Traffic Analysis for Voice in Wireless IP Networks

TONI JANEVSKI BORIS SPASENOVSKI
Department of Telecommunications, Faculty of Electrical Engineering
University “Sv. Kiril i Metodij”
Karpos 2 bb, 1000 Skopje

Abstract: - In this paper we provide an extensive analysis of voice service in wireless IP networks. Voice traffic is serviced with priority over the rest of the traffic. Hence, we concentrate on the voice traffic in the analysis. We model the packetized voice traffic with Markov Modulated Poisson Process (MMPP), and we propose an analytical framework for its analysis in wireless networks. Also, we created a simulator in Matlab, with capability for QoS analysis in different network scenarios, considering user call intensity, voice-encoding rate, link capacity and buffer sizes. The observed QoS parameters are packet loss and delay. Voice traffic is very sensitive to delay, while some low losses may be tolerated. We present overwhelming QoS analysis of IP telephony traffic at different network setups and give a concept for dimensioning wireless links for IP telephony under given constraints on the QoS parameters.

Key-Words: - Wireless, IP, Quality of Service, Traffic, Voice

1 Introduction

Voice service was mainly based on circuit-switched technology in the past, and it is still nowadays. However, development of computer industry and low cost of communication devices (palm top devices, communicators, mobile phones, lap-top computers etc.) moved the telecommunications beyond the voice service. Now, when the penetration of users in wireless cellular networks almost reached its maximum, telecommunication industry and telecom operators are facing with a challenge of extending the market by introducing more additional services besides the traditional voice service. These services that are offered now or those that will be offered in near future, have no other alternative than IP technology and Internet. In such scenario, voice will be just one of the many services offered to the end users. However, it will remain the most used one and the oldest one. Furthermore, users are used to specific quality of the voice service that came from circuit-switched networks, and they will not accept noticeable degradation of that grade of service. So, the question is how to design cellular access networks based on IP that will provide desired Quality of Service (QoS) for voice service, something to what users are already used to.

Let consider the Internet side first. Many researches and committees are addressing QoS issues in Internet. However, main considered issues are scheduling, differentiation and admission. However, not much is done considering the design of IP network carrying real-time traffic. One may justify this by claiming that if enough bandwidth is provided, then we should not worry about the capacity and the traffic. But, even in the wire Internet voice service end-to-end is still a problem due to main concept used at the development of Internet i.e. the control is left to the end entities of the communication (e.g., to the client and the server). In [1] authors provide analysis of real-time voice service through the wire access network. They have verified the MMPP model with measurements in a laboratory test-bed.

On the side of wireless cellular networks, we have provision of certain Grade of Service (GoS) by designing cellular networks with given constraints on call-level QoS parameters: new call blocking probability and call dropping probability. However, 1G and 2G wireless networks were based on circuit switching, where for each call network allocates one channel. There, adapted Erlang-B formula was used for the design of cellular network, [2] [3]. But, here we consider packet-based allocation of wireless link.

The paper is organized as follows. In Section 2 we describe the traffic and system model for our analysis. The analytical analyses are given in Section 3. Next section shows the results from simulation runs. Finally, conclusions are given in the last section.

2 Wireless Link Model

We perform design of the wireless networks by considering the capacity per cell (the bandwidth) as well as mobility of the users that results in handovers among the adjacent cells. The wireless link is based on IP. Packets from different sources are multiplexed on the wireless media, which is shared among all traffic streams.

We assume that voice over IP traffic is differentiated from data traffic, which is based on TCP. If IP telephony traffic is mixed with TCP traffic, which is long-range dependent, then it will add unmanageable packet delays and packet loss. In our previous work [4] we have proposed classification of IP traffic into two main classes: class-A, for traffic with QoS constraints, and class-B, for best-effort traffic. We further divide class-A into three subclasses: A1 for real-time traffic with constant bit-rate (IP telephony), A2 for real-time traffic with variable bit rate (e.g. video streaming), and A3 for best-effort traffic with a constraint on packet delay (e.g. browsing).

Today, there do exist mechanisms to differentiate traffic, such as differentiated services models [5]. We assume that IP telephony is differentiated from other traffic on the wireless link, and it is not mixed with TCP traffic. Packets from IP telephony are buffered into separate buffers (of course, there are also other mechanisms to bound packet delay or loss). However, we use priority scheme to differentiate IP voice traffic from the rest.

