A Methodology for Deriving VoIP Equipment Impairment Factors for a mixed NB/WB Context

Adil Raja*, R. Muhammad Atif Azad, Colin Flanagan and Conor Ryan

***Abstract*—This paper proposes a novel approach to quantifying the quality degradation of Voice over IP (VoIP) telephony in the presence of codec and network-related impairments. This approach differs from the baisc ITU-T E-Model for VoIP quality estimation [1] in that it addresses mixed narrowband/wideband scenarios. It makes novel use of instrumental models and symbolic regression via Genetic Programming (GP) to enable the evolution of degradation models from a modest set of initial parameters. Here, a two-step approach has been used. First, values of impairment factors are derived using WB-PESQ as a reference model. Secondly, a GP based symbolic regression approach has been utilized to automatically evolve the functional form of equipment impairment factors from a set of variables. Very few a priori assumptions are made about the model structure. The effectiveness of the approach is demonstrated by a number of generated models which compare favorably with WB-PESQ and outperform the traditional E-Model in terms of prediction accuracy when compared using WB-PESQ. A significant advantage of the approach is that new models are easily generated to account for continuing evolution of the VoIP standards.**

***Index Terms*—E-Model, genetic programming, symbolic regression, PESQ-WB, $I_{e,WB,eff}$.**

I. Introduction

VoIP is curently evolving rapidly towards wideband based transmission. Wideband (WB) offers more natural sounding speech than narrowband (NB), and IP networks allow the transition to occur essentially by a simple change of codecs. It is clear, however, that there will be a transitional period, with wideband and narrowband VoIP coexisting, leading to a requirement for NB/WB interoperation. An important question that arises as a consequence is how is the quality of such a mixed NB/WB system to be estimated?

VoIP quality is affected by various factors such as packet loss, end-to-end delay, jitter and codec bit-rate etc. A number of approaches and models exist that estimate speech quality as a function of such impairments. Of particular interest among these is ITU-T Recommendation G.107 [1], commonly known as the E-Model, which is an instrumental model that was initially designed for transmission planning purposes. It is based on an impairment factor principle that assumes that the degradations induced by various sources have a cumulative effect on speech quality and that they may accordingly be transformed to a *transmission rating scale (R scale)*. The E-Model was originally intended for NB speech quality estimation. Recently, in [2], Möller et al. proposed an extension of the R scale to incorporate WB codecs into E-Model, while leaving the original R scale for the NB case intact. Their main emphasis has been on deriving *equipment impairment factors ($I_{e,WB}$)*, in a mixed NB/WB context, that represent the degradation in the *listening quality* of speech in the wake of pure codec related distortions. Their derivation is based on subjective *listening only* tests [3] for a mixture of various NB and WB codecs defined by ITU-T.

In the past several authors have taken different approaches towards deriving *effective equipment impairment factors* ($I_{e,eff}$) for NB codecs. In this paper we take a novel perspective towards deriving effective equipment impairment factors for the *mixed NB/WB* case i.e., $I_{e,WB,eff}$. Here the novelty is twofold. First, we propose to use instrumental models as a means to derive reference $I_{e,WB,eff}$, as opposed to subjective tests. Secondly, the mapping between various quality affecting parameters and reference $I_{e,WB,eff}$ is achieved by employing Genetic Programming (GP) based symbolic regression [4]. This approach is based on our past work reported in [5] and [6] where we used GP to derive parsimonious speech quality estimation models. GP employs artificial evolution to automatically induce mathematical models for an otherwise unexplained data set. Since it takes a bottom up approach towards problem solving with minimum assumptions about the structure of the solution, the results can be innovative and non-intuitive.

In this research we have employed a number of state-of-the-art VoIP telephony codecs proposed by ITU-T. We have used ITU-T P.862.2 (i.e.WB-PESQ), as reported in [7], as a reference system. We follow the methodology described in [8] for deriving $I_{e,eff}$ and propose ours as an addendum to it for deriving $I_{e,WB,eff}$.

The rest of the paper is organised as follows. In section II we describe the E-model framework. There we highlight past attempts by various researchers in deriving $I_{e,eff}$ and $I_{e,WB,eff}$ and present our approach too. In section III we discuss the factors that affect $I_{e,WB,eff}$. Section IV gives a brief introduction to GP and describes the advantages of symbolic regression compared to some other machine learning and numerical methods. Section V elucidates our methodology in detail along with our VoIP simulation system describing various NB and WB codecs used in this research and the data processing procedures. Details of GP experiments, various results and models are discussed in section VI. Finally, section VII concludes the paper.

Adil Raja and Colin Flanagan are with the Department of Electronic and Computer Engineering (Phone: +353 -(0)61- 202608 Fax: +353 -(0)61- 202572) – R. Muhammad Atif Azad and Conor Ryan are with the Department of Computer Science and Information Systems, (Phone: +353 -(0)61- 202763 & 202755 Fax: +353 -(0)61- 202734) at the University of Limerick, Ireland. e-mail: $adil.raja|atif.azad|colin.flanagan|conor.ryan@ul.ie$

EDICS: 3-QUAL – Perceptual Quality and Human Factors

II. The E-Model

The E-Model, as defined by ITU-T G.107 [1], is a computational model used for assessing the combined effect of various parameters on speech quality in a conversational sense. Initially it was designed for NB handset telephony, however, its adaptation to the WB case is currently in progress. The primary output of the model is the *Rating Factor*, R. The derivation of R is based on an impairment factor principle that assumes that factors affecting speech quality are additive in nature. Thus, R is computed according to equation (1):

$$R = R_0 - I_s - I_d - I_{e,eff} + A \tag{1}$$

where R ranges from 0 (poor quality) to 100 (optimum quality) for the NB case. R_0 is the basic signal to noise ratio which, for the NB case, defaults to 93.2. I_s represents all the impairments which occur simultaneously with the voice including, for instance, overall loudness rating and non-optimum sidetone. I_d marks the effect of delay related impairments such as echo and too long end-to-end delay that may affect the call quality in a conversational sense. $I_{e,eff}$ depicts the impairments due to low bit-rate codecs in the presence of packet losses. Finally, A is the advantage factor that compensates for the above impairment factors when there are other advantages of access to the user depending on the nature of the underlying network. Thus, for instance, A may be assigned a value of 0 for a wired network and 20 for a multi-hop satellite connection. In the case where values of one or more of these factors may not be determined, default values are used from [1].

R can be converted to *Mean Opinion Score (MOS)* and vice versa using corresponding transformations given in [1]. Since we have leveraged from these transformations in this research we shall refer to them by an abstract notation given by transformation (2).

$$R \Longleftrightarrow MOS \tag{2}$$

where MOS varies on a scale ranging between 1 (bad) to 4.5 (excellent), and it is a measure of human assessment of speech quality. The relationship between MOS and R is shown in Fig. 1 with the solid curve.

The above formulations hold for the case of NB codecs. In [2] Möller et al. proposed a transformation of the R scale from the NB case (R_{NB}) to the mixed NB/WB case ($R_{NB/WB}$) based on subjective tests performed in [9]. The test results suggest that for the scenario where only NB coded samples were present, MOS scores were higher than those for the same samples evaluated in presence of additional, objectively better, WB coded stimuli. Moreover, since the MOS to R conversion represented by transformation (2) was applied, the R_{NB}, for the NB context, turned out to be higher than $R_{NB/WB}$ for the mixed NB/WB context. This would have repercussions for the validity of the original R scale in a mixed NB/WB context as it would affect the NB usage of the scale. Thus, an extension of the R scale for the NB/WB case was proposed that leaves the original R scale for the NB context unaltered. This extension is given by equation (3)[1].

$$R_{new} = a.\left(e^{R_{NB/WB}/b} - 1\right) \tag{3}$$

where a and b were found to be equal to 169.38 and 176.32 respectively, and $R_{NB/WB}$ can be calculated via (2). This extension is now an integral part of the E-Model (see Appendix II of [1]), where the new default value for R_0 for the NB/WB case is 129. Following this, $I_{e,WB}$ (i.e. impairments solely due to various low bitrate NB/WB codecs) can be calculated according to equation (4) as a difference between R-value of the *direct* channel and R-value corresponding to the codec under consideration.

$$I_{e,WB} = 129 - R_{codec} \tag{4}$$

where R_{codec} may be calculated from (3) and 129 corresponds to the value of R for the direct channel for the mixed NB/WB context. The direct channel in this context is represented by a 16-bit linear PCM with f_s=16 kHz (this also assumes that impairments due to other factors such as echo or delay are not present).

