Voicing-Aware Parametric Speech Quality Models over VoIP networks

Sofiene JELASSI, Habib YOUSSEF
Research Unit PRINCE, ISITCom
Hammam Sousse, Tunisia
Sofiene.Jelassi@infcom.rnu.tn, Habib.Youssef@fsm.rnu.tn

Christian HOENE
RI, Universität Tübingen
Tübingen, Germany
hoene@uni-tuebingen.de

Guy PUJOLLE
University of Pierre et Marrie Curie, Paris, France
Guy.Pujolle@lip6.fr

***Abstract* -- This paper describes novel parametric speech quality models which subsume the effect of packet loss distribution and voicing feature of missing signal waves. Speech quality estimate models for voiced and unvoiced loss location patterns are developed following multiple statistical regression analysis of measurements gathered from a built speech quality assessment framework. The overall speech quality is estimated by combining voiced and unvoiced speech quality estimate scores using an expression calibrated using a large number of speech samples. The input parameters namely, mean loss durations and ratios for voiced and unvoiced packets, of speech quality estimate models are extracted at run-time using a new voicing-aware packet loss Markov model. This chain, calibrated at run-time, finely models bursty packet loss behavior over voiced and unvoiced missing speech waves. Performance evaluation study shows that our voicing-aware speech quality estimate models clearly outperform voicing-agnostic speech quality models in terms of accuracy over a wide range of conditions.**

Keywords: Perceptual evaluation of voice quality, Voicing feature importance, Packet loss modeling.

I. INTRODUCTION

The accurate assessment of QoE (Quality of Experience) of VoIP service is pivotal from customers' and telecom operators' perspectives [1]. In fact, telecom operators desire to integrate Perceived Quality of Service (PQoS) in management, monitoring, and diagnosis operations of transport systems. Moreover, PQoS can be used by network planning architects to avoid undesirable configurations and optimize expected users' QoE. On the other hand, customers can use PQoS to select preferred access network under a given circumstance. Indeed, services over next-generation networks (NGN) will be likely offered to users using a multitude of networks and access technologies [2]. In such a case, customers can select the configuration that responds to their preferences in terms of quality and price for a given condition.

To accommodate such goals, objective assessment algorithms, which *automatically* estimate the PQoS without the involvement of human subjects, are needed. In particular, for the sake of managing VoIP services, single-ended packet-layer parametric model algorithms for speech quality estimate are highly desirable due to their attractive features. In fact, this category of speech assessment algorithms estimates PQoS of live VoIP conversations without intrusion at run-time. The parametric feature of such speech quality measurement algorithm enables to quantify PQoS of VoIP conversations using a set of gathered measures form the header content of voice packet stream [1, 2, 3]. This particular property enables the deployment of such a vocal quality measurement tool in intermediate- or terminal- node. Moreover, such measurement tools are often characterized by a reduced complexity [1].

Two potential proposals for standardization as a single-ended parametric model assessment algorithm of PQoS of VoIP conversations are under revision and evaluation [4, 5]. The first speech quality estimate algorithm of VoIP service, referred to as VQmon (Voice Quality Monitoring), separately calculates the PQoS over high and low loss periods. It accounts for perceptual effect experienced by users at the transition between high and low loss periods as well as recency effect in the computation of the overall perceived quality [4]. The quality model parameters such as high and low loss durations and densities are extracted from a four state Markov chain calibrated at run-time. This chain accurately enables to capture the pertinent features of packet loss behavior sustained by VoIP end-users. The *conversational perceive quality* is obtained using the additive effect of impairment factors assumed by ITU-T E-Model [6]. The second speech quality estimate algorithm, denoted as PsyVoIP, updates a set of "base" parameters at the reception of each new voice packet such as mean packet loss rate, packet delay variation, etc. The "base" parameters are transformed then combined using tailored formulas and weighting coefficients specific to each edge-device [5]. In reality, PsyVoIP developers show empirically that perceived quality significantly differs among edge-devices under identical network impairment conditions. As such, using *generic* formulas and weightings are unsuitable to tightly measure speech quality, for given network condition. As we can deduce, VQmon and PsyVoIP examine only the header content of received packets and ignore at all the payload features of carried signal wave. As such, both speech quality assessment tools assume that missing speech segments have an equal perceptual importance which is misleading [7]. The main contributions of this paper can be summarized as follows:

(1) The development of new parametric voicing-aware speech quality models, using a sophisticated assessment framework and following a multiple regression analysis, which account for both the packet loss location pattern and voicing feature of dropped voice fragments.

(2) The design of a new non-linear combination rule, calibrated using a large number of speech samples and conditions, in order to quantify in a non-intrusive way the perceptual effect of dropped voiced and unvoiced sounds simultaneously.

(3) The proposal of a new efficient sender-based strategy used in order to inform the receiver about the voicing feature of sent packets.

(4) The design of a novel Markov model, which properly accounts for voicing feature of lost packets. The conceived loss model, which is calibrated at run-time using a computationally efficient algorithm, is employed to extract pertinent information about the mean loss durations for voiced and unvoiced packets, mean loss ratios for voiced and unvoiced packets, etc.

