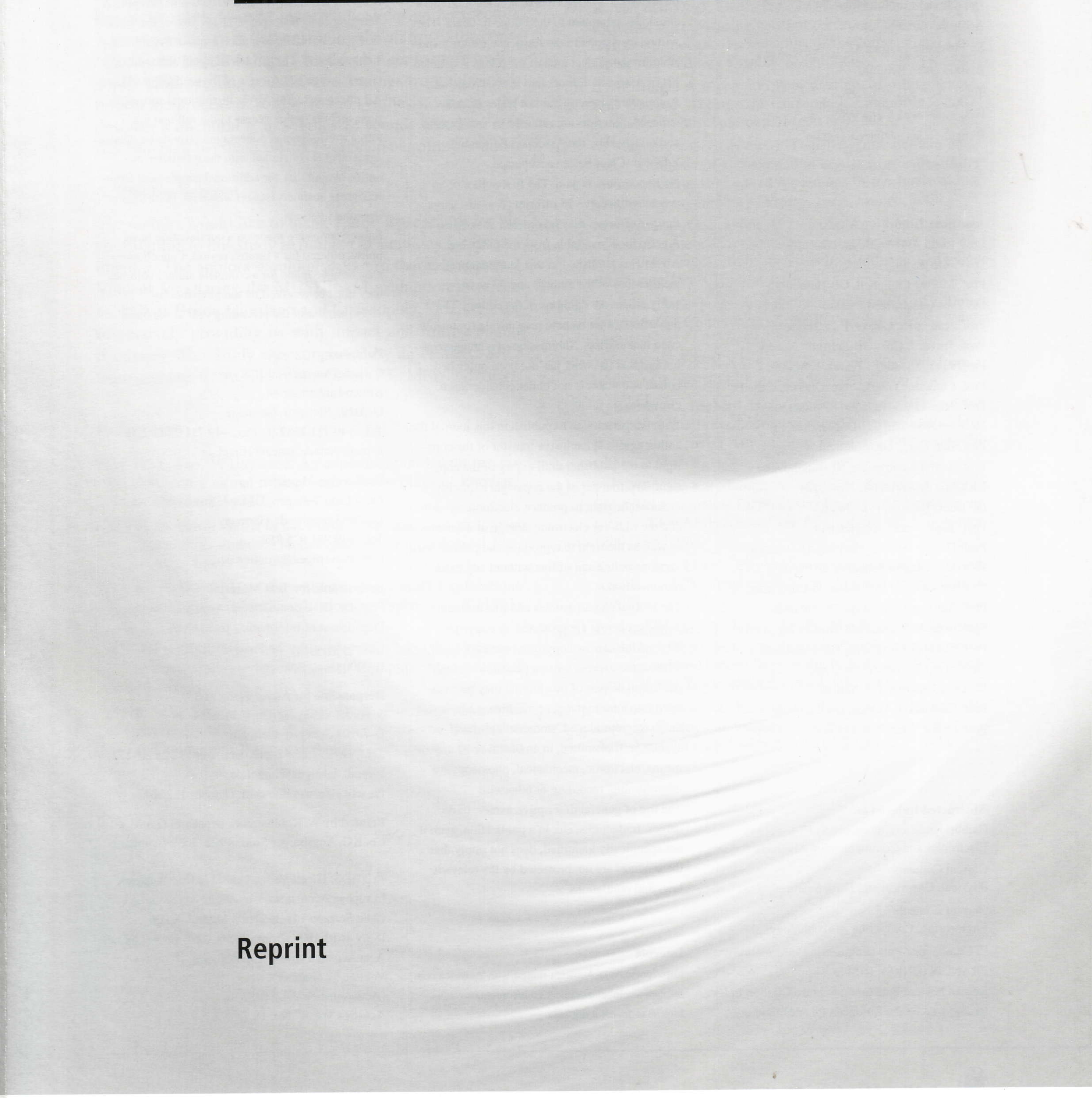
ISSN 1610-1928

Volume 97

ACTA ACUSTICA UNITED WITH ACUSTICA

The Journal of the European Acoustics Association (EAA) · International Journal on Acoustics



Reprint

S. Hirzel Verlag · P.O. Box 10 10 61 · D-70009 Stuttgart

Acta Acustica united with Acustica
Vol. 97 (2011) 852 – 868

DOI 10.3813/AAA.918465

Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models

Peter Počta[1], Jan Holub[2]

[1] Dept. of Telecommunications and Multimedia, FEE, University of Žilina, Univerzitná 1, 01026, Žilina, Slovakia. pocta@fel.uniza.sk

[2] Dept. of Measurement K13138, FEE, CTU Prague, Technická 2, 16627, Prague 6, Czech Republic. holubjan@fel.cvut.cz

Summary

This paper investigates the impact of independent and dependent losses and coding on speech quality predictions provided by PESQ (also known as ITU-T P.862) and P.563 models, when both naturally-produced and synthesized speech are used. Two synthesized speech samples generated with two different Text-to-Speech systems and one naturally-produced sample are investigated. In addition, we assess the variability of PESQ's and P.563's predictions with respect to the type of speech used (naturally-produced or synthesized) and loss conditions as well as their accuracy, by comparing the predictions with subjective assessments. The results show that there is no difference between the impact of packet loss on naturally-produced speech and synthesized speech. On the other hand, the impact of coding is different for the two types of stimuli. In addition, synthesized speech seems to be insensitive to degradations provided by most of the codecs investigated here. The reasons for those findings are particularly discussed. Finally, it is concluded that both models are capable of predicting the quality of transmitted synthesized speech under the investigated conditions to a certain degree. As expected, PESQ achieves the best performance over almost all of the investigated conditions.

PACS no. 43.71.Gv, 43.72.Gy, 43.72.Ja, 43.72.Kb

1. Introduction

In recent years, synthesized speech has reached a level of quality 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 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 developed. 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 significant 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 in order to determine the overall quality of the service achievable over the transmission network.

To determine the output quality of TTS systems (voice output devices), an application-oriented listening-only test described in ITU-T Recommendation P.85 [1] is recommended. 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 systems. 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 [2, 3, 4], it is still the most commonly

Received 5 August 2010,
accepted 29 July 2011.

used method for the overall assessment of the speech output of TTS systems but when such output is impaired by transmission degradations, a slightly modified version of this method (separating the ratings on the quality from the secondary task (collection of information (on what was understood by the subjects)) into two test sessions, see [5, 6] for details) or classical test according to ITU-T Recommendation P.800 [7] 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 the network. Such tools are available for predicting the quality of natural speech transmitted over telephone channels, e.g. the standardized "Perceptual Evaluation of Speech Quality" (PESQ) model described in [8, 9, 10], also known as ITU-T P.862 or the standardized "P.563" model defined in [11, 12]. The former one is belonging to the class of intrusive or comparison-based (full-reference) models, which are based on a comparison between the degraded output signal and the clean input signal of a 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 the quality. The difference is not calculated on the signal level but from an internal representation of the signals, consisting mainly of a non-linear frequency analysis and a loudness model. The latter is defined as a non-intrusive or single-ended model. The idea of a large class of such single-ended (reference-free) models is to generate an artificial reference (i.e., an "ideal" undistorted signal) from a 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 PESQ 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 to study the quality of synthesized speech over the phone and the performance of models for predicting and estimating the speech quality in the case of synthesized speech usage. In [5], two questions were addressed, namely whether the overall amount of degradation is similar for synthesized compared to naturally-produced speech, and how well 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 a controlled way, using a transmission simulation model. The samples were then judged 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 [6], the author extended the aforementioned work to new modelling examples with signal-based comparative measures, such as PESQ and Telekom Objective Speech Quality Assessment (TOSQA). The results show that both measures (PESQ and TOSQA) are capable of predicting the quality of transmitted synthesized speech to a certain degree. All models (i.e. the signal-based models and the 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 [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+) [14, 15], Psytechnics model, and P.563). Similar degradations to those introduced in [5] 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 the 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 [5]. That is the reason for an exhaustive investigation of their impact on the quality of synthesized speech. In particular, here we focus on an impact of independent and dependent losses and coding (focusing on current largely deployed codecs in these networks) on speech quality predictions provided by PESQ and P.563 in the case of naturally-produced and synthesized speech usage. Two synthesized speech samples generated with two different TTS systems and one naturally-produced sample are investigated. In addition, we assess the variability of PESQ's and P.563's predictions with respect to the type of speech used (naturally-produced or synthesized) and the loss conditions, as well as their accuracy, by comparing the predictions with subjective assessments. Finally, the aim of this study is three-fold: firstly, we would like to know whether the investigated models are able to provide valid predictions of perceived quality for the given application domain. Secondly, we would like to discover whether the impact of the packet loss and coding on the quality of synthesized speech is different from the impact on naturally-produced speech. Thirdly, we would like to find out which of the investigated modelling approaches is the most adequate one for the given task.

The rest of the paper is organized as follows. Section 2 introduces the experimental scenario and experiments carried out in this study. In section 3, the experimental results are presented and discussed. Section 4 concludes the paper and suggests some future studies.

