Spectroscopy of the uplink Um interface of GPRS/GSM

Jorma Kilpi

VTT Information Technology

P.O.Box 1202, 02044 VTT, FINLAND

Email: Jorma.Kilpi@vtt.fi

20th February 2004

Contents

1 Network spectroscopy

Motivation for this study originated from the following paper of Broido & al.:

> **Radon spectroscopy of packet delay**, Andre Broido, Ryan King, Evi Nemeth and kc klaffy, *Providing Quality of Service in Heterogeneous Environments, Proceedings of the* 18^{th} *International Teletraffic Congress – ITC-18*, Berlin, Germany, 31 August - 5 September 2003

Broido & al. defined network spectroscopy as

> **"object identification on the basis of delay, period and frequency spectra".**

Analysis of packet interarrival time distribution is one of the application areas of network spectroscopy.

Broido & al.:

> **"Knowledge of realistic packet interarrival delay distributions is necessary for verifying statistical theories and models of Internet traffic."**

and

> **"Identifying origins and sources of delay quantization is part of our proposed agenda for network spectroscopy."**

2 Data analysis procedure

- Our interest is on the **uplink** direction of the radio interface.
- A **sessionwise** approach: traffic from a fixed mobile station (MS).
- Measurement from an (fast) interface which is **after** the (slow) radio interface.
- "Statistical inverse problem" and "Reverse engineering": from the measured data we try to recover what happened at the Um interface.

	Definition	**Duration (ms)**
Time Slot (TS)	Basic unit	$15/26 \approx 0.577$
TDMA Frame	$8 \times$ TS	$60/13 \approx 4.6$
Multiframe	$52 \times$ TDMA Frame	240
Superframe	$8 \times$ multiframe	1920

Table 1: *The TDMA time frame structure used by GPRS.*

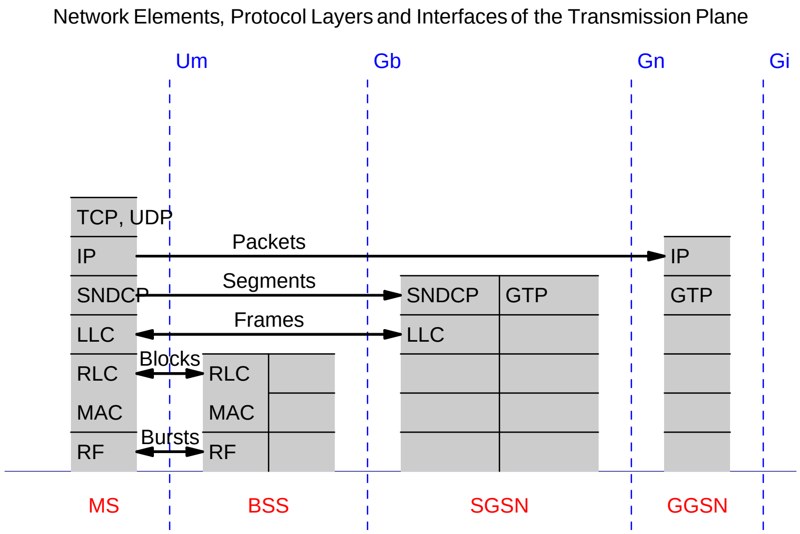

Figure 1: *Transmission plane.*

- The time slots of the TDMA frame available for GPRS use are called packet data channels (PDCHs). They can be dynamically and temporarily allocated to a MS for the transmission or receive of data.
- The physical connection between MS and BSS is called *a Temporary Block Flow (TBF)*. The TBF is a *unidirectional* concept and consist of allocation of one or more PDCHs and some number of blocks to be sent or received.
- Uplink TBFs are the main object of this study.

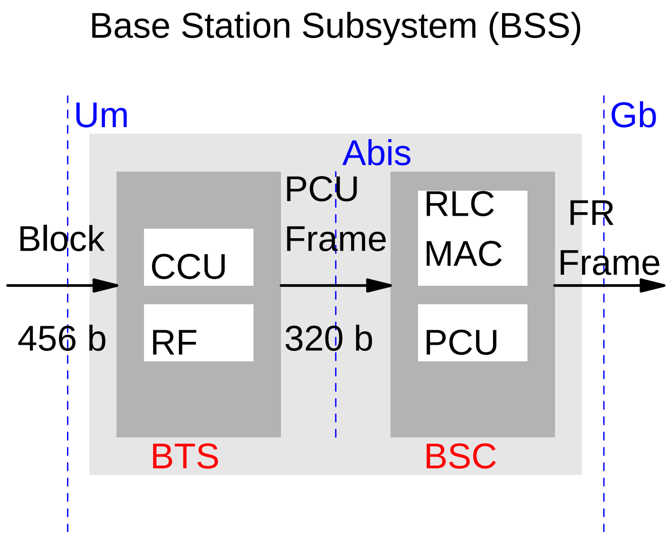

Figure 2: *The Abis interface between BTS and BSC. Each PDCH is mapped to one 16 kb/s channel, each (deconvoluted) block is put inside one PCU frame with constant size 320 bits. The transfer delay is 20 ms and one PCU frame is sent every 20 ms period.*