The performance evaluation study shows that our voicing-aware speech quality models outperform voicing-agnostic speech quality models in terms of correlation and accuracy over a wide range of conditions. Indeed, we found that our parametric models achieve an excellent correlation of as much as 0.95 coupled with a mean accuracy in the order of 0.2 for ITU-T G.729 and G.711 equipped with a PLC (Packet Loss Concealment) speech CODECs.

The remainder of this paper is organized as follows. Section 2 illustrates the importance of voicing feature in speech quality modeling and evaluation. Section 3 describes the vocal quality assessment framework used to develop and validate voicing-aware speech quality models. Section 4 presents how speech sounds are stratified according to their voicing property and describe the methodology used to develop voicing-aware speech models. In Section 5, we introduce a new voicing-aware packet loss model and present an efficient algorithm used to extract pertinent parameters. In Section 6, we compare the performance of voicing aware and agnostic speech quality models against the intrusive ITU-T PESQ algorithm. We conclude in Section 7.

II. IMPORTANCE OF VOICING FEATURE ON VOCAL QUALITY EVALUATION

There is a consensus about the requirement to subsume the features of missing speech wave parts in the estimation of PQoS of VoIP conversations [4, 7, 8]. This certainly improves the correlation and precision between human-based and machine-objective scores. In reality, speech waves can be classified into voiced sounds such as ‘a’ and ‘o’, unvoiced sounds such as ‘h’ and ‘sh’, and silence. Obviously, loss impairments which affect silence parts are negligible [7]. Further, it has been empirically seen that removed voiced sounds impair more severely the perceived quality than dropped unvoiced ones [7]. This observation is supported by Figure 1, which shows the effect of erasing either voiced or unvoiced speech fragments on the listening perceptual quality. The curves of Figure 1 are obtained through the application of a voicing-aware bursty packet loss process, which means that voice frames are dropped selectively according to their voicing feature. The *full-reference* standard speech quality estimate

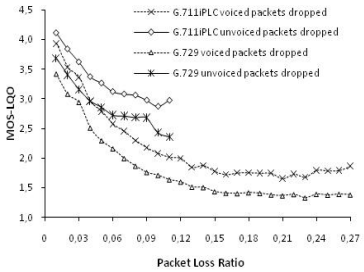


Figure 1: Importance of the voicing feature of dropped 20ms-speech segments on perceived quality for CODEC G.729 and G.711iPLC.

algorithm PESQ (Perceptual Evaluation of Speech Quality) defined in ITU-T Rec. P.862 has been used to quantify the listening perceptual quality. The output of ITU-T PESQ speech assessment algorithm is termed as MOS-LQO, which varies between 1(Poor Quality) and 4.5 (Excellent Quality). Several studies reported in literature showed that ITU-T PESQ algorithm accurately models human behavior rating under a wide range of impairments [1, 7]. Further details regarding performed empirical trials will be given later in Section 4. As we can notice, the voicing feature of removed frames greatly influences the perceptual quality regardless of the speech CODEC, G.711iPLC[1] or G.729. The deleted unvoiced frames impair much less the perceived quality than removed voiced ones. Notice that loss process likely affect voiced segments since they are statistically more frequent than unvoiced ones. Besides the effect of voicing feature on speech quality, the loss distribution, i.e., location and duration notably influence the rendered quality. In fact, it has been seen for certain model-based CODECs such as G.729 that losing a voiced frame preceded by an unvoiced one more severely harms the overall perceived quality [9]. Moreover, L. Sun et al. showed that a removed voiced frame located at the start rather than at the middle or the end of voiced sounds causes much more perpetual quality degradation [10]. That is why, speech property-based priority marking and recovering schemes of have been reported in the literature [7, 9]. This likely avoids losing perceptually important voice packets and hence improve overall perceptual quality [7].

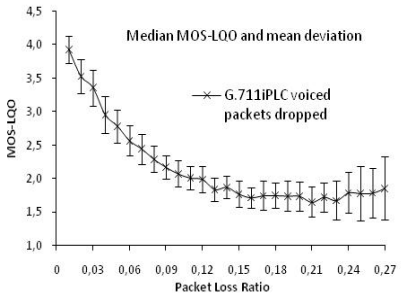


Figure 2 : Devation of perceived quality for a given PLR among sixteen voice samples using G.711PLC speech CODEC.

[1] G.711iPLC refers to G.711 speech CODEC equipped with the standard packet loss concealment algorithm defined in Appendix I of Rec. G.711

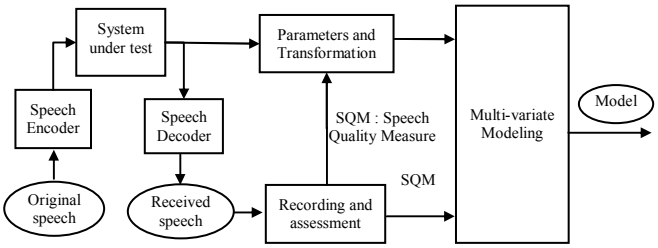


Figure 3: Speech Qauality Assessment (SQA) framework for the development of parametric speech quality estimate models.

