Multimed Tools Appl (2014) 70:2309–2329
DOI 10.1007/s11042-012-1243-5

An unified VoIP model for workload generation

Carlos Ignacio Mattos · Eduardo Parente Ribeiro · Evelio Martín García Fernandez · Carlos Marcelo Pedroso

Published online: 27 September 2012

Abstract This paper presents a new model for VoIP workload generation. The novelty of our proposal consists in modeling the sessions by characterizing both the user behavior (session level) and the packet generation for an active call (intra-session level) with easily measured parameters and low computational complexity. This approach also facilitates systematic study of changes in user behavior and voice codec. The session level was modeled by analysis of call-holding time and time interval between successive calls. The model for call-holding time, characterizing the individual user behavior, uses the Pareto type 2 probability distribution. The time interval between calls is obtained from aggregate traffic and can be modeled by exponential probability distribution. Aggregate traffic is obtained by superposition of simultaneous sessions. The data used to characterize the session level were collected at the backbone of two Brazilian telecommunication carriers. The model for intra-session level comprises the characterization of the packet size and the packet inter-arrival time. The intra-session model was based on data generated in a laboratory environment, in order to properly characterize the codec influence on packet generation and to avoid the effects of delay, jitter and loss commonly present in an operational network. Models for constant bit rate and variable bit rate codecs were considered. A simulator was implemented and the results indicate that our model properly mimics the characteristics observed in real traffic and can be used for VoIP modeling and workload generation. Additionally, an application to automate the performance analysis was developed.

Keywords Traffic models · Voice over IP · Workload generation

C. I. Mattos · E. P. Ribeiro · E. M. G. Fernandez · C. M. Pedroso (✉)
Department of Electrical Engineering, Federal University of Parana,
Curitiba, Parana, Brazil
e-mail: pedroso@eletrica.ufpr.br

1 Introduction

The appropriate design of the network infrastructure is vital to support applications' demands. The network capacity can be under or super dimensioned leading, respectively, to problems in quality of service or to an inefficient utilization of the available resources. The application traffic usage plays a fundamental role in network planning. While a Web application demands bandwidth, low round trip time and low packet loss, real time applications also require low delay and jitter.

In order to allocate resources in an accurate way and to evaluate application performance, it is necessary to use accurate traffic models. The goal of network modeling is to represent its behavior as close as possible to reality. This enables one to evaluate the Quality of Service (QoS) and correctly allocate the resources for a given application. Daniel and Virgilio [29] stated that a good model needs to represent the reality with simplicity and accuracy to make it easier to understand and to provide reliable results.

Traditional telephone systems were extensively studied and well known models are available to assist engineers in planning the network capacity of such systems. However, traditional models fails to characterize VoIP (Voice over IP) traffic accurately, as showed in [14]. This happens because the VoIP call-holding time presents a different behavior from traditional telephone calls by the fact that long calls are not a rare event in VoIP systems. Call-holding time for VoIP is modeled with heavy tailed probability distributions, and furthermore, the traffic generated by codecs in use today are not handled by traditional models, leading to practical problems in network design and performance prediction for VoIP systems.

In this paper we present a new model for VoIP traffic generation and a software for capacity planning of VoIP systems. The proposed model characterizes session and intra-session levels in order to generate the aggregated traffic. The parameters of the model were estimated by using real data collected from two Brazilian telecommunication carriers. The first one offers a pure VoIP service, in other words, users receive and originate calls using only VoIP. The second one converts the mobile phone traffic to VoIP standards in order to exchange calls between heterogeneous end user technologies. The accuracy of the proposed model is confirmed through computer simulation by comparing the model workload generated with real data. A capacity planning tool was also developed to help network administrators to predict user perceived quality of VoIP systems by simulating a queue system fed with the traffic generated by the proposed model. The capacity planning tool performs a statistical analysis of simulation results and estimates the user perceived quality using the E-Model [9].

The proposed model characterizes the behavior of session and intra-session in a separated way. This approach leads to some advantages, like simplicity and accuracy, when compared to existing models. The session model is composed by the call-holding time and the time interval between successive calls. The intra-session model is directly related to the codec (coder-decoder algorithm) in use. There are many available codecs, each one with its own particular characteristics. The codec output traffic can be CBR (Constant Bit Rate) or VBR (Variable Bit Rate) and we have modeled codecs from both approaches. The data set used to characterize intra session behavior were generated in laboratory environment, using known softphones, like Skype and Ekiga, and available ITU-T's audio database [1].

Based on the proposed model, a simulator that generates synthetic VoIP traffic was developed. The data generated on the simulator was compared with real VoIP traffic using the auto-correlation function (ACF) and the Hurst parameter. The obtained results confirm that the data generated properly mimics VoIP traffic.

The remain of this paper is structured as follows. Section 2 summarizes related works, including recent publications on VoIP modeling. Section 3 presents the proposed model and the estimated parameters. Section 4 describes the implementation of the simulator for workload generation and presents an analysis of the obtained results. Finally, conclusions and future works are presented in Section 5.

2 Related works

In the early 90's, Leland et al. [27] showed that the nature of Ethernet traffic is statistically self-similar. A process is self-similar when it keeps part of its stochastic characteristics along a certain range of scales. This result has contrasted former works that stated the data traffic was Markovian [18] or could be characterized by packet trains [24]. The work of Leland et al. [27] provides an explanation for the self-similarity of the data traffic. More recently, Abry et al. [4], provides additional explanations regarding self-similarity and heavy tails in the Internet traffic.

Following these discoveries regarding of self-similarity in network traffic, Chen et al. [14] stated that the use of the exponential distribution to model the VoIP call-holding time is inappropriate. In order to model the VoIP traffic, they used data collected from a mobile phone system and suggested a mixture of two lognormal distributions to model the call-holding time.

Studies of the VoIP traffic generated by voice codecs were initially made by Schulzrinne et al. [25] and Casilari et al. [12]. Both papers suggest to use the ON–OFF model, being the ON state the speech time and the OFF state the silence time. Schulzrinne et al. used data from VoIP calls and Casilari et al. collected data from video conferences. In both papers it was concluded that a lognormal distribution can be applied to characterize the ON and OFF states.

A deeper approach to model the codecs traffic was proposed by Menth et al. in [30] where the authors used audio sources from an international database [7] and the software Picophone [35] for coding and transmission. Regarding CBR codecs, they proposed a model that uses deterministic inter-arrival time and constant packet sizes. Menth et al. used the ON–OFF model for codecs with silence detection and a Markov Chain with Memory [2] for VBR codecs.

Huang et al. [19], made a study regarding the behavior of codecs in the presence of packet losses. Voice traffic was generated using the software Skype, which implements Forward Error Correction (FEC) mechanism. The studied codecs were G.729, iSAC (internet Speech Audio Codec) and SVOPC (Sinusoidal Voice Over Packet Coder). All those codecs implement algorithms to adapt to the packet losses, properly adjusting the packet size. The G.729 codec varies the packet size in discrete levels. They report a great variability related to the packet sizes of iSAC and a smaller variability for SVOPC.

