A new approach to speech quality assessment based on back-propagation neural networks

J. Rozhon, M. Voznak, F. Rezac, J. Slachta and J. Safarik

***Abstract*—** The paper deals with modelling network effects on the quality of speech. The packet loss modelling is based on the four-state markov chain, afterwards, the resilient back-propagation (Rprop) algorithm is applied to train a neural network. The proposed solution allows for quick and precise speech quality estimation without the need to analyze the voice signal carried and belongs to the non-intrusive models of speech quality assessment. The proposed solution is tested on G.711 A-law and further generalizes the already presented concepts of the speech quality estimation in the IP environment.

***Keywords*—** Jitter, Markov chains, Neural network, Packet Loss, Speech Quality Estimation.

I. Introduction

THE growing importance of the speech and video monitoring systems, which is mainly caused by the wider use of IP-based communication, leads to increased demand for the high precision of the monitoring algorithms as well as the low computational complexity [1]-[5]. Monitoring systems are deployed in the infrastructure of all providers and we would like to point out that it concerns also mobile and wireless networks [6], [7]. The speed of algorithms assessing quality is affected by the methodology used for the quality determination. The overall result of the measurements and estimations is always a compromise of time requirements and the precision of results. In last decade many advanced mathematical approaches have appeared to improve the precision of the output results even for the quick estimation methods. One of the possible ways is described in this paper with a major focus on modelling network features and properties and creating the database of sets that could eventually be used for neural networks training or any similar procedure [8]-[10].

This research has been supported by the Ministry of Education of the Czech Republic within the project LM2010005.

J. Rozhon and F. Rezac are assistant professors with Dept. of Telecommunications, VSB-Technical University of Ostrava (17. listopadu 15/2172, 708 33 Ostrava, Czech Republic) and also researchers with Dept. of Multimedia in Cesnet, e-mails: jan.rozhon@vsb.cz, filip.rezac@vsb.cz.

J. Slachta and J. Safarik are PhD. students with Dept. of Telecommunications, VSB-Technical University of Ostrava (17. listopadu 15/2172, 708 33 Ostrava, Czech Republic) and also researchers with Dept. of Multimedia in Cesnet, e-mails: jiri.slachta@vsb.cz and jakub.safarik@vsb.cz.

M. Voznak is an associate professor with Dept. of Telecommunications, VSB-Technical University of Ostrava (17. listopadu 15, 708 33 Ostrava, Czech Rep.) and he is also a researcher with Dept. of Multimedia in Cesnet, corresponding author provides phone: +420-603565965; e-mail: voznak@ieee.org.

In this paper, the system for estimating the speech quality in VoIP networks is to be presented. This system is built upon the neural networks and takes the generally accessible network parameters as its inputs. The output of the system is the MOS estimation, which is then compared to the output of the ITU-T P.862 PESQ (Perceptual Evaluation of the Speech Quality), which serves as the reference value.

The aim of this paper is to present the generally usable system that would allow the user to estimate the speech quality regardless of the signal being carried inside the RTP packets themselves. This system effectively estimates the impact of the packet loss on the speech quality and it can be integrated to any existing environment because it uses general network statistics and the information from the RTP headers [11]. On top of that, the system can further be augmented to employ the playout buffer model and delay model utilizing the information obtained from the external source.

For the sake of this paper, only the G.711 A-law codec is used to measure the influence of the individual network features and precision of the estimate, but the same system has also been used in conjunction with the SPEEX codec with similar results and accuracies.

II. State of the Art

The survey provided in [12] showed that the PESQ algorithm accommodates the effects of packet loss on speech quality better than the E-model, and is, therefore, better suited for the task making its estimation a sensible way for improving the precision of the estimation.

The speech quality estimation system proposed in this paper is an enhancement and generalization of the system proposed in [13]. The author in this paper uses 2-state Gilbert Model to generate the losses and tries to fit the observed packet sequence into the model. This, however, proves problematic for the networks with different packet loss distributions. This gap, as well as the fact that newer version of E-model, PESQ and training algorithms for the neural networks have been devised since the publication of this work are the main motivation for this paper.

In [14], the authors use the neural networks to map the cepstrum distance for the frame. This approach leads to a similar error rate as described here and involves the signal analysis of the speech sample, which makes the system much more complex. For this reason, the work has not been used as a basis for this paper.

For the synthesized speech the recent research [15] has been performed. The authors use neural networks and genetic algorithms to estimate the quality of speech, but again the model-specific approach for the packet loss determination is used.

III. Brief Introduction to Speech Quality Assessment

Methodologies evaluating speech quality can be subdivided into two groups according to the approach applied - conversational and listening [9], [16]. Conversational tests are based on mutual interactive communication between two subjects through the transmission chain of the tested communication system. Listening tests do not provide such plausibility as conversational tests but they are recommended more frequently. According to the method of assessment speech quality evaluation, methodologies can be subdivided into subjective methods and objective methods. To evaluate speech quality, MOS (Mean Opinion Score) scale as defined by the ITU-T recommendation P.800 is applied [17]. The basic scale of assessment as prescribed by the recommendation is depicted on Fig. 1.

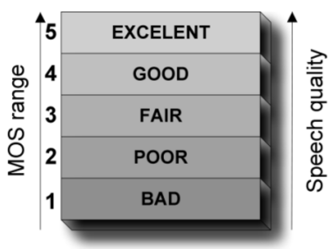

Fig. 1 MOS Scale.

In order to avoid misunderstanding and incorrect interpretation of MOS values, ITU-T published recommendation P.800.1 in 2003. This recommendation defines scales both for subjective and objective methods as well as for individual conversational and listening tests.

A. Intrusive Approach

The core of intrusive (also referred to as input-to-output) measurements is the comparison of the original sample and the degraded sample affected by a transmission chain [18]. The intrusive methods use the original voice sample as it has entered the communication system and compare it with the degraded one as it has been outputted by this transmission chain. The following list contains the most important intrusive algorithms:

– Perceptual Speech Quality Measurement PSQM,
– Perceptual Analysis Measurement System PAMS,
– Perceptual Evaluation of Speech Quality PESQ,
– Perceptual Objective Listening Quality Assessment P.OLQA.

Among these, PESQ is currently the most commonly applied algorithm [18], [19]. It combines the advantages of PAMS (robust temporal alignment techniques) and PSQM (exact sensual perception model) and is described in ITU-T recommendation P.862. The last algorithm mentioned, P.OLQA, also known as ITU-T P.863, is intended to be a successor of the PESQ. It strives to avoid the weaknesses of the PESQ's model and to incorporate a better wideband codec analysis in comparison with PESQ. As stated above, the principle of this intrusive test is the comparison of original and degraded signals, their mathematical analysis and interpretation in the cognitive model [19].