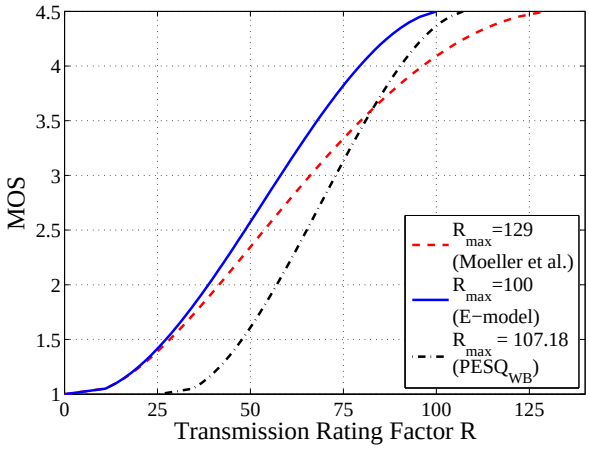


Fig. 1. Transformation rules between R and MOS. Solid line: NB case of the E-Model, dashed line, NB/WB case (Möller et al.) and dashed-dotted line forWB-PESQ

A. On Extending the R scale for WB-PESQ

Our work employs WB-PESQ as a reference for deriving $I_{e,WB}$ and $I_{e,WB,eff}$, as opposed to subjective tests. A WB version of R scale does not exist in the literature for WB-PESQ. There can be two approaches in principle to convert MOS-LQO *(MOS-Listening Quality Objective)* [10] obtained by WB-PESQ to the R scale. Both of these are discussed in this section.

1) One approach may be to extend the R scale using MOS-LQO obtained by WB-PESQ using the methodology proposed in [2] by Möller et al; as has been previously discussed. Based on this an experiment was performed to

[1]A linear version of this extension also exists that has been skipped for brevity

see whether a meaningful extension of the R scale could be made for WB-PESQ. Two test cases were prepared, each comprising 1328 pairs of reference and *coded* file pairs of experiments 1 and 3 of the ITU-T P-series supplement 23 [11]. The coded files contain various NB scenarios with distortion conditions such as low bit-rate coding, signal correlated noise, codec tandeming, bit errors and frame erasures. Conditions representing direct *(or clean)* NB channel are also present. All the files are originally coded in 16-bit linear PCM format, f_s=16 kHz. The file pairs in the first test case were evaluated with WB-PESQ. This constitutes the WB (or NB/WB) context with WB coded references and NB coded and upsampled test files. File pairs in the second test case were evaluated with NB-PESQ. To this end, all the reference and coded files were downsampled and low-pass filtered using [12] prior to evaluation. This corresponds to an NB test.

The resulting scatter plot for R_{NB} and $R_{NB/WB}$ is displayed in Fig. 2. The data was fitted using least squares regression where a linear relationship of the form of equation (5) was used.

$$R_{NB} = a.R_{NB/WB} + b \tag{5}$$

where, a=0.82, b=25.46 and $RMSE$=4.12.

According to this R_{max} was found to be 107.18. This suggests a rather small extension of the R-scale; only a 7% gain in quality due to WB coded speech. The new curve is drawn in Fig. 1 with a dashed-dotted line.

It is worth mentioning here that WB-PESQ has a number of limitations. First of all, the restricted set of training and testing databases limits the reliability of WB-PESQ in comparison with NB-PESQ. Language and codec dependence is another limitation of the algorithm [13, pp-105] [14]. It was also observed that WB-PESQ systematically underestimates speech quality in comparison with subjective tests. This was observed from a comparison made between MOS-LQO obtained by WB-PESQ and MOS-LQS *(MOS-Listening Quality Subjective)*. The results of the comparison are shown in Fig. 3. Here the MOS-LQS were obtained by performing $I_{e,WB}$ to MOS conversion for codecs under consideration using equations (4), (6)[2] and (2). Values of $I_{e,WB}$ were taken from [15].

$$R_{orig} = ln\left(\frac{R_{new} + 169.38}{169.38}\right) \times 176.32 \tag{6}$$

2) The second approach is to convert the MOS-LQO to the R scale using equations (2) and (3), in the order given. This is analogous to the methodology given in ITU-T P.834 [8]. As there are clear problems associated with reconciling the R scale in the case of subjective and objective tests, as seen during the analysis of first approach, we have chosen the second approach. This is used to derive reference values of $I_{e,WB,eff}$ in this research. We argue that our methodology would not be affected due to changes in the mathematical form of any R scale extension; the experiments that follow can be conveniently repeated for a new *(target)* R scale.

[2]Equation (6) is the inverse of equation (3)

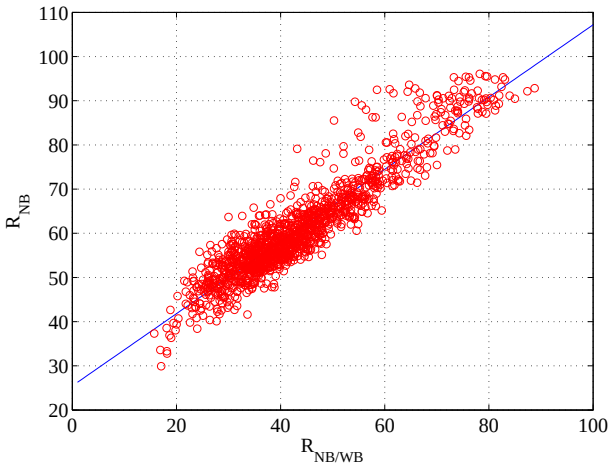


Fig. 2. Comparison between R-values obtained from a NB and a mixed NB/WB context using PESQ.

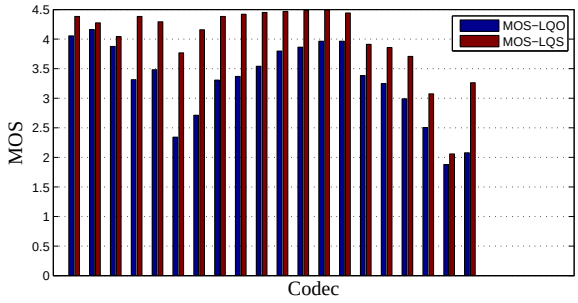


Fig. 3. Comparison between MOS-LQO and MOS-LQS for various NB and WB codecs

III. $I_{e,WB,eff}$ and Associated Quality Elements

According to the E-Model [1] $I_{e,eff}$ for a given NB codec may be computed from

$$I_{e,eff} = I_e + (95 - I_e) \times \frac{P_{pl}}{\frac{P_{pl}}{BurstR} + Bpl} \tag{7}$$

where, I_e is the impairment factor for the codec under consideration in the case of no packet loss. P_{pl} is the packet loss rate (%). BurstR is the Burst Ratio; discussed below. Bpl is the packet loss robustness factor for the codec under consideration. It describes the the robustness of the codec, including the employed packet loss concealment mechanism, against packet loss. A similar formulation for $I_{e,WB,eff}$ is given in [2] for *random* packet loss.

Given this, $I_{e,WB,eff}$, or equivalently $I_{e,eff}$, depends on two quality elements, namely *packet loss* and *codec*. In the text that follows various aspects of these two elements are discussed in detail.

A. Packet Loss

Packet loss may either be random, where loss patterns follow a Bernoulli-like distribution, or bursty in nature. In bursty loss, a lost packet tends to exhibit a temporal dependency on its immediately preceding (lost or arrived) packet, or

past n packets [13][16] [17][18]. E-Model defines a $BurstR$ parameter (*Burst Ratio*) where burstiness is modeled using a two-state Markov model, with a loss and a no-loss state, and with two transition probabilities associated with each state.

Another factor affecting quality impairment, and closely associated with packet loss, is the packetization interval (PI) (ms), i.e., the payload size of an IP packet. In order to utilize the transmission bandwidth effectively, it is desirable to increase the PI. However, larger values of PI result in larger transmission delay and possibly lower speech quality in the event of a packet loss. Current VoIP applications use values of PI ranging between 10–60 ms as a compromise [13].

The problems associated with packet loss may be circumvented to a certain extent with various *packet loss recovery* methods such as Forward Error Correction, Low-Bitrate Redundancy and Packet Loss Concealment (PLC) [19].

B. Codec

$I_{e,WB,eff}$ is a codec specific quantity and thus dependent on it. A speech codec may either belong to the class of *waveform* coders, parametric coders or hybrid coders i.e. a combination of the first two. Waveform coders perform quantization of the speech signal and parametric coders employ a suitable speech production model for reducing bandwidth requirement for speech transmission [20]. For a given class of coders the speech quality may further depend on factors such as codec's bitrate, frame size and coding algorithm. The codec's transmission bandwidth (i.e. NB or WB) also affects the quality perceived by the user. Thus WB codecs deliver better quality than their NB counterparts mainly because of the increased naturalness of speech due to the presence of higher order spectral components [2][13].

In the past various authors have tried to model speech quality as a function of coding bitrate (in addition to loss metrics) e.g. [21][22][23] and also by the authors in [5][6]. It may be argued that although coding bitrate may be used as a quality defining parameter for general predictions, it may not be able to give accurate predictions due to two main reasons.

