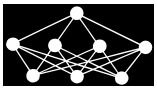

AN IMPROVED E-MODEL USING ARTIFICIAL NEURAL NETWORK VoIP QUALITY PREDICTOR

*Mousa AL-Akhras, Iman ALMomani, Azzam Sleit**

Abstract: Voice over Internet Protocol (VoIP) networks are an increasingly important field in the world of telecommunication due to many involved advantages and potential revenue. Measuring speech quality in VoIP networks is an important aspect of such networks for legal, commercial and technical reasons. The E-model is a widely used objective approach for measuring the quality as it is applicable to monitoring live-traffic, automatically and non-intrusively. The E-model suffers from several drawbacks. Firstly, it considers the effect of packet loss on the speech quality collectively without looking at the content of the speech signal to check whether the loss occurred in voiced or unvoiced parts of the signal. Secondly, it depends on subjective tests to calibrate its parameters, which makes it applicable to limited conditions corresponding to specific subjective experiments. In this paper, a solution is proposed to overcome these two problems. The proposed solution improves the accuracy of the E-model by differentiating between packet loss during speech and silence periods. It also avoids the need for subjective tests, which makes it extendable to new network conditions. The proposed solution is based on an Artificial Neural Networks (ANN) approach and is compared with the accurate Perceptual Evaluation of Speech Quality (PESQ) model and the original E-model to confirm its accuracy. Several experiments are conducted to test the effectiveness of the proposed solution on two well-known ITU-T speech codecs; namely, G.723.1 and G.729.

Key words: *Voice over IP, artificial neural network, speech quality, E-model, non-intrusive, voiced, unvoiced, perceptual evaluation of speech quality, packet loss, subjective-free*

Received: February 16, 2009
Revised and accepted: January 18, 2011

*Mousa AL-Akhras, Iman ALMomani, Azzam Sleit
King Abduallah II School for Information Technology, The University of Jordan, Amman, 11942, Jordan, E-mail: mousa.akhras@ju.edu.jo, i.momani@ju.edu.jo, asleit@ju.edu.jo

1. Introduction

Transmitting Voice over IP networks (VoIP) is an increasingly important application in the telecommunications world due to its advantages, including [6, 19]:

- Lower bandwidth requirements.
- Reduced cost for long-distance calls.
- Integration of voice and data applications into one unified network, which reduces operation and management expenses and makes the creation of new and innovative applications possible.
- Enabling live broadcasting of radio and TV channels.

Many enterprises and network operators adopted VoIP technology to achieve some or all of the above advantages and some of the potential revenue by the traditional Public Switched Telephone Networks (PSTN) operators. In order to compete with the PSTN networks, VoIP networks should achieve comparable quality to that achieved by the highly-reputable telephony networks to meet customers' quality expectations from any potential competitor.

Measuring the quality of VoIP networks is important for legal, commercial and technical reasons. Customers and companies are bound by a service level agreement usually requiring the company to provide a certain acceptable quality. Also, measuring quality allows network administrators to overcome temporal problems that may affect the quality of ongoing voice calls. It also allows service providers to evaluate their own and their competitors' service on a standard scale. It is also a strong indicator of user's satisfaction of the service provided. In doing so, a specialized speech quality measurement mechanism is needed [28].

There are many methods for measuring the quality of a voice call. The selection of a method for this task must take the characteristics of IP networks and voice calls into consideration. Such characteristics that affect the selection include the requirement to measure the quality while the network is running in a real environment during a voice call. Therefore, it is necessary to use an automated solution that measures quality without human interference depending on the received signal at the receiver side without the need for the reference (original) speech signal at the sender side, i.e. non-intrusively. The model that satisfies the above requirements is the E-model which was proposed by the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T).

However, the E-model suffers from several drawbacks that affect its usage. First, when the E-model considers the effect of packet loss on the received quality, it does not take the contents of the received signal into consideration in the estimation of the quality and packet loss is taken as a whole without considering whether packet loss occurs in the voiced or the unvoiced parts of the signal, i.e. during speech or silence periods. Also the E-model requires subjective tests to calibrate its parameters [9]. The inherent problems of subjective tests are that they are hard-to-conduct (as they require strict lab conditions), time-consuming, expensive, and lack repeatability.

In this paper, a solution is proposed to extend the E-model to any new network conditions and for newly emerging speech codecs without the need for the subjective tests. The proposed solution also improves the accuracy of the E-model by differentiating between packet loss during speech and silence periods. The proposed solution is based on an Artificial Neural Network (ANN) model and is compared against the more accurate Perceptual Evaluation of Speech Quality (PESQ) [15] and the original E-model [9] to check its prediction accuracy. The performance of the system is tested using two well-known ITU-T speech codecs G.723.1 [11] and G.729 [12].

The rest of this paper is organized as follows: Section 2 reviews the main methods used for measuring the speech quality. It also discusses the E-model and the main problems associated with it. Section 3 discusses the proposed technique to avoid the use of subjective parameters and to improve the E-model's accuracy. Section 4 studies the design and the performance of ANNs in estimating the quality, and Section 5 presents the results of applying the ANN model in quality estimation. Section 6 summarizes the paper and presents avenues for future work.

2. Assessment Technologies for Measuring VoIP Quality

Speech quality assessment techniques can be categorized into three main classes: subjective assessment techniques, intrusive objective assessment techniques, and non-intrusive objective assessment techniques.

2.1 Subjective assessment of speech quality

The user's perception of service quality or subjective quality is the primary criterion for voice and video communication. The most widely used subjective quality assessment methodology is opinion rating standardized in ITU-T Recommendation P.800 [13]. The most common metric in opinion rating is Mean Opinion Score (MOS) metric which is a five-point scale (5) Excellent, (4) Good, (3) Fair, (2) Poor, and (1) Bad [13]. MOS is internationally accepted metric as it provides a direct link to the quality, as perceived by the user. MOS score is obtained as an arithmetic mean for a collection of MOS scores for a set of subjects [13, 14, 28, 29].

The problem with MOS measurement and subjective tests in general is the difficulty in performing such tests as they require strict conditions regarding the lab settings and the subjects participating in the subjective tests [13]. The inherent problems in subjective MOS measurement are that it is: time-consuming, expensive, lacks repeatability, and inapplicable for monitoring live traffic as commonly needed for VoIP applications. This has made objective methods very attractive to estimate the subjective quality for meeting the demand for voice quality measurement in communication networks. However, subjective methods are used to calibrate objective methods as they are the most accurate methods for measuring speech quality.

2.2 Intrusive objective assessment of speech quality

Intrusive methods for measuring speech quality are full reference as they require the reference speech signal to measure speech quality. The most prominent intrusive method is Perceptual Evaluation of Speech Quality (PESQ), which is the latest ITU-T standard for objective evaluation of speech quality standardized in ITU-T Recommendation P.862. PESQ measurements are highly correlated with subjective tests with a correlation factor of 0.935 on 22 benchmark experiments which cover 9 different languages [15, 16, 17, 24, 30].

In PESQ, the reference and the degraded signals are time-aligned, then both signals are transformed to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness. After this transformation, the reference signal is compared with the degraded signal using a perceptual model and distortion during silence is given different weight and hearing threshold is taken into account [15, 24].

PESQ score lies in the range -0.5 to 4.5. A function is provided in ITU-T Recommendation P.862.1 to map these values to the range 1 to 5 representing MOS. Recommendation P.862.1 also provides a formula to move back to PESQ score from an available MOS score [16]. PESQ and similar full-reference methods provide an accurate measurement of speech quality, however, such methods are inapplicable in monitoring live traffic because it is difficult or impossible to obtain the reference signal at the receiver side.

2.3 Non-intrusive objective assessment of speech quality

Subjective methods and intrusive methods for measuring the quality cannot be used in monitoring live traffic, this makes non-intrusive methods the only available solution for monitoring the quality of live traffic in VoIP networks. One of the most widely used methods for objectively evaluating speech quality non-intrusively is opinion modeling. The most famous standard for opinion modeling is the E-model which is standardized in ITU-T Recommendation G.107 [9, 28].