B. Non-Intrusive Approach

Contrary to intrusive methods which require both the output (degraded) sample and the original sample, non-intrusive methods do not require the original sample. This is why they are more suitable to be applied in real time. Yet, since the original sample is not included, these methods frequently contain far more complex computation models. Intrusive methods are very precise but their application in real-time measurement is unsuitable because they require sending a calibrated sample and both endpoints of the examined communication. Nevertheless, we usually need to assess the speech quality in real traffic and be able to record its changes, especially degradation [20]. Non-intrusive approaches investigate the receiving signal. Two basic principles exist: a source-based approach and a priori-based. The former, the source-based approach, is based on knowledge of various types of impairments, i.e. a set of all impairments gained by comparison of original and degraded signal characteristics. The PLP (Perceptual-linear Prediction) model is a representative of this approach. PLP compares the perceptual vectors extracted from examined samples with the untainted vectors gained from original samples. As I have mentioned, it requires a database with the set of impairments and high computational complexity. Later the PLP model was modified and the computation was accelerated, nevertheless this model is not suitable for implementation in practice as its accuracy strongly depends on the quality of the database with patterns. As for the latter approach, I would like to mention the pioneer work of Zoran and Plakal [21]. They applied artificial neural networks (ANN) to determine statistical ties between a subjective opinion and a characteristic deformation in the received sample. They also investigated spectrograms (a spectrogram is defined as a two-dimensional graphical representation of a spectrum varying in time) and they were able to establish typical uniform aspects of speech in spectrograms. The important method was standardized in recommendation ITU-T P.562 (INMD) and in ITU-T G.107, so-called E-model [20]. INMD measurement (In-service Non-intrusive Measurement Devices) is applied primarily to measure voice-grade parameters such as speech, noise and echo. The output from the model is a prediction of customer opinion Y_C^B (1).

$$Y_C^B = 1 + (E^B \cdot Y_{Cpre-echo}^B) \tag{1}$$

E^B is an echo and a delay multiplier, its value is between zero and one, to modify the pre-echo opinion score to take account of echo and delay impairments. $Y_{Cpre-echo}^B$ is the calculated pre-echo opinion score, on a zero-to-four scale, which takes into account effects of noise and loss. The addition of one converts Y_C^B to a one-to-five scale. All intermediate opinion score values are based on a zero-to-four scale for ease of calculation. It is possible to generate a rating R (2) using INMD measurements for a connection which is translated into a customer opinion of E-model [20], [22]. The E-model is one of the most modern method belonging to non-intrusive methods.

$$R = R_0 - I_{OLR} - I_{DD} - I_{e-eff} - I_{DTE} \qquad (2)$$

R_0 is the signal-to-noise ratio at a 0 dB reference point. In the equations provided (2), the 0 dB reference point is at the 2-wire input to the telephone receiving system at the near end of the connection. I_{OLR} represents the impairment term for the overall loudness rating, I_{DD} the impairment term for the absolute one-way delay and I_{e-eff} is the impairment term for the low bit-rate coding under random packet loss conditions. The last parameter I_{DTE} represents the impairment term for the delayed talker echo. I_{OLR} represents the impairment term for the overall loudness rating, I_{DD} the impairment term for the absolute one-way delay and Ie-eff is the impairment term for the low bit-rate coding under random packet loss conditions. Last parameter I_{DTE} represents the impairment term for the delayed talker echo [20].

IV. Modelling the Network Impairments

Since the vast majority of the modern communications is performed using the technologies built upon the Internet Protocol (IP), the network impairments that can actually occur during the communication include Packet Loss, Jitter and Delay.

These individual network features combine their effects on the quality of call during the transmission. And since each of them has a different nature, modelling their combined effect requires the combination of two separate models.

A. Packet Loss Modelling

The packet loss, as the name suggests, affect the call by losing one or more packets. Since each packet carries multiple samples of audio or video (for G.711 codec it is 160 samples in one packet, one sample obtains 1 Byte) the loss of those samples reflects in the quality deterioration of the call. This deterioration is as severe as much the given codec is unable to reconstruct the samples from the previous and following ones. Therefore, the impact of packet loss varies highly in dependence on the chosen codec.

Throughout the development of the IP communications the various models of packet loss have been presented, starting with simple Bernoulli model, which is a two-state Markov model with a single independent probability, and ending 4-state Markov model with 5 independent transition probabilities [23]. As the models evolved, they incorporated more and more conditions of the network with the latter mentioned model incorporating the independent losses, the correlated losses, and the bursts of losses, thus being the most general model currently used [24].

Since the 4-state model is the most general model of the packet loss, the other models (Bernoulli, Gilbert, Gilbert-Elliot) are special cases of it, therefore it is the most suitable model for creating the complex packet loss modelling tool. The 4-state model is shown in Fig. 2.

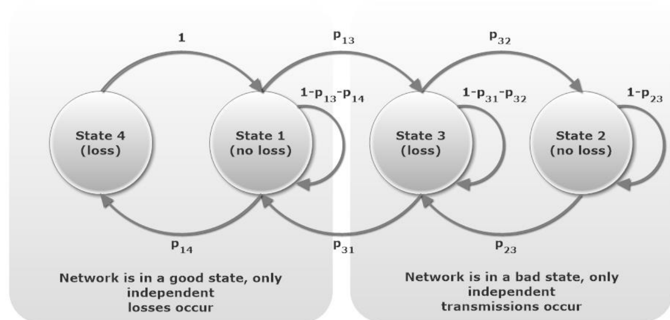

Fig. 2 The diagram of the 4-state Packet Loss Model.

The model is an actual combination of two 2-state models. The first one, which represents the ”good” state is meant to simulate only sporadic losses, therefore, the appropriate probability should be considerably low (e.g. below 20 %). On the other hand, the losses in the bad state simulate the lengthy bursts of lost packets, where the packet loss probability should be higher (e.g. above 40 %). The system in this model passes from a good to a bad state according to the desired susceptibility of the system to failures.

Let $P(x)=\{p_1, p_2, \ldots, p_n\}$ be the input sequence of packets with the length n and $P_S(x)=\{p_{S1}, p_{S2}, \ldots, p_{Sn}\}$ be the sequence of ones and zeros with same length, which is a product of the aforementioned 4-state Markov model, and where 1' represents a packet loss event and 0 represents the successful transmission (no packet loss event occurred), then the output sequence of the packets can be defined as (3)

$$P_O(x) = P(x) \wedge \neg P_S(x) \qquad (3)$$

The length of the output sequence is shortened by the number of lost packets.

