Single-Ended Parametric Voicing-Aware Models for Live Assessment of Packetized VoIP Conversations

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

The perceptual quality of VoIP conversations depends tightly on the pattern of packet losses, i.e., the distribution and duration of packet loss runs. The wider (resp. smaller) the inter-loss gap (resp. loss gap) duration, the lower is the quality degradation. Moreover, a set of speech sequences impaired using an identical packet loss pattern results in a different degree of perceptual quality degradation because dropped voice packets have unequal impact on the perceived quality. Therefore, we consider the *voicing feature* of speech wave included in lost packets in addition to packet loss pattern to estimate speech quality scores. We distinguish between voiced, unvoiced, and silence packets. This enables to achieve better correlation and accuracy between *human-based subjective* and *machine-calculated objective* scores.

This paper proposes novel *no-reference parametric* speech quality estimate models which account for the voicing feature of signal wave included in missing packets. Precisely, we develop separate speech quality estimate models, which capture the perceptual effect of removed voiced or unvoiced packets, using elaborated simple and multiple regression analyses. A new speech quality estimate model, which mixes voiced and unvoiced quality scores to compute the overall speech quality score at the end of an assessment interval, is developed following a rigorous multiple linear regression analysis. The input parameters of proposed voicing-aware speech quality estimate models, namely Packet Loss Ratio (PLR) and Effective Burstiness Probability (EBP), are extracted based on a novel *Markov model of voicing-aware packet loss* which captures properly the feature of packet loss process as well as the voicing property of speech wave included in lost packets. The conceived *voicing-aware packet loss model* is calibrated at run time using an efficient packet loss event driven algorithm. The performance evaluation study shows that our voicing-aware speech quality estimate models outperform voicing-unaware speech quality estimate models, especially in terms of accuracy over a wide range of conditions. Moreover, it validates the accuracy of the developed parametric no-reference speech quality models. In fact, we found that *predicted scores* using our speech quality models achieve an excellent correlation with *measured scores* (>0.95) and a small mean absolute deviation (<0.25) for ITU-T G.729 and G.711 speech CODECs.

Keywords: VoIP, perceptual evaluation of voice quality, voicing feature importance, packet loss modeling.

1. Introduction

Over the last few years, VoIP (Voice over IP) service has reached large popularity because of its attractive features for consumers and Telecom service providers. For consumers, the cheap and even free billing, the good perceptual quality, the mobility support, and the enriched vocal service capability constitute highly attractive features. For telecom operators, the management flexibility and handy service personalization and upgrading are highly desirable properties [1]. In fact, packetized VoIP service increasingly replaces and extends ordinary vocal telephone service in homes and enterprises [2]. To successfully integrate telephone service over IP infrastructure, customers should experience a good perceptual quality. However, ordinary *unmanaged* multi-service IP networks impair the flow of voice packets, which are often carried using the unreliable UDP transport protocol, by introducing delay, delay jitter, and packet loss, disorder, and duplication [3]. Several remedies have been reported in the literature to deal with such sources of impairments [3, 4]. Basically, there are two schools of thought to improve the perceptual quality of VoIP telephony, reactive and predictive strategies:

- *Reactive approaches*: They try to reduce introduced IP impairments through the well-engineering of adaptive applications at sender and receiver sides to account for service sensitivity to network delay and packet loss. This enables to smartly hide perceptual annoying effects caused by time-varying end-to-end bandwidth, packet loss, and one-way network delay, without the requirement to upgrade/alter the operational mode of existing network infrastructure. Actually, Skype and GoogleTalk represent two well-known distributed adaptive applications widely-used in the Internet to achieve *multimedia* and *vocal* telephony over IP and hybrid IP/PSTN networks, respectively [5]. The adaptive behavior of Skype and GoogleTalk at sender and receiver sides has been extensively studied and compared at perceptual level by B. Sat et al. [5].

- *Predictive approaches*: They reduce network IP impairments by smartly *managing* network resources to accommodate services according to their specific requirements. The telecom operators can define their proprietary management policy and network architecture to improve the quality of delivered delay-sensitive services. The intermediate nodes are equipped with suitable QoS mechanisms such as call admission, packet classification and scheduling, as well as preferential treatment of crossing streams [6, 7]. This requires upgrading the operational mode of intermediate nodes, which may be difficult in large, heterogeneous environments.

In practice, to perform VoIP conversations, application layer reactive approaches are usually used by default. If a predictive approach is presented in the transport network, then the intensity of network impairments will be significantly reduced or even removed, which greatly helps end-to-end reactive approaches to achieve a better perceptual quality. Notice the existence of some recent proposals which aim at improving the perceptual quality of VoIPoW (VoIP over wireless) using cross-layer optimisation strategy [4, 8, 9]. For instance, the source can dynamically adapt the packet duration according to the channel state, number of wireless hops, and prevailing access network delay [8]. Moreover, it can dynamically adjust the number of retransmission attempts and backoff delay at link-layer according to the perceptual importance of the outgoing voice packet [9].

In recent years, the performance of proposed adaptive behavior of VoIP application and network management policy is judged according to their achievable perceptual quality [3, 4, 5, 8, 9, 10]. Typically, the perceptual quality of an audio processing system is quantified in terms of MOS (Mean Opinion Score), which is a real number between 1 (bad quality) and 5 (excellent quality) [11]. Normally, the value of MOS score for a given configuration (application and network) is obtained using *subjective trials* [12]. Precisely, a set of human subjects, placed in a lab environment, are asked to vote either a set of heard impaired speech sequences, which is referred to as *listing quality* and termed as MOS-LQS, or a conversational task experience, which is referred to as *conversational quality* and termed as MOS-CQS. The ITU-T P.800 specification of subjective trials aims primarily at evaluating the perceptual effect of potential sources of impairments during vocal conversations over *circuit-switched telephone systems* such as loudness, side-tone, noise, echo, signal attenuation, acoustic features of edge devices, and one-way transit delay [12]. They have been subsequently adapted and extended by the research community to evaluate new sources of impairments experienced over

VoIP systems such as packet loss, low bitrate CODECs, and delay jitter [4, 13, 14]. Notice that the subjective approach, especially for large scale testing, is usually judged as time-consuming, expensive, and cumbersome [13, 14]. Moreover, it is unable to rate at run-time packet-based voice conversations in order to adapt the application and network behavior, accordingly. That is why, objective approaches, which estimate *automatically* the perceptual quality using *machine-executable* speech quality measurement (SQM) algorithms, are preferred and widely-used by telecom operators [11]. Extensive research effort within standardization bodies, academic institutions, and industry companies has improved the correlation and accuracy between objective and subjective scores to a satisfactory degree [11]. Machine-executable SQM algorithms can be classified into two categories:

- *Black box signal strategies*: They estimate the perceptual quality by analyzing speech waves without any knowledge about the features of transport systems. They can be classified as full-reference (or intrusive) approaches, which have as input the reference and degraded speech sequences, and no-reference (non-intrusive or single-ended) approaches, which only have as input the degraded speech sequence.
- *Glass box system parameter strategies*: They estimate the perceptual quality using a set of statistical measurements gathered from the network such as delay, delay jitter, echo, and packet loss ratio and features of edge-devices such as coding scheme, packet loss concealment algorithm, and de-jittering buffer strategy.

In practice, the glass box system parameter approaches are more preferable for VoIP conversations because they are able to efficiently predict at run-time speech quality scores using packet-layer statistical measurements. However, glass box system parameter approaches are relatively less accurate than black box signal approaches in the estimation of the perceptual quality scores. The development of a glass box system parameter assessment approach needs the development of suitable parametric quality models, which transform objective network and edge measurements to MOS domain. Normally, speech quality estimate models are derived following a regression analysis using a wide range of *subjective* speech quality empirical measurements [11]. However, the large number of conditions makes large scale subjective testing unreasonable in terms of cost and time. That is why, full-reference signal-layer objective approaches, which give a tight estimation of subjective scores, are used to measure the perceptual quality [13, 14, 15, 16].

Generally, the standard full-reference signal-layer ITU-T SQM algorithm, described in Rec. P.862 and denoted as PESQ (Perceptual Evaluation Speech Quality), is used to gather required SQM for parametric model development [17]. The produced score is termed as MOS-LQO (Mean Opinion Score – Listening Quality Objective).