2. Experiment description

2.1. Experimental scenario

A one-way VoIP session was established between two hosts (VoIP Sender and VoIP Receiver), via the loss simulator (Figure 1). In the case of the loss simulator, the two current most widely used models were deployed for the purpose of packet loss modeling, namely the Bernoulli and Gilbert loss models. More details about loss models can be found in section 2.2. For this experiment the ITU-T G.729AB encoding scheme [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 the small constant jitter inserted by the loss simulator during the measurement. The reference samples described in section 2.3 were utilized for transmission through the given VoIP connection. Finally, speech quality was assessed by the PESQ and P.563 algorithms. In the case of the PESQ model, the raw PESQ scores were converted to MOS-Listening Quality Objective narrow-band (MOS-LQOn) values by the equation defined in [17]. In the case of the PESQ and P.563 score calculations, some batch data processing techniques proposed in [18] were used.

For the coding experiment, the experimental scenario with the loss simulator and VoIP clients (VoIP Sender and Receiver) was replaced by coding algorithms, namely ITU-T G.729AB [16] (bit rate: 8 kbps, frame size: 20 ms), ITU-T G.711 [19] (64 kbps, 0.125 ms), GSM-FR (GSM 06.10) [20] (13 kbps, 20 ms), Internet Low Bit Rate Codec (iLBC) [21] (15.2 kbps, 20 ms), Speex [22] (4-8 kbps (variable), 20 ms) and Enhanced Variable Rate Codec version B (EVRC-B) [23] (9.6 kbps, 20 ms) but the speech quality assessment procedure was not changed and followed the aforementioned description. In the case of the EVRC-B codec, the noise suppression was disabled in comparison to the default settings. In the case of the other codecs, the default settings were applied.

2.2. Packet loss models

Packet loss is a major source of speech impairment in VoIP. Such a loss may be caused by dropped packets in IP networks (network loss) or by dropped packets at gateways/terminals due to late arrival (late loss).

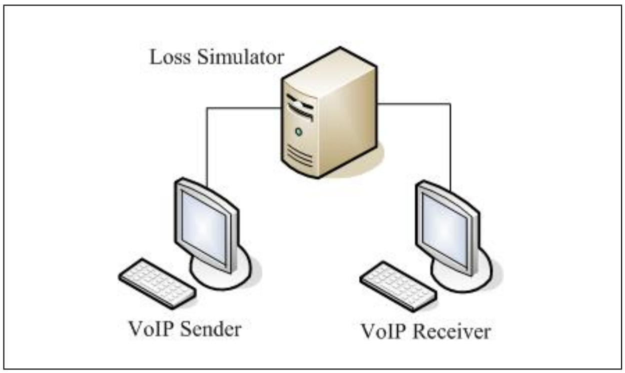


Figure 1. Experimental scenario.

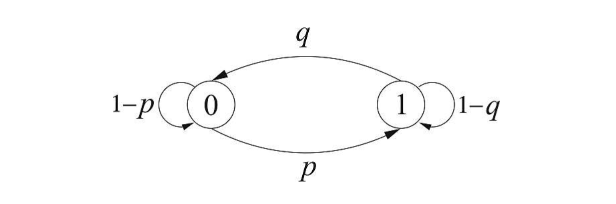


Figure 2. Gilbert model.

Several models [24, 25] have been proposed for modelling network losses, the current most widely adopted of those will be briefly discussed in the following subsections.

2.2.1. 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, the average packet loss rate, which is the number of lost packets divided by the total number of transmitted packets in a trace.

2.2.2. Gilbert model

Most research in VoIP uses a Gilbert model to represent packet loss characteristics [24, 25, 26]. In a 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 in [27]:

$$ulp = \frac{p}{q + p}. \tag{1}$$

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

Six independent loss and eleven dependent loss conditions were chosen to cover all scenarios of interest. They consisted of combinations of packet loss rate (from 0% to

15%) in the case of the independent losses and unconditional loss probability (ulp, 0%, 1.5%, 3%, 5%, 10% and 15%), conditional loss probability (clp, 70% and 80%) in the case of the dependent losses and 40 values of the initial seed parameter (The 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.

2.3. Speech material

For the purpose of this experiment and following the criteria given by ITU-T Recommendations P.830 [28] and P.800 [7], we defined three meaningful and non-technical sentences in Slovak with different lengths. In regards to those sentences, speech files were 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 a professional speaker. The decision to use a male voice was influenced by a previous study published in [29]. These tests 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 are most likely due to differences in synthesis quality between male and female voices.

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

Finally, three reference samples (namely Natural, Diphone and Unit) 12 seconds in length (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. As mentioned above, those speech samples were processed by transmission channel simulation (packet loss impact) as well as by codecs (coding impact), see the details in section 2.1 and 2.2. To avoid the differences in MOS values between the samples caused by the different perceptual impact of the same loss locations when the samples with dissimilar distributions of talkspurts are used [31], the same distributions and very similar durations of talkspurts (different talkers used) were deployed. Because the listening level has proven to be an important factor for the quality judgments of synthesized speech [32], all speech samples were 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 and stored in 16-bit, 8000 Hz linear PCM. Background noise was not present.

2.4. Subjective assessment

The subjective listening tests were performed in accordance with ITU-T Recommendation P.800 [7]. In every case, up to 9 listeners were seated in a listening chamber with a reverberation time less than 190 ms and background noise well below 20 dB SPL (A). All together, 25 listeners (11 male, 14 female, 21–30 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 a random order and dichotically presented (two loudspeakers, presentation level: 79 dB SPL(A)) to the test subjects. The results of the opinion scores from 1 (bad) to 5 (excellent) were averaged to obtain MOS-Listening Quality Subjective narrowband (MOS-LQSn) values for each sample.

Because of the big amount of very similar objective measurement data for dependent losses (clp = 70% and 80%), we had to make the decision as to which of the data set would be better to test in order to limit the number of samples used in subjective tests. In other words, which of the data sets representing dependent losses (speech samples impaired by clp = 70 or 80%) was more suitable to prove the behavior of the investigated models? In the end, we decided to use the second group of dependent losses, namely clp = 80% due to some effects related to higher burstiness of losses reported in section 3.1.1.2. Finally, the subjective tests were performed for independent losses and dependent losses with a 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 voices) 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 section 2.2) by expert listening. In addition to the loss experiment, we also realized a subjective test for the coding experiment, we investigated 6 current codecs (see section 2.1) which resulted in 18 samples (6 codecs * 3 voices) involved in this part of the subjective test. To ensure balanced sessions from impairment as well as size perspective, we combined samples from the coding experiment with the loss experiment as follows: all samples from the independent losses experiment (54 samples) and 9 samples from the coding experiment, namely samples belonging to ITU-T G.711, iLBC and ITU-T G.729 codecs (all together 63 samples) belonged to session No.1 and all samples from the dependent losses experiment (54 samples) and the rest of the samples from the coding experiment (EVRC-B, GSM-FR and Speex) belonged to session No.2 (containing 63 samples as well).

3. Experimental results

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

3.1. Impact of packet loss

3.1.1. 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 values of ulp and clp (dependent losses) and the same voice. The average MOS-LQOn scores, 95% Confidence Intervals (CI) and Mean Absolute Deviations (MAD) were calculated for both models under the study. The next subsections provide a detailed description of the experimental results for both examined types of losses.

3.1.1.1. *Independent losses*

Using a Bernoulli model gives us the possibility to analyze PESQ's and P.563's behavior from two perspectives, namely packet loss and voice (natural, diphone, and unit). Figures 3 and 5 depict differences between investigated voices in speech quality evaluation, provided by PESQ and P.563 respectively. It can be seen from the above-mentioned figures that the type of speech used (naturally-produced (Natural) or synthesized (Diphone and Unit)) has a significant impact on overall speech quality predicted by both investigated models. In particular, we can see that both models provided much higher MOS-LQOn values for synthesized voices, especially for 0% packet loss. This similar effect was obtained in [6]; see Figures 5.15 and 5.16. Unfortunately, the author did not specify the reason for this effect. Probably, this is due to some differences in 'artificiality' dimension between the naturally-produced and the synthesized speech coded by the ITU-T G.729 codec, which may be perceived as degradations by the models. In the case of synthesized speech, 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 voice and naturally considered that as a higher degradation. The reported behavior was also motivation for us to investigate the impact of other codecs (predominantly deployed codecs in current networks) on the final MOSn score (see section 3.2) in respect to objective as well as subjective assessments. Moreover, there is no difference between the investigated synthesized voices from this perspective because of similar 'artificiality' dimension introduced by both synthesizers.

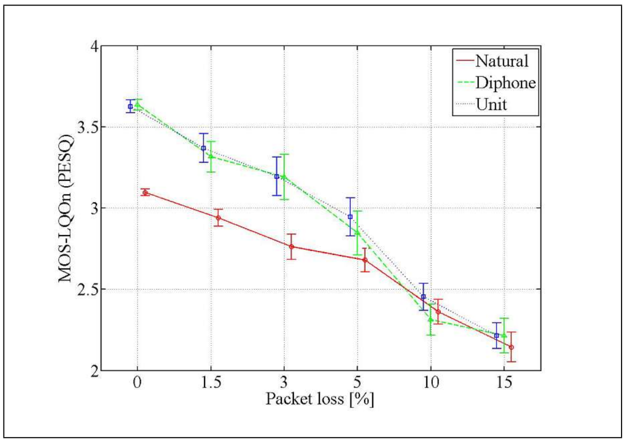


Figure 3. MOS-LQOn predicted by PESQ (MOS-LQOn (PESQ)) as a function of packet loss for individual voices in the case of independent losses. The vertical bars show 95 % CI (derived from 40 measurements) for each loss and voice.