B. Delay and Jitter Modelling

The delay and jitter are the time connected features of the packet. While delay characterizes the time needed for packet to traverse the transmission chain, the jitter is defined as the variability of the delay. The former has limited effect in today's communications, because the latencies even for long distance calls are below the 400 ms limit defined in ITU-T G.114. Moreover, the intrusive algorithms are, in case of really constant delay, unable to compute the quality impairment, because it is not possible for them to recognize that this is a network-related issue and not the early started recording.

Jitter, on the other hand, poses a great problem, because the fluent stream of packet is necessary for satisfiable communication. The jitter affects quality of call in two ways. First, countering it by de-jitter buffers increases the latencies and can contribute to exceeding the aforementioned limit. Secondly, really high jitter values (higher than the length of de-jitter buffer) lead to additional packet loss, because the late arrived packets are discarded.

General consensus simulates the jitter using the normal distribution, but any other distribution can be used when appropriate. For the sake of the jitter modelling in this paper the normal distribution has been chosen with mean value equal to the desired delay and the range of jitter equal to 2.575 standard deviations of the population, which for normal distribution covers 99 % of the population.

V. Dataset Creation

In the previous section, the individual models of packet loss, network delay and network jitter have been introduced. These models were actually implemented using the Python language and experimental set of degraded samples has been created for the analysis.

Because all the three network transmission features influence the call simultaneously, for one setting multiple rounds of modelling need to be performed with half of the attempts using the packet loss model first and timing model second, and vice versa. The main problem is the fact that neither the states and probabilities of the packet loss model nor the network delay and jitter model parameters can be measured and determined. These network transmission features can only be estimated, which means that the input data of both the models cannot be used in further analysis or the neural network inputs. Therefore, statistical substitutions have to be calculated for all the three parameters in order to be able to construct the input vectors of the neural network.

A. Packet Loss Analysis

To detect the packet loss in the stream of RTP packets, the standard procedure as defined in RFC 3550 can be used utilizing the packet sequence numbers [25].

From the received sequence of packets, the original Markov model probabilities cannot be obtained. This is because the originating state is not known, large quantity of packets can be lost in the end of the sequence and additional packet loss may have been introduced by the de-jitter buffer. However, part of the network state fingerprint is encoded in this given packet loss scheme. Since the best way to describe the loss events is using the model with the finite set of parameters, the reverse analysis of the sample in order to obtain the 4-state Markov model need to be done. This procedure is described in [23]. As the output of the Packet Loss Analysis, the parameters (transition probabilities) of the modified 4-state Markov model can be used.

B. Delay and Jitter Analysis

For the delay calculation, internal RTP timestamps can be used and the delay is therefore an easily accessible parameter, although as well as the packet loss and jitter it can only be estimated.

As for the jitter calculation, the substitution of the original jitter is the interarrival jitter [25], [26]. It is calculated in (4)

$$J(i) = J(i-1) + \frac{\left|D(i-1) - J(i-1)\right|}{16} \tag{4}$$

where $D(i,j)$ is the difference of relative transit times for the two packets (5).

$$D(i,j) = (Rj - Sj) - (Ri - Si) \tag{5}$$

where Ri and Si are arrival time of the packet and the timestamp from the packet respectively.

VI. Harnessing the Obtained Data

The previous sections of this paper have provided an insight into the development of a modelling tool that has been used to create samples needed to establish a data basis for the speech quality prediction. The whole picture of the procedure used to create the data is depicted in Fig. 3.

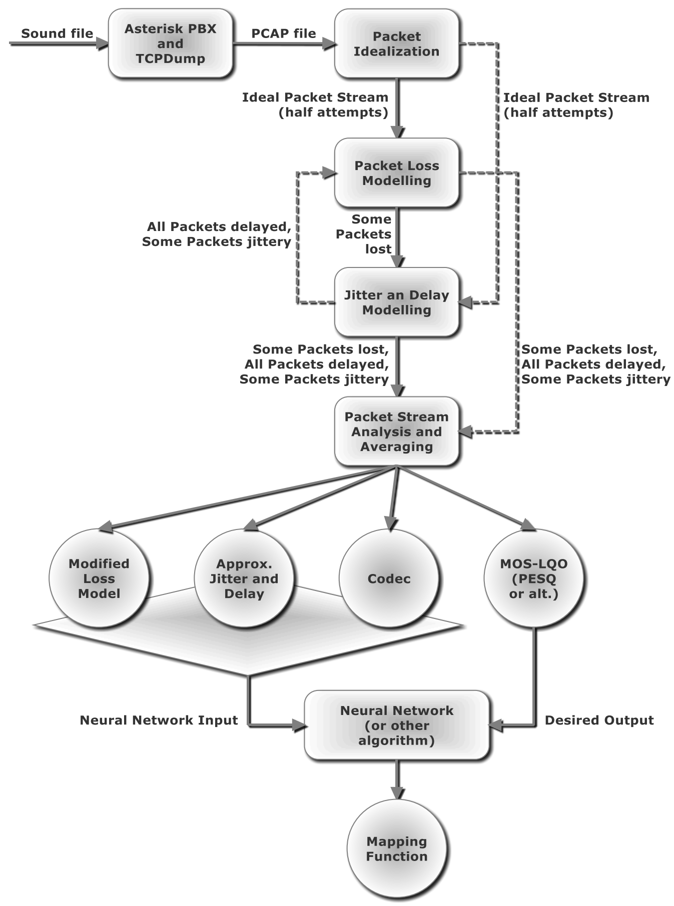

Fig. 3 Modelling algorithm with data creation and possible usage.

The prepared network samples (PCAP files) are first idealized to counter all the possible negative influences that could have been introduced during the creation of the file. This process involves checking whether no packet is missing and time alignment of the packets to precise timing in conformance with the chosen codec. These idealized packets then enter the modelling tool, that models the delay, jitter and packet loss as described above. The result is a modified PCAP file containing the *network fingerprint*'.

This PCAP file is then analyzed and the named models are then reconstructed. The reconstructed models may not resemble the original ones, but the information about the original models is lost and therefore cannot be recovered. These two new models, however, can be related to the measured quality obtained using the PESQ algorithm or any other suitable one. The relation of the network model parameters, used codec and resulting quality can be used as an input for the mathematical mapping function that can further be used for speedup in precise voice or video quality measurements.

