QuoTe
An Extensible Platform for QoE Monitoring and Benchmarking of Telemedicine Applications

Martín Varela, Toni Mäki, Juho Merilahti
VTT Technical Research Centre of Finland
PL1100, 90571 Oulu, Finland
{martin.varela,toni.maki,juho.merilahti}@vtt.fi

Eva Rodríguez Rodríguez and Arnaud Runge
European Space Agency
ESTEC - Keplerlaan 1, 2201 AZ Noordwijk, Netherlands
{eva.rodriguez,arnaud.runge}@esa.int

***Abstract*—Telemedicine applications provide many opportunities for improving health care in a variety of conditions, in particular for people living in remote or geographically isolated areas without fast access to doctors ("medical deserts"), such as some parts of Finland. In many applications, the technologies used are actually off-the shelf solutions for videoconferencing, in some cases even used in an Over The Top (OTT) fashion over best-effort networks. In these contexts, the quality (both Quality of Experience and Quality of Service) of the whole system can have a greater significance than in other contexts such as entertainment, yet there are no quality monitoring and assessment systems specifically conceived for this purpose. In this paper we present an on-going effort to develop an extensible quality monitoring and benchmarking platform designed with video-based telemedicine applications, and the particular issues associated with them, in mind.**

I. Introduction

A. Motivation and Context

Telemedicine is a large domain, spanning multiple applications in health care, including teleconsultation, telemonitoring, telediagnosis, and educational activities [1]. These applications involve the transmission of different types of media, from simple audiovisual communications to more specialized medical imaging and instrumental data transmission. The current work focuses on teleconsultation and to some degree telediagnostic (though it is not in principle limited to those). In teleconsultation, a health care professional can attend a patient in a remote location via a video conference-type system. Telediagnostic typically involves additional medical imagery[1], and commonly a medical specialist. The use of medical imagery, or other instrumental data, poses additional challenges for transmission, as commonly-used compression algorithms are usually designed to perform best for so-called "natural scenes", and may introduce different degradations in e.g. highly-detailed medical images.

Our system is designed mostly around teleconsultation and education use cases, and those are the ones we will discuss herein, but extensibility is at the core of its design, and so it can be adapted for other uses such as telediagnostic.

The work presented in this paper has its origins in a project commissioned by the European Space Agency (ESA) in 2012 under the TRP program to develop a system for benchmarking and monitoring the Quality of Experience (QoE) of telemedicine systems. The project is planned to end before the end of 2014. At the outset, we considered four use cases related to Finnish health care, namely:

Video Consultation of a Heart Patient
: At the Heart Centre of the Tampere University Hospital, some surgeries are transmitted via video conference for both educational and consultation / telementoring purposes. Technically both cases present similar issues, and thus they are treated jointly, though the educational use case is more prevalent. In these cases, both video of the surgery and of instrumental data (often composited live with a video console installed in the operating room) is transmitted.

Remote Wound Care Consultation
: The geriatric hospital and institutional care in the Kauppi Hospital of Tampere provides care for the elderly in nursing homes and patients requiring long-term therapy or institutional care. One of the hospital's groups is focused on the remote services utilizing audiovisual technologies. Remote Wound Care Consultation (RWC) is one of the services they currently offer. In RWC, local nursing homes can agree with the consulting nurse to assess their resident's state and progress of wound care. Traditionally the remote wound care consultations are organized via digital images, uploaded to the hospital's electronic patient record system. However, the consulting nurses observed that the photo-based approach limits the assessment of the patient, and thus a videoconferencing solution was adopted.