First, in the absence of any other impairments two different codecs, with differing bitrates, may deliver the same quality to users; e.g. G.722[24] (64 kbps), G.722.1[25] (32 kbps) G.722.2[26] (12.65 kbps) have their $I_{e,WB}$ equal to 13 [15].

Secondly, a high degradation of quality may be associated with a codec with low I_e (or $I_{e,WB}$) in the presence of packet loss. An example of this may be AMR-NB (12.2 kbps) and iLBC (15.2 kbps); the former offers a better quality in the absence of packet losses, whereas the latter outperforms in the presence of losses [27]. This behavior is due to the *robustness of a codec against packet loss* and may depend on several factors such as , loss distribution (random or bursty), type of packet loss recovery algorithm employed by the codec and the time taken by a decoder's state to resynchronize with that of the coder in the event of packet loss [28].

This reflects the interpretation that codec-related effects on $I_{e,WB,eff}$ may be due to $I_{e,WB}$ and the robustness of that codec against packet loss.

C. Discussion

The E-Model uses two predefined parameters to compute $I_{e,eff}$; namely I_e and Bpl (packet loss robustness factor) along with packet loss statistics as in equation (7). The significance of these parameters has been discussed in section III-B. Similarly, in [29] Cole and Rosenbluth and in [27] Sun and Ifeachor have proposed a logarithmic function of the form:

$$I_{e,eff} = x_1 + x_2 \times ln(1 + x_3 \times mlr) \tag{8}$$

where they tune x_i to compute the codec specific $I_{e,eff}$ as a function of *mean packet loss rate* (mlr). It may be argued that their formulation of $I_{e,eff}$ depends on I_e (i.e. when mlr=0) and packet loss robustness, which translates into parameters x_2 and x_3.

Janssen et al. have depicted a relationship between codec specific $I_{e,eff}$ and packet loss rate in the form of quadratic curves [30].

It follows that different codecs may have different curves for $I_{e,WB,eff}$. This effect may be seen in Fig. 4. Here, for example, $I_{e,WB,eff}$ for the *Adaptive Multi-Rate-NB (AMR-NB)* [31] codec (7.4 kbps) may be approximated with a logarithmic curve (equation (8)), whereas the best fit curve for G.722.2 (19.85 kbps) was found to be a 4^{th} order polynomial.

It is clear that currently there is no widely accepted and *clearly superior* formulation for $I_{e,WB,eff}$. We suggest an alternative, altogether different, strategy. Instead of approaching the problem with *a priori* assumptions about the analytical form of $I_{e,WB,eff}$, we allow the data to speak for themselves. We propose to *evolve* high-quality expressions for $I_{e,WB,eff}$ using GP, a brief introduction to which is given in the next section.

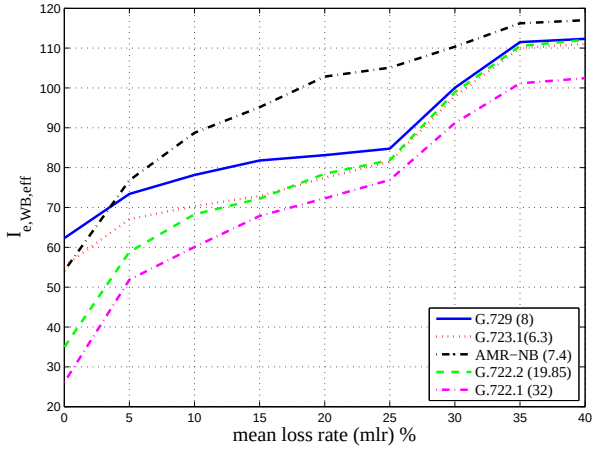


Fig. 4. $I_{e,WB,eff}$ as a function of mlr for various NB/WB codecs. values for $I_{e,WB,eff}$ were computed using WB-PESQ with random packet loss and PIs equal to one speech frame of the respective codecs.

IV. Genetic Programming

Genetic Programming (GP) [4] [32] is a biologically-inspired machine learning technique. It seeks to generate plausible approximate solutions to complex optimization problems by using concepts adopted, loosely, from natural evolution.

It has the advantage that, unlike many other optimization techniques, it can generate solutions (or quasi-solutions) to problems in symbolic form. Although, the solution representation is problem specific, it is common to use mathematical expressions or a subset of C/C++ for this purpose. GP produces human comprehensible results; an advantage when compared to approaches like Artificial Neural Networks (ANNs) where making sense out of a trained network can be quite a challenge [32, pp-85]. Another crucial advantage is that GP is not merely restricted to tuning the parameters of a pre-defined mathematical model like ANNs and other numerical optimization techniques. Instead, as in this paper, it also discovers the model itself with the primary aim of optimizing a user defined error metric. GP does not render numerical methods totally redundant, however. It has been used to advantage in conjunction with numerical optimization techniques such as linear regression [33], gradient descent [34] and quasi-Newton [35]. It has been suggested that the hybrid GP/numerical methods yield superior results by allowing GP to focus at the truly innovative aspect of the work, i.e., discovery of the model structure [33]. In this light a hybrid approach is used in this paper that combines GP with linear regression.

GP is coarsely modelled on natural evolution. Biological organisms aim to overcome environmental obstacles and compete for resources in a bid for survival and reproduction. GP evolves digital populations in a similar way. The environmental challenges are defined by an error metric that each member of the population, an *individual* or a candidate solution, seeks to minimise.

Initially, the population is created by generating a set of solutions randomly. To allow this, the syntactic constituents of an individual are pre-specified in the form of two sets: *functions* and *terminals*. Functions are exemplified by arithmetic operators, trigonometric functions and boolean functions as they require operands to produce an output. Terminals require no arguments. They may be, inter alia, numeric constants, system variables and functions with constant inputs. An initial population is generated by randomly picking from these sets, although other methods exist [36].

Each individual is tested on the given problem to assign it a measure of quality which is called its *fitness*. The fitness of a GP individual determines the chances of an individual surviving to the next *generation* or producing offspring. The offspring result from introducing some variation into the *selected* parent(s). Normally, there are two kinds of variation (or genetic) operators, *crossover* and *mutation*. Crossover involves combining the genetic material of different (normally two) individuals to produce new solutions. Sometimes completely new genetic material is introduced into the offspring, albeit with a small probability. This phenomenon is called mutation, and it is observed to be useful in GP by helping the system to work its way out of local minima (which are undesired, but inevitable, artifacts of the objective function).

Clearly, the genetic operators of crossover and mutation must work in a manner so the resulting offspring obey the syntactic constraints of the language used to represent the solutions. To facilitate this, the GP-individuals are maintained with data structures that are amenable to the carefully designed genetic operators. Abstract syntax trees are by far the most popular choice, although linear structures are also becoming common [36] [37]. Fig. 5 shows two example GP individuals undergoing crossover and the resulting offspring. Note that the crossover point is selected randomly in each individual and the subtrees rooted at those points are exchanged during the process. During mutation, a new subtree is randomly generated at the selected point, subject to a user specified *maximum depth* limit. As a result of the genetic operations the resulting offspring can be different in size and shape from their parents as is the case in the present example. This allows GP to explore a variable length solution space. However, to stop the trees growing arbitrarily large, again a maximum depth limit is employed. If the resulting offspring have larger depths, they are discarded.

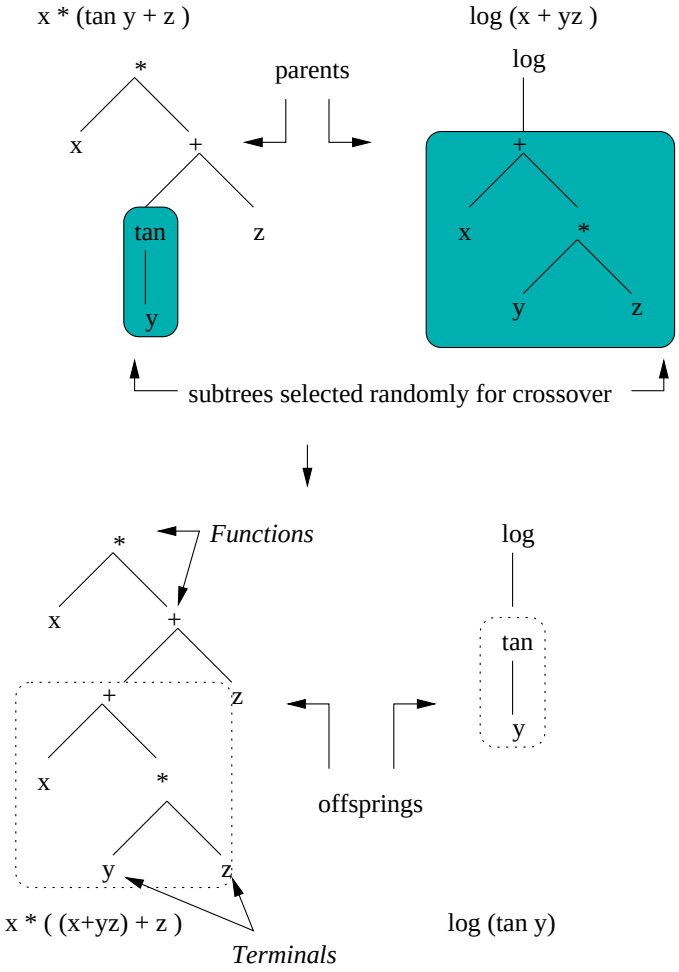


Fig. 5. Depicted are the example abstract syntax trees for GP individuals and the corresponding expressions. Functions are the internal nodes, while the terminals appear only as the leaves. The shaded portions in the upper trees represent the subtrees to be exchanged during crossover. The resulting offspring are shown underneath with dotted boundaries marking the exchanged fragments.

