Measuring the Quality of VoIP Traffic on a WiMAX Testbed

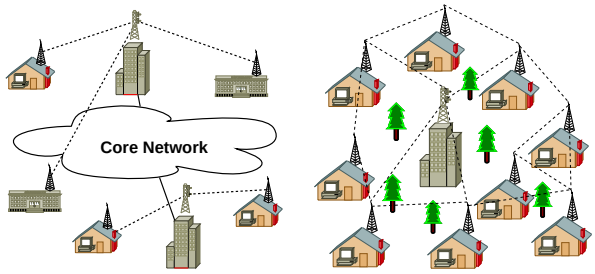
Nicola Scalabrino, Francesco De Pellegrini, Roberto Riggio,
Andrea Maestrini, Cristina Costa, and Imrich Chlamtac
CREATE-NET
Via dei Solteri 38, 38100, Trento, Italy
Email: name.surname@create-net.org

***Abstract*— The large radio coverage of the IEEE 802.16 standard, widely known as WiMAX, represents a key advantage compared to several first mile solutions proposed so far, while ensuring a rather inexpensive equipment at the subscriber side. The IEEE 802.16 standard, in practice, promises to be a flexible solution especially where cabling is not a viable choice, or as an alternative to customary leased lines. Nevertheless, modern requirements to wireless connectivity include mandatory QoS guarantees for a wide set of real-time applications, and this is the case of the ever growing trend of VoIP calls. To this aim, WiMAX supports natively real-time traffic. In this paper, we report the results of a set of measurements performed on the field on a WiMAX Alvarion testbed, located in Turin, Italy. We fed the system with synthetic VoIP traffic, real-time guaranteed, competing with concurrent best effort traffic. We obtained E-model figures, thus characterizing the operation intervals of the system, depending on the codec source and the number of calls.**

I. Introduction

Affordable high bandwidth connectivity and Internet access represents an irremissible feature of modern service provision. From the service provider perspective, though, there exist mainly two options. On one hand, several traditional first mile solutions leverage cables or fibers. But, wired solutions all pay an high entrance barrier, meaning that new network operators are usually banned from entering the market. On the other hand, the option is wireless connectivity.

The first mile service provision, in particular, is a known issue especially in rural or in remote areas: there, due to the low users density, a service provider would not gain enough return on investment from broadband connectivity. In such scenarios, in fact, Broadband Wireless Access (BWA) technologies represent the economically viable solution to Internet access. Their wireless architectures, in fact, make their deployment simpler and flexible compared to their wired counterparts. Thus, BWAs are believed a promising mean to provide Internet access to customers scattered over larger areas or in developing countries. Among BWA technologies, the IEEE 802.16 standard [**?**], promoted by the WiMAX (Worldwide Interoperability for Microwave Access) forum [1], is considered the leading technology for the provision of Internet-based broadband services in wide area networks.
A typical WiMAX deployment relies on a Point-to-MultiPoint (PMP) architecture, as depicted in Fig. 1(a): it consists of a single Base Station (BS), which interconnects several Subscriber Stations (SSs) to an Internet gateway. The IEEE 802.16 standard provides support for mesh-based architectures as well, as depicted in Fig. 1(b); even though WiMAX-based mesh deployments may contribute to the success of such technology in a more mature stage [2], the current standard provides only optional support to such architecture. Hence, in this work, the focus is devoted to the PMP configuration, as deployed in our testbed.



(a) Point-to-Multipoint (b) Mesh

Fig. 1. WiMAX network architectures.

In order to support the next-generation Internet, WiMAX will face current trends in broadband multimedia services. The critical point, there, will be the capability to support multimedia applications including VoIP, video streaming, video conferencing, online gaming, tele-reality, and so on. It is well understood that these applications pose strict constraints on throughput, packet loss and delay: to this aim, the IEEE 802.16 standard encompasses four different QoS classes and provides basic signaling between the Base Station (BS) and the Subscriber Stations (SSs), in order to support service requests/grants.

After an overview of the WiMAX technology [3], we will report on the results of a measurements campaign performed over a WiMAX-testbed. The testbed is deployed in Turin, Italy within the national experimentation on WiMAX coordinated by Fondazione Ugo Bordoni (FUB) [4]. The architecture of the testbed is sketched in Fig. 2. For the experimental setup, we adopted the Alvarion BreezeMAX [5] equipment, which operates in a 3.5 GHz licensed band, and is fully IEEE 802.16-2004 compliant. The measurements reported in this paper will

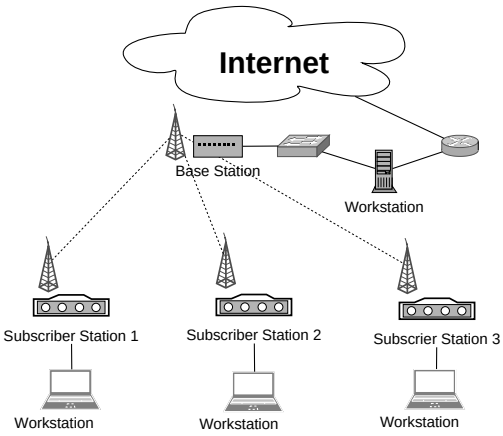


Fig. 2. The architecture of the WiMAX testbed located in Turin, Italy.

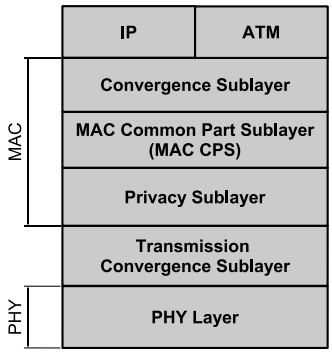


Fig. 3. The IEEE 802.16 protocol stack.

assess the ability of the current WiMAX technology to support VoIP flows. In particular, the voice quality has been evaluated through a computational scheme, namely the E-Model [6].

The remainder of this paper is organized as follows. In Sec.s II we briefly summarize the related work on the subject of performance evaluation of WiMAX networks. Section III gives an overview of the IEEE802.16 standard. In Sec. IV the Packet E-Model is introduced. In Sec. V we describe the experimental settings and the traffic patterns used for the testbed measurements. Section VI reports on the outcomes of the measurements on the WiMAX testbed under mixed VoIP calls and background traffic. In Section VII we discuss on some indications for future work. The last section contains final remarks.