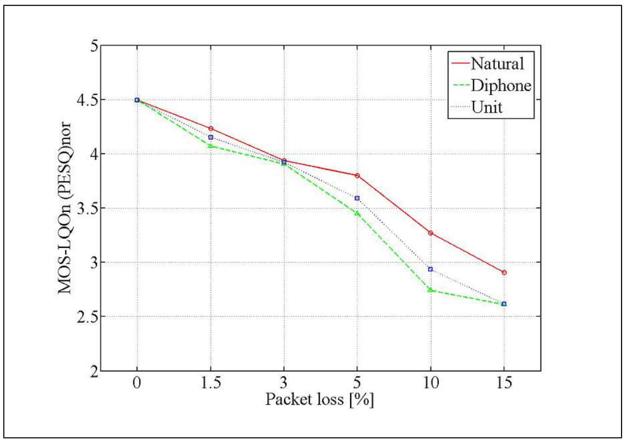


Figure 4. Normalized MOS-LQOn values predicted by PESQ (MOS-LQOn (PESQ)) for individual voices in the case of independent losses.

In addition, the higher MOS-LQOn values of the synthesized speech samples 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. To prove if the synthesized speech is really more prone to packet losses from PESQ and P.563 predictions perspective, we decided to normalize mean MOS-LQOn values to an optimum MOS-LQOn value according to formula 5.4.1 defined in [6]:

$$MOS_{\text{nor}} = (4.5 - 1)\frac{MOS - 1}{\text{topline} - 1} + 1. \tag{2}$$

The topline parameter is the MOS-LQOn value predicted by PESQ or by P.563 for the "clean" channel (namely 0%

Table I. Values of the topline parameter for the individual voices.

Prediction	Natural	Diphone	Unit
PESQ	3.10	3.64	3.63
P.563	2.42	3.26	3.09

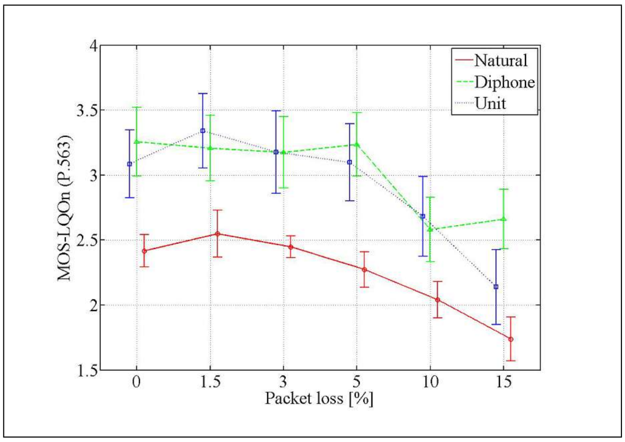


Figure 5. MOS-LQOn predicted by P.563 (MOS-LQOn (P.563)) as a function of packet loss for individual voices in the case of independent losses. Other detailed descriptions of Figure 3 apply appropriately.

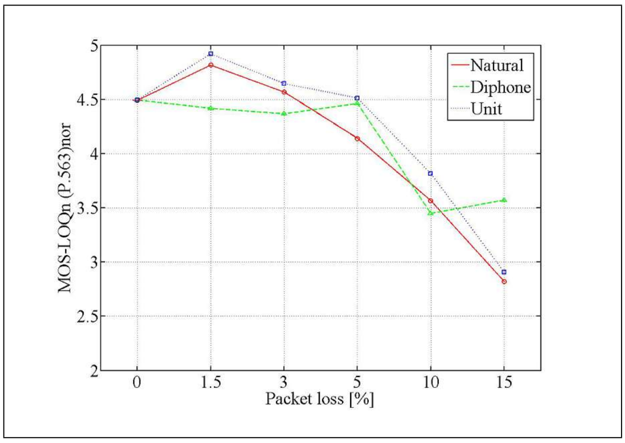


Figure 6. Normalized MOS-LQOn values predicted by P.563 (MOS-LQOn (P.563)) for individual voices in the case of independent losses.

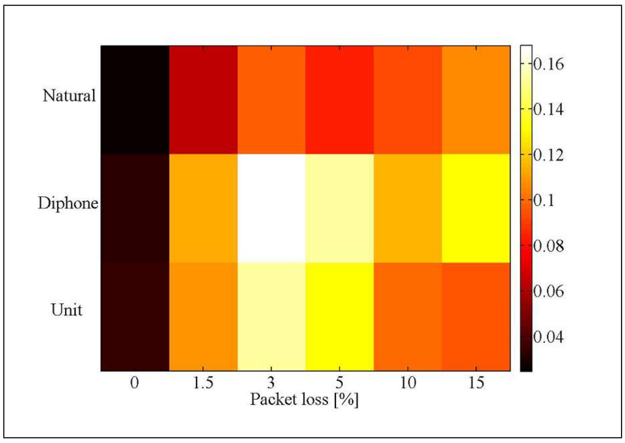


Figure 7. MAD of MOS-LQOn's predicted by PESQ at each point of loss space and for individual voices in the case of independent losses.

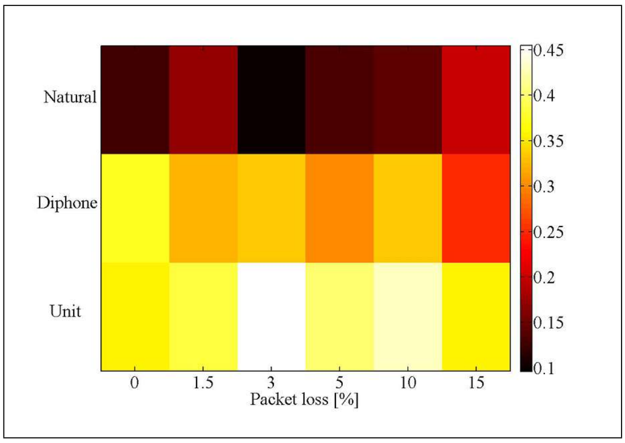


Figure 8. MAD of MOS-LQOn's predicted by P.563 at each point of loss space and for individual voices in the case of independent losses.

packet loss) with that voice. These values are given in Table I. Figures 4 and 6 show the normalized MOS-LQOn values for individual voices which have been calculated by the formula (2). It can be observed that the relative amount of degradation predicted by PESQ and P.563 due to packet loss is similar for the naturally-produced speech as for the synthesized speech (similar curves obtained). In other words, there is no evidence of higher sensitivity of the synthesized speech to packet loss (independent losses) from PESQ and P.563 predictions perspective. Moreover, a similar behavior is also expected when applying a different transformation from the one given in ITU-T Recommendation P.862.1 [17].

However, it can be seen from Figure 5 that non-monotonic results have been obtained in the case of the 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 7 and 8 show MAD's of MOS-LQOn's (PESQ) and MOS-LQOn's (P.563), which have been obtained from this experiment. It can be seen from Figures 7 and 8 that the deviations of predictions for naturally-produced speech are smaller than those for synthesized speech, especially for the P.563 model.

Two two-way analyses of variance (ANOVA) were conducted on MOS-LQOn's (PESQ) and MOS-LQOn's (P.563) using packet loss and voice as fixed factors (Appendix 5.1.1, Table VIII and IX). The highest F-ratio for the packet loss (F = 1493.55, $p* < 0.01$) in the case of PESQ usage and for the voice (F = 273.06, $p* < 0.01$) in the case of P.563 usage was determined. Moreover, the voice factor (MOS-LQOn's (PESQ)) and packet loss (MOS-LQOn (P.563)) appeared to a have a weaker effect on quality than other mentioned factors for PESQ as well as P.563 based predictions, with F = 290.96, $p* < 0.01$

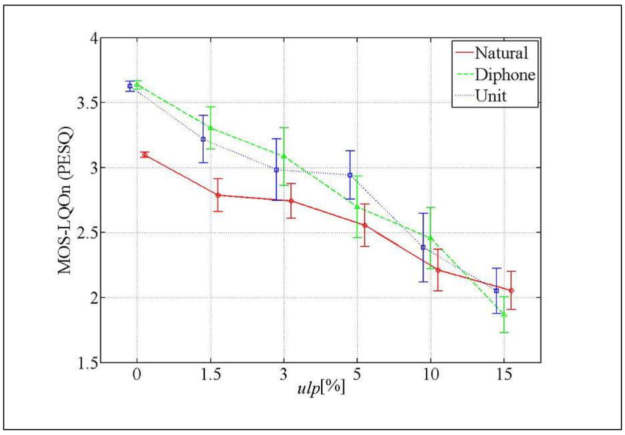


Figure 9. MOS-LQOn predicted by PESQ as a function of unconditional loss probability for individual voices in the case of dependent losses ($clp = 70\%$). Other detailed descriptions of Figure 3 apply appropriately.

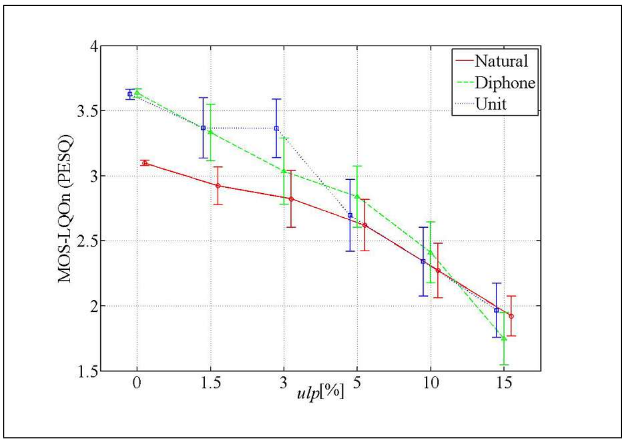


Figure 10. MOS-LQOn predicted by PESQ as a function of unconditional loss probability for individual voices in the case of dependent losses ($clp = 80\%$). Other detailed descriptions of Figure 3 apply appropriately.