Using the described network model, any network situation can be simulated providing the necessary data for subsequent analysis, which ties together the fingerprint of the network, used codec and resulting speech or video quality for further use in any suitable mathematical mapping function.

VII. System Architecture

The speech quality estimation is meant to be used with the basic characteristics of the IP networks - the packet loss, the one-way delay and jitter. Each characteristic impacts the speech quality differently. Moreover, only the packet loss can be acquired by simple packet observation. For jitter effect to be measured, the playout buffer needs to be employed. For the delay measurements the external source, such as RTCP, needs to be harnessed. The entire estimation system is divided into several modules based on the functionality each part implements. These parts are designed to simulate the effect of the above-mentioned characteristics and to perform the necessary calculations for the system to work. The complete architecture with closely discussed modules highlighted is depicted in Fig. 4. These modules are:

- Playout Buffer Module simulates the effect of the playout buffer on the receiving side.
- Packet Loss Analyzer takes the output of the playout buffer and searches for the lost and discarded packets using the RTP sequence numbers.
- Codec Type Module detects the codec used (codec type and packetization) and forwards the information in a binary encoded form to the neural network.
- Neural Network Module takes the loss characteristics and codec type and estimates the MOS_{LQO} and R-factor respectively. This module uses the Back-Propagation Algorithm (Rprop) to train the network with the topology of 5-3-1 neurons (plus one bias neuron for each layer) and Elliot activation function.

The architecture of the estimation system is depicted in the Fig. 4.

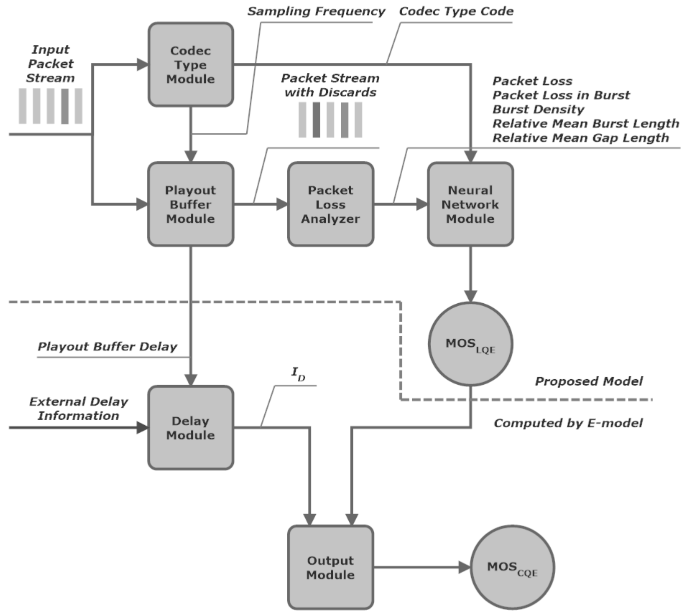

Fig. 4 The Architecture of the Speech Quality Estimation System.

A. Playout Buffer Module

There is a simplistic implementation of the playout (or de-jitter) buffer functionalities in this module. The main purpose of this module is to prove the viability and the accuracy of the original assumption that the delay variations add to both the network loss and network delay impairments.

The playout buffer implemented for the testing and measuring purposes is the fixed-length one with an unlimited data space (no early-arriving packets are discarded). It initializes the with the first arriving packet and resynchronizes the time scale after the sequence of five consecutive discards.

Let T_0 and T_i be the actual arrival times of first and i-th arriving packet (in seconds from unix epoch) and t_0 and t_i be the timestamps read from the RTP header of the respective packet (in milliseconds after recalculation using codec sampling frequency), then the ideal packet arrival time T'_i can be calculated:

$$T_i^{'} = T_0 + t_i - t_0 \tag{6}$$

By using this ideal arrival time, the packet discard happens under this condition:

$$T_i - T_i^{'} > S \tag{7}$$

where S is the playout buffer size in milliseconds.

By discarding late on arrival packets the playout buffer increases the overall statistics of the network packet loss and their characteristics causing increase in speech quality deterioration [26], [27]. Moreover, the fixed length of the buffer adds to the overall packet delay, which is also used to calculate the effect of the delay.

B. Playout Buffer Module

The Packet Loss Analyzer converts the information about the network losses and playout buffer discards into five RFC

3611 compliant statistics, which are described as follows.

Packet Loss Probability

Denoted as P_{PL}, the Packet Loss Probability is the overall percentage of the lost packets in relation to the total number of the packets sent (see eq. 8).

$$P_{PL} = \frac{N_{LOST}}{N_{SENT}} \tag{8}$$

where N_{LOST} is the number of lost packets and N_{SENT} the total number of RTP packets sent.

Packet Loss Probability in Burst

Denoted as P_{BPL}, the Packet Loss Probability in Burst is the overall percentage of packets lost in a burst (for burst definition see [rfc3611]) defined as stated in eq. 9).

$$P_{BPL} = \frac{N_{LOST\_IN\_BURST}}{N_{SENT}} \tag{9}$$

where $N_{LOST\_IN\_BURST}$ is the number of packets lost in all bursts in a packet sequence.

Burst Density

Denoted as ρ, the Burst Density is the ratio of the number of lost packets in all burst periods and the total length of all bursts in a packet sequence (see eq. 10).

$$\rho = \frac{N_{LOST\_IN\_BURST}}{N_{BURST}} \tag{10}$$

where N_{BURST} is the number of all packets lost in all bursts in a packet sequence.

Relative Mean Burst Length

Denoted as E'(B), the Relative Mean Burst Length is derived from the characteristic defined in [28], but is related to the total length of the captured packet stream (see eq. 11). This relativization is done for the neural network to be able to categorize similar characteristics of the different-length packet sequences.

$$E'(B) = \frac{N_{BURST}}{N_{SENT} \cdot K_{BURST}} \tag{11}$$

where K_{BURST} is the number of bursts in a packet sequence.

Relative Mean Gap Length

Denoted as E'(G), the Relative Mean Gap Length is the complement of the E'(B) and is defined in eq. 12.

$$E'(G) = \frac{N_{GAP}}{N_{SENT} \cdot K_{GAP}} \tag{12}$$

where N_{GAP} is the number of packets in all gaps and K_{GAP} is the number of gaps in a packet sequence. The definition of the burst relies on the G_{min} parameter [28], which in this case was set to default of 16 packets as it is recommended in the mentioned RFC.

VIII. Experimental Measurements

As a part of the conducted research several experimental measurements have been performed to solidify the assumptions stated above. In this section, the possibilities of using the neural network to estimate the quality of speech based on the packet loss data will be discussed. At first two minimalistic network designs have been created. The first one is the model of the wired network with independent losses only. The second one then models the wired network with dependent losses. Both of the models use G.711 (A-law) encoded speech data, but for the sake of this paper the former one will be discussed in greater detail to present the proposed approach.