II. Related Work

The performance evaluation of WiMAX systems is still in its infancy, and only a few authors studied the actual performances of the system. For example, the authors of [7] report the outcomes of numerical simulations assuming a WiMAX channel width of 5 MHz and a 2×2 MIMO system, showing that, under ideal channel conditions, data rates up to 18 Mb/s are feasible. In [8], a prototype simulator of a IEEE 802.16 protocol stack, including a sophisticated channel model has been developed. The simulator implemented the convergence sublayer, the MAC and PHY layer; the downlink and the uplink delay and the MAC throughput were evaluated.

In [9], the authors report the results of the performance evaluation of Internet access over BWA networks, using the NS2 simulator. The MAC layer functionality were developed in C/C++ and interfaced to NS and the probing traffic was represented by HTTP requests originating from a population of Web users.

The work in [10] analyzes the voice capacity delivered by IEEE 802.11 clusters connected to a IEEE 802.16 backhaul, showing that the capacity is limited by the WLAN bottleneck. The authors also propose a multiplex-multicast scheme to double the capacity by installing a multiplexer between the SS and the access point. For downstream VoIP traffic, the multiplexer combines multiple VoIP packets into a single multiplexed packet and the AP multicasts the multiplexed packet to the wireless end stations, showing a large multiplexing gain. Finally, in [11] the authors have assessed, via simulation, the performance of an IEEE 802.16 system using the class of latency-rate scheduling algorithms where a minimum reserved rate is the basic QoS parameter negotiated by a connection within a scheduling service.

The main contribution of our paper is an assessment of the VoIP quality supported by the IEEE 802.16 standard. We remark that, to the best of the authors' knowledge, currently there exist no related literature on testbed performance evaluation of IEEE 802.16.

III. WiMAX Overview

WiMAX is the commercial name of products compliant with the IEEE 802.16 standard. As in the case of IEEE 802.11 and Wi-Fi, an industrial organization, the WiMAX Forum has been set up to promote the adoption of such technology and to ensure interoperability among equipment of different vendors.

The protocol architecture of IEEE 802.16 is depicted in Fig. 3, and it comprises several sub-layers. The Transmission Convergence Sublayer (TCS) operates on top of the PHY layer and is specifically responsible to convert variable-length MAC PDUs into fixed length PHY blocks. The Common Part Sublayer (CPS) is responsible for the segmentation and the reassembly of MAC service data units (SDUs), and for scheduling and retransmission of MAC PDUs. It also supports the signaling mechanisms for system access, bandwidth allocation and connection maintenance. The Convergence Sublayer (CS) is placed above the MAC layer, to interface both to IP and ATM. Basic privacy support is provided at the MAC layer.

The IEEE 802.16 first release accounted a scenario with no mobility and operations in licensed frequency bands ranging between 10 and 66 GHz, with the mandatory use of directional antennas to obtain satisfactory performance figures. Later amendments to the standard (802.16a and 802.16-2004) extended IEEE 802.16 to non-line-of-sight applications in the $2 - 11$ GHz frequency band. Further amendments of the standard will encompass mobility (802.16e), multi-hopping (802.16f), handover and improved QoS (802.16g). Duplexing

is provided by means of either Time Division Duplexing (TDD) or Frequency Division Duplexing (FDD). In TDD, the frame is divided into two subframes, devoted to downlink and uplink, respectively. A Time-Division Multiple Access (TDMA) technique is used in the uplink subframe, the BS being in charge of assigning bandwidth to the SSs, while a Time Division Multiplexing (TDM) mechanism is employed in the downlink subframe. In FDD, uplink and downlink subframes overlap in time and are transmitted on separate carrier frequencies. Support for half-duplex FDD SSs is also provided, at the expense of some additional complexity. Each subframe is divided into physical slots. Each TDM or TDMA burst carries MAC Protocol Data Units (PDUs) containing data towards SSs or BS, respectively. Each SS learns the boundaries of its allocation within the current uplink subframe by decoding a UL-MAP message broadcasted by the BS at the beginning of each frame.

The IEEE 802.16 MAC protocol is connection-oriented and it is based on a centralized architecture, where each connection is uniquely identified by a 16-bit address. The core of the protocol is the bandwidth requests/grants management. A SS may request bandwidth, by means of a MAC message, to indicate to the BS that it needs (additional) upstream bandwidth. Bandwidth requests can be transmitted during the uplink subframe in either a dedicated contention period or in a contention-free period. Furthermore, bandwidth requests can be piggybacked during data packet transmission. We notice that bandwidth is granted per Subscriber Station: each individual SS, is then in charge of allocating the available resources to the currently active flows. The MAC CS provides three main functionalities:

1) *Classification*. The CS associates the traffic coming from upper layer with an appropriate *Service Flow* and *Connection Identifier*.
2) *Payload Header Suppression*. The CS may provide payload header suppression at the sending entity and reconstruction at the receiving entity.
3) *Delivery*. The resulting CS PDUs are delivered to the MAC Common Part Sublayer according to the negotiated QoS levels.

The standard defines two different CSs for mapping services to and from IEEE 802.16 MAC protocol, and regarding IP, the packets are classified and assigned to the MAC layer connections based on a set of matching criteria, including the IP source and the destination addresses, the IP protocol field, the Type-of-Service (TOS) or DiffServ Code Points (DSCP) fields for IPv4, and the Traffic Class field for IPv6.

Classified data packets are finally associated with the particular QoS level of the service flow they belong to. The QoS may be guaranteed by shaping, policing, and/or prioritizing the data packets at both the SS and BS ends. The BS allocates upstream bandwidth for a particular upstream service flow based on the parameters and service specifications of the corresponding service scheduling class negotiated during connection setup. The IEEE 802.16 standard defines four QoS service classes: Unsolicited Grant Service (UGS), Real-Time Polling Service (rtPS), Non-Real Time Polling Service (nrtPS) and Best Effort (BE) [11][3].

A. CIR and MIR

Two main parameters are used in order to support service differentiation at the higher layers: the *Committed Information Rate* (CIR) and the *Maximum Information Rate* (MIR), inherited from other existing technologies [12], [13]. Both parameters are set for a certain service class and regulate the aggregated downlink and uplink flows of a given SS connection.

The CIR parameter for a WiMAX system is the bitrate that the network agrees to accept from the user. In case of congestion, throughput reduction may occur below the CIR: thus, the word "committed" is by no mean a guarantee that the CIR will be met. A proper design of the user network, anyhow, should make this event quite rare[1]. Flows exceeding the CIR are vulnerable to packet discarding policies at the operator need: if the WiMAX network is congested, the BS will typically discard frames on connections exceeding the CIR before frames on connections that are within their CIR. Thus, the CIR provides a crude method for being fair when allocating limited capacity.