The E-model was used in an enormous number of studies for the purpose of network planning or to help network operators in designing the network or to perform live monitoring of the network. Now the E-model is being used for objectively estimating voice quality for VoIP applications using network and terminal quality parameters. The E-model is a non-intrusive method as it does not use the reference signal in the estimation of the quality as the estimation is based purely on the terminal and network parameters [9, 27, 28].

The E-model starts by calculating the degree of quality degradation due to individual quality factors on the same psychological scale, therefore the E-model is able to describe the effect of several impairments occurring simultaneously. The sum of these values is then subtracted from a reference value to produce the output of the E-model which is a single scalar value called the R-Rating factor. The R-Rating factor lies in the range of 0 and 100 to indicate the level of estimated quality where $R = 0$ represents an extremely bad quality and $R = 100$ represents a very high quality. The computed R-Rating factor can be mapped into an MOS value based on ITU-T Recommendation G.107. Recommendation G.107 also provides a

formula to move back to R-Rating factor from an available MOS score [9, 18, 20]. The R-Rating factor is calculated according to the following formula:

$$R = R_0 - Is - Id - Ie\text{-}eff + A \tag{1}$$

where

R_0	Basic signal-to-noise ratio to group the effects of noise
Is	Impairments which occur simultaneously with the voice signal
Id	Impairments due to delay, echo
$Ie\text{-}eff$	Impairments due to codec distortion, packet loss and jitter
A	Advantage (expectation) factor (e.g. 0 in landline and 10 in cellular networks)

When all parameters are set to their default values, the R-Rating factor as defined in equation (1) has the value of 93.2 which is mapped into an MOS value of 4.41.

Packet loss dependent Effective Equipment Impairment Factor ($Ie\text{-}eff$) in equation (1) characterizes quality degradation due to packet loss. In this paper, the effect of other parameters will not be considered and the default values for all the parameters except $Ie\text{-}eff$-related parameters will be used. For example, Id will be set to zero. $Ie\text{-}eff$ is calculated according to the following formula [9]:

$$Ie\text{-}eff = Ie + (95 - Ie).\frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{2}$$

where

Ie	Codec-specific Equipment Impairment Factor
Bpl	Codec-specific Packet-loss Robustness Factor
Ppl	Packet loss Probability
$BurstR$	Burst Ratio (to count for burstiness in packet loss)

$Ie\text{-}eff$ – as defined in equation (2) – is derived using codec-specific values for Ie and Bpl at zero packet-loss. The values for Ie and Bpl for several codecs are listed in ITU-T Recommendation G.113 Appendix I [10] and they were derived using subjective MOS test results. On the other hand, Ppl and $BurstR$ depend on the packet loss properties presented in the system. $BurstR$ is defined in the latest version of the E-model [9] as:

$$BurstR = \frac{\text{Average length of observed bursts in an arrival sequence}}{\text{Average length of bursts expected for the network under “random” loss}} \tag{3}$$

When packet loss is random, i.e. independent, $BurstR = 1$ and when packet loss is bursty, i.e. dependent, $BurstR > 1$.

The E-model is a good choice for the non-intrusive estimation of voice quality in VoIP networks. However, the E-model suffers from several drawbacks that affect its usage. These drawbacks concern the application of equation (2) to compute $Ie\text{-}eff$.

When the E-model considers the effect of packet loss on quality, it takes packet loss as a whole and it does not pay attention to the contents of lost packets and whether the loss occurred in voiced or unvoiced parts of the signal. Packet loss during voiced parts (speech periods) would have a greater degradation effect on the received signal than packet loss in unvoiced parts (silence periods) [23, 26]. During many experiments that led to the current paper, comparisons were made between quality estimated from the E-model and quality calculated from PESQ, it was observed that there is a deviation between quality estimation calculated according to the E-model and quality measurements according to PESQ. The work presented in this paper attempts to remove or decrease this deviation.

Also, the E-model requires subjective tests to calibrate some of the parameters used in equation (2), specifically Ie and Bpl [9]. The inherent problems of subjective tests are that they are hard-to-conduct (as they require strict lab conditions), time-consuming, expensive and lack repeatability.

In this paper, a solution is proposed to extend the E-model to new network conditions and for newly emerging speech codecs without the need for the subjective tests by avoiding the use of Ie and Bpl. The proposed solution also improves the accuracy of the E-model by differentiating between packet loss during speech and silence periods. The proposed solution is based on an ANN model and is compared with the more accurate PESQ [15] and with the original E-model [9] to check its prediction accuracy. The effectiveness of the proposed technique is tested using two speech codecs; namely, G.723.1 and G.729.

3. The Proposed Technique

Artificial Neural Networks and Genetic Algorithms can be used to solve various kinds of problems varying from function approximation and clustering problems to function maximization or minimization problems. These techniques have been used in an enormous number of studies by the authors of this paper and other authors to aid in solving various problems such as image processing, finding certain patterns and networking problems, such as load balancing, finding the best route and, in the current paper, to aid in the estimation of quality in VoIP networks [1, 2, 3, 5, 7, 8, 21, 22, 25].

Recall that in the E-model, packet loss dependent Effective Equipment Impairment Factor ($Ie\text{-}eff$) is characterized by equation (2). The equation has four parameters: Equipment Impairment Factor (Ie), Packet-loss Robustness Factor (Bpl), packet-loss probability (Ppl) and burst ratio ($BurstR$). We identified two problems with the applicability and accuracy of this equation which, in turn, affect the applicability and the accuracy of the E-model in voice quality prediction. The first problem is that the values for Ie and Bpl, which are codec-specific, are derived using the time-consuming and expensive subjective MOS tests. The values for several codecs derived using subjective tests are listed in ITU-T Recommendation G.113 Appendix I [10]. For G.723.1 speech codec, the values for Ie and Bpl are 15 and 16.1, respectively. For G.729 their values are 11 and 19, respectively. The second problem is that the other two parameters, Ppl and $BurstR$, consider the effect of packet loss collectively without testing whether packet loss occurred

during voiced or unvoiced parts of the signal as packet loss during different parts of the signal has a different perceptual effect on perceived quality.

In previous publications by the authors, only the first problem was solved as the dependency of the E-model on subjective parameters (Ie and Bpl) was avoided. This was achieved by deriving a new formula to relate *Ie-eff* with Ppl and $BurstR$ in the absence of Ie and Bpl [1, 2, 3]. As this is a function approximation problem, several methods can be used to derive such a new formula, including: Genetic Algorithms (GA) [1], Artificial Neural Networks (ANN) [2] and statistical methods, such as linear and non-linear regression [3]. Using the previous function approximation techniques, *Ie-eff* has only two input parameters which are Ppl and $BurstR$, while Ie and Bpl parameters are integrated in the derived linear and non-linear regression equations or absorbed in the form of weights and biases in the ANN, thereby, the subjectivity of the E-model is avoided.

Considering the second problem, packet-loss probability (Ppl) and burst ratio ($BurstR$) depend on packet loss properties presented in the system. Ppl and $BurstR$ represent the overall packet loss as the E-model does not look at the contents of received signal as it considers the effect of packet loss on received quality collectively; i.e. no distinction is made between packet loss during voiced or unvoiced parts of the signal during speech or silence periods.

Previous studies [26, 27] have shown that packet loss during voiced parts of the signal has a more perceptual effect on the quality than packet loss during unvoiced parts of the signal. The E-model can be modified so that it considers the content of the lost frames and whether they represent voiced parts of the signal or unvoiced parts.

To improve the accuracy of the E-model and bring its estimation closer to the PESQ measurement of quality as PESQ is more accurate due to its intrusive nature, we classified packet loss into either Voiced or Unvoiced loss to give different weights for different classes of loss, the new weights were derived using a Genetic Algorithms (GA) approach. In this case, Ppl is broken into Voiced Ppl (Ppl_{Voiced}) and Unvoiced Ppl ($Ppl_{Unvoiced}$). Similarly, the BurstR is broken into $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$.

