Estimation of Voice over IP Quality in the Netherlands

X. Zhou[1], F. Muller[1], R. E. Kooij[1,2], and P. Van Mieghem[1]

[1] Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands
{X.Zhou, F.Muller, R.E.Kooij, P.VanMieghem}@ewi.tudelft.nl
[2] TNO ICT, P.O. Box 5050, 2600 GB Delft, The Netherlands
R.E.Kooij@telecom.tno.nl

Abstract. We have analyzed the measurements of end-to-end VoIP packets by tracing UDP packets between 12 testboxes in the Netherlands. We show that voice probes experience low delay and loss in the network. Moreover, we find that the reordering of packets has virtually no impact on the voice quality in our experiments. To determine the quality of VoIP calls over time, we have monitored the end-to-end packet delay, and packet loss over 10 days. The experimental results indicate that the networks in the Netherlands can continuously achieve a high VoIP quality.

1 Introduction

Voice over IP (VoIP) is becoming an increasingly popular and a cheap alternative to public switched telephone networks (PSTNs). Moreover, VoIP technology enables the integration of both data and voice traffic in the same network; it allows to easily introduce new multimedia services, and it supports more flexibility in terms of codecs. For example, PSTNs are bound to a single codec G.711, while VoIP can use any codec supported by both user terminals. However, due to the connectionless, packet-switched character of IP networks packets may experience different delay, may arrive at the destination out-of-order and may even get lost. All of the above factors (i.e. different delay, packet reordering, and packet loss) affect the perceived quality of voice calls.

The Internet is made up of a large number of separate networks that are interconnected at exchanges hubs. If a packet is sent from one network to another, it has to pass through one of those hubs. There are four high-speed hubs in The Netherlands[3], of which the Amsterdam Internet Exchange (AIX) is the biggest. Because interactive services such as VoIP are not only increasingly important but also pose stringent requirements to the network, assessing the performance of VoIP is an important issue. Many Dutch Internet operators offer services to

[3] The Amsterdam Internet Exchange (AIX) in Amsterdam, the Netherlands Internet Exchange (NL-IX), also in Amsterdam, the Groningen Internet Exchange (GN-IX) in Groningen and the Dutch-German Internet Exchange (ND-IX) in Enschede.

small and medium enterprises. They provide email, Internet access with firewall, Windows networking and backup services, as well as national VoIP.

This paper describes an assessment of VoIP quality in the Netherlands, and the network performance is measured for VoIP packets sent between 12 Internet testboxes (servers) of a Dutch ISP. Our observations are based on packet level traces collected throughout the network. The main aim lies in understanding to what extent today's Internet (in the Netherlands) meets the quality requirements for voice calls from the perspective of users.

Several researchers have worked on the measurement and assessment of VoIP quality over Internet. The closest to our work is the work of Marsh *et al.* [5], who measured the VoIP quality on an hourly basis by tracing a pre-recorded PCM coded call between nine sites in 2002, and compared the results with those obtained from a similar study in 1998. Their results showed that the best-effort Internet is sufficient for VoIP. Our work differs from [5] in terms of the experimental setup since we analyzed real network traces using much more different encoding schemes (up to 6). In addition, we also considered the impact of the playout buffer.

2 Prediction of the Voice Quality with E-Model

The E-Model [1] was used to estimate the subjective quality of voice calls. According to ITU-T Recommendation G.107, every rating R-value calculated from the E-Model corresponds to a Mean Opinion Score (MOS) value, as shown in Table 1, to predict subjective user reactions. An R-value above 70 corresponds to PSTN quality.

R-value rang	100>R>90	90>R>80	80>R>70	70>R>60	60>R>0
MOS	4.50-4.34	4.34-4.03	4.03-3.60	3.60-3.10	3.10-1.00
Speech quality	best	high	medium	low	very poor

Table 1. Speech transmission quality classes and corresponding R-value ranges

The mapping function from an R-value to a MOS value has the following form [11]:

$$MOS = 1 + 0.035R + 7 \times 10^{-6} R(R-60)(100-R) \tag{1}$$

where the output of the E-Model is the rating factor R:

$$R = (R_0 - I_s) - I_d - I_e + A \tag{2}$$

where R_0 is the effect of background and circuit noise, while I_s captures the effect of quantization. Both R_0 and I_s describe the transmitted voice signal itself and do not depend on the transport network. I_d is the impairment caused by one-way delay of the path, and I_e is the impairment caused by losses. A is the expectation factor. Based on recommended values in [1], the rating R can be defined by

$$R = 94.2 - I_d - I_e \tag{3}$$

where I_d has the following form:

$$I_d = 0.024d + 0.11(d - 177.3)H(d - 177.3) \quad (4)$$

where d is the one-way delay in milliseconds, and $H(x)$ is the Heavyside or step function where $H(x) = 0$ if $x < 0$ and 1 otherwise.

