perceptual objective listening quality assessment".

[i.40] A. Kurittu: "Validation of ITU-T Recommendation P.563 Single-Ended Objective Speech Quality Measurement", in J. Audio Eng. Soc., vol. 54, pp. 1092-1098, 2006, ISSN 1549-4950.

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 (P.863)	MOS-LQOn predicted by P.863
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" and "P.863" model described in [i.6], [i.7], [i.8] and [i.39] or the standardized "P.563" model defined in [i.9] and [i.10]. The former ones are 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, P.863 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 the predictions with respect to the type of signal (naturally-produced or synthesized) and the loss conditions was evaluated. Moreover, the accuracy of predictions provided by the investigated models was assessed by comparing these 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 did not play any role in this experiment because of small constant jitter inserted by the loss simulator during the measurement.

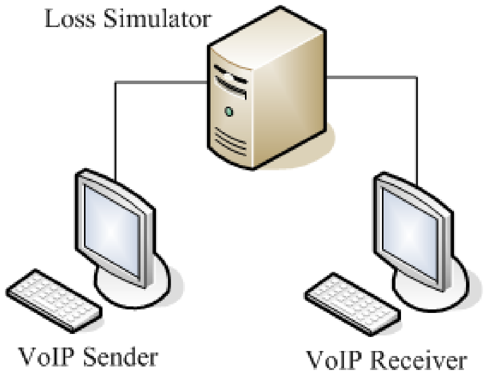


Figure 1: Experimental scenario

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] (bit rate: 8 kbps, frame size: 20 ms), ITU-T Recommendation G.711 [i.19] (64 kbps, 0,125 ms), GSM-FR (ETS 300 580-2 [i.20]) (13 kbps, 20 ms), Internet Low Bit Rate Codec (iLBC) [i.21] (15,2 kbps, 20 ms) and Speex [i.22] (4 kbps to 8 kbps (variable), 20 ms) but naturally speech quality assessment procedure was not changed and followed the description presented in clause 5.4 In all cases, the 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.23] and [i.24] 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:

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

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.23] and [i.25]. 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.26]:

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

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

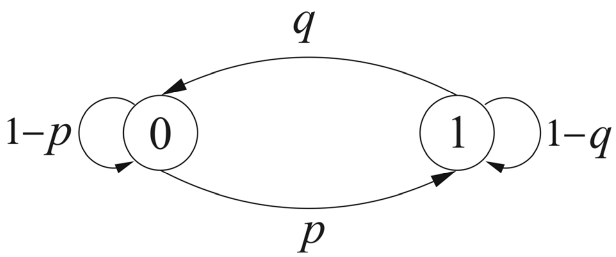


Figure 2: Gilbert model

Six independent loss and eleven dependent loss conditions were chosen to cover all 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 values of initial seed parameter (Initial seed parameter initializes the random number generator that the loss simulator uses to activate a loss generation process.) to simulate different loss locations/patterns in both cases. The same initial seed parameter values were used for all simulated loss conditions in this study in order to identically activate the loss generation process.

5.3 Reference signals

The reference signals selection should follow the criteria given by ITU-T Recommendations P.830 [i.27] 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.27], 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.28]. 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.29], 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.29], 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 (containing three sentences with different lengths uttered by one voice) were created. The text material used was the same for each voice used in this study. 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.30], 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.31], 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.38] 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.

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

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 (6 loss conditions * 3 samples representing each loss condition * 3 signals) for each type of loss investigated here. In particular, 3 samples representing each loss condition correspond to the best, average and worst speech quality obtained for a given loss condition. These samples were selected out of all recorded samples for each condition (40 samples per loss condition recorded (40 different loss patterns), see clause 5.2) 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 resulted 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 from coding experiment, namely samples belonged 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 from coding experiment (EVRC-B, GSM-FR and Speex) created 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 detail, respectively.

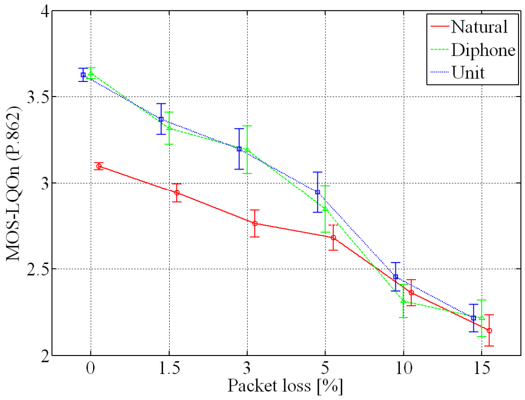
6.1 Impact of packet loss

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 detail.

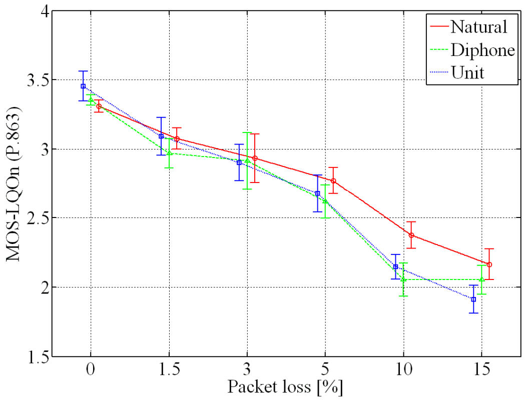
6.1.1.1 Independent losses

Using a Bernoulli model gives us the possibility to analyze P.862's, P.863'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 to 5 depict differences between investigated signal types in speech quality evaluation, provided by P.862, P.863 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 predicted by the investigated models (besides P.863 model). In particular, it can also be seen that P.862 and P.563 models provided 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 were detected by the models and the models decreased the score according to that. On the other hand, the models detected 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 codecs on final MOSn score (see clause 6.2) in respect to 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.



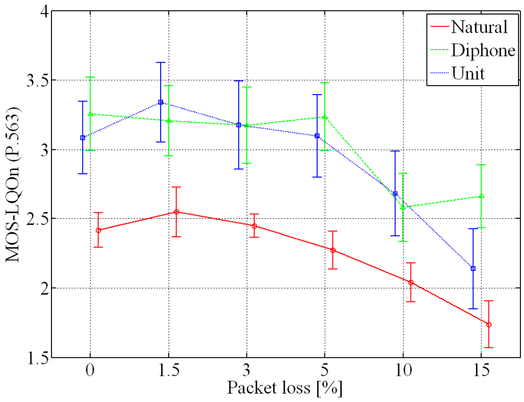
NOTE: The vertical bars show 95 % CI (derived from 40 measurements) for each loss and signal type.

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.863 [i.39] (MOS-LQOn (P.863)) 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 5: 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

In addition, the higher *MOS-LQOn* values (predicted by P.862 and P.563 models) of the synthesized signals obtained for 0 % packet loss resulted in a steeper slope for the *MOS-LQOn* curves representing synthesized speech. This steeper slope might be explained as higher vulnerability of this kind of speech to packet loss impairments. This assumption was tested in [i.35]. The results show that there is no evidence of higher sensitivity of the synthesized speech to packet loss (independent losses) from P.862 and P.563 predictions perspective. According to the additional analysis, the statement presented above is also valid for P.863 model.

However, it can be seen from Figure 4 that non-monotonic results have been obtained in case of P.563 model. At this moment, we do not have an explanation as to what could be the reason for such behavior. A detailed analysis of the P.563 model is needed to justify this behavior.

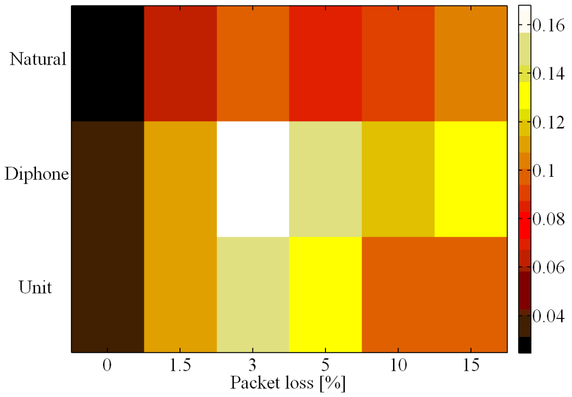


Figure 6: MAD of MOS-LQOn's predicted by P.862 [i.6] at each point of loss space and type of the signal used in case of independent losses

Figures 6 to 8 show MAD's of MOS-LQOn's (P.862), MOS-LQOn's (P.863) and MOS-LQOn's (P.563), which have been obtained from this experiment. It can be seen from Figures 6 to 8 that the deviations of predictions for naturally-produced speech are smaller than those for synthesized speech, especially for P.563 model.