The classification of lost packets into voiced or unvoiced is based on surrounding received packets based on the fact that the shape of the vocal tract and its mode of excitation change relatively slowly. Therefore, speech signal can be considered to be quasi-stationary over a short period of time, which allows it to show high degree of predictability. *Ie-eff* can be calculated using a modified equation

$$Ie\text{-}eff = Ie + (95 - Ie).\frac{newPpl}{\frac{newPpl}{newBurstR} + Bpl} \tag{4}$$

where

$$newPpl = \alpha_V.Ppl_{Voiced} + \alpha_U.Ppl_{Unvoiced} \tag{5}$$

and

$$newBurstR = \alpha_V.BurstR_{Voiced} + \alpha_U.BurstR_{Unvoiced} \tag{6}$$

In equation (4), the subjective parameters Ie and Bpl are used as in the original E-model as the purpose is to consider the effect of packet loss using voiced and

unvoiced parts of the signal to improve the accuracy of the E-model and bring its estimation closer to the PESQ measurement of the quality without avoiding the use of subjective parameters (*Ie* and *Bpl*). In other words, this would solve only the second problem presented in the original *Ie-eff* equation.

The algorithm for calculating Ppl_{Voiced}, $Ppl_{Unvoiced}$, $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$ first starts with applying a Voice Activity Detector (VAD) algorithm on the received packets to classify received packets into either voiced or unvoiced. Then, missing packets are classified into either voiced or unvoiced based on surrounding received packets utilizing the fact that speech signal has quasi-stationary characteristics. This classification is not 100% accurate, but conducted experiments indicate that the prediction accuracy is 87.35%, which is significantly better than considering the effect of packet loss collectively.

Ppl_{Voiced} and $Ppl_{Unvoiced}$ are calculated as percentages of lost packets classified as voiced and unvoiced packets over the total number of packets, respectively. $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$ are also calculated using a modified version of equation (3) to consider burstiness in voiced and voiced parts of the signal.

$$BurstR_{Voiced} = \frac{\text{Length of burst in Voiced Lost Frames in a sequence}}{\text{Length of burst under “random” loss } (Ppl_{Voiced})} \tag{7}$$

$$BurstR_{Voiced} = \frac{\text{Length of burst in Unvoiced Lost Frames in a sequence}}{\text{Length of burst under “random” loss } (Ppl_{UnVoiced})} \tag{8}$$

In the original E-model, no distinction is made between voiced and unvoiced losses and the values of α_V and α_U are 1, i.e. no difference is made between packet loss during voiced and unvoiced parts of the signal. After conducting several experiments using GA, the optimum values for α_V and α_U in equations (5) and (6) were found to be 2.364 and 0.00238, respectively. Having a value greater than 1 for α_V and a value less than 1 for α_U is consistent with the fact that loss in voiced parts of the signal has more effect on the quality than loss during unvoiced parts.

The new classified losses are integrated with the E-model using equation (4) with the same form of equation used to calculate the original *Ie-eff*. This enforces some restriction on the power of the classification extension for the E-model as the same form of equation is used with or without classification. Also, in the new equation, subjective parameters *Ie* and *Bpl* are still present.

The above two ideas aim to improve the E-model in different ways independently, to avoid the subjectivity of the parameters and to give different weights to different classes of loss. If the above two ideas are combined together to produce a non-intrusive extension for the E-model that is as accurate as PESQ, by considering the effect of loss in voiced and unvoiced parts of the signal separately. At the same time, the extension does not depend on subjective tests to calibrate its parameters, i.e. no use for *Ie* and *Bpl* parameters. This new model will solve the problems of the E-model and will have a wide applicability in estimating speech quality for real-time applications.

In this paper, a technique is proposed to offer both improvements to solve the two problems. The proposed extension is tested on several speech codecs and over

several packet loss probabilities to confirm its effectiveness. Also, the proposed extension is compared with the accurate PESQ measurement and with the original E-model.

The setup for the system is depicted in Fig. 1. In the system setup, PESQ is used as a base criterion for comparison to avoid the need for subjective tests required to retrieve the E-model's parameters. Also, by comparing the E-model's estimation of quality with PESQ's measurement of quality, the accuracy of the E-model is improved by bringing its estimation closer to the accurate PESQ measurement. The modified model satisfies the requirements of quality estimation of voice traffic in IP networks. Such requirements include having an automated, non-intrusive, and accurate solution that does not depend on subjective parameters to calibrate its parameters.

In the system setup, the reference speech signal is first encoded and then packet loss is simulated with different possible probabilities. The received stream is decoded to retrieve the degraded speech signal, and quality is measured by comparing the reference speech signal with the degraded speech signal using PESQ. This measured PESQ value is then mapped into MOS score which, in turn, is used to calculate R-Rating factor and then $Ie\text{-}eff$. The calculated $Ie\text{-}eff$ is considered an accurate measurement, as it is calculated using the accurate PESQ algorithm. Also this calculated value does not depend on subjective parameters Ie and Bpl but rather it is calculated using Ppl and $BurstR$.

At the same time, the degraded signal at the receiver side is analyzed to calculate packet loss statistics for Voiced and Unvoiced parts of the signal. These statistics include Ppl_{Voiced}, $Ppl_{Unvoiced}$, $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$. By

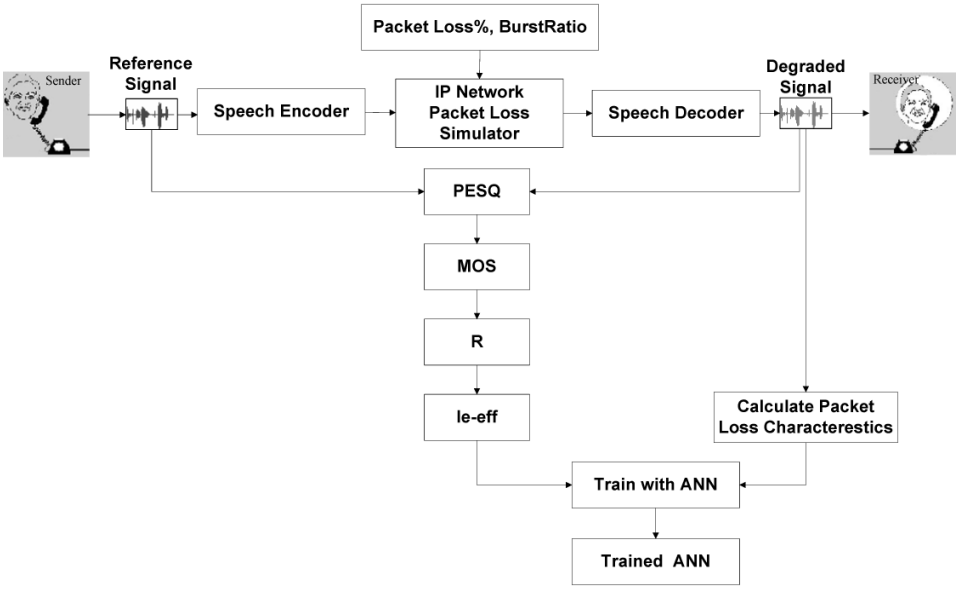

Fig. 1 *System setup for the E-model extension based on PESQ with voice classification.*

feeding these four packet loss statistics as four inputs to an ANN structure and using the calculated *Ie-eff* as target information, an ANN model can be trained to find a relation between packet loss statistics and *Ie-eff*, i.e. to predict *Ie-eff* based on input packet loss statistics. This is a function approximation problem which is a classical use of ANN.

As packet loss statistics Ppl_{Voiced}, $Ppl_{Unvoiced}$, $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$ are calculated and used as input information and *Ie-eff* from PESQ is used as output information, this scheme gives more accurate estimation of speech quality than packet loss dependent Effective Equipment Impairment Factor (*Ie-eff*) in the the original E-model which affects the E-model accuracy. Additionally, as the subjective-dependent parameters, namely *Ie* and *Bpl*, are not used as input parameters, this scheme also does not depend on subjective tests to calibrate its parameters but rather both *Ie* and *Bpl* parameters are absorbed in the form of the ANN weights and biases. The performance of the proposed improvement of the E-model system and its advantage over the original E-model is to be confirmed by statistical analysis. To check whether the proposed system gives similar performance to that provided by the intrusive PESQ but in a non-intrusive way depending on the received voice stream, a two paired-t statistical test is performed to confirm that there is no significant difference between the two methods.