Unlike I_d, which only depends on the transport network and not on the codecs, I_e is codec dependent. The following form is presented in [2]:

$$I_e = a + b\ln(1 + cP/100) \quad (5)$$

where P is the packet loss rate in percentages, while a, b and c are fitting parameters for various codecs [9].

Parameters	G.711	G.729(10ms)	G.729(20ms)	G.723.1	iLBC
bitrate(kb/s)/framesize(ms)	64/20	8/10	8/20	6.3/30	15.2/20
a	0	10	10	15	10
b	30	25.21	25.21	36.59	19.8
c	15	15	20.2	6	29.7

Table 2. Parameters for different codecs (except for GSM)

The specific values of a, b and c for different codecs (except for GSM) are shown in Table 2. For G.711, it is assumed that Packet Loss Concealment has been implemented. The codec iLBC (internet Low Bitrate Codec) [10] is a free speech codec suitable for robust voice communication over IP networks. The parameter values for G.729 and G.723.1 are derived in [3][12], while the values for G.711 are derived in [11]. To calculate the a, b and c for iLBC, we extracted the iLBC MOS versus P from GlobalIPsound [10], then converted this relationship to I_e versus P via (1) and (2). The fitting model for the iLBC codec is shown in Fig. 1(a). Note that G.729 and iLBC have the same I_e values if there is no packet loss. Thus we take for a in iLBC the same a value as that of G.729.

For GSM (13 kbit/sec and 22.5 ms), ITU-T G.107 [1] and G.113 appendix I [4] is used. The corresponding formula[4] for I_e is:

$$I_e = 5 + 90\frac{P}{P + 10} \quad (6)$$

3 Experiment Results

The locations of the twelve testboxes have been chosen uniformly over the area of the Netherlands. They are mainly 2 or 3 hops away from the high speed backbone network, and their locations are shown in the map of Fig. 1(b). The sites were connected in a full mesh. The terminal clocks were synchronized using NTP software (with an accuracy of about ± 3 ms) every half an hour. Different encoding schemes were used. The packet sizes were calculated for different codecs.

[4] However, only values for GSM 6.60 Enhanced Full Rate (EFR) are given, which has a slightly lower bit rate than the simulated packet streams (12.2 kbit/sec instead of 13 kbit/sec) that were based on GSM 6.10. So there is a small inconsistency here.

Following ITU-T P.59 recommendation [7], a sequence of alternating voice signals and silence periods (without hangover time) was used as an input signal. No voice packets were generated during silence periods.

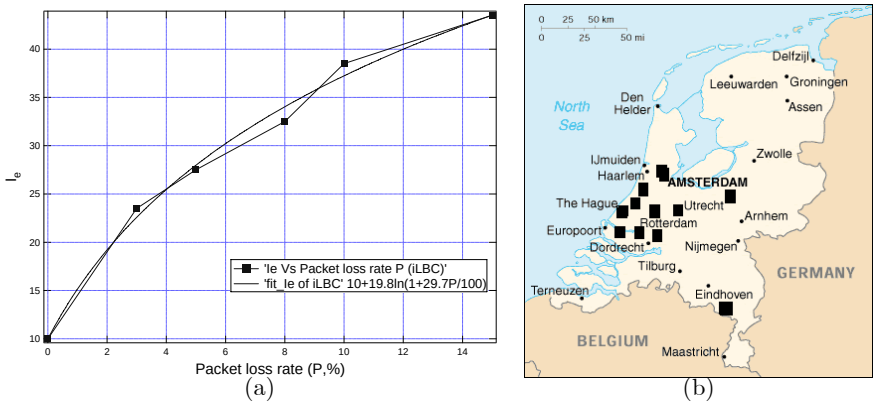

Fig. 1. (a) I_e vs. packet loss rate P for iLBC; (b) Locations of the test-boxes (black boxes) in the Netherlands

The 12 testboxes participated in two experiments. Firstly, during a 2 week period from Feb. 2, 2005 to Feb. 15, 2005, between each sender-destination pair of measurement boxes, packets generated in parallel with different codecs G.729, G.723.1 and GSM, are continuously transmitted from 7 AM to 9 PM (Local Time). Secondly, we repeated the same experiment with G.729, G.711 and iLBC packets over 10 days from June 3, 2005 to June 12, 2005. The difference between the two experiments may indicate how Internet packet dynamics change over time.

