A method for quantitative evaluation of audio quality over packet networks and its comparison with existing techniques

Samir Mohamed Gerardo Rubino Martín Varela *

May 10, 2004

Abstract

We have recently proposed a novel, non–intrusive, real–time approach to measuring the quality of an audio (or speech) stream transmitted over a packet network. The proposed approach takes into account the diversity of the factors which affect audio quality, including encoding parameters and network impairments. The goal of this method is to overcome the limitations of the quality assessment techniques currently available in the literature, such as the low correlation with subjective measurements, or the need to access the original signal, which precludes real–time applications.

Our approach correlates well with human perception, it is not computationally intensive, does not need to access the original signal, and can work with any set of parameters that affect the perceived quality, including parameters such as FEC, which are usually not taken into account in other methods. It is based on the use of a Random Neural Network (RNN), which is trained to assess audio quality as an average human being.

In this paper we compare the performance of the proposed method with that of other assessment techniques found in the literature.

1 Introduction

Several methods for objectively evaluating speech quality are currently available in the literature. Their main *raison d'être* is to provide a cheaper and more practical alternative to subjective tests (i.e. the one specified in [12]), which are complex, expensive, and take long times to be performed.

Some of the objective methods found in the literature are the Signal-to-Noise Ratio (SNR), Segmental SNR (SNRseg), Perceptual Speech Quality Measure (PSQM) [2], Measuring Normalizing Blocks (MNB) [19], ITU E–model [11], Enhanced Modified Bark Spectral Distortion (EMBSD) [21], Perceptual Analysis Measurement System (PAMS) [17] and PSQM+ [1]. These quality metrics often provide assessments that do not correlate well with human perception, and thus their use as a replacement of subjective tests is limited. Except for the ITU E–model, all these metrics propose different ways to compare the received sample *with the original one*. The E–model allows to obtain an approximation of the perceived quality as a function of several ambient, coding and network parameters, to be used for network capacity planning. However, as stated in [9] and even in its specification [11], its results do not correlate well with subjective assessments either.

The method we propose allows to obtain a good estimation of MOS for speech and audio samples, without needing access to the original sample. This allows us to have a very accurate quality assessment in real–time, which can be useful for controlling quality, or for pricing applications, for example. We believe that the good performances obtained, coupled with the ability to perform them in real–time and with low computational requirements make our approach an interesting alternative to the other methods found in the literature.

The rest of the paper is organized as follows. Section 2 presents our approach, along with some of the math behind the RNN model. In Section 3 we present some per-

*IRISA/INRIA Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France. E-mails: {Samir.Mohamed, Gerardo.Rubino, Martin.Varela}@irisa.fr

formance metrics of other objective assessment methods, which we took from the literature, and in Section 4 we present some of the results we obtained, to provide a form of comparison with the other objective methods. It should be noted that at the moment it is not possible to make a direct comparison of performances, since the data we have comes from several independent sources. We believe, however, that the information we present is representative of the relative performance of the approaches considered. Finally, we present our conclusions in Section 5.

2 Our Approach: Pseudo–subjective Quality Assessment

As discussed in the Introduction, correctly assessing the perceived quality of a speech stream is not an easy task. As quality is, in this context, a very subjective concept, the best way to evaluate it is to have real people do the assessment. There exist standard methods for conducting *subjective* quality evaluations, such as the ITU-P.800 [12] recommendation for telephony. The main problem with subjective evaluations is that they are very expensive (in terms of both time and manpower) to carry out, which makes them hard to repeat often. And, of course, they cannot be a part of an automatic process.

Given that subjective assessment is expensive and impractical, a significant research effort has been done in order to obtain similar evaluations by *objective* methods, i.e., algorithms and formulas that measure, in a certain way, the quality of a stream.