L. Ding et al. modeled the speech quality degradation caused by packet losses located at voiced and unvoiced speech signal parts [11]. The authors followed a 3-order polynomial regression analysis to develop separate parametric speech quality estimate models for missing voiced and unvoiced packets. Their models have the mean packet loss ratio solely as an input parameter. They compute the overall perceived quality at the end of an assessment period based on a simplified weighted linear combination of scores calculated using voiced and unvoiced speech quality estimate models. In our opinion, their intuitive models are unable to accurately capture the effect of bursty loss behavior. In fact, using only the mean packet loss ratio to predict the perceived quality leads likely to a wrong estimate of speech perceptual quality over bursty lossy channel [4]. The curve plotted in Figure 2 illustrates the significant deviation relative to the mean score for a given mean voiced-packet loss ratio over sixteen speech samples. To avoid such an imprecision, speech quality estimate models should consider loss location and duration pattern. Moreover, their simplified linear combination rule used to derive the overall perceived quality over voiced and unvoiced missing sounds can be greatly improved to mimic users' behavior rating. Further, the receiver-based technique proposed by authors to detect the voicing feature of missing packets introduces additional processing overhead with a high risk of wrong decisions, especially over a burst of packet loss. In our opinion, even with the additional consumed bandwidth, we believe that a sender-based strategy is more efficient.

III.FRAMEWORK FOR VOCAL QUALITY MODELING

The study of PQoS of speech needs to build suitable Speech Quality Assessment (SQA) frameworks. Several methodologies can be followed to set-up a SQA framework. Figure 3 outlines the basic components of a SQA framework which aims at the development of parametric speech quality estimate models. Basically, a large set of reference speech sequences, having specific properties such as sampling rate, sample precision, content, and duration, are delivered across a *system under test* which involves several sources of impairments such as packet losses, delay, delay jitter, echoes, side-tone, noises, etc. The system under test output is monitored to generate the degraded speech sequences. The corresponding relevant parameters of the system under test are properly measured and recorded. The quality of degraded speech sequences is either measured using human-based subjective trials or a machine-executable objective algorithm. The large scale subjective testing is beyond any reasonable allocated time and budget. This explains why subjective trials are often confined to powerful Telecom operators, standardization telecom institutes such as ITU and ETSI, and worldwide corporations specialized in PQoS. In academia, SQA algorithms, which are deemed pretty accurate are preferred and used [1]. Typically, the ITU-T signal-layer full-reference SQA algorithm PESQ is used to measure speech quality. In this work, we assume that the system under test only introduces bursty packet losses, which are generated according to a Markovian model, described later in Section 4. As illustrated in Figure 3, the potential set of parameters, that likely affects the perceived quality, is directly sampled from the system under test such as mean loss ratios for voiced and unvoiced sounds, maximal voiced and unvoiced burst durations, and the set of inter-loss gap and loss durations. Often, "base" measurements require to be transformed to increase their correlation with obtained speech quality scores. To do that, all characteristic parameters of the system under test are monitored. For certain parameters a single value is returned, e.g., PLR (Packet Loss Ratio), CLP (Conditional Loss Probability), and maximal burst duration. For other parameters several values are recorded, e.g., inter-loss gap and burst duration. For each parameter, we determine, using regression, the degree and weighting coefficients of the polynomial that maximizes the *correlation* between the measured parameter values and the corresponding MOS-LQO scores obtained using ITU-T PESQ. For multi-value parameters, we calculate at first stage the L_P-norm as follows:

$$L_p(X(k)) = \left[\frac{1}{N} \sum_{k=1}^{N} (X(k))^P \right]^{1/P} \qquad (1)$$

where, X(k) is the k^{th} measure of the parameter X and N is the number of measures. Note that L_p-norm has been classically used to model the non-linearity of human hearing system [1]. In fact, L_p-norm highlights the effect of parameter variation on perceived quality. In this work, the value of p-norm is varied in the set {1/10, 1/9... 1/2, 1, 2 … 8, 9}. The correlation factor is calculated as follows:

$$r = \frac{\sum_{i=1}^{M} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{M} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{M} (y_i - \bar{y})^2}} \qquad (2)$$

where, r is the correlation coefficient between two cardinal-equal sets, x_i is a known value of the measured quality scores, y_i is the corresponding value of the examined parameter, $\bar{x}$ and $\bar{y}$ are, respectively, the mean value of the two analyzed sets, and M represents the cardinality of handled sets. Algorithm 1 summarizes the transformation process applied to the analyzed parameters. Once the optimal-correlated transformations for all potential parameters are obtained, the overall speech quality estimate model, which combines potential transformed parameters, is derived following a

multiple linear regression analysis (see Figure 3) [12]. To do that, a factor selection technique should be followed to pick-up suitable parameters. This enables to develop stable speech quality estimate models. It is obvious to select transformed parameters which exhibit a strong correlation with measured subjective scores. In this work, we follow the *forward parameter selection* methodology which can be summarized as follows: initially, select the parameter which achieves the best correlation with the set of known values of measured quality, then, iteratively, select the most correlated parameters with the set of residual values of the measured quality after eliminating the effect of previously selected parameters [12]. This process is halted when the returned t-student value (test of significance) of the correlation coefficient between the examined parameter and residual subjective scores becomes pretty low (< 3). Finally, Notice that multicollinearty among potential independent parameters should be avoided.

