Miroslava Mrvova - Peter Pocta *

A QUALITY ESTIMATION OF SYNTHESIZED SPEECH TRANSMITTED OVER IP NETWORKS

A design of the parametric models estimating a quality of synthesized speech transmitted through IP networks is presented in this paper. A Genetic Programming and Random Neural Network as machine learning techniques were deployed to design the models. A set of the quality-affecting parameters was used as an input to the designed parametric estimation models in order to estimate a quality of synthesized speech transmitted over IP networks (VoIP environment). The performance results obtained for the designed parametric estimation models have validated both genetic programming and random neural network as powerful techniques, delivering good accuracy and generalization ability; this makes them perspective candidates for quality estimation of this type of speech in the corresponding environment. The developed parametric models can be helpful for network operators and service providers in a planning phase or early-development stage of telecommunication services based on synthesized speech.

Keywords: *Genetic programming, random neural network, speech quality estimation, synthesized speech, packet loss, speech codec.*

1. Introduction

A speech quality assessment process is useful for network operators and service providers to evaluate the quality of voice services offered by current telecommunication networks. The speech quality assessment can be performed either from the subjective or objective point of view. The traditional way how to assess speech quality is called subjective testing. In principle, subjective testing is based on a large enough group of human subjects, who listen to given samples and assign an opinion score on a scale ranging from 1 "bad quality" to 5 "excellent quality" (i.e. MOS (Mean Opinion Scale) scale). This approach is impractical in real conditions, because of the number of subjects, that have to participate in a test, time-consumption, high costs, etc.

In contrast to subjective testing, the objective testing employs a computer program or mathematical model to approximate an average user behavior associated with a perception of speech quality. There are two kinds of objective testing. Firstly, the intrusive models (e.g. ITU-T PESQ) are characterized by comparing two types of signals. They evaluate the quality of a degraded (output) speech signal by comparing it with a corresponding reference (input) speech signal. These methods are very accurate but not suitable for monitoring real-time traffic in telecommunication networks. Secondly, non-intrusive models can be classified into two groups, namely signal-based and parametric. Signal-based non-intrusive models are based (as the name implies) on a speech signal and do not use the reference speech sample in comparison to the intrusive models. A hidden reference speech sample is created internally and compared to degraded signal deploying similar approaches as used in the intrusive models. On the other hand, parametric non-intrusive models are based on estimating the quality of speech transmission using input parameters characterizing this transmission from a quality point of view [1] and [2].

Parametric models based on machine learning techniques, such as a neural network or genetic programming, are considered as a novel approach of speech quality estimation. This approach has become popular because of its high correlation with a human perception and relatively easy implementation. Moreover, it is able to offer a real-time and continuous evaluation of the speech quality in comparison with the intrusive and non-intrusive signal-based methods.

In recent years, the interest in a synthesized speech (speech generated by computer) has grown extremely. Up to now, synthesized speech has reached a quality level which allows it to become a part of modern daily life [3]. Furthermore, voice-based telecommunication services are on decline. In order to make some of voice-based services more economically effective, informational services involving call center operators have been replaced by services based on the synthesized speech. Although,

* **Miroslava Mrvova, Peter Pocta**
Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia,
E-mail: miroslava.mrvova@fel.uniza.sk

the services based on this kind of speech are broadly used, there is no parametric model designed for estimating a speech quality of these services. Such a model could be helpful for network operators and service providers implementing them in a planning phase or early-development stage of telecommunication services based on synthesized speech.

In this article, novel parametric non-intrusive models estimating the quality of synthesized speech transmitted through IP networks are addressed. These models make use of two different biologically inspired machine learning techniques, namely Genetic Programming (GP) and Random Neural Network (RNN). The decision to deploy these two approaches in this study was mostly motivated by very promising results obtained by GP and RNN for very similar tasks as here, see for instance [4 and 5]. It should be noted here that GP employs an evolutionary process inspired by Darwinian evolution to automatically derive mathematical model (function) from initial parameters; it requires only minimal assumptions about a structure of solution and it provides promising results in terms of accuracy and computational efficiency [4]. On the other hand, RNN consists of neurons, which interact with each other by exchanging excitatory and inhibitory signals in a network, similar to a biological neural network. Each state of neuron is characterized by a neuron potential as a non-negative integer value, which yields to a more detailed state representation. The positive potential means that neuron is in an excited state, sending excitatory and inhibitory signals at random intervals to other neurons or outside of the network. It was found in [5] that a well trained RNN model is able to give reasonable results even for parameter values outside the range defined in training phase, i.e. it provides a good extrapolating ability. In other words and contrary to ANN models, RNN has also good ability to generalize for new inputs. More detailed information about GP and RNN can be found in [6 - 9].