The method used here [15, 14] is a hybrid between subjective and objective evaluation, which can be applied to speech, high-quality audio and even video. The idea is to have several distorted samples evaluated subjectively, and then use the results of this evaluation to teach a Random Neural Network (RNN) the relation between the parameters that cause the distortion and the perceived quality. In order for it to work, we need to consider a set of P parameters (selected *a priori*) which may have an effect on the perceived quality. For example, we can select the codec used, the packet loss rate of the network, the end–to–end delay and/or jitter, etc. Let this set be $\mathcal{P} = \{\pi_1, \ldots, \pi_P\}$. Once these *quality–affecting* parameters are defined, it is necessary to choose a set of representative values for each π_i, with minimal value $\pi_{\min}$ and maximal value $\pi_{\max}$, according to the conditions under which we expect the system to work. Let $\{p_{i1}, \cdots, p_{iH_i}\}$ be this set of values, with $\pi_{\min} = p_{i1}$ and $\pi_{\max} = p_{iH_i}$. The number of values to choose for each parameter depends on the size of the chosen interval, and on the desired precision. For example, if we consider the packet loss rate as one of the parameters, and if we expect its values to range mainly from 0% to 5%, we could use 0, 1, 2 and 5% as the selected values, or in a more conservative way, the set $\{0\%, 1\%, 2\%, 3\%, 5\%, 10\%\}$. In this context, we call *configuration* a set with the form $\gamma = \{v_1, \ldots, v_P\}$, where v_i is one of the chosen values for p_i.

The total number of possible configurations is usually very large. For this reason, the next step is to select a subset of the possible configurations to be subjectively evaluated. This selection may be done randomly, but it is important to cover the points near the boundaries of the configuration space. It is also advisable not to use a uniform distribution, but to sample more points in the regions near the configurations which are most likely to happen during normal use. Once the configurations have been chosen, we need to generate a set of "distorted samples", that is, samples resulting from the transmission of the original media over the network under the different configurations. For this, we use a testbed, or a network simulator.

Formally, we must select a set of M media samples (σ_m), $m = 1, \cdots, M$, for instance, M short pieces of audio (subjective testing standards advise to use sequences having an average 10 sec length). We also need a set of S configurations denoted by $\{\gamma_1, \cdots, \gamma_S\}$ where $\gamma_s = (v_{s1}, \cdots, v_{sP})$, v_{sp} being the value of parameter π_p in configuration γ_s. From each sample σ_i, we build a set $\{\sigma_{i1}, \cdots, \sigma_{iS}\}$ of samples that have encountered varied conditions when transmitted over the network. That is, sequence σ_{is} is the sequence that arrived at the receiver when the sender sent σ_i through the source-network system where the P chosen parameters had the values of configuration γ_s.

Once the distorted samples are generated, a subjective test (e.g. as in [12]) is carried out on each received piece σ_{is}. After statistical processing of the answers, the sequence σ_{is} receives the value μ_{is} (often, this is a *Mean Opinion Score*, or MOS). The idea is then to associate

each configuration γ_s with the value

$$\mu_s = \frac{1}{M} \sum_{m=1}^{M} \mu_{ms}.$$

At this step we have a set of S configurations $\gamma_1, \ldots, \gamma_S$. Configuration s has value μ_s associated with it. We randomly choose S_1 configurations among the S available. These, together with their values, constitute the "Training Database". The remaining $S_2 = S - S_1$ configurations and their associated values constitute the "Validation Database", reserved for further (and critical) use in the last step of the process.

The next step is to train a statistical learning tool (in our case, a RNN) to learn the mapping between configurations and values as defined by the Training Database. Assume that the selected parameters have values scaled into [0,1] and the same with quality. Once the tool has "captured" the mapping, that is, once the RNN is trained, we have a function $f()$ from $[0,1]^P$ into $[0,1]$ mapping now any possible value of the (scaled) parameters into the (also scaled) quality metric. The last step is the validation phase: we compare the value given by $f()$ at the point corresponding to each configuration γ_s in the Validation Database to μ_s; if they are close enough for all of them, the RNN is validated (in Neural Network Theory, we say that the tool *generalizes well*). In fact, the results produced by the RNN are generally closer to the MOS than that of the human subjects (that is, the error is less than the average deviation between human evaluations). As the RNN generalizes well, it suffices to train it with a small (but well chosen) part of the configuration space, and it will be able to produce good assessments for any configuration in that space. The choice of the RNN as an approximator is not arbitrary. We have experimented with other tools, namely Artificial Neural Networks, and Bayesian classifiers, and found that RNN perform better in the context considered. ANN exhibited some problems due to over-training, which we did not find when using RNN. As for the Bayesian classifier, we found that while it worked, it did so quite roughly, with much less precision than RNN. Besides, it is only able to provide discrete quality scores, while the NN approach allows for a finer view of the quality function.