During the experiments, about 10 gigabytes experimental data (such as sending time, arrival time, and sequence numbers) were collected in a central point. A packet is classified as a reordered or out-of-order packet if it has a sequence number smaller than its predecessors. We examined each arrived packet by checking its arrival sequence order to calculate the total number of reordered packets.

We also executed traceroutes every 6 minutes during each test to determine the route taken during the tests. The resulting lists of intermediate routers of the paths were checked with the RIPE database [6]. The traceroutes provide some insight into the structure of the Internet in The Netherlands and they are useful to verify the changes in the delay during the measurements. Our results indicate that almost all traffic (99.2%) is routed through the AIX. Of the remaining 0.78% the routing is unclear. Only a few routes are very inefficient. Traffic on these routes is either routed through a router in London or through a router in Frankfurt (via the AIX).

3.1 Network Delay Performance

It is well-known that VoIP will not perform well if delays between the communicating parties exceed a certain QoS delay threshold (i.e. 150 ms). In this section, we will discuss the delay measured between the 12 testboxes. Fig. 2 summarizes the complementary cumulative distribution function (CCDF) of the median, average, 97.5 percentile and 99 percentile delays with different codecs in our experiments. Each data point corresponds to a pair of peers. The results of our experiments indicate that packets experience low delays. About 95% of all experimental pairs have median delays less than 40 ms, and the same holds for the average delays. About 95% of all experimental pairs have a 97.5 percentile delays less than 74 ms, and 95% of all experimental pairs have a 99 percentile delays less than 114 ms. Moreover, we also observe that those delay distributions do not vary significantly from codec to codec.

The end-to-end delay for voice over PSTN is less than 100 ms (reported by Bennett *et al.* [8]). Our results of the voice delays in the Netherlands are lower compared to the numbers reported for both voice over IP and voice over PSTN in [8].

We observe that few paths suffer from large delay, and this is mainly caused by a heavy load (with large packet loss) at the links connecting these pairs in the rush hours, and system updates in the testboxes.

Fig. 2 shows that the CCDFs of the delays D exhibit heavy tails: most individual pairs have a relatively small delay, but that large outliers are not uncommon. This suggests that networks in the Netherlands in 2005 can achieve a high performance. The heavy tail is fitted by a power law defined as $\Pr[D > x] \simeq cx^{-b}$, where the number b is the power law exponent (i.e. the slope in a log-log plot). Fig. 2 shows the exponents $2.45 \leq b \leq 4.68$ in the distributions of the medians delay, while $2.53 \leq b \leq 4.17$ in the average delays, $1.59 \leq b \leq 2.24$ in the 97.5 percentile delays, $b \approx 1.41$ in the 99 percentile delays.

3.2 Network Packet Loss Percentage

The packet loss percentage P is the percentage of unreceived packets in the data network. Unlike applications like email or ftp, which can simply request a retransmission when data is lost, VoIP just discards those voice samples that are lost or arrive too late. Packet loss results in a degradation of the conversational voice quality. According to industry standards, the maximum packet loss tolerable is about 3%.

Fig. 3 plots the CCDF of the percentage of lost packets. The CCDFs of the packet loss exhibit very heavy tails: This suggests that most (above 70%) individual pairs for different codecs have virtually no packets loss, while about 99.5% of all the pairs for different codecs have the percentage of packet loss less than 1%. We also observed that few paths suffered from large packet loss ($> 5\%$), and this is mainly caused by the system updates of the testboxes. The experimental results suggest that in our experiments, the packet loss is low enough to satisfy the industry standards.