The second parameter, the MIR, regulates the maximum allowed peak rate of a connection. If the transmission rate exceeds the MIR, all the MAC frames violating the MIR will be discarded automatically; usually, the details on the BS discarding policies is proprietary to the hardware vendor.

B. Parameters Setting

MIR and CIR are specified for each SS according to the negotiated Service level Agreement (SLA); the compliance to the negotiated SLA is assessed over a reference window, called Committed Time (CT). In what follows we assume that n SSs make MIR and CIR requests to the BS. We let R_{max} the maximum traffic rate available at the WiMAX Downlink Air Interface, and denote CIR_k and MIR_k the request of the k-th SS[2], where $0 \leq \text{CIR}_k \leq \text{MIR}_k \leq R_{\text{max}}$.
The BS dynamically allocates the BE Service Rate R_{BE} (bit/s) and the Real Time (RT) Service Rate R_{RT} (bit/s) with a cumulative upper bound of R_{max}, making sure that the RT service traffic has a higher priority than the BE service traffic: $R_{\text{RT}} + R_{\text{BE}} \leq R_{\text{max}}$. The residual capacity is allocated as R_{BE}. Let N_{tot} be the total number of downstream service flows consisting of N_{VoIP} VoIP flows and N_{TCP} TCP persistent connection, so that $N_{\text{tot}} = N_{\text{TCP}} + N_{\text{VoIP}}$.
Let $R_{\text{TCP}}(m)$ be the service rate that the BS can provide to the m-th TCP service flow, the aggregated BE service rate is $R_{\text{BE}} = \sum_{m=1}^{N_{\text{TCP}}} R_{\text{TCP}}(m)$; similarly, if $R_{\text{VoIP}}(m)$ is the service rate that the BS provides to the m-th VoIP service flow, the

[1]If the customer has negotiated a Service Level Agreement with the service provider, the service provider should pay a penalty for missing a CIR commitment.

[2]The Alvarion BreezeMAX device does not allow to set the MIR parameter for real-time traffic

aggregated RT service rate becomes: $R_{\text{RT}} = \sum_{m=1}^{N_{\text{VoIP}}} R_{\text{VoIP}}(m)$. The Alvarion equipment used in the testbed provides resource allocation mechanisms corresponding to three cases.

In the first case, the downlink bandwidth is over provisioned, meaning that the aggregated traffic service rate for the WiMAX network is *deterministically* lower than R_{max}, i.e. $\sum_{m=1}^{N_{\text{TCP}}} \text{MIR}(m) + \sum_{n=1}^{N_{\text{VoIP}}} \text{MIR}(n) \leq R_{\text{max}}$, and no congestion occurs: the allocation in this case is fairly simple and the BS sets $R_{\text{VoIP}}(n) = \text{MIR}(n)$ and $R_{\text{TCP}}(m) = \text{MIR}(m)$.

The opposite case occurs when the aggregate of the CIR requested by VoIP subscribers exceeds R_{max}, i.e. $\sum_{n=1}^{N_{VoIP}} \text{CIR}(n) > R_{\text{max}}$: then the BS sets $R_{\text{VoIP}}(n) = \frac{R_{\text{max}}}{N_{VoIP}}$ and $R_{\text{TCP}}(m) = 0$ for every SS $n = 1, 2, \ldots, n$.

The remaining case is such that :

$$\sum_{m=1}^{N_{\text{TCP}}} \text{MIR}(m) + \sum_{n=1}^{N_{\text{VoIP}}} \text{MIR}(n) > R_{\text{max}};$$
$$\sum_{n=1}^{N_{\text{VoIP}}} \text{CIR}(n) \leq R_{\text{max}}. \quad (1)$$

This is the case when the BS guarantees the minimum service rate for the VoIP traffic and can reallocate the remaining bandwidth to the BE services, namely

$$R_{\text{VoIP}}(n) = \text{CIR}(n); \quad (2)$$
$$R_{\text{TCP}}(m) = \frac{(R_{\text{max}} - R_{\text{RT}})}{N_{\text{TCP}}}.$$

This is also the case that was considered for our measurement, since it is the probing case when QoS guarantees must be provided in spite of concurrent data traffic.

Notice that the actual implementation of the resource allocation depends on the scheduling implemented at the BS and vendors usually do not disclose such a critical detail to customers. Nevertheless, with appropriate probing, we could get some insight into the system behavior (see Sec. VI-A).

IV. Packet-E-Model

The quality of conversation in VoIP systems is traditionally assessed by mean of the Mean Opinion Score (MOS). MOS is a numerical measure and is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality, and 5 is the highest perceived quality. Being based on a listening test, evaluating the MOS rate for a VoIP solution can be a time consuming process. For this reason, we made our probes through synthetic traffic generation, and we resorted to the E-Model [6], which provides an objective method to evaluate speech quality in VoIP systems (for a thorough description see [14], [15]). The outcome of an E-Model evaluation is called R-factor (R). The R-factor is a numerical measure of voice quality, ranging from 0 to 100. The reference values of the R-factor are categorized as shown in Table I.

In the E-Model several different parameters affecting the quality of a conversation are taken into account. The main assumption is that various impairments at the physiological scale have an additive behavior (dB-like behavior),

TABLE I
R-factors, quality ratings and the associated MOS

R-factor	Quality of voice rating	MOS
$90 < R \leq 100$	Best	4.34 - 4.5
$80 < R \leq 90$	High	4.03 - 4.34
$70 < R \leq 80$	Medium	3.60 - 4.03
$60 < R \leq 70$	Low	3.10 - 3.60
$50 < R \leq 60$	Poor	2.58 - 3.10

TABLE II
The typical R factor values of some reference cases.

Scenario	R
PSTN/PSTN	82
ISDN/ISDN	92
PSTN/Mobile	64
VoIP	68

$$R = R_0 - I_s - I_d - I_e + A. \quad (3)$$

In particular, R_0 is the basic signal-to-noise ratio (environmental and device noises), I_s accounts for the impairments on the coded voice signal (loud connection and quantizations), I_d represents the effect of delay, I_e the effect of low bit rate codecs and A is the advantage factor, corresponding to the user allowance due to the convenience in using a given technology. We reported in Table II some sample values for the R factor for different scenarios.

The main advantage of the E-model is that, for a given codec, i.e. given I_e, only delays and losses are needed for speech quality estimation.

