

Field Test of Mobile Wi-Fi Terminals in a Wireless City

Petter Stray, Poul E. Heegaard, Thomas Jelle

Norwegian University of Science and Technology

petter.stray@gmail.com,

{poul.heegaard, thomas.jelle}@item.ntnu.no

Abstract

We evaluate the achievable voice quality using VoIP on hybrid Wi-Fi/GSM/3G mobile terminals in the citywide wireless network deployed in Trondheim. Using the network analysis tool IxChariot from Ixia a set of Qtek and HTC terminals are tested under different conditions and network loads. Performance metrics are gathered from each test and compared up against the already ubiquitous GSM service. Our findings indicate that the quality of a voice service over a shared citywide wireless network is not yet good enough for it to be a competitor to the GSM phone service.

1 Introduction

Over the last two decades the mobile phone has gone from being a tool for the few to a necessity for the public. More and more people terminate their land line subscriptions, and completely rely on a cell phone for communication. This even though the experienced voice quality on a mobile phone is substantially lower than that on the public switched telephone network (PSTN). People accept this degradation because they value the gain in mobility higher than the loss in quality.

Today many people mean that Voice over the Internet Protocol (VoIP) poses a threat to mobile services as we know them. As entire cities become giant wireless zones it is expected that these also support Voice over Wireless Local Area Networks (VoWLAN). If these large scale networks provide the necessary resources to obtain voice quality at a sufficient level, people are likely to use this as their primary communication network. The advantages are many; "limitless" downloading, free Voice over IP using Skype, GoogleTalk a.o. and the possibility to simultaneously send multiple types of traffic over the same network.

In Trondheim such a citywide wireless network has been deployed on the initiative of the Norwegian University of Science and Technology (NTNU). The network is built with special weight on providing more than enough bandwidth so high capacity services are providable. However, since wireless cities are just starting to be deployed, little research and testing is done on the services and equipment to be used in these networks. Testing different mobile Wi-Fi terminals in Wireless

Trondheim under various network conditions can help give an indication of what works best in a system as the one deployed here. Comparing the quality of a mobile voice service up against the quality of the already ubiquitous GSM service can be useful to see if the potential is present to take voice over citywide wireless networks into use. This paper is written on the basis of [9].

The remaining of this paper is organized as follows. Section two gives an introduction to how voice quality is measured. Section 3 describes how the tests are carried out throughout our research. Section 4 presents some of the most important results, while section 5 discusses the obtained results. Finally, section 6 concludes the paper.

2 Evaluating Voice Quality

Comparing the quality of data networks is a complex task, since there are many factors to consider. In wireless networks this is specifically difficult since different security mechanisms as well as infrastructure solutions can greatly affect a specific terminals performance in a network. Different applications weight a networks performance metrics in different ways. Some applications may require high bandwidth, but are less sensitive to variations in the end-to-end delay. For voice services, a single metric has been established to rate call quality - the Mean Opinion Score (MOS). For Voice over IP, the E-model provides a number comparable to MOS.

Mean Opinion Score

Assessing call quality has traditionally been a subjective task; picking up a telephone and listening to the quality of the voice. The most widespread subjective voice quality metric is the Mean Opinion Score described in the International Telecommunications Union (ITU) recommendation P.800 [8]. The MOS for a call service is calculated from letting a large number of people listen to audio and give their opinion of the call quality on a scale from 1 to 5. Each score has a description related to it as illustrated in table 1. However, if every little tuning adjustment to a phone service requires a large amount of people listening and rating the quality it would be both expensive and time-consuming. As a consequence objective measurement techniques have been established, one of them being the E-model.

Rating	Definition	Description
5	Excellent	a perfect speech signal recorded in a quiet booth
4	Good	intelligent and natural like PSTN telephone quality
3	Fair	communication quality, but requires some hearing effort
2	Poor	low quality and hard to understand the speech
1	Bad	unclear speech, breakdown

Table 1: ITU P.800 MOS conversation opinion scale [8].