In this paper, we present an analysis of VoIP traffic by decomposing the aggregated traffic according to the individual user behavior, henceforth referred as session

modeling, and intra-session, at packet level, for VBR and CBR codecs. Our model is simpler than the proposed in [30] because it uses probability distributions and a time series model (Auto Regressive Moving Average, ARMA) to model the performance metrics instead of detecting voice activity by windowing process or other heuristics methods. This study complements previous work presented in [16] with a deeper analysis and adding the VBR traffic analysis.

3 Model description

The general idea of the proposed model for VoIP traffic characterization is based on the Scalable URL Reference Generator (SURGE) model, presented by Crovella et al. in [6], originally designed for Web servers. The SURGE model is different from others because it was developed based on the user behavior and the application characteristics. SURGE model applies the idea of *user equivalent*, according to Crovella et al. defined as follows: "The workload generated by SURGE should roughly correspond to that generated by a population of some known number of users. Thus, the intensity of service demand generated by SURGE can be measured in user equivalents (UEs). A user equivalent is defined as a single process in an endless loop that alternates between making requests for Web files, and lying idle. Both the Web file requests and the idle times must exhibit the distributional and correlational properties that are characteristic of real Web users. Each UE is therefore an ON/OFF process". In our model, we do not use the concept of user equivalent and instead of it we use the aggregated inter arrival time, for two main reasons: (a) the aggregated inter arrival time at session level is simpler to obtain than the individual user behavior, by simply observing the VoIP signaling messages, which do not exist in Web systems, and (b) the aggregated inter arrival time seems to fit very well with an exponential distribution, leading to good analytical perspectives.

The proposed model for VoIP systems uses five variables, as illustrated by Fig. 1. The variables are classified in two categories: *session level* representing the user behavior and *intra-session level* which represents the packet flow generated by one session. The variables are described as follows:

1. Session level
   (a) Time Interval Between Calls (δ_s): Represents the time interval between successive VoIP sessions.
   (b) Call Holding Time (ω_s): Describes the user call-holding time. From this point, the time from the call initiation until its end will be referred as *session active*.
2. Intra-session level
   (a) Packet Size (l_p): Represents the packet size for a given session.
   (b) Time Interval Between Packets (δ_p): This is the time interval between successive packets for one user session, observed at the sender, as generated by a certain codec algorithm without the effects of packet processing or network delays.
   (c) Deviation in Time Interval Between Packets (ε_p): This is the variable delay added to each packet in a sender due to packet processing.

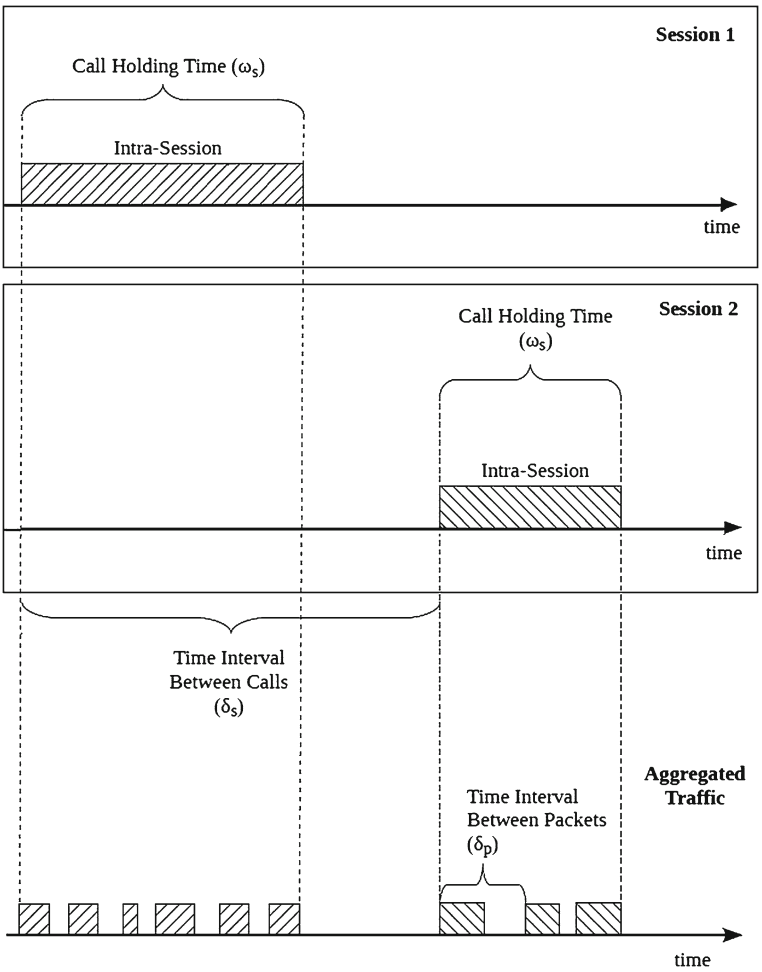

Fig. 1 Aggregated traffic produced by two non simultaneous voice conferences

The call-holding time and the time interval between calls are variables at the session level and are related to the user behavior. When a session is active, the packet generation starts according to the three variables: l_p, δ_p and w_p; these variables are strongly related to the codec in use.

3.1 Modeling the session level

In order to identify the models to characterize the variables at session level, it was necessary to capture real traffic from VoIP systems. The following sections describe the results from data collected in two major telecommunication carriers using VoIP in Brazil.

3.1.1 Data set

Data were collected from two telecommunications carriers. Carrier 1 offers a pure VoIP service and Carrier 2 offers a mobile phone service but internally converts the traffic to VoIP. On both carriers, data collection were carried out at the network backbone, which consists of a non congested Ethernet network in both cases. Packets belonging to the same session were identified based on the IP address (source and destination) and the sequence number from RTP protocol (Real Time Protocol [33]). The latter is necessary because the same user could have different destination calls simultaneously, as in a conference for example. In the following paragraphs a detailed description of the data collection for each carrier is presented.

Carrier 1 The VoIP service offered by Carrier 1 uses SIP protocol (Session Initiation Protocol) [32] for signaling and RTP protocol for data transport. The SIP protocol was designed to interact with other Internet protocols, initializing, modifying and ending sessions, independently of the media or application. When a session begins, voice is coded/decoded by a codec and transmitted with RTP protocol. The codecs in use are ITU G.711 [22] and ITU G.729 [21]. G.729 is used by 93 % of the sessions and the remaining sessions use G.711. The G.711 has a sampling frequency of 8 kHz and 8 bits per sample resulting in a rate of 64 kbps. This codec guarantees a high quality for the voice and it is used frequently as a reference standard. G.729 codec tries to achieve a good voice quality with lower transmission rate. G.729 possible rates vary among 6.4 kbps, 8 kbps and 11.8 kbps, depending on the desired voice quality. Both codecs in use at this carrier are CBR.

The analyzed network had about 10,000 users by the time of the data collection, in September 2007. The access network was formed mainly by ADSL (Asymmetric Digital Subscriber Line) links. The VoIP traffic generated by users is transported by a non congested Gigabit Ethernet network. In order to collect the data, the Ethernet switch ports that serve as backbone were mirrored in a way that the total VoIP system traffic was captured using the open source protocol analyzer Wireshark [31]. Call-holding time was calculated with the analysis of the time interval between SIP INVITE and the respective BYE messages. The time interval between calls is the time between two consecutive INVITE messages.