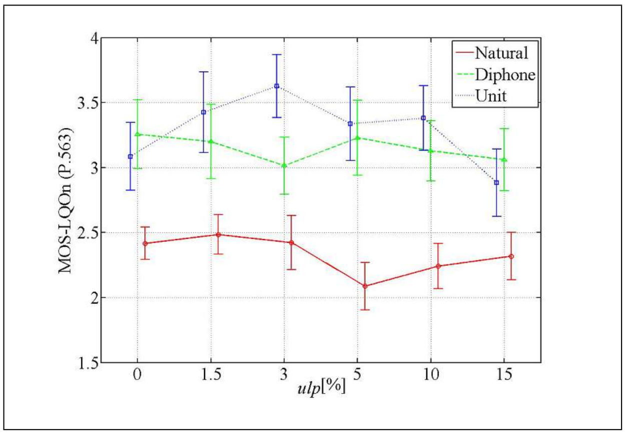


Figure 11. MOS-LQOn predicted by P.563 as a function of unconditional loss probability for individual voices in the case of dependent losses ($clp = 70\%$). Other detailed descriptions of Figure 3 apply appropriately.

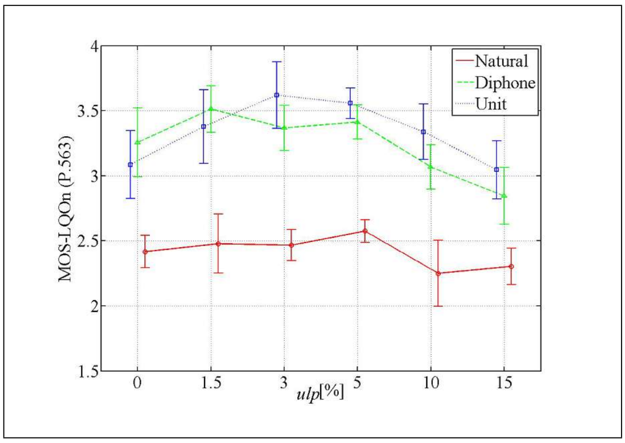


Figure 12. MOS-LQOn predicted by P.563 as a function of unconditional loss probability for individual voices in the case of dependent losses ($clp = 80\%$). Other detailed descriptions of Figure 3 apply appropriately.

and $F = 87.73$, $p* < 0.01$, respectively. The ANOVA tests reveal that a different factor affected the average MOS-LQOn values for each model investigated. In particular, the P.563 model seems to be more sensitive to voice than PESQ. It has to be emphasized that the 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 a big impact of voice factor on P.563's predictions, as observed in this experiment.

3.1.1.2. *Dependent losses*

Using a Gilbert model makes it possible to investigate PESQ's and P.563's behavior from three perspectives, namely ulp, clp and voice. The experimental results for all investigated clp's are depicted in Figures 9–12. We can observe that the kind of speech used (naturally-produced or synthesized) could also seriously influence the quality in the case of dependent losses. Similarly as in the previous case, the normalization to optimum MOS-LQOn has been performed. For reasons of similarity, we refrained from including the figures displaying the normalized MOS-LQOn curves for dependent losses. The same effect as that in first case (independent losses) was evidently obtained. This means that there is no difference between the vulnerability of synthesized and naturally-produced speech to packet loss impairments (dependent losses) from the PESQ and P.563 predictions perspective.

Moreover, the higher burstiness of losses (expressed by the clp parameter) leads to higher non-monotonicity of predictions provided by the P.563 model (see Figures 11–12) than for independent losses. As mentioned above, a detailed investigation of the P.563 model is required to rationalize this behavior.

In Figures 13–14, we can see the MAD of MOS-LQOn's (PESQ) and MOS-LQOn's (P.563) for a 70% clp. Unsurprisingly, PESQ's and P.563's predictions deviation behavior is similar to that obtained in the previous case.

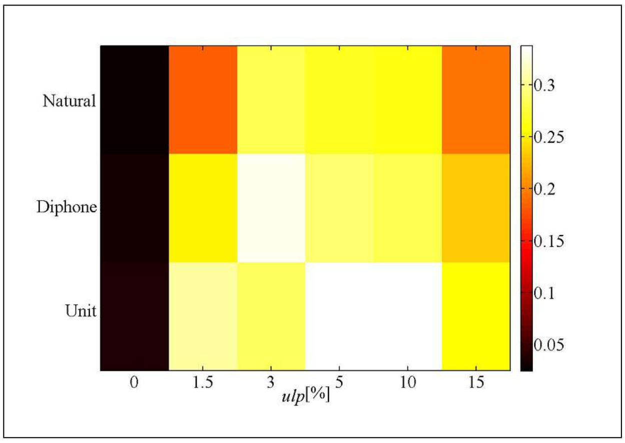


Figure 13. MAD of MOS-LQOn’s predicted by PESQ at each point of loss space and for individual voices in the case of dependent losses (clp = 70 %).

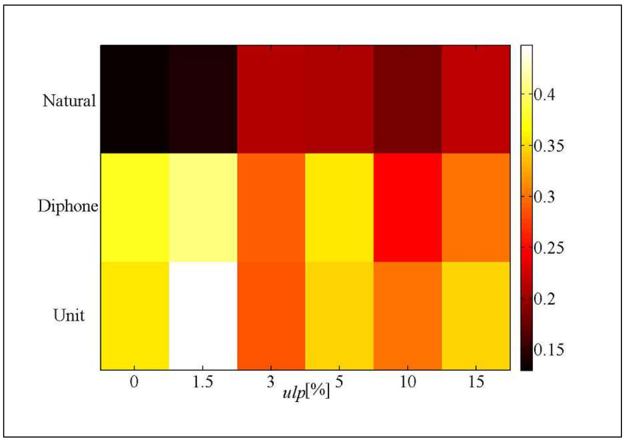


Figure 14. MAD of MOS-LQOn’s predicted by P.563 at each point of loss space and for individual voices in the case of dependent losses (clp = 70%).

Moreover, the MAD was increased for dependent losses and all voices but only in the case of PESQ predictions.

Likewise as for independent losses, four two-way ANOVA’s were carried out on MOS-LQOn’s (PESQ) and MOS-LQOn’s (P.563) for all investigated clp’s, using the ulp and the voice as fixed factors (Appendix 5.1.2, Tables X–XIII). In principle, we obtained similar results as for independent losses. However, a higher impact of voice (expressed by the 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 IX and XII–XIII. Contrariwise, the loss impact (expressed by packet loss (independent losses) or ulp (dependent losses)) was decreased by higher burstiness in the case of PESQ (see Tables VIII and X–XI) but still remains the most influencing factor.

3.1.2. Comparison between auditory and predicted quality scores

In the following subsections, auditory MOS values (MOS-LQSn) will be compared to the predictions of the two investigated models, namely intrusive PESQ and non-intrusive P.563. The comparison will be performed for all experimental conditions (independent and dependent losses), i.e. all combinations of voice (source speech) 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 a clp = 80% in this case due to similarities in the results obtained for both types of dependent loss conditions, as described in section 2.4. 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 a range of conditions that are part of the respective test [33]. This may be done, for example, by using a monotonic 3-rd order mapping function. Such monotonic functions (3-rd 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 the Pearson correlation coefficient R, the respective root mean square error ($rmse$) and epsilon-insensitive root mean square error ($rmse*$) as follows [34, 35]:

$$R = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}} \qquad (3)$$

and

$$rmse = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}\left(X_i - Y_i\right)^2}, \qquad (4)$$

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 the case of 3-order mapping function, $d = 1$ in the case of no regression). On the other hand, the epsilon-insensitive root mean square error can be described as

$$Perror_i = \max\left(0, |X_i - Y_i| - ci_{95_i}\right), \qquad (5)$$

where the ci_{95_i} represents the 95% confidence interval and is defined by [35]

$$ci_{95_i} = t(0.05, M)\frac{\delta_i}{\sqrt{M}}, \qquad (6)$$

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 the $Perror$ with the formula (5)

$$rmse* = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N} Perror_i^2}. \qquad (7)$$

Table II. Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (PESQ) as well as MOS-LQOn (P.563) before regression for independent losses.

	Voice	R	$rmse$	$rmse*$
PESQ	Natural	0.9366	0.1740	0.1148
	Diphone	0.8915	0.2959	0.2320
	Unit	0.9471	0.0712	0.0476
P.563	Natural	0.7356	0.2708	0.2064
	Diphone	0.5197	0.3251	0.2679
	Unit	0.6474	0.1480	0.0934

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 a similar measure to 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 the 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 voices, in order to get an indication of the characteristics of the individual models on different types of source data. To also provide the information on the significance of the differences between presented R, $rmse$ and $rmse*$ values for the PESQ and the P.563 models, the statistical significance tests according to [36] will be performed.