The E-model

The E-model was introduced in the ITU recommendation G.107. The E-model uses measured delays and equipment impairment factors to calculate a single scalar, the "R factor". The R factor is given by the equation [1]:

$$R = R_0 - I_s - I_d - I_e (+A) \quad (1)$$

where:

- R_0 : unaltered signal, expresses the basic signal-to-noise ratio (SNR)
- I_s : impairments that occur simultaneously with the voice signal, such as too loud speech level
- I_d : delays introduced from end-to-end
- I_e : impairment introduced by the equipment
- A : advantage factor, willingness to trade voice quality for convenience

These delay and equipment impairments are influenced by the data networks one-way delay, jitter and data loss. Implicitly the codec used also influences the delay and impairments, especially if a compression codec is used. Since the voice-critical metrics delay and jitter are regarded by the R factor calculations, specific delay and jitter measurements are not presented in this paper. The advantage factor is in parenthesis in equation (1) because it is not regarded by the network analysis tool used.



Figure 1: Translation from objective R factor to subjective MOS value[5].

The R factor ranges from 0 to 100, and can easily be translated into a corresponding MOS value. When a voice conversation is converted to a digital signal and back there is an inherent degradation. This reduces the theoretical maximum R factor with no impairments from 100 down to 93.2 [5]. The translation from R factor values to MOS values is illustrated in figure 1. In the figure you see the R factor values from the E-model to the left, the likely opinion of human listeners in the middle and MOS values to the right.

City	MOS down-link % over 3.0	MOS up-link % over 3.0
Ålesund	98.88	99.40
Bergen	98.36	98.85
Stavanger	99.71	100.00
Stor-Oslo	97.96	98.18
Tromsø	99.41	99.41
Trondheim	98.52	98.71

Table 2: MOS values from Telenors GSM network [11].

Performance Criteria

Generally voice quality is considered to be good if it exceeds a MOS value of 3.5. However, according to [4], a normal GSM network only delivers audio with a MOS score between 2.9 and 4.1. The lower bound of this interval is relatively far down in the "Nearly All Users Dissatisfied" section of the MOS scale in figure 1. People clearly value the mobility they get from a cellular phone a lot since they accept such a large degradation in voice quality.

Telenor, Norway's largest GSM provider, has provided some voice quality statistics from their mobile network in six major Norwegian cities. They have measured the share of conversations that obtain a MOS score of over 3.0, their results are summarized in table 2.

As you can see from the table a very high percentage of the calls in the GSM network qualify for a MOS score over 3, as many as 98.95% of the conversations on average.

Considering that most people find the quality achieved by their GSM phones to be sufficient, a lower quality than that in GSM will probably be acceptable for the majority of potential users of a citywide wireless network for voice communication. Since the lower MOS value in GSM is around 3, we assume that a value as low as 2.6 on the MOS score might be acceptable for the basically free voice service that is possible over the packet switched network. We have chosen the lower quality limit 2.6 since this also is the lower border of the "Nearly All Users Dissatisfied" group of MOS scores in figure 1.

3 Test Implementation

IxChariot [6], the analysis tool used in this research, directly gives you the MOS value a conversation achieves. The software was installed on a server at Uninett¹ in Trondheim, while small software parts, called *performance endpoints*, were installed on all of the terminals being tested. Two laptops were also used in the tests. One was used to start and stop tests through a remote desktop connection to the IxChariot server, while the other one was used to generate background traffic or act as an endpoint in conversations with the mobile terminals. A desktop computer placed at NTNU's largest campus, Gløshaugen, was used in the same way. Figure 2 illustrates how the software works.

¹Uninett is developer and operator of the Norwegian research network which provides high capacity Internet connectivity to over 200 educational and research institutions in Norway.

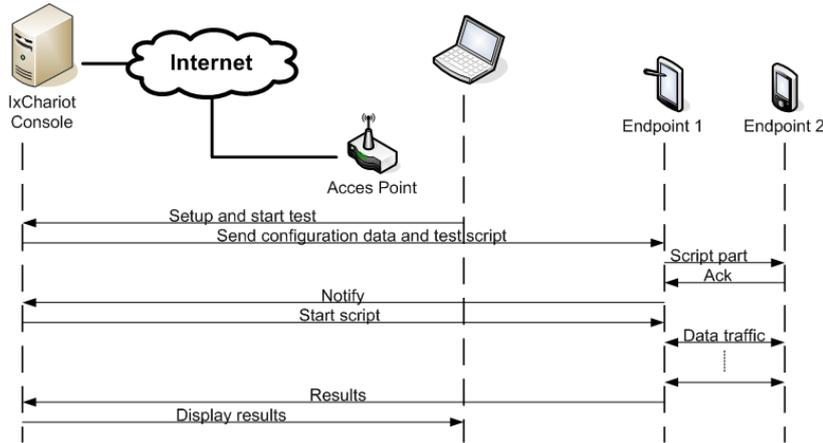


Figure 2: Example of WLAN test setup with Ixia's IxChariot.