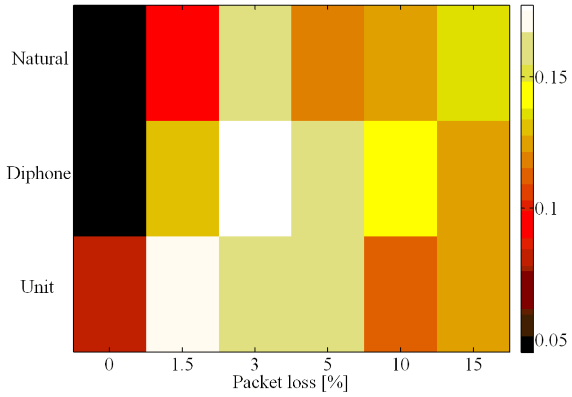


Figure 7: MAD of MOS-LQOn's predicted by P.863 [i.39] at each point of loss space and type of the signal used in case of independent losses

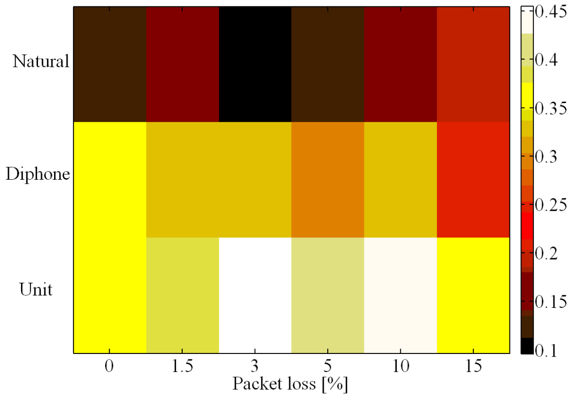


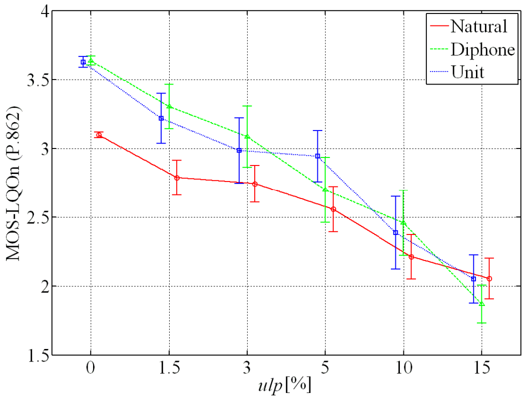
Figure 8: MAD of MOS-LQOn's predicted by P.563 [i.9] at each point of loss space and type of the signal used in case of independent losses

Two two-way analyses of variance (ANOVA) were conducted on MOS-LQOn's (P.862), MOS-LQOn's (P.863) and MOS-LQOn's (P.563) using packet loss and type of the signal as fixed factors (Tables A.1-3). The highest *F*-ratios for the packet loss (*F* = 1 493,55, *p** < 0,01) in the case of P.862 usage, for the packet loss again (*F* = 932,66, *p** < 0,01) in the case of P.863 usage and for the type of the signal used (*F* = 273,06, *p** < 0,01) in case of P.563 usage was determined. Moreover, the signal factor (MOS-LQOn's (P.862) and MOS-LQOn's (P.863)) and packet loss (MOS-LQOn (P.563)) appeared to have a weaker effect on quality than other mentioned factors for P.862, P.863 as well as P.563 based predictions, with *F* = 290,96, *p** < 0,01, *F* = 13,45, *p** < 0,01 and *F* = 87,73, *p** < 0,01, respectively. The realized *ANOVA* tests reveal that different factor 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 and P.863 models. It has to be emphasizes that P.563 model was built for monitoring the quality degradation produced by a 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, as reported in this experiment. In addition, it is worth to note that P.863 seems to be less sensitive to different voices involved in this analysis than P.862 and P.563 (lowest F-ratio obtained). This fact can also be clearly seen in Figure 4.

6.1.1.2 Dependent losses

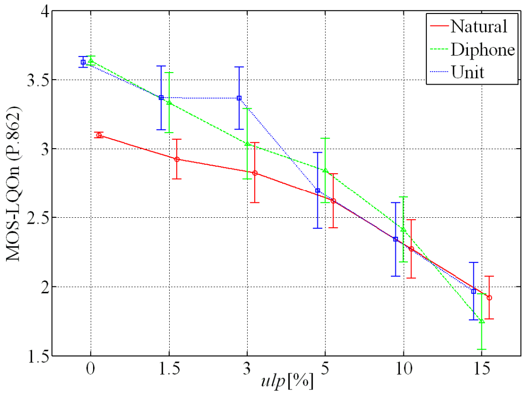
Using a Gilbert model extends the possibilities to investigate P.862's, P.863's, and P.563's behavior to three perspectives, namely ulp, clp and naturally type of the signal used. The experimental results for all investigated clp's are depicted in Figures 9 to 14. We can observe that the kind of signal used (naturally-produced or synthesized) could also seriously influence the quality in the case of dependent losses (again besides P.863 model). Regarding higher vulnerability of synthesized speech, the same effect as that in first case (independent losses) was evidently obtained, see more information in [i.35]. This means that there is no difference between the vulnerability of synthesized and naturally-produced speech to packet loss impairments (dependent losses) from the P.862 and P.563 predictions perspective. According to the additional analysis, the statement presented above is also valid for P.863 model.

Moreover, the higher burstiness of losses (expressed by clp parameter) leads to higher non-monotonicity of predictions provided by P.563 model (see Figures 13 and 14) than for independent losses. As mentioned above, a detailed investigation of the P.563 model is needed to rationalize this behavior.



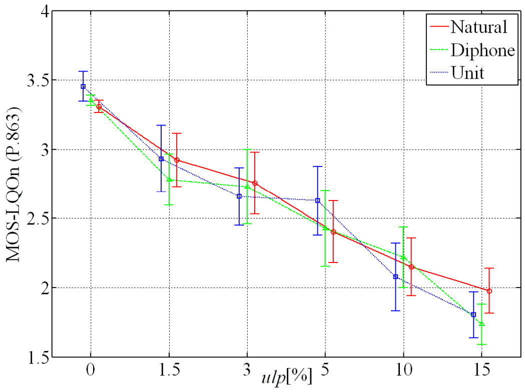
NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 9: 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 %)



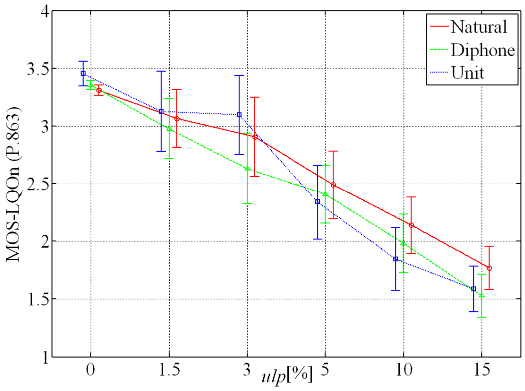
NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 10: 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 %)



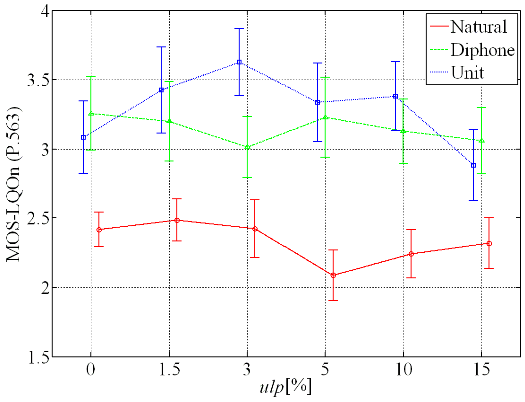
NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 11: MOS-LQOn predicted by P.863 [i.39] as a function of unconditional loss probability for different types of signal used in case of dependent losses (clp = 70 %)



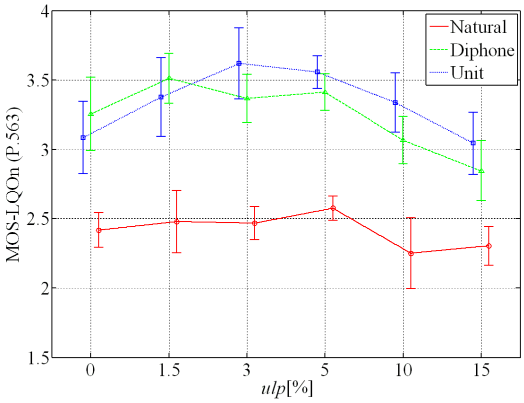
NOTE: Other detailed descriptions of Figure 3 apply appropriately.

Figure 12: MOS-LQOn predicted by P.863 [i.39] 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 13: 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 14: 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 15 to 17, the MAD of MOS-LQOn's (P.862), MOS-LQOn's (P.863) and MOS-LQOn's (P.563) for 70 % clp can be seen. Unsurprisingly, P.862's, P.863's and P.563's predictions deviation behavior is similar to that obtained in the previous case. Moreover, the *MAD* was increased for dependent losses and all voices but only in the case of *P.862* and P.863 predictions.