A well-known glass box parametric speech quality model, denoted as E-Model, has been defined in the ITU-T G.107. E-Model has been conceived to predict speech quality over *telephone systems* [18]. The goal of E-Model was to give a general picture about the degree of satisfaction of a set of users for a given network configuration. The system characterization parameters are stratified into simultaneous, delay, and equipment impairment factors. For the sake of simplicity, the perceptual effect of impairment factors is assumed additive on psychological scale [18]. Notice that recent subjective experiences indicate that additive property of impairment factors can lead to inaccurate prediction of the conversational perceptual quality under several circumstances [19]. This constitutes the major reason of confining the utilisation of ITU-T E-Model for planning purposes only [19]. The values of parameters of E-Model are measured from the planned/existing configuration, then combined using a set of models to produce a rating factor, denoted as R and ranging from 0 to 100. Notice that the rating factor R can be transformed to MOS scale using standard functions [4]. As such, E-Model is unable to accurately evaluate at run-time the perceptual quality on call-by-call basis. Moreover, it is unable to evaluate a VoIP conversation given that input parameters over IP networks are time-varying. That is why the E-Model has been adapted and upgraded by several researchers to be able to predict on call-by-call basis the perceptual quality of VoIP conversations [15, 16, 19, 20, 21, 22]. Accordingly, extended E-Model can act as a *single-ended packet-layer parametric* SQM tool of VoIP conversations. To do that, impairment characterization parameters which are independent from transport network, such as room and circuit noises, and the acoustic features of edge devices are set to their default values. Moreover, several delay impairment models, which accept as input the experienced mean end-to-end delay, have been rigorously developed and extensively evaluated in the literature [20, 21, 22]. Furthermore, new equipment impairment models specific for VoIP conversations, which quantify the perceptual effect of *packet loss* and coding scheme, have been reported in the literature [14, 20, 21, 22]. Notice that equipment impairment factor, denoted as I_e, can be transformed to a MOS-LQO score using suitable functions [22]. Actually, new extensions of E-Model are in progress to consider new configurations and scenarios experienced by customers over next generation networks such as vertical and horizontal handover over last-hop wireless data networks, route changes over multi-hop wireless networks (MANETs), wideband and multiple description speech CODEC schemes, on-line switching of coding scheme and rate, features of loss process, etc.

It is well-recognized that packet loss over wide area IP networks is bursty and time-varying [21]. Thus, using mean packet loss ratio alone as a characterization parameter for quality prediction can lead to an inaccurate estimation of experienced quality. Recently, research work has been reported in the literature to accurately quantify the perceptual effect of time-varying bursty packet loss behavior. In [21], author estimates *separately* the perceptual quality at high and low packet loss periods and subsume the perceptual artefacts at transition between high and low loss periods as well as the temporal location of high loss period in the calculation of the overall equipment impairment factor. The parameters of developed equipment impairment model such as mean packet loss densities and durations for high and low loss periods are extracted from a four state packet loss Markov model which efficiently and finely captures the global features of packet loss process. The conversational speech quality is calculated using the additive effect of impairment factors adopted by ITU-T E-Model. In [23], the author describes a SQM tool which calculates a set of “base” parameters at the reception of each new voice packet such as packet loss ratio, packet delay variation, mean burst duration, maximal burst duration, etc. Each “base” parameter is transformed by a non-linear function to subsume network impairment factor that influences the perceptual quality in a non-linear way. The transformation functions and weighting coefficients are adapted for each edge-device and CODEC used. The calibration of speech quality estimation models is performed through a large scale training process, which covers a wide range of conditions evaluated using ITU-T PESQ algorithm. In [14], authors proposed new speech quality estimation models that account for the bursty nature of packet loss process over IP networks. To do that, speech quality regression models, which accept as parameters inter-loss and loss durations, are developed and validated for several CODECs. At run-time, the perceptual quality is estimated for each (inter-loss, loss) pairs, then linearly combined at the end of an assessment period to produce the overall perceptual quality.

The goal of the previously described single-ended packet-layer SQM algorithm was to accurately evaluate the speech quality by properly capturing the bursty nature of packet loss process over IP networks. They solely rely on information included in the standard header content of received packet stream and do not account for the features of payload content. As such, they assume that conveyed voice packets have an equal impact on perceptual quality. However, it has been clearly shown in the literature that voice packets have different effect on the perceptual quality according to their temporal location and the content features [4, 24]. This can result in an inaccurate estimation of listening perceptual quality, especially when the evaluation process is performed on a sequence-by-sequence basis. Hence, for the sake of accuracy improvement, new speech quality models that account for the features of lost packets in the calculation of the perceptual quality are needed. Notice that the payload content itself is not needed, but its features or characterization information (metadata) are crucial for the evaluation of the perceptual effect of missing packets.

By considering the voicing feature of wave signal included in lost packets during the assessment of live VoIP conversations, this paper proposes the following contributions:

(1) The development of new parametric voicing-aware speech quality estimate models, using a sophisticated assessment framework and multiple regression analysis, which account for both the packet loss location pattern and the voicing feature of signal waves included in dropped voice fragments. The receiver is notified about the voicing feature of dropped voice packets by the sender.

(2) The design of a new 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 speech wave fragments simultaneously.

(3) The design of a novel Markov model, which properly accounts for voicing feature of speech wave included in lost packets. The conceived loss model, which is calibrated at run-time using a computationally efficient algorithm, is employed to extract pertinent characterization parameters of packet loss process such as the mean loss durations for voiced and unvoiced packets, mean loss ratios for voiced and unvoiced packets.

(4) The proposal of a new efficient sender-based notification strategy used in order to inform the receiver about the voicing feature of sent packets. An analytical study is conducted to accurately quantify the additional bandwidth overhead and a practical configuration is recommended.

The performance evaluation study shows that our voicing-aware speech quality estimate models outperform voicing-unaware speech quality estimate models in terms of correlation and accuracy over a wide range of conditions. Indeed, we found that our parametric models achieve an excellent correlation above 0.95

and a mean absolute deviation in the order of 0.2 for ITU-T G.729 and G.711, equipped with a standard receiver-based Packet Loss Concealment algorithm, 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 voice 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 unaware speech quality models against the intrusive ITU-T PESQ algorithm. We conclude in Section 7.

2. Importance of voicing feature on speech quality evaluation

Basically, speech waves can be divided into voiced sounds such as ‘a’ and ‘o’, unvoiced sounds such as ‘h’ and ‘sh’ or silence, which is referred to as *voicing* feature [4]. Several studies reported in the literature have shown that the voicing feature of missing packets greatly influences the perceptual quality of delivered packet stream [4, 24, 25]. In accordance, besides the pattern of missing voice segments, a single-ended packet-layer SQM algorithm should account for the voicing features of lost packets. In [21], A. Clark indicates in the description of his widely employed SQM tool the existence of some outliers which can likely be removed by the consideration of the voicing feature of lost packets. Often, the sender checks the vocal source activity using a Voice Activity Detection (VAD) algorithm and ceases temporarily the transmission process upon the detection of a silence [3, 4]. In such a case, packet loss process can only affect voiced or unvoiced voice segments. Obviously, if the VAD mechanism is disabled then perceptual effect of lost packets, which occur during silences, are negligible [4].

For the sake of illustration, we plotted in Figure 1 the MOS-LQO calculated using sixteen standard 8s-speech sequences impaired by dropping either voiced or unvoiced 20ms-speech segments. The patterns of dropped packets are obtained using a voicing-aware bursty packet loss generator, which signifies that speech frames are dropped selectively according to their voicing feature. The listening quality scores are *automatically* estimated using the full-reference ITU-T SQM PESQ algorithm. Two standard speech CODECs, which are often used as reference, have been considered: G.729 (model-based coding scheme) and

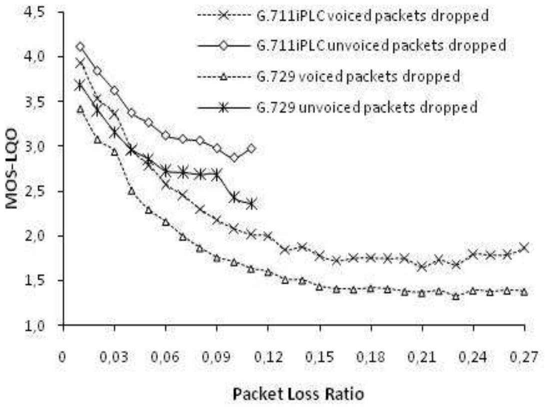


Figure 1 : Importance of the voicing feature of dropped 20ms-speech segments on perceived quality for speech CODEC G.729 and G.711iPLC.

G.711iPLC (sample-based coding scheme), which refers to the ITU-T speech CODEC G.711 equipped with the standard receiver-based Packet Loss Concealment (PLC) algorithm described in ITU-T Rec. G.711 Appendix I [26, 27]. The data rate generated by G.729 and G.711iPLC are respectively equal to 8 kbps and 64 kbps. Further details about performed empirical trials will be given later in Section 4. As we can see from Figure 1, the voicing feature of dropped voice frames significantly influences the quality scores regardless of the speech CODEC in use. Moreover, we clearly observe that dropped unvoiced 20ms-speech segments impair much *softly* the perceptual quality than dropped voiced 20ms-speech segments. Notice that the packet loss occur more frequently during voiced segments than unvoiced segments because they are statistically more frequent than unvoiced ones.