With this background we can now briefly describe the overall GP algorithm. The purpose is to breed better and better individuals as the evolution progresses through several generations until some user specified criterion is met. It may

be that some success criterion is fulfilled e.g. the squared error is reduced below a threshold or a maximum number of generations (a GP system parameter) have elapsed. Each generation typically entails the following steps (although variations exist):

1) if it is the first generation, an initialisation procedure is invoked [36, pp118-122] to produce the initial population of a fixed size;
2) choose two parents through a process termed *selection*. Different selection schemes exist e.g. tournament selection randomly picks n individuals and the best of them is the winner, while in roulette wheel selection the chances of getting picked are directly proportional to the fitness of an individual;
3) crossover is applied to yield two offspring which are then subjected to mutation. Typically crossover is used with a high probability (e.g. 90%) while mutation is used sparingly (1%);
4) evaluate the fitness of the two offspring;
5) if the number of offspring generated so far have reached a user specified limit, follow on to the next step. Otherwise, go to step 2;
6) in this study, the offspring and parent populations are considered together to keep the best performers for the next generation. The rest of the individuals from either pool are discarded. Other schemes may keep all the offspring as the population members for the next generation.

A number of generational cycles constitutes a GP *run*. Due to the stochastic nature of the evolutionary process, each run of GP can produce individuals that are different from those of the other runs with the same system parameters and fitness criteria. Therefore, it is a regular practice to conduct several runs in order to have different results of competitive quality and also to have a statistical justification of the behavior of GP. Detailed accounts of various aspects of GP can be found in [4] [36].

V. The new Methodology

In what follows we first describe our methodology for deriving $I_{e,WB,eff}$ as a function of VoIP traffic parameters. Next, we list the details of our data preparation procedure and of the VoIP simulations undertaken.

A. Methodology

Our methodology is based on our research presented in [5] and [6], with the main difference being that there the objective was to compute MOS for an NB context whereas here the main focus is on deriving equipment impairment factors, $I_{e,WB,eff}$, for a mixed NB/WB context. The schematic in Fig. 6 depicts a conceptual diagram of our approach for deriving $I_{e,WB,eff}$ for VoIP. An initial requirement is to have a database consisting of clean speech signals. These signals are subjected to degradations typical of VoIP traffic; coding distortions and packet loss. The degraded VoIP stream is eventually converted back to linear PCM format using a decoder corresponding to the encoder. In the process of doing so the values of various VoIP traffic parameters, such as packet loss rate, are calculated and the decoded speech signal is sent to a viable instrumental model that may report its results in terms of human assessment of speech quality i.e. MOS-LQO. Moreover, the model should be able to evaluate both NB and WB coded speech. An example of such a model is WB-PESQ, which has been used as a reference system in this research. The resulting MOS-LQO is converted to $I_{e,WB,eff}$ using equations (2), (3) and eventually (4). We call this the *target* $I_{e,WB,eff}$. The process is repeated for a large number of speech signals with varying degrees of network distortion conditions. Once the target $I_{e,WB,eff}$ for all the speech signals have been computed and the values of corresponding VoIP network traffic parameters gathered, GP based evolution is performed to derive a suitable mapping. More specifically, the VoIP network traffic parameters serve as the input domain variables during evolution and the corresponding $I_{e,WB,eff}$ values form the *target* output values.

A linear interpolation between the $I_{e,WB}$ obtained by the instrumental model (WB-PESQ) and subjective tests may be performed as suggested by [8, pp-9] to adjust the target $I_{e,WB,eff}$. To this end, interpolation was performed between $I_{e,WB}$ values for 20 (14 WB and 6 NB) codecs obtained using WB-PESQ, and from subjective tests reported in [15]. Here, for each codec, $I_{e,WB}$ corresponding to WB-PESQ was obtained by averaging the evaluations of 30 speech file pairs. Clean speech files were taken from experiment-1 of [11]. The slope and intercept were found to be equal to 0.6730 and 35.7881 respectively. It must be noted that a large intercept indicates an experimental bias [38, pp12]. This is possibly due to the fact that WB-PESQ underestimates the speech quality as compared with subjective tests, as discussed in section II-A.

B. Input Domain Variables

mlr, PI and *mean burst length* (mbl) were chosen as the input domain variables related to packet loss. $I_{e,WB}$ and a *coarse* estimate of loss robustness factor were computed for each codec separately as other independent parameters. It was expected of GP to make efficient use of these parameters during evolution. It was discussed in section III-C, and shown in Fig. 4, that the functional form of $I_{e,WB,eff}$ may vary for different schemes and codecs. Given this, the gradient of $I_{e,WB,eff}$ for mlr ranging between 0–0.3 was computed according to equation (9) as a *coarse* estimate of packet loss robustness factor, assuming that GP would use it effectively during evolution. This range of mlr is chosen because $I_{e,WB,eff}$ varies the most for mlr=0.0–0.3. After this the change is only gradual, as can be seen from Fig. 4. Moreover, the data presented by Sun and Ifeachor [27] imply the same for $I_{e,eff}$, where maximum mlr=0.3.

$$grad = \frac{I_{e,WB,eff}(mlr = 0.3) - I_{e,WB,eff}(mlr = 0.0)}{0.3} \tag{9}$$

Values of $I_{e,WB}$ and gradients of $I_{e,WB,eff}$ with respect to mlr for the codecs under consideration are listed in Table. I.

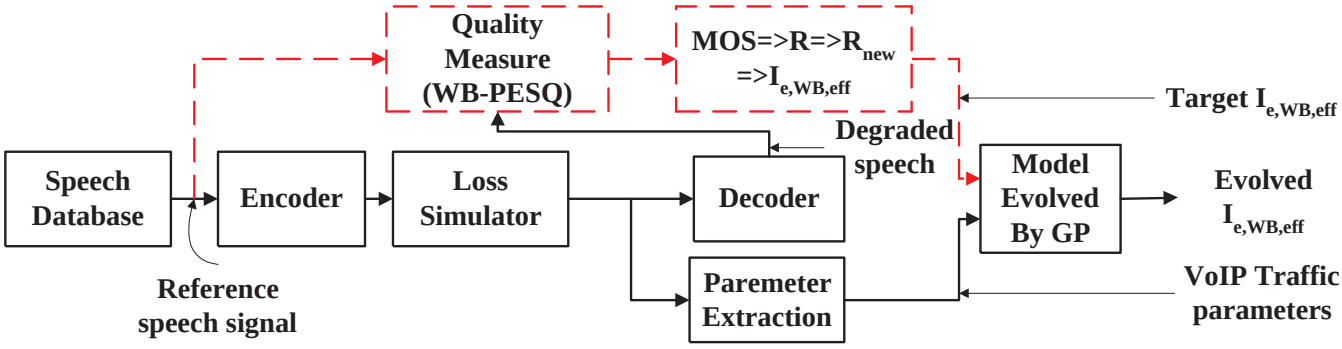


Fig. 6. Simulation system for derivation of $I_{e,WB,eff}$

TABLE I
VALUES FOR $I_{e,WB}$ AND COARSE ESTIMATES OF LOSS ROBUSTNESS FACTOR

Codec	bitrate	$I_{e,WB}$	gradient
G.722.1	32	26.12	216.88
G.722.1	24	29.04	208.36
G.722.2	6.6	68.13	104.25
G.722.2	8.85	58.64	139.67
G.722.2	12.65	43.91	187.62
G.722.2	14.25	41.19	196.13
G.722.2	15.85	39.59	201.50
G.722.2	18.25	36.09	212.81
G.722.2	19.85	34.97	213.20
G.722.2	23.05	32.09	225.27
G.722.2	23.85	33.88	221.27
G.729	8	62.33	125.66
G.723.1	6.3	55.27	142.14
AMR-NB	7.4	63.9	151.30
AMR-NB	12.2	54.12	187.48

C. VoIP Simulation

A simulation based approach was pursued in this research, where distortions typical of a VoIP network were induced on a large number of clean speech signals before decoding the corresponding coded bitstreams. Clean speech samples from experiments 1-A and 1-D of ITU-T P-series supplement 23 were used. The NB codecs include: ITU-T G.729 CS-ACELP (8 kbps) [39], ITU-T G.723.1 MP-MLQ/ACELP (5.3/6.3 kbps) [40] and AMR-NB codec [31]. AMR-NB was used in its 6.7 and 12.2 kbps modes whereas G.723.1 was used in its 6.3 kbps mode. The WB codecs include ITU-T G.722.1 [25] (24/32 kbps) and ITU-T G.722.2 [26], *Adaptive Multi-Rate (AMR-WB)* codec. AMR-WB can operate in 9 different coding/decoding modes, each targeting a different bit-rate: all the coding modes were utilized in this research.