The neural network model used has some interesting mathematical properties, which allow, for example, to obtain the derivatives of the output with respect to any of the inputs, which is useful for evaluating the performance of the network under changing conditions (see next section).

The method proposed produces good evaluations for a wide range variation of all the quality affecting parameters, at the cost of one subjective test.

2.1 Random Neural Network (RNN) description

Let us briefly describe the way we can use a specific class of queuing networks as a very efficient statistical learning tool. The mathematical object and its use in learning was introduced and developed in [4, 6, 7].

An RNN is an open Markovian queuing network with positive and negative customers, also called a G-network. We have N nodes (or neurons) which are $./M/1$ queues (the service rate of node i is denoted by ν_i), interconnected, receiving customers from outside and sending customers out of the network. Customers are "positive" or "negative"; the arrival flow of positive (respectively negative) customers arriving at node i from outside is Poisson with rate λ_i^+ (respectively λ_i^-). After leaving neuron (queue) i, a customer leaves the network with probability d_i, goes to queue j as a positive customer with probability r_{ij}^+ and as a negative customer with probability r_{ij}^-. When a negative customer arrives at a node i (either from outside or from another queue) it disappears, removing the last customer at i, if any. Transfers between queues are, as usual with queuing network models, instantaneous. This means that negative customers can not be observed; at any point in time there are only positive customers in the network; negative customers act only as *signals*, modifying the behavior of the system.

Let us denote by X_t^i the number of customers in queue i at time t. Then, it was proved in [5, 6] that when the (Markov) process $\vec{X}_t = (X_t^1, \cdots, X_t^N)$ is stable, its stationary distribution is of the product-form type: that is, assuming that $(\vec{X}_t)$ is stationary, we have

$$\Pr(\vec{X}_t = (k_1, \cdots, k_N)) = \prod_{i=1}^{N} (1 - \varrho_i) \varrho_i^{k_i}.$$

The factors $\varrho_1, \cdots, \varrho_N$ in this expression are the loads of the nodes in the network. The specificities of these networks make that these loads are not obtained by solving a

linear system (as in the Jackson case) but by solving the following non-linear one:

$$\varrho_i = \frac{\lambda_i^+ + \sum_{j=1}^{N} \varrho_j \nu_j r_{ji}^+}{\nu_i + \lambda_i^- + \sum_{k=1}^{N} \varrho_k \nu_k r_{ki}^-}.$$

It can then be proved that when this system has a solution $\varrho_1, \cdots, \varrho_N$ such that for each node i it is $\varrho_i < 1$, then the process is stable, and the product-form result holds (see [5]).

To use such a queuing network as a learning tool, we perform the following mapping: the input variables (codec, FEC offset, loss rate, etc.) are scaled into [0,1] and then associated with the external arrival rates of positive customers at P specific nodes of the network $\lambda_1^+, \cdots, \lambda_P^+$. The remaining external arrival rates of positive customers are set to 0; we also set to zero the external rates of negative customers. The quality of the sequence after being also normalized into [0,1], is mapped to the load of a specific node *o* in the system. The problem now is to find a network such that when $\lambda_1^+ = v_{1s}, \cdots, \lambda_P^+ = v_{Ps}$, then the load of the chosen node *o* is close to μ_s. This is an optimization problem where the control variables are now the remaining parameters of the network: the service rates ν_i and the routing probabilities r_{ij}^+ and r_{ij}^-.

For all neurons i such that $d_i < 1$ (that is, for all neuron that does not send all its signals (its customers) out of the network), we denote $w_{ij}^+ = \nu_i r_{ij}^+$ and $w_{ij}^- = \nu_i r_{ij}^-$. These w_*^* factors are called *weights* as in the standard neural network terminology, and they play a similar role in this model. Instead of optimizing with respect to the service rates and the transition probabilities, the standard approach is to do it with respect to the weights, and just to keep constant the service rates of the "output" neurons (those neurons j where $d_j = 1$).

The optimization problem can be solved using standard techniques such as gradient descent (observe that we are able to compute any partial derivative of the output, using the non-linear system of equations satisfied by the occupation rates).