In next section, we adopt the described strategy in order to develop parametric speech quality estimate models for VoIP conversations. The conceived speech quality estimate models consider both the voice frame feature and loss location pattern.

Algorithm 1: Determination of optimal polynomial regression models for each potential parameters

OV: matrix which contains the original values of analyzed parameters.
LPV: matrix which contains L_p-norm values of analyzed parameters.
MOS: array which contains the MOS value of each (speech sequence, condition) pair.
RC: matrix which contains polynomial regression coefficients of analyzed parameters.
RCO: matrix which contains optimal regression coefficients of analyzed parameters.
OP: matrix which contains for each parameters optimal polynomial degree and p-norm.
CM: matrix which stores correlation for each analyzed parameters.

1: **for each** par belongs to the set of potential parameters **do**
// Vary the polynomial degree from 1 to 6
2: **for** m_i from 1 to 6 **do**
// Vary the norm from p_1 (= 1/9) to p_N (= 9)
3: **for** p_j from p_1 to p_N **do**
// Compute and record L_p-norm of each analyzed parameter
4: LPV[par] = Lp-norm (OV[par], p_j)
// Apply regression process of degree m_i
5: RC[par] = polynomial-regression (LPV[par], MOS, m_i)
// Measure the correlation between estimated and measured scores
6: CM[m_i, p_j] = correlation(regress(LPV[par], RC[par]), MOS)
// Update the regression model if correlation is higher than previously founded
7: **if** MC[m_i, p_j] > r_{max} **then**
8: OP[par] = {m_i, p_j}; r_{max} = CM[m_i, p_j];
RCO[par] = RC[par] ;
9: **end if**
10: **end for**
11: **end for**
12: **end for**

IV. SPEECH QUALITY MODELS FOR DROPPED VOICED AND UNVOICED FRAMES

The development of voicing-aware speech quality models needs to stratify speech signals into fragments according to their voicing feature. In this work, we use the simple, yet efficient sender-based algorithm SUVING to discriminate between speech wave segments [13]. In short, SUVING uses zero-crossings coupled with short-term energy to identify the type of each examined fragment [13]. The zero-crossing metric represents the number of times in a speech fragment where the amplitude of sound wave changes its sign. For instance, for a 10 ms clean voice segment, the zero-crossing rate is roughly equal to 12 for voiced speech and 50 for unvoiced speech [13]. SUVING calculates short-term energy as follows:

$$E_n = \sum_{m=n-N+1}^{n} (x(m)w(n-m))^2 \qquad (3)$$

where, x(m) corresponds to the energy of the m^{th} sample, w is a hamming window of size N samples and centered between the $(n-N+1)^{th}$ and n^{th} sample. The short-term energy is higher for voiced than unvoiced speech, and should be equal to zero for silence regions in clean speech signal recordings. However, unavoidable background noise, which is characterized by high zero-crossing rate and low short-term energy, induces inaccuracy in S/V/U discrimination process which is properly treated by SUVING using a set of thresholds and additional rules [13].

A classical Gilbert/Elliot Markov model (see Figure 4) has been used to mimic packet loss behavior experienced by users over a bursty lossy channel [9]. As illustrated in Figure 4, a Gilbert/Elliot model has 2 states, NON-LOSS and LOSS which represent respectively a successful and failed voice packet delivering. The mean sojourn durations under states NON-LOSS and LOSS are, respectively, equal to $1/p$ and $1/q$ where p and q are the transition probabilities from NON-LOSS to LOSS state, and conversely. The model is calibrated using ULP (Unconditional Loss Probability), CLP (Conditional Loss Probability), and EBP (Effective Burstiness Probability) which are calculated as follows:

$$ULP = \frac{p}{p+q} \qquad CLP = 1-q \qquad EBP = ULP \times CLP \qquad (4)$$

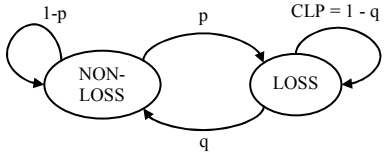


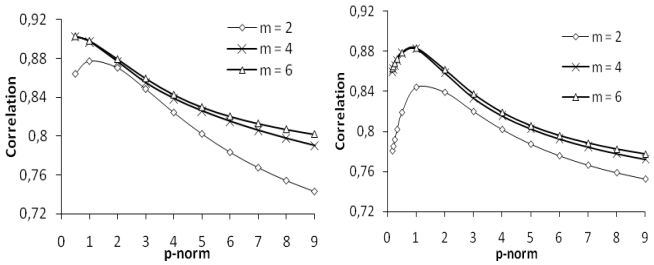
Figure 4: Gilbert/Elliot Markov loss model.