The rest of the paper is organized as follows: In Sections 2 and 3, the experimental setup and results (a performance of the designed parametric models for predicting the speech quality of synthesized speech impaired by packet loss and coding) are presented and discussed respectively. Finally, Section 4 concludes the paper and suggests a future work.

2. Experimental setting

The GP and RNN are approaches, which employ supervised learning, i.e. they require values of inputs together with values of their relevant outputs in a training process. In general, their main task is to learn the numerical or logical relationship between input and actual output parameters during the training process. Due to a fact that the aim of this article is to design parametric model estimating a quality of synthesized speech in a VoIP environment, following parameters were considered as input parameters:

- a quality-affecting parameters representing an impact of IP networks, i.e. parameters characterizing packet loss process: unconditional loss probability (ulp) and conditional loss probability (clp)),
- other quality-affecting parameters, like speech codec type and a type of synthesized speech signal.

The input parameters defined above were used together with their corresponding MOS values expressing the speech quality ranging from 1 “bad quality” to 5 “excellent quality” (actual output parameter). GP and RNN were fed with the above mentioned parameters in order to provide the MOS values as a function of parameters typical for IP networks. Fig. 1 shows a diagram of the designed parametric estimation model.

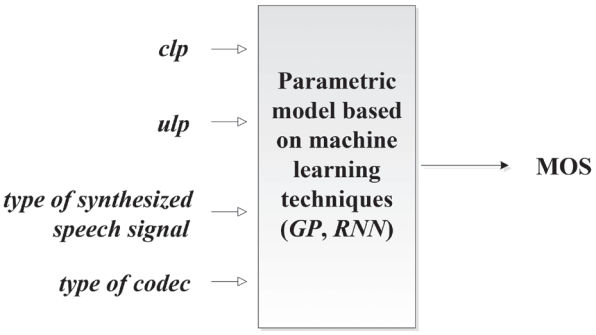

Fig. 1 A diagram of the designed parametric estimation model

A. Database creation

4320 speech samples representing different conditions affecting the speech quality were recorded and assessed for the purpose of this research. At the beginning, the speech samples with a length of 12 seconds were synthesized by two state-of-the-art Slovak synthesizers, namely Diphone synthesizer (Kempelen 1.6 [10]) and Unit-selection synthesizer (Kempelen 2.1 [10]). Subsequently, these synthesized speech samples were coded by three different encoding schemes, namely ITU-T G.729AB [11], ITU-T G.711 [12] and iLBC [13]. After a transmission of the particular speech signals through the IP network, where we simulated different packet loss situations, their speech quality scores were predicted by ITU-T P.862 PESQ (Perceptual Evaluation of Speech Quality) [14]. The decision to use PESQ model (an objective model designed for predicting a quality of naturally-produced speech) as a predictor of speech quality was based on the experiments carried out in [15 - 17], which have proven that PESQ is able to provide accurate predictions of the quality of synthesized speech impaired by the impairments used in this study. For example, the correlation coefficient reported in [15] was above 0.9.

Regarding the packet loss situations, we have concentrated on parameters, which precisely capture all aspects of a packet loss process, namely conditional loss probability (as an indicator of packet loss burstiness, denoted as clp) and unconditional loss

probability (as an indicator of packet loss rate, denoted as ulp). In fact, we simulated different values of clp: 0% (theoretically) - Bernoulli loss model (modeling independent losses), 70% and 80% - Gilbert loss model (modeling dependent losses) and of ulp: 0%, 1.5%, 3%, 5%, 10% and 15%. By using these 4320 synthesized speech samples, 108 average speech quality scores (6 ulp x 3 clp x 2 types of synthesized speech signals x 3 types of codec) called configurations were formed.

B. Design of parametric models

The 108 average configurations served to create two databases needed for the purpose of designing parametric models. The first one (denoted as D1) had a knowledge of an overall range of the configurations during a training process, whereas the second one (denoted as D2) did not cover any boundary condition during a training process.