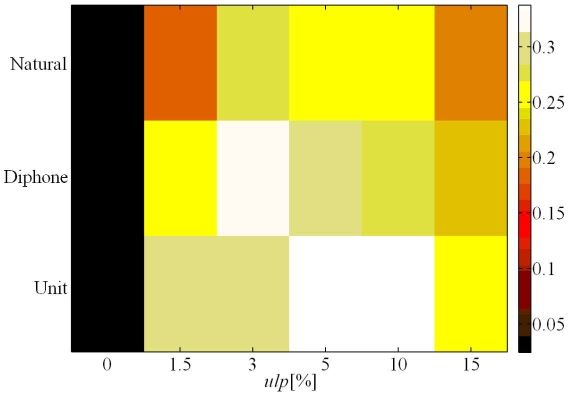


Figure 15: MAD of MOS-LQOn's predicted by P.862 [i.6] at each point of loss space and type of the signal used in case of dependent losses (clp = 70 %)

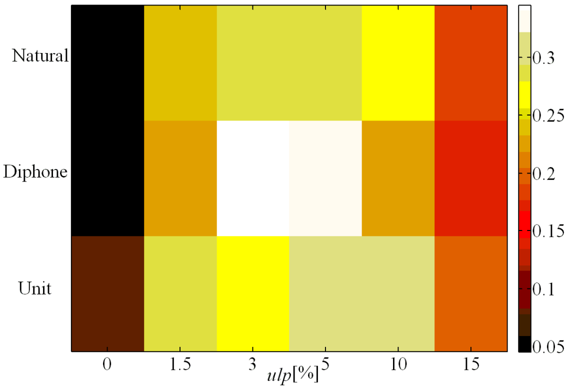


Figure 16: MAD of MOS-LQOn's predicted by P.863 [i.39] at each point of loss space and type of the signal used in case of dependent losses (clp = 70 %)

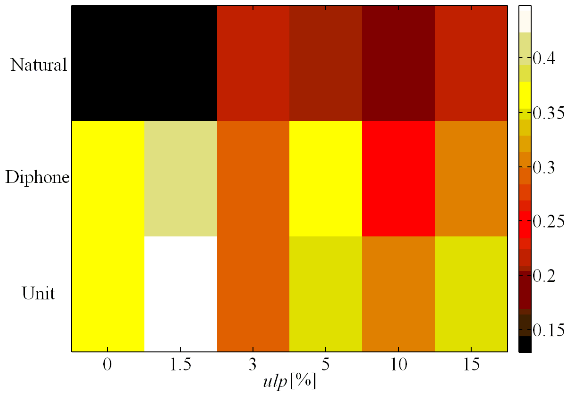


Figure 17: MAD of MOS-LQOn's predicted by P.563 [i.9] at each point of loss space and type of the signal used in case of dependent losses (clp = 70 %)

Similarly as for independent losses, six two-way ANOVA's were carried out on MOS-LQOn's (P.862), MOS-LQOn's (P.863) and MOS-LQOn's (P.563) for all investigated *clp*'s, using *ulp* and signal type as fixed factors (Tables A.4 to A.9). 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 the case of P.563, see Tables A.3, A.8 and A.9. Contrariwise, the loss impact (expressed by packet loss (independent losses) or *ulp* (dependent losses)) was decreased by higher burstiness in the case of P.862 and P.863 (see Tables A.1, A.4, A.5 and A.2, A.6, A.7) 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.32]. This may be done for example, by using a monotonic 3rd order mapping function. Such monotonic function (3rd order monotonic functions if possible) have been determined for each model 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.33] and [i.34]:

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

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.34]:

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

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 (non-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 non-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 18, 20 and 22 compare the MOS-LQSn values and the raw model predictions, namely MOS-LQOn (P.862), MOS-LQOn (P.863) 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, 0,8437 and 0,9178 for P.863 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. Regarding P.863, the correlation coefficient is higher for 'natural' signal than for unit and diphone type of synthesized speech. In addition, the smallest rmse and rmse* were obtained for synthesized speech generated by unit selection synthesizer. Similarly as in previous case, the correlation is higher for naturally-produced speech in the case of P.563 but interestingly the smallest rmse and rmse* have been again attained for 'unit' signal.

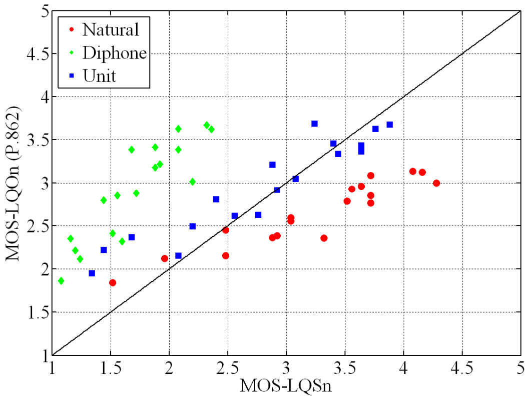


Figure 18: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (non-regressed)

On the other hand, Figures 19, 21 and 23 depict the subjective MOSn values (MOS-LQSn) and the regressed model predictions (MOS-LQOn (P.862), MOS-LQOn (P.863) and MOS-LQOn (P.563)). As attempted to use 3rd order regression (as mentioned above) has occasionally lead to non-monotonic results, the 1st order regression was used instead that finally led to monotonic results with acceptable accuracy of the final quality prediction as shown in Table 2. Table 2 also shows that the correlation coefficients belonging to P.863 model were only affected by the transformation. In addition, the root mean square errors and epsilon-insensitive root mean square errors are slightly reduced in some cases (see Table 2), after applying mapping functions.

Comparing the performance of the three investigated models, the P.862 model achieves the slightly higher correlations than P.863 and P.563 for all voices in this study. On the other hand, the P.863 attains the lowest root mean square errors and epsilon-insensitive root mean square errors for most of voices in this study.

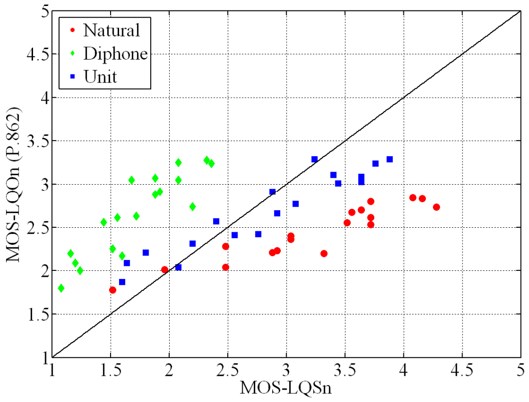


Figure 19: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (regressed)

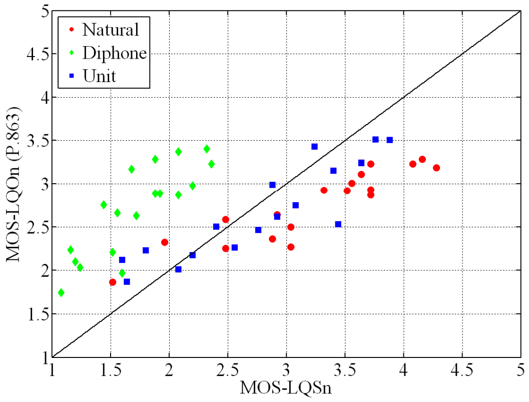


Figure 20: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for independent losses (non-regressed)

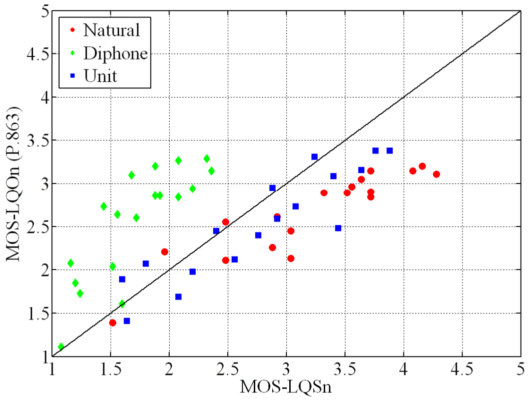


Figure 21: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for independent losses (regressed)

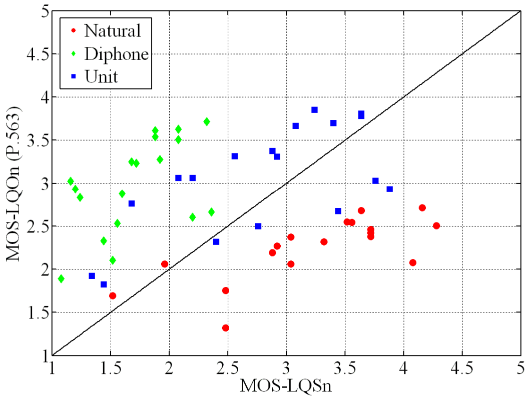


Figure 22: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (non-regressed)