Various network traffic simulation conditions were chosen in the light of ITU-T Recommendation G.1050 [41], which entails a model for evaluating multimedia transmission performance over an IP network. Bursty packet loss was emulated using a 2-state Markov model; with probabilities p, for transitioning from a no-loss state to a loss state and q, for the converse. It was assumed that $jitter$ also maps to packet loss and that it can be modeled using this 2-state model as in [21]. Packet loss for twelve different values of (target) mlr was simulated; [0,2.5,..., 15, 20, ..., 40]%. For each value of mlr, *conditional loss probability (clp) (i.e. 1-q)* was set to 10, 50, 60, 70 and 80%. It is worth mentioning that higher values of clp model higher degrees of loss burstiness and vice versa. Moreover, PI (packetization interval) was varied between 10–60 ms.

Since the clean speech samples are coded at a 16 kHz sampling rate, they were downsampled before encoding in the case of NB codecs. Subsequently, the corresponding decoded speech samples were upsampled before evaluation by WB-PESQ.

In all, 2,820 combinations of network distortion conditions were emulated. A given combination of network distortion conditions was applied to four speech samples. Moreover, each speech sample under consideration was subjected to the same combination of network distortion conditions 30 times to produce as many test samples by pseudo-randomly generating different loss patterns each time. This was done to negate the effect of packet loss locations as in [27] by eventually aggregating the MOS for all test samples corresponding to one source sample. Thus, a total of 338,400 distorted speech files were created. These distorted speech files were subsequently evaluated by WB-PESQ on a Beowulf cluster with respect to corresponding reference files. Values of the network traffic parameters for all files and the corresponding MOS were averaged to form a total of 11,280 input/output patterns, that would later be utilised during symbolic regression.

VI. EXPERIMENTS AND RESULTS

A. Experimental Details

Two GP experiments were performed to evolve models for $I_{e,WB,eff}$ using the input/output data patterns. The accumulation of data patterns has already been discussed in section V-C. GPLab was used for evolution which is a GP toolbox for Matlab developed by Sara Silva [3]. Previously in [6] we performed four GP experiments with different maximum tree depths and error measures with different results. In this work we chose the experimental conditions that produced superior results in terms of quality to perform the two GP experiments. The common parameters of both experiments are listed in Table II.

In both experiments scaled mean squared error (MSE_s) was used as the fitness criterion and is given by equation (10).

[3]http://gplab.sourceforge.net/

TABLE II
COMMON GP PARAMETERS AMONG ALL EXPERIMENTS

Parameter	Value
Initial Population Size	300
Initial Tree Depth	6
Selection	LPP
Tournament Size	2
Genetic Operators	Crossover and Subtree Mutation
Operators Probability Type	Adaptive
Initial Operator probabilities	0.5 each
Survival	Half Elitism
Generation Gap	1
Function Set	plus, minus, multiply, divide,sin, cos, log_2, log_{10}, log_e, sqrt, power
Terminal Set	Random real-valued numbers between 0.0 and 1.0. Integers (2-10). mlr, mbl, PI, $I_{e,WB}$, $grad$

$$MSE_s(y,t) = 1/n \sum_{i}^{n} (t_i - (a + by_i))^2 \qquad (10)$$

where y is a GP evolved function of the input parameters in this case (a mathematical expression), y_i represents the output value produced by y for the input case i and t_i represents the corresponding target value of $I_{e,WB,eff}$. a and b adjust the slope and y-intercept of the evolved expression to minimise the squared error. They are computed as follows:

$$a = \overline{t} - b\overline{y}, b = \frac{cov(t,y)}{var(y)} \qquad (11)$$

where $\overline{t}$ and $\overline{y}$ represent the mean values of the corresponding entities whereas var and cov are their variance and covariance respectively. This is known as *linear scaling* and has been found to be beneficial for the symbolic regression with GP [33].

Tournament selection with Lexicographic Parsimony Pressure (LPP) [42] was used in both experiments. In this selection strategy a group of G (G $\geq$ 2) individuals is picked randomly from the current population. The individual with the highest fitness in the group is selected as a parent. In the case of a tie between two or more individuals, their expression sizes are compared with the smaller individual winning out. Moreover, the selection criteria in both the experiments was also adapted to the one proposed by Gustafson et al. in [43] for symbolic regression problems. This requires that when the two parents are selected through tournament selection, they should be of different fitness values. This discourages parents with similar fitness and hence, possibly, of similar constitution producing offspring identical to themselves.

Whenever input values outside the domain of the functions *log, sqrt, division* and *pow* are encountered, NaN (undefined) values are generated. This results in the individual concerned being assigned the worst possible fitness value and minimising its chances of being selected as a parent.

As mentioned in section IV, it is typical to conduct several independent runs of GP. In this case, both experiments entailed 50 independent runs each spanning 50 generations.

The only difference between the two experiments was that in the first experiment the maximum tree depth was 17. This was reduced to 7 in the second experiment to see if parsimonious individuals with performance comparable to those of the first experiment can be obtained.

B. Results and analysis

Of 11,280 input/output patterns reported in section V-C, 1,440 patterns corresponding to AMR-NB 7.4 kbps and G.722.1 32 kbps were separated for model validation on *unseen* codecs. Of the remaining 9,840 patterns, 70% were used for training and 30% for testing the evolved models. Various VoIP traffic parameters have been discussed in section V-C. More specifically, these include, $I_{e,WB}$, mlr, PI, mean burst length (mbl) and $grad$, as in equation (9), as a coarse estimate of codec specific loss robustness factor.

The statistics pertaining to $RMSE_s$ (square root of the scaled MSE) of training and testing data of both GP experiments are listed in Table III(a). The table also lists various statistics related to the tree sizes of GP individuals, in terms of the number of nodes. The results of both experiments in the final generations were also treated to a Mann-Whitney Wilcoxon test to assay the significance of differences in various respects. The significance analysis is reported in Table III(b) where a value of '1' confirms a significant difference, at a 5% confidence level, whereas a '0' implies otherwise. It was found that the overall results of the two experiments are not significantly different from each other in terms of fitness over training and testing data. However, the difference in terms of tree size is significant, with experiment 2 having individuals with smaller trees.

In this paper we present three models resulting from the experiments. Two of these correspond to individuals with minimum $RMSE_s$ over the testing data in each of the experiments. These are represented by equations (12) and (13) and they belong to experiments 1 and 2 respectively. The third model, represented by equation (14) corresponds to the most parsimonious individual of both the experiments and is derived from experiment 2. The $RMSE_s$ and Pearson's product moment correlation coefficient (σ), corresponding to $I_{e,WB,eff}$ for these models are compared with each other in Table III(c). The values of $RMSE_s$ corresponding to *MOS-LQO* are also listed as another comparison. These were computed by converting the target values of $I_{e,WB,eff}$ and those obtained by the models under consideration to the MOS scale. This may be done by obtaining the values of R corresponding to $I_{e,WB,eff}$ from equation (4). The result can then be transformed to the original R scale for the NB-only context using equation (6); the inverse of equation (3). The resulting values of R can be converted to the MOS scale using transformation (2). The significance of all of the models can be judged by observing that the values of $RMSE_s$ on the MOS scale in all cases range between 0.098–0.12. This presents a considerably minute difference for a human subject to detect.

Equation (13) has the best statistics among all. Fig. 7 shows the scatter plots of equation (13) versus WB-PESQ, where it can be seen that the data points produced by both are firmly glued to the 45 degrees reference line.