In order to acquire a relation between input and output, 90 configurations (from overall number of 108 configurations) were employed during a training process. In order to verify a performance of the designed parametric estimation models, remaining 18 configurations (common ratio of 80:20) were used during a testing process.

i. Training process of GP-based parametric model

GPlab (Matlab toolbox for a genetic programming developed by Sara Silva [18]) was used for a computational purpose in this research. Some of the GPlab parameters and their values used in the simulations are listed in Table 1. Regarding other parameters, we used default values defined in the GPlab.

The initial population size consisted of 1000 individuals with an initial maximum depth of each tree equals to 6. Ramped half-and-half method was chosen to generate initial population, because it generates trees with a wide variety of sizes and shapes in a ratio of 50:50; which means 50% of identical trees (with the same initial maximum tree depth for all branches) and 50% of different trees (with the different tree depth for all branches considering initial maximum tree depth). Selection of parents needed for a recombination was performed according to a lexicographic parsimony pressure tournament, which prefers shorter individuals to longer ones when their fitness is identical. This strategy helps to reduce a production of complicated individuals and consequently a code growth. A standard tree crossover and mutation were used with probabilities adjusted to equal value. The reproduction rate was set to 0.1. It means that each selected parent has 10% chance of being copied to a next generation without modifications in a tree structure [19]. The maximum tree depth in the case of a creation of offspring by genetic operators was set to 32. New populations were composed of newly-generated offspring only, i.e. replace survival. A function set was defined according to the following studies [4, 16 and 20]. Due to a stochastic nature of the genetic programming, it is required to execute several runs of GP to obtain statistically stable and reliable results. For that reason, 5 independent runs for each database were performed. It would be noted that each run spanned 30 generations with 1000 individuals.

Values of the GP parameters used in the simulations Table 1

GP Parameters	Defined values
Initial population size	1000 individuals
Initial tree depth	6
Initialization of population	Ramped half-and-half method
Selection	Lexicographic parsimony pressure tournament
Genetic operators	crossover; mutation; reproduction
Operator probabilities	0.5, 0.5; 0.1
Survival	Replace
Function set	plus, minus, times, divide, sqrt, power, log, log10, log2, sin, cos
Terminal set	X1, X2, X3, X4 - inputs
Fitness	the sum of the absolute values of the differences between obtained and actual outputs

ii.Training process of RNN-based parametric model

RNNSIM v.2 developed by Abdelbaki [21] was used to implement RNN in MATLAB. We applied two different RNN feed-forward architectures (without circuit in the layers, i.e. signal cannot return back to neuron which has already visited) in a design process of the estimation model. Figure 2 depicts RNN architectures used in our research, namely 3-layer architecture with one hidden layer and 9 neurons on that layer and 4-layer architecture with 5 and 4 hidden neurons situated in two hidden layers respectively. The decision to use these architectures with the corresponding configurations (e.g. number of hidden layers and nodes per hidden layer) was based on other experiment run by the authors and published in [22], where the selected architectures and their particular configurations have achieved the best performance from all investigated conditions. Gradient descent method was used as a learning algorithm because of its simplicity and strong generalization capability even for small training data sets. Due to a stochastic nature of the random neural network, five simulations were conducted for each network architecture and training database in order to obtain statistically stable and reliable results.

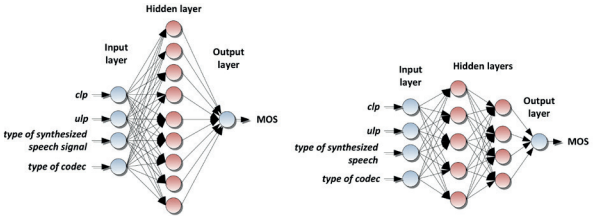

Fig. 2 Architectures of RNN used in the simulations

C. Design of parametric model - testing process

As already mentioned above, the randomly chosen 18 testing configurations (naturally not used in a training process) were used to verify a performance of the designed parametric estimation models during a testing process. The MOS values provided by these models (ranging from 1 "bad quality" to 5 "excellent quality") were compared with the actual MOS values predicted by the PESQ model. On the basis of this comparison, the performance of the designed parametric estimation models was 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 defined in [23].