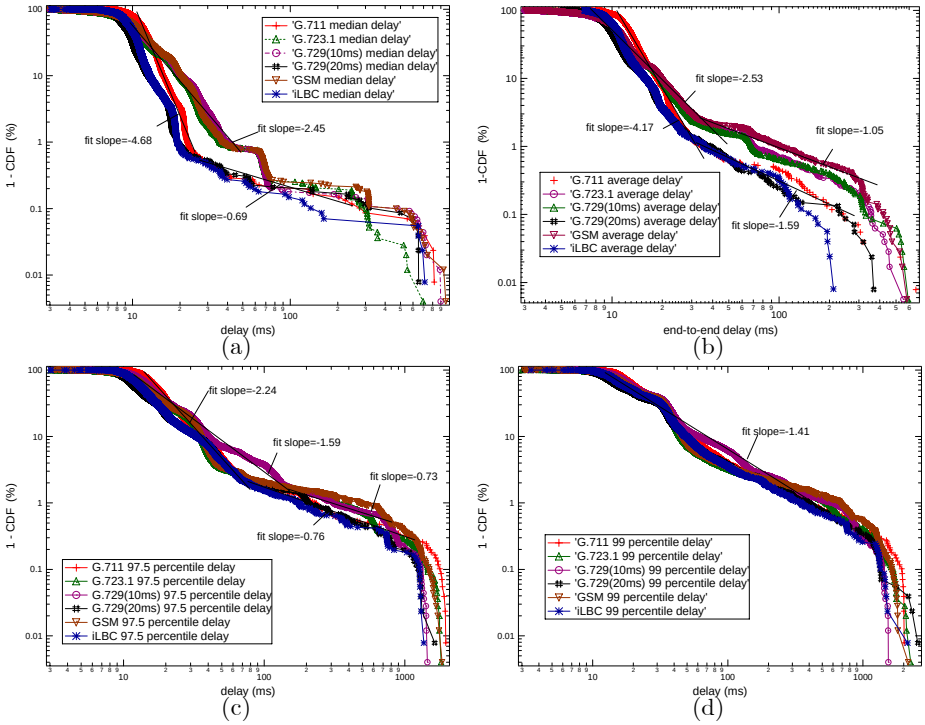

Fig. 2. The CCDFs of the (a) median, (b) average, (c) 97.5 percentile and (d) 99 percentile delays with different codecs in our experiments, and their corresponding power law fits.

3.3 Reordering Packets

Reordering of packets may impact the performance of applications on the Internet. In a TCP connection, the reordering of three or more packets within a flow may cause fast retransmission and fast recovery multiple times resulting in a reduced TCP window size and consequently in less throughput for the application. For delay-sensitive services which use UDP as transport protocol (such as VoIP or video conference), the ability to restore the order of packets at the destination has finite limits. The deployment of a real-time service necessitates certain reordering constraints to be met. For example, in case of VoIP, to maintain the high quality of voice, packets need to be received in order, and also within 150 ms. To verify whether these QoS requirements can be satisfied, knowledge about reordering in the Internet is desirable. To measure the number of reordered packets, for each source-destination pair with different codecs, we examine each arrived packet by checking its arrival sequence order, and calculate the total number of reordered packets for different codecs by summarizing the reordered packets for different codecs measurement.

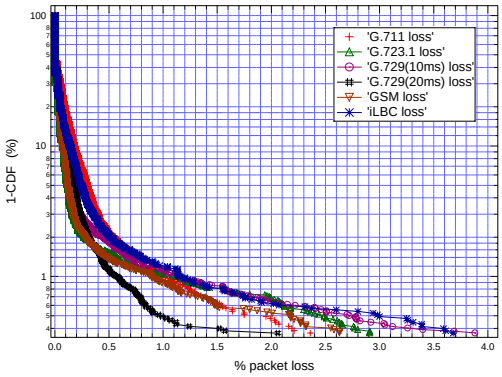

Fig. 3. The CCDF of the packet loss percentage.

codecs	G.711	G.723.1	G.729(10ms)	G.729(20ms)	GSM	iLBC
Total Nr. of reordering	3	2	2	4	7	1

Table 3. Total number of reordered packets received

Table 3 shows the total number of reordered packets observed for different codecs during the period of our experiments. Of all the packets sent successfully in our experiments, only very few (<0.0001%) are reordered. This result suggests that reordering is negligible.