Besides the influence of voicing feature, the duration and location of loss runs effect notably the perceptual speech quality. Typically, the larger the duration of loss runs is, the bigger is the quality degradation. Moreover, it has been observed for certain model-based CODECs such as G.729 that dropping a single voiced frame located at the start rather than the middle or the end of a voiced sound entails much more perceptual quality degradation [4, 24, 25]. In fact, model-based coding schemes find major difficulty to recover such a missing voiced packet because the lack of suitable information to reconstruct the original voice frame. In addition, such CODECs entail lengthy error propagation period, which can lead sometimes to impair the whole subsequent voiced sound [25]. That is why, Speech Property-Based (SPB) priority marking and recovering schemes of speech fragments have been reported in the literature to improve the perceptual quality [4, 25, 28, 29]. This likely avoids losing perceptually important voice packets and hence improves the overall perceptual quality.

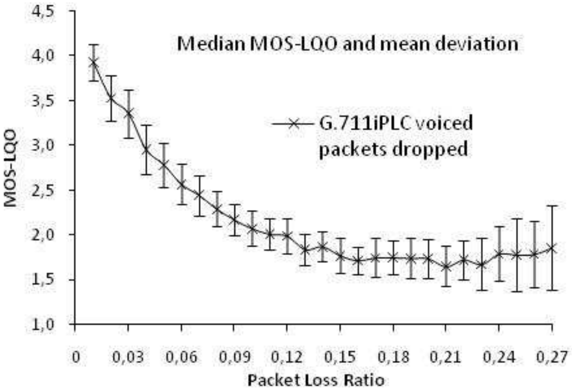


Figure 2 : Devation of perceptual quality for a given mean PLR among sixteen voice samples using G.711iPLC speech CODEC.

Recently, L. Ding et al. conceived new single-ended packet-layer SQM models that account for the voicing feature of signal wave included in missing voice packets [30]. Their voicing-aware SQM models were derived following a third-order regression polynomial model. The sole input parameter of the proposed voiced (resp. unvoiced) SQM model, which captures the effect of missing voiced (resp. unvoiced) packets, is the mean ratio of lost voiced (resp. unvoiced) packets. As such, their SQM models are unable to accurately capture the effect of bursty packet loss behavior. In such a case, the pattern of missing packets, i.e., duration and distribution of loss instances should be properly considered to accurately estimate the perceptual quality. This feature is supported by Figure 2, which represents the measured average of MOS-LQO scores and standard mean deviation for a given mean voiced-packet loss ratio using G.711iPLC. The SQM are performed using sixteen 8s-speech sequences and a voicing-aware bursty packet loss generator. This curve indicates that per-sequence speech quality scores can significantly *deviate* from the average score for a given mean packet loss ratio. Hence, the building of speech quality prediction models, which only use the mean packet loss ratio as predictor, leads likely to an inaccurate estimation of experienced listening speech quality. To reduce inaccuracy, speech quality prediction models should consider the location and duration of missing voice parts. Notice that the input parameter and polynomial degree have been selected *intuitively*, i.e., without a thorough statistical analysis investigation. In our opinion, the predictors and regressive model should be rigorously selected through an elaborated statistical analysis. Moreover, L. Ding et al. estimated the overall speech quality score of a packet loss impaired-speech sequence is calculated through a linear combination of the two scores produced by developed voiced and unvoiced speech quality models. In our opinion, the linear combination model can be greatly improved to tightly mimic the behavior rating of users by accounting for eventual interaction between voiced and unvoiced speech quality scores. Further, the receiver-based methodology adopted by authors to detect the voicing feature of lost packets introduces additional processing overhead with a high risk of wrong decisions, especially over a burst of packet loss. However, even with an additional consumed bandwidth, we believe that a sender-based strategy is more suitable and efficient.

3. Framework for speech quality modeling

The development of parametric speech quality models needs to set-up suitable speech quality assessment (SQA) frameworks. There are several approaches to develop a SQA framework, which is dependent on intended goals, e.g., evaluation of adaptive behavior of application or transport network policy, calibration and tuning of speech quality models, measurements of voice quality over existing voice transport systems, etc [11]. Particularly, for speech quality modeling, software-based SQA frameworks, rather than emulation-based test-beds or existing voice transport systems, are more suitable because of their price- and time-effectiveness and their ability to generate speech quality measurements under specific and controlled scenarios. It has been widely used in the recent few years to *evaluate* and *develop* parametric speech quality models over a wide range of packet-based network impairments [4, 14, 15, 16, 22, 23, 24].

Figure 3 gives the basic components of a software-based SQA framework which aims at modeling of the listening speech quality according to a set of signal- and packet- layer measurements. Basically, a set of standard reference speech sequences, that have specific properties such as sampling rate, sample precision, content, and duration, are encoded, packetized, then delivered through a system under test and that involves several sources of impairments such as packet loss, bit error, delay, echoes, and noises. Notice that for a software-based framework, the system under test can be a generic network simulator such as Network Simulator (NS2), a dedicated voice transport system simulator such as Message Automation and Protocol Simulation (MAPS) tool, or analytical network impairment models [31, 32, 33]. The system output is used to generate impaired versions of reference speech sequences. In our case, the corresponding relevant *packet-layer parameters* of delivered packet stream are properly measured and recorded. The quality of degraded speech sequences is either evaluated by human subjects or an accurate signal-layer full-reference vocal quality assessment algorithm such as ITU-T P.862 [17]. As mentioned previously, impractical subjective trials over a large scale testing are circumvented by using machine-executable speech quality assessment algorithms, which accurately mimic users' behavior rating [4].

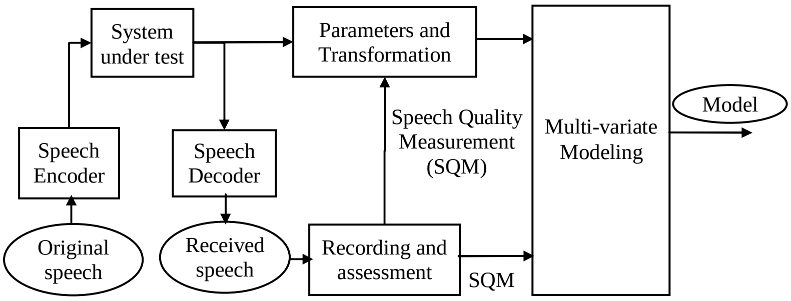


Figure 3: Vocal assessment framework for no-reference speech quality model developement.

In this work, we assume that the system under test, which is imitated based on a widely-used analytical Markovian model, only introduces bursty packet loss to the flow of sent packet stream. As illustrated in Figure 3, the potential set of parameters, that likely affects the perceived quality, is directly measured from the system under test such as mean loss ratios for voiced and unvoiced packets, maximal voiced and unvoiced burst durations, and the set of inter-loss gap and loss durations. Often, “base” measurements require to be transformed to precisely reflect human perception of experienced distortions. This helps to precisely account for the sensitivity of the overall listening perceptual quality score to each single base parameter variation. The developed framework enables monitoring and recording all characteristic parameters of packet loss process. For certain parameters a single value is returned, e.g., PLR (Packet Loss Ratio), CLP (Conditional Loss Probability), and maximal burst duration (MaxBD). For other parameters several values are recorded, e.g., inter-loss gap and burst loss durations. For each parameter, we determine, using regression, the *degree* and *fitting coefficients* of the polynomial that maximizes the *correlation* between the measured parameter values and MOS-LQO scores. For multi-value monitored parameters, we compute at the first stage the L_p-norm as follows:

$$L_p\big(X(k)\big) = \left[\frac{1}{M}\sum_{k=1}^{M}\big(X(k)\big)^p\right]^{1/p} \tag{1}$$

where, X(k) is the k^{th} measure of the parameter X and M is the total number of measured samples over an examined sequence. Notice that L_p-norm has been classically used to model the non-linearity behavior of human hearing system [4, 17]. In fact, L_p-norm highlights the effect of parameter variation on perceived quality. In this work, the value of p is varied in the set {1/10, 1/9... 1/2, 1, 2 … 8, 9}. The correlation factor is calculated as follows:

$$R = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \qquad (2)$$

where, R is the correlation coefficient between two cardinal-equal sets, x_i is known value of the measured quality score, y_i represents a value of the examined parameter, $\overline{x}$ and $\overline{y}$ represent, respectively, the mean value of the two examined sets, and N corresponds to the cardinality of each set. Algorithm 1 summarizes the transformation process applied to analyzed parameters and included in the developed framework.

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

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

When the optimal-correlated transformations for all potential parameters (predictors) are determined, the voicing-aware speech quality estimate models, which combines potential transformed parameters, is derived using multiple linear regression analysis (see Figure 3) [34]. To do that, a parameter/factor selection procedure should be followed to pick-up the parameters which exhibit a strong dependence with speech quality measurements. Basically, there are three techniques which can be used to select suitable parameters: forward regression, backward elimination, and stepwise regression [34]. The *backward elimination* technique initially subsumes all parameters and eliminates iteratively those with negligible fitting coefficients. The *forward regression* technique initially selects the parameter that achieves the best correlation factor with the set of known scores of the measured quality, then, iteratively, selects the most correlated one with the set of residual scores of the measured quality after the elimination of the effect of selected variables. 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 too low. The *stepwise regression* technique, which has been used in this work, is a combination of forward and backward technique. The selection of the suitable model is made step-by-step after examination of several combinations. Note that multicollinearty or dependence among potential parameters should be avoided and removed to obtain stable speech quality models.