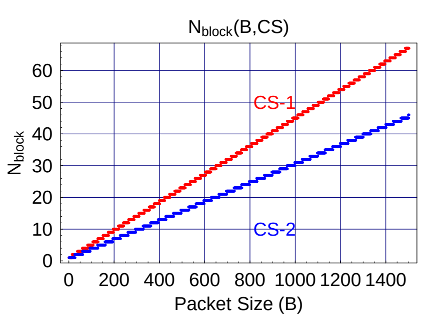

Figure 3: *Approximate number of blocks given the packet size.*

- We need to calculate the number of blocks that one packet of size B bytes most typically is assumed to generate, $N_{block}(B, CS)$.
- Here CS is a parameter with value $CS = 181$ in case of CS-1 and $CS = 268$ in case of CS-2.
- These are approximate numbers of user bits within one block.
- We approximate $N_{block}(B, CS)$ by formula

$$\left\lceil \frac{8 \times B + 30}{CS} \right\rceil . \qquad (1)$$

Here $\lceil x \rceil = \min\{n \mid n \geq x\}$.

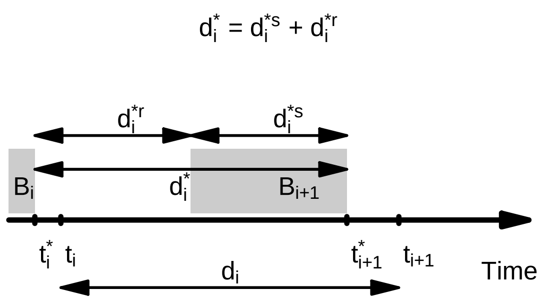

Figure 4: *The grey boxes refer to the "length" of the packet at the slow Um interface.*

- t_i^* refers to the time at the SGSN when the IP packet is completely reassembled.
- t_i refers to the time stamp at the fast interface.
- $d_i^* = t_{i+1}^* - t_i^*$
- $d_i = t_{i+1} - t_i$
- The data consist of pairs

$$(B_{i+1}, d_i),\ i = 1, \ldots, n-1.$$

- One of the basic assumptions is that the size dependent **distortion** defined by $D = d_i - d_i^*$ is small.

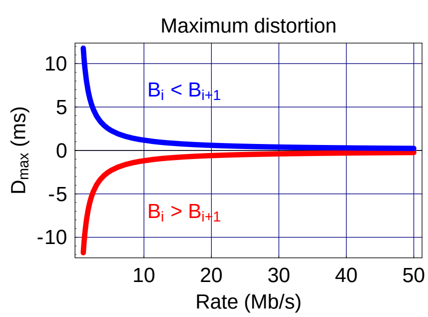

Figure 5: *Maximum distortions as a function of constant bit rate.*

- Distortion can be written as

$$D = (t_{i+1} - t_{i+1}^*) - (t_i - t_i^*)$$

- Maximum increase occurs, when $B_i = 28B$ and $B_{i+1} = 1500B$.
- Maximum decrease occurs, when $B_i = 1500B$ and $B_{i+1} = 28B$.
- This picture shows that size dependent distortion should not be a problem.

The more serious problem is **buffering**:

- At BTS there should not be buffering of RLC blocks: Erroneous blocks need not be stored and correctly received blocks are relayed to the BSC.
- Many BTSs are connected to one BSC. At BSC there must be some buffering capability of RLC blocks and LLC frames. Due to RLC layer retransmissions the RLC blocks may not come in order to BSC. Hence BSC must also have reordering capabilities of RLC blocks.
- Many BSSs are connected to one SGSN. The SGSN is router like machine and must have buffering capacity for LLC frames.
- Many SGSNs are connected to one GGSN. The GGSN acts as a gateway between mobile packet routing and fixed IP routing of the Internet. Hence it is assumed to have buffering capabilities like a router.
- Buffering at the measurement equipment before the time stamp is attached to the packet.