Testing in IxChariot is script-based, so at startup the IxChariot console sends the script to be run to the involved terminals. The endpoints involved can either collect the performance information and transfer it to the console when the test is finished, batch-mode, or they can constantly update the performance to the console, real-time-mode. In this research batch-mode was used while performing stationary tests, and real-time-mode was used in tests involving roaming. This because batch-mode provides the most accurate results since no performance information is transferred in the network during a test. Real-time-mode is used in roaming because if the connection to an access point is lost during a test no information on the performance up till that point is reported.

If VoIP is tested, the traffic between Endpoint 1 and 2 in figure 2 follows standard VoIP procedure with Session Initiation Protocol (SIP) control plane followed by Real-Time Transport Protocol (RTP) stream and teardown [6].

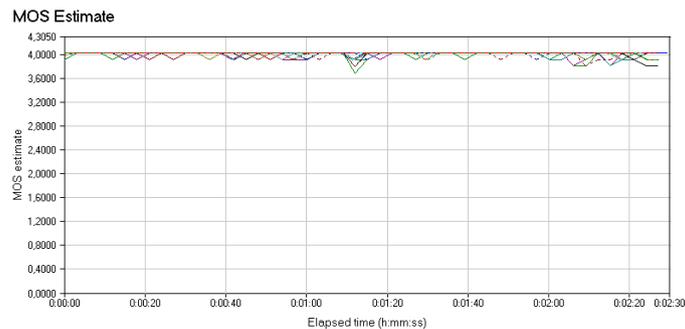
Test Setup

The tests in this research were performed at different locations in Trondheim, and different combinations of terminals were used. A total of five mobile Wi-Fi terminals were tested, one Qtek 9000, one Qtek 8310, two Qtek 8300's, and one HTC TyTN. Voice conversations were set up between the mobiles, between mobiles and a laptop, and between mobiles and a desktop computer. A few tests were also run only involving a laptop and a desktop, in order to see if limitations were located at the terminal or network side. Varying amounts of background traffic were also used to see how the voice quality was affected by different traffic types. The background traffic was generated using both built in scripts in IxChariot and the synthetic Internet traffic generator GenSyn [3]. The background traffic was generated to go between a laptop connected to the same access point as the tested terminals and the desktop computer at Gløshaugen. The background traffic had both the shape of high capacity applications as well as ordinary web browsing.

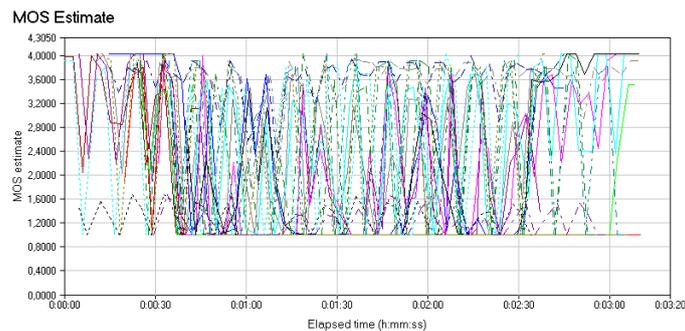
Tests were performed using two of the most common voice codecs, ITU-T G.711 and G.729. The two codecs have fairly different characteristics. G.711 is the preferred codec in PSTN/ISDN [13], merely digitizing voice and providing a 64 kbps stream. G.729 uses compression as well and provides 8 kbps of output. Since G.729 provides low bitrate output it is preferred in wireless voice networks [10].

4 Results

Most infrastructure producers don't recommend more than six or seven active voice conversations at once on their access points. To test the capacity of the access points in Wireless Trondheim up to ten concurrent conversations were tested at one access point. Figure 3 shows the achieved quality using only a laptop, and how the quality was after switching to mobile Wi-Fi terminals. In the figures the differentiation of the plots is not important, it is the trend of the curves that show the changes in voice quality.



