VoIP Speech Quality Estimation in a Mixed Context with Genetic Programming

Adil Raja
ECE Department
University of Limerick
Ireland
adil.raja@ul.ie

R. Muhammad Atif Azad
CSIS Department
University of Limerick
Ireland
atif.azad@ul.ie

Colin Flanagan
ECE Department
University of Limerick
Ireland
colin.flanagan@ul.ie

Conor Ryan
CSIS Department
University of Limerick
Ireland
conor.ryan@ul.ie

ABSTRACT

Voice over IP (VoIP) speech quality estimation is crucial to providing optimal Quality of Service (QoS). This paper seeks to provide improved speech quality estimation models with better prediction accuracy by considering a richer set of input features than the current International Telecommunications Union-Telecommunication (ITU-T) recommendations. It addresses a transitional phase, where wideband (WB) networks are becoming available. However, they have to co-exist with the existing narrowband (NB) setups for the time being. Quality estimation becomes a challenge in such a *mixed* context. The ITU-T recommendation (termed E-Model) has recently been extended to deal with the mixed context. However, it evaluates the speech degradation in the WB scenario based solely on *codec related distortions* (only a subset of factors affecting the speech quality on a VoIP network). The extension is derived out of speech signals evaluated by human subjects: an expensive and difficult to reproduce exercise. This paper innovates by considering a number of other network distortion types as well to produce generalised models that predict the quality degradation to a higher accuracy. To this end, an extensive set of speech samples is subjected to a wide variety of distortions. The degraded signals are evaluated by the currently best available *algorithmic approximation* of human evaluation of speech to produce quality scores. Using the distortions as the input features and targeting the quality scores, we employ Genetic Programming to produce parsimonious models that show considerable prediction gain compared to the E-Model. As against some existing approaches, where the models are tailored to various telephony codecs, the evolved models generalise across a variety of modern codecs.

GECCO'08, July 12-16, 2008, Atlanta, Georgia, USA.

Categories and Subject Descriptors

I.2.2 [**Artificial Intelligence**]: Automatic Programming—*Program Synthesis*; I.2.6 [**Artificial Intelligence**]: Learning—*Induction*; I.5.1 [**Pattern Recognition**]: Models

General Terms

Reliability, Standardisation, Performance, Algorithms

Keywords

E-Model, Genetic Programming, Symbolic Regression, PESQ-WB, Speech Quality, VoIP, $I_{e,WB,eff}$.

1. INTRODUCTION

VoIP has started observing a drift towards wideband (WB) transmission of speech due to certain advantages. Firstly, it offers superior quality due to bandwidth extension from 300-4000 Hz (in the case of traditional narrowband telephony) to 50-7000 Hz. This bandwidth extension is believed to make the speech sound smoother and more natural. Secondly, as opposed to the traditional circuit-switched systems, the underlying IP network supports this leisure to some convenience. Before VoIP can be transformed into a WB only telecommunications system, it will have to coexist with the current narrow-band (NB) based systems. This gives rise to a transition phase in which WB systems would either operate in parallel or in cascade with NB systems. The first case is instantiated when the participants of a given conversation are equipped with NB codecs whereas another call uses WB codecs. The second case (i.e. the cascade case) represents scenarios where one of the participants is using a WB codec and the other is using an NB one. The coded speech frames, in this case, would have to be trans-coded into a format acceptable to the recipient. Both of these cases represent the so called *mixed NB/WB* scenarios. These developments also require speech quality estimation models to operate effectively in a mixed NB/WB context.

VoIP quality is affected by various factors such as packet loss, end-to-end delay, jitter and codec bit-rate etc. Different approaches and models estimate speech quality as a function of such impairments. ITU-T Recommendation G.107 [12], commonly known as the E-Model is of special

interest, however. It assumes that the degradations induced by various sources have a cumulative effect on speech quality and suggests the degradation transformation into a *transmission rating scale (R scale)*. The E-Model was originally intended for NB speech quality estimation. Recently, Möller et al. [22], 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. They derive a model for estimating the degradation in the listening quality of speech, termed as *equipment impairment factors* ($I_{e,WB}$), in a mixed NB/WB context in the wake of *pure codec related distortions*. Their derivation is based on subjective *listening only* tests [7], (where human subjects evaluate the quality) 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. However, this paper differs from the past endeavours as it deals with the *mixed NB/WB* case i.e., it derives $I_{e,WB,eff}$ and entails following novelties.