In next section, we adopt the described strategy in order to develop parametric voicing-aware quality estimate models for packetized voice conversation over IP networks. The conceived vocal quality estimate models account for both the voicing feature and pattern of packet loss.

4. Speech quality models for dropped voiced and unvoiced frames

Obviously, the development of voicing-aware speech quality models needs to discriminate between voiced and unvoiced speech signals. In this work, we use the simple, yet efficient sender-based SUVING algorithm to distinguish between speech wave segments [35]. The SUVING algorithm utilizes *zero-crossings* (ZC) and *short-term energy* (STE) to identify the type of each examined speech fragment [35]. The zero-crossing metric represents the number of times in a speech fragment where the amplitude of sound wave changes its sign. The short-term energy of a speech fragment is calculates as follows:

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

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} samples. The energy is higher for voiced than unvoiced speech, and should be equal to zero for silent regions in clean speech signal recordings. Moreover, the zero-crossing rate is higher for unvoiced speech fragment than voiced one. The standard values of zero-crossing metric, for 10 ms clean voice segment, are roughly equal to 12 and 50 for voiced and unvoiced speech, respectively [35].

TABLE I: Voicing decision rules

	Zero-crossings (ZC)	**Short-Term Energy (STE)**	**Decision**
Rule 1	≈ 0	≈ 0	Silence
Rule 2	HIGH	LOW	Unvoiced
Rule 3	LOW	HIGH	Voiced
Rule 4	≈ 0	HIGH	Voiced
Rule 5	HIGH	HIGH	Voiced
Rule 5	LOW	LOW	Voiced
Rule 6	≈ 0	LOW	Unvoiced
Rule 7	LOW	≈ 0	Silence
Rule 8	HIGH	≈ 0	Background noise
If (ZC < zTh_1) Then ≈ 0		If (STE < eTh_1) Then ≈ 0	
Else If (ZC ≤ zTh_2) Then LOW		Else If (STE < eTh_2) Then LOW	
Else HIGH		Else HIGH	
$zTh_1 = 5$	Lower threshold of zero-crossings	$eTh_1 = 2\times10^{-5}$	Lower threshold of short-term energy
$zTh_2 = 35$	Upper threshold of zero-crossings	$eTh_2 = 10^{-2}$	Upper threshold of short-term energy

The presence of unavoidable background noise, which is typically characterized by high zero-crossing rate and low short-term energy, induces inaccuracy in S/V/U (Silence/Voiced/Unvoiced) discrimination process. To reliably identify the voicing feature of speech segments, a set of additional rules has been defined by SUVING developers which are summarized in Table I. The upper and lower thresholds, given in Table I, are used to classify metric as ≈ 0, LOW, and HIGH have been tuned and calibrated according to the properties of our processed speech materials.

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 [13]. 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.

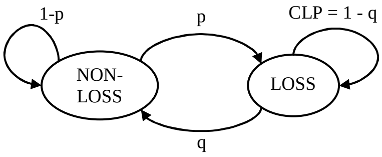


Figure 4: Gilbert/Elliot chain Markov loss model.

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. Notice that normally the value of p + q is less than one [36]. If p + q = 1 then the Gilbert/Elliot model is reduced to a Bernoulli model. The model is calibrated using ULP (Unconditional Loss Probability), which represents the PLR (Packet Loss Ratio), CLP (Conditional Loss Probability), and EBP (Effective Burstiness Probability), which are calculated as follows:

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

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 (8-20s) [37]. The value of EBP should be less than ULP according to the definition given in (4). This property should be considered during the design of SQM trials to produce realistic and accurate loss patterns. The developed Gilbert/Elliot model, which mimics the distortion introduced by the system under test (see Figure 3), has as input ULP and EBP, which have been finely varied to cover a wide range of conditions.

TABLE II: Experimental conditions for packet loss behavior using Gilbert Model

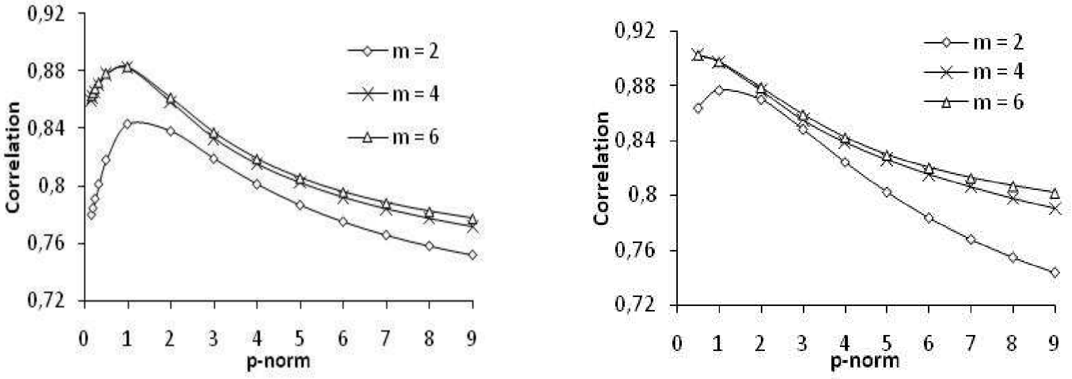
Parameters	Conditions	Instances
CODEC	G.711iPLC, G.729	2
Mean Packet loss ratio (PLR)	1, 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 45 %	12
Ratio of burstiness, R (EBP = PLR/R)	2, 4, 6, 8	4
Dropped frame feature	Voiced, Unvoiced	2
Audio sample	8 male, 8 female	16
Total number of combinations	2×12×4×2×16	3072

Table II summarizes the series of conducted empirical speech quality measurement trials. The speech materiel contains a total of sixteen standard 8s-speech sequences, spoken by eight male and eight female English speakers. For each speech sequence, we drop voice packets according to Gilbert/Elliot model loss generator while considering the voicing speech wave feature included in removed voice packets. In reality, original Gilbert/Elliot model drops media packets regardless the voicing feature of speech wave included in them. To enable a voicing-aware packet loss process, we monitor the voicing feature of presumed dropped packets in order to ignore loss instances which affect unsuitable packets. The degraded version of original sample is generated then the MOS-LQO score is calculated using the ITU-T full-reference SQM PESQ algorithm. In addition, the *effective* ULP, EBP, maximum burst duration (MaxBD), and the sets of inter-loss gap and loss durations are properly recorded for each (speech sequence, condition) pair. The total number of evaluated samples and conditions is equal to 3072.

TABLE III: Best correlation between measured transformed parameters and measured speech quality scores of G.711iPLC and G.729

CODEC	**Parameter**	**Voiced**			**Unvoiced**		
		m	**p**	**R**	**m**	**p**	**R**
G.711iPLC	ULP	2	-	0.952	2	-	0.906
	EBP	5	-	0.787	2	-	0.638
	MaxBD	4	-	0.491	2	-	0.540
	{inter-loss}	3	0.50	0.900	3	1	0.880
	{loss}	5	0.25	0.866	4	0.16	0.905
G.729	ULP	3	-	0.965	2	-	0.832
	EBP	5	-	0.790	6	-	0.556
	MaxBD	4	-	0.501	2	-	0.466
	{inter-loss}	4	0.2	0.924	3	0.5	0.825
	{loss}	4	0.11	0.951	1	0.11	0.836

The obtained measurements based on empirical trials are statistically analysed using Algorithm 1. For the sake of illustration, we plot in Figures 5a and 5b the result of application of Algorithm 1 to inter-loss gap duration metric for G.711iPLC speech CODEC. As we can note, the perceived quality is optimized for a specific combination, p-norm and polynomial degree, m, which is recorded and used during the application of the multiple variable regression analysis. Notice that some authors refer to such a process as parameter linearization with respect to the response variable [38].



(a) G.711 voiced ignored

(b) G.711 unvoiced ignored

Figure 5: Illustration of the application of Algorithm 1 to inter-loss gap duration metric.

Table III summarizes the optimal settings for G.711iPLC and G.729 speech CODECs which achieve the best correlation factor, R, between examined parameters and measured quality scores. As we can see, the parameter transformation of ULP, EBP, and MaxBD is independent of p-norm because they are single-value parameters. The transformed ULP, {inter-loss}, and {loss} exhibit high correlation with MOS-LQO, whereas, transformed EBP and MaxBD are relatively less correlated with MOS-LQO (see Table III).

As outlined in Section 3, *stepwise regression technique* has been adopted to derive suitable speech quality estimate models for bursty missing voiced and unvoiced packets. The proposed voicing-aware parametric speech quality estimate models for G.711iPLC and G.729 Speech CODECs, which are given in (5) and (6), have been selected after examination of several combinations of investigated packet loss process characterization parameters. In our statistical analysis, we found a strong correlation between ULP and L_p({inter-loss}) measures. Therefore, to assure the stability of speech quality model, either the ULP or the L_p({inter-loss}) parameter has to be eliminated from the final model.