The EBP metric, which has been initially defined by F. Hammer et al., is used to introduce packet loss burstiness in accurate way over a short period of time (8-20 sec) [14]. The developed Gilbert/Elliot model, which mimics the distortion introduced by the system under test (see Figure 3), has as input parameters ULP and EBP which have been finely varied to produce controlled conditions. Table I summarizes the tested conditions. In these empirical trials, we use a total of 16 speech sequences, pronounced by eight male and eight female English speakers. For each speech sample, we generate voicing-aware packet loss pattern according to S/V/U pattern of each speech sequence and voicing-agnostic packet loss pattern produced by the Gilbert/Elliot model. In reality, classical Gilbert/Elliot model produces a packet loss pattern which does not account at all for voicing feature of missing packets. To enable the generation of voicing-aware packet loss pattern, presumed lost packets are *ignored* when a loss event affects unsuited voicing packets. The degraded version of original sample is produced and evaluated using ITU-T PESQ algorithm. In addition, the effective ULP, EBP, maximum burst duration, and the sets of inter-loss packet gaps and loss durations are properly recorded for each (speech sequence, condition) pair. The total number of evaluated samples and conditions is equal to 1536.

TABLE I: Experimental conditions for packet loss behavior using Gilbert/Elliot Model

Parameters	Conditions	Instances
CODEC	G.711iPLC, G.729	2
Mean Packet loss ratio (ULP)	1, 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 45 %	12
Ratio of burstiness, R (EBP = ULP/R)	2, 4, 6, 8	4
Total number of combinations	2×12×4	96

At the end of the empirical trials, we statistically analyse using Algorithm 1 the obtained results. Curves plotted in Figure 5a and 5b graphically illustrate the result of application of Algorithm 1 for the inter-loss gap duration metric. As we can note, the perceived quality is optimized for a specific combination p-norm and polynomial degree, m, which is recorded and used subsequently in the application of the multiple regression process.



(a) G.711 voiced ignored (b) G.711 unvoiced ignored

Figure 5: Application of Algorithm 1 on inter-loss metric.

TABLE II: Best correlation between measured transformed parameters and measured speech quality scores of G.711iPLC

	voiced			unvoiced		
	m	**p**	**r**	**m**	**p**	**r**
ULP	2	-	0.95	2	-	0.90
EBP	5	-	0.78	2	-	0.63
MaxBD	4	-	0.49	2	-	0.54
$L_p(\{\text{inter-loss}\})$	3	0.50	0.90	3	1	0.88
$L_p(\{\text{loss}\})$	5	0.25	0.86	4	0.16	0.90

Table II summarizes the individual parameter settings which achieve the best correlation factor with measured quality scores. As described in Section 3, parameters forward selection methodology coupled with a multiple linear regression analysis are used to obtain adequate quality models for missing voiced and unvoiced sounds. The following speech quality models for G.711iPLC and G.729 speech CODEC has been obtained with an overall correlation factors above 0.85:

$$\begin{cases} MOS_V(ULP,\ \{loss\}) = 0.80 \times f_2(ULP) + 0.23 \times f_5\left(L_{1/4}(\{loss\})\right) \\ MOS_U(ULP,\ \{loss\}) = 0.48 \times f_2(ULP) + 0.52 \times f_4\left(L_{1/8}(\{loss\})\right) \end{cases} \text{If CODEC} = \text{G.711iPLC} \quad (5)$$

$$\begin{cases} MOS_V(ULP,\ \{loss\}) = 0.74 \times f_3(ULP) + 0.25 \times f_4\left(L_{1/9}(\{loss\})\right) \\ MOS_U(ULP,\ \{loss\}) = 0.31 \times f_2(ULP) + 0.68 \times f_1\left(L_{1/9}(\{loss\})\right) \end{cases} \text{If CODEC} = \text{G.729} \quad (6)$$

where, f_m refers to the polynomial of degree m that maximizes the correlation between measured speech quality and examined parameter. After modeling the effect of loss process that only affects voiced or unvoiced frames, now it is required to develop a speech quality model which quantifies the effect of voicing-agnostic packet loss process that affects both voiced and unvoiced packets. To this end, we generate a packet loss location pattern using Gilbert/Elliot model which affects indifferently voiced and unvoiced packets. Using generated voicing-agnostic packet loss location and speech sequence S/V/U patterns, we create voiced and unvoiced loss location patterns which affect either voiced or unvoiced frames. The speech samples used in previous trials are *separately* impaired using voicing-agnostic and the resulting voiced and unvoiced packet loss location patterns. Hence, for each (speech sequence, condition) pair, three quality scores are computed that quantify perceived quality of impaired speech sequences under voiced-unvoiced, voiced, and unvoiced loss patterns. The overall speech quality model, which captures the effect of removed voiced and unvoiced frames, is obtained using multiple regression analysis. The model predictors are speech quality scores calculated after the application of voiced and unvoiced loss location patterns, separately. Specifically, the overall quality speech quality estimate models have the following form:

$$MOS_{UV} = w_{v1} \times MOS_V + w_{u1} \times MOS_U + w_{v2} \times MOS_v^2 + w_{u2} \times MOS_u^2 + w_{uv} \times MOS_U \times MOS_V \quad (7)$$

where, w_{v1}, w_{u1}, w_{v2}, w_{u2}, and w_{uv} are the fitting coefficients which are obtained based on the minimisation of RMSE technique. The value of coefficients that achieves the best correlation is summarized in Table III. Notice that at run-time the values of MOS_U and MOS_V are calculated based on Equation (5) and (6).