Considering the voice service in a given cell, we may have active users (with ongoing voice connections) and idle users. Traffic from all active users within the cell is multiplexed onto wireless link. We illustrate such scenario in Figure 1.

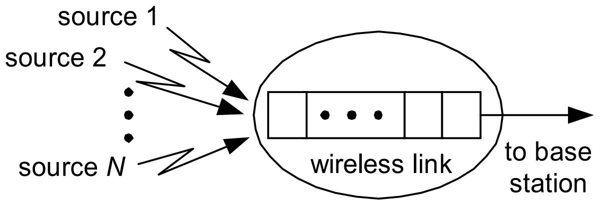

Figure 1. Multiplexing voice sources over wireless link

The design issue is how many subscribers we can serve with given network resources, or how much bandwidth we need to have service for predicted number of users of the voice service.

There are two different level of consideration:

1. Connection level; and
2. Packet level.

On a connection level voice service is well modeled by using Poisson process, i.e. exponential distribution of inter-arrival times between consecutive arrivals as well as exponential distribution of call duration time. Poisson process is well proven for voice service through the usage of Erlang-B formula, which is in fact M/M/c/c queue, where *c* is the number of channels (servers) on the link.

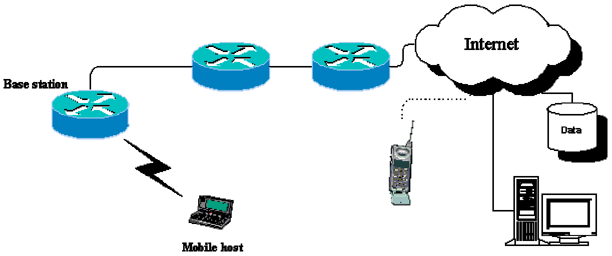

Figure 2. Wireless IP network architecture

On packet level we primarily consider packets from voice telephony streams. However, different traffic types are multiplexed on the wireless link, as shown in a typical wireless IP network architecture given in Figure 2. When a connection is established for a user, we may have a talk period (when the user is talking and packets are generated onto the link) and an idle period. During talk periods IP packets are transmitted back-to-back. We assume that there is no packet generation in idle periods.

We do not discuss the medium access layer and physical layer. But, we suppose that these layers provide medium access solutions to avoid collisions from different sources when transmitting packets, based on division of time into time steps, i.e. semi-permanent logical channels. Because we model the sources as on-off sources, we assume that during on period each source has dedicated time interval for packet transmission over the wireless link.

3 Traffic model

A. Call level properties

Let consider a single cell in the network. We assume Poisson arrival processes in the cell, which is well proven for the voice connections. With λ_n and λ_h we denote new call and handover arrival rate in the cell. An ongoing voice connection completes at rate μ_c or departs the cell at rate μ_h. But, in contrary to the circuit-switched cellular networks we do not have a limited number of channels in the networks. We do use statistical multiplexing of voice sources and other non-voice traffic over the wireless link. If we denote with λ the total call arrival rate, then:

$$\lambda = \lambda_n + \lambda_h \qquad (1)$$

Similarly, we denote with μ the total call departure rate:

$$\mu = \mu_c + \mu_h \qquad (2)$$

Then we obtain a birth-death process as shown in Figure 3, where *N* is the maximum number of active users in the cell.

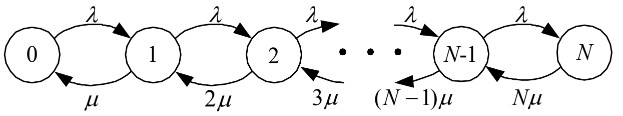

Figure 3. Markov chain model

So, a user may be an active voice source or idle considering voice service (in this case are included situations when users are using non-voice services).

B. Single voice-source properties

As for traffic models, voice connections arrive according to a Poisson process. Once a connection (or call) is established, the voice source is modeled as two-state Markov chain with one state representing the talk spurt (ON) and other state representing the silent period (OFF) [6]. A simple ON-OFF model accurately models the behavior of a single voice source. During ON (talk) periods the source is transmitting IP packets. Most encoding schemes have fixed bit rate and fixed packetization delay. During off (silence) periods the source sends no packets. We assume that ON and OFF periods are exponentially distributed, which is well analyzed in [1]. The voice sources can be viewed as two state birth-death processes with birth rate α_{on} (arrival rate for on-periods) and death rate α_{off} (ending rate for on-periods). Then, $1/\alpha_{on}$ and $1/\alpha_{off}$ are durations of talk period and silent-period of a voice source, respectively. Average of talk spurt duration is 0.352 s, and average of silent state duration is 0.650 s. The talk spurt has a constant bit rate that we denote as b_{ts}. Then, average voice connection bi rate is:

$$b_{av} = \frac{T_{on}}{T_{on} + T_{off}} b_{ts} = \frac{\alpha_{off}}{\alpha_{on} + \alpha_{off}} b_{ts} \qquad (3)$$