Information regarding the data set used to model the user behavior (session level) can be found in Table 1. The modeling for this carrier was performed using data sets 1, 2, 3 4 and 5. Data sets 6 and 7 were used for model validation. The analysis was done with busy hour traffic (BHT) and the data were carefully analyzed to ensure that all metrics present stationarity.

Table 1 Data used to model the user behavior (session level)

Data set 1	May 4, 2006
Data set 2	September 17, 2007
Data set 3	September 18, 2007
Data set 4	September 19, 2007
Data set 5	September 20, 2007
Data set 6	September 21, 2007
Data set 7	September 22, 2007
Data set 8	August 3, 2009
Data set 9	October 26, 2009

Carrier 2 It is a mobile phone carrier that uses VoIP to transport the calls originated in the mobile phones destined to other telecommunication carriers. The traffic is transfered from one carrier to another via an interconnection gateway. The voice coding is made by AMR (Adaptive Multi-Rate) codec [3] and the resulted data are sent through media gateways with RTP protocol. The signaling is made by BICC (Bearer Independent Call Control) protocol [20]. BICC messages IAM (Initial Address Message) and RLC (Release Complete Message) were used to identify the beginning and ending of a VoIP call. The time interval between calls was identified as the time between two IAM messages. The call-holding time is the time between IAM and RLC messages for a VoIP terminal. Carrier 2 has 80 % of its clients with prepaid charging plans and the remaining clients have postpaid plans. Data were obtained in two collections of eight hours done in weekdays in August and October 2009 (Data sets 8 and 9). The transport network is also a non congested Gigabit Ethernet.

In following sections, the probability distributions that characterize the session are shown. The methods used to verify the goodness of fit were Quantile-Quantile Plot (QQ-Plot) and Kolmogorov–Smirnov test. The QQ-Plot consists of plotting one distribution against another. When points gather on the 45° line there is a good adherence between both distributions. All statistic analysis were made using the R statistic software [34].

3.1.2 Modeling the call-holding time

Carrier 1 It was observed a heavy tailed behavior which can be modeled by Pareto type 2 probability density function. Pareto distribution is a heavy tail distribution that foresees extreme events [15]. Pareto type 2 or Lomax distribution [26] was used because traditional Pareto distribution can not properly represent required time values because it does not generate values lower than the scale parameter. Pareto type 2 can generate values lower than the scale parameter and have the same heavy tail behavior. Its cumulative distribution function is defined as

$$F(x) = 1 - \left(1 + \frac{x}{\beta}\right)^{-\alpha}, \tag{1}$$

where α is the shape parameter and β is the scale parameter.

Figure 2a illustrates the QQ-Plot which confirms the adherence of empirical to theoretical data. The parameters α and β were estimated by the maximum likelihood method. The obtained results were 2.16 for α and 166 for β. For $\alpha \leq 1$, the mean is not convergent and for $1 < \alpha \leq 2$, the mean converges but the variance does not. The α value obtained is near of the non-convergence region for variance, leading to a significant variability of the data. The average time of ω_s was 143.70 s and the standard deviation was 490.41 s. For this reason, we applied the QQ-Plot to verify the adherence between empirical and theoretical distributions. Results indicate that Pareto type 2 distribution can properly characterize ω_s.

Carrier 2 The analysis performed regarding Carrier 2 data was similar to the previous carrier and the empirical data also showed a good adherence to Pareto type 2 distribution. Figure 2b presents graphically the adherence between the theoretical and empirical distributions using the QQ-Plot. The estimated values for the distribution parameters were: $\alpha = 2.50$ and $\beta = 60$. This α value indicates that

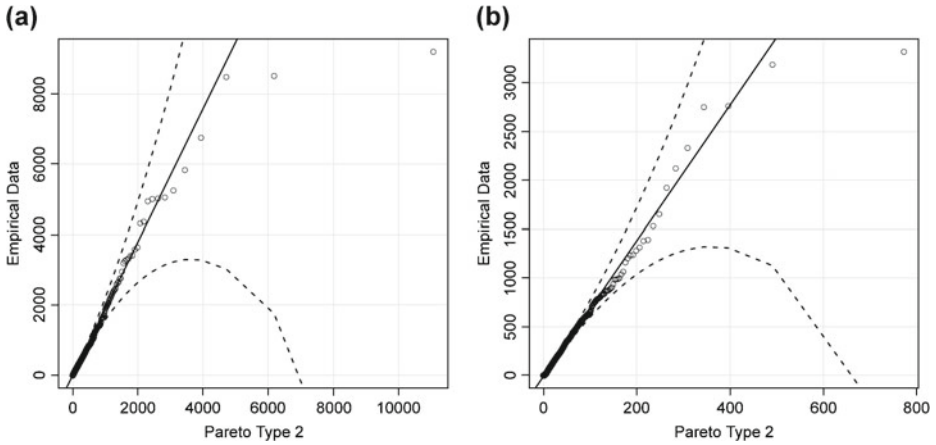

Fig. 2 Quantile-Quantile plot of Pareto type 2 distribution compared with ω_s at Carrier 1 (**a**) and Carrier 2 (**b**), with confidence level of 95 % (*dashed lines*)

the variability of the Carrier 2 call-holding time is lower than Carrier 1. In Fig. 2b, because of a greater α value, less points appear in the tail of the distribution.

3.1.3 Time interval between calls

Carrier 1 Time interval between calls in telephone systems is usually described by an exponential distribution [17]. In order to confirm that hypothesis, the Kolmogorov–Smirnov goodness of fit test was carried out with the empirical data and the theoretical exponential distribution. The obtained p value was 0.84. For rejection of the adherence hypothesis, the p value must be less than 0.05 [23]. The exponential distribution is parametrized only by the average time δ_s which was 1.125 s. As illustrated in Fig. 1, δ_s represents the time between two successive calls.

Carrier 2 The interval between calls from Carrier 2, δ_s, also presented an exponential distribution. The Kolmogorov–Smirnov test resulted in $p = 0.83$ which confirms mathematically the good adherence. The average δ_s observed was 0.506 s.

In both cases, a fast decay of the auto-correlation function (ACF) for δ_s series was observed, confirming the independence of successive call arrivals.

3.2 Modeling the intra-session level

The data set used to model intra-session level were generated in a laboratory environment in order to avoid effects of network impairments such as network delays and packet loss. These distortions could lead to a model that does not characterize only the codec, but also a specific network behavior. For each studied codec, 64 calls were generated using the available audio files from ITU-T [1]. The packet flow resulting from each voice conference was modeled separately and a general model was proposed for each codec.

The transmission of voice streams uses RTP protocol to transport voice encoded by a particular codec algorithm. For the two VoIP systems under study, different

codecs are used, mainly G.711 and G.729. Both generate a constant bit rate transmission. However, as there is a trend to use VBR codecs and in order to make a more extensive study, the modeling of intra-session was carried out using voice samples obtained from International Telecommunications Union for the English language encoded with the G.711, G.729, iSAC and SILK codecs [1].

3.2.1 Packet size (l_p)

The packet size is determined by the payload (coded voice) and the Ethernet, IP, UDP and RTP headers. The payload varies along the sessions but the headers have a fixed size of 58 bytes (due to protocols Ethernet, IP and UDP and RTP).