Table III: Regression coefficients of UV models.

Coefficients	G.711iPLC	G.729
w_{v1}	0.825	-0.243
w_{u1}	0.044	0.591
w_{v2}	-0.028	-0.023
w_{u2}	-0.025	-0.177
w_{uv}	0.075	0.375
correlation	0.998	0.999

As we can see, the developed speech quality models require vital meta-data about voicing feature of missing packets. To do that accurately, a sender-based notification scheme can be adopted. Precisely, this is performed by piggybacking meta-data about voicing feature of recent sent voice packets to the receiver entity. There exists an important trade-off between additional consumed bandwidth and timeliness. For instance, if voicing feature of each voice packet is coded using one bit, where 0 indicates an unvoiced packet and 1 indicates a voiced packet, then ([T/F]/8) additional bytes are required to notify the receiver about the voicing pattern of [T/F] previous F-sec voice packets, where T represents the monitoring period duration. The value of T can be fixed in advance or adjusted dynamically according to packet loss behavior or talk-spurt duration. To reduce consumed bandwidth, we propose piggybacking voicing meta-data solely at the detection of a transition from a voiced to an unvoiced sound, and conversely. Notice that modern speech CODECs such as G.729, G.726, and iLBC generate a very small and fixed payload size of as much as 20 Bytes to encode 20 ms of speech waves. As such, in our opinion, additional bytes piggybacked occasionally are not considered as a critical overhead. The receiver entity can implicitly identify data packets which contain meta-data voicing information by checking the packet length.

V.VOICING AWARE PACKET LOSS BEHAVIOR MODEL

To efficiently extract required voicing-aware measures for speech quality estimation, we propose using a novel model of packet loss behavior which accounts for voicing feature of missing fragments. The developed model constitutes a relevant extension to classical Gilbert/Elliot model (see Figure 6). It accurately enables characterizing loss behavior over voiced and unvoiced frames. The conceived model has three states, NON-LOSS, $LOSS_{voiced}$, and $LOSS_{unvoiced}$ which represent, respectively, the successful reception of a voice packet and failed delivering of a voiced and unvoiced voice packet.

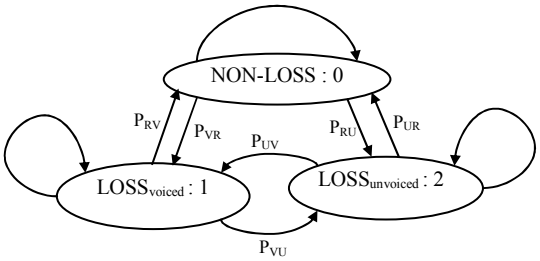


Figure 6: Voicing feature aware packet loss model.

The packet loss model illustrated in Figure 6 is calibrated at run-time according to the flow of received / lost packets and voiced / unvoiced pattern piggybacked with data packet stream. An efficient voicing-aware loss driven algorithm is used to update at run-time a set of counters which are used at the end of a monitoring period to calculate the transition probabilities. Hence, parameters such as mean packet loss ratio for voiced and unvoiced sounds and mean burst durations for voiced and unvoiced sounds can be formally computed. During the voicing-aware monitoring period, the set of inter-loss gap and unvoiced and voiced packet loss durations are properly recorded.

Algorithm 2: Calibration and parameters estimation at run-time

```
1: if (new V/U packets is received) then
2: vu = read-vu-pattern(last-seq, cur-seq)
3: rcv = read-loss-pattern(last-seq, cur-seq)
4:   for i from last-seq to cur-seq do
5:       if (rcv[i] = "1") then // voice packet is received
6:          if (sate = "0") then
7:          c00++,  ac00++;
8:          elseif(state = "1") then
9:          if (ac11 > maxv) then maxv = ac11 end if
10:         record(ac11);  c10++, state = "0"; ac11 = 0;
11:         elseif(state = "2") then
12:         if (ac22 > maxu) then maxu = ac22 end if
13:         record(ac22);  c20++, state = "0"; ac22 = 0;
14:     else // voice packet is lost
15:         if (vu[i] = "V" and state = "0") then
16:         c01++, state = "1"; record(ac00); ac00 = 0; ac11 = 1;
17:         elseif (vu[i] = "V" and state = "2") then
18:         if (ac22 > maxu) then maxu = ac22 end if
19:         record(ac22);  c21++; state = "1"; ac22 = 0;  ac11 = 1;
20:         elseif (vu[i] = "V" and state = "1") then
21:         c11++;  ac11++;
22:         elseif (vu[i] = "U" and state = "0") then
23:         c02++, state = "2"; record(ac00); ac00 = 0; ac22 = 1;
24:         elseif (vu[i] = "U" and state = "1") then
25:         if (ac11 > maxv) then maxv = ac11; end if
26:         record(ac11);  c12++; state = "2"; ac11 = 0;  ac22 = 1;
27:        elseif (vu[i] = "U" and state = "2") then
28:        c22++; ac22++;
29:      end if
30:    end for
31: end if
```