3. Experimental results

A. Experimental results for GP-based parametric model

As already stated above, five simulations (runs) were conducted for each database. Each simulation spanned 30 generations with 1000 individuals. The estimated MOS values provided by the designed parametric model within a testing phase were compared with the MOS values predicted by PESQ. The best results selected from all realized simulations according to the root mean square error and epsilon-insensitive root mean square error are presented in Table 2. A comparison of actual (predicted by PESQ model) and MOS values estimated by the designed parametric model based on GP approach for both used databases (D1 and D2) is depicted in Figs. 3 and 4, respectively.

The performance results presented in Table 2 obtained for the designed parametric model confirmed very good accuracy provided by GP approach, which is proven by very high values of Pearson correlation coefficient (94% and higher) and very low values of root mean square error (0.104MOS) and epsilon-insensitive root mean square error (0.052MOS). Surprisingly, the performance results are comparable with the results obtained in our previous study [24], where we trained the designed GP-based parametric models separately for each of the three speech codecs (ITU-T G.729AB, ITU-T G.711 and iLBC) involved in this study using smaller training databases (30 configurations). On the basis of this fact, we can conclude that a bigger training database did not help to improve an accuracy of the parametric model based on the GP approach.

Chosen best-of-run individuals — Table 2

Database	D1	D2
R [%]	94.37	98.22
RMSE	0.1106	0.1037
*RMSE **	0.0517	0.0560

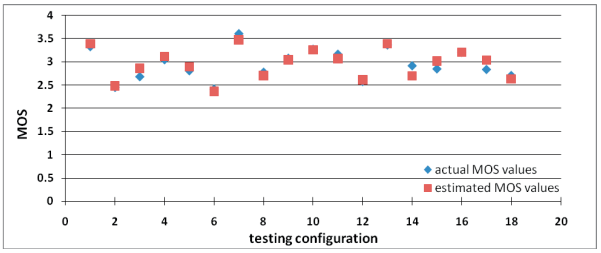

Fig. 3 The actual and estimated MOS values obtained for D1 database and GP approach

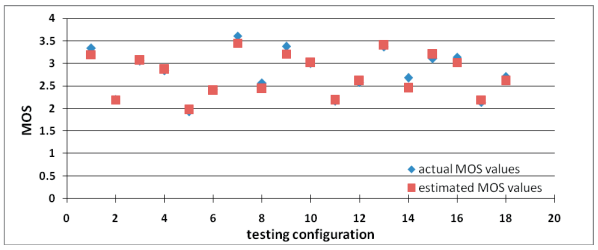

Fig. 4 The actual and estimated MOS values obtained for D2 database and GP approach

B. Experimental results for RNN-based parametric model

Similar to the GP approach, five simulations were conducted for two network architectures of RNN and two experimental databases. The estimated MOS values provided by the designed parametric model were compared with the MOS values predicted by PESQ. The best results selected from all realized simulations according to the root mean square error and epsilon-insensitive root mean square error are presented in Table 3. A comparison of the actual (predicted by PESQ model) and MOS values estimated by the designed parametric model based on RNN for both databases (D1 and D2) are depicted in Figs. 5 and 6, respectively.

The performance results presented in Table 3 also confirmed very high estimation accuracy. Comparing the individual network architectures, 3-layer RNN with 9 hidden neurons seems to provide more accurate results than a 4-layer RNN. This fact is confirmed by lower values of root mean square error and epsilon-insensitive root mean square error. As expected, a higher number of the configurations in a training database provides more accurate estimations in comparison with the estimations obtained in [25] for the RNN-based parametric model designed for speech codec G.729AB (using smaller training database (30

configurations)). It should be noted here that an expansion of the training database was only useful for the RNN-based parametric model. In other words, as can be seen in a previous subsection, a size of the training database did not influence an accuracy of the parametric model based on the GP approach.