A. Modelling and Estimating the Independent Losses

For the model presented in the Fig. 2 to behave as a simple Bernoulli model with just one independent variable, it is necessary to set the transition probabilities as follows (13):

$$p_{41} = p_{23} = p_{31} = 1, \quad p_{13} = p_{32} = 0 \tag{13}$$

These constraints come from the fact that only two of four model states are used and that the starting state can be any given one. Therefore, setting the probabilities as it is shown ensures that after many transitions the model will converge into the Bernoulli one. The only transition probability not listed p_{14} is the independent variable of this model. Since the model needs to be in a steady state to generate the expected outcome, the first 1000 transitions are discarded. As long as the model works with the precision in percents, this value is sufficient. On the listening side (output of the model) however, the information about the model's transition probabilities cannot be obtained. Therefore, the generally used term - packet loss probability P_{pl} is observed. For this experiment the range for the P_{pl} is set from 0 to 0.15 with the step of 0.01, which ensures the results in whole spectrum of the MOS range. From the balance equations for the given model the p_{14} can then be calculated as follows (14):

$$p_{14} = \frac{\pi}{1-\pi_4} \wedge \pi_4 = P_{PL} \tag{14}$$

With the model set, input data have been modified based on the models output resulting in degraded voice samples, which have then been used for the MOS_{LQO} calculations. The results of these calculations are in the chart in the Fig. 5.

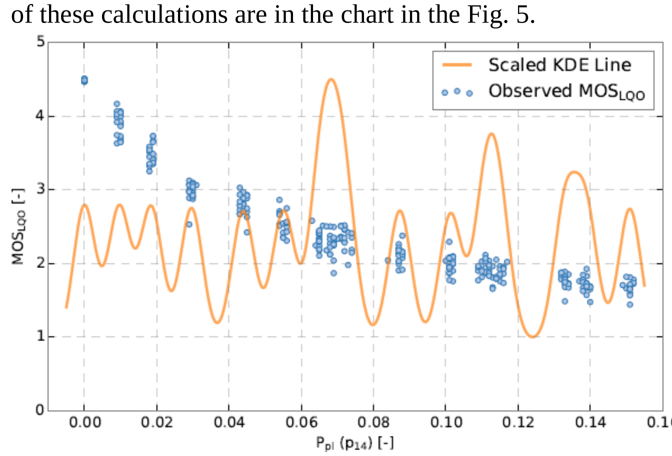

Fig. 5 MOS_{LQO} vs. PL Probability and respective KDE Line.

These observations display an obvious exponential decrease in speech quality. However, due to the probabilistic nature of the model, the observations are shifted from the input coordinate P_{pl} has been chosen with a 0.01 step, therefore, the data cannot be easily clustered. To cluster the data for the further analysis, histograms or kernel density estimation can be used, since the distribution of the observations on the x-axis is the set of 1D data. Due to the obvious feature of the observations, which form wider and tighter clusters, the histogram approach could lead to wrong classification. For this reason, kernel density estimation have been used with gaussian-shaped kernels. The resulting curve is shown in the scaled and shifted manner in the Fig. 5. By exploring this curve, the clusters of observations can easily be identified when looking on the maxima of the curve as the clusters' centers of gravity and minima as the clusters' boundaries.

Since the clusters of observations have been identified, the descriptive parameters of these clusters can be calculated, such as the outliers, means and standard deviations. Based on these parameters the neural network can be trained. However, to achieve higher precision one more step of data preprocessing have been added and the fitting curves for the mean (15) and standard deviations have been calculated.

$$MOS_{LQO} = 2.805 \cdot e^{-20.989 \cdot P_{PL}} + 1.622 \qquad (15)$$

The outliers, means, standard deviations and fitting curves are shown in the Fig. 6.

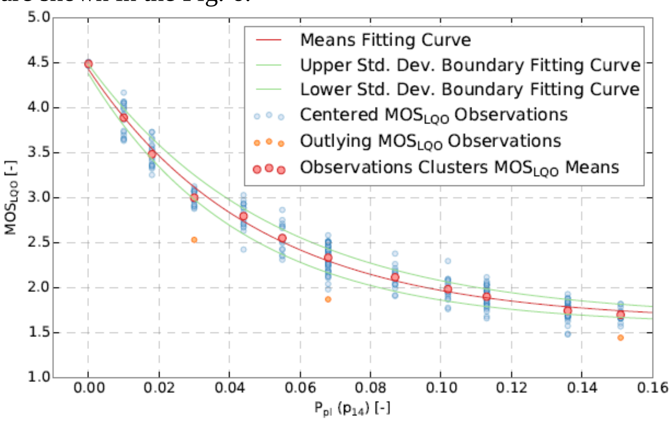

Fig. 6 MOS_{LQO} Observations Clusters and Fitting Curves.

Although the fitting curves provide sufficient way to estimate the MOS_{LQO} based on the percentage of the independently lost packets, the estimation system is meant to be much more robust. Therefore, the estimation is done using the neural networks. In this simple case, where the exponential function is to be estimated, the network with only two hidden neurons can perform well. This is because of the similarities between exponential curve and the sigmoid function used in the neurons. The comparison between the fitting curve of the means of the observations and the neural network estimate of it is shown in the Fig. 7.

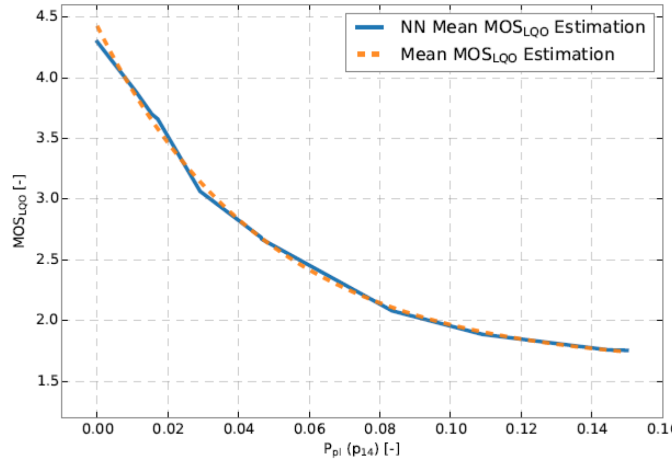

Fig. 7 The Comparison of the NN based MOS_{LQO} Estimation with Fitting Curve of Means.