3.4 Estimation of the Voice Quality

In order to apply the E-model to assess the perceived voice quality, we need to estimate the end-to-end delay and the packet loss. Both end-to-end delay and packet loss consist of a part induced by the network and a part originating in the VoIP terminals. The network performance has been discussed in the previous sections. The sending terminal contributes to the end-to-end delay through packetisation and coding delay. Typical values per codec can be found in [4]. The receiving terminal also adds delay through the operation of the playout buffer. The playout buffer is a buffer at the receiver side that compensates the effects of delay jitter by holding the first packet in a voice call for some time T before it is being decoded. The dejittering delay T adds to the end-to-end delay. In this paper we assume that T is fixed at $40ms$. Packets that arrive too late in the playout buffer to be decoded are considered lost. Hence, the playout buffer also contributes to the end-to-end packet loss. In our case the packet loss ratio induced by the playout buffer equals the ratio of packets that experience a network delay exceeding the minimum delay plus $40ms$. We now apply the E-model. Fig. 4 shows the voice call ratings and MOS values for different codecs in our two experiments. From Fig. 4 we can see that G.711 gives the highest call rating, followed by GSM, while the G723.1 gives the lowest call rating. The results for iLBC are almost as good as G.729. In general, the quality of calls in

different codecs is very high: with 99% of all calls experiencing a quality above 74 (3.7 in MOS). These results confirm that high VoIP quality can be achieved in the Netherlands. However, few paths achieved low MOS value ($MOS < 3.7$), this is due to high delay and loss.

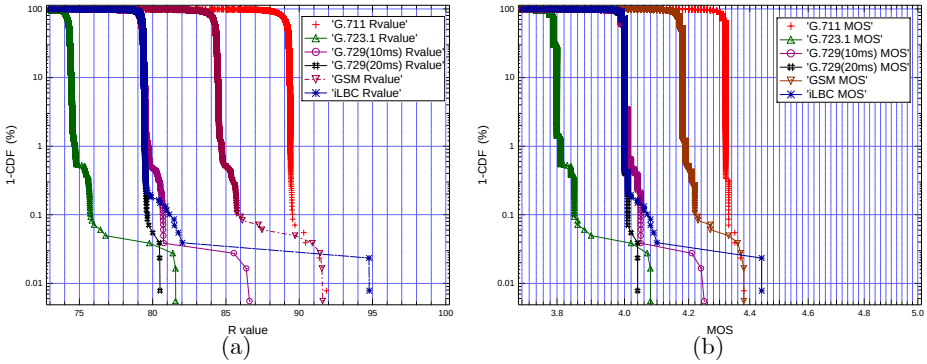

Fig. 4. The CCDF of call quality for different codecs

To determine the time varying quality of the calls, we calculated the average delay, average packet loss rate, R and MOS values for every 1.5 hours of daily experiment (thus 6 sample points per day).

Fig. 5 shows these values versus time. Fig. 5(a) shows that all the packets in different codecs had a very low end-to-end delay, and mainly in a range of 9-25 ms (below the noticeable 100-150 ms). There are higher delays in the rush hours compared with morning (7 AM-8 AM) and night (7:30 PM-9 PM). The corresponding traceroutes indicate that most of the source-destination pairs often followed fixed paths, indicating that the delay variation may be caused by queueing. Fig. 5(b) shows average packet loss versus time. The experimental results indicate that almost all the pairs experience consistently very small (or even no) loss ratios during our two experimental periods. Fig. 5(c) and Fig. 5(d) indicate that the current Internet in the Netherlands can continuously achieve satisfying results ($MOS \geq 3.7$). Our measurements suggest that the paths with low delay and loss can achieve an excellent MOS ($4 \leq MOS < 4.4$) at all times except for the rare cases when outages occur (*i.e.* system updates in the testboxes). We repeated the experiments by calculating the average delay, average packet loss rate, R and MOS values for a smaller time scale (1 minute voice of daily experiment) and observed similar results, which are not shown here.