Telestroke
: Finland sees about 11000 ischemic stroke cases per year, for which a rapid diagnosis is critical for treatment to be effective (intravenous thrombolysis is most effective when started within roughly 4 hours [2] of the stroke happening). Given the population distribution in Finland, it is not always possible for a neurologist to examine the patient *in-situ*, so in 2006 the Finnish Telestroke network was introduced to allow for remote diagnosis of possible stroke cases. The service is manned by the neurology department of the Helsinki Central University Hospital (HUCH). The

[1]This is, however, not always the case, as for example in the Telestroke application described below.

system allows the neurologist to examine the patient via a video conference link, as well as providing access to computed tomography (CT) and laboratory data. The CT images are sent over a dedicated network, separate from that used for the videoconferencing system. The assessment includes the following items:

- Level of consciousness
- Face appearance and gestures
- Eye movements and appearance
- Skin color
- Wounds and injuries
- Movements and sensory functions of the limbs
- Speech and cognition

Co-operative Care Negotiation for Detoxification Patient
: The A-Clinic Foundation operates to reduce alcohol, drug and other addiction problems by providing versatile professional services. Through its regional units, the Foundation provides treatment, detoxification and rehabilitation services in order to improve the quality of life for both people with addiction problems and their families. Järvenpää Addiction Hospital (JAH) — special unit of the A-Clinic Foundation — is the only specialized hospital for the treatment of addictions in Finland. The patients came from 73 municipalities all over the country. Co-operative care negotiations by video conference allow the nurses and social workers to assess the state of the patient before admission and mid-way through the treatment.

Three of the four Finnish use cases rely on commercial-off-the-shelf (COTS) videoconferencing systems by Cisco and Tandberg, operated by third parties, whereas the fourth (RWC) relies on a proprietary tablet PC-based integrated solution.

A fifth use case was provided by ESA, and it stems from the T4MOD project[2]. It concerns a system under development to allow medical experts to assist remote sites in a variety of medical acts, including e.g. ultrasound and neurosurgery. The system runs on satellite links, using a PC-based COTS videoconferencing system in the context of a larger e-Health system. This use case presents some additional technical restrictions due to the limited bandwidth available on the satellite link, and the use of resources reservation on the link.

B. Quality of Experience of Telemedicine Services

Quality of Experience (QoE) is currently a trending topic in multiple research communities, including signal processing, network communications, psychology and others. The topic itself has been under study for many years, but efforts to establish it as a stand-alone, multidisciplinary domain have surged in the past few years. One such effort concerns the very definition of QoE [3] which states that QoE is:

> *the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state.*

A more utilitarian definition has been previously given by the ITU [4], stating that it is

> *The overall acceptability of an application or service, as perceived subjectively by the end-user.*

The ITU definition comes accompanied by two notes, as follow:

1) Quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.)
2) Overall acceptability may be influenced by user expectations and context.

It can be seen that while both definitions have a similar goal, the first definition has a clear focus on the user, whereas the second is centered around the service itself.

In the context of telemedicine services, we have different concerns than in other domains, such as entertainment or personal communications, and therefore the definition of QoE needs to be considered in the right context. For starters, we have a specialized user — a medical doctor — who has very specific requirements from the service under consideration, mostly of an utilitarian nature, rather than a hedonic one. Secondly, we have an objective concern in terms of whether the service provides enough quality to be useful, concretely, whether the medical act under consideration can be performed properly with the quality currently provided by the service, and with what level of confidence we can say so. These things are critical, and they go beyond the user's subjective perception of quality. However, the latter is also important, as it affects the practitioner's confidence, and potentially the outcome of the medical act.

II. The QuoTe Platform

A. Goals

As mentioned in the introduction, the QuoTe platform needs to accomplish two main tasks:

- Benchmark the quality of the telemedicine systems under consideration via so-called Full-Reference (FR) quality assessments, and
- Monitor the quality of live telemedicine systems in real-time, based on network QoS metrics and knowledge about the system in question. This monitoring is done with a so-called No-Reference (NR) approach.