The comparison of the two curves shows the neural network approximates the fitting curve almost perfectly and can therefore be used for the MOS_{LQO} calculations in the network with independent losses.

B. Modelling Dependent Losses

Unlike the independent losses the dependent ones are much more complex problem. When talking in the terms of the presented 4-state model, to model the dependent losses only the states 1 and 3 are to be used. Similarly to previous subsection, the constraints defining the model are as follows.

$$p_{41} = p_{23} = 1, \quad p_{14} = p_{32} = 0 \qquad (16)$$

The remaining transition probabilities p_{13} and p_{31} are the independent variables of the model. The observer, as well as for the previous case, cannot determine these probabilities based on the output packet stream, therefore the analytically obtained variables have to be used. These include P_{pl} - packet loss probability, P_{bpl} - packet loss probability in burst, ρ - density of packet loss in bursts.

These variables are conform to ones defined in [23]. In the experiment the total percentage of lost packets again have ranged from 0 to 0.15 with a step of 0.01, but since there are two independent variables the input sets were chosen (from the whole interval 0-1) so that both probabilities are in whole percents. The equation specifying the relation between probability of packet loss P_{pl} and the variables p_{13} and p_{31} can again be derived from the balance equations of the model and looks as follows (17).

$$\pi_3 = P_{PL} = \frac{p_{13}}{p_{13} + p_{31}} \qquad (17)$$

Based on this input data, the wired network with dependent losses has been modeled resulting in the set of observations as shown in Fig. 8. In this chart the MOS_{LQO} values are encoded in color of the scatter points. It can be observed that with increasing burstiness of the losses the voice quality increases, thus showing the evident relation between speech quality and burstiness of the packet losses.

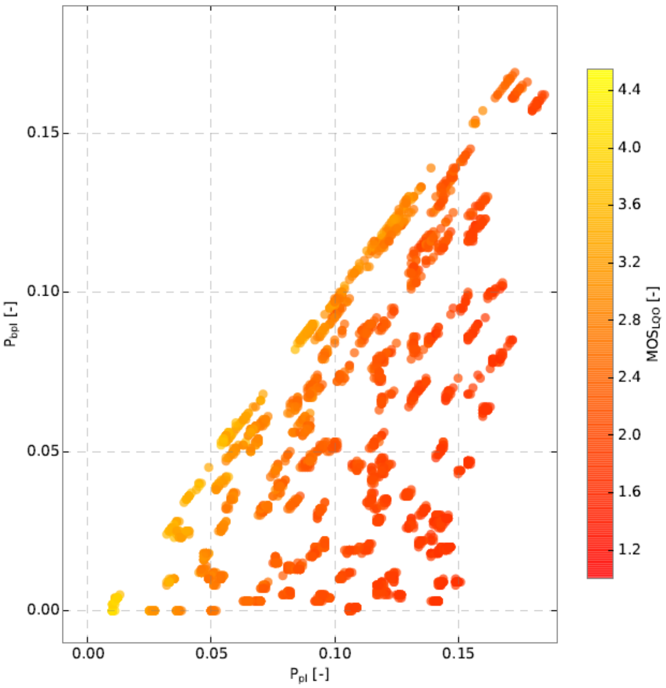

Fig. 8 The color-encoded MOS_{LQO} values in relation to overall and burst Packet Loss.

This relation can again be encoded in the structure of the neural network. However, there are several obstacles on the way and therefore this is the main aim of our current research.

IX. Network Characteristics Modelling and Achieved Results

The network characteristics were, for the purposes of this paper, generated using two models. The first one is the 4-state Markov Model generating the network losses. For this model, the transition probabilities were set to cover the whole range of the MOS_{LQO} scale. The second model is the model simulating the jitter effect. For this purpose, the simple Normal Distribution model has been used. The model was implemented with just one parameter - the jitter threshold, which is the value equal to 2.575 ς of the model. This setting results in 99% of the observations falling into the interval [-threshold, threshold]. There were 10 sound samples encoded in mono-channel linear 16-bit PCM as the simulation inputs, which were taken from the P.862 recommendation. These sound files were transformed into G.711 A-law encoded RTP packets, which were idealized in time (exact 20 ms long packet spacing) and then stored in a PCAP file. These files were then manipulated using the models mentioned above and reconstructed to form the input for the PESQ algorithm.

First of all, the simulation consisting in pure packet loss manipulations was performed. The 4-state loss model was set with the combinations of transition probabilities stated in Tab. 1. All unique combinations were used 4 times to smoothen the effect of possible outliers, which resulted in 12,200 observations for training and validation and 3,840 observations for testing of neural network.

Tab 1. Settings of individual probabilities in a 4-state loss model.

Probability	Training Values [%]	Testing Values [%]
p_{13}	0, 1, 3, 5, 10	1, 3, 7, 15
p_{14}	0, 1, 3, 5, 10	1, 3, 7, 15
p_{32}	0, 1, 3, 5, 10	1, 3, 7, 15
p_{31}	60, 75, 90	50, 80
p_{23}	100	100
p_{41}	100	100

The given settings produced the network statistics with statistical features of the training set as they are depicted in Fig. 9. For the purposes of readability the MOS is scaled to fit in the range from 0.2 (equals to 1) to 1 (equals to 5).

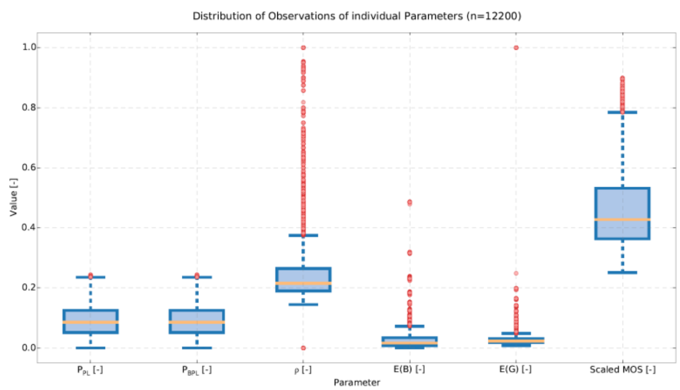

Fig. 9 The Statistical Features of the Observed Training Set.

Approximately 20% of randomly chosen observations from the training set then formed a validation set, which was then used to confirm the ability of the neural network to estimate the speech quality of the samples, the features of which resemble those of the training samples.

The remaining training set was then preprocessed in a way, so that the groups of same input vectors had the same output. This step is necessary because the neural network is not capable of learning multitude of outputs for the single input.