G.711 The data used to model the packet size of the G.711 codec were generated in Ekiga software. We used G.711 A-Law due to the fact that this version is employed in most of the world, including Brazil. The packet size is 214 bytes for the entire session, because it is a CBR codec. The FEC mechanism was not used.

G.729 The packet sizes are 16 bytes and 20 bytes for transmission rates of 6.4 kbps and 8 bps, respectively. The FEC mechanism was disabled in this test.

iSAC The packet size for iSAC codec varies over the voice conference, which characterizes it as a VBR codec. This variation is illustrated in Fig. 3, which shows the time series of consecutive packet sizes.

Considering the size of packets as a time series, we discovered that it could be characterized using an ARMA (Auto Regressive Moving Average) model. This model was studied by Jenkins and Box in [11] and is based on the dependence of the Z_t in function of the past Z_{t-k} elements and was assembled from the union of the model $AR(p)$ and $MA(q)$. In our proposed model for VoIP, the index t represents the order of arrival and Z_t is the packet size. The $AR(p)$ model is given

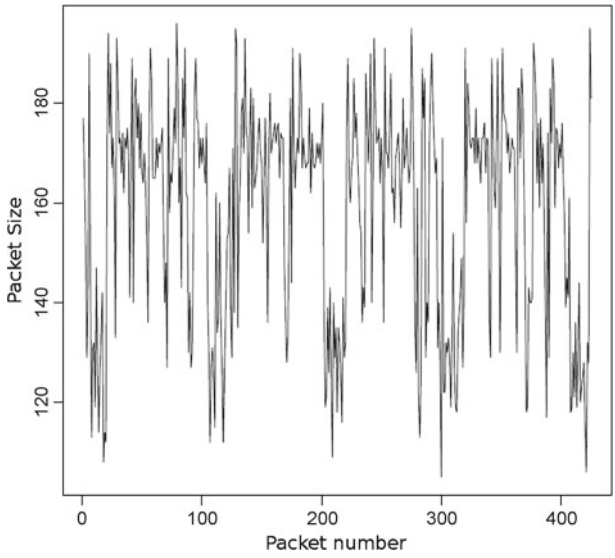

Fig. 3 Size of packets in a voice conference coded by iSAC

Table 2 ARMA(2,1) typical parameters for modeling packet size l_p for iSAC

ϕ_1	ϕ_2	θ_1	μ	a_t std
1.117	−0.190	−0.631	159	22

by $Z_t = \mu + \phi_1.Z_{t-1} + \ldots + \phi_p.Z_{t-p} + a_t$, where μ is the mean of the process, a_t represents white noise and $\phi_1, \ldots, \phi_p$ are parameters of the model. The $MA(q)$ model is given by $Z_t = \mu + a_t - \theta_1.a_{t-1} + \ldots + \theta_q.a_{t-q}$, where the current value of Z_t is composed by a weighted sum of present and precedent random noises. The values of $\theta_1, \ldots, \theta_q$ are model parameters. In $ARMA(p, q)$ model the values p and q respectively indicate the number of parameters in the AR and MA model.

Once the model is identified, the next step is to estimate the parameters and test the goodness of fit. The estimation method used was based on maximum likelihood function. Details on the method can also be found in [11]. The estimated parameters for the model are summarized in Table 2. Figure 4 shows the ACF of residuals, the QQPlot demonstrating a good fit for one voice conference packet sizes modeled by

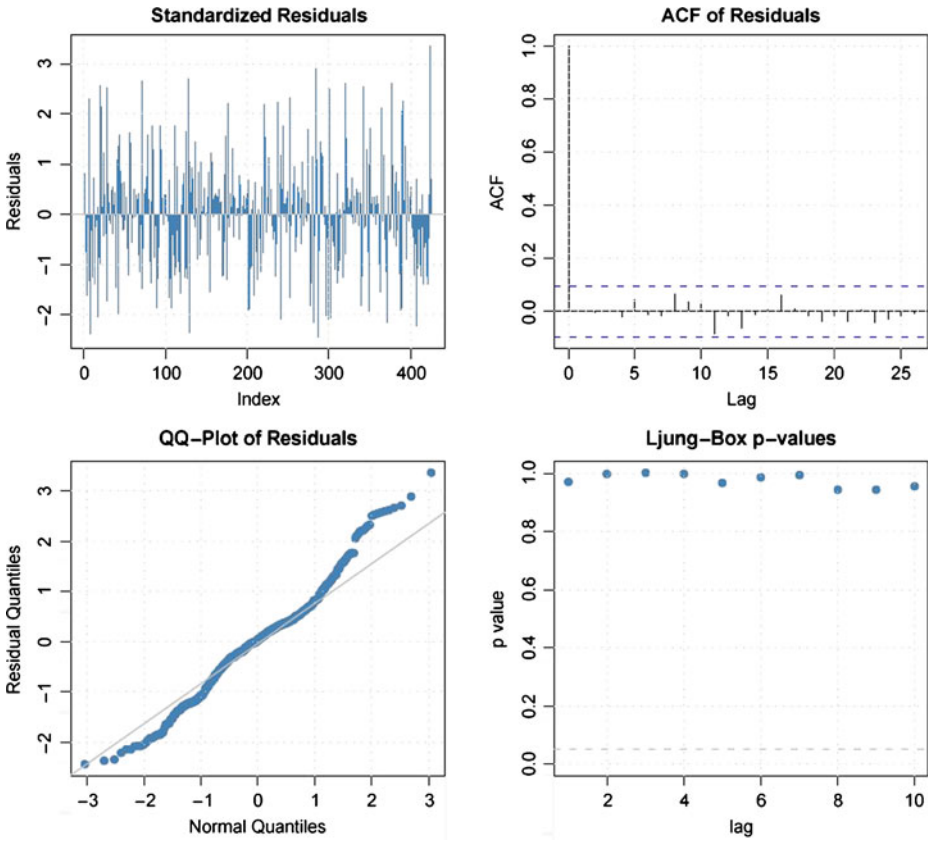

Fig. 4 Summary of ARMA(2,1) goodness of fit for packet size of a voice conference coded with iSAC

Table 3 ARMA(2,1) typical parameters for modeling l_p for SILK codec

	ϕ_1	ϕ_2	θ_1
Typical parameters 1 (76 %)	1.281	−0.332	−0.600
Typical parameters 2 (24 %)	0.240	0.462	0.548

ARMA(2,1) and the LJung–Box test. The LJung–Box test is commonly used for evaluation of randomness of residuals at distinct lag of a time series modeled by an ARMA model [10]—if the p values for each lag is greater than a limit value, marked with dashed lines, the hypothesis of randomness of residuals can not be rejected.

SILK The SILK is a VBR codec developed by Skype to replace SVOPC [28]. For this codec, the ARMA(2,1) model fits well the sequence of packet sizes generated in voice conferences. The ARMA parameters are summarized in Table 3. Two typical parameterizations were identified. Tests for goodness of fit are shown in Fig. 5, which confirms that ARMA(2,1) characterizes well the data series of packet sizes. Figure 5 shows the ACF, the QQPlot of residuals and the LJung–Box test.