For the FR benchmarking, the system takes as input a media sample captured before being sent through the system, and captured during reception, thus having suffered degradations incurred by encoding[3] and transmission. In simple terms, a measure of a "perceptually relevant" distance between the two signals is performed. For the monitoring case, the use of a FR approach is not feasible, as the original signal is not available at the receiver's end, nor is it possible in most cases to capture it off the network due to the use of encryption during transmission. A parametric, NR approach is therefore used, which is based on a mapping of measurable (or

[2] http://artes-apps.esa.int/projects/t4mod

[3] Depending on the concrete configuration, encoding-related degradations may be already present during capture.

known) parameters (e.g. packet loss rate in the network, video resolution, etc.) into a perceptual quality scale. The mapping needs to correlate strongly with subjective perception in order to be valid.

Moreover, as opposed to commonly-used quality models, the ones used within the QuoTe platform need to be adapted to the concrete medical applications at hand. This also means that the system must be easily adaptable to cope with new use cases and applications as they come along in the future.

Beyond those high-level goals, a series of detailed requirements were outlined by ESA concerning the behavior of the QuoTe platform and the information it provides for both usage scenarios, related to the actual metrics and estimates output by the system, the ability to export results, be accessed remotely and be easily extensible, among others.

B. Design

The QuoTe platform is divided in two main components. One handles the FR assessment, and the other handles the NR/monitoring part.

1) Quality Benchmarking (Full-Reference Assessment): The FR component wraps around a standards-based commercial solution, namely Opticom's PEXQ[4], and it provides audio and video quality assessments (POLQA [5] and PEVQ[6], respectively), as well as audiovisual (A/V) synchronization metrics and video-specific distortion metrics (blockiness and blurriness, number of frozen frames, etc.). While currently based on PEXQ, the QuoTe platform is designed as an extensible platform, and can be easily modified to add (or substitute) the FR back-end system.

A further step (currently under implementation) is involved in the FR assessment: the values provided by the underlying FR model will be mapped into an use-case specific score. These mappings will be based on the results of subjective assessment campaigns yet to be done, where medical experts will assess the quality of the telemedicine systems under study in a number of different network conditions (cf. Section III below for more details).

2) Quality Monitoring (No-Reference Estimation): The monitoring component is architecturally more complex than the benchmarking one, as it deals with real-time quality monitoring and alerting in a distributed manner. In order to estimate the quality of the audiovisual streams in the telemedicine applications based on the network performance (and encoding parameters), the QuoTe platform must be able to measure network QoS between the end-points of the system under study. Network QoS can be measured either *actively* or *passively*. In active network measurements, specially-crafted traffic is injected into the network at one end, and after reception at the other end, statistics related to e.g. losses and delays can be extracted. In passive measurements, no extraneous traffic is injected into the network, but the existing traffic is observed instead. This poses additional challenges, since some network elements (such as routers performing Network Address Translation — NAT) or intermediate servers can modify the packets in transit, making it difficult to match the received packets with those that were observed on the sending side. In the context of use of the QuoTe platform, active measurement is undesirable, since it can significantly alter the performance of the underlying network (e.g. in the ESA-provided use case, where a low-bandwidth satellite link is used, active network measurements can be very disruptive). The QuoTe platform uses a highly sophisticated passive measurement component[5] that can trace IP flows even when they go through a NAT or in some cases even after an intermediate server modifies them to some degree.

The monitoring component is based on an distributed, agent-based architecture, developed within the CELTIC QuEEN project [7], which allows for simple integration of different types of probes[6] and QoE models, providing a way for the QuoTe platform to grow and adapt to new applications and use cases as they come along, as well as interacting with existing probes via a standard protocol (SNMP). We have augmented the QuEEN agent architecture with a rule-based engine which allows, given a drop in quality (as estimated by a NR model) and a set of QoS measurements, to infer the most likely causes of the problem (e.g. network congestion, insufficient forward error correction, etc.), and include them in an email alert to the system administrator, along with network QoS data (throughput, loss rate and distribution, delay and jitter).

Figures 1 and 2 show the results views for the Full-Reference analysis and the Monitoring components in action with test data.

The core functionalities of the QuoTe platform are exposed via a RESTful API, which allows for simple integration with other monitoring or control systems, expanding its potential uses. The UI is web-based, simplifying remote access to the QuoTe platform, and runs on top of this API.