(a) Laptop only.



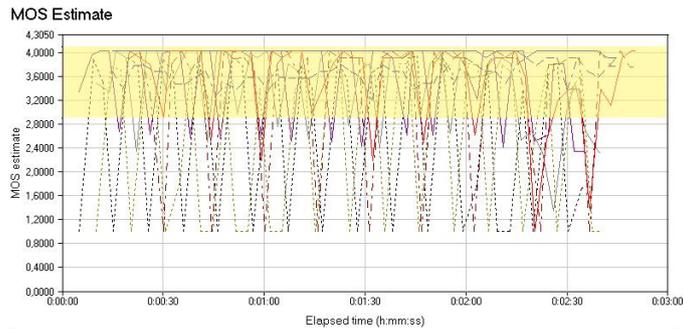
(b) After introducing mobiles.

Figure 3: 10 G.729 conversations.

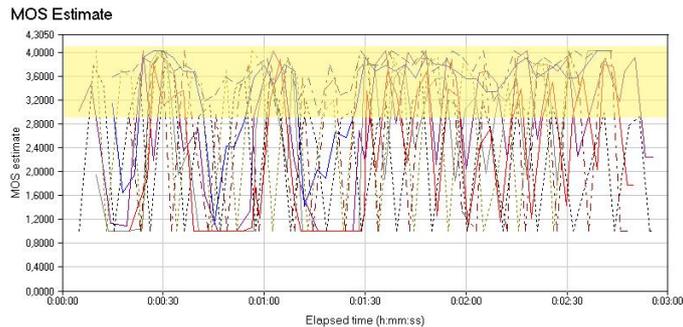
An increasing amount of consecutive calls was also used to test the network cell call capacity. Testing with from three to ten conversations, you could clearly see that the voice quality dropped below acceptable values when seven calls were active at once. This was the same for both voice codecs.

The effect of adding background application data traffic to the access point was also tested. Tests were performed using both bandwidth demanding applications as well as ordinary web-browsing traffic. While a high throughput application generating a little over 3 Mbps of traffic caused the voice quality in ongoing conversations to drop with 25 %, simple web-browsing generating a little over 230 kbps of traffic caused a drop of 15 % in voice quality.

Figure 4 shows how voice quality in the network deteriorates after adding background traffic. The average voice quality in the ongoing conversations drops from 3.34 to 2.82 after introducing background traffic in figures 4(a) and 4(b). The yellow-shaded area in both figures indicate the quality that is obtained in GSM networks.



(a) Without background traffic.



(b) With background traffic.

Figure 4: 4 G.729 conversations with background traffic.

Roaming

Roaming is the action where a terminal switches from being connected to one access point to being connected to another access point. Tests on roaming were performed walking with the roaming terminal out of the coverage area of the associated access point and into the coverage area of another AP. Roaming was also performed using both codecs, and the two codecs performed quite differently. While conversations using G.711 generally held the conversations going longer than those using G.729, the conversations using G.729 maintained higher voice quality.

Terminal Differences

The tested terminals had fairly different performance characteristics. The Qtek 9000 was the most powerful terminal tested with a 520 MHz processor. This was also reflected in the results, where the Qtek 9000 generally outperformed all the other terminals. The HTC TyTN was the terminal with the second best prerequisites to perform well with a 400 MHz processor. However, the TyTN was unable to provide good voice quality on the streams going to it. While the streams originating at the TyTN mostly achieved high voice quality, the streams terminating at the terminal would have been completely inaudible. The two Qtek 8300's and the 8310 all performed very similar. On average these three terminals performed better than the TyTN, but worse than the Qtek 9000. Also these terminals had greatly varying quality on streams originating and terminating at the mobile. However, not as differing as the TyTN. Figure 5 shows how the tested terminals vary in performance.