3.1.2.1. *Independent losses*

Initially, it should be noted that 95% confidence intervals for MOS-LQSn values presented in this comparison computed according to (6) were on average 0.2955 MOS (for Natural voice), 0.2625 MOS (for Diphone voice), and 0.2847 MOS (for Unit voice). Figures 15 and 17 compare the MOS-LQSn values and the raw model predictions, namely MOS-LQOn (PESQ) 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 II. The correlation calculated over all test conditions varies between 0.8915 and 0.9471 for the PESQ and 0.5197 and 0.7356 for the P.563 model (see Table II). For PESQ, the correlation coefficient is higher for unit voice (synthesized speech generated by unit selection synthesizer) than for naturally-produced voice and the diphone type of synthesized speech. Moreover, the smallest $rmse$ and $rmse*$ were also obtained for synthesized speech generated by the unit selection synthesizer. Contrariwise in the case of P.563, the correlation is higher for naturally-produced speech but interestingly the smallest $rmse$ and $rmse*$ were attained for unit voice.

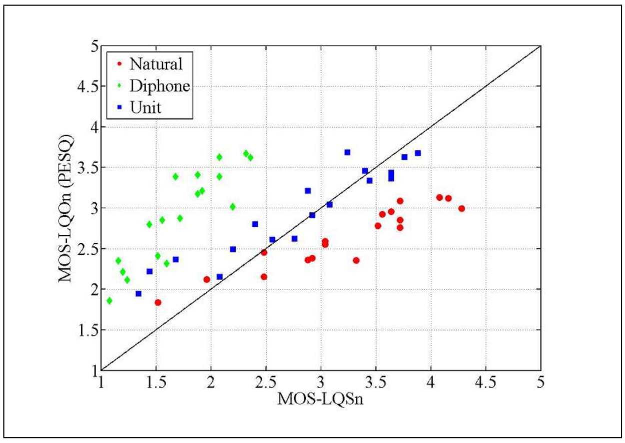


Figure 15. Subjective results (MOS-LQSn) versus MOS-LQOn (PESQ) scores for independent losses (non-regressed).

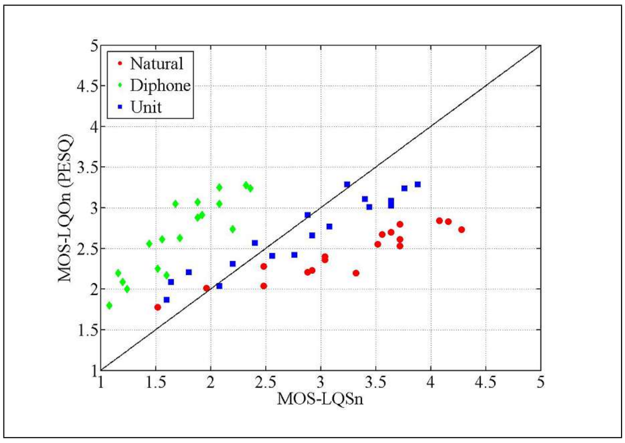


Figure 16. Subjective results (MOS-LQSn) versus MOS-LQOn (PESQ) scores for independent losses (regressed).

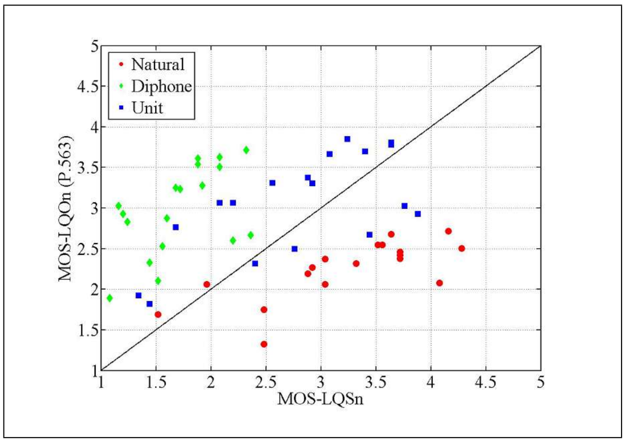


Figure 17. Subjective results (MOS-LQSn) versus MOS-LQOn (P.563) scores for independent losses (non-regressed).

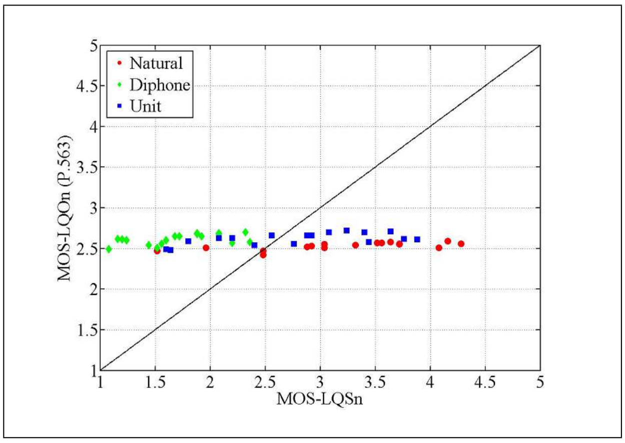


Figure 18. Subjective results (MOS-LQSn) versus MOS-LQOn (P.563) scores for independent losses (regressed).

On the other hand, Figures 16 and 18 depict the subjective MOSn values (MOS-LQSn) and the regressed model predictions (MOS-LQOn (PESQ), MOS-LQOn (P.563)). The 3-rd order regression as recommended in [33] leads, in this case, to non-monotonic results. There are several options available to assure final regression monotonicity in such cases (e.g. outliers influence weighting, polynomial order change or non-polynomial function regression). To stick to common polynomial regression and to avoid sometimes questionable outlier penalization, we choose the 1-st order polynomial regression that finally led to monotonic results with acceptable accuracy of the final quality prediction as shown in Table III. Table III also shows that the correlation coefficients were not affected by this transformation. In addition, the root mean square errors and epsilon-insensitive root mean square errors are slightly reduced in some cases (see Table III), after applying mapping functions.

Comparing the performance of the two investigated models, the PESQ model achieves the highest correlations and lowest root mean square errors and epsilon-insensitive root mean square errors for all voices in this study, as expected.

However, it can be observed from Figure 18 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 [37] 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 III - are really meaningless.

To specify the significance of the differences between the presented R, $rmse$ and $rmse*$ values for PESQ and P.563, statistical significance tests were performed. The results of such tests for independent losses are displayed in Table IV. Table IV shows that most of the differences are statistically significant. It means that the models are statistically different in such cases.

Table III. Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (PESQ) as well as MOS-LQOn (P.563) after regression for independent losses.

	Voice	R	$rmse$	$rmse*$
PESQ	Natural	0.9366	0.2295	0.1658
	Diphone	0.8915	0.2399	0.1758
	Unit	0.9471	0.0904	0.0454
P.563	Natural	0.7356	0.2474	0.1887
	Diphone	0.5197	0.2421	0.1929
	Unit	0.6474	0.1774	0.1194

Table IV. Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors for independent losses. Note: “1” indicates that the difference is statistically significant. “0” indicates that the difference is not statistically significant.

	Before regression			After regression		
Voice	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

One two-way ANOVA was conducted on MOS-LQSn’s using packet loss and voice as fixed factors (Appendix 5.2.1, Table XIV). We found that the highest F-ratio was clearly that of the voice factor (F = 350.72, $p* < 0.01$). Moreover, the packet loss factor had a weaker effect on quality than the former factor, with F = 99.31, $p* < 0.01$. The results of the ANOVA test revealed that subjects were more sensitive to the voice than to the independent losses. This behavior is in line with P.563’s behavior, as can be clearly seen in Table IX (Appendix 5.1.1). Most probably, that was due to differences between the investigated voices, especially from a 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 also according to the voice (different ‘artificiality’ dimensions of the investigated voices) and not only according to the amount of impairments heard from the 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 section 2.4). In addition, we also found that one of the synthesized voices, namely the diphone voice (sounds less natural than unit and natural voices) was particularly disliked (on average over all conditions diphone samples were rated by approx. 1.11 MOS-LQSn less than the sam-

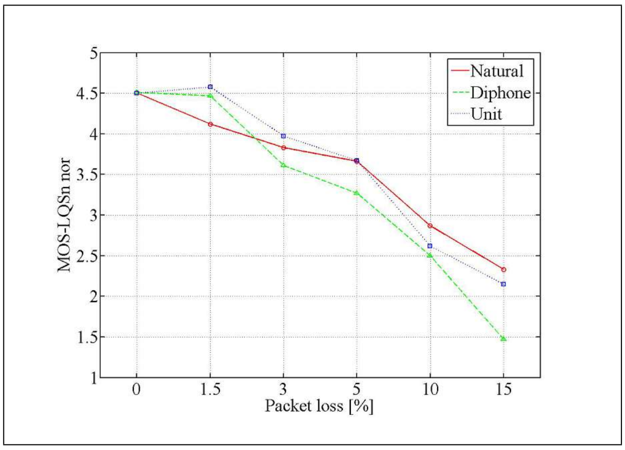


Figure 19. Normalized MOS-LQSn values for individual voices in the case of independent losses.