III. The Quality Models

As discussed in the introduction, the notion of quality, in particular QoE, needs to be considered in the context of the applications under study. Traditionally, QoE has been studied mainly in the context of entertainment and personal communications (e.g. IPTV, Web-based streaming, VoIP, Videoconferencing etc.). Those types of applications have different characteristics than telemedicine ones, even if technically they might be very similar, or even the same[7]. Whereas in an application such as IPTV, even minor quality degradations of short duration can result in bad ratings, it is likely than in some of the use cases considered (e.g. the RWC case) they might be overlooked or at least tolerable, if the medical act can proceed unimpeded. Conversely, for some other telemedicine applications like robot-assisted telesurgery, even the tiniest quality degradations can have serious implications.

This tells us that generic quality models (e.g. the FR model provided by the PEXQ suite used in the QuoTe platform, or others that may be used for other cases [8]) will need to be

[4]http://www.opticom.de/products/pexq.html

[5]Cf. http://www.cnl.fi/qosmet.html for details.

[6]For the use cases in this work considered, QuoTe uses software-based network and system performance probes, but other options are possible, should other use cases have different requirements.

[7]In fact, the video conference systems used in some of the use cases considered in this work are actually the same as can be found in many companies' meeting rooms

Fig. 1. FR analysis view — reference (non-medical) test content. The analysis view enables showing the original and degraded media, as well as audio and video mean opinion score estimations, video PSNR, A/V synchronization and other overall video quality metrics such as blurriness and blockiness, skipped frames, etc.

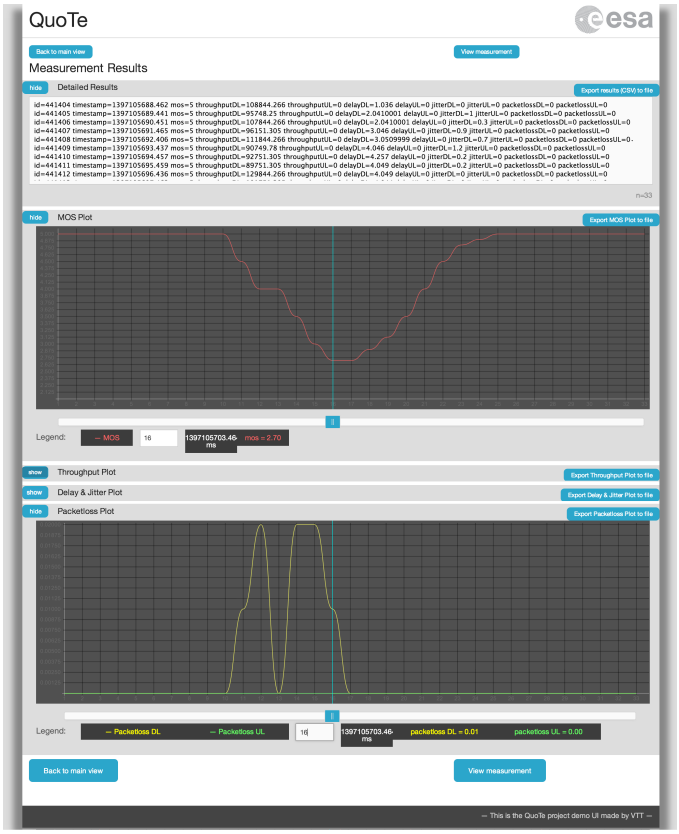

Fig. 2. Monitoring view — synthetic test data used for debugging. The monitoring view allows showing an A/V mean opinion score estimate, as well as throughput, delay, jitter and loss measurements, along with a textual representation of the measured data.

adjusted according to the application considered. Likewise, the NR models used for quality monitoring will need to be tailored to the use cases in question.