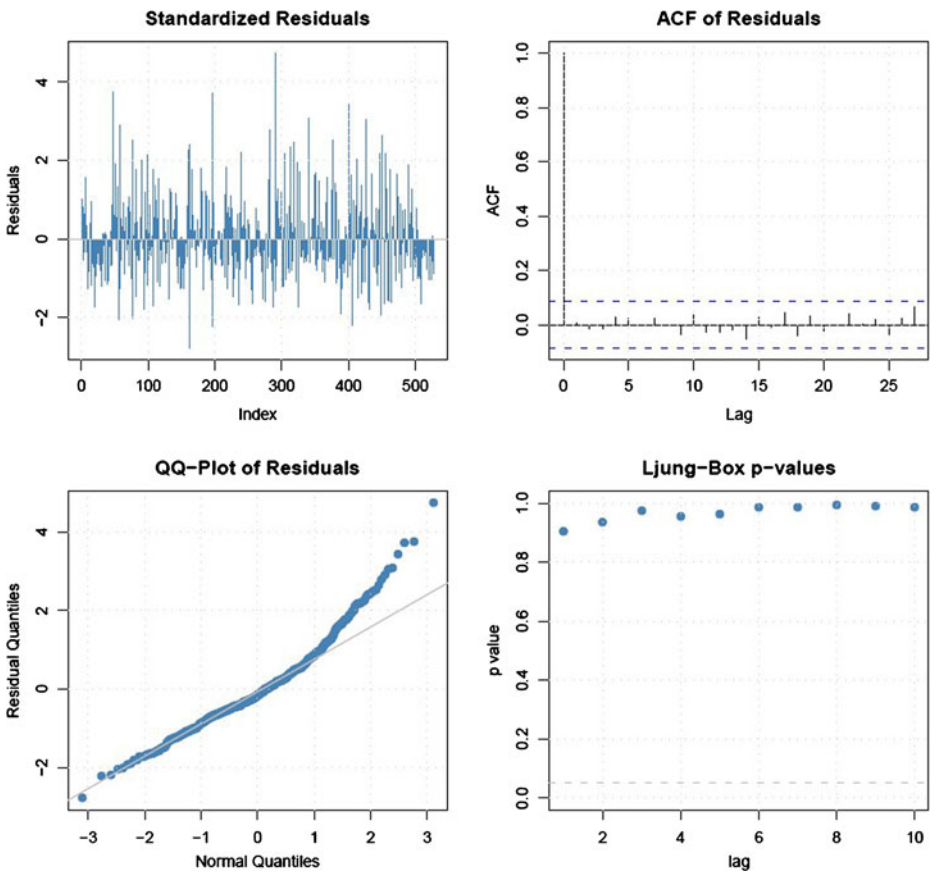

Fig. 5 Summary of ARMA(2,1) goodness of fit evaluation for packet size of a voice conference coded with SILK

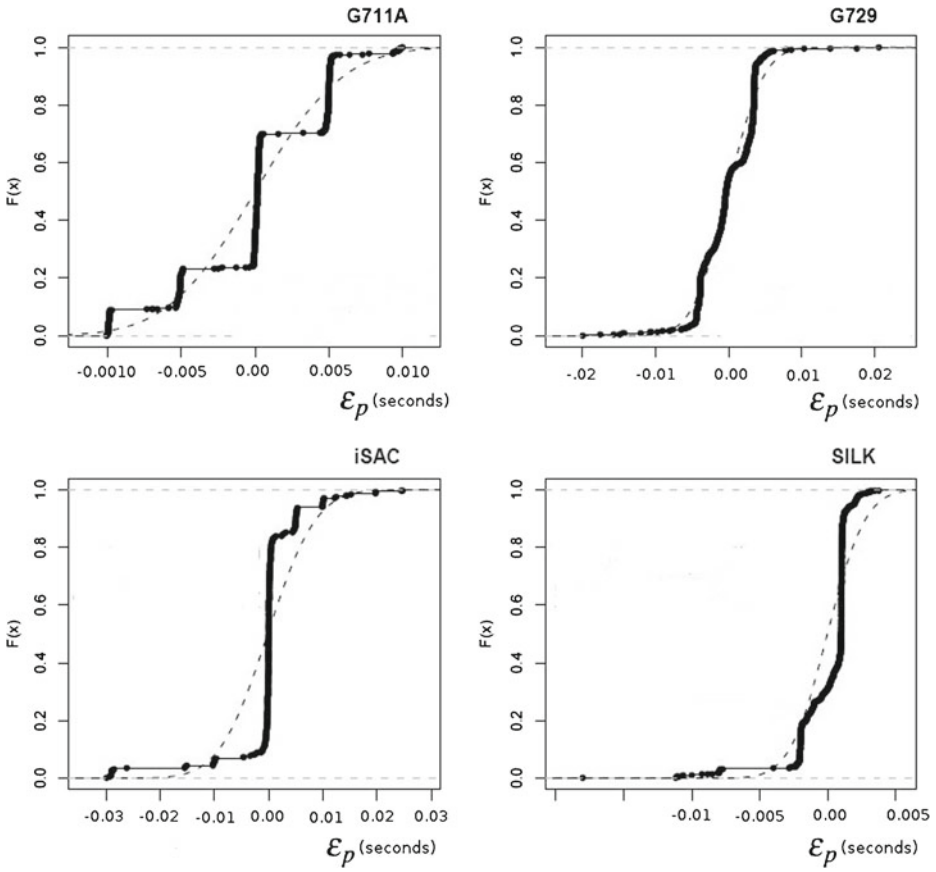

Fig. 6 Cumulative empirical distribution of ε_p compared to a Gaussian probability distribution

3.2.2 Time interval between packets (δ_p)

The VBR codecs usually produce packets of variable size at constant time intervals, in order to reduce the delay and jitter. If the time intervals between packets were variable depending on information, the delay and jitter could severely affects the quality perceived by the users. For all the codecs under consideration, the δ_p is constant for the session, and the size of packet varies depending on the data. For G.711, G.729 and SILK typical observed value was $\delta_p = 20$ ms, and for the iSAC was $\delta_p = 30$ ms. FEC mechanism was inactive.

Table 4 Parameters for ε_p when modeled with a Gaussian distribution

	Mean (s)	Standard deviation
G711	0.02	0.0047
G729	0.02	0.0038
iSAC	0.03	0.0070
G729	0.02	0.0022

3.2.3 Deviation in time interval between packets (ε_p)

The ε_p variation occurs mainly as a consequence of processing time, due to the multi-task nature of operational systems, and enqueue delay, generated by packetization algorithm. We recognize that ε_p could strongly affect the burstiness of the aggregate traffic, and influences the time interval between packets. ε_p was calculated from the data set and analyzed. The time series of successive observations of ε_p shows no auto-correlation, allowing modeling with probability distributions. Although there was no exact match with common probability distributions, their values seem to follow a Gaussian distribution. Figure 6 illustrates the cumulative distribution of empirical data compared with the Gaussian theoretical distribution (dashed line), for all codecs under consideration. Table 4 shows the parameters for ε_p when modeled with the Gaussian probability distribution.