The training and testing was repeated 10 times. As it is obvious from the Fig. 10, the proposed system achieves high correlation with Pearson's Correlation Coefficient around the 0.95 and RMSE of 0.2 (MOS), which corresponds to an error of approximately 7% (related to the middle of the MOS Scale).

Due to the fact that packet loss has a great impact on speech quality, the most of the observations in all sets are below tolerable value of MOS and can be discarded as unusable. For the threshold of 2.5 (MOS; all observations below this are discarded), the 3,819 observations in training set fit this condition and resulting RMSE gets higher to approximately 0.25. This estimation error is the same as presented in [sun], but with the system being able to work with any possible loss model (not just the Gilbert Model).

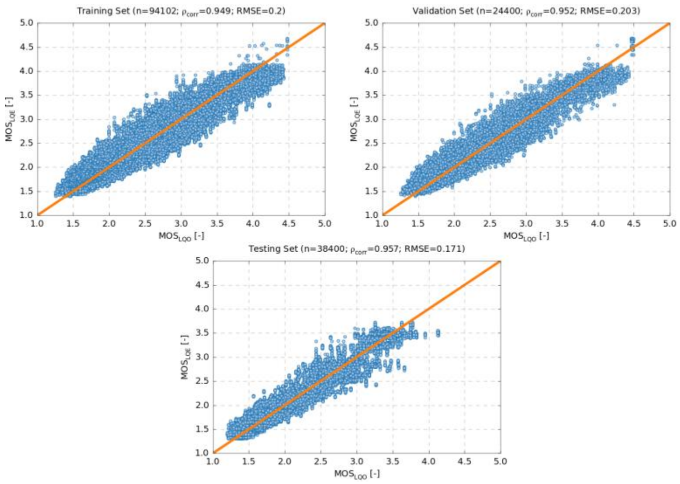

Fig. 10 The Correlation Diagrams for the Estimations Training, Validation and Testing Sets. MOS_{LQO} is the reference from PESQ and MOS_{LQE} is the estimation.

As regards jitter, the effect of jitter can be split into the packet loss part and the delay part. Since the delay impact can be calculated using E-model, only the packet loss effect is to be studied. For this purpose the jitter threshold (as it is described above) was set from 0 to 100 ms. The jitter buffer size was set to 30 ms. And each simulation was performed again on 10 unique sound samples and repeated 10 times. This way 10 000 observations were made. For the purposes of estimation, the best performing (in terms of RMSE) neural network from the previous subsection was used to confirm the validity for entirely different loss model. The Fig. 12 shows the appropriate correlation diagram.

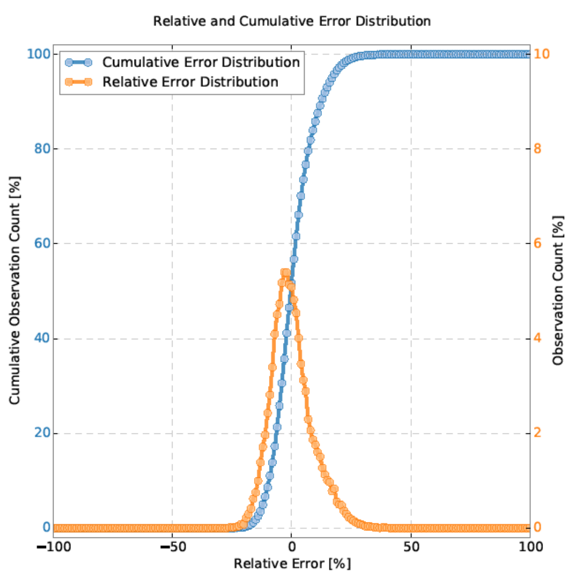

Fig. 11 The Relative Error Distribution and the Cumulative Error Distribution.

The performance in this case is 0.989 for Pearson's correlation coefficient and 0.185 for RMSE, which proves the model network is trained well. The error distribution is again very similar to the Normal Distribution.

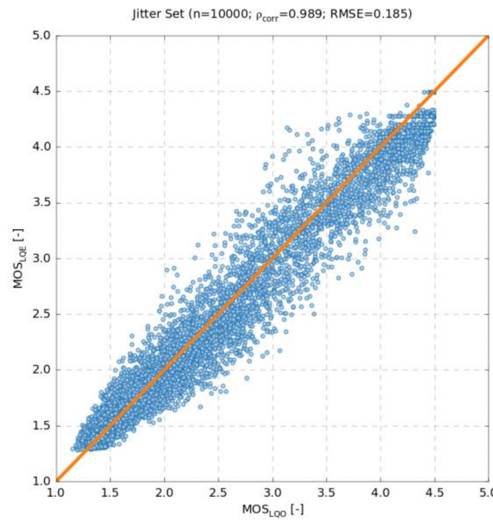

Fig. 12 The Correlation Diagram for the Jitter Observations.

X. Conclusion

In this paper the speech quality estimation system has been presented. This system takes the observed network loss statistics compliant to those defined in RFC 3611 and perform the MOS_{LQE} estimation using the neural network model. The system has been successfully tested with two codecs - G.711 A-law (no internal PLC) and SPEEX (internal PLC).

The speech quality estimation can be done with RMSE around the 0.25 MOS, which equals to approximately 8% error, which is the level achieved in similar solutions.

The training time for the network has always been below 30 seconds. Moreover, with more intense data preprocessing the training time can be decreased to approximately 3 seconds. On the other hand, however, no statistically significant improvement in speech quality estimation has been achieved with any kind of data preprocessing.

Due to the general nature of the system, it can be deployed in any existing environment. If there is a need to incorporate the delay effect as well, the E-model calculation can be added to the system seamlessly. This approach, however, requires an external source of the measured one-way delay.

References

[1] H. Toral, D. Torres, L. Estrada, "Simulation and modeling of packet loss on VoIP traffic: A power-law model," *WSEAS Transactions on Communications*, Vol. 8, No. 10, pp. 1053-1063, 2009.

[2] L. Estrada, D. Torres, H. Toral, "Analytical description of a parameter-based optimization of the quality of service for voIP communications," *WSEAS Transactions on Communications*, Vol. 8, No. 9, pp. 1042-1052, 2009.

[3] F. De Rango, P. Fazio, F. Scarcello, F. Conte, "A new distributed application and network layer protocol for voip in mobile ad hoc networks," *IEEE Transactions on Mobile Computing*, Vol. 13, No. 10, pp. 2185-2198, 2014.

[4] H. Toral, D. Torres, L. Estrada, Simulation and modeling of packet loss on α-stable VoIP traffic," in Proc. *9th WSEAS Int. Conf. Multimedia, Internet and Video Technologies, MIV '09*, Budapest, 2009, pp. 198-203.