TABLE III
STATISTICAL ANALYSIS OF THE GP EXPERIMENTS AND DERIVED MODELS

(a) MSE Statistics for Best Individuals of 50 Runs for Experiments 1 & 2

Stats	$Experiment1$			$Experiment2$		
	$RMSE_{tr}$	$RMSE_{te}$	$Size$	$RMSE_{tr}$	$RMSE_{te}$	$Size$
Mean	8.9478	32.5851	28.3617	8.9861	23.9743	19.02
Dev.	0.1890	113.2837	12.2144	0.2740	105.2397	6.3326
Max.	9.3624	655.5639	77	9.8275	753.2457	38
Min.	8.3941	8.5057	13	8.3552	8.4605	10

(b) Results of Mann-Whitney-Wilcoxon Significance Test

	$Experiment1$		
	$RMSE_{tr}$	$RMSE_{te}$	$Size$
$Experiment2$	0	0	1

(c) Performance Statistics of the Proposed Models

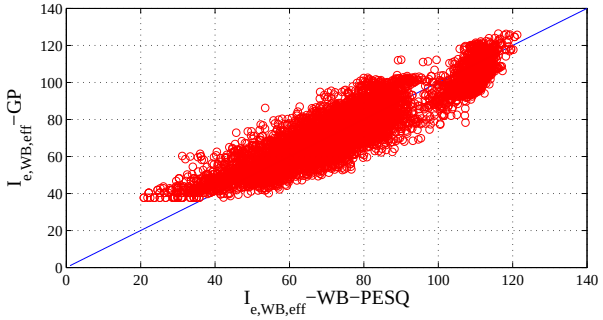
Model	Training			Testing		
	$RMSE_{sMOS}$	$RMSE_s\ I_{e,WB,eff}$	$\sigma\ I_{e,WB,eff}$	$RMSE_s\ MOS$	$RMSE_s\ I_{e,WB,eff}$	$\sigma\ I_{e,WB,eff}$
Equation (12)	0.0990	8.3941	0.9236	0.1007	8.5057	0.9240
Equation (13)	0.0975	8.3552	0.9243	0.0990	8.4605	0.9248
Equation (14)	0.1183	9.1749	0.9080	0.1207	9.3145	0.9080

$$I_{e,WB,eff} = \{11 - mbl + ln(grad) + grad \times mlr + I_{e,WB} - 2.log_2(PI)\} \times 0.8619 + 9 \quad (12)$$

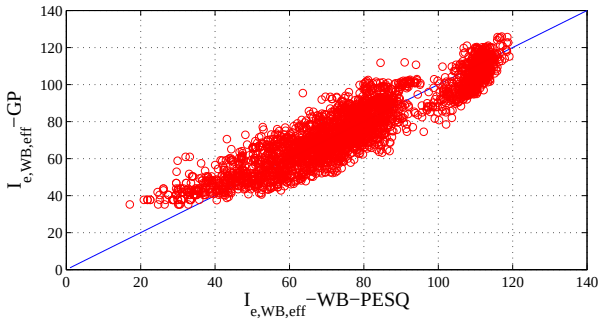
$$I_{e,WB,eff} = \left\{ln\left(\frac{9 \times (I_{e,WB} + mlr \times grad^2)}{mbl^5 - mlr}\right) + mlr + I_{e,WB} + grad \times mlr\right\} \times 0.8303 + 8.9977 \quad (13)$$

$$I_{e,WB,eff} = (log_{10}(log_{10}(log_2(I_{e,WB} - 2 \times mbl) + mlr))) \times 321.7017 + 95.3708 \quad (14)$$

A significance analysis of the various VoIP traffic parameters, in terms of their appearance in the best individuals of 50 runs of each of the two experiments, was done. The results are graphed in Fig. 8. According to this $I_{e,WB}$ and mlr had the highest utility, and appeared in 92–94% of the individuals. The third most sought-after parameter was $grad$, appearing in 36–38% of the best individuals of both experiments. mbl appeared in between 24–26% whereas, PI appeared in only 12% of the best individuals. The last two observations have also been reported by other researchers, such as [44] [45], who note that PESQ does not model the effect of burstiness on speech quality. We reported similar results in [5]. Fig. 9 illustrates similar behavior, but for WB-PESQ and a WB codec (AMR-WB 23.85 kbps). It is obvious that a correlation between $I_{e,WB,eff}$ and mbl does not exist. A similar comparison for the case of G.729 is shown in Fig. 10 where absence of any correlation between $I_{e,WB,eff}$ and PI may also be observed.



(a)



(b)

Fig. 7. $I_{e,WB,eff}$ predicted by equation (13) vs target $I_{e,WB,eff}$ for: (a) training data (b) testing data

C. Comparison with the E-Model

Finally, a comparison of equation (13) was made with the E-Model's formulation of the $I_{e,WB,eff}$, as in [2]. This is represented by:

$$I_{e,WB,eff} = I_{e,WB} + (129 - I_{e,WB}) \times \frac{P_{pl}}{P_{pl} + Bpl} \quad (15)$$

The equation is similar to equation (7) differing in the constant term, 95, which is replaced with the new R_{max}=129. The $BurstR$ parameter is also absent here. Bpl values for this equation were computed separately for each of the codecs over

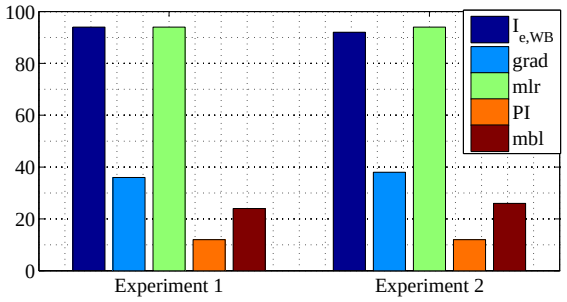


Fig. 8. Percentage of the best individuals employing various input parameters in acceptable runs of each of the two experiments.

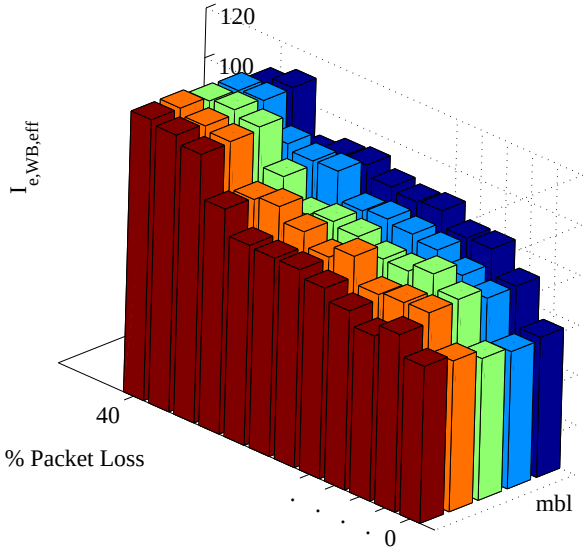


Fig. 9. Variation of $I_{e,WB,eff}$ against mlr (%) and $mbl = [1, \cdots, 5]$ for AMR-WB 23.85 kbps, PI=1.

the training data, and the performance was analysed using the testing data. Loss distributions were assumed to be random, which may be thought to be a reasonable assumption since it was shown in Figs. 9 and 10 that WB-PESQ estimates are oblivious of the effect of burstiness and varying PIs. The results are reported in Table IV for each codec. The table also shows the RMSE of equation (13) for AMR-NB (7.4 kbps) and G.722.1 (32 kbps). These codecs were not represented in the training data during evolution. Percentage *Prediction Gain (PG)* of 16.36 % was observed for unseen data in an RMSE sense. This is calculated according to equation (16)

$$\%PG = \frac{RMSE_e - RMSE_p}{RMSE_e} \times 100 \qquad (16)$$

where, $RMSE_e$ and $RMSE_p$ represent the $RMSE$ of equations (15) and (13) respectively.

However, equation (15) (i.e. the E-model formulation) does not account for bursty packet losses and also for PI spanning multiple speech frames. Moreover, all the models proposed in this research (by equations (12)–(14)) are functions of mbl. Given this, a comparison between E-model and the proposed models over datasets that include various degrees of burstiness and PIs is somewhat unfair. To ensure fairness a different

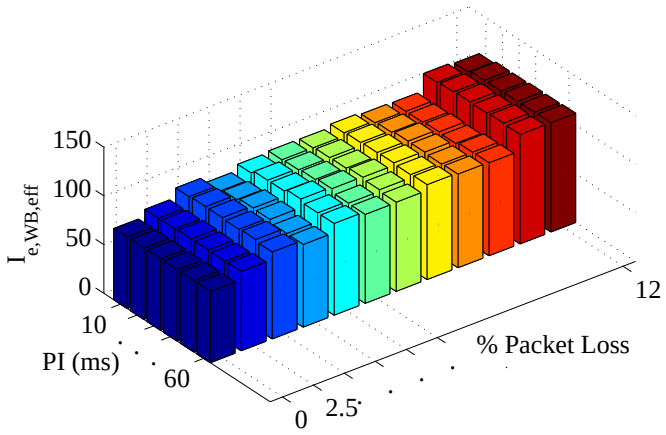


Fig. 10. Variation of $I_{e,WB,eff}$ against mlr (%) and $PI = [10, \cdots, 60ms]$ for G.729