4 Workload generation

In order to test the quality of the workload generation, a discrete event simulator was implemented, using the approach described in [5] and [13]. The software was implemented in Java language to make it independent of the platform. The operating principle is as follows. The interval between sessions was generated according to an exponential distribution. It is possible to have many active sessions simultaneously and each active session is handled by a separate thread in the simulator. The call-holding time was generated using the Pareto type 2 distribution. While the session is active, packets are generated according to ARMA(2,1), as described previously, and the user can freely configure its parameters. Besides the workload generation, the simulator also implements a FIFO queue whose maximum queue size can be adjusted. The heavy tail behavior of ω_s seems to be typical for VoIP systems, according to our findings and works from other authors [12, 14]. The alpha parameters usually lies between 2.1 and 2.6. Using these assumptions allows the determination of the beta parameter by calculating the mean, which can facilitate the modeling process.

4.1 Simulation results

The simulator was parametrized according to Carrier 1 data set to test the workload generation of a pure VoIP traffic. The queue size was considered large enough to prevent any packet drop. Package delay, queue occupancy and link utilization were computed from simulation results. In order to verify the model, the resulting synthetic traffic was compared with real data. A new data set was collected to perform the validation. Aggregated traffic at scale of 100 ms from three hours of traffic in the busy hour traffic (BHT) was considered. It is known that the time series representing aggregated traffic commonly presents a slow decay of its auto-correlation function, which is an indication of traffic self similarity. In a previous work [16], it was showed that aggregated traffic of Carrier 1 presents long range dependence, leading to a slow decay of the auto-correlation function. Figure 7 shows the ACF for real and synthetic data. Note that both curves are quite similar and exhibit long range dependence.

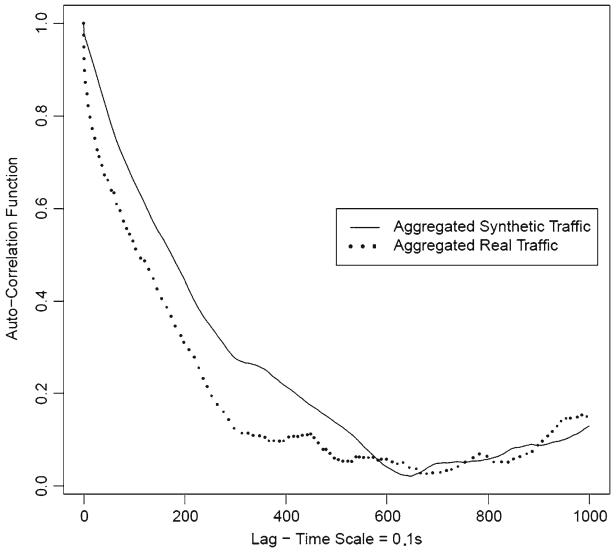

Fig. 7 Autocorrelation function of the aggregated traffic for the real and synthetic data

Another way to evaluate the presence of long range dependence is through the Hurst parameter (H). For self-similar series with long range dependence, $1/2 < H < 1$. When the Hurst parameter is closer to 1 the degree of the self-similarity and long range dependence is higher [15]. The Hurst parameter was estimated by Wavelet method. It was found to be 0.655 for synthetic traffic and 0.667 for real traffic. This result denotes that both traffics present self-similar characteristics with long range dependence.

The fact that both curves in Fig. 7 are in agreement and the Hurst parameters for both synthetic and real data traffics were very close and in the range of a self-similar process are a good indication that the proposed model properly characterizes the VoIP traffic.

As an example of model utilization the queue simulator was configured with a fixed service rate, FIFO discipline and infinity queue capacity, in order to verify the queue occupancy. The workload generator was used to produce the input traffic with a model parametrized for the SILK codec and average δ_s of 120 s, resulting in a aggregated traffic of about 3 Mbps. Each simulation was made with a generation of 10 million packets and the link rate was progressively increased. Figure 8 shows the queue size as a function of queue utilization with a dashed line representing the limits for a confidence interval of 95 %. The figure also shows the queue size when the call-holding time is modeled using an exponential probability distribution. The queue occupancy was quite different—the Markovian model tends to underestimate the average queue size.

Additionally, a software to automate the analysis was implemented. This application estimates the MOS (Mean Opinion Score) of VoIP streams using the E-Model, as described in [9]. The simulated network was configured with a dumbbell topology, with a bottleneck link between two routers and a workload generated by the proposed model. Two type of analysis were implemented: (a) evaluate the MOS for a specific workload condition over a single link with constant service time,

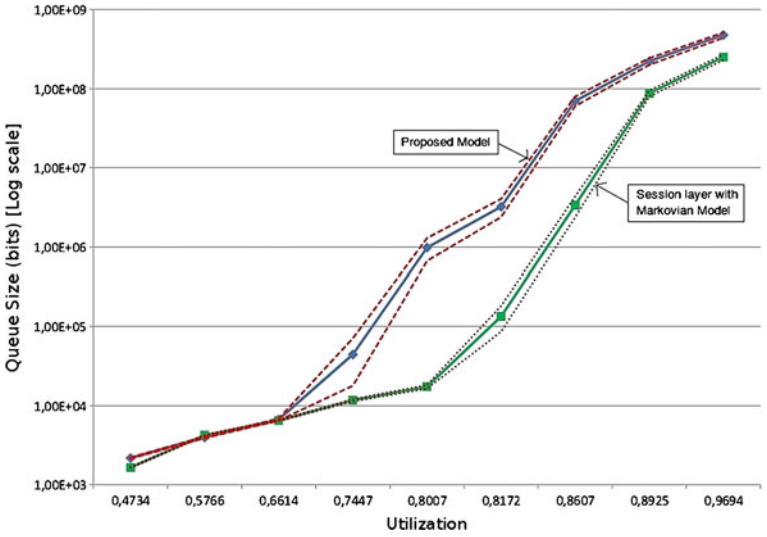

Fig. 8 A queue performance with FIFO, infinite buffer and fixed service rate

in bits per second, and (b) search for a link capacity to fulfill a desirable MOS, given a workload condition. Regarding the first option (a), the user can configure the bottleneck link capacity (bps), link delay (seconds) and the proposed model variables, including the codecs—the output is the estimated MOS, the delay and jitter for packets, with a configurable confidence level. There are graphical options to examine the response, as shown in Fig. 9, which can help the user to visually confirms the system stability. In the second option (b), the user sets the parameters for VoIP workload generation and a desired MOS level. The system starts a simulation with an over-estimated bottleneck link capacity, and from the simulation results the MOS