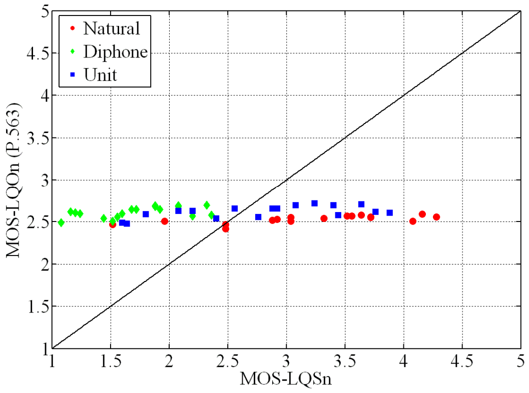


Figure 23: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (regressed)

However, it can be observed from Figure 23 that *P.563* compresses the *MOS-LQSn* range quite substantially. The samples have *MOS-LQSn* values ranging from about 1 to 4,3. The corresponding *MOS-LQOn (P.563)* range is from 2,5 to 2,7. This apparently helps to slightly decrease the root mean square error and epsilon-insensitive root mean square error for natural and diphone voice. Similar compression of the *MOS-LQSn* scale has been reported in [i.40] but for speech samples coded by the *AMR-NB* codec and containing radio channel errors. Finally, it should be noted that such predictions - despite the correlation values reported in Table 2 - are really meaningless.

To yet again prove if the synthesized speech has the same sensitivity to packet loss impairments (independent losses) from a *MOS-LQSn* perspective as natural speech, the same test as for MOS-LQOn values was performed (see details in [i.35]). The results indicate that from a MOS-LQSn perspective, packet loss has the same impact on naturally-produced speech as synthesized speech.

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]), MOS-LQOn (P.863 [i.39]) and 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.863	Natural	0,9178	0,1505	0,0914
	Diphone	0,8437	0,2452	0,1841
	Unit	0,8998	0,0878	0,0445
P.563	Natural	0,7356	0,2708	0,2064
	Diphone	0,5197	0,3251	0,2679
	Unit	0,6474	0,1480	0,0934

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]), MOS-LQOn (P.863 [i.39]) and MOS-LQOn (P.563 [i.9]) after regression for independent losses

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9366	0,2295	0,1658
	Diphone	0,8915	0,2399	0,1758
	Unit	0,9471	0,0904	0,0454
P.863	Natural	0,9136	0,1793	0,1139
	Diphone	0,8247	0,2444	0,1790
	Unit	0,9131	0,1065	0,0524
P.563	Natural	0,7356	0,2474	0,1887
	Diphone	0,5197	0,2421	0,1929
	Unit	0,6474	0,1774	0,1194

To specify the significance of the differences between the presented R, rmse and rmse* values for P.862, P.863 and P.563, statistical significance tests were performed. The results of such tests for independent losses are displayed in Table 3 (P.862 vs. P.563) and 4 (P.862 vs. P.863). Table 3 shows that most of the differences are statistically significant. It means that the models (P.862 and P.563) are statistically different in such cases. On the other hand, Table 4 shows that none of the differences are statistically significant. It means that the models (P.862 and P.863) are statistically equivalent in such cases.

Table 3: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.563 [i.9] for independent losses

	Before regression			After regression		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	1	1	1	1	0	0
Diphone	1	0	0	1	0	0
Unit	1	1	1	1	1	1
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

Table 4: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.863 [i.39] for independent losses

	Before regression			After regression		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	0	0	0	0	0	0
Diphone	0	0	0	0	0	0
Unit	0	0	0	0	0	0
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

Figures 24, 26 and 28 compare the MOS-LQSn values and the raw model predictions, namely MOS-LQOn (P.862), MOS-LQOn (P.863) and MOS-LQOn (P.563) when diphone signal has been excluded from the analysis. The corresponding correlations R and root mean square errors (rmse) and epsilon-insensitive root mean square errors (rmse*) are given in Table 5. Naturally, the correlation values, rmse and rmse* values obtained for natural and unit signal are same as reported above.

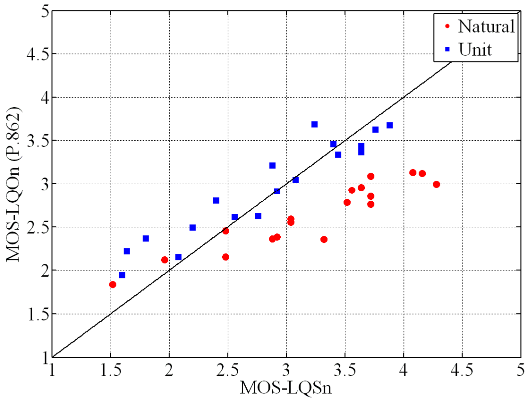


Figure 24: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (non-regressed) excluding diphone signal

On the other hand, Figures 25, 27 and 29 depict the subjective MOSn values (MOS-LQSn) and the regressed model predictions (MOS-LQOn (P.862), MOS-LQOn (P.863) and MOS-LQOn (P.563)) when diphone signal has been excluded from the analysis. As attempted to use 3rd order regression (as mentioned above) has lead to non-monotonic results, the 1st or 2nd order regression was used instead that finally led to monotonic results with acceptable accuracy of the final quality prediction as shown in Table 6. Table 6 also shows that the correlation coefficients belonging to P.863 and P.563 model were affected by the transformation. In addition, the root mean square errors and epsilon-insensitive root mean square errors are markedly reduced in most of the cases (see Table 6), after applying mapping functions.

However, it can be observed from Figure 29 that *P.563* still compresses the *MOS-LQSn* range. The samples have *MOS-LQSn* values ranging from about 1.5 to 4.3. The corresponding *MOS-LQOn (P.563)* range is from 2.6 to 3.7. In comparison to the previous case, the compression is not so strong. This apparently helps to decrease the root mean square error and epsilon-insensitive root mean square error markedly for natural voice and slightly for unit voice. Similar compression of the *MOS-LQSn* scale has been reported in [i.40] but for speech samples coded by the *AMR-NB* codec and containing radio channel errors. Finally, it should be noted that such predictions - despite the correlation values reported in Table 6 - are really meaningless.

Comparing the performance of the three investigated models, the P.862 model attains the slightly higher correlations than P.563 and P.863 for both voices involved in this part of the study before regression. Contrary to non-regressed situation, the P.563 model achieves the slightly higher correlations than P.863 and P.862 for all voices involved in this part of the study and for regressed data. On the other hand, the P.863 attains the lowest root mean square errors and epsilon-insensitive root mean square errors for most of the voices in this study.

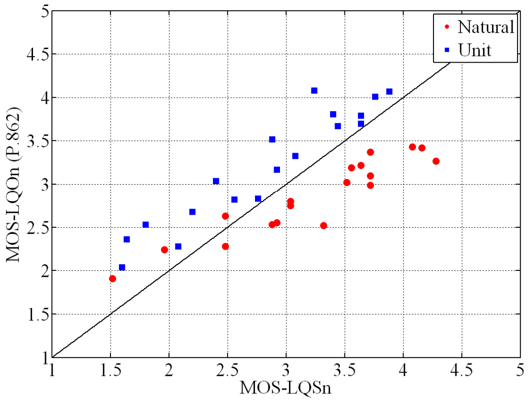


Figure 25: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (regressed) excluding diphone signal

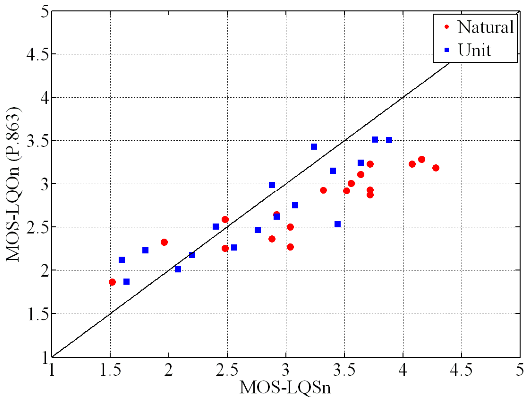


Figure 26: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for independent losses (non-regressed) excluding diphone signal

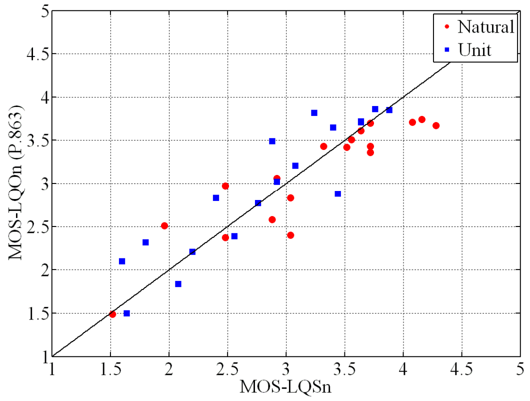


Figure 27: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for independent losses (regressed) excluding diphone signal