According to [14], (3) can be further simplified to the following expression:

$$R = 93.4 - I_d(T_a) - I_e(codec, loss\_rate). \quad (4)$$

The relation between I_d and the one-way delay, T_a, is expressed as

$$I_d = 0.024T_a + 0.11(T_a - 177.3)H(T_a - 177.3), \quad (5)$$

where $H(x)$ is the step function and I_{ef} is the equipment impairment (non-linear codecs and packet losses). I_{ef} is calculated as [16][3]

$$I_{ef} = I_{e_{opt}} + C_1 \ln(1 + C_2 \cdot loss\_rate). \quad (6)$$

In the case of the GSM 6.10 codec, the formula for the impairment factor is given as

$$I_{ef} = I_{e_{opt}} + (95 - I_{e_{opt}}) \frac{loss\_rate}{loss\_rate + B_{pl}}, \quad (7)$$

where B_{pl} is the packet loss robustness factor of the audio codec [17]. Clearly, $I_{e_{opt}}$, C_1, C_2 and B_{pl} are codec specific parameters: table III reports the values for the codecs employed in our tests.

[3]The calculation is accurate up to $loss\_rate = 0.1$, for higher values it may prove optimistic.

V. TESTBED CONFIGURATION

Our testbed targets a residential broadband access, where the system operates in the $2-11$ GHz band. The experimental data has been collected exploiting a 4-nodes wireless testbed deployed in a rural environment, implementing a PMP structure, as sketched in Fig. 2. The BS is equipped with a sectorial antennas with a gain of 14 dBi covering all the 3 SSs. The default maximum output power at antenna port is 36 dBm for both the BS and the SS. The distance between the BS and SS1, SS2 and SS3 is 8.4 km, 8.5 km and 13.7 km, respectively. The average signal-to-noise ratio is above 30 dB, thus enabling the higher modulation, i.e. 64 QAM, for each connection. The SSs work in line-of-sight conditions under FDD half-duplex. All nodes run a Linux distribution based on a 2.4.31 kernel. The measurements are performed exploiting an Alvarion BreezeMAX platform operating in the 3.5 GHz licensed band and using a 3.5 MHz wide channel in FDD mode. Each node is attached through an Ethernet connection to the WiMAX equipment.

A. Alvarion BreezeMAX settings

The Alvarion BreezeMAX platform support a *per-user* QoS model where performance parameters are enforced over a pool of connections between the BS and the SS. User QoS requirements are supported using the following parameters:

- *Committed Information Rate (CIR)*. The CIR is defined for rtPS and nrtPS traffic only. The range is from 0 to 12 Mbps that is the maximum (MAC) throughput of Alvarion BreezeMAX equipment.
- *Maximum Information Rate (MIR)*. The MIR is defined for nrtPS and BE QoS types and the rate is averaged as in the case of the CIR.
- *Committed Time (CT)*. The CT defines the time window over which the information rate is averaged to ensure compliance with the CIR or MIR parameter.

CIR, MIR and CT allowed values are reported in Tab V. The IP's DSCP [**?**] field is exploited in order to enforce a certain QoS class service. Traffic flows belonging to different service categories are tagged using the `iptables` software [18]. During our measurements, all SSs share the same QoS, as summarized in Tab. IV.

B. Traffic Patterns

Data flows and CBR VoIP flows were generated by means of the Distributed Internet Traffic Generator (D-ITG), a freely available software tool [19]. D-ITG can generate and inject different traffic patterns over TCP and/or UDP sockets. Instead VBR VoIP flows were generated using Jugi's Traffic Generator (JTG) [20]. We decided to use JTG for our experiments since it can read the information about packet transmission intervals and packet sizes from files, allowing us to create an exact duplicate of a trace starting from a pre-recorder stream. Traffic is then collected at the receiver side where suitable tools are available for analysis. In our settings, we assumed that several concurrent VoIP flows use a SS as their gateway towards a peer terminal (this would be the typical case of several voice stations multiplexed at a VoIP gateway). We measured the performances of the uplink and the downlink separately, thus neglecting interference effects.

Four commonly used codecs have been considered for our experimentation, whose parameters are reported in Tab. VI. Also, the considered scenario was homogeneous and the background data traffic (in our case persistent TCP connections) was modeled considering a TCP socket working in saturation regime, according to the parameters reported in Tab. VII. rtPS services are used for VoIP connections, while TCP-controlled traffic is mapped in the BE class. The mapping of CBR sources into the rtPS class made much easier trace the behavior of the system, since the actual scheduling policies were unknown on our side.

In order to collect reliable measure of delays, before each experiment we synchronized each node with a common reference using NTP [21]. All SSs sustain the same traffic, consisting in an increasing number of VoIP session plus one persistent TCP connection (aimed at modeling background traffic). All measurements were performed over 5 minutes intervals; results are averaged over 10 runs. In the next section we report on the performance of the testbed described above.

TABLE III
PARAMETERS OF THE EQUIPMENT IMPAIRMENT FACTOR FOR G.729.2, G.723.1 AND GSM 6.10 CODECS

Parameter	G.729.2	G.723.1	GSM 6.10
$I_{e_{opt}}$	10	15	20
C_1	47.82	90	-
C_2	0.18	0.05	-
B_{pl}	-	-	43

TABLE IV
MAPPING RULES OF ALVARION BREEZEMAX

Traffic Class	DSCP	CIR [Kbps]	MIR [Kbps]	CT [ms]
BE	1	n.a.	12000	100
nrtPS	2-31	3000	12000	100
rtPS	32-63	300	n.a.	50

TABLE V
ALLOWED VALUES FOR THE CT PARAMETER

CT (ms)	BE	nrtPS	rtPS
Short	50	50	50
Medium	100	100	100
Long	1000	1000	200

VI. PERFORMANCE MEASUREMENTS

In the first set of measurements, we determined the voice capacity, i.e. the maximum number of sustained VoIP calls with high quality ($70 < R < 80$) and related parameters: VoIP throughput, delay and packet loss. Here, we report only the downlink behavior, since we found that the downlink was

TABLE VI

REFERENCE CODECS FOR THE VOIP TRAFFIC SYNTHESIS

	G.729.2	G.723.1	G.711	GSM 6.10
Rate (Packets/sec)	50	26	100	50
Payload length (Bytes)	32	42	92	33

TABLE VII

PARAMETERS OF FTP FLOWS

	Best Effort (FTP)
Rate (Packets/sec)	2000
Payload length (Bytes)	1460

actually the bottleneck.