- *Instrumental models* are used to propose reference values for quality degradation ($I_{e,WB,eff}$), as opposed to the expensive subjective tests that are hard to reproduce. The reference values are required as target values for training and testing new models. Instrumental models are computational models which provide reproducible results saving the trouble of hiring human subjects. However, they may require both the clean and the noisy signal to evaluate the quality. Thus, such models are not employed for real time evaluation and are primarily useful for producing reference values to facilitate model induction, as in this study.

- The mapping between various quality affecting parameters and reference $I_{e,WB,eff}$ is achieved by employing Genetic Programming (GP) [19]. This approach is inspired by the research reported in [26, 27] where GP has been used to derive parsimonious speech quality estimation models. However, those studies are restricted to NB only context.

- The mapping is not limited to just the codec related distortions as in [22]. Instead, it considers several other network distortion conditions as well, as detailed in section 4.2. The benefit is reflected in the improved prediction accuracy compared to the ITU-T E-Model formulation in spite of the fact that the models are not separately evolved for each codec. This is a significant achievement.

In this research we have employed a number of state-of-the-art VoIP telephony codecs proposed by ITU-T. The instrumental model used for producing reference values for training and testing purposes is ITU-T P.862.2 (i.e.WB-PESQ) [15]. Despite its limitations [25, pp-105] [23], it is currently regarded as the best approximation of human speech quality estimation based on parametric distortions. However, the experimental approach is not tied down to WB-PESQ. Should better quality instrumental models arise or a comprehensive database of subjective tests materialise, the evolutionary experiments can be conveniently repeated. We follow the methodology described in [9] 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 2 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 3 we discuss the factors that affect $I_{e,WB,eff}$. Section 4 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 5. Finally, section 6 concludes the paper.

2. THE E-MODEL

The E-Model, as defined by ITU-T G.107 [12], is a computational model used for assessing the combined effect of various parameters on speech quality in a conversational sense. Initially designed for NB handset telephony, its adaptation to the WB case is work under progress. The primary output of the model is the *Rating Factor*, R under the assumption that factors affecting speech quality are additive in nature.Thus,

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

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 such as 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, 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 [12].

R can be transformed back and forth into *Mean Opinion Score (MOS)* as specified in [12]. We employ the transformations in this work and refer to them by an abstract notation given in (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 above formulations hold for the case of NB codecs. In [22] 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 the subjective tests performed in [1]. This transformation was required because the *same* NB coded samples were graded better by the human subjects in the absence than in the presence of WB coded samples. This results from the human tendency to judge the speech samples in a *relative* sense. Correspondingly, MOS to R conversion (2) yielded R_{NB} (purely NB context) 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).

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

where $a = 169.38$ and $b = 176.32$ and $R_{NB/WB}$ can be calculated via (2). This extension is now an integral part of the E-Model (see Appendix II of [12]), 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 bit rate 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.

$$I_{e,WB} = 129 - R_{codec} \quad (4)$$

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).

However, in our work WB-PESQ replaces subjective tests as a reference for deriving $I_{e,WB}$, that considers only codec related distortions and $I_{e,WB,eff}$ which considers several other impairment factors to be described later. A WB version of R scale does not exist in the literature for WB-PESQ. Thus, we propose to convert the MOS-LQO *(MOS-Listening Quality Objective)* [10] obtained by WB-PESQ 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 [9] and is used to derive reference values of $I_{e,WB,eff}$ in this research.

3. $I_{E,WB,EFF}$ AND ASSOCIATED QUALITY ELEMENTS

According to the E-Model [12] $I_{e,eff}$ for a given NB codec may be computed from

$$I_{e,eff} = I_e + (95 - I_e) \times \frac{P_{pl}}{\frac{P_{pl}}{BurstR} + Bpl} \quad (5)$$

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 (%). 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 [25][28] [16][2]. 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. Bpl is the packet loss robustness factor for the codec under consideration. It describes the robustness of the codec, including the employed packet loss concealment mechanism, against 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 [30]. A similar formulation for $I_{e,WB,eff}$ is given in [22] for *random* packet loss.

Another factor closely associated with packet loss is the packetization interval (PI) (ms), i.e., the acoustic content of an IP packet. An increase in PI leads to efficient use of bandwidth. 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 [25].

Given this, $I_{e,WB,eff}$, or equivalently $I_{e,eff}$, depends on two quality elements, namely *packet loss* and *codec*. However, currently there is no widely accepted and *clearly superior* formulation for $I_{e,WB,eff}$ encompassing both these elements. Therefore, we propose the data to speak for themselves. We choose to *evolve* high-quality expressions for $I_{e,WB,eff}$ using GP and discuss our methodology in the next section.