Chosen best simulations Table 3

Database/ *RNN* architecture	D1		D2	
	4_9_1	4_5_4_1	4_9_1	4_5_4_1
R [%]	93.99	94.10	98.58	97.84
RMSE	0.1204	0.1278	0.1072	0.1169
*RMSE **	0.0507	0.0521	0.0510	0.0599

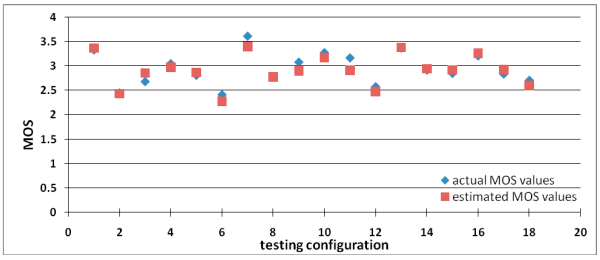

Fig. 5 The actual and estimated MOS values obtained for D1 database and 3-layer RNN architecture (4_9_1)

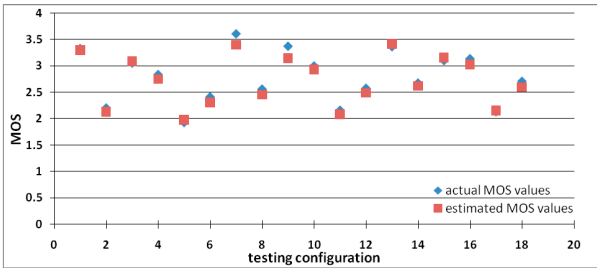

Fig. 6 The actual and estimated MOS values obtained for D2 database and 3-layer RNN architecture (4_9_1)

In addition to a relatively very high accuracy of both designed parametric quality estimation models based on the machine learning techniques, the performance results even confirmed their good generalization ability for new inputs not included in a training process. Comparing both databases (D1 and D2) for both approaches (GP and RNN), the latter one provides (in some cases) a bit more accurate results (except for RMSE* values). It is necessary to note that the D1 database had a knowledge of an overall range of the configurations during the training process, whereas the D2 database did not cover any boundary condition during the training process. This can be explained by a fact that a discovered function by the GP approach and curve characterizing relationship between the perceived quality of the synthesized speech and quality-affecting parameters learned by the RNN are well-defined even beyond the specified interval. In fact, the designed parametric models are able to provide good results even outside the area defined in the training process.

4. Conclusion

In this article, we presented the novel parametric models for a non-intrusive estimation of the speech quality based on biologically inspired machine learning techniques, like a Genetic Programming (GP) and Random Neural Network (RNN). It is worth reiterating that the designed parametric models estimate the quality of synthesized speech transmitted over IP networks (VoIP environment). Therefore, the quality-affecting parameters characterizing packet loss process, speech codec type and a type of synthesized speech signal were considered as an input to the designed parametric estimation models. Outputs of the designed parametric models represent estimated MOS values (ranging from 1 “bad quality” to 5 “excellent quality”). When comparing the performance results obtained for the designed parametric models with the predictions provided by PESQ model, we can conclude that the designed parametric models represent promising candidates for estimating the quality of synthesized speech transmitted over IP networks. Furthermore, the designed parametric models are computationally very efficient and useful for real-time speech quality estimation. The developed parametric models can be helpful for network operators and service providers in a planning phase or early-development stage of telecommunication services based on synthesized speech. Finally, as a performance of both designed parametric models is roughly the same in terms of all monitored performance indicators, we leave up to the intended users to decide, which one of them, they will deploy in their implementations.

Future work will focus on an optimization of the complexity of the function defined by GP approach, as the function has been found very complex.

Acknowledgement

This contribution is the result of the project implementation: Centre of excellence for systems and services of intelligent transport II., ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

„Podporujeme vyskumne aktivity na Slovensku/Projekt je spolufinancovany zo zdrojov EU.”

References

[1] DE RANGO, F., TROPEA, M., FAZIO, P., MARANO, S.: Overview on VoIP: Subjective and Objective Measurement Methods, *IJCSNS Intern. J. of Computer Science and Network Security*, vol.6, No.1B, 2006.

[2] MAHDI, A. E., PICOVICI, D.: Advances in Voice Quality Measurement in Modern Telecommunications, *Digital Signal Processing* 19 (2009), pp.79-103.

[3] MOELLER, S.: *Quality of Telephone-based Spoken Dialogue Systems,* Springer, New York, 2005, ISBN 0-387-23190-0.

[4] RAJA, A., ATIF AZAD, R. M., FLANAGAN, C., RYAN, C.: A Methodology for Deriving VoIP Equipment Impairment Factors for a mixed NB/WB Context, *IEEE Transactions on Multimedia*, vol.10, No. 6, 2008.

[5] RUBINO, G., VARELA, M.: *A New Approach for the Prediction of End-to-end Performance of Multimedia Streams*, Proc. of the First Intern. Conference on the Quantitative Evaluation of Systems (QEST'04), 2004.