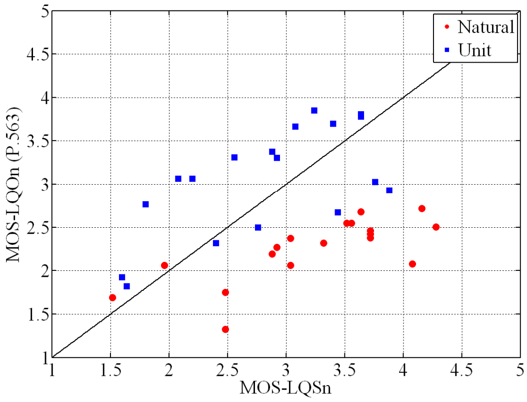


Figure 28: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (non-regressed) excluding diphone signal

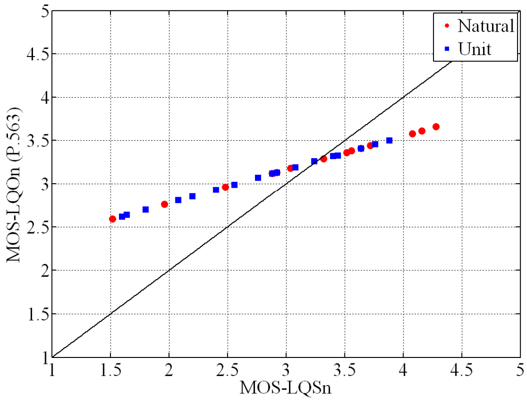


Figure 29: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (regressed) excluding diphone signal

Table 5: Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]), MOS-LQOn (P.863 [i.39]) and MOS-LQOn (P.563 [i.9]) before regression for independent losses excluding diphone signal

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9366	0,1740	0,1148
	Unit	0,9471	0,0712	0,0476
P.863	Natural	0,9178	0,1505	0,0914
	Unit	0,8998	0,0878	0,0445
P.563	Natural	0,7356	0,2708	0,2064
	Unit	0,6474	0,1480	0,0934

Table 6: Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]), MOS-LQOn (P.863 [i.39]) and MOS-LQOn (P.563 [i.9]) after regression for independent losses excluding diphone signal

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9366	0,1309	0,0741
	Unit	0,9471	0,1116	0,0648
P.863	Natural	0,9101	0,0866	0,0539
	Unit	0,9130	0,0852	0,0575
P.563	Natural	0,9999	0,1130	0,0726
	Unit	0,9999	0,1297	0,0903

To specify the significance of the differences between the presented R, rmse and rmse* values for P.862, P.863 and P.563 excluding diphone signal from the analysis, statistical significance tests were performed. The results of such tests for independent losses are displayed in Table 7 (P.862 vs. P.563) and 8 (P.862 vs. P.863). Table 7 shows that most of the differences are statistically significant. It means that the models (P.862 and P.563) are statistically different in such cases. On the other hand, Table 8 shows that only two of the differences are statistically significant. It means that the models (P.862 and P.863) are mostly statistically equivalent in such cases.

Table 7: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.563 [i.9] for independent losses excluding diphone signal

	Before regression			**After regression**		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	1	1	1	0	0	0
Unit	1	1	1	0	1	1
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

Table 8: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.863 [i.39] for independent losses excluding diphone signal

	Before regression			**After regression**		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	0	0	0	0	0	0
Unit	0	0	0	0	1	1
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

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.10). 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. This behavior is in line with P.563 behavior, as can be clearly seen in Table A.3 (clause A.1.1). Most 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 equal to or slightly higher than those impairments caused by independent losses and forced the listeners to change their opinions according to type of the signal used (different 'artificiality' dimensions of the investigated signals) and not according to amount of impairments heard from speech sample assessed. A diagnostic analysis of the test data exposed that this effect mainly occurred in the case of listeners without any 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). By 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.11. This supports our findings mentioned above.

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 30, 32 and 34 show the MOS-LQSn values and the raw model predictions for dependent losses, and Table 9 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 9). 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. Regarding P.863, the correlation coefficient is higher for 'unit' signal than for natural and diphone type of synthesized speech. In addition, the smallest rmse and rmse* were obtained for synthesized speech generated by unit selection synthesizer. On the contrary in the case of P.563, the correlation is higher for 'diphone' signal but interestingly the smallest rmse and rmse* were obtained for 'natural' signal.

The 3rd order regression as recommended in [i.32] led, in this case, to non-monotonic results. The 2nd order regression was used instead that finally led to monotonic results with an acceptable accuracy of the final quality prediction as shown in Table 10. The related scatter plots are depicted in Figures 31, 33 and 35. When transforming the *MOSn* predictions with the monotonic mapping function, the correlations increase rapidly for predictions provided by the *P.563* model and slightly for P.863 model and root mean square errors, epsilon-insensitive root mean square errors decrease in half of the cases. The corresponding values for R and $rmse$, $rmse^*$ are given in Table 10. The compression of *MOS-LQSn* as reported in the previous case has also been obtained here but not to such an extent as before. Currently the *MOS-LQSn* and *MOS-LQOn (P.563)* ranges are 1 to 4.3 and 2 to 2.8, respectively.

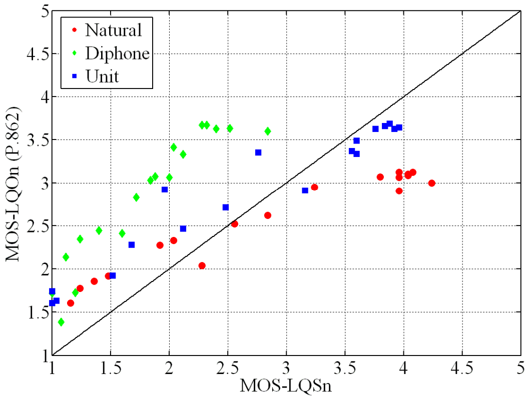


Figure 30: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for dependent losses (non-regressed)

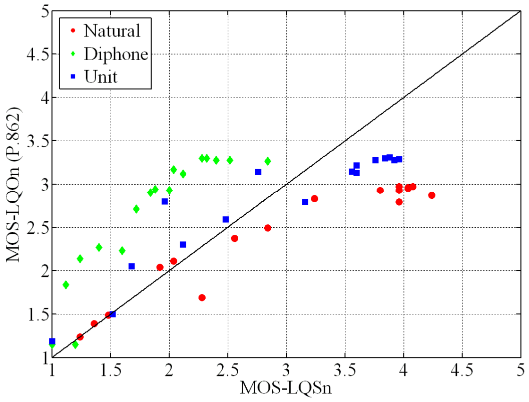


Figure 31: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for dependent losses (regressed)

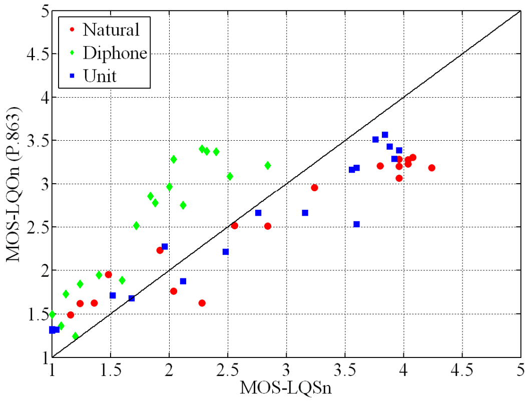


Figure 32: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for dependent losses (non-regressed)

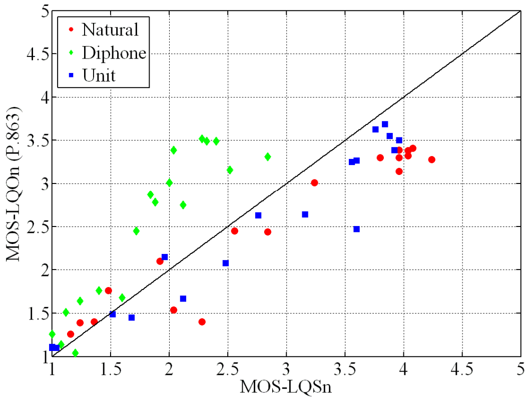


Figure 33: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for dependent losses (regressed)