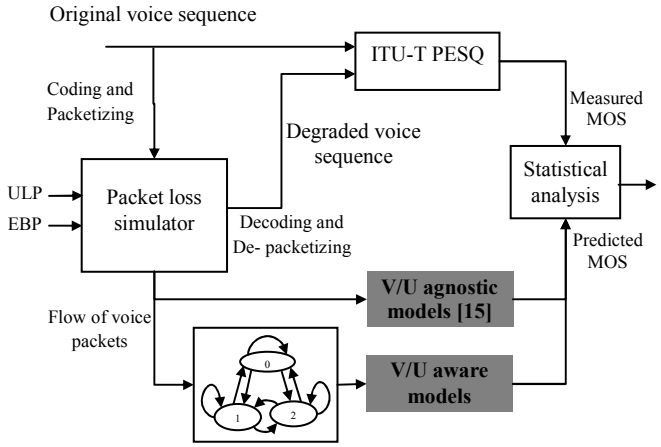


Figure 7: Evaluation framework of voicing-aware SQA models.

Algorithm 2 summarizes the calibration process of voicing-aware loss model and how suitable parameters are extracted and recorded. In Algorithm 2, state number 0, 1, and 2 represent respectively NON-LOSS, $LOSS_{voiced}$, and $LOSS_{unvoiced}$ states. Algorithm 2 uses a set of counters denoted as c_{ij} where indexes i and j refer to the state number. Basically, the developed algorithm triggers the calibration process upon the reception of a new, in-sequence, voice packet including voicing feature of the recent sent voice packets. Algorithm 2 extracts V/U and loss patterns from the received packet and the history of lost packets (lines 2 and 3). The algorithm updates measurements from the last processed packet to the current one identified using their sequence numbers. Moreover, the algorithm determines the maximal voiced and unvoiced burst durations using the variables *maxv* and *maxu*, respectively. It keeps track of the inter-loss gap and voiced and unvoiced loss durations using variables ac_{00}, ac_{11}, and ac_{22}.

At the end of a monitoring period, the mean loss packet ratio, ULP, and degree of burstiness, EBP, for voiced and unvoiced packets can be computed as follows:

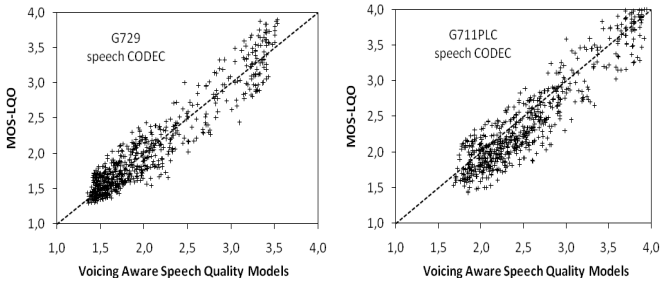
$$ULP_v = \frac{c_{01} + c_{11} + c_{21}}{nbt} \qquad ULP_u = \frac{c_{02} + c_{22} + c_{12}}{nbt} \tag{8}$$

$$EBP_v = ULP_v \frac{c_{11}}{c_{11} + c_{10} + c_{12}} \qquad EBP_u = ULP_u \frac{c_{22}}{c_{22} + c_{20} + c_{21}} \tag{9}$$

where, nbt refers to the total number of sent packets during the sensing period. Notice that under a continuous quality evaluation requirement, all variables, counters, and arrays are re-initialized at the start of a new monitoring period.

VI.PERFORMANCE EVALUATION

In order to evaluate the performance of our voicing-aware speech quality models, we set-up the vocal quality assessment framework depicted in Figure 7. The framework includes a bursty packet loss simulator which follows the Gilbert/Elliot model (see Figure 4). The reference and resulting degraded voice sequences are evaluated using ITU-T PESQ assessment algorithm. On the other hand, speech quality score is estimated



(a) (b)

Figure 8: Validation of voicing aware speech quality models.

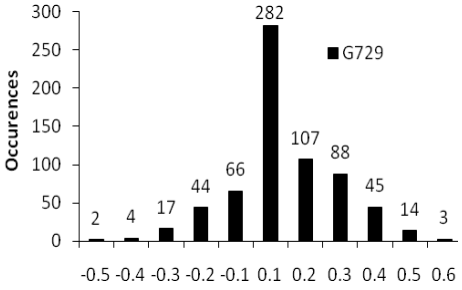
using voicing -agnostic and -aware speech quality estimate models. We compare our voicing-aware models against the voicing-agnostic speech quality estimate models reported in [15]. During these trials, a set of eight voice sequences which are pronounced by four male and four female English speakers are impaired and evaluated. The degree of burstiness is properly parameterized using ULP and EBP. Specifically, we varied the ULP value from 1% to 30% with an increase step of 3%. The value of EBP is calculated as a ratio of the ULP value which is varied from 2 to 8 with an increase step of 2.

Table IV compares the performance of voicing-aware speech quality estimate models and voicing-agnostic ones for G.711iPLC and G.729 in terms of correlation and precision. As we can note, voicing-aware speech quality estimate models achieve a correlation factor above 0.95 for both speech CODECs which is quite satisfactory. Moreover, our voicing-aware speech quality estimate models reduce notably, relative to voicing-agnostic speech quality estimate models, the mean deviation between measured MOS scores using ITU-T PESQ and estimated MOS scores using our models for both speech CODECs. The achieved accuracy is in the order of 0.2 which constitutes an excellent precision. The scatter-plots shown in Figure 8 prove the correlation and accuracy of developed speech quality estimate models.