The choice of ANNs over linear or non linear regression models to find a relation between packet loss statistics and *Ie-eff* comes from the fact that ANN performed the best in modeling *Ie-eff* with *Ppl* and *Burst* [1, 2, 3]. Also by choosing linear regression, we assume the underlying relation to be linear, which may not be true. In case of non-linear regression, the form and the degree of the non-linear polynomial need to be determined while the underlying relation could be better modeled by a non-polynomial function.

Based on the above, the combined scheme offers an automatic solution for monitoring live traffic accurately, non-intrusively and without the need for subjective tests to calibrate the parameters. As such, this model has wide applicability in estimating speech quality for real-time applications. Fig. 2 shows how the new scheme can be used as a production system to monitor conversational speech quality non-intrusively.

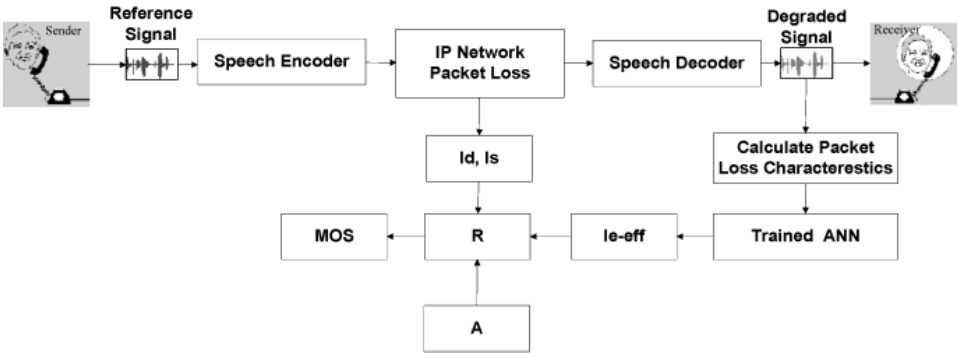

Fig. 2 *Application of the new system in monitoring live systems non-intrusively.*

In Fig. 2 the degraded speech signal is analyzed to extract packet loss statistics non-intrusively, the extracted four statistics (Ppl_{Voiced}, $Ppl_{Unvoiced}$, $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$) are fed into the trained ANN to estimate *Ie-eff*. This *Ie-eff* is then combined with *Id*, *Is* and the Advantage factor (A) are added to calculate R-Rating factor according to equation (1). This R-Rating factor is then mapped into a corresponding MOS, which is an accurate estimation of speech quality that is highly-correlated with PESQ estimation, as confirmed by the conducted experiments and analyzed statistically.

4. Performance of ANN in Estimating *Ie-eff*

Packet loss is simulated using the 2-state Markov model, also known as the Gilbert Model. In the Gilbert model, the system moves between two states "found" and "loss", as shown in Fig. 3. The system suffers from burst loss when it remains in "loss" state.

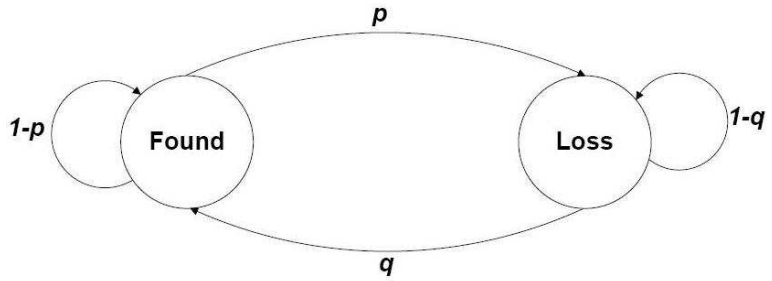

Fig. 3 *2-State Markov Model.*

The 2-state Markov model depends on two parameters p and q, where p is the probability of transition from "found" state to "loss" state and q is the probability of transition from "loss" state to "found" state. p and q are derived from Ppl and $BurstR$ using the following equation [9]:

$$BurstR = \frac{1}{p+q} = \frac{\frac{Ppl}{100}}{p} = \frac{1-\frac{Ppl}{100}}{q}. \tag{9}$$

Several values for Ppl and $Burst$ are attempted. Ppl in the range 0 to 20 and $BurstR$ in the range 1 to 2. These ranges are chosen in order to be able to compare the results with those of the E-model, as the original E-model was defined over these two ranges. For each combination of Ppl and $BurstR$, the experiment is repeated for 30 times to obtain degraded signals with several loss locations, which makes up a total of 1320 runs. During each run, packet loss is simulated using the Gilbert model constructed based on corresponding Ppl and $BurstR$ to retrieve a degraded signal. The degraded signal is compared against the original signal to calculate PESQ score. This PESQ score is then used to calculate MOS score, then R-Rating factor and *Ie-eff* can be computed. These are the empirical values to be used as targets for the ANN structure.

The degraded signal is also used for further calculations. The received packets are classified into either Voiced or Unvoiced using the VAD algorithm that comes as part of G.729 standard [12]. Then, these packets are used to classify the surrounding missing packets and to calculate statistics about the missing packets. This yields 1320 vectors, each vector contains packet loss statistics for Voiced and Unvoiced losses (Ppl_{Voiced}, $Ppl_{Unvoiced}$, $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$) as well as estimation of quality according to PESQ, mapped MOS value, R-Rating factor, and *Ie-eff*, as depicted in Fig. 1. The above experiments are repeated for two speech codecs, G.723.1 and G.729.

A standard 10-fold cross validation is applied, where in each fold 10% of the data is chosen as testing subset and the remaining data is divided as 80% training subset and 10% validation subset. This corresponds to 1056 training, 132 validation and 132 testing vectors. The chosen testing subsets in the 10-folds have no overlapping. The use of validation subset is to improve generalization accuracy and avoid over fitting the trained network into the training data. Training, validation and test subsets were picked as equally spaced points throughout the original data to avoid bias in the training set.

To obtain an ANN network that acts as function approximator between Ppl_{Voiced}, $Ppl_{Unvoiced}$, $BurstR_{Voiced}$ and $BurstR_{Unvoiced}$ input data and *Ie-eff* target, a two-layer neural network with sigmoid transfer function in the first layer and linear transfer function in the output layer is used, and the network is trained using the Levenberg-Marquardt (LM) algorithm. The sigmoid function is able to model non-linear relations between the input and the output while it squashes the output to the range -1 to 1. The linear transfer function gives values outside this range.

Different number of neurons in the hidden layer are attempted, ranging from 1 neuron to 100 neurons. With one neuron, the total number of weights and biases in the network equals 7, which could be not enough for generalization. With 100 neurons the network will have the capability fully to remember the training set which may result in overfitting. For each setting the experiment is repeated for 30 different trials, where different random initial weights are used in each trial. This adds up to 3000 experiments in total (100x30).

During the experiments, each network was allowed to be trained as far as 10000 epochs. In all cases, training stopped before reaching this number due to the error in the validation set exceeding the error in the training set. For speech codec G.723.1, the best network in terms of performance of the test set was found to be a network with 5 neurons in the hidden layer. This network has 31 weights and biases in total. For speech codec G.729, the best network in terms of performance of the test set was found to be a network with 3 neurons in the hidden layer, this network has 19 weights and biases in total. These two networks will be used for subsequent derivations in the next section.

5. Results of Applying ANN in Quality Estimation

Using the best networks obtained in the last section for G.723.1 and G.729, *Ie-eff* from the ANN can be compared with *Ie-eff* obtained experimentally and with *Ie-eff*

obtained from the original E-model over the whole data set. Having 1320 vector in total and 132 test vector in each fold, comparisons are made between the test data with the output of ANN trained by the corresponding training data in that fold. The total number of comparable pairs equal 1320 for the 10-folds. Using the empirical *Ie-eff* and the ANN predicted *Ie-eff*, the corresponding R-Rating factor and MOS values are calculated for more meaningful comparisons. The conducted experiments and the comparison are for two well-known speech codecs, G.723.1 and G.729. The same methodology and comparisons can be performed as well on other speech codecs.