4. EVOLVING $I_{E,WB,EFF}$

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.

4.1 Parametric $I_{e,WB,eff}$ Formulation

Our approach is similar to [26] and [27]. However, there the objective is to compute MOS for an NB context whereas here the focus is on deriving equipment impairment factors, $I_{e,WB,eff}$, for a mixed NB/WB context. Figure 1 presents a conceptual diagram of our approach for deriving $I_{e,WB,eff}$ for VoIP. A set of clean speech signals from a database are treated to distortions typical of VoIP traffic i.e. coding (through the encoder) distortions and packet loss (using the Gilbert Loss simulator). The values of various VoIP traffic parameters, such as packet loss rate, are also calculated from the degraded samples. The degraded VoIP stream is then decoded for quality estimation with an instrumental model using the decoder corresponding to the encoder. The model should be able to report its results (both for NB and WB coded samples) in terms of human assessment of speech quality i.e. MOS-LQO. WB-PESQ is such a model which has been used as a reference system in this research. It can, however, be replaced should a more suitable model arise.

MOS-LQO is converted to $I_{e,WB,eff}$ using equations (2), (3) and eventually (4). This forms 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 $I_{e,WB,eff}$ values have been computed for all the speech signals and the corresponding VoIP network traffic parameters evaluated, GP takes over to evolve a suitable mapping. The VoIP network traffic parameters serve as the input variables during evolution and the corresponding $I_{e,WB,eff}$ values are the *target* output values for GP.

Clean speech files were taken from experiment-1 of [8].

4.2 Input Variables for the Evolved Models

Mean loss rate 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 are computed for each codec separately as other independent parameters. Due to the dependence on the codecs, $I_{e,WB,eff}$ varies for different loss rates and codecs. Given this, the gradient of $I_{e,WB,eff}$ for *mlr* ranging between 0–0.3 was computed according to equation (6) as a *coarse* estimate of packet loss robustness factor. $I_{e,WB,eff}$ fluctuates the most (details omitted due to space restrictions) during this range of *mlr* whereas after this the change is only gradual.

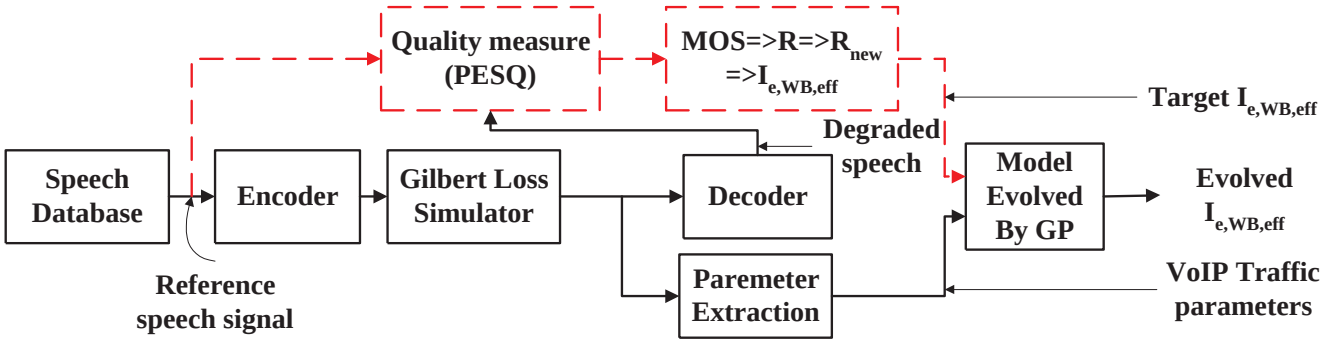

Figure 1: Simulation system for derivation of $I_{e,WB,eff}$

This claim is corroborated by the data presented by Sun and Ifeachor [30] for $I_{e,eff}$, where maximum mlr=0.3.

$$grad = \frac{I_{e,WB,eff}(mlr = 0.3) - I_{e,WB,eff}(mlr = 0.0)}{0.3} \quad (6)$$

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. 1.

Table 1: 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

4.3 VoIP Traffic Simulation

VoIP traffic was simulated and 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 ITU-T P-series supplement 23 were used. The NB codecs include: ITU-T G.729 CS-ACELP (8 kbps) [5], ITU-T G.723.1 MP-MLQ/ACELP (5.3/6.3 kbps) [6] and AMR-NB codec [3]. 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 [13] (24/32 kbps) and ITU-T G.722.2 [11], *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 as per ITU-T Recommendation G.1050 [14], 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 [20]. 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 [30] 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.