We first measured the average throughput of VoIP calls, at the increase of the number of VoIP flows. It turned out that the performance for the G.711 codec is far too low to be acceptable and a SS could not support more than 2 (high-quality) calls, as depicted in Fig. 4. Hence, in the following we reported on the comparison for the two remaining codecs. Fig. 5 and Fig. 6 depict the results for the delay and the packet loss, respectively. The delay, in particular, saturates at 300 ms, whereas, after the saturation point, packet loss increases almost linearly. The G.723.1 codec outperforms clearly G.729.2; such a difference is due to the higher G.729.2 packet generation rate, coupled to the large overhead of packet headers of the RTP/UDP/IP/MAC protocol stack ($\simeq 45\%$ for the G.729.2). Such effect is very well known in VoIP over WLANs [22] or, more in general, for bandwidth-limited connections [15]. In practice, it is convenient to employ larger speech trunks per packet and consequently larger packet generation intervals.

Packet loss, as depicted in Fig. 6, as soon the rtPS traffic at each SS exceeds the CIR, severely impair the performance of VoIP. Basically, an increasing fraction of packets are discarded at the BS side, and, in the same region, the delay value stabilizes around a saturation value. This is clearly due to the fact that VoIP packets arriving at the rtPS queue are likely to find a full buffer.

Finally, Fig. 7 provides a comprehensive picture in terms of the R-Factor. We notice that there exist roughly three regions: in the leftmost region, G.729.2 provides a fairly good quality, but as soon as the network starts saturating around 10 calls, G.723.1 obtains much better performance. In the end, we could assess that with G.723.1, the system under exam can support up to 17 VoIP calls per SS with a high quality. Conversely, the use of a G.729.2 codec reduces voice capacity to 10. Clearly, the voice capacity figures which can be obtained for larger values of the CIR parameter will be scaled accordingly.

A. Delay: uplink and downlink comparison

In order to determine the voice capacity of the system, we restricted our focus on the downlink, because it poses the most stringent constraints; in this section we justify this statement showing the outcomes of the uplink and downlink. As reported in Fig. 8, the quality of the perceived speech is

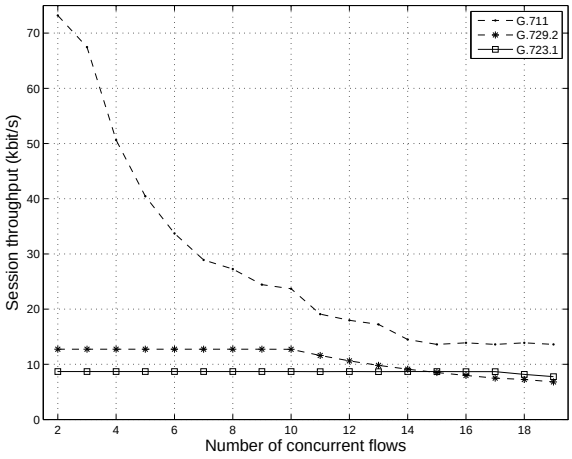


Fig. 4. Average Throughput per VoIP session versus an increasing number of VoIP calls per SS.

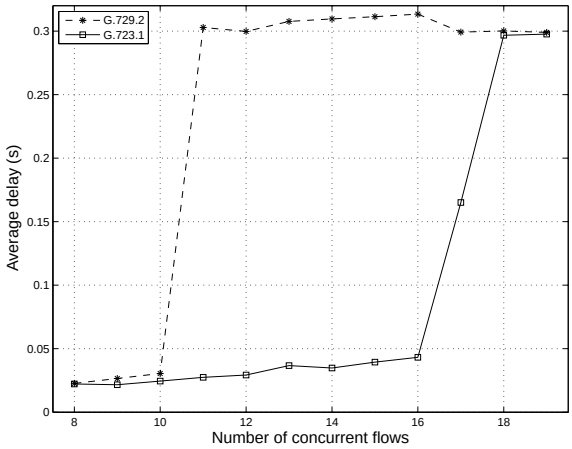


Fig. 5. Average delays versus an increasing number of concurrent VoIP flows per SS for different codecs.

better for the uplink, irrespective of the index of the SS VoIP flow considered, and of the codec considered.

Also, the packet loss was always slightly better for the the uplink compared to the downlink, for this case and the following ones.

In order to have a better understanding of the behavior of the system, we sampled the first order probability density function (pdf) of of the packet delay, both for the uplink and the downlink in some critical cases, i.e. for a number of sessions around the voice capacity. In particular, Fig. 9, Fig. 10 represent the sample delay pdf for downlink VoIP flows. Even though the scheduling policy is undisclosed, it is apparent that it is not simply the average delay to degrade a the increase of the offered VoIP traffic, but the whole delay distribution is shifted around higher delay values. This proves that the BS operates a very strict threshold control policy: in case a SS exceeds a certain threshold above the CIR, the system basically penalizes any violating SS, since packets are discarded and no further capacity is assigned to

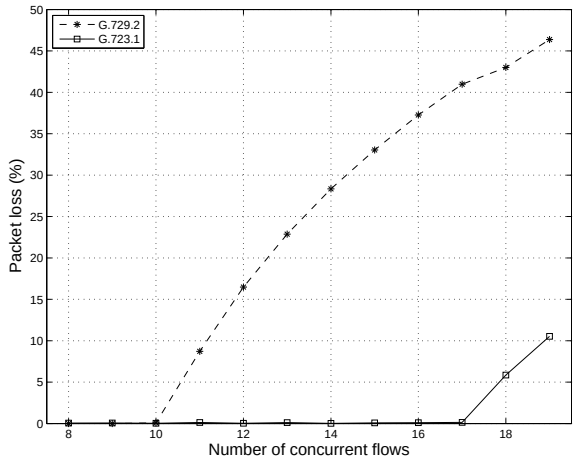


Fig. 6. Packet loss rate of VoIP flows per SS using different codecs.

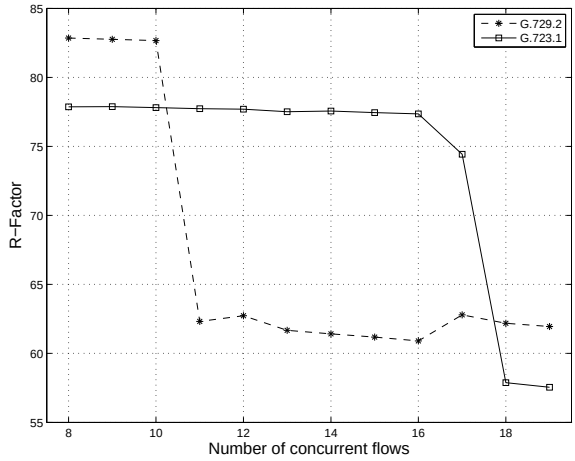


Fig. 7. Average R-factor versus number of concurrent VoIP flows per SS using different codecs.