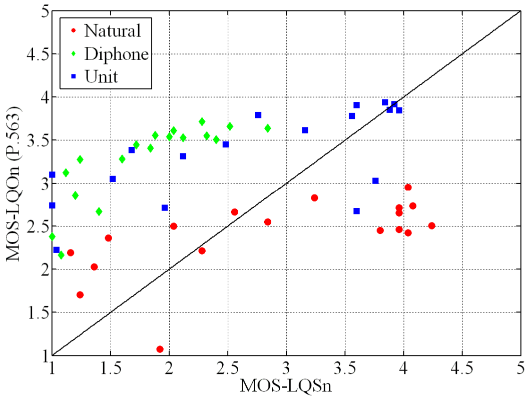


Figure 34: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for dependent losses (non-regressed)

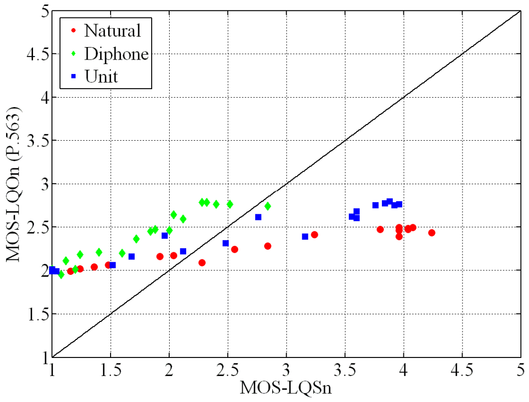


Figure 35: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for dependent losses (regressed)

Comparing the performance of the two investigated models, the P.863 model attains slightly higher correlations for all voices after regression. In addition, the smallest root mean square errors and epsilon-insensitive root mean square errors for all voices used in this study are mostly reported by the P.863 model.

To again define the significance of the differences between the presented *R*, *rmse* and *rmse** values for P.862, P.863 and P.563, statistical significance tests were performed. The results of such tests for dependent losses are displayed in Table 11 (P.862 vs. P.563) and 12 (P.862 vs. P.863). It can be seen from Table 11 that only one half of the differences is statistically significant. It means that the models (P.862 and P.563) are statistically different in such cases. On the other hand, Table 12 shows that none of the differences are statistically significant. It means that the models (P.862 and P.863) are statistically equivalent in the investigated cases.

Table 9: Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]), MOS-LQOn (P.863 [i.39]) and 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.863	Natural	0,9606	0,1455	0,0871
	Diphone	0,9213	0,1857	0,1286
	Unit	0,9673	0,1046	0,0640
P.563	Natural	0,6260	0,2535	0,1953
	Diphone	0,8114	0,3625	0,3060
	Unit	0,6751	0,2549	0,2255

Table 10: Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]), MOS-LQOn (P.863 [i.39]) and MOS-LQOn (P.563 [i.9]) after regression for dependent losses

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9583	0,1962	0,1394
	Diphone	0,8888	0,2174	0,1539
	Unit	0,9406	0,1131	0,0628
P.863	Natural	0,9620	0,1466	0,0848
	Diphone	0,9243	0,2080	0,1443
	Unit	0,9659	0,1071	0,0630
P.563	Natural	0,9766	0,2894	0,2267
	Diphone	0,9592	0,1710	0,1285
	Unit	0,9669	0,2207	0,1728

Table 11: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.563 [i.9] for dependent losses

	Before regression			After regression		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	1	0	1	0	0	1
Diphone	0	0	1	0	0	0
Unit	1	1	1	0	1	1
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

Table 12: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.863 [i.39] for dependent losses

	Before regression			After regression		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	0	0	0	0	0	0
Diphone	0	0	0	0	0	0
Unit	0	0	0	0	0	0
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

In order to once more demonstrate that the synthesized speech has the same sensitivity to packet loss impairments (dependent losses) from a *MOS-LQSn* perspective as natural speech, the same test as for independent losses (see [i.35] for details) was performed. The results again proves that there is no difference between the impact of dependent losses on naturally-produced speech and synthesized speech from a MOS-LQSn perspective.

Figures 36, 38 and 40 compare the MOS-LQSn values and the raw model predictions, namely MOS-LQOn (P.862), MOS-LQOn (P.863) and MOS-LQOn (P.563) when diphone signal has been excluded from the analysis. The corresponding correlations R and root mean square errors (rmse) and epsilon-insensitive root mean square errors (rmse*) are given in Table 13. Naturally, the correlation values, rmse and rmse* values obtained for natural and unit signal are same as reported above.

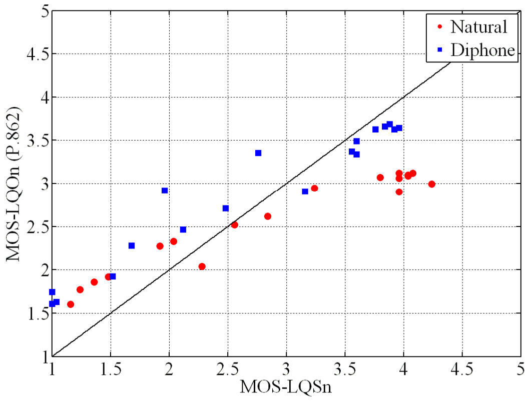


Figure 36: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (non-regressed) excluding diphone signal

On the other hand, Figures 37, 39 and 41 depict the subjective MOSn values (MOS-LQSn) and the regressed model predictions (MOS-LQOn (P.862), MOS-LQOn (P.863) and MOS-LQOn (P.563)) when diphone signal has been excluded from the analysis. As attempted to use 3rd order regression (as mentioned above) has lead to non-monotonic results, the 2nd order regression was used instead that finally led to monotonic results with acceptable accuracy of the final quality prediction as shown in Table 14. Table 14 also shows that all correlation coefficients were more or less affected by the transformation. In addition, the root mean square errors and epsilon-insensitive root mean square errors are markedly reduced in most of the cases (see Table 14), after applying mapping functions.

However, it can be observed from Figure 41 that *P.563* still compresses the *MOS-LQSn* range. The samples have *MOS-LQSn* values ranging from about 1.2 to 4.3. The corresponding *MOS-LQOn (P.563)* range is from 2.1 to 3.5. In comparison to the previous case, the compression is milder. Similar compression of the *MOS-LQSn* scale has been reported in [i.40] but for speech samples coded by the *AMR-NB* codec and containing radio channel errors. Finally, it should be noted that such predictions - despite the correlation values reported in Table 14 - are really meaningless.

Comparing the performance of the three investigated models, the P.863 model attains the slightly higher correlations than P.563 and P.862 for both voices involved in this part of the study before regression. Contrary to non-regressed situation, the P.862 model achieves the slightly higher correlations than P.563 and P.863 for all voices involved in this part of the study and for the regressed data. On the other hand, the P.863 attains the lowest root mean square errors and epsilon-insensitive root mean square errors for most of the voices in this study before and after regression.

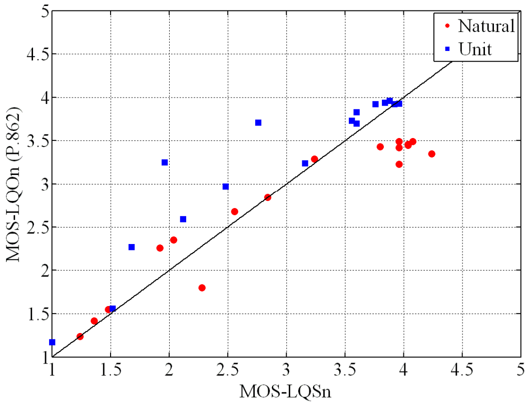


Figure 37: Subjective results (MOS-LQSn) versus MOS-LQOn (P.862 [i.6]) scores for independent losses (regressed) excluding diphone signal

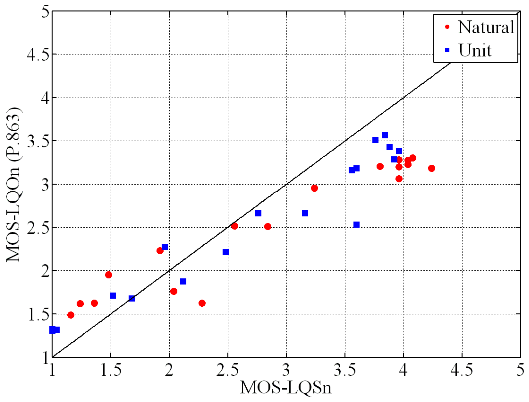


Figure 38: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for independent losses (non-regressed) excluding diphone signal

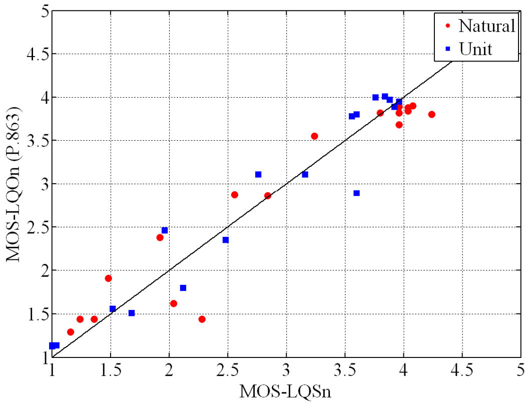


Figure 39: Subjective results (MOS-LQSn) versus MOS-LQOn (P.863 [i.39]) scores for independent losses (regressed) excluding diphone signal