For example, if talk spurt is encoded with 32 kbps, then average bit rate is 11.24 kpbs. Considering the voice packet sizes there are also different decoders, i.e. in [1] are used encoders with 64 bytes payload and 40 bytes IP overhead, while in [7] authors use voice stream of 180 byte packets, with 160 bytes payload and 20 bytes IP overhead. Considering the fact that IP telephony is based on RTP/UDP/IP protocol stack, the accurate model should include 20 bytes overhead for IP, and 20 bytes overhead for RTP/UDP (12 bytes for RTP and 8 bytes UDP [1]).

Under the assumption of fixed packet size for voice packets, we may divide the ON-period into timesteps with duration T_p, where T_p is time that voice coder needs to generate one voice packet with length l_p. Then, one may write:

$$T_{on} = \frac{1}{\alpha_{on}} = T_p N_p \qquad (3)$$

where N_p is the number of IP packets during the ON-period when the packets are sent onto the wireless link back to back. So, during the talk spurt packets are generated at rate b_{ts}:

$$b_{ts} = \frac{l_p}{T_p} \qquad (5)$$

The probability that a user is in ON-period is given by:

$$P_{user\_on} = P_{call\_on} P_{source\_on} \qquad (6)$$

where:

$$P_{call\_on} = \frac{\lambda_1}{\mu} \qquad (7)$$

$$P_{source\_on} = \frac{\alpha_{off}}{(\alpha_{on} + \alpha_{off})} \qquad (8)$$

where λ_1 is average call arrival rate for a single source. Considering Figure 3, we may write:

$$\lambda = N\lambda_1 \qquad (9)$$

The probability that a source is OFF is P_{off}=1-P_{on}.

C. Superposition of voice sources

The superposition of the voice sources can be also viewed as birth-death process, where total incoming rate is sum of incoming rates of individual sources. A convenient model in teletraffic theory for a superposition of many on-off voice sources is Markov Modulated Poisson Process (MMPP). For voice sources with talk spurts and silent periods (without packets on link) it is more convenient to use the special case of MMPP, i.e. IPP (Interrupted Poisson Process). When the process is in state j that means j sources are on. Below we show transition-state diagram for superposition of k active voice sources.

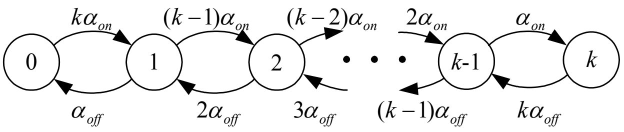

Figure 4. State transition diagram for a superposition of k voice sources (established voice connections)

In Figure 4 we assume that all k connections are previously established. Now, we may link this chain with Markov chain for calls given in Figure 3. To provide desired Quality of Service (QoS) for the voice calls we need to determine maximum number of active voice calls at a time. This number is denoted with N in Figure 3. So, at a given moment of time we may have k voice connections, where $k \leq n$. If we have k active connections then we have the situation given in Figure 4 considering ON and OFF periods of sources. IP packets arrive from k input lines/sources. Each of the k calls occupying the input lines i.e. wireless link alternates between talk spurts and silent periods. Assuming state-equilibrium in Figure 4, we may derive the distribution of active voice sources, i.e. the probability that j sources are ON out of k active voice connections is:

$$b_j = \frac{\binom{k}{j}\left(\frac{\alpha_{off}}{\alpha_{on}}\right)^j}{\left(1+\frac{\alpha_{off}}{\alpha_{on}}\right)^k} \tag{10}$$

Superposition of many Poisson processes is also a Poisson process. So, for the wireless link we may consider one Poisson process where arrival rate for ON periods is a sum of ON arrival rates of individual sources. One may notice that the same statement is valid for call arrival process from users. Considering the assumption for exponential distribution of call arrivals and call durations, as well as exponential distribution of ON and OFF periods during one call, and by using (5) and (6), one may write:

$$P\{\text{at least one source is ON}\} = \frac{\lambda}{\mu} \frac{\alpha_{off}}{(\alpha_{on} + \alpha_{off})} \tag{11}$$

4 Simulation analysis

We implemented our model in Matlab programming environment. As input parameters we use number of IP voice sources, bandwidth of the wireless link, buffer length in the network nodes, voice encoding rate, call arrival and departure rates. We observe packet loss and mean packet delay.