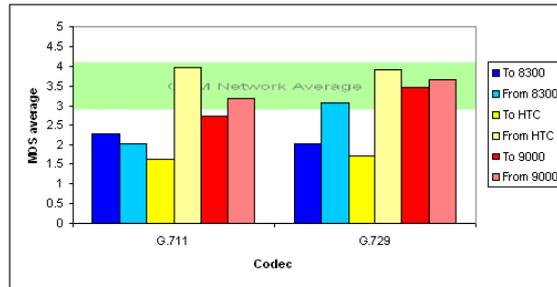


Figure 5: Comparison of terminal performance.

Test in Indoor Enterprise Network

The results obtained in the tests in the Wireless Trondheim network were below acceptable values. The tests performed could not unveil if the cause of poor performance was the terminals or if problem lay in the compatibility between the terminals and the network. Trying to figure out what was the cause of poor performance tests were also performed in an indoor enterprise network at Uninett. The results obtained here were far better. As you can see from figure 6 the voice quality was more uniform at Uninett. Figure 6(a) shows that the voice quality in Wireless Trondheim is constantly jumping between 1 and 4 on the MOS scale, while figure 6(b) shows how the quality obtained at Uninett is only shifting between 3.8 and 4. In both figures the lines averaging between 1.2 and 1.5 are the streams terminating at the HTC TyTN.

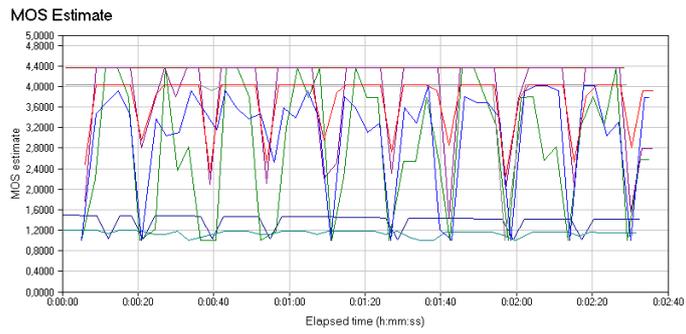
These tests indicate that the poor performance in Wireless Trondheim at least isn't caused by poor processing power at the terminal side.

Isolated Access Point Test

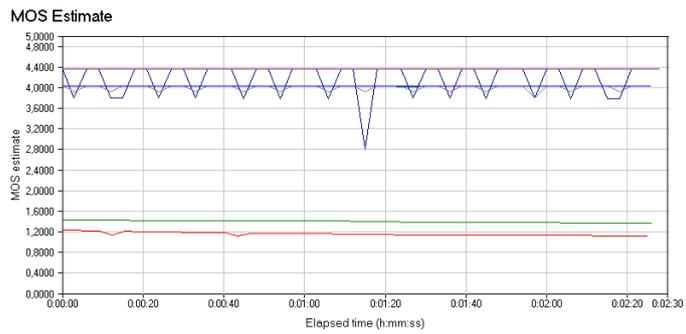
A few final tests were performed at an isolated access point in Wireless Trondheim. The access point is placed in a park, and is long enough away from other access points for interference to be a problem. In these tests a spectrum analysis was performed as well to monitor traffic on the different channels in the 2.4 GHz band. As you can see from figure 7 there was only traffic on channel 11 in these tests, which is the channel the access point is operating within. In figure 7(b) you can clearly see that the voice conversation started after approximately 40 seconds where the channel 11 area turns completely white.

All the terminals performed far better connected to the isolated access point. Although the performance varied when the terminal was far away from the access point at close to it, the overall performance was more similar to the results obtained at Uninett than those obtained in the city.

These tests rule out that compatibility issues between the tested terminals and the Wireless Trondheim network are to blame for the performance weaknesses.

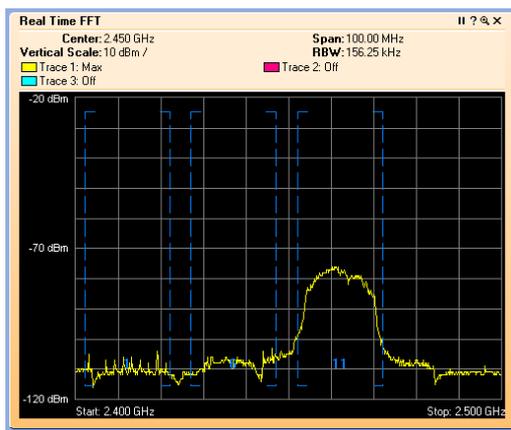


(a) Results from Wireless Trondheim.