3 Performance of other Objective Speech Quality Measures

In order to have an idea about the performance of the known objective speech quality measures (namely the ones that have been introduced in Section 1), we have collected some data and figures from the literature.

We present the results found in the literature in two parts: the first one shows the performance of these metrics when assessing speech quality with only encoding impairments considered; the second one when they are used with both encoding and network impairments. Depending on the available data, the compared metrics are SNR, SNRseg, BSD, MBSD, EMBSD, PSQM, PSQM+, MNB(1,2), E-model and PAMS.

Table 1 shows the correlation coefficients for the metrics considered when assessing the impact of encoding, and MOS. It is easy to see that the simple metrics like SNR and SNRseg yield relatively poor correlation with subjective quality tests (ranging from 0.22 to 0.52 for SNR and from 0.22 to 0.52 for SNRseg). BSD, MBSD and EMBSD give better results than SNR or SNRseg.

PSQM and its enhanced version PSQM+, exhibit better performances, comparable to that of PAMS and MNB. The correlation coefficient can reach up to 0.98 for certain metrics. The variation in the correlation values is due to different levels of distortion being considered.

However, when networking parameters are also taken into account (cf. Table 2), the performance of these metrics suffers a sometimes very significant drop. The most comprehensive comparative study of the performance of objective assessment metrics we found is the one in [21]. Note, however, that the authors did not disclose all the measured performances explicitly. This does not hinder our study, since we are interested in knowing the best performances for the other metrics, and so these values are sufficient. We can see that there is a general drop in performance when some form of network impairment is considered, even if these impairments are not necessarily those specific to VoIP applications (the impairments considered where temporal shifting, front clipping, bit errors, frame erasures and level variations, which correspond better to wireless networks, such as GSM or CDMA). It should be noted that the correlation coefficients were calculated for regression curves used to fit the results and not

Objective Measures	Correlation with Subjective Measure (MOS)
SNR	0.22-0.52
SNRseg	0.22-0.52
BSD	0.36-0.91
PSQM	0.83-0.98
PSQM+	0.87-0.98
MNB2	0.74-0.98
PAMS	0.64-0.89
MBSD	0.76
EMBSD	0.87

Table 1: Correlation coefficient of some existing objective speech quality measures with MOS. Results are taken from the literature. Only the encoding impairments are considered. Sources are [21, p. 103], [19, p. 84], [18, p. 1517] and [17, p. 10/7].

for the metrics themselves, which may lead to a slightly higher correlation than that of the actual output of the metrics and MOS values.

Table 3 shows results taken from [21, P. 103] and [9], which show the correlation coefficients obtained for the listed metrics and MOS values when used with real VoIP traffic (the E-model values are taken from [9], since the E-model is not considered in [21]). The higher values correspond to the the results obtained in [21]. The authors did not specify the network conditions used, but only stated that it was voice traffic recorded over a real network. In [9], the network parameters considered are loss rate with values 0%, 1% and 5% (the distribution of losses within the streams was not specified), and jitter values of 0, 50 and 100ms. We believe that the difference in performance of the different metrics for both tests is due to differences in network conditions (which may have been harsher in [9]), resulting in more damaged samples, which may have affected the metrics' performance.

From the objective speech quality measures considered, only the ITU E-model does not need access to the original signal to compute the quality. Thus, it is the only available measure which is computationally simple [9] and can be used in real-time applications. However, as stated in its specification [11], and from the results reported in [9], the E-model was designed as a network planning tool, and not as a precise quality metric. Besides, it is still in development; for instance, as of March 2003, explicit consideration of network losses was a recent addition, and then it was only independent losses, with no provision for loss bursts.

As a final example of the performance problems currently found in objective quality metrics, in Figure 1 we present two scatter plots (taken from [9]) for MNB2 and the E-model. In these plots, there are a very important number of values that are clearly inconsistent with MOS results. For example, there are points that correspond to the same values of MNB or the E-model R rating, but have very different MOS values (there are differences of 2 and 3 MOS points, which are highly significant),and vice-versa.

4 Performance of Our Approach

In this section we present the results we obtained with our approach for two different VoIP test campaigns we have performed.

For the first battery of tests we considered the parameters listed in Table 4.