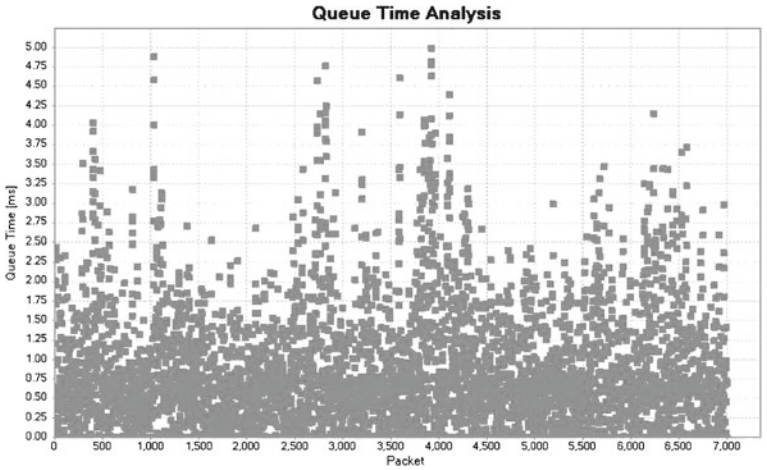

Fig. 9 Example of output of the software for automated analysis: queue delay

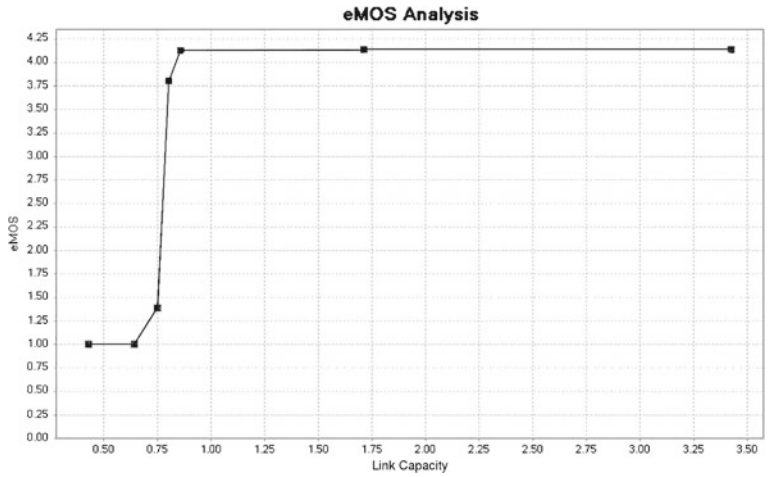

Fig. 10 Example of output of the software for automated analysis—link capacity search

is calculated. If the obtained value is greater than the desired MOS, the bottleneck link capacity is reduced by half of the previous value and the process is repeated—the next link rate to be tested is obtained with a binary search, until the desired MOS is attained. Figure 10 shows an example of the application output for the second type of analysis. By examining the output, the network designer can predict the performance and set the appropriate link capacity or the appropriate reservation of network resources for the system to work properly. All tasks are performed through computer simulations, with confidence level configured by the user.

4.2 Comparison with other models

In the systems under study, the use of CBR codecs dropped from 90 % in the first year of measurements to less than 10 % at the last year. Traditional models for telephone systems fail to capture the autocorrelation structure of VoIP traffic. Many authors have reported the self similar behavior of VoIP traffic by analyzing the auto-correlation function of aggregated traffic which can indicate whether the traffic exhibit long range dependence. Additionally, the increasing availability of more powerful hardware is popularizing the use of VBR codecs.

ON–OFF models with heavy tail probability distributions, as proposed by [25] and [12], can be used to produce synthetic traffic at packet level that mimics the long range dependence of real traffic. However, the parameters of these models are not easily related to operational conditions of the system, such as call arrival rate, call-holding time and codecs, therefore, their practical use for network design and performance prediction is difficult.

The model proposed by Menth et al. [30] addressed the problem of capturing long range dependence in VoIP VBR codecs, with good results. The authors claim to be the first model for VBR audio codecs and the employed model is the Markovian Chain with Memory [2]. However, its parametrization for specific traffic conditions is not easily performed—the authors published a table with the parameters in their *website* and do not consider how the call-holding time affects the aggregated VoIP

traffic. It is important to note that VBR codecs present both short and long range dependence—the first caused by temporal and spatial redundancy explored by the codec to improve its performance and the second caused by user behavior.

In order to compare the performance of the proposed model we collected only the VBR sessions of real traffic, coded with SILK and iSAC codecs. We captured 3.33 h of traffic (61,552,138 packets). In order to illustrate the performance of proposed model, we also generated synthetic traces using the self similar series and using the Markovian model. The self similar series was generated employing the Fractional Gaussian Noise (FGN) method [8]. For the FGN, the Hurst parameter of aggregated traffic was estimated using the Wavelet and R/S methods. Figure 11 shows the ACF for the aggregated traffic at one second scale, compared with the proposed model, FGN and Markovian. One can see that the proposed method presents a closer match to the real traffic. The self similar series, generated with FGN, can mimic the long range dependence, but the short range dependence of real trace was not properly reproduced.

The model we developed splits the problem of modeling VoIP traffic into two layers: the session layer, modeling the user behavior, and the intra-session layer, modeling the codec behavior. This approach presents the advantage of enabling the study of changes in user behavior, for example, as a result of changes on charging plans or changes from business users to home users. Furthermore, the session layer can be easily parametrized with aggregated arrival rate of calls and with the call-holding time for the individual sessions, which is modeled by Pareto type 2 distribution—the alpha parameter affects the long range dependence of traffic. One could use the typical values or estimate the parameters with own data—all the parameters can be obtained from signaling protocol analysis. The call-holding time is affected by the alpha parameter, and this variable is related to the long range dependence of the aggregated traffic. We conjecture that the alpha value could be typical for specific set of users, which is useful for network planing and pricing.

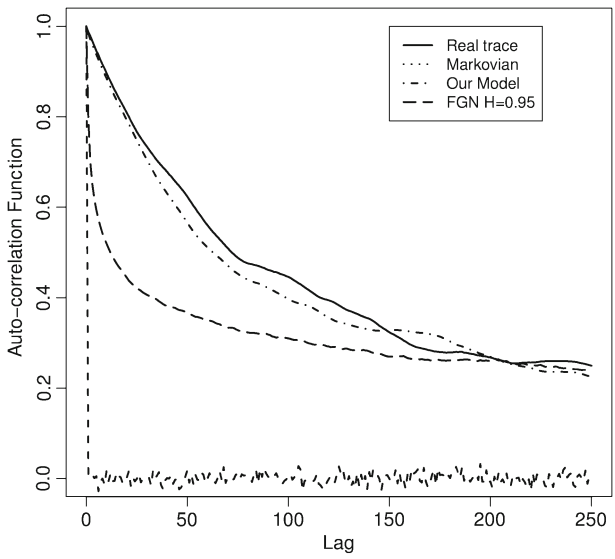

Fig. 11 Auto correlation function of aggregated traffic generated by several models (time scale one second) considering only VBR codecs, compared with real traffic

With this model it is possible to predict the consequences in the traffic caused by modifications of user behavior or by changes in the codecs independently. To model the intra-session level we apply the ARMA Model, with parameters that depend on the codec in use. The resulting traffic mimics the short and long range dependence of VoIP VBR codecs. Also, the proposed model presents low complexity, and can be used in computer simulations or by traffic generators. We are not aware of any other model to represent the VoIP traffic behavior that allows systematic study of such changing in operational conditions and that relate the model variables with easily measured values.

5 Conclusion and future work

The proposed model for VoIP workload generation employs simultaneously user behavior and voice coding algorithms properties to produce the aggregated packet flow that mimics the real traffic. User behavior was characterized by observing the data sets from two commercial VoIP systems in Brazil, along three years. The aggregated arrival rate for call initiation was consistently modeled using the exponential probability distribution and the call-holding time was modeled using the Pareto type 2 probability distribution. The intra-session layer was modeled with conversations data sets in English language, publicly available, for CBR and VBR codecs.

Example of how to employ the model to predict user quality of experience in terms of MOS was presented, using computer simulation. Three simulation softwares were implemented: (a) for workload generation, (b) for queue simulation and (c) for automating the analysis and capacity planing. The workload generation model was validated by comparing the synthetic trace with real one, collected exclusively for validation. The results show that the proposed model can mimic the short and long range behavior of real traffic.