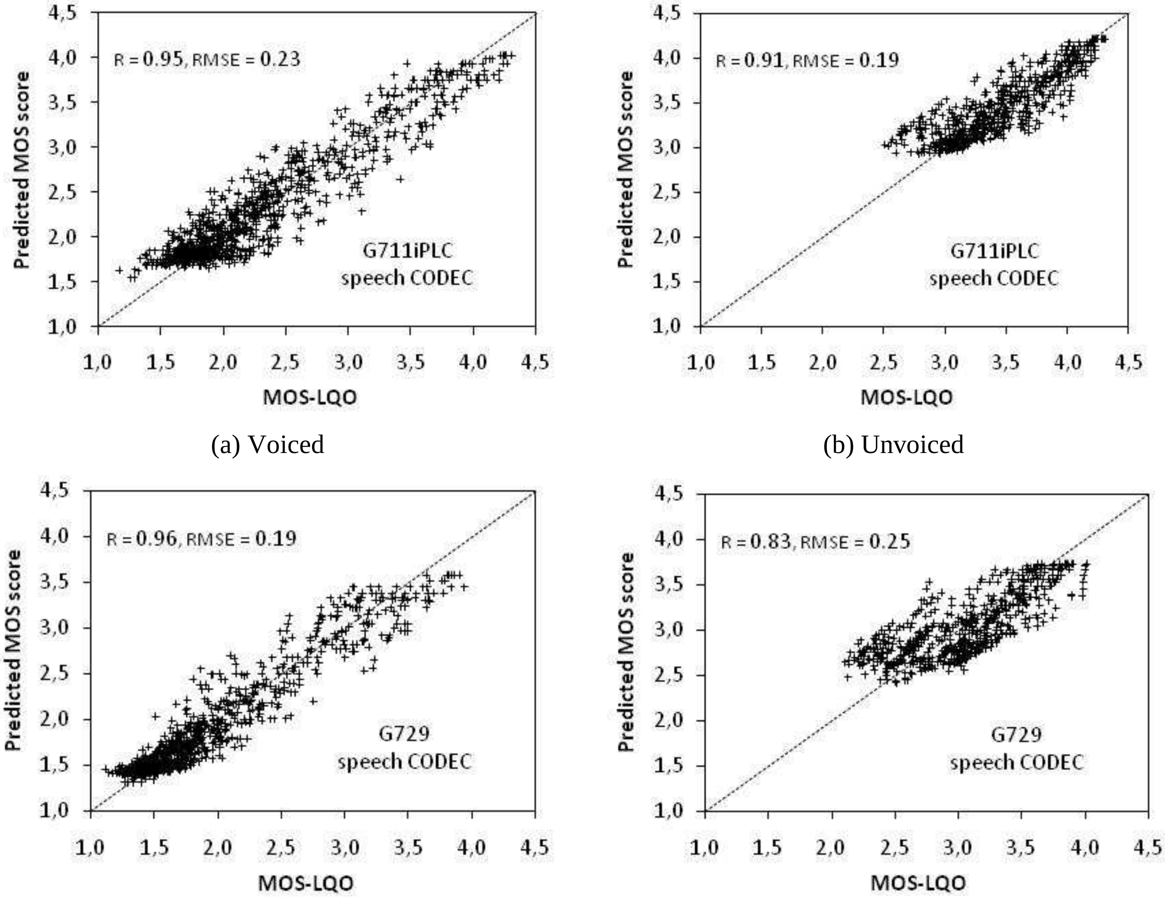
$$\begin{cases} \mathrm{MOS}_{\mathrm{V}}\left(\mathrm{ULP}, \{\mathrm{loss}\}\right) = 0.80 \times \mathrm{P}^{2}_{\text{v-G.711}}\left(\mathrm{ULP}\right) + 0.23 \times \mathrm{P}^{5}_{\text{v-G.711}}\left(\mathrm{L}_{1/4}\left(\{\mathrm{loss}\}\right)\right) \\ \mathrm{MOS}_{\mathrm{U}}\left(\mathrm{ULP}, \{\mathrm{loss}\}\right) = 0.48 \times \mathrm{P}^{2}_{\text{u−G.711}}(\mathrm{ULP}) + 0.52 \times \mathrm{P}^{4}_{\text{u-G.711}}\left(\mathrm{L}_{1/8}\left(\{\mathrm{loss}\}\right)\right) \end{cases} \quad \begin{matrix}\text{If CODEC} \\ = \text{G.711iPLC}\end{matrix} \qquad (5)$$

$$\begin{cases} \mathrm{MOS}_{\mathrm{V}}\left(\mathrm{ULP}, \{\mathrm{loss}\}\right) = 0.74 \times \mathrm{P}^{3}_{\text{v-G.729}}\left(\mathrm{ULP}\right) + 0.25 \times \mathrm{P}^{4}_{\text{v-G.729}}\left(\mathrm{L}_{1/9}\left(\{\mathrm{loss}\}\right)\right) \\ \mathrm{MOS}_{\mathrm{U}}\left(\mathrm{ULP}, \{\mathrm{loss}\}\right) = 0.31 \times \mathrm{P}^{2}_{\text{u-G.729}}(\mathrm{ULP}) + 0.68 \times \mathrm{P}^{1}_{\text{u-G.729}}\left(\mathrm{L}_{1/9}\left(\{\mathrm{loss}\}\right)\right) \end{cases} \quad \begin{matrix}\text{If CODEC} \\ = \text{G.729}\end{matrix} \qquad (6)$$

where, MOS_V and MOS_U stand for speech quality estimate models when loss process only affects voiced and unvoiced packets, respectively, P is the polynomial transformation applied to each selected parameter, the exponent of P refers to the polynomial degree. Table IV gives the optimal fitting coefficients of polynomials used in (5) and (6).

TABLE IV: Coefficients of polynomial regressive models

CODEC	**Quality model**	**Parameter**	**Fitting coefficients for each degree**					
			0	**1**	**2**	**3**	**4**	**5**
G.711iPLC	*Voiced*	ULP	3.992	-26.974	77.053	*	*	*
		L_p({loss})	4.064	1.224	-7.161	5.832	-1.783	0.186
	Unvoiced	ULP	4.244	-29.045	166.470	*	*	*
		L_p({loss})	4.195	-1.556	3.790	-6.752	3.287	*
G.729	*Voiced*	ULP	3.637	-35.898	194.417	-351.605	*	*
		L_p({loss})	3.414	-0.841	-3.260	2.931	-0.659	*
	Unvoiced	ULP	3.804	-28.789	173.329	*	*	*
		L_p({loss})	3.700	-1.300	*	*	*	*



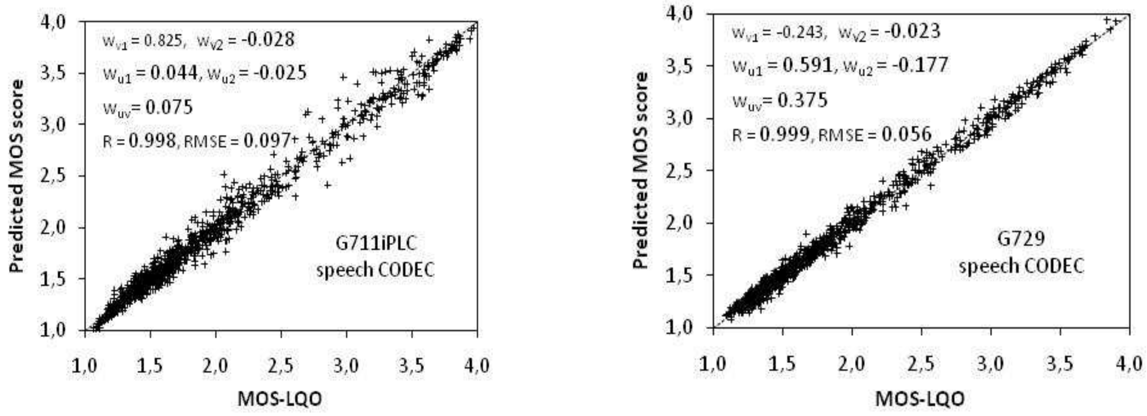
(a) Voiced (b) Unvoiced

(c) Voiced (d) Unvoiced

Figure 6: Accuracy of developed voicing-aware speech quality estimate models.

In order to verify the accuracy of developed models, we plot in Figure 6 the scatter-plots which show the relation between measured MOS-LQO and predicted MOS scores using models given in (5) and (6). As we can see, the predicted scores exhibit good correlation with measured one for all configurations (>0.83). Moreover, we see that our models achieve a very low Root Mean-Squared Error (RMSE) below 0.25.