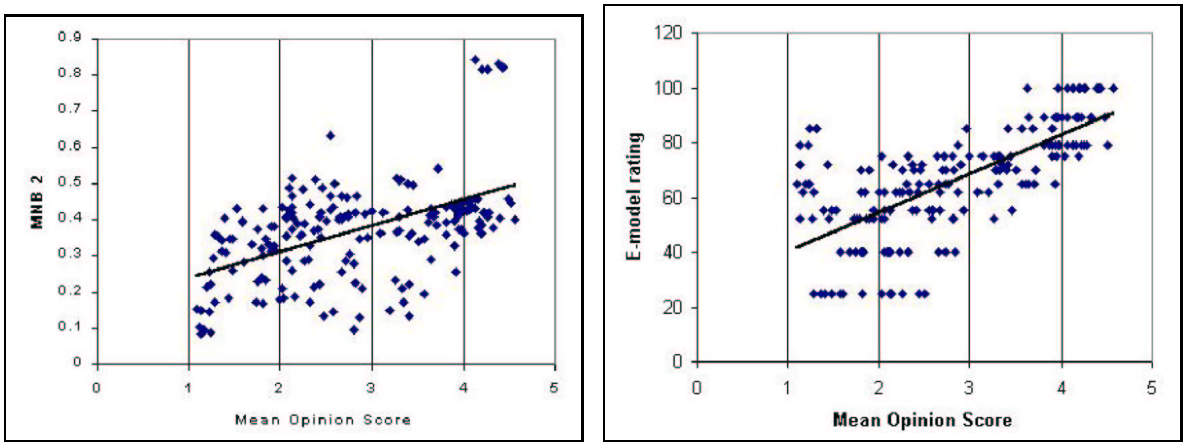
With these parameters, we simulated the network effects on encoded files, and used these files to conduct MOS tests (as specified in [12]) in three languages (French, Spanish and Arabic). Once the MOS results were screened, we proceeded as described in Section 2, and trained three RNN, one for each language considered. The results obtained were very good, with correlation coefficients of 0.99 for Spanish and Arabic, and 0.98 for French (using only validation data). Figure 2 shows scatter plots for these tests.

Objective Measures	Correlation with Subjective Measure (MOS)
A	0.87
B	0.85
C	0.56
D	0.86
E	0.90
F	0.86
MBSD	0.24
EMBSD	0.54

Table 2: Correlation coefficient of some existing objective speech quality measures with MOS for encoding impairments and some network impairments which can be found on GSM or CDMA networks, and some of the effects found on IP networks. Source is [21, P. 106]. The letters A to F represent the objective quality measures mentioned in Sec. 3. Note that not all the metrics evaluated were explicitly named in the study.

Objective Measures	Correlation with Subjective Measure (MOS)
EMBSD	0.39 – 0.87
MNB1	0.61 – 0.83
MNB2	0.63 – 0.74
E-model	0.62 – 0.86

Table 3: Correlation coefficients for EMBSD, MNB(1 & 2) and E-model with MOS for VoIP impairments, taken from [21, P. 103] and [9].



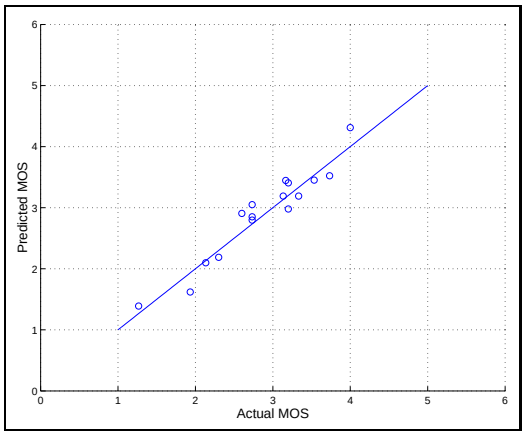
(a) Performance of MNB2

(b) Performance of the E-model

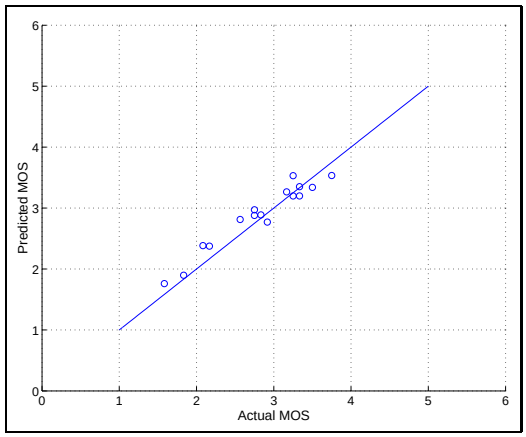
Figure 1: MNB2 and E-model results against the MOS subjective values in evaluating a set of speech samples distorted by both encoding and network impairments (taken from [9]).