5.1 Results & comparison for speech codec G.729

After performing a standard 10-fold cross validation where in each fold 10% of the data is chosen as testing subset and the remaining data is divided as 80% training subset and 10% validation subset. This corresponds to 1056 training, 132 validation and 132 testing vectors. For each fold, an ANN is trained, using the training set where the input for the training was packet loss statistics to predict *Ie-eff* and the performance is tested using the testing subset.

To check whether there is a difference between the empirical values and the values obtained through ANN, two paired-t statistical test is performed for the produced classifier in each fold to compare between empirical *Ie-eff* values and *Ie-eff* values produced by ANN. The produced t-values are listed in Tab. I where the degree of freedom is 131 as there are 132 testing pair of empirical and ANN produced values in each fold.

Fold	1	2	3	4	5	6	7	8	9	10
t-value	-0.0024	-0.0054	-0.0008	-0.0055	0.0043	0.0044	0.0007	-0.0015	-0.0125	-0.0003

Tab. I *t-values between empirical and ANN predicted* Ie-eff *in each of the 10-folds for speech codec G.729.*

Looking at Tab. I, it can be concluded that there is no significant difference between the empirical and the ANN prediction of *Ie-eff* at the 1% significance level (99% confidence level) for the 10 ANNs as all the t-values are included in the interval $[-2.5758, 2.5758]$ which confirms the effectiveness of the 10 ANNs in different folds in predicting an accurate *Ie-eff* that corresponds to the empirical *Ie-eff* produced by the accurate PESQ method. When different test sets (1320 vector) are grouped together to compare the empirical value with the predicted values produced by the corresponding ANN, the produced t-value was -0.00062, which also indicates that there is no significant difference between the empirical and the ANN values at 1% significance level.

Performing comparison between empirical and ANN values over the whole data sets yields multiple correlation coefficient (R) of value 0.9457 between *Ie-eff* obtained empirically and *Ie-eff* from the ANN model for speech codec G.729, which indicates a strong positive correlation and a good fit. The R^2, the coefficient of determination has the value of 0.8943, which indicates that 89.43% of the time the variation in the independent variable is explained by the model.

Boxplot of differences between *Ie-eff* obtained empirically and *Ie-eff* obtained from ANN for speech codec G.729 over the whole data sets is shown in Fig. 4 where it appears that the values of prediction error are clustered in the lower range with the first quartile below 1.14 *Ie-eff* and the lower two quartiles (median value) below 2.35 *Ie-eff*. More than 75% of the data are below 4 *Ie-eff*. There are few outliers (out of 1320) with high prediction error. The maximum absolute *Ie-eff* difference equals 11.9131 while the average absolute difference equals 2.7880.

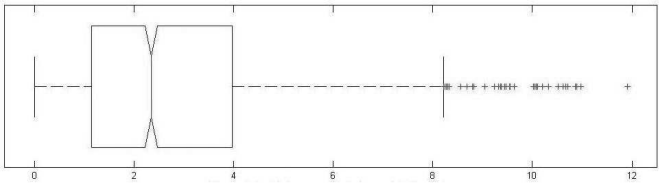

Fig. 4 *Boxplot of the error in* Ie-eff *between empirical values and Artificial Neural Network prediction for speech codec G.729.*

Fig. 5 shows the scatter diagram between *Ie-eff* values obtained empirically and *Ie-eff* from ANN model to visualize the correlation between the corresponding values. Most of the points are located near the perfect fit line due to the very high correlation.

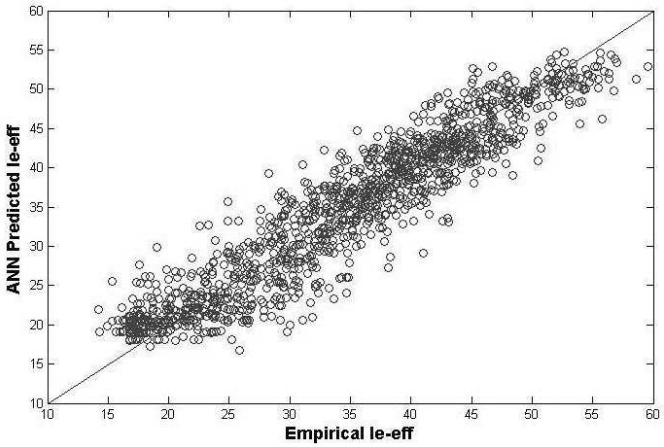

Fig. 5 *G.729 scatter diagram of Ie-eff quality prediction.*

From the *Ie-eff*, *R*-Rating factor can be calculated, which can then be mapped into MOS score. Comparisons in terms of MOS differences are of more interest as it is more plausible to the listener since it represents the output according to the user's perception.

The *t*-values produced by two paired-t statistical test to check whether there is a difference between the empirical MOS values and the MOS values obtained through

ANN are listed in Tab. II. The t-test is performed for the produced classifier in each fold to compare between empirical MOS values and MOS values produced by ANN. The produced t-values are listed in Tab. II, where the degree of freedom is 131 as there are 132 pairs of empirical and ANN produced values.

Fold	1	2	3	4	5	6	7	8	9	10
t-value	0.0022	0.0038	-8.6210E-05	0.6498	-0.0049	-0.0046	-0.0019	0.0007	0.0114	-0.0005

Tab. II *t-values between empirical and ANN predicted MOS in each of the 10-folds for speech codec G.729.*

From Tab. II, it can be concluded that there is no significant difference between the empirical and the ANN prediction of MOS at the 1% significance level (99% confidence level), which confirms the effectiveness of the 10 ANNs in different folds in predicting an accurate MOS score that is close to the MOS produced by the PESQ method. The produced t-value, when different test sets (1320 vector) are grouped together, was 0.000364046, which indicates there is no significant difference between the empirical and the ANN values at 1% significance level.

The multiple correlation coefficient (R) has the value of 0.9466 between MOS obtained empirically and MOS obtained from the ANN model for speech codec G.729, which indicates a strong positive correlation and a good fit. The R^2, the coefficient of determination has the value of 0.8961, which indicates that 89.61% of the time the variation in the independent variable is explained by the model. Boxplot of differences between MOS obtained empirically and MOS obtained from the ANN model for speech codec G.729 over the whole data set is shown in Fig. 6, where it appears that the values of prediction error are clustered in the lower range with the first quartile below 0.0566 MOS and the lower two quartiles (median value) below 0.1112 MOS. More than 75% of the data are below 0.1953 MOS. The maximum absolute MOS difference equals 0.6199 while the average absolute MOS difference equals 0.1389 MOS.

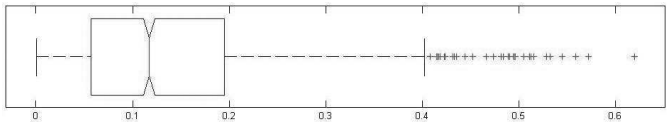

Fig. 6 *Boxplot of the error in MOS between empirical values and Artificial Neural Network prediction for speech codec G.729.*

A scatter diagram between the ANN prediction and the empirical PESQ-derived MOS scores is shown in Fig. 7 to visualize the correlation between the corresponding values. Most of the points are located near the perfect fit line due to the very high correlation.

Tab. III compares the accuracy of the derived ANN with the empirical values obtained through the accurate PESQ measurement, with those values obtained by the original E-model and with the results obtained in [2] for speech codec G.729

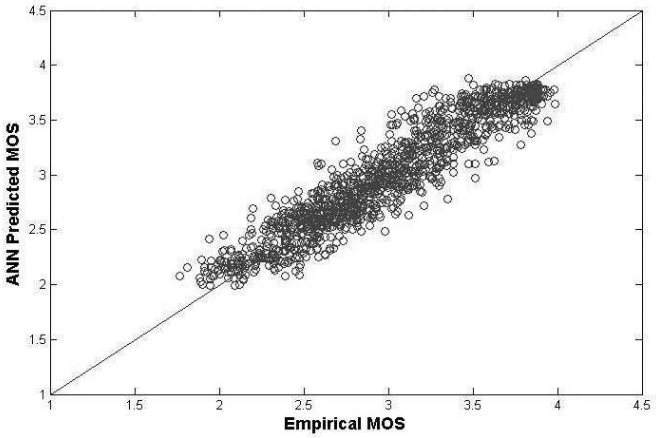

Fig. 7 *G.729 scatter diagram of MOS quality prediction.*

to prove the effectiveness of the proposed technique. It should be noted that the aim of the work proposed in [2] by the authors was to avoid the subjective parameters not to improve the accuracy, while the aim of this work is to avoid the subjectivity and to improve the accuracy. The values in the Table are average values over 30 different iterations performed for each combination of $BurstR$ and Ppl.