After modeling of the effect of packet loss process that only affects voiced or unvoiced frames, actually it is required to develop a speech quality estimate model which quantifies the effect of *voicing-unaware* packet loss process that drops indifferently voiced and unvoiced packets. To do that, we drop media packets according to the original Gilbert/Elliot Markov model, which results in the deletion of both voiced and unvoiced packets. Three degraded speech sequences are generated for each treated clean speech sample. The first (resp. second) degraded speech sequence includes only missing packets that contain voiced (resp. unvoiced) speech wave. The third degraded speech sequence includes missing packets that contain voiced and unvoiced speech wave. The quality scores of the three produced distorted speech sequences are obtained using



(a) G.711 (b) G.729

Figure 7: Accuracy of developed *global* speech quality estimate models.

ITU-T SQM PESQ algorithm. The *overall* speech quality estimate model, which captures the effect of missing voiced and unvoiced speech packets, is obtained following a multiple linear regression analysis. The primary factors of overall speech quality estimate model are speech quality scores measured after the deletion of either voiced or unvoiced speech packets for a given speech sequence and loss pattern. After examination of several models, we found that the following expression achieves an excellent correlation and precision in the estimation of overall speech quality scores for the two considered CODECs:

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

where, w_i are the weighting coefficients which are obtained based on the minimisation of RMSE. The correlation factor, R, the RMSE, and the values of model coefficients are given in Figure 7. The scatter-plots (see Figure 7) prove the suitability of proposed speech quality models to estimate the overall speech quality scores. Notice that at run-time the values of $\mathrm{MOS_U}$ and $\mathrm{MOS_V}$ are calculated based on (5) and (6).

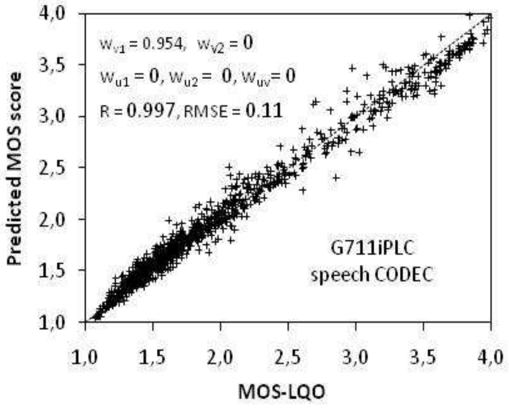


Figure 8: Optimized voicing-aware global speech quality estimate models.

According to the regression statistical analysis, we see the existence of negligible fitting coefficients in the proposed *generic* speech quality estimate model (< ±0.1). Variables with such negligible coefficients can be dropped, leading to simpler overall speech quality estimate models. This is illustrated in Figure 8, which shows the scatter-plot linking MOS-LQO and predicted MOS score for G.711iPLC, when only MOS_V factor is considered. As we can see, the obtained overall speech quality estimate model achieves strong correlation with an insignificant increase of RMSE. This suggests that it would be beneficial to seek a simple model for each speech CODEC rather than a complex more generic model for all CODECs.

5. A sender-based voicing feature notification strategy

The developed speech quality models need vital meta-data about the voicing feature of lost packets. To do that reliably, a sender-based notification scheme can be adopted. This is performed by piggybacking voicing feature of recent sent media packets toward the opposite end. Such a voice packet will be referred hereafter to as *media-voicing-report packet*. Three important factors should be considered to optimize the performance of such a scheme, framing duration (F), inter-delay between two consecutive sent media-voicing-report packets, denoted as T, and temporal window covered by the included voicing report (see Figure 9). The framing delay refers to the required delay to fill one voice packet, which is often set between 20 ms and 50 ms according to the one-way network delay, network workload, and loss severity. The larger the framing delay is, the smaller is the amount of meta-data information inserted in a media-voicing-report packet. In fact, an increase of framing delay results in a decrease of total number of sent packets. Moreover, the lower (resp. larger) the inter-media-voicing-report delay (resp. window duration) is, the greater is the additional consumed bandwidth. For the sake of reliability, overlapping windows should be used. Notice that the overlapping duration is dependent on the inter-media-voicing-packet and window durations. As such, a missing voicing pattern fragment can be recovered later at the reception of the next media-voicing-report packet. If overlapping is disabled then the window duration should be set equal to inter-media-voicing-report duration.

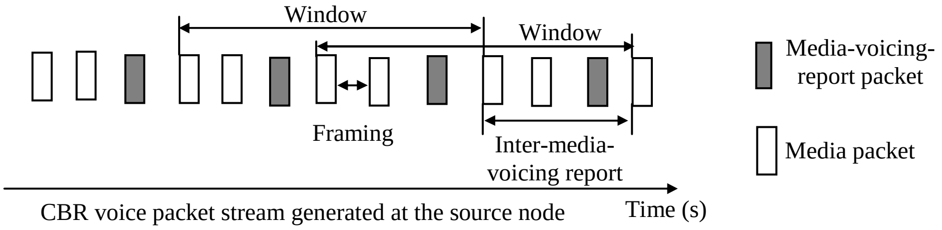
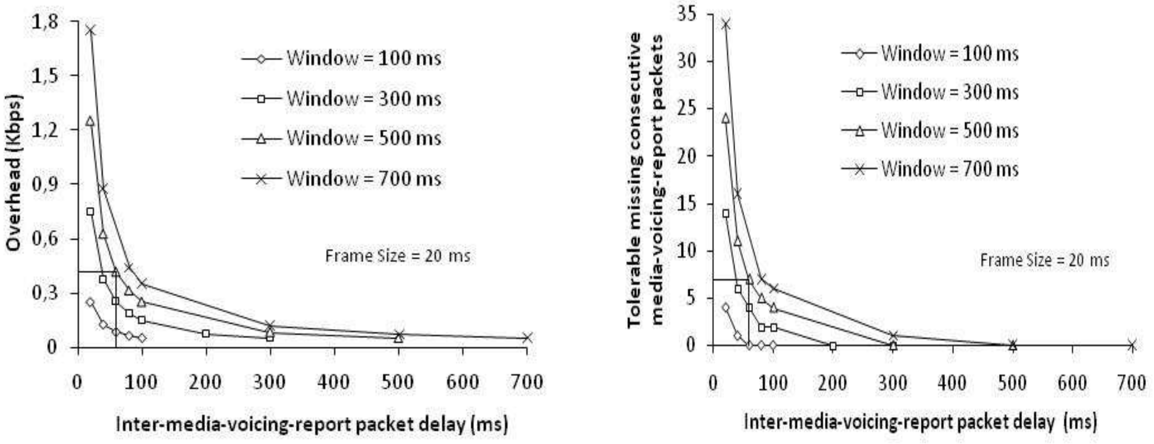


Figure 9: Temporal relations of sender-based voicing feature notification strategy.



(a) (b)

Figure 10: Overhead due to sender based notification strategy to send voicing information about sent packet stream.

To properly quantify the additional overhead, we assume for instance that the voicing feature of each media packet is coded using one bit, where 0 indicates an unvoiced packet and 1 indicates a voiced packet. In such a case, $\lfloor (W/F)/8 \rfloor$ additional bytes are required to tell the receiver about the voicing pattern of $\lfloor W/F \rfloor$ previous F-sec. voice packets, where W represents the window duration. In Figure 9, the selected temporal setting enables dropping a single media-voicing-report packet without losing a fragment of voicing pattern because the window size is equal to twice of inter-media-voicing-report packet delay. Generally speaking, if the goal is to tolerate losing X successive media-voicing-report packets, then the window duration should be set to X multiplied by the inter-media-voicing-report packet delay. Figure 10a shows the additional overhead in terms of consumed bandwidth under several window and inter-media-voicing-report packet delay settings. As we can see, the supplementary overhead in all investigated situations remains pretty low (<2 kbps). Figure 10b illustrates that a decrease of inter-voicing-report packet delay results in an increase of consecutive tolerable consecutive lost media-voicing-report packets. A good configuration under normal condition of packet loss consists of setting inter-media-voicing-report and window durations, respectively, to 60 ms and 500 ms, which results in an insignificant overhead equal to 0.42 kbps and a good tolerance of successive media-voicing-report packet losses of as much as 7 (see Figure 10b). Notice that the overhead is also related to voice source activity. The longer the activity duration is, the bigger is the consumed bandwidth.

In reality, the value of W and T can be fixed in advance or adjusted dynamically according to packet loss behavior. An optimisation strategy consists of adjusting T according to the prevailing channel state which can be either BAD or GOOD. Under BAD (resp. GOOD) network state the value of T should be decreased (resp.

increased) properly. To calibrate T at run-time, the mean inter-loss gap duration metric is relevant. In fact, an increase of mean inter-loss gap duration enables increasing T which reduces the additional overhead. On the other hand, a decrease of mean-loss gap duration needs to reduce T. Practically, to avoid losing media-voicing-report packets with high probability, the value of T should be set at least equal to the half of mean-loss inter gap. Obviously, a maximal tolerable threshold of T set to 500 ms for example should be used since the receiver requires the reception of voicing data as soon as possible. Notice that the receiver end often uses a non-overlapping *assessment window* which lies between 9s and 20s. The mean inter-loss gap duration can be either determined implicitly by monitoring the flow of received packets or explicitly through adequate feedback formulate and sent from the opposite end. The implicit strategy is less accurate than the explicit one because it assumes that transport routes to deliver packet stream are symmetric, which may be invalid under several scenarios.