[6] KOZA, J. R.: *Genetic Programming: On the Programming of Computers by Means of Natural Selection,* A Bradford book, 1998, ISBN 2-262-11170-5.

[7] POLI, R., LANGDON, W. B., MCPHEE, N. F., KOZA, J. R.: *A Field Guide to Genetic Programming,* 2008, Published via http://dces.essex.ac.uk/staff/rpoli/gp-field-guide/A_Field_Guide_to_Genetic_Programming.pdf. .

[8] ABDELBAKI, H. E.: *Random Neural Network Simulator* (RNNSIM v. 2), for use with MATLAB, September 1999, online: http://www.cs.ucf.edu/~ahossam/rnnsimv2/rnnsimv 2.pdf.

[9] GELENBE, E.: Random Neural Networks with negative and positive Signals and Product Form Solution, *Neural Computation* 1 (4), 1989, pp. 502-510.

[10] DARJAA, S., RUSKO, M., TRNKA, M.: *Three Generations of Speech Synthesis Systems in Slovakia,* Proc. of XI Intern. Conference Speech and Computer (SPECOM 2006), Sankt Peterburg, 2006, pp. 297-302, ISBN 5-7452-0074-X.

[11] ITU-T Rec. G.729: *Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Exited Linear Prediction (CS-ACELP)*, Intern. Telecommunication Union, Geneva (Switzerland), 2007.

[12] ITU-T Rec. G.711: *Pulse Code Modulation (PCM) of Voice Frequencies*, Intern. Telecommunication Union, Geneva, 1988.

[13] IETF RFC 3951: *Internet Low Bit Rate Codec (iLBC)*, Internet Engineering Task Force, 2004.

[14] ITU-T 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,* Intern. Telecommunications Union, Geneva, 2001.

[15] POCTA, P., HOLUB, J.: Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models, *Acta Acustica united with Acustica*, vol. 97, No. 5, pp. 852-868, 2011, ISSN 1610-1928.

[16] RAJA, A., ATIF AZAD, R. M., FLANAGAN, C., PICOVICI, D., RYAN, C.: *Non-Intrusive Quality Evaluation of VoIP Using Genetic Programming,* Bio-Inspired Models of Network, Information and Computing Systems, 2006, pp.1-8.

[17] BASTERRECH, S., RUBINO, G., VARELA, M.: *Single-sided Real-time PESQ Score Estimation*, Proc. of Measurement of Speech, Audio, and Video Quality in Networks (MESAQIN'09), Prague, 2009.

[18] SILVA, S.**:** GPLAB: *A Genetic Programming Toolbox for MATLAB,* Published via http://gplab.sourceforge.net/download.html.

[19] VANNESCHI, L., CASTELLI, M., SILVA, S.: *Measuring Bloat, Overfitting and Functional Complexity in Genetic Programming*, GECCO 2010, pp. 877-884.

[20] RAJA, A., ATIF AZAD, R. M., FLANAGAN, C., RYAN, C.: *Real-Time, Non-intrusive Evaluation of VoIP,* EuroGP'07, LNCS, vol. 4445, Springer, Heidelberg 2007, pp. 217-228.

[21] ABDELBAKI, H. E.: *Random Neural Network Simulator* (RNNSIM v. 2), 1999, online: http://www.cs.ucf.edu/~ahossam/rnnsim.

[22] MRVOVA, M.: *Quality Estimation of Synthesized Speech Signals Transmitted through a Telecommunication Channel*, Ph.D. thesis (available only in Slovak), University of Zilina, 2013.

[23] ITU-T Rec. P.1401: *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models,* Intern. Telecommunication Union , Geneva, 2012.

[24] MRVOVA, M., POCTA, P.: Novel Parameter-based Models Estimating Quality of Synthesized Speech Transmitted over IP Network Based on Genetic Programming Approach, *Microwave and Radio Electronics Week,* 2013, Pardubice, pp. 361-366, ISBN 978-1-4673-5517-9.

[25] MRVOVA, M.: *Novel Parameter-based Model Estimating Quality of Synthesized Speech Transmitted over IP Network Based on Different RNN Architectures*, 10th European Conference of Young Research and Scientific Workers TRANSCOM 2013, 2013, Zilina, Slovakia, pp. 81-84, ISBN: 978-80-554-0692-3.