In order to create the required NR models, and the mappings needed to adjust the FR assessments to the different medical acts studied, a number of subjective quality assessments will be carried out with medical experts involved in different use cases. Normally, subjective quality assessments involve having a panel of users rate a set of media samples, and then (after statistical filtering) averaging those ratings into a Mean Opinion Score (MOS). This is a costly, time-consuming process, and even more so when the test subjects are highly-specialized professionals as in the case of telemedicine services. This is partly why good, specialized quality models are needed; subjective assessment cannot often be used for benchmarking new systems, and never for monitoring purposes. As of this writing, the test campaigns are still in the planning phase. We plan to use a slightly different approach than we would take for a more "traditional" quality assessment campaign for audiovisual media (i.e. following for example the ITU-T P.911 [9]). For the applications considered we not only care about the multimedia perceptual quality in and of itself, but also of whether it is sufficient to perform the medical act in question, and the degree of confidence with which the practitioner would be able to perform his or her task. To this end, we will use modified questionnaires with a binary "acceptable / non-acceptable" rating, and a confidence score in an absolute category rating (ACR) scale. We expect this to provide us with clear boundaries for determining whether at any given moment the system is usable to carry out the medical act considered. It should be noted, that the acceptability and confidence estimates are based on the opinions of a group of medical practitioners during subjective testing (as opposed to statistical data from real independent medical acts).

For the modeling part, we will perform a mapping from the mean opinion scores (MOS) provided by the FR assessment into the acceptability and confidence ratings obtained from the subjective assessments, in order to augment the quality ratings with this information. For the monitoring component, we will build NR models using the Pseudo-Subjective Quality Assessment [10] (PSQA) methodology. PSQA provides a general approach to developing parametric NR models for estimating media applications' QoE. It has been successfully used for creating listening quality models for VoIP [11], conversational VoIP quality [12], video quality [13], [14] and audiovisual quality for IPTV-like services [15]. The main idea behind PSQA is to establish a robust mapping between quality-affecting parameters and subjective perception. This is achieved by creating a carefully chosen set of test conditions, which is then used to create a set of degraded media samples (or to tune a test-bed if doing live assessments, as is the case

for conversational applications). After conducting a subjective assessment campaign using those conditions, a statistical estimator (commonly a Random Neural Network [16]) is trained and verified. This is then used to provide the quality estimations, which usually have high correlation with subjective perception.

IV. Conclusion

In this paper we have described QuoTe, an extensible QoE benchmarking and monitoring platform for telemedicine applications that is currently under development. The platform has several applications, such as benchmarking telemedicine systems before acquisition, auditing existing systems, or real-time monitoring of the systems' quality. The monitoring allows alerting the system administrators when problems arise and help them with diagnosing possible causes for them.

The QuoTe platform has been designed with extensibility in mind, by using a flexible, distributed agent-based architecture. This allows to "plug in" different QoE models adapted for the concrete telemedicine applications under consideration, namely:

- Video consultation of a heart patient
- Remote wound care consultation
- Satellite-based telemedicine systems for remote expert consultations
- Telestroke
- Co-operative care negotiations for detoxification patients

Besides being able to work with several quality models, the system can be integrated with industry-standard network probes and system monitors with little development effort by using standard protocols such as SNMP and HTTP. With respect to the latter, a RESTful API provides a simple way to integrate the platform with other applications (e.g. existing monitoring or control applications).

The quality models within the platform will be adapted to the concrete telemedicine applications under consideration, being based on subjective assessments carried out by specialized medical experts using real content relevant to the medical acts considered in the use cases. The proposed models will go beyond perceptual quality and include acceptability thresholds and confidence ratings in order to let users know whether the system being monitored is really usable under the current usage conditions.

At the moment, to our best knowledge, no widely accepted telemedicine QoE dataset nor algorithm exists that could be used as a reference (i.e. "gold standard"). Therefore, the generalizability of the models should be validated by independent and comparable future studies. Within-study verification of the proposed models will be done with data captured from heart surgery and wound care use cases.