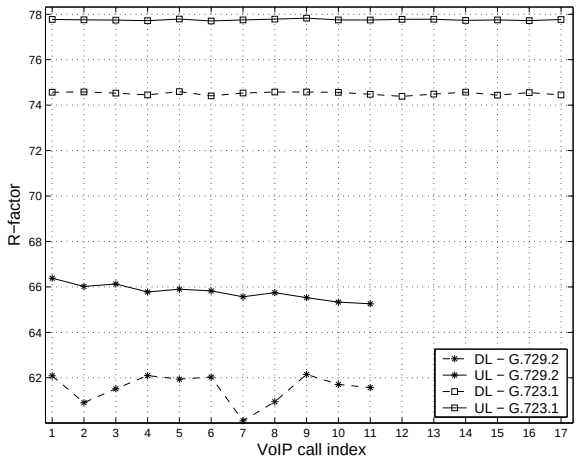


Fig. 8. Uplink and Downlink R-factor vs SS VoIP session index, using 11 and 17 concurrent calls with the G.729.2 and G.723.1 codecs, respectively.

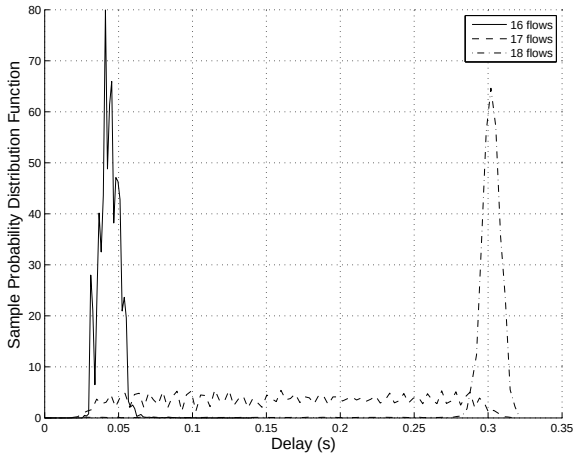


Fig. 9. Sample probability density function of G.723.1 codec in downlink direction.

the VoIP flows. In this way, the delay of packets which are not discarded is concentrated around a value which accounts for the transmission time and the queuing delay in a full buffer. In fact, only for 17 concurrent G.723.1 VoIP calls the excess above the CIR appears evenly redistributed over the interval, the rationale being that in such case the smaller throughput of the codec might bring oscillations above and below the limit.At the SS side, this strict BS policy indeed suggests to employ suitable admission control for outgoing and incoming VoIP flows, in order not to incur into major service degradation. We repeated the same measurements in the case of the uplink, and, as reported in Fig. 11 and Fig. 12, the results are similar. As emerged from the R-factor measurements, the uplink performs better than the downlink and, in fact, the delay distribution of the uplink around the VoIP capacity appears in all cases centered at lower values compared to the downlink.

We remark that the uplink measurements contradict the simulation results obtained in [11], where larger delays in the uplink, compared to the downlink, were ascribed to the bandwidth request mechanism and to the PHY overhead. In the case at hand, the uplink delay due to bandwidth request did not prove significant. We ascribe this to the activation of piggybacking for bandwidth reservation provided by WiMAX.

B. The case of VBR traffic

Many commercial voice codecs employ CBR coding. If this is the case, dimensioning and testing the system can be efficiently performed using a procedure similar to the one showed above. Nevertheless, several voice codecs can optionally employ Voice Activity Detection (VAD). VAD is a technique typically used in speech processing that aims at detecting the presence or absence of human speech. Under VAD, the application stops packet transmissions when the user is not speaking until new voice activity is detected. Clearly, since the CBR packet source generates packets only during active periods, consistent bandwidth savings are possible.

We tested the system under VAD enabled voice traffic. Traffic traces have been generated using Ekiga [23], an open source VoIP and video conferencing application, running over a wired

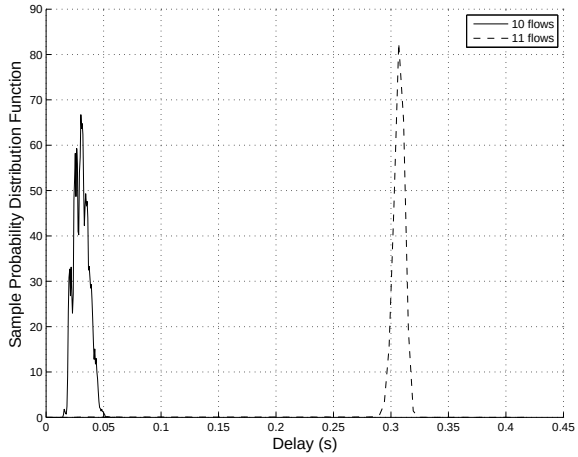


Fig. 10. Sample probability density function of G.729.2 codec in downlink direction.

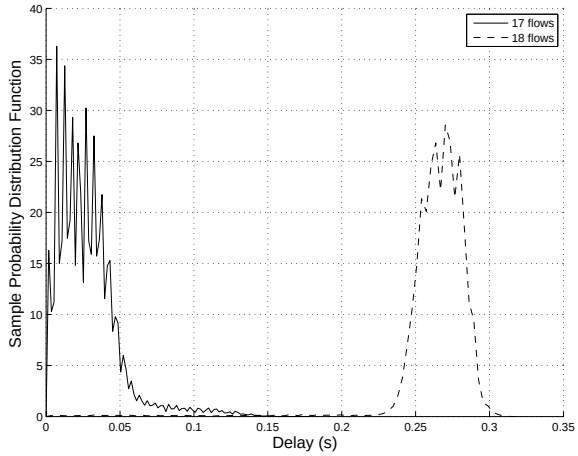


Fig. 11. Sample probability density function of G.723.1 codec in uplink direction.

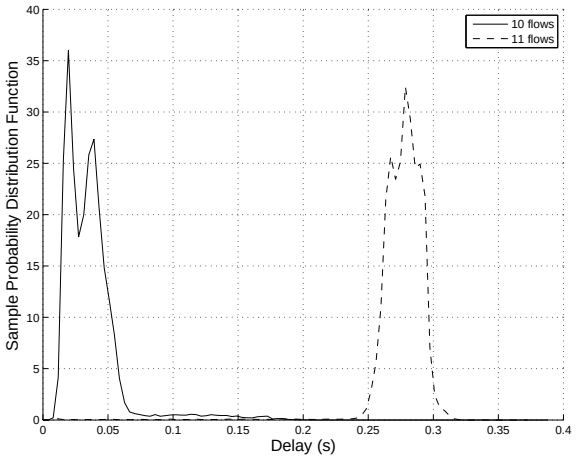


Fig. 12. Sample probability density function of G.729.2 codec in uplink direction.