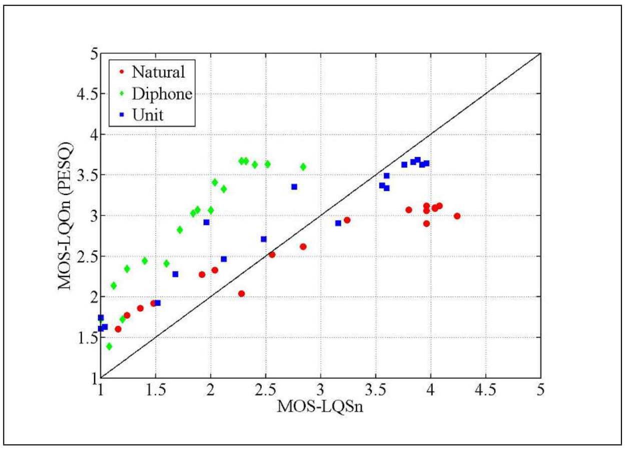


Figure 20. Subjective results (MOS-LQSn) versus MOS-LQOn (PESQ) scores for dependent losses (non-regressed).

ples generated by unit selection synthesizer and by approx. 1.5 MOS-LQSn less than naturally-produced samples). By excluding diphone voice from the analysis, the influence of the voice was decreased and packet loss became the dominant factor (F (packet loss) = 87.99, $p* < 0.01$; F (voice) = 42.02, $p* < 0.01$), more details can be seen in Table XV (Appendix 5.2.1). This supports our findings mentioned above.

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, we performed a normalization according to the formula (2) but the MOS-LQOn values were naturally replaced by MOS-LQSn values. It should be noted that the topline values for natural voice, diphone voice and unit voice are 4.05, 2.09 and 3.48, respectively. Figure 19 displays the normalized MOS-LQSn values for all investigated voices in the case of independent losses. It can be seen in this figure that the curves are very similar which indicates that, from a MOS-LQSn perspective, packet loss has the same impact on naturally-produced speech as synthesized speech.

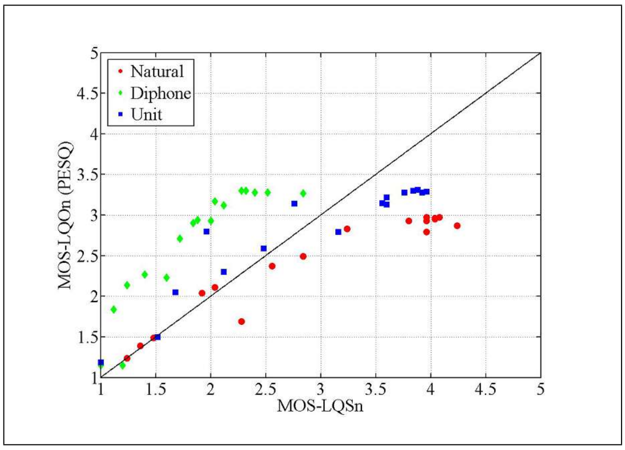


Figure 21. Subjective results (MOS-LQSn) versus MOS-LQOn (PESQ) scores for dependent losses (regressed).

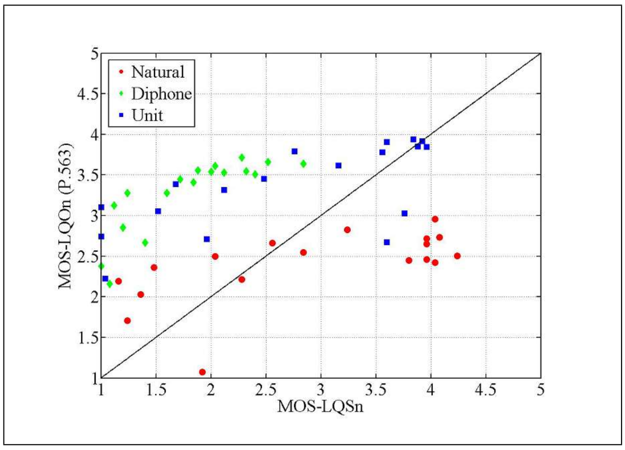


Figure 22. Subjective results (MOS-LQSn) versus MOS-LQOn (P.563) scores for dependent losses (non-regressed).

3.1.2.2. *Dependent losses*

Firstly, it should be noted that the 95% confidence intervals of MOS-LQSn values for natural voice, diphone voice and unit voice computed according to (6) were on average 0.22827 MOS, 0.2532 MOS and 0.2299 MOS, respectively. Figure 20 and 22 show the MOS-LQSn values and the raw model predictions for dependent losses, and Table V lists the respective correlations, root mean square errors and epsilon-insensitive root mean square errors. As observed for the independent loss test, the correlation between auditory judgements and instrumental predictions varies considerably between voices and models (see Table V). For PESQ, the correlation coefficient is highest for naturally-produced speech. Moreover, the smallest *rmse* and *rmse*∗ were attained for synthesized speech generated by the unit selection synthesizer, similarly for independent losses. On the contrary in the case of P.563, the correlation is higher for diphone voice but interestingly the smallest *rmse* and *rmse*∗ were obtained for natural voice.

The 3-rd order regression as recommended in [33] leads, in this case, to non-monotonic results. There are several methods to assure final regression monotonicity

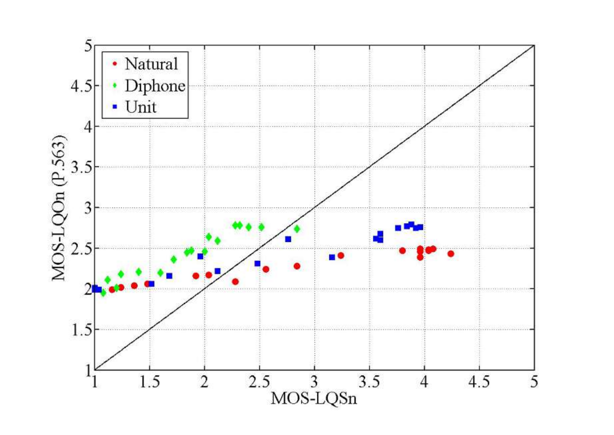


Figure 23. Subjective results (MOS-LQSn) versus MOS-LQOn (P.563) scores for dependent losses (regressed).

in such cases (e.g. outliers influence weighting, polynomial order change or non-polynomial function regression). To stick with common polynomial regression and to avoid sometimes questionable outlier penalization, we have chosen the 2-nd order polynomial regression that finally led to monotonic results with an acceptable accuracy of the final quality prediction as shown in Table VI. The related scatter plots are depicted in Figures 21 and 23. When transforming the MOSn predictions with the monotonic mapping function, the correlations rapidly increase for predictions provided by the P.563 model and root mean square errors, epsilon-insensitive root mean square errors decrease in most cases. The corresponding values for R and $rmse$, $rmse*$ are given in Table VI. 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. Comparing the performance of the two investigated models, the P.563 model attains slightly higher correlations for all voices after regression. However 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 PESQ model. To again define the significance of the differences between the presented R, $rmse$ and $rmse*$ values for PESQ and P.563, statistical significance tests were performed. The results of such tests for dependent losses are displayed in Table VII. It can be seen from Table VII that only one half of the differences is statistically significant.

Likewise as in previous case, one two-way ANOVA was conducted on MOS-LQSn's using ulp and voice as fixed factors (Appendix 5.2.2, Table XVI). In practice, we obtained similar results as for independent losses. However, a smaller impact of voice (expressed by F-ratio; F = 145.46, $p* < 0.01$) was obtained for dependent losses rather than independent losses (F = 350.72, $p* < 0.01$) in the case of all voices involved in the analysis, see Tables XIV and XVI. On the other hand, the loss factor is currently more influential than before but still does not outweigh the voice factor. As in the previous case, we also tried to exclude diphone voice from the analysis, see the reasons above in section 3.1.2.1. The loss impact (expressed by packet loss (independent losses) or ulp (dependent losses)) again proved to be dominant factor, when diphone voice was excluded from the analysis, see Table XVII. Moreover, the impact of the voice factor was considerably decreased in comparison to the previous case.

Table V. Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (PESQ) as well as MOS-LQOn (P.563) before regression for dependent losses.

	Voice	R	$rmse$	$rmse*$
PESQ	Natural	0.9723	0.1690	0.1130
	Diphone	0.9430	0.2590	0.1972
	Unit	0.9660	0.1099	0.0831
P.563	Natural	0.6260	0.2535	0.1953
	Diphone	0.8114	0.3625	0.3060
	Unit	0.6751	0.2549	0.2255

Table VI. Pearson correlation coefficient, root mean square error, epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (PESQ) as well as MOS-LQOn (P.563) after regression for dependent losses.

	Voice	R	$rmse$	$rmse*$
PESQ	Natural	0.9583	0.1962	0.1394
	Diphone	0.8888	0.2174	0.1539
	Unit	0.9406	0.1131	0.0628
P.563	Natural	0.9766	0.2894	0.2267
	Diphone	0.9592	0.1710	0.1285
	Unit	0.9669	0.2207	0.1728

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, we performed the same normalization as for independent losses (section 3.1.2.1). It should be noted that the topline values for natural voice, diphone voice and unit voice were 4.08, 2.24 and 3.75, respectively. Figure 24 shows the normalized MOS-LQSn values for individual voices. It can be seen that the curves are again very similar which proves that there is no difference between the impact of dependent losses on naturally-produced speech and synthesized speech from a MOS-LQSn perspective.

3.2. Impact of coding on objective and subjective scores

The codecs investigated here cover a wide range of different types of degradations and can be considered as the predominant codecs in current networks. In particular, the ITU-T G.729AB, Speex, iLBC, GSM-FR and EVRC-B introduce 'artificiality' dimension, unnatural sounding whereas the ITU-T G.711 produces no perceptual degradation (natural sounding), (informal expert judgements).