Various voice coders for IP networks exist on the market today. All of them are using RTP/UDP/IP protocol stack. For the input parameter values, considering the simulator setup, we used the same values as found in [1], i.e., 32 kb/s ADPCM voice encoding with 16 ms packet inter-arrival time, which results in 64 bytes of voice payload per packet, a protocol overhead of 40 bytes (12 bytes for RTP, 8 for UDP and 20 bytes for IP). Link headers are not included. So, the total packet size is 104 bytes. We use fixed size for all voice packets. The duration of talk-spurts and silence periods is exponentially distributed on the positive integers with a mean of 0.351s for ON-periods and 0.650s for the OFF-periods.

The results from the simulations are divided into two main groups. The first group considers voice sources that are permanently active (during the whole simulation period). This assumption does not reflect the real situations in the network. However, it is useful for calculating the limits of the network capabilities. The number of sources, the number of buffers and the bandwidth of the output link are the input parameters while the average packet loss and the average packet delay are the output parameters of the simulation process.

Figures 5 and 6 show the average packet loss and the average packet delay as a function of the number of buffers (measured in packets). The number of sources is used as a parameter. The bandwidth of the output link is 2 Mb/s and the voice sources have bit rates of 52 kb/s in ON periods. The simulation period is 1 hour. We can see from these graphs that the average packet loss, regardless of the number of sources, rapidly decreases until certain buffer length. For 40 and 80 voice sources, the number of buffers that is of particular importance is 20 and 40 buffers, respectively. Further enlargement of buffer length does not lead to better results, and the average packet delay increases with increasing the number of buffers. This means that the previous buffer length gives optimal results, regarding the average packet loss and average packet delay. When the network link is loaded with 110 voice sources, after reaching buffer length of 40 packets, the average packet loss stays pretty much the same while the average packet delay increases linearly. According to the recommended QoS parameters [8] and considering the results of the simulation runs, the optimal buffer length for this particular situation is around 45.

With proper analysis of the simulation results, we can give certain guarantees for the quality of the voice services that are being offered by the providers to their consumers. Also, we can determine the margins of the consumers that can receive adequate quality regarding the network infrastructure.

The second set of results reflects the behavior of the real systems. This means that the activity of the voice sources is modeled considering statistical analysis of the real voice traffic in the packet-based networks. We consider the performance model of a single cell in a cellular wireless communication network. We use Poisson arrival stream of new calls at the rate λ_n and the Poisson stream of handover arrivals at the rate λ_h. An ongoing call (new or handover) completes service at the rate μ_c and the mobile engaged in the call departs the cell at the rate μ_d. These rates are used as input parameters in the simulation. The bandwidth of the output link, the number of the buffers and the sources bit rates are also used as input parameters. The output parameters are the same as in the previous set of results.

Our main task was to create a set of graphs that can be used for dimensioning wireless cellular networks for IP telephony. Figures 7 and 8 show the average packet loss and the average packet delay as a function of the number of total calls (new and handover). The number of buffers is used as a parameter. The bandwidth of the output link is 2 Mb/s and the voice sources have bit rates of 52 kb/s in active periods. From these graphs we can see that buffer length of 20 packets is completely unusable. Also, we can determine the range of the input voice traffic that is eligible to certain, desirable quality.

After we determined this range, in order to find the optimal buffer length, we made two more graphs, i.e. Figures 9 and 10, which show the average packet loss and the average packet delay as a function of the number of buffers. The number of total calls is used as a parameter. With proper analysis of these graphs, according to the required quality of service, we can easy pinpoint the most effective buffer length that can be used in the particular network. Usually we have given constraints on both parameters, i.e. packet loss and delay. While real-time applications (such as IP telephony) have stringent demands on packet delay, they can tolerate reasonable packet loss (e.g., losses < 1%). We should choose the buffer size as a balance to either of the two constraints. Of course, one should apply admission control mechanism that will satisfy the constraints on loss and delay, i.e. reject the incoming calls to provide the desired quality to the ongoing calls.