Parameter	Values
Loss rate	0%...40%
Loss burst size (constant, as in [10])	1...5
Codec	PCM Linear-8, G726 and GSM-FR
Packetization interval	20, 40, 60 and 80ms

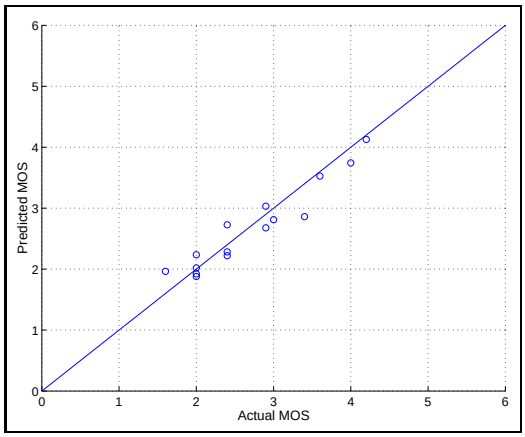
Table 4: Network and encoding parameters and values used for the first test set



(a) Spanish samples – Correlation Coefficient = 0.99



(b) Arabic samples – Correlation Coefficient = 0.99



(c) French samples – Correlation Coefficient = 0.98

Figure 2: Scatter plots for the first series of tests. Estimations are for validation data (never seen before by the RNN).

For the second set of tests, we refined our network model, substituting the simple loss model suggested in [10] for a simplified Gilbert model [8], like the one suggested in [20, 3]. The distorted speech samples were generated on a live network using the Robust Audio Tool (RAT [13]) and a proxy that generated the losses as specified (this software is an adaptation of O. Hodson's packet reflector [16]) on a live network. A MOS test was performed and the results screened as per [12]. We tested several RNN architectures and various combinations of training/validation database sizes, and found good results using about 100 samples for training, and 10 for validation. We also considered Forward Error Correction (FEC) parameters in these tests. The parameters considered for our experiment are listed on table 5. The results obtained varied with the different sizes of training/validation databases, and yielded correlation coefficients between 0.73 and 0.93 with actual MOS values. It is interesting to see that even when using relatively small sets of training samples, very good results can be obtained, and this allows for a trade-off between cost and performance for our method (since its main cost is that of performing the subjective tests to train the RNN). Figure 3 shows a scatter plot for the validation data of the second set of tests.

5 Conclusions

In this paper we present a novel approach to assessing VoIP quality in a non–intrusive way, in real–time if necessary, and with very good results with respect to subjective quality assessment. Our approach is based on the use of a Random Neural Network trained with the results of a MOS test performed on a suitable set of distorted speech samples.

We provide an evaluation of our method's performance when applied to one–way speech flows, both on a simulated environment and on a live network, with real audio-conferencing software. We found that this performance is better than the performance of other objective assessment methods found in the literature under similar conditions. Although no direct comparison is possible with the data currently available, we believe that the figures we provide give a good estimation of the relative performance of the metrics considered.

Our method has other benefits, such as being easily applicable to other kinds of media, such as *hi-fi* audio or video, and the ability of providing a good assessment in real–time, since it does not need to access the original media in order to work.

References

[1] J. Beerends. Improvement of the p.861 perceptual speech quality measure. ITU-T SG12 COM-34E, December 1997.

[2] J. Beerends and J. Stemerdink. A perceptual speech quality measure based on a psychoacoustic sound representation. *Journal of Audio Eng. Soc.*, 42:115–123, December 1994.

[3] J-C. Bolot and A. Vega Garcia. The case for FEC-based error control for packet audio in the Internet. In *ACM Multimedia Systems*, 1996.