TABLE IV
COMPARISON BETWEEN THE PREDICTION ACCURACIES OF THE E-MODEL AND THE PROPOSED MODEL

Codec (kbps)	E-Model			Equation (13)	
	Bpl	RMSE train	RMSE test	RMSE train	RMSE test
G.722.1 (24)	20.32	8.6824	8.8958	8.1701	8.9118
G.722.2 (6.6)	40.75	9.6225	8.9933	8.0938	7.6603
G.722.2 (8.85)	28.74	10.0175	9.9919	8.0185	7.8304
G.722.2 (12.65)	21.58	10.5538	10.4088	8.2188	8.0678
G.722.2 (14.25)	21.03	10.4684	11.2854	8.3031	8.5836
G.722.2 (15.85)	19.98	10.599	11.5020	8.3257	9.1166
G.722.2 (18.25)	19.48	11.2017	10.92	8.6862	9.0266
G.722.2 (19.85)	18.86	10.5502	11.3529	8.2338	8.7685
G.722.2 (23.05)	18.44	11.4079	11.1663	9.1417	8.7729
G.722.2 (23.85)	17.92	10.789	11.1948	8.6125	9.3168
G.729 (8)	28.43	8.95	9.1631	7.3888	7.4943
G.723.1 (6.3)	29.19	10.83	10.3630	8.8116	8.5259
AMR-NB (12.2)	13.50	8.0689	7.2947	9.4549	8.7322
G.722.1 (32)	18.93	8.9112	–	–	8.4775
AMR-NB (7.4)	15.71	7.1335	–	–	8.6188
Average	–	9.8527	10.1946	8.42	8.5269
% PG	–	–	–	14.54	16.36

simulation study was performed in which speech files were subjected to random packet losses with loss rates ranging between [0,2.5,..., 15, 20, ..., 40]% for each of the encoding conditions. The results are reported in Table V. To this end, the data was split into training and testing datasets as previously. Bpl values were recalculated for each of the codecs and the RMSE was noted with respect to WB-PESQ. The performance of equation (13) was found to be inferior to the traditional E-model formulation initially. Since random loss conditions were alien to the GP training conditions, the performance degradation was not unexpected when compared to a retuned E-model.

However, upon merely re-scaling equation (13) using training data a prediction gain of approximately 36% was obtained. This shows the robustness of the model produced by GP as evolutionary re-training was not required. Linear re-scaling resulted in new *slope* and *intercept* terms which were found to be 0.5085 and 46.7468 respectively. Linear re-scaling was done by treating equation (13) with equations (10) and (11).

TABLE V
COMPARISON BETWEEN THE PREDICTION ACCURACIES OF THE E-MODEL AND THE PROPOSED MODEL FOR RANDOM LOSS CONDITIONS

Codec (kbps)	E-Model			Equation (13)		Equation (13) after re-scaling	
	Bpl	RMSE train	RMSE test	RMSE train	RMSE test	RMSE train	RMSE test
G.722.1 (24)	11.9699	12.2622	13.1168	14.9678	14.8504	6.6551	6.9227
G.722.2 (6.6)	24.0580	8.5488	8.5060	8.2690	7.8553	6.9593	7.2537
G.722.2 (8.85)	14.6072	9.7573	9.8179	11.2923	11.3162	5.9292	6.2118
G.722.2 (12.65)	10.6167	11.2011	11.2734	14.4629	14.7035	6.6388	6.7783
G.722.2 (14.25)	10.0051	11.2489	11.1616	15.0003	14.8100	6.2532	6.5631
G.722.2 (15.85)	9.8967	11.7606	12.1703	15.5678	15.6983	6.4557	6.9346
G.722.2 (18.25)	9.3617	12.3315	12.7594	16.5918	17.0715	6.9336	7.3102
G.722.2 (19.85)	9.0622	12.3594	12.1367	16.8570	16.8927	6.9994	6.9165
G.722.2 (23.05)	8.3718	13.1165	12.9824	18.1885	17.9847	7.1816	7.3181
G.722.2 (23.85)	8.5254	12.6131	12.4198	17.7068	17.3806	7.0750	6.6818
G.729 (8)	16.8150	8.5228	8.7014	9.6080	9.7882	5.5697	5.5535
G.723.1 (6.3)	14.9484	9.5020	10.3181	10.7527	11.1905	6.4791	6.6315
AMR-NB (12.2)	5.7637	8.6698	8.3887	17.9201	18.4502	9.2798	9.6566
G.722.1 (32)	11.6870	14.0378	13.6078	16.3347	15.7470	7.4048	7.8482
AMR-NB (7.4)	7.5927	7.8755	8.5562	14.7356	14.7646	7.8499	7.7822
Average	–	10.9205	11.0611	14.5504	14.5669	6.9109	7.0909
% PG	–	–	–	-33.2393	-31.6949	36.7163	35.8934

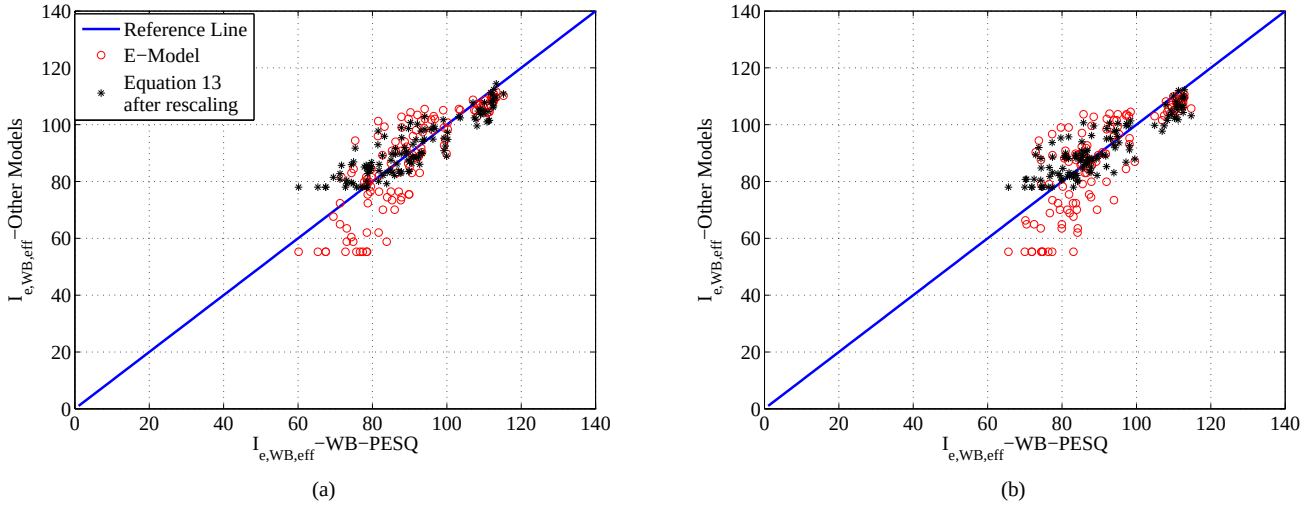


Fig. 11. $I_{e,WB,eff}$ predicted by equation(15) (i.e. the E-Model) and equation (13) vs target $I_{e,WB,eff}$ obtained from WB-PESQ for random loss: (a) training data (b) testing data.

A pictorial comparison similar to Fig. 7 is also done between equations (13) and (15) with respect to WB-PESQ in Fig. 11 for the case of ITU-T G.723.1 codec. It can be observed that the points produced by equation (13) are more firmly glued to the 45% reference line as compared to those produced by equation (15).

VII. CONCLUSIONS

In this paper we have proposed a novel methodology for determining NB/WB equipment impairment factors, $I_{e,WB,eff}$, for a mixed NB/WB context. It is based on using GP to perform symbolic regressions which generate simple formulae for $I_{e,WB,eff}$. It is advantageous in the sense that the derived models do not result from human bias, but as a direct consequence of program evolution. Moreover, parameter optimization is done in parallel with evolution for every model using linear scaling. The derived models are applicable for the network distortion conditions under observation. Our approach utilizes WB-PESQ for deriving reference values of $I_{e,WB,eff}$ as opposed to subjective tests. This is suitable for fast and inexpensive derivation of reference $I_{e,WB,eff}$. We have demonstrated the utility of our approach by generating three models for $I_{e,WB,eff}$ from different GP runs. The proposed models were thoroughly tested on a wide variety of VoIP traffic scenarios including a blend of modern IP telephony codecs.

A comparison of equation (13), which has the best performance among the proposed models, with the E-Model, equation (15), has also been done, where it is shown that our approach outperforms the E-Model with a significant margin in terms of prediction accuracy. Even though we have used WB-PESQ in this research, the proposed methodology is independent of it and simply requires a generic instrumental model of this kind. The methodology may also be augmented

with subjective tests.

REFERENCES

[1] ITU-T, *The E-Model, a computational model for use in transmission planning*, International Telecommunications Union, Geneva, Switzerland, 2005, ITU-T Recommendation G.107.

[2] S. Moller, A. Raake, N. Kitawaki, A. Takahashi, and M. Waltermann, "Impairment factor framework for wide-band speech codecs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1969–1976, November 2006.

[3] ITU-T, *Methods for subjective determination of transmission quality*, International Telecommunications Union, Geneva, Switzerland, 1996, ITU-T Recommendation P.800.