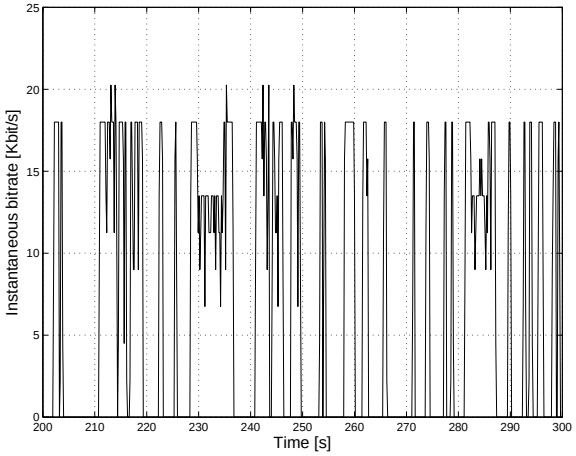


Fig. 13. Instantaneous bitrate of a GSM 6.10 encoded voice trace with VAD enabled.

LAN. At one end, we registered voice traces corresponding to a VAD enabled GSM 6.10 device, dumping the resulting packet trace with WireShark [24]. The instantaneous bitrate of the recorded voice trace pattern over a 100 seconds interval is reported in Fig 13, where we can clearly distinguish the active and silent periods modulated by VAD.

We assessed preliminarily that a non-VAD enabled GSM 6.10 source has a voice capacity of 10 voice sessions. In the VAD enabled case, according to our voice recorded traces, the codec was detected inactive for a fraction of time $a = 53\%$. The guess is then that, since the employed CIR is far from the system capacity, we should expect roughly to double the number of VoIP flows compared to the plain CBR case. We measured the R-factor of the multiplexed voice flows, which provided the quality of the perceived speech as detailed in Fig. 14. The gain of the VAD technique brings voice capacity to 22 voice sessions, which is in line with our guess[4]. Notice that, as expected, the degradation of the system performance is much smoother than in the case of pure CBR sources.

The smooth degradation of the R-factor is confirmed by the delay and packet loss figures, as depicted in Fig. 15 and Fig. 16 respectively. Even in this case, once the CIR threshold is exceeded, the system gracefully degrades to saturation.

VII. General discussion

From the results presented in the previous section, it is clear that the VAD technique yields a multiplexing gain, in terms of number of VoIP calls, larger than a pure CBR source. However, the codec's choice is not controlled by WiMAX system designer/network administrator, since the customer starts the audio conversation using her/his favorite application (e.g., Skype/Softphone) whenever he wishes. Although the IEEE 802.16 specifications define the multiple access signaling mechanisms [**?**], the radio resource management issues such as bandwidth allocation and connection admission control are still open. Connection admission control [26], in particular, is used to limit the number of connections in the network

[4]In principle, we could employ the *effective bandwidth* formula [25] in order to make such estimate precise, but here the buffer size is unknown.

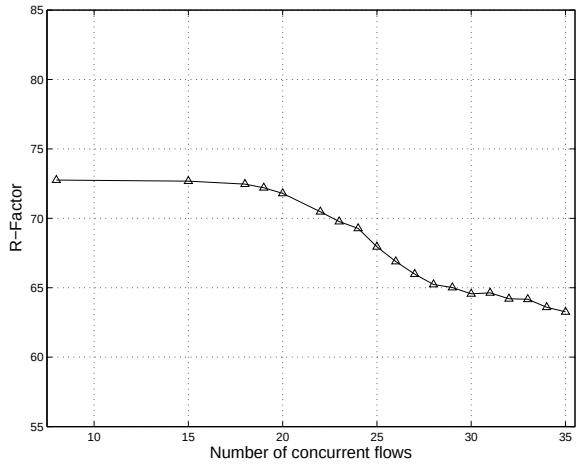


Fig. 14. Average *R*-factor versus number of concurrent VoIP flows using a VAD enabled GSM 6.10 codec.

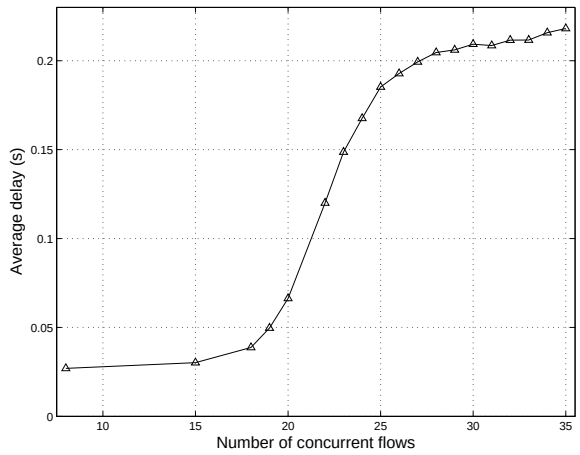


Fig. 15. Packet delay versus number of concurrent VoIP flows using a VAD enabled GSM 6.10 codec.

and it works jointly with the bandwidth allocation mechanism, which allocates available radio resources among outgoing and incoming connections so that the QoS performances of both types of connections can be maintained at the target level.

Indeed the R-factor is a parameter that a system designer has to take into account for a correct design of the WiMAX scheduler. For example authors of [27] compute an R-factor estimate of the low rate probing traffic on each available path using running averages of the delay and loss in a wireless mesh network. When the averaged R-factor of the probing traffic stays under 70 for more than a few seconds, the actual voice traffic is re-reouted to a better path.

An interesting direction, in the authors' opinion, would then check wheter the same principle is also applicable to a WiMAX architecture, where the BS is in charge to the centralized-control of the wireless resources. Basically, each connection ID is associated to the running average of the R-factor. Since the R-factor should be maximized, a right scheduling algorithm is needed to be developed in order to obtain the amount of allocated bandwidth for all of the ongoing

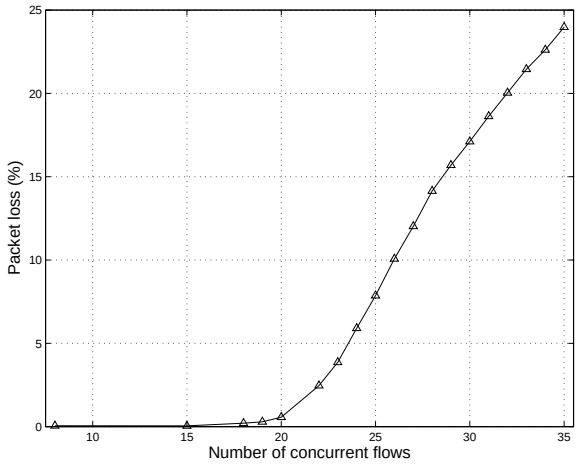


Fig. 16. Packet loss rate versus number of concurrent VoIP flows using a VAD enabled GSM 6.10 codec.

and the newly arriving connections (assuming that they are admitted in the WiMAX system). Specifically, a new connection is admitted if, upon admission of that connection, the QoS requirements of all the connections can be satisfied [28].