- The data that we will use for analysis consist of pairs

$$(B_{i+1}, d_i),\ i = 1, \ldots, n-1, \tag{2}$$

  made from the measured values (t_i, B_i), $i = 1, \ldots, n$, of a single session.
- We will analyze empirical probabilities

$$p(B, d) = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{1}_{\{(B,d)=(B_{i+1},d_i)\}}. \tag{3}$$

- Writing the packet interarrival delay as $d = d^s + d^r$ we interpret
  - d^s as the *size dependent* or *deterministic* part
  - d^r as the *residual* or *random* part

  of the interarrival delay d. Our interest is on the cases where d is *small*, otherwise d^s could be ignored. Then d is typically quantized to some discrete set of possible values.

	Packet Counts	Durations	Volumes (kB)	Mean Bit Rates (kb/s)	% of TCP Packets
Example 1	1155	19 min. 1 sec.	111.5	0.8	94.8%
Example 2	2259	65 min. 3 sec.	290.7	0.6	9.5%
Example 3	3360	10 min. 0 sec.	270.1	3.7	99.1%
Example 4	2525	29 min. 7 sec.	156.2	0.7	92.4%
Example 5	5529	29 min. 2 sec.	215.6	1.0	99.2%
Example 6	2808	11 min. 5 sec.	152.9	1.8	98.6%

Table 2: *Upstream characteristics of those sessions that are used as examples.*

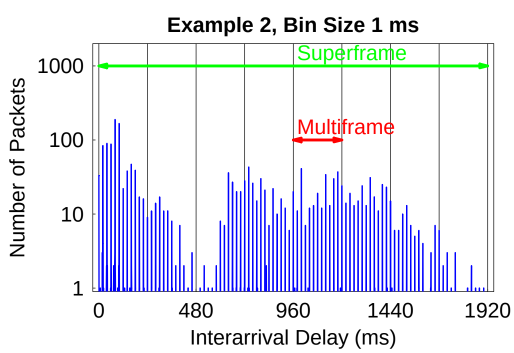

Figure 6: *Example 2.*

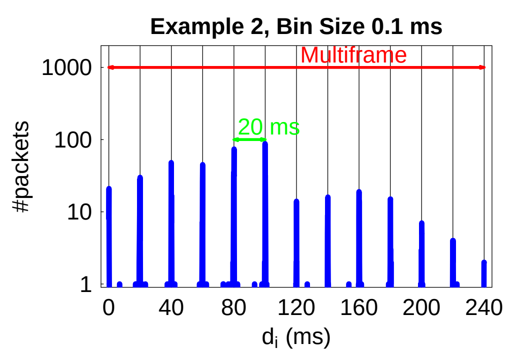

Figure 7: *The 20 ms delay quantization due to the Abis interface is rather exact.*