[4] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[5] A. Raja, R. M. A. Azad, C. Flanagan, D. Picovici, and C. Ryan, "Non-intrusive quality evaluation of $voip$ using genetic programming," in *First International Conference on Bio Inspired Models of Network, Information and Computer Systems*, vol. 4, 2006, pp. 2573–2577.

[6] A. Raja, R. M. A. Azad, C. Flanagan, and C. Ryan, "Real-time, non-intrusive evaluation of VoIP," in *Proceedings of the 10th European Conference on Genetic Programming*, ser. Lecture Notes in Computer Science, M. Ebner, M. O'Neill, A. Ekárt, L. Vanneschi, and A. I. Esparcia-Alcázar, Eds., vol. 4445. Valencia, Spain: Springer, 11 - 13Apr. 2007, pp. 217–228.

[7] ITU-T, *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, International Telecommunications Union, Geneva, Switzerland, 2005, ITU-T Recommendation P.862.2.

[8] ——, *Methodology for the derivation of equipment impairment factors from instrumental models*, International Telecommunications Union, Geneva, Switzerland, 2002, ITU-T Recommendation P.834.

[9] V. Barriac, J. Y. Sout, and C. Lockwood, "Discussion on unified objective methodologies for the comparison of voice quality of narrowband and wideband scenarios," in *In. Proc. Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction*, 2004.

[10] ITU-T, *Mean opinion score (MOS) terminology*, International Telecommunications Union, Geneva, Switzerland, 2003, ITU-T Recommendation P.800.1.

[11] ——, *coded-speech database*, International Telecommunications Union, Geneva, Switzerland, 1998, ITU-T P.Supplement 23.

[12] ——, *Software tools for speech and audio coding standardization*, International Telecommunications Union, Geneva, Switzerland, September 2005, ITU-T Recommendation G.191.

[13] A. Raake, *Speech Quality of VoIP Assessment and Prediction*. John Wiley and Sons Inc, 2006.

[14] C. Morioka, A. Kurashima, and A. Takahashi, "Proposal on objective speech quality assessment for wideband telephony," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2004.

[15] ITU-T, *New Appendix IV – Provisional planning values for the wideband equipment impairment factor Ie,wb*, International Telecommunications Union, Geneva, Switzerland, June 2006, ITU-T Recommendation G.113.

[16] H. Sanneck and G. Carle, "A framework model for packet loss metrics based on loss runlengths," in *SPIE/ACM SIGMM Multimedia Computing and Networking Conference*, January 2000.

[17] W. Jiang and H. Schulzrinne, "Modeling of packet loss and delay and their effect on real-time multimedia service quality," in *In Proc. NOSSDAV*, June 2000.

[18] A. D. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality," in *2nd IP-Telephony Workshop*, Columbia University, New York, April 2001.

[19] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, pp. 40–48, Sep/Oct 1998.

[20] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Codecs*. Cambridge, MA, USA: John Wiley and Sons Inc, May 2003.

[21] Lingfen and E. C. Ifeachor, "perceived speech quality prediction for voice over ip-based networks," in *IEEE International Conference on Communications (ICC)*, vol. 4, 2002, pp. 2573–2577.

[22] S. Mohamed, G. Rubino, and M. Varela, "A method for quantitative evaluation of audio quality over packet networks and its comparison with existing techniques," in *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, 2004.

[23] C. Hoene, H. Karl, and A. Wolisz, "A perceptual quality model intended for adaptive VoIP applications," *International Journal of Communications Systems*, vol. 99, no. 7, pp. 1–20, August 2005.

[24] ITU-T, *7 kHz Audio coding within 64 kbit/s*, International Telecommunications Union, Geneva, Switzerland, November 1988, ITU-T Recommendation G.722.

[25] ——, *Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss*, International Telecommunications Union, Geneva, Switzerland, May 2005, ITU-T Recommendation G.722.1.

[26] ——, *Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*, International Telecommunications Union, Geneva, Switzerland, July 2003, ITU-T Recommendation G.722.2.

[27] L. Sun and E. C. Ifeachor, "Voice quality prediction models and their application in VoIP networks," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 809– 820, August 2006.

[28] J. D. Rosenberg, "G.729 error recovery for internet telephony," Columbia University Computer Science Technical Report CUCS-016-01, Tech. Rep., 19 December 2001.

[29] R. G. Cole and J. H. Rosenbluth, "Voice over ip performance monitoring," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 2, pp. 9–24, 2001.

[30] J. Janssen, D. D. Vleeschauwer, M. Buchli, and G. H. Petit, "Assessing voice quality in packet-based telephony," *IEEE Internet Computing*, vol. 6, no. 3, pp. 48–56, 2002.

[31] ETSI EN 301 704 V7.2.1, *Digital cellular telecommunications system; Adaptive Multi-Rate (AMR) speech transcoding*.

[32] T. Mitchell, *Machine Learning*. New York: McGraw Hill, 1997.

[33] M. Keijzer, "Scaled symbolic regression," *Genetic Programming and Evolvable Machines*, vol. 5, no. 3, pp. 259–269, September 2004.

[34] A. Topchy and W. F. Punch, "Faster genetic programming based on local gradient search of numeric leaf values," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, L. Spector, E. D. Goodman, A. Wu, W. B. Langdon, H.-M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. H. Garzon, and E. Burke, Eds. San Francisco, California, USA: Morgan Kaufmann, 7-11Jul. 2001, pp. 155–162. [Online]. Available: http://www.cs.bham.ac.uk/ wbl/biblio/gecco2001/d01.pdf

[35] E. M. Mugambi, A. Hunter, G. Oatley, and L. Kennedy, "Polynomial-fuzzy decision tree structures for classifying medical data," *Knowledge-Based Systems*, vol. 17, no. 2-4, pp. 81–87, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/B6V0P-4C4VYG9-2/2/8ee7c8541e99bf3c8c22922dad2ebfbf

[36] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, *Genetic Programming – An Introduction; On the Automatic Evolution of Computer Programs and its Applications*. Morgan Kaufmann, dpunkt.verlag, 1998.

[37] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 4, pp. 349–358, 2001.

[38] ITU-T, *Methodology for derivation of equipment impairment factors from subjective listening-only tests*, International Telecommunications Union, Geneva, Switzerland, 2001, ITU-T Recommendation P.833.

[39] ——, *Coding of Speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)*, International Telecommunications Union, Geneva, Switzerland, March 1996, ITU-T Recommendation G.729.

[40] ——, *Dual rate speech coder for multimedia communication transmitting at 5.3 and 6.3 kbit/s*, International Telecommunications Union, Geneva, Switzerland, March 1996, ITU-T Recommendation G.723.1.

[41] ——, *Network model for evaluating multimedia transmission performance over internet protocol*, International Telecommunications Union, Geneva, Switzerland, November 2005, ITU-T Recommendation G.1050.

[42] S. Luke and L. Panait, "Lexicographic parsimony pressure," in *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference*, W. B. L. et. al., Ed., New York, 2002, pp. 829–836.

[43] S. Gustafson, E. K. Burke, and N. Krasnogor, "On improving genetic programming for symbolic regression," in *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, D. C. et. al., Ed., vol. 1. Edinburgh, UK: IEEE Press, 2-5Sep. 2005, pp. 912–919.

[44] L. Sun and E. C. Ifeachor, "Subjective and objective speech quality evaluation under bursty losses," in *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, January 2002.

[45] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm," in *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, January 2002.

Adil Raja received his B.Eng. in metallurgical engineering and materials science in 2000 from the University of Engineering and Technology, Lahore, Pakistan and MS in computer sciences from the Lahore University of Management Sciences (LUMS), Pakistan. Currently he is a candidate for PhD in the University of Limerick, Ireland where the topic of his research is non-intrusive estimation of speech quality of VoIP.

His other research interests include computer aided diagnosis, machine learning and evolutionary algorithms.

R. Muhammad Atif Azad is a post doctoral researcher with the Biocomputing and Developmental Systems Group, University of Limerick, Ireland.

He received B.E. in Computer Software Engineering from the National University of Sciences and Technology, Pakistan in 1999. He then moved to University of Limerick to complete his doctorate in Grammatical Genetic Programming and its variants in 2003. His research interests include theory and applications of Genetic Algorithms, Genetic Programming, Grammatical Evolution and its variants.

Colin Flanagan received the B.Eng. degree in electronic engineering, the M.Eng degree in computer engineering and the Ph.D. degree from the University of Limerick, Limerick, Ireland, in 1986, 1988 and 1991 respectively.

He is a Senior Lecturer in the Department of Electronic and Computer Engineering, University of Limerick. His research interests include QoS in VoIP, computer architecture and network processors.

Conor Ryan received his BA degree in Computer Science and Economics from University, College Cork in Ireland in 1992 and his PhD from there in 1996.

He is a Senior Lecturer in the University of Limerick where he heads the Biocomputing and Developmental Systems research group, who introduced the popular Grammatical Evolution automatic programming system.

He has been a Science Foundation of Ireland funded Principal Investigator since 2002.