Figures 11 and 12 show the average packet loss and the average packet delay as a function of the voice source's bit rate (this means different voice-encoding techniques). The number of buffers is 40 and the link is loaded with 1.0 calls/s (the bandwidth of the output link is 1.5 Mb/s). We may notice that the average packet loss decreases up to 90 % going from 52 kb/s to 28 kb/s rate of the voice-encoding scheme, as given in Figure 11. Also, the average packet delay also decreases significantly, Figure 12. According to these results we may conclude that by reducing the encoding bit rate we can either increase the quality of the offered services (the network QoS) for the same number of consumers, or increase the number of the consumers that will be able to receive the previous quality of service.

5 Conclusions

In this paper we analyzed the QoS for voice service over the wireless IP networks. However, the analysis may be extended to the wired ones. We assumed that voice is serviced with priority over other services on the link.

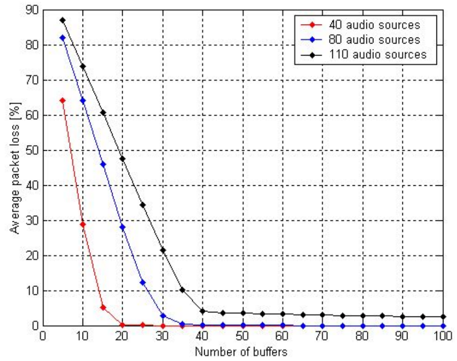

Figure 5 Average packet loss vs. number of buffers for different number of sources

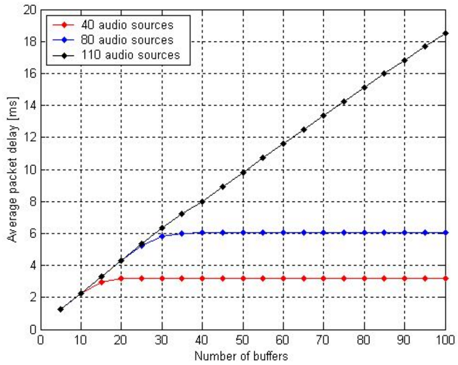

Figure 6 Average packet delay vs. number of buffers for different number of sources

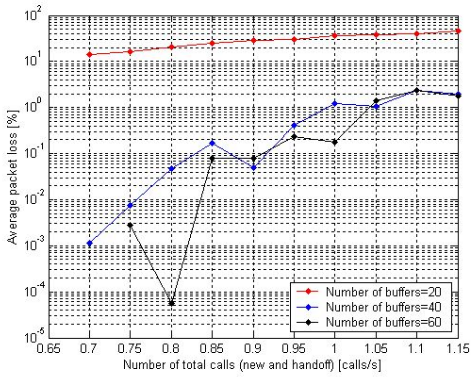

Figure 7 Average packet loss vs. call intensity for different buffer sizes

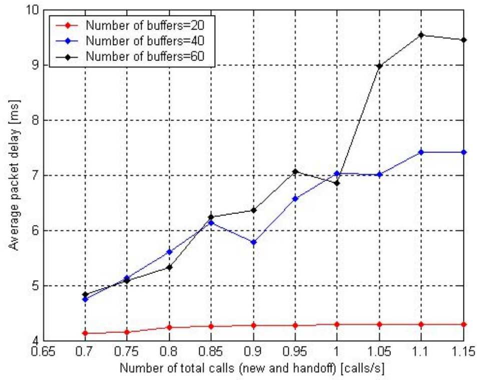

Figure 8 Figure 7 Average packet delay vs. call intensity for different buffer sizes

One model that provides such differentiation of voice traffic is differentiated services model. We presented an analytical framework for the analysis of the voice service in wireless IP networks. We used ON-OFF MMPP model, which is well proven for modeling packetized voice traffic [1].

We implemented our model in Matlab. Using the simulator we performed analyses of packet loss behavior and delay for different traffic and network scenarios, i.e. by using different input parameters, such as call intensities of the users, number of voice sources multiplexed on the link, voice-encoding rates, link capacity, buffer sizes, to obtain their influence on the QoS parameters.

Considering the results shown in the previous Section, this model can be applied in several different situations. One typical scenario is when we have a given bandwidth and we need to provide certain QoS to the users, according to the delay requirements specified in [8] as well as a given constraint on the packet losses. In such case, we may use the simulation results for access network dimensioning wireless links for IP telephony, or, as an input to an admission control algorithm.