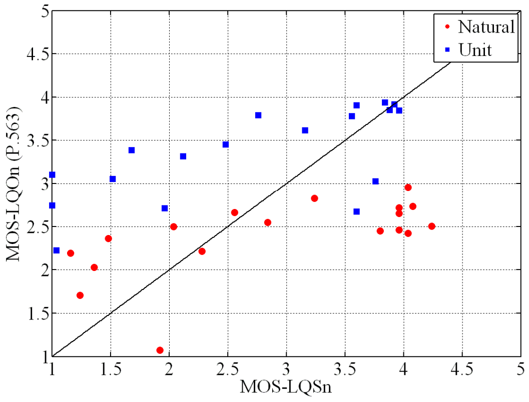


Figure 40: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (non-regressed) excluding diphone signal

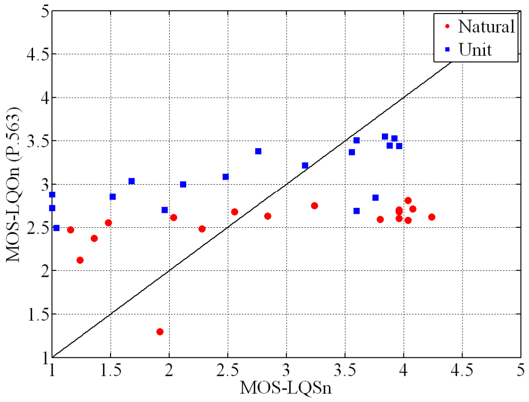


Figure 41: Subjective results (MOS-LQSn) versus MOS-LQOn (P.563 [i.9]) scores for independent losses (regressed) excluding diphone signal

Table 13: Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]), MOS-LQOn (P.863 [i.39]) and MOS-LQOn (P.563 [i.9]) before regression for independent losses excluding diphone signal

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9723	0,1690	0,1130
	Unit	0,9660	0,1099	0,0831
P.863	Natural	0,9606	0,1455	0,0871
	Unit	0,9673	0,1046	0,0640
P.563	Natural	0,6260	0,2535	0,1953
	Unit	0,6751	0,2549	0,2255

Table 14: Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (P.862 [i.6]), MOS-LQOn (P.863 [i.39]) and MOS-LQOn (P.563 [i.9]) after regression for independent losses excluding diphone signal

	Type of the signal	R	rmse	rmse*
P.862	Natural	0,9598	0,1152	0,0663
	Unit	0,9441	0,1155	0,0903
P.863	Natural	0,9575	0,0873	0,0561
	Unit	0,9716	0,0707	0,0470
P.563	Natural	0,5022	0,2846	0,2209
	Unit	0,7134	0,2572	0,2239

To specify the significance of the differences between the presented R, rmse and rmse* values for P.862, P.863 and P.563 excluding diphone signal from the analysis, statistical significance tests were performed. The results of such tests for dependent losses are displayed in Table 15 (P.862 vs. P.563) and 16 (P.862 vs. P.863). Table 15 shows that most of the differences are statistically significant. It means that the models (P.862 and P.563) are statistically different in such cases. On the other hand, Table 16 shows that none of the differences are statistically significant. It means that the models (P.862 and P.863) are statistically equivalent in such cases.

Table 15: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.563 [i.9] for independent losses excluding diphone signal

	Before regression			After regression		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	1	0	1	1	1	1
Unit	1	1	1	1	1	1
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

Table 16: Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors between P.862 [i.6] and P.863 [i.39] for independent losses excluding diphone signal

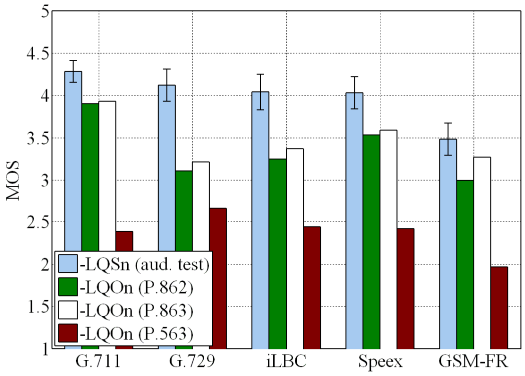
	Before regression			After regression		
Type of the signal	**R**	**rmse**	**rmse***	**R**	**rmse**	**rmse***
Natural	0	0	0	0	0	0
Unit	0	0	0	0	1	1
NOTE: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.						

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.12). 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 A.10 and A.12. 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.13. Moreover, the impact of the signal factor was considerably decreased in comparison to the previous case.

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 introduce 'artificiality' dimension, unnatural sounding whereas the ITU-T Recommendation G.711 [i.19] produce no perceptual degradation (natural sounding), (informal expert judgements).

Figures 42 to 44 show a fundamental difference in the quality judgements for natural speech and synthesized speeches provided by auditory test, P.862, P.863 and P.563, when processed by those codecs. In particular, a comparison of P.862, P.863 and P.563 predictions to the auditory MOSn values is shown in Figure 42 for naturally-produced speech. It is possible to see from the mentioned figure that 'artificially sounding' codecs are rated significantly worse in all models' predictions compared to the auditory test. For the ITU-T Recommendation G.711 codec (natural sounding codec), the predicted quality is in better agreement with the auditory results (at least for P.862 and P.863). Furthermore, the results provided by P.863 are closer to auditory ratings than the results provided by P.862 in all cases.

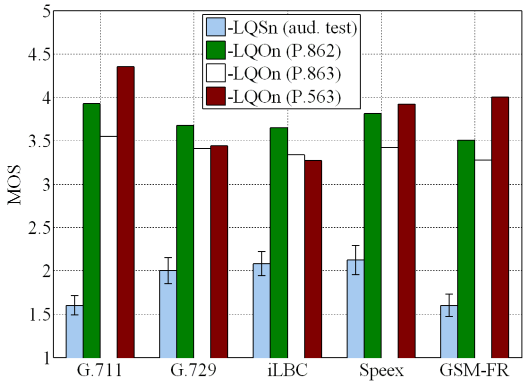


NOTE: The vertical bars show 95 % CI.

Figure 42: Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by P.862 [i.6], P.863 [i.39] as well as by P.563 [i.9] for naturally-produced speech

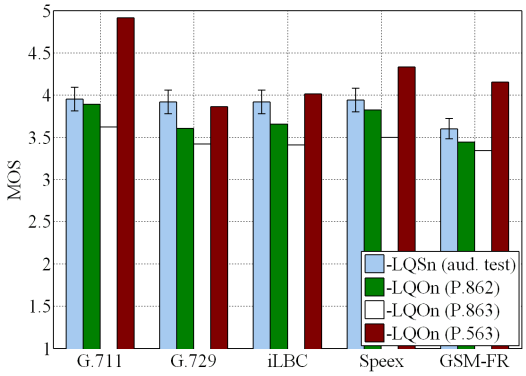
The picture is quite different for synthesized signals, see Figures 43 and 44. In Figure 43, the comparison of the auditory ratings with the predictions provided by the three investigated models for 'diphone' signal can be seen. As discussed above (see clause 6.1.2.1), 'diphone' signal (sounds less natural than 'unit' and 'natural' signals) was particularly disliked by test subjects. This is probably the reason for such small ratings provided by subjects. It appears that 'artificiality' dimension introduced by the diphone synthesizer might markedly prevail over coding impairments in this case. In general, it would be expected that its 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, Figures 9 to 12, clauses 6.1.1.1 and 6.1.1.2). On the basis of the presented facts, 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 serves as an example of how higher unnaturalness of the signal can affect the opinions of the test users. Figure 44 depicts the effect of the investigated codecs on MOS-LQSn and MOS-LQOn predicted by P.862, P.863 and P.563 models for 'unit' signal. In contrast to naturally-produced speech (see Figure 42), the predictions of all models are more or less 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*, a detailed investigation of this model is needed to determine the reasons for such a prediction.

Moreover, when comparing the behavior of the synthesized speech with the behavior of naturally-produced speech from auditory ratings perspective see Figure 45 (excluding 'diphone' signal from this comparison because this signal 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 (see unit signal in Figure 45) is insensitive to degradations introduced by the investigated codecs - except for GSM-FR codec - (almost the same *MOS-LQSn* values obtained for unit signal for almost all codecs investigated) because of higher degree of 'artificiality' dimension introduced by synthesizer than by the codecs. Regarding the GSM-FR codec behavior, it is probable that this codec introduces 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.12], 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 it should be noted that 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.12] mainly focuses on the different types of codecs and their combinations. This study can serve as an extension of the study published in [i.12].