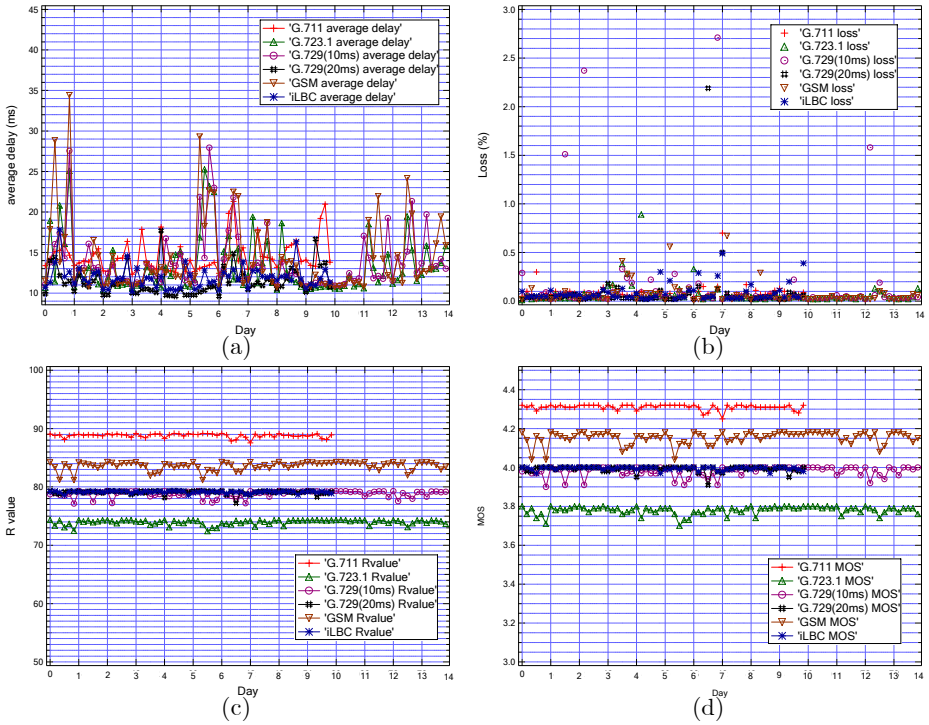

Fig. 5. Call quality statistics for different codecs for every 1.5 hours of each experimental day.

4 Conclusions and Future Work

In this paper, we have estimated the quality of VoIP as experienced by users by tracing UDP packets sent between 12 testboxes in the Netherlands. Our results lead to several observations with respect to network and VoIP in the Netherlands:

- Packet reordering hardly ever occurs.
- Paths can continuously achieve low delay and low packet loss.
- Networks can continuously support satisfying VoIP quality.

An important part of the future work will be to study the impact of perceived VoIP quality in a larger network environment. The RIPE TTM [13] infrastructure, which consists of about 100 testboxes located in different countries, can be used for this purpose. We also want to study the relation between the quality as experienced by users on different time scales, e.g. on an hourly basis and based upon measurements averaged over 1 minute. Moreover, the impact of different playout buffer schemes (like fixed playout and adaptive scheme) will also be a subject of the future work.

Acknowledgement: We thank L. Ding and C. Hoene for their useful discussions to identify the Equipment impairment I_e fitting function of codec iLBC. We thank H. de Bokx, D. Andriessen and R. van Meer from 1A First Alternative for their help with the data measurement.

References

1. ITU-T Recommendation G.107, "The E-Model, a computational model for use in transmission planning", March 2003.
2. R. G. Cole and J. Rosenbluth, "Voice over IP performance monitoring", Journal on Computer Communications Review, vol. 31, April 2001.
3. L. Ding and R. A. Goubran. "Speech quality prediction in voip using the extended e-model", IEEE GLOBECOM, December 2003.
4. ITU-T Recommendation G.113 appendix I, "Provisional Planning Values for the Equipment Impairment Factor Ie and Packet-loss Robustness Factor Bpl", 2002
5. I. Marsh, F. Li, and G. Karlsson, "Wide Area Measurements of Voice over IP Quality", QofIS 2003: 93-101
6. RIPE WHOIS Database, http://www.ripe.net/
7. Telephone Transmission Quality Objective Measuring Apparatus, Artificial Conversational Speech, ITU-T Recommendation P.59, 1993.
8. U. Varshney, A. P. Snow, M. McGivern, and C. Howard, "Voice over IP", Commun. ACM 45(1): 89-96 (2002)
9. ITU-T P.833, Methodology for derivation of equipment impairment factors from subjective listening-only tests, 2001
10. GlobalIPsound, http://www.ilbcfreeware.org/
11. R.G. Cole and J. H. Rosenbluth, "Voice over IP Performance Monitoring", SIGCOMM Comput. Commun. Rev., 31(2):9-24, 2001.
12. L. Ding, R. Goubran, "Assessment of effects of packet loss on speech quality in VoIP", Proc. of the 2nd IEEE HAVE 2003, Canada, September 2003, pp. 49-54
13. RIPE Test Traffic Measurement, http://www.ripe.net/ttm