The column that comes after MOS value for each of the three methods lists the difference between the MOS in that method and the empirical MOS values for each combination of $BurstR$ and Ppl. At the end of the Table, the average differences are computed for each method. When $BurstR = 1$, the average differences are **0.313**, **0.271**, and **0.061** for the E-model$_{MOS}$, the work in [2], and the improved E-model proposed in this paper, respectively. When $BurstR = 2$, the average differences are **0.357**, **0.379**, and **0.053** for the E-model$_{MOS}$, the work in [2], and the improved E-model proposed in this paper, respectively.

It is clear from the figures that the proposed method outperforms the the original E-model and the work presented in [2] in terms of quality prediction accuracy.

5.2 Results & comparisons for speech codec G.723.1

Similar set of experiments and tests are conducted for speech codec G.723.1 as for speech codec G.729. A two paired-t test is conducted to check whether there is a difference between the empirical values and the values obtained through ANN. The test is repeated for each fold to compare between empirical *Ie-eff* values and *Ie-eff* values produced by ANN in that fold, where in each fold the degree of freedom is 131, as there are 132 pair of empirical and ANN produced values. The t-values are listed in Tab. IV.

From Tab. IV, it can be concluded that there is no significant difference between the empirical and the ANN prediction of *Ie-eff* at the 1% significance level (99% confidence level) for the 10 ANNs. When different test sets (1320 vector) are grouped together to compare the empirical values with the predicted values

	BurstR=1							BurstR=2						
Ppl	$PESQ_{MOS}$	E-model$_{MOS}$	Diff	[2]	Diff	Improved E-model	Diff	$PESQ_{MOS}$	E-model$_{MOS}$	Diff	[2]	Diff	Improved E-model	Diff
0	3.883	4.100	0.217	4.090	0.207	3.756	0.127	3.883	4.100	0.217	4.030	0.147	3.756	0.127
0.5	3.854	4.030	0.177	4.020	0.167	3.754	0.099	3.818	4.020	0.202	3.940	0.122	3.702	0.116
1	3.815	3.950	0.135	3.950	0.135	3.751	0.065	3.733	3.940	0.208	3.840	0.108	3.700	0.033
2	3.715	3.790	0.075	3.810	0.095	3.720	0.005	3.612	3.770	0.158	3.650	0.038	3.553	0.059
3	3.648	3.630	0.018	3.670	0.022	3.694	0.046	3.561	3.590	0.029	3.450	0.111	3.411	0.150
4	3.573	3.480	0.093	3.540	0.033	3.657	0.084	3.362	3.410	0.048	3.260	0.102	3.320	0.042
5	3.522	3.340	0.182	3.410	0.112	3.605	0.083	3.229	3.240	0.011	3.080	0.149	3.184	0.045
6	3.375	3.210	0.165	3.290	0.085	3.510	0.135	3.166	3.060	0.106	2.910	0.256	3.076	0.089
7	3.304	3.080	0.224	3.180	0.124	3.455	0.151	2.981	2.890	0.091	2.760	0.221	3.018	0.037
8	3.272	2.960	0.312	3.070	0.202	3.362	0.090	2.923	2.730	0.193	2.620	0.303	2.909	0.014
9	3.197	2.850	0.347	2.970	0.227	3.277	0.080	2.846	2.580	0.266	2.490	0.356	2.816	0.029
10	3.134	2.750	0.384	2.880	0.254	3.155	0.021	2.808	2.430	0.378	2.370	0.438	2.729	0.079
11	3.006	2.650	0.356	2.780	0.226	3.032	0.026	2.685	2.290	0.395	2.260	0.425	2.642	0.043
12	2.972	2.560	0.412	2.690	0.282	3.025	0.053	2.581	2.160	0.421	2.150	0.431	2.625	0.044
13	2.929	2.470	0.459	2.590	0.339	2.942	0.014	2.536	2.030	0.506	2.050	0.486	2.545	0.009
14	2.900	2.400	0.500	2.490	0.410	2.901	0.001	2.522	1.920	0.602	1.940	0.582	2.434	0.089
15	2.791	2.320	0.471	2.380	0.411	2.832	0.042	2.330	1.810	0.520	1.830	0.500	2.353	0.023
16	2.753	2.250	0.503	2.270	0.483	2.726	0.028	2.409	1.710	0.699	1.730	0.679	2.366	0.043
17	2.623	2.190	0.433	2.160	0.463	2.689	0.066	2.241	1.620	0.621	1.620	0.621	2.270	0.030
18	2.586	2.130	0.456	2.070	0.516	2.634	0.048	2.252	1.540	0.712	1.520	0.732	2.248	0.004
19	2.532	2.070	0.462	1.980	0.552	2.587	0.055	2.197	1.460	0.737	1.430	0.767	2.180	0.017
20	2.519	2.020	0.499	1.910	0.609	2.540	0.021	2.115	1.390	0.725	1.360	0.755	2.150	0.035
Difference Average			**0.313**		**0.271**		**0.061**		**0.357**		**0.379**			**0.053**

Tab. III *Comparison of G.729 speech codec empirical results obtained from the accurate PESQ measurements with the E-model results, results in [2] and the work proposed in this paper. The difference between the empirical values and each of the 3 methods is listed after each method.*

Fold	1	2	3	4	5	6	7	8	9	10
t-value	-0.0057	-0.0037	-0.0055	-0.0053	-0.0044	-0.0023	0.0016	-0.0059	-0.0044	0.0041

Tab. IV *t-values between empirical and ANN predicted* Ie-eff *in each of the 10-folds for speech codec G.723.1.*

produced by the corresponding ANN, the produced t-value was -0.0010, which also indicates that there is no significant difference between the empirical and the ANN values at 1% significance level.

The multiple correlation coefficient (R) between empirical *Ie-eff* and *Ie-eff* from the ANN model for speech codec G.723.1 has the value of 0.9520, which indicates a strong positive correlation and a good fit. The R^2, the coefficient of determination has the value of 0.9064, which shows that 90.64% of the variation in the independent variable is explained by the ANN model.

Boxplot of differences between empirical *Ie-eff* and *Ie-eff* obtained from ANN for speech codec G.723.1 over the whole data set is shown in Fig. 8, where it appears that the values of prediction error are clustered in the lower range with the first quartile below 1.098 *Ie-eff* and the lower two quartiles (median value) below 2.219 *Ie-eff*. More than 75% of the data are below 3.8 *Ie-eff*. There are few outliers (out of 1320) with high prediction error. The maximum absolute *Ie-eff* difference equals 14.2291, while the average absolute difference equals 2.7226.

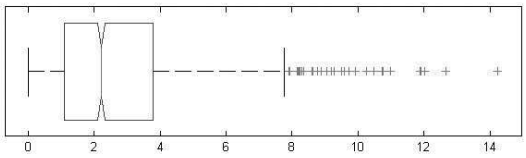

Fig. 8 *Boxplot of the error in* Ie-eff *between empirical values and Artificial Neural Network prediction for speech codec G.723.1.*

The scatter diagram between *Ie-eff* values obtained empirically and *Ie-eff* from ANN is shown in Fig. 9 to visualize the correlation between the corresponding values. Most of the points are located near the perfect fit line due to the very high correlation.

Empirical and ANN predicted *Ie-eff* values are used to calculate the corresponding R-Rating factor, which are then mapped into MOS score. Comparisons in terms of MOS differences are more readable to the listeners, as it represents the output according to the user's perception.