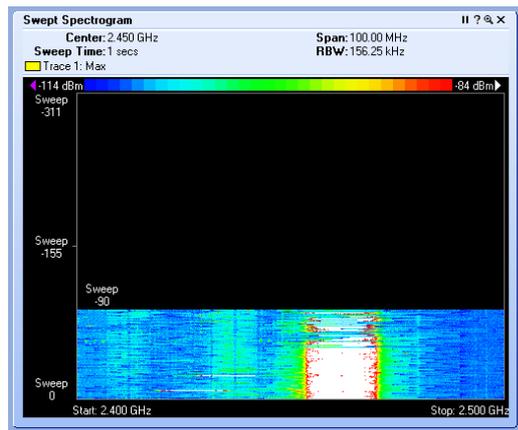


(b) Results from Uninett.

Figure 6: Two G.711 and two G.729 conversations.



(a) Radio Frequency (RF) power as a function of frequency.



(b) RF power per 1 second sweep.

Figure 7: 2.4 GHz band activity.

5 Discussion

The call capacity in each network cell is a constraint. As the test results show each access point can handle approximately six conversations at once. If application traffic is introduced to the access point as well, calls may be dropped and/or the voice quality in the ongoing conversations will drop significantly.

An access point usually has a signaling radius of about 50 meters [7]. This means that when walking down a street in Trondheim you will roam from one access point to another approximately every 100 meters. Walking in one of the main streets in Trondheim on a sunny Saturday afternoon you will most likely see many more than six people every hundred meters talking in their cell phones. Although the density of access points in the city is much higher in such high traffic areas, there will probably be many situations where an access points capacity is breached.

Reducing the transmit power on the access points can be one solution to the cell capacity problem. By reducing the transmit power each access point covers a smaller area, which lowers the risk of overloading an AP. In a wireless network covering an entire city this would be a very expensive solution since lowering the range on each of the access points would require the deployment of a significant amount of new access points.

Another solution could be using a seamless roaming system such as the Generic Access Network (GAN). GAN, also known as Unlicensed Mobile Access (UMA), enables a terminal to seamlessly roam between GSM/UMTS and a WLAN if a WLAN comes within reach [12]. If this is combined with a capacity check of the WLAN before connecting to it, this could greatly prevent access point overloads. Quite a few terminals with GAN/UMA support has already reached the market from large cell phone producers such as Nokia and Samsung. For further reading about UMA the reader is referred to [12].

Comparison with GSM

As previously stated, it is unlikely that people will start to use voice applications over the Wireless Trondheim network unless a certain voice quality is maintained. Since almost everybody uses a cellular phone every day, it is natural to evaluate the quality of VoWLAN up against the GSM phone service.

As you saw in table 2 as many as 98.95 % of the calls made in Telenors GSM network qualify for a MOS score over 3. Using the G.729 codec on a set of terminals in the Wireless Trondheim network only 77.46 % of the conversations score over 3.0 on the MOS scale. This means that almost $\frac{1}{4}$ th of the time the voice quality using voice over wireless is below the quality you get using GSM. In the same case only a little over 80 % of the conversations maintain MOS values over 2.6. If however you are unlucky and somebody is downloading a file while connected to the same access point as you it is likely that only 58.16 % of the conversations maintain a voice quality over 3.0. At the same time, only 62 % of the time the conversations maintain voice quality above the minimum MOS level of 2.6 set in section 2.

As the tests from Uninett show the terminals behave differently in an isolated indoor environment where interference is deliberately counteracted upon. The tests performed connected to an isolated access point in Wireless Trondheim support this as well. It seems obvious that the mobile terminals' radios are far to poor to enable time critical applications to function well in a noisy environment, with interference caused by private access points, cordless phones and microwaves [2].

Data applications may well function because retransmission of lost packets is acceptable, but for the time critical application retransmissions are seldom an option.

The situation will presumably not improve until mobile hand held devices' radios improve. This can be by improving the current 802.11b or g radios sitting in them already, or that the terminals get support for IEEE 802.11a, while, at the same time, private actors do not install private 802.11a networks. If all 802.11a access points are in the control of Wireless Trondheim a mostly noise free environment can exist in the 5 GHz band, hopefully improving the performance of mobile terminals in a voice over WLAN session.