[5] G. Yasin, S. F. Abbas, S. R. Chaudhry, "MANET routing protocols for real-time multimedia applications," *WSEAS Transactions on Communications*, Vol. 12, No. 8, pp. 386-395, 2013.

[6] P. Pocta, P. Kortis, M. Vaculik, "Impact of background traffic on speech quality in VoWLAN," *Advances in Multimedia*, Vol. 2007, art. no. 57423, 2007.
[7] A. Roy, M. I. Islam, M. R. Amin, "Performance evaluation of voice-data integrated traffic in IEEE 802.11 and IEEE 802.16e WLAN," *WSEAS Transactions on Communications*, Vol. 12, No. 7, pp. 352-365, 2013.
[8] Q. Fu, K. Yi, M. Sun, "Speech quality objective assessment using neural network," in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, 2000, pp. 1511-1514.
[9] M. Voznak, "Recent advances in speech quality assessment and their implementation," *Lecture Notes in Electrical Engineering*, Vol. 282 LNEE, 2014, pp. 1-14.
[10] P. Marius-Constantin, V.E. Balas, L. Perescu-Popescu, N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, Vol. 8, No. 7, pp. 579-588, 2009.
[11] R. Burget, D. Komosny, K. Ganeshan, "Topology aware feedback transmission for real-time control protocol," *Journal of Network and Computer Applications*, Vol. 35, No. 2, pp. 723-730, 2012.
[12] H.A. Khan, L. Sun, "Assessment of Speech Quality for VoIP Applications using," *PESQ and E-Model. Advances in Communications, Computing, Networks and Security*, Vol. 7, pp. 263-273, 2008.
[13] L. Sun, E.C. Ifeachor, "Voice quality prediction models and their application in VoIP networks," *IEEE Transactions on Multimedia*, Vol. 8, No. 4, pp. 809-820, 2006.
[14] M. Meky, T. Saadawi, "Prediction of speech quality using radial basis functions neural networks," in Proc. *2nd IEEE Symposium on Computers and Communications*, Alexandria, 1997, pp. 174-178.
[15] M. Mrvova, M., Pocta, P.,"Quality estimation of synthesized speech transmitted over IP channel using genetic programming approach," in Proc. *International Conference on Digital Technologies*, Zilina, pp. 39-43, 2013.
[16] A. E. Mahdi, D. Picovici, "Advances in voice quality measurement in modern telecommunications," *Digital Signal Processing*, Vol. 19, No. 1, pp. 79-103, 2009.
[17] Methods for subjective determination of transmission quality, ITU-T Recommendation P.800, Geneva, 1996.
[18] A. Rix, M. Hollier, A. Hekstra, J. Beerends, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment," *AES: Journal of the Audio Engineering Society*, Vol. 50, No. 10, pp. 755-764, 2002.
[19] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T Recommendation P.862, Geneva, 2001.
[20] M. Voznak, "Non-intrusive Speech Quality Assessment in Simplified E-Model, " *WSEAS Transactions on Systems*, Vol. 11, No. 8, pp.315-325, 2012.
[21] J. Palakal, M. Zoran, "Feature extraction form speech spectrograms using multi-layered network models," In Proc. *IEEE International Workshop on Tools for Artificial Intelligence, Architectures, Languages and Algorithms*, 1989, pp. 224–230.
[22] M. Voznak, "E-model modification for case of cascade codecs arrangement," *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol.5, No. 8, pp. 1439-1447, 2011.
[23] S. Salsano, F. Ludovici, A. Ordine, A. Giannuzzi, "Definition of a general and intuitive loss model for packet networks and its implementation in the Netem module in the Linux kernel," University of Rome Tor Vergata, Technical Report, 2012.
[24] A. Jurgelionis, J. Laulajainen, M. Hirvonen, A. Wang, "An empirical study of NetEm network emulation functionalities," in Proc. *International Conference on Computer Communications and Networks*, ICCCN, art. no. 6005933, 2011.
[25] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, RFC3550 - RTP: A Transport Protocol for Real-Time Applications. IETF, 2003.
[26] A. Kovac, M. Halas, M. Orgon, M. Voznak, "E-model Mos estimate improvement through jitter buffer packet loss modelling," *Advances in Electrical and Electronic Engineering*, Vol. 9, No. 5, pp. 233-242, 2011.
[27] D. Komosny, M. Voznak, K. Ganeshan, H. Sathu,"Estimation of Internet Node Location by Latency Measurements - The Underestimation Problem," *Information Technology and Control*, Vol. 44, No. 3, pp. 279-286, 2015.
[28] T. Friedman, R. Caceres, A. Clark, RTP Control Protocol Extended Reports (RTCP XR), IETF RFC 3611, 2003.

J. Rozhon is an Assistant Professor with the Dpt. of Telecommunications, Technical University of Ostrava. He is also a researcher with Dpt. of Multimedia in CESNET. His main topic of interest is the research in the field of performance testing and benchmarking of SIP network elements. He is also active in the IP telephony in general focusing on the various implementations of Asterisk PBX and its alternatives.

M. Voznak holds position as an associate professor with Department of Telecommunications, Faculty of Electrical Engineering and Computer Science (FEECS) VSB-Technical University of Ostrava, Czech Republic. He received his M.S. and Ph.D. degrees in telecommunications, dissertation thesis “Voice traffic optimization with regard to speech quality in network with VoIP technology” from the Technical University of Ostrava, in 1995 and 2002, respectively. The topics of his research include next generation networks, IP telephony, speech quality and network security. He is a member of the editorial boards of several journals, conference committees of international scientific conferences and IEEE Communications Society.

F. Rezac is an Assistant Professor with Dpt. of Telecommunications, Technical University of Ostrava, Czech Republic. He received his MSc in 2009 in study branch “Mobile Technology” from Technical University of Ostrava and currently continues in the doctoral study. His research is focused on Voice over IP technology, Network Security and Call Quality in VoIP. He is also a researcher with Dpt. of Multimedia in CESNET..

J. Slachta received his M.S. degree in telecommunications from VSB – Technical University of Ostrava, Czech Republic, in 2014 and he continues in studying Ph.D. degree at the same university. His professional activities are focused on embedded systems, networks and application development for mobile systems.

J. Safarik received his M.S. degree in telecommunications from VSB – Technical University of Ostrava, Czech Republic, in 2011 and he continues in studying Ph.D. degree at the same university. His research is focused on IP telephony, computer networks and network security.