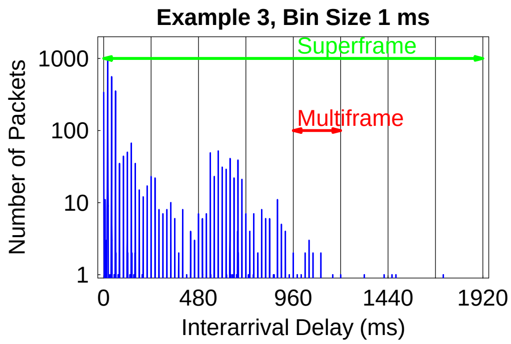

Figure 8: *Example 3*

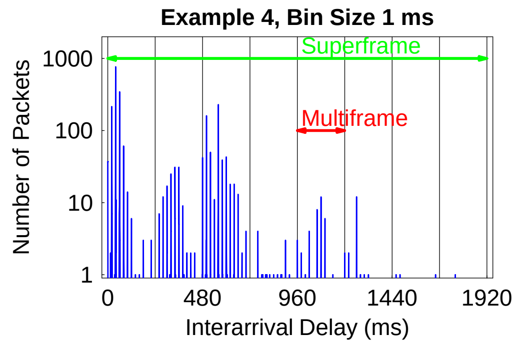

Figure 9: *Example 4*

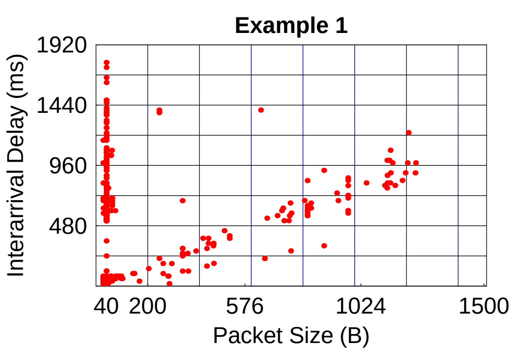

Figure 10: *The (B, d)-plane. At these points $p(B, d) > 0$.*

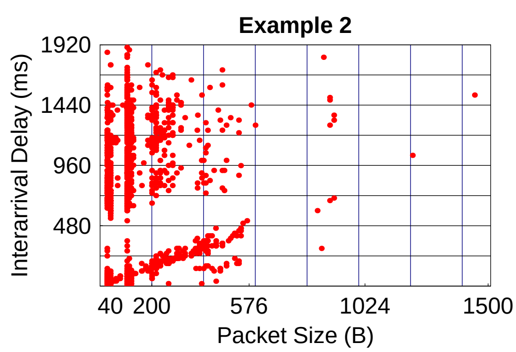

Figure 11: *Example 2 has similar but more clear structure.*

- The analysis continues like in Broido & al. by taking a suitable *Radon transform* of $p(B,d)$, which in our case is thus not used for the direct detection of the bit rate of the Um interface, but to the simultaneous detection of the transfer rate of Gb interface and the parameters CS and TS where TS is the number of uplink PDCHs used.
- The Radon transform in our case is defined as

$$p_R(r,v) = \sum_B p\left(B, r + \frac{18.5+20}{TS} N_{block}(B, CS) + \lceil (8\times B)/v \rceil\right) \tag{4}$$

- Like in Broido & al. we detect the value of the (deterministic) parameter v by looking the minimum of the Shannon-Kolmogorov entropy

$$-\sum_r p_R(r,v) \log p_R(r,v). \tag{5}$$

Here the sum is over all those values of r such that $p_R(r,v) \neq 0$.

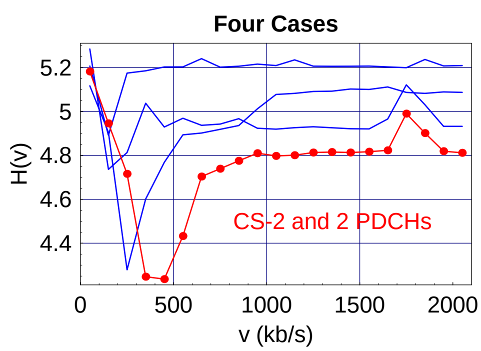

Figure 12: *From all 4 different alternatives CS-2 with 2 PDCHs is the most promising one.*

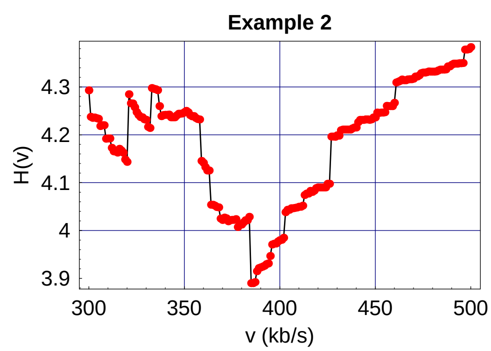

Figure 13: *A closer analysis with CS-2 and 2 PDCHs give the rate 384 kb/s, which is one of the rates offered by the Frame Relay.*

- Note that the use of formula like $(8 \times B)/v$ assumes a flow of bits with constant bit rate v, which in general is far from reality. It worked in our case since the real Gb interface was not congested due to other simultaneous sessions at the same time.
- We could have used the Radon transform like in Broido & al

$$p_R(r,t) = \sum_B p\left(B, r + \left(\frac{18.5+20}{TS}\right) \times N_{block}(B, CS) + t \times B\right) \tag{6}$$

  and choose t which minimize $H(t) = -\sum_r p_R(r,t) \log p_R(r,t)$.
- If we know the transfer rate of the physical Gb interface beforehand, then (6) could perhaps be used to detect whether the BSC or the Gb interface was congested or not.

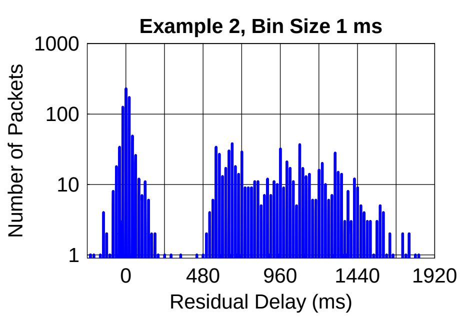

Figure 14: *Histogram of residual delays.*

- After removing d_i^s from d_i we get **residual** values

$$d_i^r = d_i - d_i^s.$$

- First block of each multiframe of one of the *downlink* PDCHs is reserved to signalling by default. If no other downlink signalling is used, it explains partially the gap between 240 ms and 480 ms.