[4] E. Gelenbe. Random neural networks with negative and positive signals and product form solution. *Neural Computation*, 1(4):502–511, 1989.

[5] E. Gelenbe. Stability of the random neural network model. In *Proc. of Neural Computation Workshop*, pages 56–68, Berlin, West Germany, February 1990.

[6] E. Gelenbe. G-networks: new queueing models with additional control capabilities. In *Proceedings of the 1995 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 58–59, Ottawa, Ontario, Canada, 1995.

[7] E. Gelenbe and K. Hussain. Learning in the multiple class random neural network. *IEEE Trans. on Neural Networks*, 13(6):1257–1267, 2002.

[8] E. Gilbert. Capacity of a burst–loss channel. *Bell Systems Technical Journal*, 5(39), September 1960.

[9] T. A. Hall. Objective speech quality measures for Internet telephony. In *Voice over IP (VoIP) Technology, Proceedings of SPIE*, volume 4522, pages 128–136, Denver, CO, USA, August 2001.

Parameter	Values
Loss rate	0%...15%
Mean loss burst size	1...2.5
Codec	PCM Linear 16 bits, GSM
FEC	ON(GSM)/OFF
FEC offset	1...3
Packetization interval	20, 40, and 80ms

Table 5: Network and encoding parameters and values used for the second test set

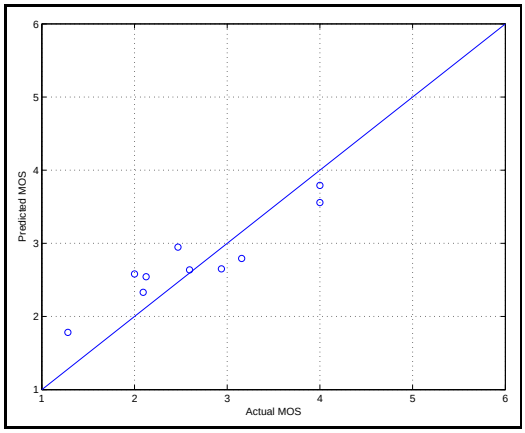


Figure 3: Scatter plot for the second series of tests – Correlation Coefficient = 0.93. Estimations are for validation data (never seen before by the RNN).

[10] D. Hands and M. Wilkins. A study of the impact of network loss and burst size on video streaming quality and acceptability. In *Interactive Distributed Multimedia Systems and Telecommunication Services Workshop*, October 1999.

[11] ITU-T Recommendation G.107. The E-model, a computational model for use in transmission planning.

[12] ITU-T Recommendation P.800. Methods for subjective determination of transmission quality.

[13] University College London. Robust Audio Tool website. http://www-mice.cs.ucl.ac.uk/multimedia/software/rat/index.html.

[14] S Mohamed and G. Rubino. A study of real–time packet video quality using random neural networks. *IEEE Transactions On Circuits and Systems for Video Technology*, 12(12):1071 –1083, December 2002.

[15] Samir Mohamed, Gerardo Rubino, and Martín. Varela. Performance evaluation of real-time speech through a packet network: a random neural networks-based approach. *Performance Evaluation*, 57(2):141–162, 2004.

[16] Hodson, O. Packet Reflector.

[17] A. Rix. Advances in objective quality assessment of speech over analogue and packet-based networks. In *the IEEE Data Compression Colloquium*, London, UK, November 1999.

[18] A. Rix and M. Hollier. The perceptual analysis measurement system for robust end-to-end speech assessment. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP*, pages 1515–1518, Istanbul, Turkey, June 2000.

[19] S. Voran. Estimation of perceived speech quality using measuring normalizing blocks. In *IEEE Workshop on Speech Coding For Telecommunications Proceeding*, pages 83–84, Pocono Manor, PA, USA, September 1997.

[20] M. Yajnik, S. Moon, J.F. Kurose, and D.F. Towsley. Measurement and modeling of the temporal dependence in packet loss. In *Proccedings of IEEE INFOCOM '99*, pages 345–352, 1999.

[21] W. Yang. *Enhanced Modified Bark Spectral Distortion (EMBSD): an Objective Speech Quality Measrure Based on Audible Distortion and Cognition Model*. PhD thesis, Temple University Graduate Board, May 1999.