NOTE: The vertical bars show 95 % CI.

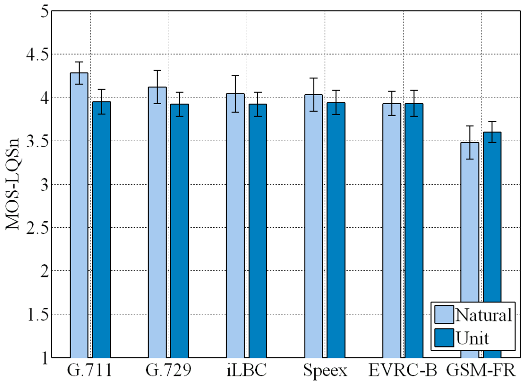
Figure 43: Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by P.862 [i.6], P.863 [i.39] as well as by P.563 [i.9] for synthesized speech generated by Diphone synthesizer



NOTE: The vertical bars show 95 % CI.

Figure 44: Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by P.862 [i.6], P.863 [i.39] as well as by P.563 [i.9] for synthesized speech generated by Unit selection synthesizer

In addition, it looks like both models have serious problems correctly predicting the quality of natural speech impaired by present 'unnatural sounding' codecs like ITU-T Recommendation G.729 [i.16] (they predict the quality slightly more pessimistically than was judged in the test), see Figure 42. The P.563 model is even more pessimistic than the P.862 model in this case. It should be noted that the test stimuli was composed of naturally-produced samples and a large amount of synthesized speech samples (two third of the test stimuli, see clause 5.5). One reason for such under-prediction of both investigated models reported here might be that the synthesized speech samples in the auditory test may have influenced the subjective ratings, in the sense that the large amount of synthesized data might have put the focus of the test subjects onto the 'artificiality' dimension and not only the impairments presented in the samples. This means that test subjects might have given the higher subjective ratings to natural samples (containing less artificiality than the rest of stimuli) than in the subjective test involving only naturally-produced stimuli. Naturally, the subjective ratings influenced in such way might cause the big differences between *MOS-LQSn* values and *MOS-LQOn* values predicted by both investigated models for naturally-produced speech, as reported here. Moreover, the problem with the diphone voice, as pointed out above (clause 6.1.2.1), supports this theory. It should be noted that the correlations reported for the loss experiment might have also been influenced by 'artificiality' dimension in a similar way as in the coding experiment because both kinds of samples (impaired by coding and packet loss) were mixed in the subjective test (see clause 5.5).



NOTE: The vertical bars show 95 % CI.

Figure 45: Comparison of the subjective ratings for naturally-produced speech with the ratings for synthesized speech generated by Unit selection synthesizer *in the case of coding impairments*

Comparing the performance of the three investigated models from coding impairments perspective, P.863 model out-performs P.563 and P.862 models, 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 two comparison-based (P.862 and P.863) and one single-ended (P.563) quality prediction models. The main goal of this study was to gain a better understanding of behavior of three 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.35], [i.36] and [i.37]. It has to be again emphasized that none of the instrumental models investigated here (P.862, P.863 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 and P.863 model seem to be in line with the auditory ratings. On the other hand, P.563 is less accurate than both comparison-based models (P.862 and P.863) and for some parameters provides meaningless predictions and therefore should not be applied. Finally, it is possible to pronounce that some models initially designed for predicting quality of natural speech are capable of predicting the quality of the transmitted synthesized speech under most of 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, only the coding impairments have a different impact on the quality of naturally-produced speech and synthesized speech. More precisely, the impact seems to depend on the perceptual type of degradation which is linked to the specific codec. An 'artificiality' dimension introduced by the investigated 'unnatural sounding' codecs is an additional degradation for the naturally-produced speech. This is not the case for the synthesized speech, which already carries a certain degree of artificiality. Moreover, the synthesized speech seems to be insensitive to most of the coding impairments investigated here. Comparison of all models seems to confirm that, the P.863 model copes best with both degradations investigated here (question 3).

8 Implication, advices to users and future work

The results have one potential implication for designers of telecommunications networks (speech communication systems) and providers of voice services based on TTS systems. As mentioned before, the results of the second experiment (coding impact) show that the synthesized speech seems to be insensitive to the coding impairments provided by present codecs, like ITU-T Recommendation G.729 [i.16], Speex, etc. If it will also be proven by other studies, the designers of speech communication systems and telecommunications networks and providers of voice services based on TTS systems will be allowed to select an arbitrary codec from the current codecs available in the market without any impact on the ultimate speech quality.

As discussed above, none of the investigated models was trained on synthesized speech. On the basis of this fact, one would expect that the models will not be able to predict a quality of synthesized speech transmitted over the phone (impaired by packet loss and coding) to a desirable degree. Contrary to this assumption, the results displayed in the present document have revealed that some investigated models are capable to predict the quality of transmitted synthesized speech under most of the investigated conditions to a certain degree. Some of them (P.862 and P.863) provide accurate predictions but. P.563 is less accurate and for some conditions provides meaningless results. On the basis of the findings described in the present document, it is not recommended to use P.563 in this context. Of course, there are still some limitations related to usage of synthesized speech in this context. Firstly, as can be seen in clause 6, all models have a big problem to predict a quality of diphone speech due to its high unnaturalness. Anyhow, it seems that P.863 can cope better with unnaturalness of this kind of synthesized speech than P.862 and P.563 (see Figure 43). If synthesized speech is generated by currently mostly deployed synthesis, namely unit-selection TTS system (naturalness of output generated by this kind of synthesis is very close to natural speech), the predictions provided by all models are more or less in good agreement with the auditory ratings (see Figure 44). Secondly, as can be seen above, P.563 model provides less accurate predictions than both comparison-based models. There is one assumption, why this model has a big problem with predicting the quality of synthesized speech impaired by packet loss and coding but it has to be proved by detailed analysis planned as a future work. The assumption is following: each type of synthesized speech regardless of type of synthesis used to generate the speech sometimes contains some sharp transitions or even small gaps between consecutive speech units in generated words, mainly unaudible to human being. Those gaps or sharp transitions can be considered as not properly concealed losses by prediction model. This fact posses big challenge (see the behavior of P.563 predictions above) to speech reconstruction and distortion detection modules commonly deployed in signal-based non-intrusive models to properly detect real amount of packet loss induced by network in the sample or to make proper internal reference (used as basis for assessment process).

Future work will focus on the following issues. Firstly, 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). Secondly, the extension of the E-model towards the synthesized speech impaired by the time-varying and coding impairments would be also desirable. Thirdly, a detailed analysis (using modified version of P.563 providing all internal parameters as output) of the P.563 model with regard to its non-monotonic predictions for packet loss as well as its higher predictions for some codecs, reported in this study would be performed.

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 A.1 and A.3 provide the results of ANOVA carried out on the independent losses test results (Dependent variable: MOS-LQOn (P.862), MOS-LQOn (P.863) 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.863 [i.39]) in case of independent losses

Effect	SS	df	MS	F	p*
Packet loss (1)	148,994	5	29,7987	932,66	0,0000
Type of the signal (2)	0,86	2	0,4299	13,45	0,0000
(1)*(2)	2,751	10	0,2751	8,61	0,0000
Error	22,429	702	0,032		
Total	175,034	719			

Table A.3: 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 A.4 to A.9, the results of ANOVA for the dependent losses test results and the all investigated *clp*'s (Dependent variable: MOS-LQOn (P.862), MOS-LQOn (P.863) and MOS-LQOn (P.563)) are shown. More details about this can be found in clause 6.1.1.2.

Table A.4: 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.5: 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.6: Summary of ANOVA conducted on the MOS-LQOn's (P.863 [i.39]) in case of dependent losses (clp = 70 %)

Effect	SS	df	MS	F	p*
ulp (1)	194,737	5	38,9474	414,65	0,0000
Type of the signal (2)	2,463	2	1,2316	13,11	0,0000
(1)*(2)	2,014	10	0,2014	2,14	0,0195
Error	65,937	702	0,0939		
Total	265,152	719			

Table A.7: Summary of ANOVA conducted on the MOS-LQOn's (P.863 [i.39]) in case of dependent losses (clp = 80 %)

Effect	SS	df	MS	F	p*
ulp (1)	213,247	5	42,6495	278,79	0,0000
Type of the signal (2)	0,538	2	0,2691	1,76	0,1730
(1)*(2)	7,533	10	0,7533	4,92	0,0000
Error	107,392	702	0,153		
Total	328,711	719			

Table A.8: 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.9: 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 A.10 and 11 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.10: 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.11: 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 A.12 and 13 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.12: 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.13: 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
V1.2.1	December 2012	Publication