Given the redundant distribution of voiced and unvoiced segments, it is likely possible that classical statistical lossless compression schemes can reduce bandwidth overhead. Note here 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. Therefore, the receiver entity can implicitly identify data packets which contain meta-data voicing information by only checking the packet length.

6. Voicing aware packet loss behavior model

To extract efficiently required voicing-aware measures of input parameters of previously developed speech quality models, we propose using a novel Markov model of packet loss process which accounts for voicing feature of lost fragments. The developed model constitutes a relevant extension to classical Gilbert/Elliot model (see Figure 11). It enables the accurate capturing the characteristic of the overall packet loss process over voiced and unvoiced speech wave 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 the failed delivering of a voiced and unvoiced voice packet.

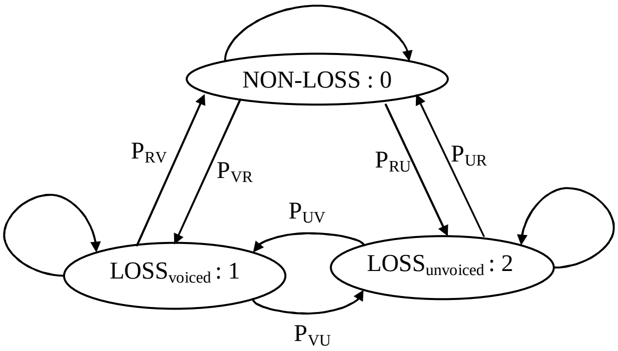


Figure 11: Markov model of voicing-aware packet loss process.

The packet loss model illustrated in Figure 11 is calibrated at run-time according to the flow of received and dropped media and media-voicing-report packets. An efficient voicing-aware packet loss driven algorithm is developed to update at run-time a set of counters which are used at the end of a monitoring period to calculate the transition probabilities. Therefore, parameters such as mean packet loss ratios and mean burst durations for voiced and unvoiced speech wave frames can be formally computed. Moreover, during the voicing-aware monitoring period, the set of inter-loss gap and unvoiced and voiced packet loss durations are properly recorded.

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, media-voicing-report packet. Algorithm 2 extracts V/U and loss patterns from the received media-voicing-report 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 max_v and max_u, respectively. It keeps track of the inter-loss gap and voiced and unvoiced loss durations using variables ac_{00}, ac_{11}, and ac_{22}.

Algorithm 2: Calibration and parameters estimation at run-time the voicing-aware packet loss model shown in Figure 11

```
if (new media-voicing-report packets is received) then
vu = read-vu-pattern(last-seq, cur-seq)
rcv = read-loss-pattern(last-seq, cur-seq)
  for i from *last-seq* to *cur-seq* do
     if (rcv[i] = "1") then // voice packet is received
        if (sate = "0") then
        $c_{00}$++,  $ac_{00}$++;
        elseif(state = "1") then
        if ($ac_{11}$ > maxv) then maxv = $ac_{11}$ end if
        record($ac_{11}$);  $c_{10}$++, state = "0"; $ac_{11}$ = 0;
        elseif(state = "2") then
        if ($ac_{22}$ > maxu) then maxu = $ac_{22}$ end if
        record($ac_{22}$);  $c_{20}$++, state = "0"; $ac_{22}$ = 0;
     else // voice packet is lost
        if (vu[i] = "V" and state = "0") then
        $c_{01}$++, state = "1"; record($ac_{00}$); $ac_{00}$ = 0; $ac_{11}$ = 1;
        elseif (vu[i] = "V" and state = "2") then
        if ($ac_{22}$ > maxu) then maxu = $ac_{22}$ end if
        record($ac_{22}$);  $c_{21}$++; state = "1"; $ac_{22}$ = 0;  $ac_{11}$ = 1;
        elseif (vu[i] = "V" and state = "1") then
        $c_{11}$++;  $ac_{11}$++;
        elseif (vu[i] = "U" and state = "0") then
        $c_{02}$++, state = "2"; record($ac_{00}$); $ac_{00}$ = 0; $ac_{22}$ = 1;
        elseif (vu[i] = "U" and state = "1") then
        if ($ac_{11}$ > maxv) then maxv = $ac_{11}$; end if
        record($ac_{11}$);  $c_{12}$++; state = "2"; $ac_{11}$ = 0;  $ac_{22}$ = 1;
       elseif (vu[i] = "U" and state = "2") then
       $c_{22}$++; $ac_{22}$++;
     end if
   end for
end if
```

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

$$\mathrm{ULP}_v = \frac{c_{01} + c_{11} + c_{21}}{\mathrm{nbt}} \qquad \mathrm{ULP}_u = \frac{c_{02} + c_{22} + c_{12}}{\mathrm{nbt}} \tag{8}$$

$$\mathrm{EBP}_v = \mathrm{ULP}_v \frac{c_{11}}{c_{11} + c_{10} + c_{12}} \qquad \mathrm{EBP}_u = \mathrm{ULP}_u \frac{c_{22}}{c_{22} + c_{20} + c_{21}} \tag{9}$$

where, ULP_v and ULP_U are mean packet loss ratios for voiced and unvoiced packets, EBP_V and EBP_U are the effective burstiness probabilities for voiced and unvoiced packets, and nbt refers to the total number of sent packets during the assessment period. Note that for a *continuous quality assessment* purposes, all variables, counters, and arrays are re-initialized at the start of a new assessment period.

7. Performance evaluation and models validation

To evaluate the performance of our voicing-aware speech quality estimate models, we set-up the voice quality assessment framework depicted in Figure 12. 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 the full-reference signal-layer ITU-T PESQ assessment algorithm. On the other hand, speech quality is predicted using voicing -unaware and -aware speech quality estimate models. Our voicing-aware speech quality estimate models are compared against the voicing-unaware models reported in [22]. During these empirical trials, a new 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.

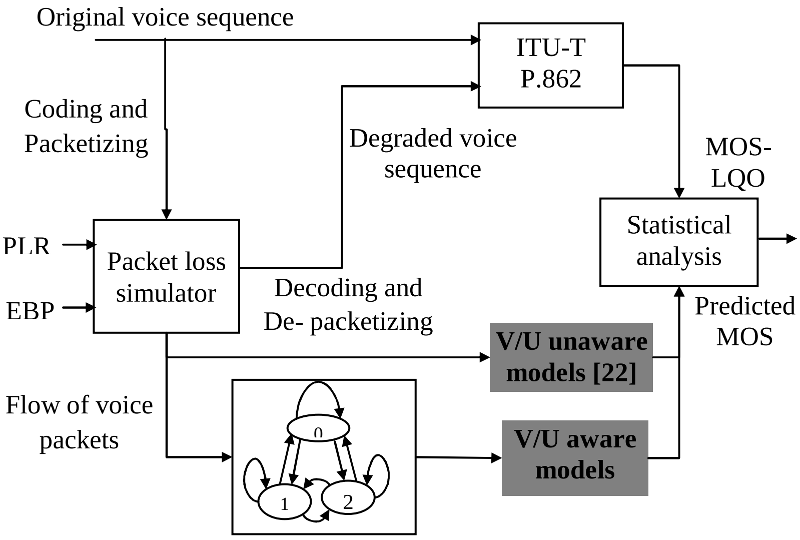
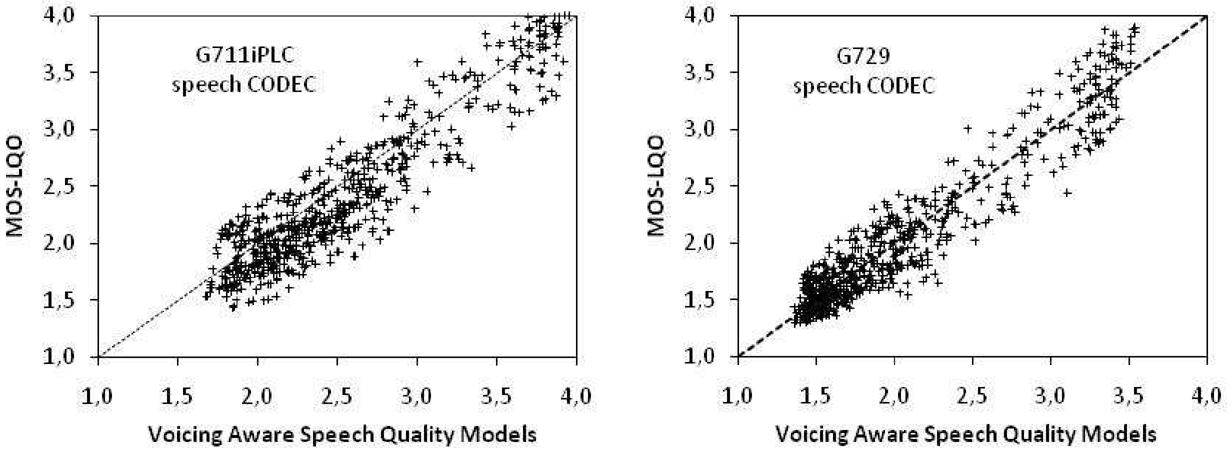


Figure 12: Evaluation framework of voicing aware speech quality estimate models.