5. EXPERIMENTS AND RESULTS

5.1 GP Experimental Setup

Two sets of experiments were performed to evolve models for $I_{e,WB,eff}$ using the input/output data patterns gathered as mentioned earlier. GPLab, a GP toolbox for Matlab [1] was employed for evolutionary runs. Previously [27], four GP experiments were conducted with different maximum tree depths and error measures with varying results. This work

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

chooses the two most fruitful experimental conditions for superior output quality. The common parameters of both experiments are listed in Table 2.

Table 2: Common GP Parameters among all experiments

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

Both the experiments used scaled mean squared error (MSE_s) as the fitness criterion which is given by equation (7).

$$MSE_s(y,t) = 1/n \sum_{i}^{n} (t_i - (a + by_i))^2 \quad (7)$$

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:

$$a = \overline{t} - b\overline{y}, b = \frac{cov(t,y)}{var(y)} \quad (8)$$

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 [18].

Tournament selection with Lexicographic Parsimony Pressure (LPP) [21] was used in both experiments. Moreover, the selection criteria in both the experiments was also adapted to the one proposed by Gustafson et al. in [4] 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 assigns the responsible individual the worst possible fitness value and minimises its chances of acting as a parent. This approach is also used in [17] where the *protected* operators are blamed for overfitting and asymptotic anomalies.

Both the experiments entailed 50 independent runs each spanning 50 generations. The only difference between the two experiments was that of maximum tree depth: the first experiment had it at 17 whereas the second restricted it to 7. This was to see if parsimonious expressions with comparable performance could be obtained.

5.2 Results and Analysis

Of 11,280 input/output patterns reported in section 4.3, 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 4.3. Specifically, these include, $I_{e,WB}$, mlr, PI, mean burst length (mbl) and $grad$, as in equation (6), 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 3(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 3(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 (9) and (10) and they belong to experiments 1 and 2 respectively. The third model, represented by equation (11) 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 3(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 by inverting 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.

Overall equation (10) has the best statistics.

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

Table 3: 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		
	$RMSE_{sMOS}$	$RMSE_s$ $I_{e,WB,eff}$	σ $I_{e,WB,eff}$	$RMSE_s$ MOS	$RMSE_s$ $I_{e,WB,eff}$	σ $I_{e,WB,eff}$
Equation (9)	0.0990	8.3941	0.9236	0.1007	8.5057	0.9240
Equation (10)	0.0975	8.3552	0.9243	0.0990	8.4605	0.9248
Equation (11)	0.1183	9.1749	0.9080	0.1207	9.3145	0.9080

$$I_{e,WB,eff} = \left\{ ln\left(\frac{9 \times (I_{e,WB} + mlr \times grad^2)}{mbl^5 - mlr}\right) + mlr + I_{e,WB} + grad \times mlr \right\} \times 0.8303 + 8.9977 \quad (10)$$

$$I_{e,WB,eff} = (log_{10}(log_{10}(log_2(I_{e,WB} - 2 \times mbl) + mlr))) \times 321.7017 + 95.3708 \quad (11)$$

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 plotted in Figure 2. It shows the highest utility of $I_{e,WB}$ and mlr appearing 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 was used 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 [29] [24], who note that PESQ does not model the effect of burstiness on the quality. Similar frequencies were also observed in [26].

5.3 Comparison with the E-Model

Finally, a comparison of equation (10) was made with the E-Model's formulation of the $I_{e,WB,eff}$, as in [22]. This is represented by:

$$I_{e,WB,eff} = I_{e,WB} + (129 - I_{e,WB}) \times \frac{P_{pl}}{P_{pl} + Bpl} \quad (12)$$

The equation is similar to equation (5) differing in the constant term, 95, which is replaced with the new R_{max}=129.

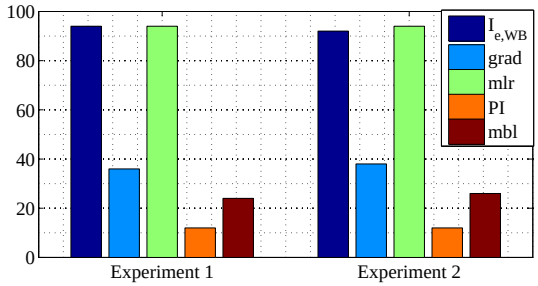

Figure 2: Percentage of the best individuals employing various input parameters in acceptable runs of each of the two experiments.