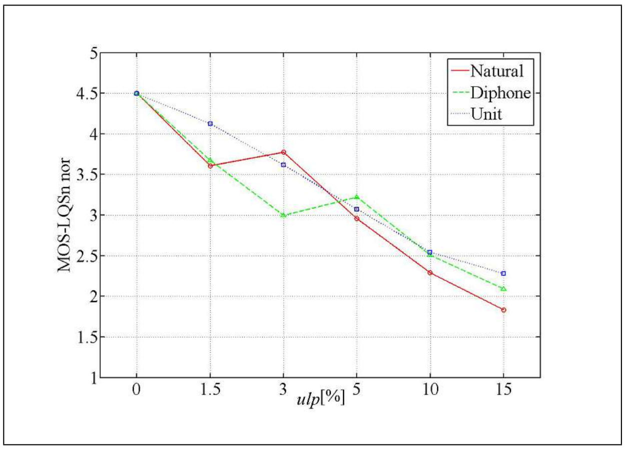


Figure 24. Normalized MOS-LQSn values for individual voices in the case of dependent losses.

Table VII. Results of statistical significance tests for the correlations coefficients, root mean square errors and epsilon-insensitive root mean square errors for dependent losses. Note: "1" indicates that the difference is statistically significant. "0" indicates that the difference is not statistically significant.

	Before regression			After regression		
Voice	*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

Figures 25–27 show a fundamental difference in the quality judgements for natural speech and synthesized speech provided by auditory test, PESQ and P.563, when processed by those codecs. In particular, a comparison of PESQ's and P.563's predictions to the auditory MOSn values is shown in Figure 25 for naturally-produced speech. It is possible to see from the mentioned figure that 'artificially sounding' codecs are rated significantly worse in both models' predictions compared to the auditory test. For the ITU-T G.711 codec (natural sounding codec), the predicted quality, especially provided by PESQ, is in better agreement with the auditory results. Furthermore, the P.563 model under-predicts the quality much more than PESQ in all cases.

The picture is quite different for synthesized voices, see Figures 26 and 27. In Figure 26, we can see the comparison of the auditory ratings with the predictions provided by the two investigated models for diphone voice. As discussed above (see section 3.1.2.1), diphone voice (sounds less natural than unit and natural voices) was particularly disliked by the 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, we would expected that its behavior would be in line with the behavior of the second synthesized voice, namely the unit voice because of similar behavior attained for objective results (see Figures 3 and

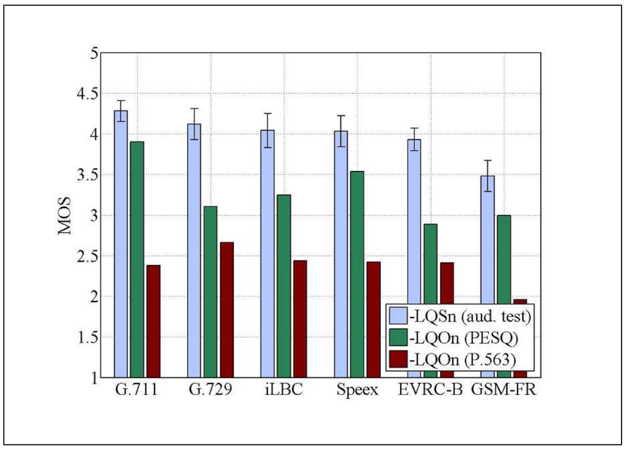


Figure 25. Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by PESQ as well as by P.563 for naturally-produced speech. The vertical bars show 95 % CI.

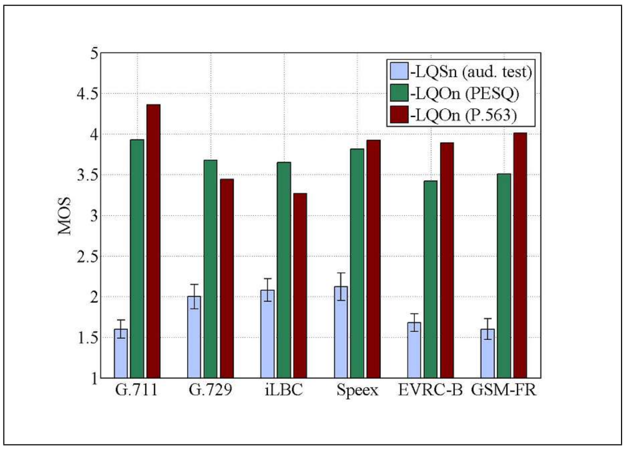


Figure 26. Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by PESQ as well as by P.563 for synthesized speech generated by a diphone synthesizer. The vertical bars show 95 % CI.

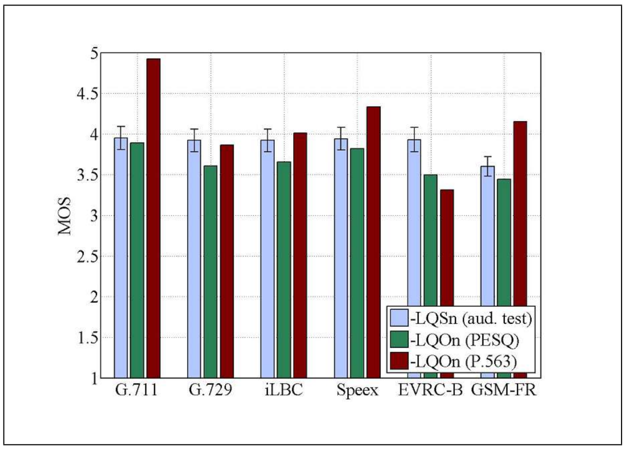


Figure 27. Effect of codecs on MOS-LQSn and MOS-LQOn's predicted by PESQ as well as by P.563 for synthesized speech generated by a unit selection synthesizer. The vertical bars show 95 % CI.

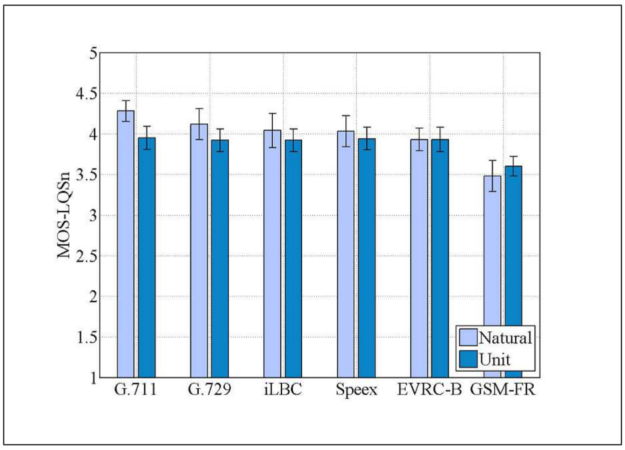


Figure 28. Comparison of the subjective ratings for naturally-produced speech with the ratings for synthesized speech generated by unit selection synthesizer in the case of coding impairments. The vertical bars show 95 % CI.

5, 9-12, Sections 3.1.1.1 and 3.1.1.2). On the basis of the presented facts, we decided to omit the diphone voice from further analysis of the behavior of synthesized speech under coding impairments. On the other hand, the behavior of the diphone voice serves as a good example of how higher unnaturalness of the voice can affect the opinions of test users. Figure 27 depicts the effect of the investigated codecs on MOS-LQSn and MOS-LQOn as predicted by PESQ as well as P.563 for the unit voice. In contrast to naturally-produced speech (see Figure 25), the predictions of both models are in good agreement - with the exception of some predictions provided by the P.563 model, such as for the ITU-T G.711 codec, etc. - with the auditory ratings. Regarding the behavior of P.563 for ITU-T G.711, a detailed investigation of this model is required 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 the auditory ratings perspective see Figure 28 (excluding diphone voice from this comparison because this voice was disliked by subjects in the test), there are some differences between subject ratings for the unit voice and the natural voice. 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 'artificially sounding' codecs is an 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 appears that the synthesized speech (see unit voice in Figure 28) is insensitive to degradations introduced by the investigated codecs - except for the GSM-FR codec - (almost the same MOS-LQSn values obtained for unit voice for almost all codecs investigated) because of the higher degree of 'artificiality' dimension introduced by the 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 the reason for lower scores for synthesized as well as naturally-produced speech. Our results are well in line with the results described in [6]. The synthesized speech is assessed a little more pessimistically than natural speech for the ITU-T G.729 codec, which is shown in Figure 5.12 (p.225, [6]). On the other hand, the synthesized speech is rated a bit more optimistically by subjects than naturally-produced speech for the IS-54 codec and its combinations. The effect is much more dominant for its combinations. Unfortunately, we did not investigate this codec as well as its combinations in this study but it should be noted that the GSM-FR codec was involved in this study and belongs to a similar family of codecs. The same behavior as for IS-54 in [6] was also reported here for GSM-FR, probably because of very similar special techniques deployed in both codec-families. Regarding the predictions of PESQ (see Figures 5.15–5.16 [6]), which were also investigated in the discussed study, they are more or less in line with our results, particularly for the ITU-T G.729 codec (see Figures 25 and 27). Unfortunately, the study published in [6] mainly focuses on the different types of codecs and their combinations. This study can serve as an extension of the study published in [6].

In addition, it looks like both models have serious problems correctly predicting the quality of natural speech impaired by present 'artificially sounding' codecs like ITU-T G.729 (they predict the quality slightly more pessimistically than was judged in the test), see Figure 25. The P.563 model is even more pessimistic than the PESQ model in this case. It should be noted that the test stimuli were composed of naturally-produced samples and a large amount of synthesized speech samples (two third of the test stimuli, see section 2.4). 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 (section 3.1.2.1), 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 section 2.4).