The model is easily parameterized as its parameters can be obtained by simple measurements on the network. This is the main advantage of our model if compared with other available models. Additionally, if the parameters of Pareto distribution are not known one could use the typical shape parameter. The scale parameter can be evaluated based on the average of the call-holding time.

The effects of deviation in time interval between packets, ε_p, is not fully understood and it could play a vital role in the quality of service in small devices. The analysis of main causes of ε_p and its effects is a topic of future research.

Acknowledgements We wish to thank the Electrical Engineers Rafael Alesi and Willian Mattos for their dedication during the year of 2010 in implementation the application to automatize the analysis of simulation results, the Computer Engineer Jeferson Caldeira for programming the scripts to analyze the large amount of data, the Electrical Engineer M.Sc. Edgard Massahiro for providing of the data from Telecommunication Carrier 2 and to Electrical Engineer M.Sc. Mateus Cruz for providing the data from Telecommunication Carrier 1.

References

1. 12 ISG (2009) ITU-T test signals for telecommunication systems. Test vectors associated to recommendation ITU-T P.50 appendix I

2. 16th Int Teletraffic Congr (ITC) (1999) A memory Markov chain model for VBR traffic with strong positive correlations, pp 827–836
3. 3GPP (1999) 3GPP recommendation TR26.075: performance characterization of the AMR speech codec
4. Abry P, Borgnat P, Ricciato F, Scherrer A, Veitch D (2009) Revisiting an old friend: on the observability of the relation between long range dependence and heavy tail. Telecommun Syst (Special issue on Traffic Modeling, its Computations and Applications) 43(3–4):147–165
5. Banks J, Carson J, Nelson BL, Nicol DM (2004) Discrete-event system simulation, 4th edn. Prentice Hall
6. Barford P, Crovella M (1998) Generating representative Web workloads for network and server performance evaluation. SIGMETRICS Perform Eval Rev 26(1):151–160
7. Bavarian archive for speech signals (BAS) verbmobil 6.1 (1996) http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html. Accessed September 2011
8. Beran J (1994) Statistics for long-memory processes. Chapman and Hall, New York
9. Bergstra JA, Middelburg CA (2003) ITU-T recommendation G.107: the e-model, a computational model for use in transmission planning
10. Box GEP, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 65(332):1509–1526
11. Box G, Jenkine G, Reineel G (1994) Time series analysis, 3th edn. Prentice-Hall, New York
12. Casilari E, Montes H, Sandoval F (2002) Modelling of voice traffic over IP networks. In: Proc of communication systems, networks and digital signal processsing (CSNDSP)
13. Chandy KM, Misra J (1981) Asynchronous distributed simulation via a sequence of parallel computations. Commun ACM 24:198–206
14. Chen WE, Hung HN, Lin YB (2007) Modeling VoIP call holding times for telecommunications. IEEE Netw 21:22–28
15. Crovella ME, Bestavros A (1997) Self-similarity in World Wide Web traffic: evidence and possible causes. IEEE/ACM Trans Netw 5(6):835–846
16. de Mattos CI, Ribeiro EP, Pedroso CM (2010) A new model for VoIP traffic generation. In: International telecommunication symposium. Brazilian Telecommunication Society, Manaus, Brazil
17. Flood J (1997) Telecommunications networks, 2nd edn. The Institution of Eletrical Engineers
18. Heffes H, Lucsmtoni DM (1986) Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer peformance. IEEE J Sel Areas Commun SAC-4:856–868
19. Huang TY, Huang P, Chen KT, Wang PJ (2010) Could Skype be more satisfying? A QoE-centric study of the FEC mechanism in an internet-scale VoIP system. IEEE Netw 24(2):42–48
20. ITU-T (2000) Q.1901 bearer independent call control
21. ITU-T (1996) Recommendation G.729. Coding of speech at 8kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)
22. ITU-T (1988) Recommendation G.711. Pulse code modulation (PCM) of voice frequencies
23. Jain R (1991) The art of computer system performance analysis: techniques for experimental design, measurement, simulation and modeling. Wiley, New York
24. Jain R, Routhier SA (1986) Packet trains: measurements and a new model for computer network traffic. IEEE J Sel Areas Commun 4:986–995
25. Jiang W, Schulzrinne H (2000) Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation. In: Proceedings of 9th IEEE international conference on computer communication networks
26. Klugman SA, Panjer HH, Willmot GE (2004) Loss models from data to decisions, 2nd edn
27. Leland WE, Taqqu, MS, Willinger W, Wilson DV (1994) On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Trans Netw 2(1):1–15
28. Lindblom J (2005) A sinusoidal voice over packet coder tailored for the frame-erasure channel. IEEE Trans Speech Audio Process 13(5–2):787–798
29. Menascé DA, Almeida VA (1998) Capacity planning for Web performance. Prentice Hall
30. Menth M, Binzenhöfer A, Mühleck S (2009) Source models for speech traffic revisited. IEEE/ACM Trans Netw 17(4):1042–1051
31. Orebaugh A, Ramirez G, Burke J, Pesce L (2006) Wireshark & Ethereal network protocol analyzer toolkit (Jay Beale's open source security). Syngress Publishing
32. Rosenberg J, Schulzrinne H, Camarillo G, Johnston A, Peterson J, Sparks R, Handley M, Schooler E (2002) RFC 3261: session initiation protocol

33. Schulzrinne H, Casner S, Frederick R, Jacobson V (1998) RTP: a transport protocol for real-time applications
34. Team RDC (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org. Accessed September 2011
35. Vitez M (2011) Picophone. Available at http://www.vitez.it/picophone/. Accessed September 2011

Carlos Ignacio Mattos conclude his undergraduate in Electrical Engineering in 2008 from Federal University of Parana, working with mathematical modeling and frequency response of network analyzers. Received Master of Science in Electrical Engineering in 2011 from Federal University of Parana. His interests include modeling, coding and network technologies. Carlos Mattos is currently with Volvo do Brasil.

Eduardo Parente Ribeiro is a Professor at Electrical Engineering Department of Federal University of Parana, Brazil. He received Ph.D. degree in Electrical Engineering from Pontifícia Universidade Católica do Rio de Janeiro in 1996. He did research stage at Vanderbilt University in 1995 and a post-doctoral stage at The University of British Columbia in 2005. His interests include data communication, multimedia modeling and signal processing.

Evelio Fernandez received the B.Eng. degree in Electrical Engineering from the Central University of Las Villas (UCLV), Cuba, in 1985. He received the M.Sc. degree in electrical engineering and the Ph.D. degree in electrical engineering from the State University of Campinas, Brazil in 1997 and 2001 respectively. He is currently an associate professor at the department of Electrical Engineering at the Federal University of Parana. His current research interests include channel coding techniques, digital communications and wireless networks.

Carlos Marcelo Pedroso received the B. Eng. degree in Computer Engineering from Pontifical Catholic University of Parana in 1994 and Ph.D. degree in Electrical Engineering from Federal Technological University of Paraná, in 2006. He is currently at the department of Electrical Engineering at Federal University of Parana. His interests include data communication, modelling and performance evaluation, multimedia systems and Internet technologies.