3 Long TBFs

(TBF = Temporary Block Flow)

- If a TBF carries only of one packet, then we cannot say much about it.
- When we can infer that the TBF consists *at least* of 2 packets we call it a *long TBF*.
- The longer the TBF, *i.e.* the more packets it carries, the more we can say about it.

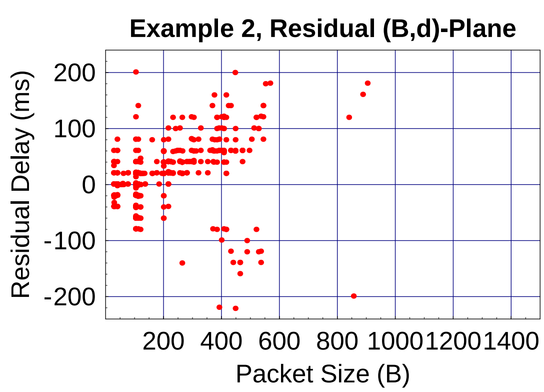

Figure 15: *Plot of* (B_{i+1}, d_i^r), $i = 1, \ldots, n-1$.

- How do we explain negative values?
- How do we explain horizontal positioning?

Theoretical example:

- Consider the case where 3 packets, A, B and C, are sent from an MS in this order and within one TBF (= long TBF).
- For simplicity we assume first that one block carries bits of one IP packet only.
- assume that the packet A is carried by three blocks whereas packets B and C are carried by two blocks.
- Assume that the second block that carry bits of user packet A, is not received correctly at BSS and retransmission is done *as soon as possible* when the MS gets a negative acknowledgement.

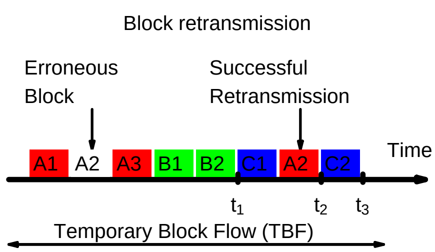

Figure 16: *Block retransmission within the same long TBF.*

- At the BSC the packets are reassembled and relayed to the SGSN in the order B, A and C.
- The data that would be measured in this case consist of two pairs

$$(A, t_2 - t_1) \text{ and } (C, t_3 - t_2).$$

- In this case d can be as small as the time of 1 radio block which, due to the Abis interface is always rather exactly 20 ms (unless the BSC, the Gb interface or SGSN is congested), hence
- d^r can be negative!
- Moreover, the order of packets may change. (LLC layer not in acknowledged mode)

- One block can carry bits of two different IP packets when the SNDCP segment consist of more than one packet or when 2 consecutive LLC frames are segmented simultaneously into blocks.
- None of these packets is forwarded before all blocks are successfully received, and when they are forwarded the measured packet interarrival time is almost zero.
- In this case the residual delay of the first packet is the same than its size dependent delay, but negative:
$$d^r = -d^s.$$
- Due to block (and frame) retransmissions the rationale of multiplexing many packets into one segment is more than questionable since it may delay several packets simultaneously and significantly. (Would the performance be improved if blocks of such muliplexed segment would be sent always using CS-1, like the signalling blocks?)

- In our theoretical example, given the time stamps and packet sizes we can infer that the length of the TBF has been at least the duration of 8 blocks (8 × 20 ms = 160 ms) instead of the optimal duration of 7 blocks (140 ms).
- Moreover, we can reconstruct the TBF up to some level: we know the positions of blocks $A2$, $B2$ and $C2$ and we know that $A1$ and $A3$ were before $A2$, $B1$ was before $B2$ and $C1$ was before $C2$.
- If we know that the original order of the transmitted packets, we could completely reconstruct our theoretical TBF.
- Reconstruction of measured long TBFs is also possible up to surprisingly high level.
- With only handful of examples we do not yet give any statistics about TBFs.

4 Conclusions and further work

- We have shown that it is possible to some extent to detect the behavior of the TDMA based radio interface of GPRS from the IP level packet data that is measured after the radio interface.
- More precisely, it is possible to extract TBFs, number of PDCHs and channel coding scheme used within long TBFs and detect retransmissions of blocks within long TBFs.
- Although we do not yet have any examples, it should also be possible (with optimal measurement) to detect whether multiplexing of users at the Um interface has been the case or whether the BSC or the real Gb interface has been congested.
- Of course, these are statistical results and valid only with some probability but we believe that this probability is rather high.

- Natural concept for further work is to study whether similar approach, namely the analysis of the empirical distribution $p(B, d)$, yields interesting results from WCDMA based UMTS and CSMA based WLANs.