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

4. Conclusions and outlook

In this paper, auditory MOSn values for the naturally-produced and synthesized speech samples transmitted over different telephone channels were predicted with one comparison-based (PESQ) and one single-ended (P.563) quality prediction model. The main goal of this study was to gain a better understanding of the behavior of both model's predictions under different types of losses, coding schemes and voices as well as to assess their accuracy by comparing the predictions with subjective assessments. It has to be emphasized that none of the instrumental models investigated here (PESQ and P.563) were verified for synthesized speech, the presented analysis is an out-of-domain use case for these models.

Three specific questions were addressed in our investigation (see section 1). The first question can be answered in a positive way. All in all, the predictions provided by the PESQ model seem to be in line with the auditory ratings. On the other hand, P.563 is less accurate than PESQ for independent losses and coding impairments. Finally, we can state that both models are capable of predicting the quality of the transmitted synthesized speech under the investigated conditions to a certain degree. 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 'artificially 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 both models seems to confirm that, the PESQ model copes best with both degradations investigated here (question 3).

Future work will focus on the following issues. Firstly, we would like to investigate the performance of a brand new ITU-T intrusive model for predicting speech quality, namely POLQA under the same conditions as investigated here (as a part of the characterization phase of this model). Secondly, on the basis of the results obtained for the P.563 model, we have decided to try to design a new non-intrusive model for such conditions (synthesized speech and IP impairments). Thirdly, we would like to perform a detailed analysis 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.

5. Appendix

5.1. ANOVA for objective results

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

Table VIII. Summary of ANOVA conducted on MOS-LQOn's (PESQ) in the case of independent losses. "Loss": Packet loss.

Effect	SS	df	MS	*F*	*p*∗
Loss (1)	141.477	5	28.2954	1493.55	0.0000
Voice (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	172.420	719			

Table IX. Summary of ANOVA conducted on MOS-LQOn's (P.563) in the case of independent losses. "Loss": Packet loss.

Effect	SS	df	MS	*F*	*p*∗
Loss (1)	57.65	5	11.5301	87.73	0.0000
Voice (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			

Table X. Summary of ANOVA conducted on the MOS-LQOn's (PESQ) in the case of dependent losses ($clp = 70\%$).

Effect	SS	df	MS	*F*	*p*∗
ulp (1)	175.701	5	35.1402	503.71	0.0000
Voice (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 XI. Summary of ANOVA conducted on the MOS-LQOn's (PESQ) in the case of dependent losses ($clp = 80\%$).

Effect	SS	df	MS	*F*	*p*∗
ulp (1)	174.971	5	34.9942	358.24	0.0000
Voice (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			

5.1.1. Independent losses

Tables VIII and IX provide the results of ANOVA carried out on the independent losses test results (Dependent variable: MOS-LQOn (PESQ) and MOS-LQOn (P.563)) described in more detail in section 3.1.1.1.

5.1.2. Dependent losses

In Tables X–XIII, the results of ANOVA for the dependent losses test results and the all investigated *clp*'s (Dependent variable: MOS-LQOn (PESQ) and MOS-LQOn (P.563)) are shown. More details about this can be found in section 3.1.1.2.

Table XII. Summary of ANOVA conducted on the MOS-LQOn's (P.563) in the case of dependent losses ($clp = 70\%$).

Effect	SS	df	MS	F	$p*$
ulp (1)	13.105	5	2.6211	23	0.0000
Voice (2)	129.218	2	64.6089	494.78	0.0000
(1)*(2)	4.707	10	0.4707	3.6	0.0000
Error	91.667	702	0.1306		
Total	238.697	719			

Table XIII. Summary of ANOVA conducted on the MOS-LQOn's (P.563) in the case of dependent losses ($clp = 80\%$).

Effect	SS	df	MS	F	$p*$
ulp (1)	11.02	5	2.204	20.07	0.0000
Voice (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			

Table XIV. Summary of ANOVA conducted on the MOS-LQSn's in the case of independent losses. "Loss": Packet loss.

Effect	SS	df	MS	F	$p*$
Loss (1)	388.73	5	77.747	99.31	0.0000
Voice (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 XV. Summary of ANOVA conducted on the MOS-LQSn's in the case of independent losses excluding diphone voice. "Loss": Packet loss.

Effect	SS	df	MS	F	$p*$
Loss (1)	368.57	5	73.715	87.99	0.0000
Voice (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			

5.2. ANOVA for subjective results

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

5.2.1. Independent losses

Tables XIV and XV provide the results of ANOVA carried out on the independent loss test results (Dependent variable: MOS-LQSn) described in more detail in section 3.1.2.1.

5.2.2. Dependent losses

Tables XVI and XVII show the results of ANOVA carried out on the dependent loss test results (Dependent variable: MOS-LQSn) described in more detail in section 3.1.2.2.

Table XVI. Summary of ANOVA conducted on the MOS-LQOn's in the case of dependent losses ($clp = 80\%$).

Effect	SS	df	MS	F	$p*$
ulp (1)	427.06	5	85.413	74.77	0.0000
Voice (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 XVII. Summary of ANOVA conducted on the MOS-LQSn's in the case of dependent losses ($clp = 80\%$) excluding diphone voice.

Effect	SS	df	MS	F	$p*$
ulp (1)	455.93	5	91.187	68.88	0.0000
Voice (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			

Acknowledgement

This work has been supported by the Slovak VEGA grant agency, Project No.1/0313/08, "The investigation of the methods of detection of the critical conditions in telecommunication networks from the speech quality point of view", and the Czech ministry of Education: MSM 6840770014 "Research in the Area of the Prospective Information and Navigation Technologies". In addition, we would like to thank Marian Trnka (Slovak Academy of Sciences) for kindly providing the synthesized speech files, all subjective participants for rating our samples and last but not least the anonymous reviewers for their helpful comments and suggestions on the earlier versions of this paper.

References

[1] ITU-T Rec. P.85: A method for subjective performance assessment of the quality of speech voice output devices. International Telecommunication Union, Geneva, Switzerland, 1994.

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

[3] M. Viswanathan, M. Viswanathan: Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. Computer Speech and Language **19** (2005) 55–83.

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

[5] S. Moeller: Telephone transmission impact on synthesized speech: quality assessment and prediction. Acta Acustica united with Acustica **90** (2004) 121–136.

[6] S. Moeller: Quality of telephone-based spoken dialogue systems. Springer, New York, USA, 2005. Chapter 5, pp. 201-236.

[7] ITU-T Rec. P.800: Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva, Switzerland, 1996.

[8] ITU-T Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, Geneva, Switzerland, 2001.

[9] 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. J. Audio Eng. Soc. **50** (2002) 755–764.

[10] 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. J. Audio Eng. Soc. **50** (2002) 765–778.

[11] ITU-T Rec. P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications. International Telecommunication Union, Geneva, Switzerland, 2004.

[12] L. Malfait, J. Berger, M. Kastner: P.563 – The ITU-T standard for single-ended speech quality assessment. IEEE Transaction on Audio, Speech and Language Processing **14** (2006) 1924–1934.

[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. Acta Acustica united with Acustica **94** (2008) 21–31.

[14] D.-S. Kim: ANIQUE: An auditory model for single-ended speech quality estimation. IEEE Transaction on Speech and Audio Processing **13** (2005) 821–831.

[15] D.-S. Kim, A. Tarraf: ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality. Bell Labs Technical Journal **12** (2007) 221–236.

[16] ITU-T Rec. G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-exited linear prediction (CS-ACELP). International Telecommunication Union, Geneva, Switzerland, 2007.

[17] ITU-T Rec. P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO. International Telecommunication Union, Geneva, Switzerland, 2003.

[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.

[19] ITU-T Rec. G.711: Pulse code modulation (PCM) of voice frequencies. International Telecommunication Union, Geneva, Switzerland, 1988.

[20] ETS 300 580-2: Digital cellular telecommunications system (Phase 2); Full rate speech; Part 2: Transcoding (GSM 06.10 version 4.2.1). European Telecommunications Standards Institute, 2000.

[21] IETF RFC 3951: Internet low bit rate codec (iLBC). Internet Engineering Task Force, 2004.

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

[23] 3GPP2 C.S0014-C: Enhanced variable rate codec. Speech service options 3, 68, and 70 for wideband spread spectrum digital systems. Third Generation Partnership Project 2, 2007.

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

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

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

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

[28] ITU-T Rec. P.830: Subjective performance assessment of digital telephone-band and wideband digital codecs. International Telecommunication Union, Geneva, Switzerland, 1996.

[29] J. Mullennix, S. Stern, S. Wilson, C. Dyson: Social perception of male and female computer synthesized speech. Computers in Human Behavior **19** (2003) 407–424.

[30] S. Darjaa, M. Rusko, M. Trnka: Three generations of speech synthesis systems in Slovakia. Proceedings of XI International Conference Speech and Computer (SPECOM 2006), Sankt Peterburg, Russia, 2006, 297–302.

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

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

[33] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, O. Ghitza: Objective assessment of speech and audio quality. Technology and applications. IEEE Transaction on Audio, Speech and Language Processing **14** (2006) 1890–1901.

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

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

[36] Final report from video quality experts group on the validation of objective models of multimedia quality assessment. Phase I. http:// www.its.bldrdoc.gov /vqeg /projects /multimedia/, 2008.

[37] A. Kurittu: Validation of ITU-T P.563 single-ended objective speech quality measurement. J. Audio Eng. Soc. **54** (2006) 1092–1098.