6 Conclusion

The objective of this research was to perform a series of field experiments on Voice over IP in the Wireless Trondheim network. Using the network analysis tool IxChariot, Voice over IP was tested on a set of Wi-Fi enabled Qtek and HTC terminals under varying network conditions. Through tests of the available terminals, under different network compositions and loads, the overall quality of the service was assessed. The quality of the service was also compared up against results from GSM in order to see if a voice service over the Wireless Trondheim network can compete with GSM.

One of the main findings through the experiments was that in a city like Trondheim there is far too much interference for a mobile terminal to obtain a stable voice conversation with good quality. Today's mobile Wi-Fi terminals operate in the 2.4 GHz band, alongside private wireless networks, cordless phones and microwave ovens. A densely populated city center therefore contains a lot of interfering signals. On average, each of the over 100 access points in the Wireless Trondheim network sees 12 interfering access points not belonging to the network. This naturally influences the clarity of the received signal at the terminals as well.

The obtained voice quality has been measured on a five point scale, called the mean opinion score (MOS), where 1 is poor and 5 is excellent quality. Nearly all calls made in the GSM network score over 3.0 on the MOS scale. Taking this, among other factors, into consideration the criteria for acceptable quality for the voice over Wireless Trondheim service was set to 2.6 on the MOS scale. The tested terminals have performed differently, however, the obtained quality in the tests has been constantly lower than the criteria set for the service. The voice quality in the network degrades further when adding data traffic to the network. The presence of ordinary web browsing traffic causes the voice quality of ongoing conversations to drop almost 0.5 on the MOS scale.

An indirect factor also affecting the experience of voice over the Wireless Trondheim network was the battery life of the mobile terminals. Connected to a wireless network the terminals' batteries only last a few hours. Considering that most people have wireless networks available both at work and at home, it is likely that a practically free voice over wireless service will be desired to use as one's primary communication medium. If this involves recharging your terminal several times a day it is not likely to catch on in the near future.

Without any prioritizing of latency sensitive traffic the shared wireless network seems unsuitable for VoIP services. After testing several different terminals in the Wireless Trondheim network the conclusion is quite clear, voice over the citywide wireless network is not yet ready for the masses.

References

- [1] ITU-T Studygroup COM12. The E-model. (online), 2007. URL: <http://www.itu.int/ITU-T/studygroups/com12/emodelv1/introduction.htm>.
- [2] Jim Geier. The state of Wireless LANs. *NetworkWorld's Special Report*, 2004. Sponsored Exclusively by intel.
- [3] Poul E. Heegaard. GenSyn - A Generator of Synthetic Internet Traffic used in QoS Experiments. *Presentation at 15th Nordic Teletraffic Seminar, Lund, Sweden*, 2000.
- [4] inline Systems AB. Voice Quality. (online), 2005. URL: http://www.inline.se/products_vq_voice_quality.html.
- [5] Ixia. Assessing VoIP Call Quality Using the E-model. White Paper, 2005. URL: http://www.ixiacom.com/library/white_papers/display?skey=voip_quality.
- [6] Ixia. IxChariot. (online), 2007. URL: <http://www.ixiacom.com/products/display?skey=ixchariot>.
- [7] Thomas Jelle. Wireless Trondheim - 1:2006. Norwegian University of Science and Technology (NTNU), Jan 2006.
- [8] Telecommunication Standardization Sector of ITU. Methods for subjective determination of transmission quality. ITU Recommendation, 1998.
- [9] Petter Stray. Field Test of Mobile Wi-Fi Terminals in a Wireless City. Master's thesis, Norwegian University of Science and Technology (NTNU), 2007.
- [10] Ready Technology. Open Source G.729 and G.723.1. (online), 2007. URL: <http://www.readytechnology.co.uk/open/ipp-codecs-g729-g723.1/>.
- [11] Telenor. Tale statistikk telenor area mai (Norwegian). Private communication, 2007.
- [12] UMA Today. UMA Technology. (online), 2007. URL: <http://www.umatoday.com>.
- [13] Trond Ulseth and Finn Stafsnes. VoIP speech quality - Better than PSTN? *Telektronikk 1.2006*, 2006.