The t-values produced by two paired-t statistical test conducted to check if there is a difference between the empirical MOS values and the MOS values obtained through ANN are listed in Tab. V. The t test is conducted for each fold with the degree of freedom being 131, as there are 132 pair of empirical and ANN produced values.

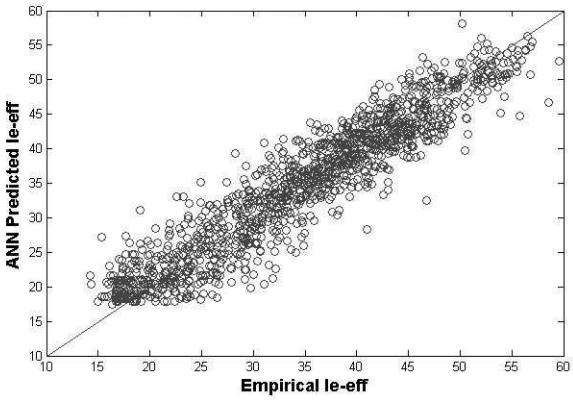

Fig. 9 *G.723.1 scatter diagram of Ie-eff quality prediction.*

Fold	1	2	3	4	5	6	7	8	9	10
t-value	0.0053	0.0025	0.0045	0.0045	0.0032	0.0015	-0.0029	0.0048	0.0039	-0.0050

Tab. V *t-values between empirical and ANN predicted MOS in each of the 10-folds.*

From Tab. V, it can be concluded that there is no significant difference between the empirical and the ANN prediction of *Ie-eff* at the 1% significance level (99% confidence level), which confirms the effectiveness of the 10 ANNs in different folds in predicting an accurate MOS score that is close to the MOS produced by the PESQ method. The produced t-value, when different test sets (1320 vector) are grouped together, is 0.00073, which indicates there is no significant difference between the empirical and the ANN values at 1% significance level.

The multiple correlation coefficient (R) has the value of 0.9519 between MOS obtained empirically and MOS obtained from the ANN model for speech codec G.723.1, which indicates a strong positive correlation and a good fit. The R^2, the coefficient of determination has the value of 0.9060, which indicates that 90.60% of the variation in the independent variable is explained by the ANN model. Boxplot of differences between MOS values obtained empirically and MOS values obtained from the ANN model for speech codec G.723.1 over the whole data set is shown in Fig. 10, where it appears that the values of prediction error are clustered in the lower range with the first quartile below 0.05437 MOS and the lower two quartiles (median value) below 0.1114 MOS. More than 75% of the data are below 0.1906 MOS. The maximum absolute MOS difference equals 0.7484, while the average absolute MOS difference equals 0.1357 MOS.

The scatter diagram between the ANN prediction and the empirical PESQ-derived MOS scores is shown in Fig. 11 to visualize the correlation between the corresponding values. Most of the points are located near the perfect fit line due to the very high correlation.

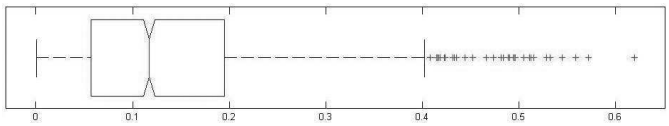

Fig. 10 *Boxplot of the error in MOS between empirical values and Artificial Neural Network prediction for speech codec G.723.1.*

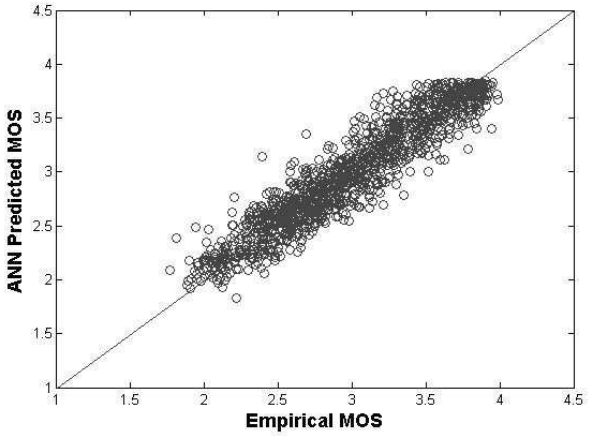

Fig. 11 *G.723.1 scatter diagram of MOS quality prediction.*

Tab. VI compares the accuracy of the derived ANN with the empirical values obtained through the PESQ measurement. The values obtained through the original E-model are also compared with the empirical PESQ measurements for speech codec G.723.1 to prove the effectiveness of the proposed technique. The values in the Table are average values over 30 different iterations performed for each combination of $BurstR$ and Ppl. It should be noted that these results are not compared with the results obtained in [2], as the results of [2] were for speech codec G.729 only, and no results were obtained for G.723.1, as the experiments conducted in this paper are more comprehensive.

The differences between the values obtained using the original E-model and the values obtained using the improved E-model are listed for each combination of $BurstR$ and Ppl. The table also lists the average differences. When $BurstR = 1$, the average differences are **0.524** and **0.051** for the E-model$_{MOS}$ and the improved E-model proposed in this paper, respectively. When $BurstR = 2$, the average differences are **0.571** and **0.053** for the E-model$_{MOS}$ and the improved E-model proposed in this paper, respectively. It is clear from the differences that the improved E-model presented in this paper outperforms the original E-model, as it has smaller differences from the empirically obtained values from the accurate PESQ measurement.

	BurstR=1					BurstR=2				
Ppl	$PESQ_{MOS}$	E-model$_{MOS}$	Diff	Improved E-model	Diff	$PESQ_{MOS}$	E-model$_{MOS}$	Diff	Improved E-model	Diff
0	3.883	3.950	0.067	3.779	0.104	3.883	3.950	0.067	3.779	0.104
0.5	3.854	3.860	0.006	3.778	0.075	3.818	3.850	0.032	3.701	0.117
1	3.815	3.760	0.055	3.771	0.045	3.733	3.750	0.018	3.666	0.067
2	3.715	3.570	0.145	3.729	0.014	3.612	3.540	0.072	3.539	0.073
3	3.648	3.390	0.258	3.699	0.052	3.561	3.330	0.231	3.380	0.181
4	3.573	3.220	0.353	3.643	0.070	3.362	3.130	0.232	3.318	0.044
5	3.522	3.060	0.462	3.572	0.050	3.229	2.930	0.299	3.194	0.035
6	3.375	2.920	0.455	3.488	0.113	3.166	2.740	0.426	3.078	0.088
7	3.304	2.780	0.524	3.416	0.113	2.981	2.560	0.421	3.004	0.023
8	3.272	2.660	0.612	3.335	0.063	2.923	2.390	0.533	2.889	0.035
9	3.197	2.550	0.647	3.237	0.040	2.846	2.230	0.616	2.824	0.021
10	3.134	2.450	0.684	3.139	0.005	2.808	2.080	0.728	2.733	0.075
11	3.006	2.350	0.656	3.025	0.019	2.685	1.940	0.745	2.633	0.053
12	2.972	2.270	0.702	3.026	0.054	2.581	1.820	0.761	2.627	0.045
13	2.929	2.190	0.739	2.934	0.005	2.536	1.700	0.836	2.568	0.032
14	2.900	2.110	0.790	2.896	0.004	2.522	1.600	0.922	2.484	0.038
15	2.791	2.050	0.741	2.841	0.051	2.330	1.500	0.830	2.366	0.035
16	2.753	1.980	0.773	2.731	0.022	2.409	1.420	0.989	2.381	0.028
17	2.623	1.930	0.693	2.701	0.078	2.241	1.340	0.901	2.275	0.034
18	2.586	1.870	0.716	2.636	0.050	2.252	1.280	0.972	2.246	0.005
19	2.532	1.820	0.712	2.596	0.064	2.197	1.220	0.977	2.170	0.027
20	2.519	1.780	0.739	2.560	0.041	2.115	1.170	0.945	2.106	0.009
Difference Average			**0.524**		**0.051**			**0.571**		**0.053**

Tab. VI *Comparison of G.723.1 speech codec empirical results obtained from the accurate PESQ results with E-model results and results of the work proposed in this paper. The difference between the empirical values and the other 2 methods is listed after each method.*

6. Conclusions

Measurement of voice quality is an important aspect in VoIP networks. Several methods have been proposed for the task of quality measurement. The proposed methods can be categorized into three main classes: subjective assessment techniques, intrusive objective assessment techniques and non-intrusive objective assessment techniques. The E-model standardized by the ITU-T in Recommendation G.107 is a well-known non-intrusive objective method for estimating speech quality [9].