(a) (b)

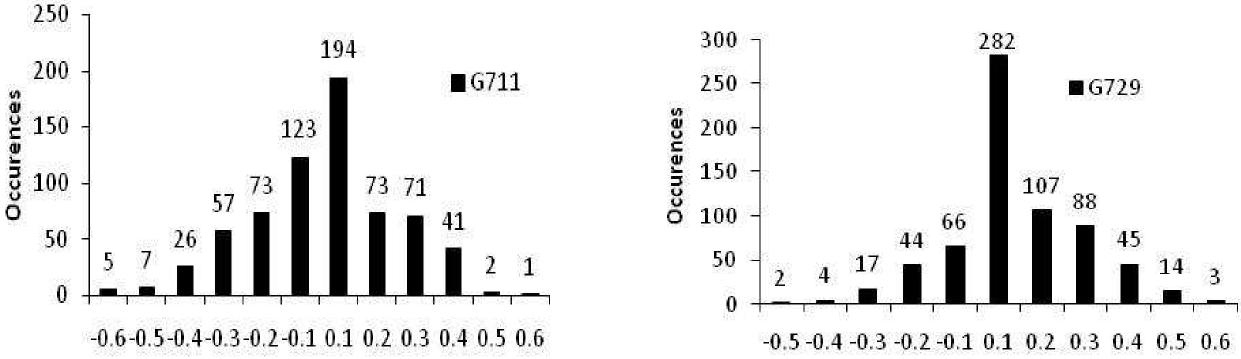
Figure 13: Validation of voicing-aware speech quality models.

Table V compares the performance of voicing -aware and -unaware speech quality estimate models for G.711iPLC and G.729 in terms of correlation and precision. As we can note, our voicing-aware speech quality estimate models achieve a better correlation factor above 0.95 for both considered speech CODECs which is pretty satisfactory. Moreover, our voicing-aware speech quality estimate models reduce notably, compared to voicing-unaware speech quality estimate models, the mean absolute deviation between measured MOS-LQO and predicted MOS scores using our models for both speech CODECs. The achieved accuracy is in the order of 0.2, which constitutes an excellent precision.

Table V: Performance comparison between voicing aware and unaware speech quality estimate models

	Voicing-Unaware Models [22]		**Voicing-Aware Models**	
	G.711iPLC	**G.729**	**G.711iPLC**	**G.729**
Correlation	0.927	0.910	0.954	0.961
Absolute mean deviation	0.61	0.92	0.22	0.17

Histograms shown in Figures 14 illustrate the distribution of predicted MOS scores with respect to measured MOS-LQO scores for the G.729 and G.711iPLC speech CODECs. These histograms prove the accuracy of our voicing-aware speech quality models to estimate MOS scores. 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 in practice given parametric, non-intrusive, and low complexity features of our developed speech quality models.



(a) (b)

Figure 14: Distribution of deviation between MOS-LQO measures and voicing aware model-based estimates of speech quality.

8. Conclusion

This paper extends conventional parametric no-reference speech quality models by accounting for the voicing feature of signal wave included in missing packets. An adequate software-based speech quality assessment framework has been set-up to develop voicing-aware speech quality models that enable to accurately quantify the effect of lost packets according to the feature of included signal wave in the payload. The overall speech quality model, which estimates the score at the end of an assessment interval, was properly developed following a multiple regression analysis. Two input parameters are used by the overall speech quality models to estimate the final score, which are the perceptual scores estimated when packet loss process affects either voiced or unvoiced media packets. The input set of parameters of speech quality estimate models were efficiently calculated based on a new Markov model of voicing-aware packet loss process calibrated at run-time. The performance evaluation study proves that our voicing-aware speech quality estimate models outperform voicing-unaware speech quality estimate models in terms of correlation and mean absolute deviation with MOS-LQO scores. Moreover, they exhibit high correlation and accuracy in the estimation of voice quality.

References

[1] Technology Marketing Corporation: TMCNet, Official Website: http://www.tmcnet.com/, visited on April 2009.

[2] G. Scheets, M. Parperis, and R. Singh, “Voice over the Internet: A Tutorial Discussing Problems and Solutions Associated with Alternative Transport”, IEEE Communications Surveys & Tutorials, pp. 22-31, Second Quarter 2004.

[3] H. Melvin, “The use of synchronized time in voice over Internet Protocol (VoIP) applications” PhD dissertation, University College Dublin, Ireland, 2004.

[4] C. Hoene, “Internet Telephony over Wireless Links”, PhD dissertation, Technical University of Berlin, Germany, December 2005.

[5] B. Sat and B. W. Wah, “Analysis and Evaluation of the Skype and Google-Talk VoIP system”, In Proceedings of IEEE international conference on Multimedia and Exposition, 2006.

[6] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, “Architecture for Differentiated Services”. IETF RFC 2475, December, 1998.

[7] R. Braden, D. Clark, and S. Shenker, “Integrated services in the Internet architecture: an overview”. IETF RFC 1633, 1994, June.

[8] Z. Li, “Improving Perceived Speech Quality for Wireless VoIP by Cross-Layer Designs”, Master thesis report, School of Computing, Communication and Electronics, University of Plymouth, September 2003.

[9] S. Madhani, S. Shah, and A. Gutierrez, “Optimized Adaptive Jitter Buffer Design for Wireless Internet Telephony”, In the Proceedings of IEEE GLOBECOM 2007, 26-30 November 2007.

[10] P. Hu, “The Impact of Adaptive Play-out Buffer Algorithm on Perceived Speech Quality Transported over IP Networks”, Master thesis report, School of Computing, Communication and Electronics, University of Plymouth, September 2003.

[11] 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.

[12] ITU-T Recommendation P.800, “Methods for Subjective Determination of Transmission Quality”, 1996

[13] L. Sun and E. C. Ifeachor, “Subjective and Objective Speech Quality Evaluation under Bursty Losses”, In Proceedings of Measurement of Speech, Audio and Video Quality (MESAQIN’02), June 2002.

[14] L. Roychoudhuri, E. Al-Shaer, and R. Settimi, “Statistical Measurement Approach for On-line Audio Quality Assessment”, In Proceedings of Passive and Active Measurements (PAM’06), 2006.

[15] 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.

[16] M. Masuda and 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.

[17] ITU-T Recommendation P.862, “Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To- End Speech Quality Assessment of Narrowband Telephone Networks and Speech CODECs”, February, 2001.

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

[19] A. Takahashi, “Opinion Model for Estimating Conversational Quality of VoIP”, in Proceedings of ICASSP'04, Vol. III, pp. 1072-1075, 2004.

[20] R.G. Cole and J. H. Rosenbluth, “Voice over IP Performance Monitoring”, Computer Communication Review, ACM SIGCOMM, 31(2), (2001)

[21] 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.

[22] 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.

[23] 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.

[24] L Sun, G Wade, B Lines and E. Ifeachor, “Impact of Packet Loss Location on Perceived Speech Quality”, in proceedings of 2nd IP-Telephony Workshop (IPTEL '01), Columbia University, New York, pp.114-122, April 2001.

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

[26] ITU-T Recommendation G.729, “Coding of Speech at 8 Kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)”, 2007

[27] Recommendation G.711 Appendix I, ITU-T, “A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711”, Sept. 1999.

[28] J. C. De Martin, “Source-driven packet marking for speech transmission over differentiated-services networks,” in Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing, Salt Lake City, UT, May 2001, pp. 753–756.

[29] Z Li, L Sun, Z Qiao and E Ifeachor, “Perceived Speech Quality Driven Retransmission Mechanism for Wireless VoIP”, Proceedings of IEE Fourth International Conference on 3G Mobile Communication Technologies, London, UK, June 2003, pp. 395 - 399.

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

[31] K. Fall and K. Varadhan, “The ns Manual”, VINT Project, November, 2001.

[32] GL Communications, “Protocol Simulation / Conformance Testing of SS7 & ISDN Protocols”, official website http://www.tmcnet.com/, visited on April, 2009.

[33] U. Jain, Y. Yokoyama, and A. Kumar, “Study of Factors Influencing QoS in Next Generation Networks” [Online], Available at http://www.eng.auburn.edu/department/csse/classes/comp8700 /index.html, visited on January, 2009.

[34] 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.

[35] 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.

[36] W. Jiang and H. Schulzrinne, “QoS Measurement of Internet Real-Time Multimedia Services”, Technical Report CUCS-015-99, Department of Computer Science, Columbia University, December 1999.

[37] 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.

[38] M. Werber, K. Kamps, U. Tuisel, J. G. Beerends, and P. Vary, “Parameter-Based Speech Quality Measures For GSM”, 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC2003), Beijing, China, September, 7-10, 2003.