Acknowledgments

M. Varela, T. Mäki and J. Merilahti's work was partly financed by an ESA commissioned project, and partly by Tekes the Finnish agency for research innovation, in the context of the CELTIC project QuEEN.

References

[1] C. Cavaro-Ménard, Z. G. Lu, and P. Le Callet, "QoE for Telemedicine: Challenges and Trends," in *Proceedings of SPIE 8856, Applications of Digital Image Processing XXXVI*, vol. 8856, 2013.

[2] G. Silva, S. Farrell, E. Shandra, A. Viswanathan, and L. Schwamm, "The Status Of Telestroke In The United States: A Survey Of Currently Active Stroke Telemedicine Programs," *Stroke*, vol. 43, Aug. 2012.

[3] P. Le Callet, S. Möller and A. Perkis, Eds., "Qualinet White Paper on Definitions of Quality of Experience (2012)," Jun. 2012, http://www.qualinet.eu/images/stories/whitepaper_v1.1_dagstuhl_output_corrected.pdf.

[4] ITU-T, " Recommendation P.10/G.100 Amendment 2 - New definitions for inclusion in Recommendation ITU-T P.10/G.100," Jul. 2008.

[5] ——, "Recommendation P.863 – Perceptual Objective Listening Quality Assessment," Jan. 2001.

[6] ——, "Recommendation J.247 – Objective Perceptual Multimedia Video Quality Measurement In The Presence Of A Full Reference," Aug. 2008.

[7] F. Guyard, M. Varela, L. Skorin-Kapov, A. Cuadra, and P. Sevilla, "Quality of Experience Estimators in Networks," in *Quality of Experience Engineering for Customer Added Value Services: From Evaluation to Monitoring*, A. Mellouk and A. Cuadra, Eds. Iste / Wiley & Sons (in press), 2014.

[8] A. Raake, J. Gustafsson, S. Argyropoulos, M. Garcia, D. Lindegren, G. Heikkila, M. Pettersson, P. List, and B. Feiten, "IP-Based mobile and fixed network audiovisual media services," *IEEE Signal Processing Magazine*, vol. 28, pp. 68–79, Nov. 2011.

[9] ITU-T, "Recommendation P.911 – Subjective Audiovisual Quality Assessment Methods for Multimedia Applications," Dec. 1998.

[10] M. Varela, "Pseudo-Subjective Quality Assessment of Multimedia Streams and its Applications in Control," Ph.D. dissertation, INRIA/IRISA, univ. Rennes I, Rennes, France, Nov. 2005.

[11] S. Mohamed, G. Rubino, and M. Varela, "Performance Evaluation of Real-time Speech Through a Packet Network: a Random Neural Networks-Based approach," *Performance Evaluation*, vol. 57, no. 2, pp. 141–162, May 2004.

[12] A. C. da Silva, M. Varela, E. de Souza e Silva, R. Leão, and G. Rubino, "Quality assessment of interactive real time voice applications," *Computer Networks*, vol. 52, p. 11791192, Apr. 2008.

[13] D. D. Vera, P. Rodríguez-Bocca, and G. Rubino, "Automatic quality of experience measuring on video delivering networks," *SIGMETRICS Performance Evaluation Review*, vol. 36, no. 2, pp. 79–82, 2008.

[14] A. C. da Silva, P. Rodriguez-Bocca, and G. Rubino, "Optimal Quality-of-Experience design for a P2P Multi-Source video streaming," in *Communications, 2008. ICC '08. IEEE International Conference on*, May 2008, pp. 22 –26.

[15] T. Mäki, D. Kukolj, D. Dordević, and M. Varela, "A Reduced-Reference Parametric Model for Audiovisual Quality of IPTV Services," in *Proceedings of QoMEX 2013*, Klagenfurt, Austria, Jul. 2013.

[16] E. Gelenbe and J. Fourneau, "Random neural networks with multiple classes of signals," *Neural Computation*, vol. 11, no. 3, pp. 953–963, 1999.