The $BurstR$ parameter is also absent here. Bpl values for this equation were computed separately for each of the codecs over 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 WB-PESQ estimates are oblivious of the effect of burstiness varying PIs. The results are reported in Table 4 for each codec. The table also shows the RMSE of equation (10) 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 (13)

$$\%PG = \frac{RMSE_e - RMSE_p}{RMSE_e} \times 100 \quad (13)$$

where, $RMSE_e$ and $RMSE_p$ represent the $RMSE$ of equations (12) and (10) respectively.

To evaluate the significance of difference between equations (12) and (10) the results were treated to a Mann-

Whitney-Wilcoxon test at $p = 0.05$. A value of 1 confirms a significant difference, whereas 0 represents otherwise. The results show that for all but for AMR-NB (7.4 kbps) equation (10) is significantly superior to E-Model.

6. 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. three functional models. 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 (10), which has the best performance among the proposed models, with the E-Model, equation (12), 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.

7. REFERENCES

[1] 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.

[2] 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.

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

[4] S. Gustafson, E. K. Burke, and N. Krasnogor. On improving genetic programming for symbolic regression. In D. C. et. al., editor, *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 912–919, Edinburgh, UK, 2-5Sept. 2005. IEEE Press.

[5] ITU-T. *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.

[6] ITU-T. *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.

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

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

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

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

[11] ITU-T. *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.

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

[13] ITU-T. *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.

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

[15] 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.

[16] 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.

[17] M. Keijzer. Improving symbolic regression with interval arithmetic and linear scaling. In C. Ryan, T. Soule, M. Keijzer, E. Tsang, R. Poli, and E. Costa, editors, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 70–82, Essex, 14-16 Apr. 2003. Springer-Verlag.

[18] M. Keijzer. Scaled symbolic regression. *Genetic Programming and Evolvable Machines*, 5(3):259–269, September 2004.

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

[20] Lingfen and E. C. Ifeachor. perceived speech quality prediction for voice over ip-based networks. In *IEEE International Conference on Communications (ICC)*, volume 4, pages 2573–2577, 2002.

[21] S. Luke and L. Panait. Lexicographic parsimony pressure. In W. B. L. et. al., editor, *GECCO 2002: Proceedings of the Genetic and Evolutionary*

Table 4: Comparison between the Prediction Accuracies of the E-Model and the Proposed Model

Codec (kbps)	E-Model			Equation (10)		Significance Test	
	Bpl	RMSE train	RMSE test	RMSE train	RMSE test	train	test
G.722.1 (24)	20.32	8.6824	8.8958	8.1701	8.9118	0	0
G.722.2 (6.6)	40.75	9.6225	8.9933	8.0938	7.6603	1	1
G.722.2 (8.85)	28.74	10.0175	9.9919	8.0185	7.8304	1	1
G.722.2 (12.65)	21.58	10.5538	10.4088	8.2188	8.0678	1	1
G.722.2 (14.25)	21.03	10.4684	11.2854	8.3031	8.5836	1	1
G.722.2 (15.85)	19.98	10.599	11.5020	8.3257	9.1166	1	1
G.722.2 (18.25)	19.48	11.2017	10.92	8.6862	9.0266	1	1
G.722.2 (19.85)	18.86	10.5502	11.3529	8.2338	8.7685	1	1
G.722.2 (23.05)	18.44	11.4079	11.1663	9.1417	8.7729	1	1
G.722.2 (23.85)	17.92	10.789	11.1948	8.6125	9.3168	1	1
G.729 (8)	28.43	8.95	9.1631	7.3888	7.4943	1	1
G.723.1 (6.3)	29.19	10.83	10.3630	8.8116	8.5259	1	1
AMR-NB (12.2)	13.50	8.0689	7.2947	9.4549	8.7322	1	0
G.722.1 (32)	18.93	8.9112	–	–	8.4775	–	0
AMR-NB (7.4)	15.71	7.1335	–	–	8.6188	–	1
Average	–	9.8527	10.1946	8.42	8.5269	–	–
% PG	–	–	–	14.54	16.36	–	–

Computation Conference, pages 829–836, New York, 2002.

[22] 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*, 16(6):1969–1976, November 2006.

[23] 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.

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

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

[26] 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*, volume 4, pages 2573–2577, 2006.

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

[28] 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.

[29] 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.

[30] L. Sun and E. C. Ifeachor. Voice quality prediction models and their application in VoIP networks. *IEEE Transactions on Multimedia*, 8(4):809– 820, August 2006.