References

[1] B. Ahlgren at el., “Dimensioning Links for IP Telephony”, *SICS, CNA Laboratory*, Sweden.

[2] L. Gavrilovska, T. Janevski, “Modeling Techniques for Mobile Communications Systems”, *GLOBECOM’98*, Sydney, Australia, November 1998.

[3] G. Harring, “Loss Formulas and Their Application to Optimization for Cellular Networks”, *IEEE Transactions on Vehicular Technology*, Vol. 50, No. 3, May 2001.

[4] T. Janevski, B. Spasenovski, “QoS Provisioning for Wireless IP Networks with Multiple Classes Through Flexible Fair Queuing”, *GLOBECOM 2000*, San Francisco, USA, November 2000.

[5] C. Dovropolis et al, “Proportional Differentiated Services: Delay Differentiation and Packet Scheduling”, *SIGCOMM*, 1999.

[6] M. Schwartz, “Broadband Integrated Networks”, Prentice Hall, 1996.

[7] D. Cavendish, A. Dubrovsky, M. Gerla, G. Reali, S.S. Lee, “Statistical Internet QoS Guarantees for IP Telephony”, UCLA CSD TR#990053.

[8] ITU-T, “One-way transmission time”, Recommendation G.114, Geneva, Switzerland, March 1993.

[9] T. Janevski, B. Spasenovski, “Admission Control for QoS Provisioning in Wireless IP Networks”, *European Wireless 2002*, February 2002.

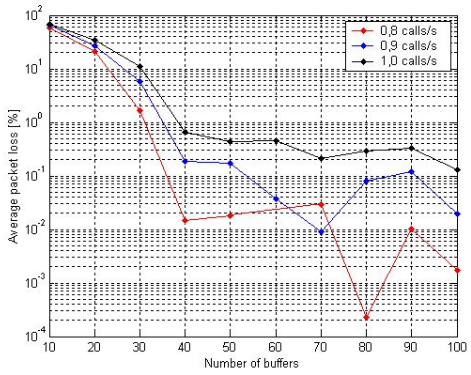

Figure 9 Average packet loss vs. buffer sizes for different call intensities

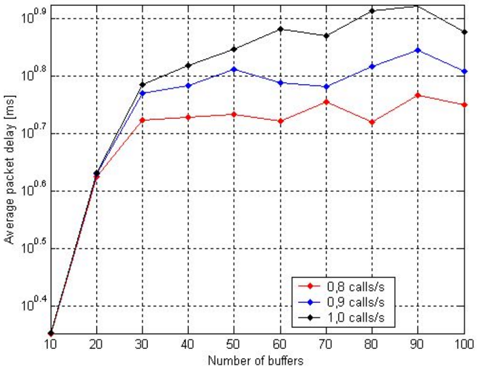

Figure 10 Average packet delay vs. buffer sizes for different call intensities

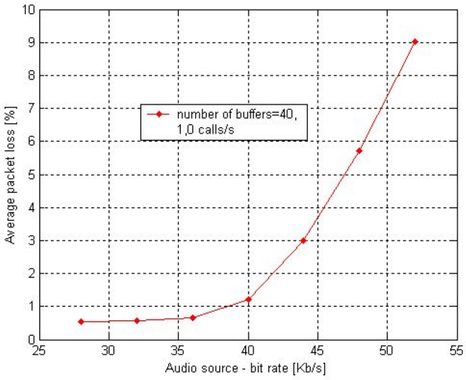

Figure 11 Average packet loss at different voice-encoding rates on 2 Mb/s link

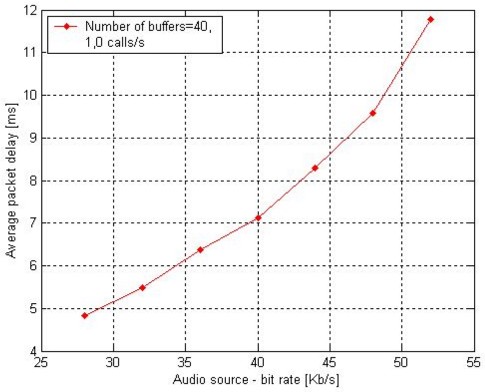

Figure 12 Average packet delay at different voice-encoding rates on 2 Mb/s link