The authors proposed in this paper improvements to the E-model to avoid two drawbacks that hinder its applicability. The proposed model improves the E-model as it avoids the hard-to-conduct, time-consuming and expensive subjective tests required to estimate the E-model's parameters by using Perceptual Evaluation of Speech Quality PESQ (PESQ) to find a model (Artificial Neural Network, ANN in this case) that does not use the subjective test related parameters Ie and Bpl.

The new model also offers more accurate estimation to speech quality by considering the class of lost packets, as statistics have shown that loss during voiced parts of the signal has a different perceptual effect than loss during unvoiced part of the signal, depending on the more accurate intrusive-based PESQ method standardized by the ITU-T in Recommendation P.862 [15] as a baseline criterion. Packet loss in the E-model is treated collectively without differentiating between voiced and unvoiced loss. An algorithm is provided to calculate different packet loss statistics. By calculating packet loss statistics and finding a relation with the PESQ measurement, the E-model estimation can be improved and the subjectivity of parameters can be avoided. The relation between packet loss statistics and the quality is derived, using an ANN structure that works as function approximator.

Several experiments are conducted on two speech codecs G.723.1 and G.729 with 10-fold cross validation. To check the accuracy of the proposed ANN structure in predicting quality, several statistical tests are performed to check if there is a difference between the proposed method and the empirical values. Comparisons are also made with the original E-model. Experimental results indicate that the proposed method improves quality prediction over the original E-model. The same principles presented in this paper can be tested with other speech codecs by running the required experiments and training an appropriate ANN model to gain the same advantages.

The proposed model is an attractive and accurate solution for measuring speech quality objectively and non-intrusively in live networks. Therefore, it has wide applicability in estimating speech quality for voice applications over IP networks, which increases its significance as it provides better features than the E-model for VoIP traffic by being more accurate in estimating the quality and being able to avoid the subjectivity in estimating the E-model's parameters.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments which indicate strong understanding of the proposed work. Their comments improved the quality and the evaluation methods presented in this paper.

References

[1] AL-Akhras M.: A Genetic Algorithm Approach for Voice Quality Prediction. In: Proceedings of 5th IEEE International Multi-Conference on Systems, Signals & Devices, Proceedings of International Conference on Communication & Signal Processing (CSP 2008), Amman, Jordan, July 20-24, 2008.

[2] AL-Akhras M., Zedan H., John R., ALMomani I.: Non-Intrusive Speech Quality Prediction in VoIP Networks Using a Neural Network Approach. In: Neurocomputing (ISSN 0925-2312), 72, 2009, pp. 2595–2608.

[3] Almomani I., Al-AKhras M.: Statistical Speech Quality Prediction in VoIP Networks. In: Proceedings of the 2008 International Conference on Communications in Computing (CIC'8), Las Vegas, July 14-17, 2008.

[4] Allnatt J.: Subjective Rating and Apparent Magnitude. In: International Journal Man-Machine Studies, vol. **7**, 1975, pp. 801–816.

[5] Brunner J., Koutnik J.: SiMoNNe – Simulator of Modular Neural Networks. In: Neural Network World, vol. **12**, no. 3, 2002, pp. 267–278.

[6] Collins D.: Carrier Grade Voice over IP. 2nd edition, McGraw-Hill Companies, 2003.

[7] Frolov A. A., Husek D., Polyakov P., Rezankova H.: New Neural Network Based Approach Helps to Discover Hidden Russian Parliament Voting Patterns. In: Proceedings of International Joint Conference on Neural Networks, 2006 (IJCNN '06).

[8] Frolov A. A., Husek D., Muraviev I. P., Polyakov P. Yu.: Boolean Factor Analysis by Attractor Neural Network. In: IEEE Transactions on Neural Networks, vol. **18**, no. 3, 2007, pp. 698–707.

[9] ITU-T. Recommendation G.107 – The E-model, a computational model for use in transmission planning. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), April 2009.

[10] ITU-T. Recommendation G.113 Appendix I – Provisional planning values for the equipment impairment factor Ie and packet-loss robustness factor Bpl. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), May 2002.

[11] ITU-T. Recommendation G.723.1 – Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), March 1996.

[12] ITU-T. Recommendation G.729 – Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), March 1996.

[13] ITU-T. Recommendation P.800 – Methods for subjective determination of transmission quality. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), August 1996.

[14] ITU-T. Recommendation P.800.1 – Mean Opinion Score (MOS) terminology. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), March 2003.

[15] ITU-T. Recommendation P.862 – Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), February 2001.

[16] ITU-T. Recommendation P.862.1 – Mapping function for transforming P.862 raw result scores to MOS-LQO. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), March 2005.

[17] ITU-T. Recommendation 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, Amendment 2: Revised Annex A: Reference implementations and conformance testing for Recommendations P.862, P.862.1 and P.862.2. International Telecommunication Union – Telecommunication Standardisation Sector (ITU-T), November 2003.

[18] James J. H., Chen B., Garrison L.: Implementing VoIP: A Voice Transmission Performance Progress Report. In: IEEE Communications Magazine, vol. **42**, July 2004, pp. 36–41.

[19] Khasnabish B.: Implementing Voice over IP. 2nd edition, Wiley-Interscience, 2003.

[20] Markopoulou A. P., Tobagi F. A., Karam M. J.: Assessment of VoIP Quality over Internet Backbones. In: Proceedings of IEEE Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002), vol. **1**, 23-27 June 2002, pp. 150–159.

[21] Mashford J. S.: A neural network image classification system for automatic inspection. In: proceedings of IEEE International Conference on Neural Networks, vol. **2**, Nov/Dec, 1995.

[22] Pajares G.: A Hopfield Neural Network for Image Change Detection. In: IEEE Transactions on Neural Networks, vol. **17**, no. 5, 2006, pp. 1250–1264.

[23] Prasad R. V., Sangwan A., Jamadagni H. S., Chiranth M. C., Sah R., Gaurav V.: Comparison of Voice Activity Detection Algorithms for VoIP. In: Proceedings of Seventh International Symposium on Computers and Communications, 2002 (ISCC 2002), 1-4 July 2002, pp. 530–535.

[24] Rix A., Beerends J., Hollier M., Hekstra A.: Perceptual Evaluation Of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In: proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 01), vol. **2**, 7-11 May 2001, pp. 749–752.

[25] Sleit A., Al-Mbaideen W., Alzabin N., Dawood H., Alqarute K.: Efficient query processing over mirror servers using genetic algorithms. In: Neural Network World, vol. **17**, no. 4, 2007, pp. 311–320.

[26] Sun L., Wade G., Lines B., Ifeachor E. C.: Impact of Packet Loss Location on Perceived Speech Quality. In: Proceedings of 2nd IP-Telephony Workshop (IPTEL 01), Columbia University, 6-10 April, 2001, pp. 114–122.

[27] Sun L.: Speech Quality Prediction for Voice over Internet Protocol Networks. PhD Thesis, School of Computing, Communications and Electronics, University of Plymouth, U.K., Jan. 2004.

[28] Takahashi A., Yoshino H., Kitawaki N.: Perceptual QoS Assessment Technologies for VoIP. In: IEEE Communications Magazine, vol. **42**, no. 7, 2004, pp. 28–34.

[29] Takahashi A.: Opinion Model for Estimating Conversational Quality of VoIP. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 04), vol. **3**, 17-21 May 2004, pp. iii–1072–1075.

[30] Zurek E. E., Leffew J., Moreno W. A.: Objective Evaluation of Voice Clarity Measurements for VoIP Compression Algorithms. In: Proceedings of the Fourth IEEE International Caracas Conference on Devices, Circuits and Systems, 17-19 April 2002, pp. T033–1–T033–6.