VIII. Conclusions

In this paper, we presented the results of a series of measurements assessing the performance of VoIP applications on a WiMAX testbed deployed in Turin, Italy. We employed an objective performance evaluation technique feeding the system with synthetic traffic flows reproducing VoIP traffic and elaborated the outcome through the E-model, to evaluate the corresponding perceived voice quality.

We showed that the QoS support of WiMAX fits quite well the requirements for VoIP applications, even in presence of background best effort traffic. The voice capacity, in particular, was strongly dependent on the codecs adopted, and confirming that the most stringent parameter is the codec packet generation rate, despite the VoIP packet length. Furthermore, we found that, at least in the case of the testbed considered, the uplink performance is slightly better than the downlink, so that the downlink determines the voice capacity of the system.

We could also get some insight into the QoS control policy implemented by the equipment vendor, showing that it performs a very strict per connection control on the volume of traffic generated: this suggests the adoption of per-flow admission control at the SS side, in order not to incur into penalties. With further measurements we considered the effect of VBR voice sources, as those resulting from the activation of VAD devices. We could assess the gain in the voice capacity obtained through the adoption VBR VoIP codecs, which represent the natural candidates for rtPS traffic.

IX. Acknowledgements

The authors of this paper wish to thank CSP for the valuable support during the testing phase and the experimental setup.

REFERENCES

[1] "Wimax forum." [Online]. Available: http://www.wimaxforum.org

[2] R. Bruno, M. Conti, and E. Gregori, "Mesh Networks: Commodity Multihop Ad Hoc Networks," *IEEE Communications Magazine*, vol. 43, no. 3, pp. 123 – 131, Mar. 2005.

[3] C. Eklund, R. Marks, K. Stanwood, and S. Wang, "IEEE standard 802.16: a technical overview of the WirelessMAN air interface for broadband wireless access," *IEEE Communications Magazine*, vol. 40, no. 6, pp. 98–107, June 2002.

[4] "Fondazione Ugo Bordoni." [Online]. Available: http://wimax.fub.it/

[5] "Alvarion." [Online]. Available: http://www.alvarion.com/

[6] *Recommendation G.107: The E-model, a computational model for use in transmission planning*, ITU-T Std., 2005.

[7] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen, "Broadband wireless access with WiMax/8O2.16: Current performance benchmarks and future potential," *IEEE Communications Magazine*, vol. 43, no. 2, pp. 129–136, February 2005.

[8] C. Hoymann, "Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16," *Computer Networks*, vol. 49, no. 3, pp. 341–363, 2005.

[9] I. Stojanovic, M. Airy, D. Gesbert, and H. Saran, "Performance of TCP/IP over Next Generation Broadband Wireless Access Networks," in *Proc. of WPMC*, Aalborg, Denmark, 2001.

[10] P. C. Ng, S. C. Liew, and C. Lin, "Voice over wireless LAN via IEEE 802.16 wireless MAN and IEEE 802.11 wireless distribution system," in *Proc. of IEEE WIRELESSCOM*, Hawaii, USA, 2005.

[11] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network Magazine*, vol. 20, no. 2, pp. 50–55, March 2006.

[12] N. J. Muller and R. P. Davidson, *The Guide to Frame Relay & Fast Packet Networking*. Flatiron Pub, 1991.

[13] P. Tzerefos, V. Sdralia, C. Smythe, and S. Cvetkovic, "Delivery of low bit rate isochronous streams over the DOCSIS 1.0 cable television protocol," *IEEE Transactions on Broadcasting*, vol. 45, no. 2, pp. 206–214, June 1999.

[14] R. G. Cole and J. H. Rosenbluth, "Voice over ip performance monitoring," in *Proc. of ACM SIGCOMM*, San Diego, CA, USA, 2001.

[15] C. Hoene, H. Karl, and A. Wolisz, "A perceptual quality model intended for adaptive voip applications: Research articles," *Int. J. Commun. Syst.*, vol. 19, no. 3, pp. 299–316, 2006.

[16] *Recommendation G.113: Transmission impairments due to speech processing*, ITU-T Std., 2005.

[17] R. Kwitt, T. Fichtel, and T. Pfeiffenberger, "Measuring perceptual voip speech quality over umts," in *Proc. of IPS-MoMe*, Salzburg, Austria, 2006.

[18] "The netfilter.org project." [Online]. Available: http://www.netfilter.org/

[19] "D-ITG, Distributed Internet Traffic Generator." [Online]. Available: http://www.grid.unina.it/software/ITG/

[20] "Jtg." [Online]. Available: https://hoslab.cs.helsinki.fi/savane/projects/jtg/

[21] "Simple Network Time Protocol (SNTP) Version 4." [Online]. Available: http://www.apps.ietf.org/rfc/rfc2030.html

[22] D. P. Hole and F. A. Tobagi, "Capacity of an IEEE 802.11b Wireless LAN Supporting VoIP," in *Proc. of IEEE ICC*, Paris, 20-24 June 2004.

[23] "Ekiga." [Online]. Available: http://www.gnomemeeting.org

[24] "Wireshark." [Online]. Available: http://www.wireshark.org

[25] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968 – 981, Sept. 1991.

[26] G. Boggia, P. Camarda, L. A. Grieco, and S. Mascolo, "Feedback-based bandwidth allocation with call admission control for providing delay guarantees in IEEE 802.11e networks," *Computer Communications*, vol. 28, pp. 325–337, Feb. 2005.

[27] S. Ganguly, N. V, K. Kim, A. Kashyap, D. Niculescu, R. I. S. Hong, and S. Das, "Performance optimization for deploying voip services in mesh networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 2147–2158, Nov. 2006.

[28] D. Niyato and E. Hossain, "A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks," *IEEE Trans. Comput.*, vol. 55, no. 11, pp. 1473–1488, 2006.