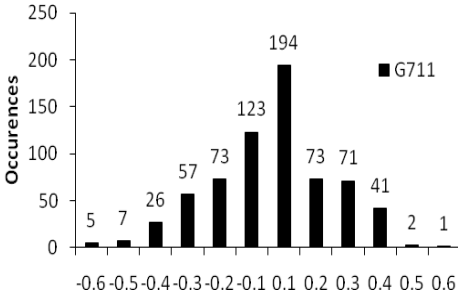
Table IV: Performance comparison between voicing aware and voicing agnostic models

	Voicing Agnostic Models [15]		**Voicing Aware Models**	
	G.711PLC	**G.729**	**G.711iPLC**	**G.729**
Correlation	0.927	0.910	0.954	0.961
Deviation	0.61	0.92	0.22	0.17

Histograms shown in Figures 9 illustrate the distribution of predicted MOS scores with respect to measured MOS scores for the G.729 and G.711 speech CODECs. These histograms prove the accuracy of our voicing-aware speech quality models to estimate MOS score. Indeed, 75% of estimated MOS score for G.729 and 70% for estimated MOS scores for G.711iPLC falls in the range [-0.2, 0.2] which is quite satisfactory given parametric, non-intrusive, and low complexity features of our developed speech quality models.



(a) : Speech CODEC G.729



(b) : Speech CODEC G.711iPLC

Figure 9: Distribution of deviation between MOS-LQO scores and voicing-aware model-based estimate models of speech quality.

VII.CONCLUSION

This paper extends conventional parametric speech quality estimate models by considering voicing feature of lost packets. A sophisticated speech quality assessment framework has been set-up to build voicing-aware speech quality models that enable to accurately quantify the effect of lost packets according to their voicing property. An overall speech quality estimate model was properly developed using multiple regression analysis to quantify the effect of dropped voiced and unvoiced fragments. The relevant parameters of speech quality models were efficiently calculated using a new voicing-aware loss model calibrated at run-time. The accuracy evaluation study proves that our voicing-aware speech quality models outperform voicing-agnostic speech quality models in terms of correlation and RMSE with objective scores.

REFERENCES

[1] A. Rix, J. Beerends, D. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality: Technology and Applications". IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 6, pp. 1890:1901, November 2006.

[2] U. Jain, Y. Yokoyama, and A. Kumar, "Study of Factors Influencing QoS in Next Generation Networks", [on-line], www.eng.auburn.edu/department/csse/classes/comp8700 /index.html, visited in April 2009.

[3] A. Takahashi, N. Egi, and A. Kurashima, "QoE Estimation Method for Interconnected VoIP Networks Employing Different Codecs", IEICE Transactions on Communication Vol. E90-B, No 12, December 2007.

[4] A.D. Clark, "Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality", In Proceedings of IP Telephony Workshop, Columbia, USA, 2001.

[5] S. R. Broom, "VoIP Quality Assessment: Taking Account of the Edge-Device". IEEE Transactions on Audio, Speech, And Language Processing, Vol. 14, No 6, pp.1977:1983, November 2006.

[6] ITU-T Recommendation G.107, "The E-Model a Computational Model for Use in Transmission Planning", March 2003.

[7] C. Hoene, "Internet telephony over wireless links", PhD dissertation, Technical University of Berlin, Germany, December 2005.

[8] M. Masuda, T. Hayashi, "Non-Intrusive Quality Monitoring Method of VoIP Speech Based on Network Performance Metrics", IEICE Transactions on Communication Vol. E89-B, No. 2, February 2006.

[9] H. Sanneck, "Packet Loss Recovery and Control for Voice Transmission over the Internet", PhD dissertation, Technical University of Berlin, Germany, December 2000.

[10] L. F. Sun, G. Wade, B. M. Lines, E. C. Ifeachor, "Impact of Packet Loss Location on Perceived Speech Quality", in Proceedings of 2nd IP-Telephony Workshop (IPTEL '01), Columbia University, New York, April 2001, pp.114-122.

[11] L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, R. A. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP", Elsevier Speech Communication Journal, No. 49, pp: 477–489, 2007.

[12] R. Jain, "The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling", Wiley- Interscience, New York, NY, April 1991, ISBN: 0471503361.

[13] M. Greenwood and A. Kinghorn, "SUVing: automatic silence / unvoiced / voiced classification of speech", Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK, 1999.

[14] F. Hammer, P. Reichl, and T. Ziegler, "Where packet traces meet speech samples: an instrumental approach to perceptual QoS evaluation of VoIP", in Proceedings of 12th International Workshop IWQoS, pp: 273-280 , Montreal, Canada, June 7-9, 2004.

[15] L. Sun and E. Ifeachor, "Voice Quality Prediction Models and Their Application in VoIP Networks", IEEE Transactions on Multimedia, Vol. 8, No. 4, pp 809:820, August 2006.